Delft University of Technology

Rapid Calibration of Raman Spectroscopy Models for Bioreactor Monitoring

Klaverdijk, M.

DOI
10.4233/uuid:733d19d8-ec47-46ed-880a-ec09751b708b

Publication date
2025

Document Version
Final published version

Citation (APA)
Klaverdijk, M. (2025). *Rapid Calibration of Raman Spectroscopy Models for Bioreactor Monitoring*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:733d19d8-ec47-46ed-880a-ec09751b708b

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

RAPID CALIBRATION OF RAMAN SPECTROSCOPY MODELS FOR BIOREACTOR MONITORING

MAARTEN KLAVERDIJK

Rapid Calibration of Raman Spectroscopy Models for Bioreactor Monitoring

Rapid Calibration of Raman Spectroscopy Models for Bioreactor Monitoring

Dissertation

For the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates,
to be defended publicly on
Friday 28th of November 2025 at 10:00 o'clock

by

Maarten Klaverdijk

Master of Science in Biotechnology,
Wageningen University & Research, the Netherlands

born in Arnhem, the Netherlands

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	Chair person
Prof. dr. ir. M. Ottens	Delft University of Technology, promotor
Dr. M.E. Klijn	Delft University of Technology, copromotor

Independent members:

Prof. dr. M.J. Barbosa	Wageningen University & Research
Prof. dr. ing. M. H. M. Eppink	Delft University of Technology
Prof. dr. K. V. Gernaey	Danmarks Tekniske Universitet
Prof. Dr. F. Hollman	Delft University of Technology
Dr. B. Hunyadi	Delft University of Technology

The research described in this thesis was performed at and funded by the Department of Biotechnology, Delft University of Technology, The Netherlands.

Printed by Proefschriftspecialist

Cover design: Maarten Klaverdijk

Copyright © 2025 by Maarten Klaverdijk

ISBN: 978-94-6384-879-4

An electronic version of this dissertation is available at https://repository.tudelft.nl/

Table of Contents

Summary **7**

Samenvatting **11**

1 Introduction **15**

1.1 Bioreactor monitoring 17
1.2 Raman spectroscopy 20
1.3 Pre-processing and Partial Least Squares models 23
1.4 Motivation and aim of the thesis 29
1.5 Thesis outline 30

2 Single compound data supplementation to enhance transferability of fermentation specific Raman spectroscopy models **37**

2.1 Introduction 39
2.2 Materials & Methods 42
2.2.1 Fermentation 42
2.2.2 Single compound solutions 43
2.2.3 Reference data 44
2.2.4 Raman spectroscopy 44
2.3 Results and Discussion 46
2.3.1 Quantitative analysis of base model performance for batch data 46
2.3.2 Base model performance on fed-batch data 47
2.3.3 Qualitative assessment of base model 48
2.3.4 Impact of single compound spectra data supplementation 51
2.4 Conclusion 56
2.5 Supplementary 60
2.5.1 Fermentation time evolution graphs 60
2.5.2 Overview of pure component measurement concentrations 60
2.5.3 Influence of acquisition time on model accuracy 61
2.5.4 RMSEC vs RMSECV plots for base and supplemented models 61
2.5.5 Cross-correlations between process compounds in the base and supplemented datasets 62
2.5.6 Performance of supplemented models on batch validation dataset 63

3 Impact of bioreactor process parameters and yeast biomass on Raman spectra **65**

3.1 Introduction 67
3.2 Materials & Methods 71
3.2.1 Measurement setup 71
3.2.2 Bioreactor parameter experiments 71
3.2.3 Fermentations and biomass measurements 72
3.2.4 Data analysis 73
3.3 Results and discussion 74
3.3.1 Spectral effects of temperature 75

3.3.2 Spectral effects of bubbles 78
3.3.3 Effects of viscosity 81
3.3.4 Effects of yeast biomass 82
3.4 Conclusion 87
3.5 Supplementary 92
3.5.1 Temperature effects 92
1.1.1. Bubble effects 92
3.5.2 Viscosity effects 94
3.5.3 Biomass effects 95
4 Towards rapid calibration of bioprocess quantification models using single compound Raman spectra: a comparison of four approaches 97
4.1 Introduction 99
4.2 Materials & Methods 102
4.2.1 Experimental methods 102
4.2.2 Computational methods 103
4.3 Results and Discussion 106
4.3.1 Approach 1: PLS model calibration with single compound spectra 108
4.3.2 Approach 2: Indirect Hard Model (IHM) calibrated with single compound spectra 111
4.3.3 Approach 3: PLS model calibration with synthetic spectra 112
4.3.4 Approach 4: PLS model calibration with augmented process spectra 114
4.3.5 Discussion on modelling approaches 116
4.4 Conclusion 119
4.5 Supplementary 123
4.5.1 Approach 1: PLS model calibration with single compound spectra 123
4.5.2 Approach 2: Indirect Hard Model (IHM) calibrated with single compound spectra 125
4.5.3 Approach 3: PLS model calibration with synthetic spectra 126
4.5.4 Approach 4: process spectra augmentation 127
4.5.5 Discussion on modelling approaches 128
5 Conclusions & Outlook 133
5.1 Conclusions 135
5.2 Outlook 137
5.2.1 Leveraging rapid model development 137
5.2.2 Raman miniaturization for accelerated implementation 141
5.2.3 Democratization of Raman spectroscopy 145
Acknowledgements 149
Curriculum Vitae 161
List of Publications 162

Summary

Bioprocess engineering involves the controlled cultivation of cells for the production of specialized products. These cells function as living factories, and their environment must be carefully controlled to optimize metabolic activity and productivity. Cell cultures are operated in bioreactor systems which aim to maintain optimal environmental conditions and provide optimal mass transfer. Monitoring bioreactor conditions such as nutrient levels or cell growth is typically dependent on manual sampling. This involves an operator extracting a small sample and analysing it on an external device, which provides a delayed and partial view of the process. To overcome the challenges of manual sampling, the bioprocessing industry is adopting Process Analytical Technology (PAT) tools like Raman spectroscopy, which can monitor the molecular composition of the system in real-time. However, accurate monitoring by Raman spectroscopy is dependent on chemometric models that typically require extensive calibration with process data, leading to process-specific models which do not transfer to related processes. This often necessitates repeating the extensive collection of process data for each new process that must be monitored. Therefore, this thesis focuses on investigating alternative approaches to calibration data collection and chemometric model calibration, while studying how varying measurement conditions affect spectral integrity.

The work in **Chapter 2** addresses common challenges in calibrating Partial Least Squares (PLS) models for bioreactor processes, focusing on glucose, ethanol, and biomass quantification during *Saccharomyces cerevisiae* batch fermentation. While the models calibrated with batch fermentation data performed well under the same batch conditions, prediction accuracy declined significantly when applied to a related fed-batch process. This study shows how standard model performance metrics do not reflect model specificity and highlights how cross-correlations between analytes hinder model transferability. To address this, the chapter emphasizes the importance of qualitative model evaluation to assess compound specificity. To overcome the transferability issues, the original calibration datasets were supplemented with single compound spectra to improve both specificity and transferability of the models. This approach improved both model specificity and robustness to changes in operating conditions, offering a simple and effective alternative to collecting new fermentation process data.

Chapter 3 investigates how varying measurement conditions in bioreactor systems affect Raman spectra, while focusing on scattering effects caused by particles such

as bubbles and *S. cerevisiae* cells. Spectral pre-processing is essential to correct for spectral distortion during chemometric model development, and a deeper understanding of the sources of distortion allow for effective pre-processing strategies. The study evaluated the impact of changing temperature, bubble quantity, and viscosity on the position and intensity of Raman peaks, and investigated their effects on the spectral baseline. Increasing the medium temperature shifted peaks to lower wavenumbers, while higher bubble concentrations led to the attenuation of Raman peaks. A biomass concentration of 5 g/L led to peak extinctions of up to 44.6%, demonstrating how the spectral contribution of yeast is dominated by scattering effects. Despite the strong signal extinction, spectral features related to cell proteins and lipids were detectable after spectral baseline correction and normalization. This work highlights the importance of understanding sources of spectral distortion when developing robust chemometric models, and emphasizes the importance of spectral alignment when combining data from different experimental setups.

The work in **Chapter 4** addresses the challenge of requiring process data for model calibration by exploring alternative calibration approaches using single compound measurements. By utilizing a small dataset of 16 single compound spectra quantification models were calibrated through four different approaches. Direct calibration via PLS and Indirect Hard Modelling (IHM) led to accurate prediction models for fermentation processes, demonstrating how targeted measurements can lead to robust models without the need for process data. The IHM approach generated compound hard models, enabling the generation of de novo synthetic spectra simulating custom fermentation conditions, allowing the calibration of highly specific PLS models. As a final step, the compound hard models were used to artificially adjust the concentration of glucose and ethanol in a small dataset of process spectra, offering a data augmentation approach to expand limited datasets. This work highlights how computational approaches can generate or augment spectral calibration data, by which small datasets can be leveraged to calibrate robust quantification models.

In summary, this thesis demonstrates alternative approaches to calibration data collection utilizing simple measurements that can be used to directly calibrate or improve chemometric quantification models for Raman spectroscopy. In addition, the characterization of spectral distortion and yeast biomass spectra improves signal understanding leading to targeted pre-processing strategies. The knowledge from this thesis contributes to the rapid implementation of Raman spectroscopy into new bioprocesses, and provides stepping stones to flexible model calibration approaches.

Samenvatting

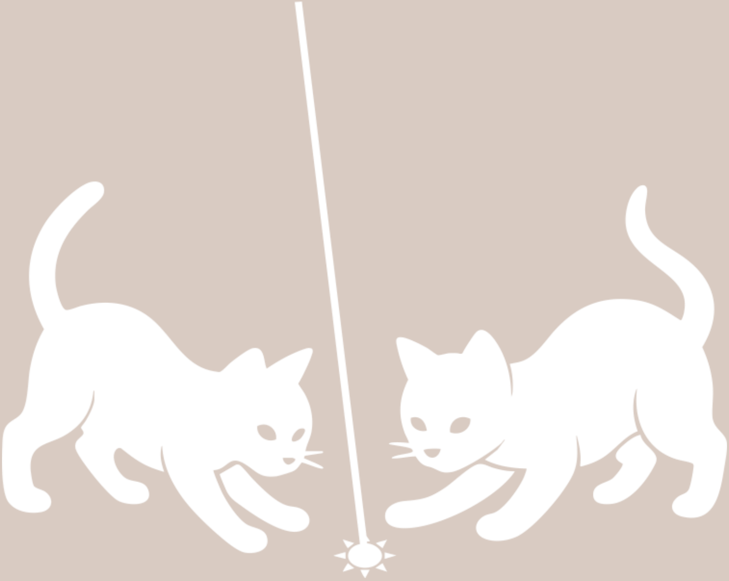
Biotechnologie betreft de gecontroleerde kweek van cellen voor de productie van gespecialiseerde producten. Deze cellen functioneren als levende fabrieken, waarvan de omgeving nauwkeurig moet worden gecontroleerd om metabolische activiteit en productiviteit te optimaliseren. Celkweekprocessen vinden daarom plaats in bioreactoren, die zijn ontworpen om de ideale omstandigheden te bieden en een efficiënte massaoverdracht te garanderen. Het monitoren van de omstandigheden in deze bioreactoren, zoals nutriëntconcentraties of celgroei, is vaak afhankelijk van handmatige monsterafname. Dit betekent dat een laborant een monster neemt en dit meet op een extern apparaat, wat een vertraagd en gedeeltelijk beeld van het proces oplevert. Om de nadelen van handmatige monstername te elimineren onderzoekt de biotechnologische industrie het gebruik van Process Analytical Technology (PAT) tools, zoals Raman spectroscopie, die in staat zijn om de moleculaire samenstelling van de inhoud van de bioreactor in real-time te meten. Nauwkeurige metingen met Raman spectroscopie zijn echter afhankelijk van chemometrische modellen die de complexe spectrale data vertalen naar bruikbare meetwaardes. Deze modellen worden doorgaans gekalibreerd met uitgebreide datasets gemeten in het proces zelf, wat resulteert in proces-specifieke modellen die slecht overdraagbaar zijn naar gerelateerde processen. Dit heeft vaak als gevolg dat voor elk nieuw bioreactorproces opnieuw kalibratiedata verzameld moet worden van het proces zelf, wat een grote hindernis vormt voor de implementatie van Raman spectroscopie. Daarom richt deze thesis zich op het onderzoeken van alternatieve methodes voor het verzamelen van kalibratiedata en chemometrische modelkalibratie, terwijl ook de invloed van variërende meetomstandigheden op de spectrale integriteit wordt onderzocht.

Het werk in **Hoofdstuk 2** behandelt de veelvoorkomende uitdagingen bij het kalibreren van Partial Least Squares (PLS) modellen voor het meten van bioreactor processen, met de focus op glucose, ethanol, en biomassa kwantificatie tijdens een *Saccharomyces cerevisiae* batch-fermentatie. Hoewel modellen die gekalibreerd zijn met batch fermentatiedata goed presteren wanneer ze worden toegepast onder dezelfde omstandigheden, nam de kwantificatie-nauwkeurigheid aanzienlijk af toen de modellen werden toegepast op een gerelateerd fed-batch proces. Dit werk toont aan hoe standaard prestatieparameters de specificiteit van modellen niet weerspiegelen en benadrukt hoe correlaties tussen componenten de modeloverdraagbaarheid belemmeren. Om dit aan te pakken, legt het hoofdstuk de nadruk op het belang van kwalitatieve modelbeoordeling om de specifiteit te evalueren. Om de overdraagbaarheidsproblemen aan te pakken, werden de oorspronkelijke kalibratiedataset gecomplementeerd met spectra van enkele componenten om zowel de specificiteit als de overdraagbaarheid van de modellen te verbeteren. Deze methode verbeterde zowel de specificiteit van de modellen als de robuustheid ten opzichte van veranderingen in de procescondities, en bood een eenvoudig alternatief tegen het verzamelen van nieuwe fermentatiegegevens.

In **Hoofdstuk 3** wordt onderzocht hoe variërende meetomstandigheden in bioreactoren Raman spectra beïnvloeden, met de nadruk op lichtverstrooiingseffecten veroorzaakt door deeltjes zoals bellen en *S. cerevisiae* cellen. Spectrale pre-processing is essentieel om spectrale vervorming tijdens de ontwikkeling van chemometrische modellen te corrigeren. Meer kennis over de bronnen van deze vervormingen maakt het mogelijk om effectievere pre-processing toe te passen. De studie onderzocht hoe temperatuurveranderingen, het aantal bellen en de viscositeit invloed hadden op de positie en intensiteit van Raman-pieken, evenals op de spectrale baseline. Een verhoging van de mediumtemperatuur verschoof de Raman-pieken naar langere golflengtes, terwijl een hogere bellenconcentratie leidde tot een afname in piekintensiteit. Een biomassa concentratie van 5 g/L veroorzaakte afnames in piekintensiteit tot wel 44.6%, wat aantoont dat de voornaamste spectrale bijdrage van gist bestaat uit lichtverstrooiingseffecten. Ondanks de sterke signaalextinctie waren er spectrale kenmerken van cel proteïnen en lipiden detecteerbaar na baseline correctie en normalisatie. Dit werk benadrukt hoe cruciaal het is om de bronnen van spectrale vervorming te begrijpen bij de ontwikkeling van robuuste chemometrische modellen, en onderstreept daarnaast de noodzaak van nauwkeurige pre-processing bij het combineren van spectra uit verschillende experimentele opstellingen.

Het werk in **Hoofdstuk 4** richt zich op de afhankelijkheid van procesdata tijdens modelkalibratie, en verkent alternatieve kalibratiemethodes op basis van simpele metingen. Op basis van een kleine dataset van 16 pure spectra werden kwantificatiemodellen gekalibreerd met vier verschillende methoden. Directe kalibratie via PLS en Indirect Hard Modelling (IHM) resulteerde in nauwkeurige kwantificatiemodellen voor fermentatieprocessen, wat aantoont dat gerichte metingen robuuste modellen kunnen opleveren zonder dat procesdata nodig is. De IHM-methode genereerde compound hard-modellen, waarmee de novo synthetische spectra konden worden gemaakt. Deze methode stelde ons in staat om op maat gemaakte fermentatiecondities te simuleren, wat resulteerde in een gemakkelijke kalibratie van zeer specifieke PLS-modellen. Als laatste stap werden de compound hard-modellen gebruikt om de concentratie van glucose en ethanol kunstmatig aan te passen in een kleine dataset van processpectra, wat een manier bood om beperkte datasets te verrijken en uit te breiden. Dit werk benadrukt hoe computationele methodes spectrale kalibratiedata kunnen genereren of verrijken, waardoor kleine datasets kunnen worden verbeterd voor de kalibratie van robuuste modellen.

Deze thesis biedt een overzicht van alternatieve methodes voor het verzamelen van kalibratiegegevens, waarbij eenvoudige metingen worden gebruikt voor de kalibratie van chemometrische kwantificatiemodellen voor Raman spectroscopie. Daarnaast leidt meer kennis over de oorzaken van spectrale vervorming en de spectrale contributie van gist biomassa tot efficiëntere en gerichte pre-processing strategieën. De kennis uit deze thesis draagt bij aan de snelle implementatie van Raman spectroscopie in nieuwe bioprocessen, en biedt aanknopingspunten voor flexibele model kalibratiemethodes.



Chapter 1

Introduction

1.1 Bioreactor monitoring

Fundamental to bioprocess engineering is controlled cell cultivation with the objective to generate a product that consistently meets a set of quality attributes. In most bioprocesses cells are used as living factories, and their environment dictates growth and productivity. This means the environment needs to be tightly regulated to achieve optimal cell metabolism and product formation, despite the inherent variability of biological systems [1]. The bioreactor is designed to maintain optimal environmental parameters as well as maximize mass transfer and mixing to ensure homogeneous nutrient availability while preventing shear damage to the cells. Moreover, the objective of bioreactor process development is to define optimal conditions that enable consistent and reproducible production over a wide range of volumes to enable large scale manufacturing.

Several environmental process parameters are directly measured in state-of-the-art bioreactor systems through the use of immersion probes. This includes liquid-based measurements such as medium temperature, pH, and dissolved oxygen (DO), thereby providing continuous information on the state of the system [2]. Fluctuations in these parameters can be detrimental to process performance and control systems are therefore used to automatically maintain a desired setpoint. Next to liquid-based measurements, off-gas analyzers allow for continuous monitoring of the consumption of O_2 and the production of CO_2 by the cell culture. The combination of liquid-based and off-gas analytical tools allow for a basic level assessment of the state of the bioreactor. When information is required on key process variables or product quality attributes, such as nutrient concentrations, product titres, or cell density and viability, manual sampling is still necessary. These samples are extracted from the bioreactor and subsequently analyzed off-line on external analytical machines. This form of discrete process monitoring leads to a delayed and limited view of the bioreactor process, thereby hindering the implementation of automated control loops to act on the obtained information, and is prone to measurement variations caused by sample handling and preparation steps. Moreover, the low sampling frequency and monitoring resolution may cause operators to miss critical process phases, resulting in manufacturing mistakes [3].

To improve bioprocess development and control, adequate real-time information on key process parameters is required. The implementation of process analytical technology (PAT) tools that extract this information during bioreactor operation can advance process monitoring and understanding. The Food and Drug Administration (FDA) published a series of guidelines on PAT, which is defined 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" [4]. These guidelines are intended to direct industry away from delayed in-process and post-production quality control, and encourage the adoption of modern analytical tools for process understanding, control, and quality assurance. This incentivized the implementation of advanced PAT tools capable of providing real-time and accurate information on critical process parameters, thus supporting the transition towards data-driven manufacturing approaches.

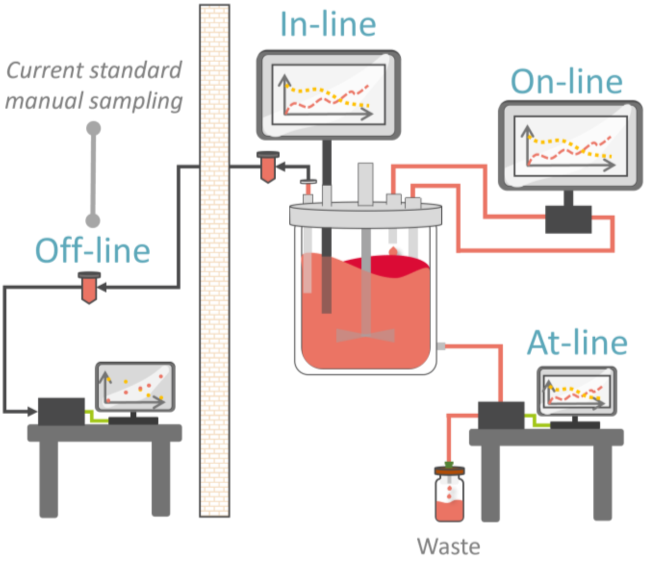


Figure 1.1: Schematic of possible measurement modes. In-line: Sensors or probes are inserted directly into the process, in-situ analysis. On-line: The analyzer is directly connected to the process in a closed loop and sample returns to the process after analysis. At-line: The analyzer is close to the process but disconnected, automated sampling is required. Off-line: The analyzer is placed remotely from the process, manual sampling and sample preparation is often required

Where manual sampling and external sample analysis can be categorized as off-line monitoring, PAT can be configured in an in-line, on-line, and at-line setup (Figure 1.1) [5]. In-line sensors measure directly into the system and circumvent the need for sampling, sample preparation, and complex sampling loops [6]. This enables straightforward measurements of a sample most representative to the process but reduces the ability to control measurement conditions. On-line configurations use a sample loop that extracts a sample from the system automatically, analyses the sample in an external device, and returns it to the process [7]. While this setup offers more control over the measurement conditions, it adds complexity to the measurement setup. In addition, extracting a sample and measuring it in an external loop may not accurately reflect the conditions of the bioreactor itself. During at-line analysis, an auto-sampler system extracts samples from the bioreactor and delivers it to external analytical devices for analysis [8]. After the measurement, the sample is not returned to the system. This approach essentially automates off-line sample analysis, making it compatible with high-precision equipment, but can introduce relatively long delays in sample analysis. Ideally, PAT devices for real-time monitoring of bioreactors should be fast, robust, and utilize non-destructive mechanisms, requiring no sample preparation and generate multi-analyte data without the consumption of these analytes [9]. Of the many analytical techniques available, automated spectroscopy techniques fit these requirements well as their optical nature allows for continuous, and non-destructive analysis of sample material within seconds.

Spectral PAT provides comprehensive information on the biochemical composition of the process, typically supporting multiplexed compound measurements, and it can be implemented in-line through immersion probes or glass windows, or at-line in a combination with a of flow-cell [10-12]. For this reason a steep rise in popularity is observed over the last decade for techniques such as ultraviolet-visible (UV-Vis), fluorescence, near-infrared (NIR), and Raman spectroscopy [13, 14]. For example, UV-Vis spectroscopy was used to quantify cell density and viability based on absorbance characteristics [15], while fluorescence spectroscopy can offer increased specificity to protein structures [16]. NIR spectroscopy can quantify a wide range of process compounds and parameters, ranging from metabolites and amino acids to complex proteins and cell density [17-19]. However, NIR spectroscopy measures molecular overtones and combinations of vibrational modes, and water has strong absorption in the NIR region because of its large dipole moment. This can interfere with measurements in aqueous systems, as seen in bioreactors [20]. Raman spectroscopy measures inelastic light scattering and is suitable for aqueous environments as water has a low polarizability and produces weak Raman scattering (see section 1.2). The reduced interference of spectral features of water combined with commercialization of autoclavable immersion probes and flow-cell assemblies have supported the integration of Raman spectroscopy in a wide range bioprocessing applications [13]. Initial applications of in-line Raman spectroscopy for bioreactor monitoring focused on the real-time quantification of substrates, products, cell density and cell viability [21, 22]. This reduced the dependency on manual sampling and improved process understanding through continuous monitoring. Real-time quantification of process compounds could drive automated dynamic feeding strategies, maintaining concentration setpoints of substrates such as glucose and amino acids [23, 24]. Recent studies have demonstrated the successful use of in-line Raman spectroscopy to monitor complex product critical quality attributes, including real-time monitoring of antibody glycosylation in Chinese Hamster Ovary cell processes, potentially substituting complex and time-consuming off-line analysis steps [25, 26]. These studies highlight the necessity and applicability of Raman spectroscopy as a real-time monitoring tool for a wide range of key process parameters and show that the continuous real-time signal can form the basis for advanced process control strategies.

1.2 Raman spectroscopy

Raman spectroscopy is based on the measurement of inelastic light scattering discovered by Nobel Prize winner C.V. Raman in 1928. When incident light passed through a transparent material, a small fraction was scattered in different directions and at different wavelengths from the incident beam, indicating an exchange of energy between the light source and the vibrational modes of molecules in the sample [27]. The strength of Raman scattering is determined by how much a molecule's electron cloud can be distorted by an external electric field, called the polarizability. Vibrational modes with high polarizability provide strong Raman signals, while molecules with tightly bound electrons scatter weakly, as is the case with water. In modern day spectroscopes, stable monochromatic lasers are used as incident light, and the inelastic scattering is measured with Charge Coupled Devices (CCD). During the measurements, molecules in the sample interact with the laser light, where they are excited to a virtual vibrational state (Figure 1.2A). This virtual state persists for less than a picosecond, after which most molecules fall back to their original vibrational state (elastic Rayleigh scattering) [28]. A small fraction of the interactions (about 1 in 10^9-10^{10}) will lead to inelastic scattering by either losing or gaining energy, resulting in Raman scattering events. These inelastic events can be subdivided into Stokes and Anti-Stokes scattering. During Stokes scattering a molecule is raised from the ground to an excited vibrational state, thus leading to an increase in wavelength of the scattered photon. During Anti-Stokes scattering a molecule that was in an elevated vibrational state will fall back to a lower state, by which the scattered photon leaves the interaction with a shorter wavelength. Stokes scattering is monitored for biological conditions and temperature ranges as most molecules are in their ground vibrational state, leading to a stronger Stokes signal. For both Stokes and Anti-Stokes scattering, the difference between the excitation and inelastically scattered photons provides information on the structure of the molecule, leading to a specific fingerprint signal (Figure 1.2B). The resulting Raman spectrum or molecular fingerprint arises from the accumulation of individual scattering events, each concerning different vibrational modes of a molecule.

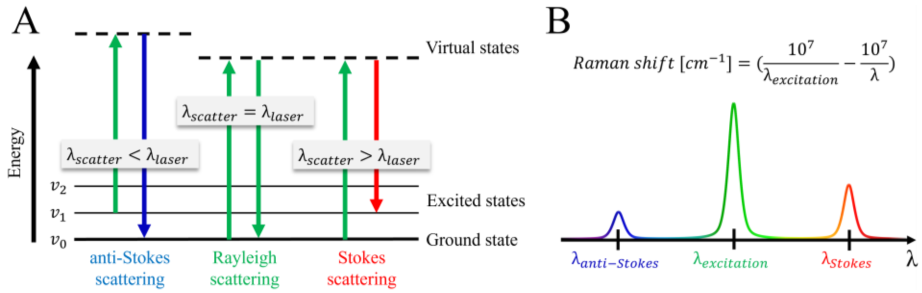

A
Energy
Virtual states
$\lambda_{scatter} = \lambda_{laser}$
$\lambda_{scatter} < \lambda_{laser}$
$\lambda_{scatter} > \lambda_{laser}$
v_2
v_1
v_0
Excited states
Ground state
anti-Stokes scattering
Rayleigh scattering
Stokes scattering
B
$Raman\ shift\ [cm^{-1}] = (\frac{10^7}{\lambda_{excitation}} - \frac{10^7}{\lambda})$
$\lambda_{anti-Stokes}$
$\lambda_{excitation}$
λ_{Stokes}
λ

Figure 1.2: The mechanism of Raman scattering (A) and a depiction of the Raman shift (B). The laser light interacts with the sample molecules and in most cases leads to elastic Rayleigh scattering (green). During Raman scattering energy is exchanged with the molecule, either absorbing (Stokes, red) or releasing (anti-Stokes, blue) energy depending on the molecule's vibrational state. Raman spectra are typically plotted in wavenumbers [cm^{-1}] representing the energy difference between scattered photons and the excitation wavelength.

Measured Raman spectra are typically shown with the frequency on the x-axis and the measured intensity at each frequency on the y-axis (Figure 1.3). The frequency is displayed in Raman shift (cm^{-1}), which provides the energy difference between the excitation light source and the inelastically scattered light. This unit allows for the direct comparison of measurements acquired with different laser wavelengths as the energy differences are maintained. When measuring Raman spectra in aqueous conditions, the spectral baseline mainly consists of contributions from water vibrational modes, resulting in high tails in the low (<800 cm^{-1}) and high (>3000 cm^{-1}) wavenumber regions (Region I and Region IV in Figure 1.3). Most biochemical information is located in the fingerprint region of approximately 500-1800 cm^{-1}, containing highly specific fingerprints of molecules (Regio II in Figure 1.3). The 1800-2800 cm^{-1} region contains little to no biochemical information as organic molecules do not display peaks in this region, and it is often called the silent region (Region III in Figure 1.3). The 2800-3000 region contains strong CH-stretching vibrations, while the rest of the high wavenumber region (>3000 cm^{-1}) mainly displays strong water vibrational modes. Although water produces strong spectral features in the low and high wavenumber regions, it causes little interference with key spectral features of other molecules in the fingerprint region.

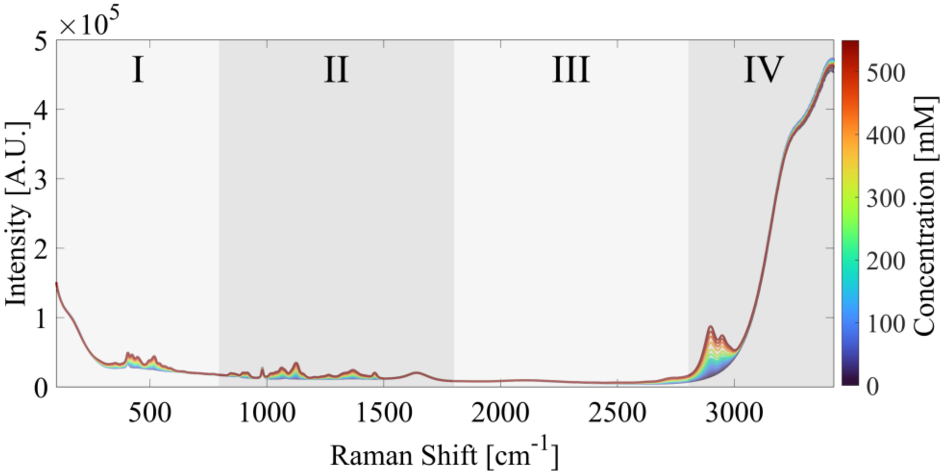


Figure 1.3: Raman spectra of a glucose concentration range from 0 to 550 mM (colour bar). The spectrum is divided into the low wavenumber region (I), fingerprint region (II), biological silent region (III), and the high wavenumber region (IV).

For process monitoring applications, a Raman spectroscope is typically operated in a continuous acquisition mode. The exposure time for each individual spectrum has to be carefully optimized to achieve an appropriate CCD saturation. If the exposure time is too short, the measurement can be dominated by noise inherent to the detector, leading to a poor signal-to-noise ratio. High exposure times can potentially lead to oversaturation of the CCD by which information is lost. Common sources of spectral noise include shot noise and thermal noise. Shot noise is caused by the statistical fluctuations in the number of photons reaching the detector, with short exposure times typically resulting in higher noise levels. When single exposures are not free of shot noise despite a good CCD saturation, multiple exposures can be averaged to reduce signal noise [29, 30]. Thermal noise originates from thermally excited electrons in the CCD that mimic incoming photons, thereby leading to noisy measurements. The level of thermal noise greatly reduces at lower temperatures and becomes relatively constant once the CCD has reached a stable temperature through cooling [31]. Thermal noise can be captured by acquiring a dark measurement with the laser blocked, allowing for its subtraction from sample spectra to improve signal accuracy.

In addition to spectral noise, two major challenges for measuring Raman scattering in biological processes are (1) the intrinsic weakness of the Raman signal and (2) the risk of fluorescence overwhelming this weak signal. The Raman signal strength can be increased by using a more powerful laser, but this comes at the cost of a higher risk to sample fluorescence. Higher power lasers can excite molecules beyond the virtual state into electronic states and the subsequent relaxation generates fluorescence much stronger than Raman scattering [32]. Raman spectroscopes can be equipped with varying laser wavelengths to optimize the ratio between Raman signal strength and risk of fluorescence, and commonly used wavelengths are 532 nm, 785 nm, and 1064 nm [33, 34]. A near-infrared 785 nm laser is widely used for monitoring bioreactor processes to provide a good balance between reducing fluorescence and maintaining sufficient Raman signal strength. In addition to appropriate laser wavelength selection, other methods have been developed to prevent signal fluorescence. While the relaxation of a molecule from the virtual vibrational state is almost instantaneous ($<10^{-11}$ s), relaxation from an excited electronic state can take significantly longer (10^{-9} to 10^{-7} s) [32]. This temporal difference is leveraged through time-resolved techniques, where pulsed lasers and time-gated detectors capture scattering within the specific timeframe before fluorescence occurs [35, 36]. However, time-resolved techniques lead to more complex spectroscopic configurations, and in most cases moderate background fluorescence can be computationally removed from the measured spectra.

1.3 Pre-processing and Partial Least Squares models

Raman spectra of single compounds yield a fingerprint of peaks that display a linear correlation between signal intensity and compound concentration, allowing for quantitative measurements in real-time. However, when the number of molecular species in the sample increases, the enhanced system complexity causes spectral features to overlap. This results in high-dimensional data that prevents straightforward quantitative analysis. Next to the overlapping spectral features of molecular compounds in the sample, the presence of particles can cause spectral scattering effects, for example due to bubbles and cells in a bioreactor [37, 38]. In addition, fluorescent compounds produced during cell cultures can cause strong background fluorescence leading to baseline shifts [32].

The spectral features of process compounds typically of interest for quantification can be distorted by scattering and baseline effects, and removal of this distortion improves quantification accuracy in subsequent data analysis and modelling steps. Spectral pre-processing is therefore an essential part of applying Raman spectroscopy, as it emphasizes variation linked to the identity and abundance of compounds of interest [39]. One pre-processing approach is to reduce spectral information through region selection. Depending on the equipment type and resolution, Raman spectra usually consist of roughly 3000 spectral variables where the low, high, and silent wavenumber regions contain little biochemical information.

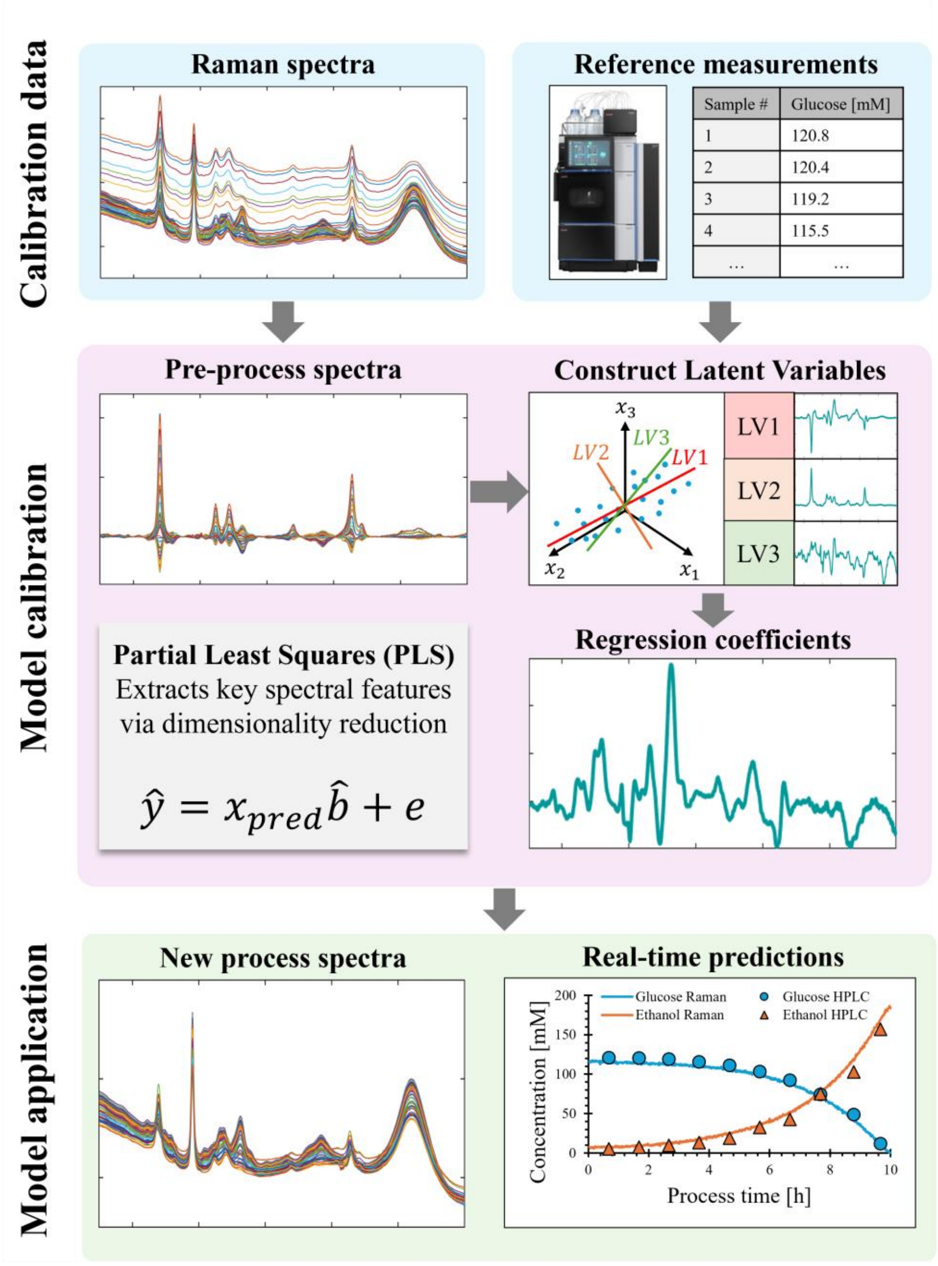
Spectra are often truncated to exclude these redundant regions and focus on the fingerprint region (500-1800 cm^{-1}), sometimes also including the CH-stretching region at 2800-3000 cm^{-1}. To deal with scattering and fluorescence effects that may disrupt the linearity between molecule abundance and signal intensity, a wide range of baseline corrections, scatter corrections, and normalization steps can be applied [40]. A typical order of Raman spectrum pre-processing consists of (optional) region selection, baseline correction, scatter correction, normalization, and mean-centering or scaling [39, 41]. The optimal pre-processing strategy varies per application of Raman spectroscopy, as each process and measurement setup is subject to different types of spectral distortions.

Besides dealing with spectral distortion, the high dimensionality of spectral data complicates the direct interpretation of overlapping features arising from multiple compounds. For this reason, Raman spectroscopy requires chemometric modelling, which is the application of statistical and machine learning techniques to extract meaningful information from complex chemical data. Data-driven chemometric models are commonly used for this. Here, models are calibrated (i.e., trained) with spectral data representative of the target process and reference measurements on the compound of interest.

Raman spectra contain many spectral variables (wavenumbers) with high collinearity and the number of response parameters is often limited or singular (e.g., concentration of a single compound). Partial Least Squares (PLS) regression is a multivariate regression method designed to handle datasets containing highly collinear predictor variables and one or more response variables, making it highly suitable for spectral data [42]. The input to PLS models is a pre-processed predictor data matrix X (e.g., Raman spectra) and the output is a response matrix Y (e.g., chemical compound concentrations). The main objective of PLS is to extract latent variables (LVs) that most accurately predict Y-data using X-data as input. These LVs are constructed as linear combinations of the original predictor variables and are designed to capture the variation in X that is most relevant for explaining Y. In mathematical terms, PLS decomposes the predictor and response matrices as shown in Equation 1.1 and Equation 1.2.

$$X = TP^T + E \quad (1.1)$$

$$Y = UQ^T + F \quad (1.2)$$



Sample #	Glucose [mM]
1	120.8
2	120.4
3	119.2
4	115.5
…	…

Figure 1.4: Workflow for spectral pre-processing and calibration of chemometric quantification models through Partial Least Squares (PLS) regression. A calibration dataset containing measured spectra and reference concentration measurements is pre-processed and used to construct Latent Variables (LV) that maximize covariance between the predictor (X) and response (Y) data. The resulting PLS model applies the same pre-processing to new spectra, projecting them onto the LV space. The model then uses the resulting scores to predict concentration values.

Here, T and U are the latent variable score matrices, summarizing the information contained in the LVs. The matrices P and Q are the corresponding loadings which indicate the contribution of each variable to the latent variables. Finally, E and F are residuals, representing the unexplained variance in X or the difference between the predicted and measured Y, respectively. To extract LVs, PLS implementations typically use an iterative algorithm, such as nonlinear iterative partial least squares (NIPALS) [43]. During each iteration, weight vectors for X and Y are estimated to maximize the covariance between X and Y, and the corresponding LV T and U are calculated. The data variance explained by the LV is removed from the data before the next LV is extracted, and this continues until the desired number of LVs is obtained. The optimal number of LVs is chosen to minimize prediction error while avoiding overfitting, typically determined using cross-validation.

During model calibration, a data set consisting of X data (spectra) and known Y responses (e.g., chemical compound concentrations) is provided, which the PLS model uses to extract LVs that maximize covariance between the provided X data and Y responses (Figure 1.4). This model calibration approach minimizes the need for prior process knowledge, as only representative Raman spectra (X data) and reference measurements of the compound of interest (Y-data) are required. Once the latent variables are extracted, the regression coefficient vector B relating X to Y can be expressed as shown in Equation 1.3.

$$B = W(P^T W)^{-1} Q^T \quad (1.3)$$

Here, W contains the weights defining the latent variables, P and Q are loadings of X and Y, respectively. A calibrated PLS model can then be used to predict the response variables (e.g., concentration) of new spectral data according to two approaches, namely approach 1 and approach 2. During full PLS projection the new pre-processed spectra X_{new} are projected onto each LV to obtain scores as described in Equation 1.4.

$$T_{new} = X_{new} W(P^T W)^{-1} \quad (1.4)$$

T_{new} can be used to calculate the predicted responses ($\hat{Y}$) as shown in Equation 1.5.

$$\hat{Y} = T_{new} Q^T \quad (1.5)$$

The full PLS projection approach provides intermediate scores of the new spectra on each LV, which can be useful for process monitoring and model interpretation. Alternatively, the new pre-processed spectra can be directly multiplied with the regression coefficient B to obtain predictions (Figure 1), as described in Equation 1.6.

$$\hat{Y} = X_{new}B \quad (1.6)$$

Equation 1.5 is a simplified and computationally efficient approach for obtaining predictions but does not provide scores on individual LVs, thereby losing the ability to interpret sample positioning within the calibration dataset distribution or to perform outlier detection. The flexibility and computational simplicity allows for PLS models to be easily applied for real-time process monitoring. The compatibility with complex spectral data, interpretability of the model, computational efficiency, and widely available software packages made PLS one of the most popular modelling approaches for Raman spectroscopy [13, 44].

Both during calibration and validation PLS models are evaluated using several performance statistics and indicators to ensure that they have strong predictive performance. During calibration the number of LVs to include in the model is chosen based on the percentage of explained variance in X and Y and through the root mean square errors of calibration (RMSEC) and cross-validation (RMSECV). Cross-validation estimates model performance by iteratively leaving out subsets of the calibration data, calibrating a model on the remaining samples, and predicting the excluded samples thereby providing an internal measure for predictive error. This procedure is also important to prevent overfitting on the calibration data as additional LVs are only included when they improve predictions on excluded samples. After calibration PLS models are validated on an independent dataset from which a root mean square error of prediction (RMSEP) is derived, indicating the predicting performance of the model on unseen data. The specificity of a PLS model can be evaluated by inspecting the variable weights in the regression coefficient vector B which summarizes the importance of each X variable for predicting Y [45]. If strong regression coefficients correspond to known spectral features of the analyte of interest, it suggests that the model predictions are chemically meaningful and not driven by unrelated variation. A robust PLS model therefore combines low errors (RMSEC, RMSECV, RMSEP), stable performance across calibration and validation, and chemically meaningful regression coefficients that demonstrate specificity to the target compound.

Next to PLS, another notable modelling approach is support vector regression (SVR), which is more effective at handling non-linear relationships between signal intensity and compound concentrations. The SVR technique has been successfully applied for compound quantification in bioreactor processes, but it requires more fine-tuning of model parameters compared to PLS [44, 46]. Artificial neural networks (ANNs) play an increasingly significant role in the application of Raman spectroscopy for measuring biological processes, where ANNs excel at classification problems and at capturing complex non-linear relationships [47]. However, the lower interpretability of a neural network as well as the requirement of large process datasets for calibration are considered disadvantages. Recent use of data augmentation for spectral dataset expansion may improve the performance of ANNs for classification problems [48].

Despite the growing popularity of other chemometric modelling techniques, PLS remains a staple technique for modelling Raman spectroscopic data due to its simplicity and interpretability. However, the data-driven nature hinders the implementation of Raman spectroscopy in bioprocesses, where calibration typically cannot be finalized until sufficient process data has been collected. This is time- and labor-intensive, and during early-stage process development often not feasible. Furthermore, calibration with process data leads to highly process-specific models, meaning PLS models perform well on processes operated under identical conditions but can be highly sensitive to changes in process parameters. Furthermore, PLS models often do not transfer to different operational modes with similar cell lines and cultivation media [49, 50]. This results in the need to repeat the extensive collection of process data for every new process to be monitored. All in all, this slows down the adoption of Raman spectroscopy as the calibration efforts are high and the applicability of these models is limited.

Alternative routes to translate Raman spectra to quantitative output are hard models. A recent example is the model-driven Indirect Hard Model (IHM) approach that describes spectral data based on known components [51]. In this approach, each process compound is individually modelled by fitting Pseudo-Voigt peak functions to pure spectra, or to mixture spectra using complemental hard modelling [52]. In this way, each compound is described by a defined set of peak shapes, and a mixture model is constructed from these individual hard models. The mixture model is calibrated using a spectral dataset with known concentrations, adjusting the peak weights of each compound to fit the data and enabling the prediction of compound concentrations in new process spectra. Because the model does not have to learn new spectral variation during calibration, this can be performed using a few simple single compound or mixture spectra. This results in a model based on physical principles, and has increased flexibility over the PLS approach as the mixture model can be expanded with new compound hard models when required [10]. However, the level of process knowledge required for this approach is difficult to achieve for many bioreactor processes, and the reported literature on these methods is currently limited. The major advantage of the IHM approach is that actual process data is not required for model calibration, meaning that a monitoring approach can be set up before the actual process is operated. This tackles one of the main hurdles for the implementation of Raman spectroscopy through data-driven approaches: the dependency on process data for calibration. Process data independence enables fast implementation of Raman spectroscopy in early process development as calibration can be performed with simple single compound or mixture measurements. In addition, the mixture model bases predictions on physical knowledge instead of statistical relationships, and is therefore more robust to operational deviations.

1.4 Motivation and aim of the thesis

The three major challenges defined in this thesis are (1) the requirement for extensive collection of process calibration data, (2) the low transferability of process-specific calibration models, and (3) the dependency on data-driven calibration modes. To overcome these challenges, the aim of this thesis is to compare alternative approaches to calibration data collection and model calibration, and study the influence of changing measurement conditions on spectral integrity. More knowledge on these subjects leads to efficient and flexible modelling strategies that decrease the dependency on process data for model calibration. This opens the door to rapid implementation of real-time monitoring and advanced automated process control during the early stages of process development.

1.5 Thesis outline

Chapter 1 introduces challenges of state-of-the-art bioreactor monitoring and control, and demonstrates the need for automated real-time monitoring technologies. It provides an overview of the potential of Raman spectroscopy for in-line quantification of key process parameters, as well as the current challenges related to chemometric model development that lies at the foundation of using Raman spectroscopy for real-time quantification.

The work in **Chapter 2** applied a common workflow for PLS model calibration using bioreactor-based process data to showcase the limitations arising from cross-correlations and compares an alternative calibration approach to increase model robustness. This is done by performing qualitative model assessment and extracting quantitative evaluation parameters often used to determine model accuracy. This comparison demonstrates that quantitative data parameters alone cannot identify severe cross-dependencies in the model, resulting from solely using process data in the calibration dataset. This chapter shows how supplementing the calibration dataset with single compound spectra improved model specificity and performance, also when applied to a related process. This work indicates how simple data collection strategies can overcome process specific dependencies in calibration datasets without the need for more process data.

Robust PLS models depend not only on the composition of the calibration dataset but also on the pre-treatment of the spectral data to remove undesired variation. Therefore, spectral pre-processing is essential in chemometric modelling to ensure signal linearity and prevent spectral misalignment, and a deeper understanding of spectral distortion sources allows for an improved separation between relevant signals and noise. **Chapter 3** dives into the influence of measurement conditions on the integrity of Raman spectra acquired in bioreactors. This work reports on the influence of temperature, bubble size and quantities, viscosity, and *Saccharomyces cerevisiae* biomass on Raman spectra. This was done by systematically investigating the influence of these parameters on baseline shifts, peak position, and peak intensity. The observed spectral changes caused by these process parameters can potentially disrupt model prediction accuracy and data pre-processing methods to minimize these influences are discussed. Improved understanding of different sources of spectral distortion contributes to the development of robust chemometric models applicable across different experimental setups with varying measurement conditions.

Chemometric model calibration with process data, as seen in Chapter 2, leads to process-specific models with poor transferability, necessitating repeated data collection for each new process which delays the adoption and implementation of Raman spectroscopy. Chapter 3 demonstrated that the molecular composition of yeast can be detected using in-line Raman spectroscopy, although the signals are easily overshadowed by extinction effects contributions from other process compounds. The work in **Chapter 4** demonstrates four approaches by which a simple dataset of single compound spectra can be used to calibrate highly specific quantification models for monitoring glucose, ethanol, and biomass concentration in *S. cerevisiae* fermentation. The effectiveness of calibration with single compound spectra is compared for data-driven PLS models and a model-driven IHM approach, and the areas of application for these strategies are discussed. The third approach includes a workflow for distinct peak feature isolation and applies these features for de novo generation of synthetic spectra, effectively simulating fermentation conditions. Lastly, the isolated peak features are used to augment process spectra with the aim of enhancing compound variability in the calibration dataset. A comparison of these four approaches highlights potential strategies for rapid model development and data augmentation, contributing to the efforts of model calibration without the need of process data.

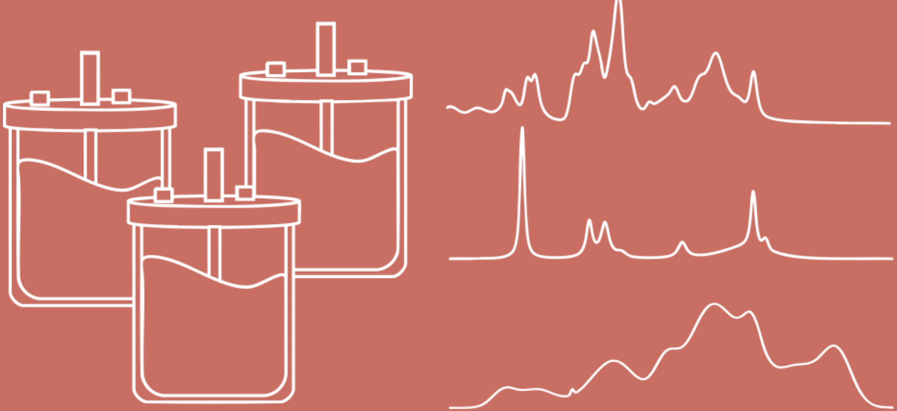
Finally, **Chapter 5** provides an overview of how the key findings from this thesis address the defined challenges, and reflects on how these contributions aid in streamlining Raman spectroscopy implementation for real-time bioreactor monitoring. An outlook is provided to continue the discussion on alternative methods for rapid model development. The value of miniaturized Raman spectroscopy setups is examined in the context of sample exploration and high-throughput data collection. And finally, the future position of Raman spectroscopy in the bioprocessing industry is emphasized.

References

1. Ündey, C., et al., *Applied advanced process analytics in biopharmaceutical manufacturing: Challenges and prospects in real-time monitoring and control.* Journal of Process Control, 2010. **20**(9): p. 1009-1018.
2. O'Mara, P., et al., *Staying alive! Sensors used for monitoring cell health in bioreactors.* Talanta, 2018. **176**: p. 130-139.
3. Rathore, A.S. and H. Winkle, *Quality by design for biopharmaceuticals.* Nature biotechnology, 2009. **27**(1): p. 26-34.
4. FDA, *Guidance for industry: PAT—A framework for innovative pharmaceutical development, manufacturing, and quality assurance.* Food and Drug Administration, Rockville, MD, 2004.
5. Minnich, C., S. Hardy, and S. Krämer, *Stopping the babylonian confusion: An updated nomenclature for process analyzers in PAT applications.* Chemie Ingenieur Technik, 2016. **88**(6): p. 694-697.
6. Shek, C.F. and M. Betenbaugh, *Taking the pulse of bioprocesses: at-line and in-line monitoring of mammalian cell cultures.* Current opinion in biotechnology, 2021. **71**: p. 191-197.
7. Mishra, A., et al., *Spectroscopic Advances in Real Time Monitoring of Pharmaceutical Bioprocesses: A Review of Vibrational and Fluorescence Techniques.* Spectroscopy Journal, 2025. **3**(2): p. 12.
8. Dahotre, S., et al., *Real-time monitoring of antibody quality attributes for cell culture production processes in bioreactors via integration of an automated sampling technology with multi-dimensional liquid chromatography mass spectrometry.* Journal of Chromatography A, 2022. **1672**: p. 463067.
9. Vojinović, V., J. Cabral, and L. Fonseca, *Real-time bioprocess monitoring: Part I: In situ sensors.* Sensors and Actuators B: Chemical, 2006. **114**(2): p. 1083-1091.
10. Müller, D.H., et al., *Bioprocess in‐line monitoring using Raman spectroscopy and Indirect Hard Modeling (IHM): A simple calibration yields a robust model.* Biotechnology and Bioengineering, 2023.
11. Graf, A., et al., *A novel approach for non-invasive continuous in-line control of perfusion cell cultivations by Raman spectroscopy.* Frontiers in bioengineering and biotechnology, 2022. **10**: p. 719614.
12. Romann, P., et al., *Advancing Raman model calibration for perfusion bioprocesses using spiked harvest libraries.* Biotechnology Journal, 2022: p. 2200184.
13. Esmonde-White, K.A., M. Cuellar, and I.R. Lewis, *The role of Raman spectroscopy in biopharmaceuticals from development to manufacturing.* Analytical and Bioanalytical Chemistry, 2021: p. 1-23.
14. Buckley, K. and A.G. Ryder, *Applications of Raman spectroscopy in biopharmaceutical manufacturing: a short review.* Applied spectroscopy, 2017. **71**(6): p. 1085-1116.
15. Drieschner, T., et al., *Direct optical detection of cell density and viability of mammalian cells by means of UV/VIS spectroscopy.* Analytical and Bioanalytical Chemistry, 2020. **412**: p. 3359-3371.
16. Hisiger, S. and M. Jolicoeur, *A multiwavelength fluorescence probe: Is one probe capable for on-line monitoring of recombinant protein production and biomass activity?* Journal of biotechnology, 2005. **117**(4): p. 325-336.
17. Li, M., et al., *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, 2018. **137**: p. 205-213.

18. Pontius, K., et al., *Monitoring yeast fermentations by nonlinear infrared technology and chemometrics—understanding process correlations and indirect predictions.* Applied microbiology and biotechnology, 2020. **104**(12): p. 5315-5335.
19. Vann, L. and J. Sheppard, *Use of near-infrared spectroscopy (NIRs) in the biopharmaceutical industry for real-time determination of critical process parameters and integration of advanced feedback control strategies using MIDUS control.* Journal of Industrial Microbiology and Biotechnology, 2017. **44**(12): p. 1589-1603.
20. Rowland-Jones, R.C., et al., *Comparison of spectroscopy technologies for improved monitoring of cell culture processes in miniature bioreactors.* Biotechnology Progress, 2017. **33**(2): p. 337-346.
21. Abu-Absi, N.R., et al., *Real time monitoring of multiple parameters in mammalian cell culture bioreactors using an in-line Raman spectroscopy probe.* Biotechnology and bioengineering, 2011. **108**(5): p. 1215-1221.
22. Iversen, J.A., R.W. Berg, and B.K. Ahring, *Quantitative monitoring of yeast fermentation using Raman spectroscopy.* Analytical and bioanalytical chemistry, 2014. **406**(20): p. 4911-4919.
23. Domján, J., et al., *Raman-based dynamic feeding strategies using real-time glucose concentration monitoring system during adalimumab producing CHO cell cultivation.* Biotechnology Progress, 2020. **36**(6): p. e3052.
24. Domján, J., et al., *Real-time amino acid and glucose monitoring system for the automatic control of nutrient feeding in CHO cell culture using Raman spectroscopy.* Biotechnology Journal, 2022. **17**(5): p. 2100395.
25. Gibbons, L., et al., *Raman Based Chemometric Model Development for Glycation and Glycosylation Real Time Monitoring in a Manufacturing Scale CHO Cell Bioreactor Process.* Biotechnology Progress: p. e3223.
26. Li, M.Y., et al., *Real-time monitoring of antibody glycosylation site occupancy by in situ Raman spectroscopy during bioreactor CHO cell cultures.* Biotechnology progress, 2018. **34**(2): p. 486-493.
27. Raman, C.V. and K.S. Krishnan, *A new type of secondary radiation.* Nature, 1928. **121**(3048): p. 501-502.
28. Dietzek, B., et al., *Introduction to the fundamentals of Raman spectroscopy*, in *Confocal Raman Microscopy*. 2010, Springer. p. 21-42.
29. André, S., et al., *Mammalian cell culture monitoring using in situ spectroscopy: Is your method really optimised?* Biotechnology Progress, 2017. **33**(2): p. 308-316.
30. Yang, N., et al., *Raman spectroscopy applied to online monitoring of a bioreactor: Tackling the limit of detection.* Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2024. **304**: p. 123343.
31. Jahn, I.J., et al., *Noise Sources and Requirements for Confocal Raman Spectrometers in Biosensor Applications.* Sensors, 2021. **21**(15): p. 5067.
32. Wei, D., S. Chen, and Q. Liu, *Review of fluorescence suppression techniques in Raman spectroscopy.* Applied Spectroscopy Reviews, 2015. **50**(5): p. 387-406.
33. Esmonde-White, K.A., et al., *Raman spectroscopy as a process analytical technology for pharmaceutical manufacturing and bioprocessing.* Analytical and bioanalytical chemistry, 2017. **409**(3): p. 637-649.
34. Tuschel, D., *Selecting an excitation wavelength for Raman spectroscopy.* 2016.
35. Knorr, F., Z.J. Smith, and S. Wachsmann-Hogiu, *Development of a time-gated system for Raman spectroscopy of biological samples.* Optics express, 2010. **18**(19): p. 20049-20058.
36. Kostamovaara, J., et al., *Fluorescence suppression in Raman spectroscopy using a time-gated CMOS SPAD.* Optics express, 2013. **21**(25): p. 31632-31645.

37. Lee, H.L., et al., *In situ bioprocess monitoring of Escherichia coli bioreactions using Raman spectroscopy.* Vibrational Spectroscopy, 2004. **35**(1-2): p. 131-137.
38. Sinfield, J.V. and C.K. Monwuba, *Assessment and correction of turbidity effects on Raman observations of chemicals in aqueous solutions.* Applied spectroscopy, 2014. **68**(12): p. 1381-1392.
39. Engel, J., et al., *Breaking with trends in pre-processing?* TrAC Trends in Analytical Chemistry, 2013. **50**: p. 96-106.
40. Storey, E.E. and A.S. Helmy, *Optimized preprocessing and machine learning for quantitative Raman spectroscopy in biology.* Journal of Raman Spectroscopy, 2019. **50**(7): p. 958-968.
41. Ryabchykov, O., S. Guo, and T. Bocklitz, *Analyzing Raman spectroscopic data.* Physical Sciences Reviews, 2019. **4**(2): p. 20170043.
42. Wold, S., M. Sjöström, and L. Eriksson, *PLS-regression: a basic tool of chemometrics.* Chemometrics and intelligent laboratory systems, 2001. **58**(2): p. 109-130.
43. Wold, H., *Path models with latent variables: The NIPALS approach*, in *Quantitative sociology*. 1975, Elsevier. p. 307-357.
44. Rafferty, C., et al., *Analysis of chemometric models applied to Raman spectroscopy for monitoring key metabolites of cell culture.* Biotechnology progress, 2020. **36**(4): p. e2977.
45. Seasholtz, M.B. and B.R. Kowalski, *Qualitative information from multivariate calibration models.* Applied spectroscopy, 1990. **44**(8): p. 1337-1348.
46. Nadadoor, V.R., et al., *Online sensor for monitoring a microalgal bioreactor system using support vector regression.* Chemometrics and Intelligent Laboratory Systems, 2012. **110**(1): p. 38-48.
47. Luo, R., J. Popp, and T. Bocklitz, *Deep learning for Raman spectroscopy: A review.* Analytica, 2022. **3**(3): p. 287-301.
48. Wu, M., et al., *Deep learning data augmentation for Raman spectroscopy cancer tissue classification.* Scientific reports, 2021. **11**(1): p. 23842.
49. Berry, B., et al., *Cross-scale predictive modeling of CHO cell culture growth and metabolites using R aman spectroscopy and multivariate analysis.* Biotechnology progress, 2015. **31**(2): p. 566-577.
50. Yousefi-Darani, A., et al., *Generic Chemometric Models for Metabolite Concentration Prediction Based on Raman Spectra.* Sensors, 2022. **22**(15): p. 5581.
51. Echtermeyer, A., et al., *Inline Raman spectroscopy and indirect hard modeling for concentration monitoring of dissociated acid species.* Applied spectroscopy, 2021. **75**(5): p. 506-519.
52. Kriesten, E., et al., *Identification of unknown pure component spectra by indirect hard modeling.* Chemometrics and Intelligent Laboratory Systems, 2008. **93**(2): p. 108-119.



Chapter 2

Single compound data supplementation to enhance transferability of fermentation specific Raman spectroscopy models

---

This chapter has been published as:

Klaverdijk, M., Ottens, M., & Klijn, M. E. (2025). Single compound data supplementation to enhance transferability of fermentation specific Raman spectroscopy models. *Analytical and Bioanalytical Chemistry*, 1-12 (https://doi.org/10.1007/s00216-025-05768-5)

Abstract

Raman spectroscopy is a valuable analytical tool for real-time analyte quantification in fermentation processes. Quantification is performed with chemometric models that translate Raman spectra into concentration values, which are typically calibrated with process data from multiple comparable fermentations. However, process-specific models underperform for minor process variation(s) or different operation modes due to the integration of cross-correlations, resulting in low target analyte specificity. Thus, model transferability is poor and labor-intensive (re-)calibration of models is required for related processes. In this work, Partial Least Squares models for glucose, ethanol, and biomass were calibrated with *Saccharomyces cerevisiae* batch fermentation data and subsequently transferred to a fed-batch operation. To enhance model transferability without additional process runs, single compound data supplementation was performed. The supplemented models increased overall target analyte specificity and demonstrated sufficient prediction accuracy for the fed-batch process (root mean squared errors of prediction (RMSEP) of 3.06 mM, 8.65 mM, and 0.99 g/L for glucose, ethanol, and biomass), while maintaining high prediction accuracy for the batch process (RMSEP of 1.71 mM, 4.20 mM, and 0.17 g/L for glucose, ethanol, and biomass). This work showcases that process data in combination with single compound spectra is a fast and efficient strategy to apply Raman spectroscopy for real-time process monitoring across related processes.

Keywords: *Raman spectroscopy, chemometrics, Partial Least Squares (PLS), Saccharomyces cerevisiae, real-time monitoring*

2.1 Introduction

Bioprocess development, optimization, and control is dependent on various information sources during fermentation, ranging from pH and temperature to cell viability and product concentration. State-of-the-art bioreactor processes already include automated measurements and coupled control loops for decades, but these are often limited to parameters such as temperature, pH, dissolved oxygen (DO), and off-gas analysis [1]. For information on metabolite concentration, product quality, or the biomass, labor-intensive manual sampling and analysis is performed by trained process technicians. Accurate monitoring of all process parameters during fermentation is considered essential to ensure fast bioprocess development and stable manufacturing. Although off-line analytics are the golden standard for investigating complex parameters, manual sampling required for these measurements causes a limited and delayed perspective of the process. This leads to retrospective decision-making and the inability to pro-actively make process adjustments to ensure a successful run.

The bioprocessing industry is actively investigating process analytical technologies (PAT) that provide detailed information on a wide(r) range of critical process and product parameters, allowing the transition to data-driven bioprocess development and automated decision-making [2]. Robust real-time quantification of metabolite, product, and biomass concentrations can lead to more efficient bioprocess development and serve as response values required for automated bioprocess control. There are many different analytical techniques available for abovementioned parameters, such as enzyme-based biosensors, impedance-based probes, capacitance probes, and optical spectroscopy tools [3-5]. Among the available technologies, in-line optical spectroscopy tools that capture molecular vibrations have the advantage of being non-invasive, non-destructive, and provide continuous measurements [6]. Raman spectroscopy is a suitable choice for aqueous systems as seen in bioreactors, due to the low signal interference from water. Moreover, Raman spectroscopy captures spectral contributions of multiple analytes present in a bioreactor in a single spectrum, thereby facilitating efficient multiplexed quantification. The obtained Raman spectra are correlated to reference measurements from off-line analytics to translate the complex information into quantitative metrics required for bioprocess development and control. This is achieved using chemometric methods, which focus on studying the relationship between chemical measurements and the properties of interest during a process [7]. Chemometric methods in combination with Raman spectroscopic data have been employed for in-line monitoring and quantification of substrates, products, waste-products, and cell density in a wide range of processes, such as ethanol production by yeast and antibody production with CHO cell lines [8-10].

Although different chemometric methods are reported on, Partial Least Squares (PLS) regression is the most used technique for calibrating Raman spectroscopy quantification models in bioreactor applications [10-12]. Raman spectra contain many spectral variables (wavenumbers) while there is often only a single response parameter (reference measurement) per analyte [12, 13]. PLS models are considered appropriate for such datasets, described as systems containing high numbers of collinear predictor variables and limited response values [14]. The method is widely available in statistical software packages and allows the development of quantification models without needing extensive process knowledge. By using Raman spectra obtained during fermentation and orthogonal off-line measurements on an analyte of interest as input data, the PLS model is calibrated by assigning weights to the relevant spectral variables for its quantification. The calibrated model can be interpreted through the regression coefficients and loadings of the generated latent variables. However, an appropriate calibration dataset should capture data over the full process range, contain process and biological variability, and preferably contains an even sample distribution to prevent possible accuracy biases. As the biological processes which we want to monitor are subject to inherent cross-correlations between the changes of substrate, product, and biomass concentrations, cross-correlations are directly incorporated into the calibration datasets [15]. This is a challenge for calibrating robust PLS models, as it is an implicit modeling technique that maximizes covariance between the X (Raman spectra) and Y (reference method) data and has no knowledge of the system. The strong cross-correlations between the response values can cause a PLS model to become non-specific, meaning that information from one compound (e.g. substrate) is used to quantify a cross-correlated compound (e.g. product or biomass). This subsequently leads to non-specificity of models towards the analyte of interest. This lack of analyte specificity may be less important in cases where the relationship between the cross-correlated parameters does not change (e.g., monitoring identical processes), or when the goal is to monitor solely process evolution. However, as these models are tailored to a specific process, the predictive capability is compromised when these models are applied to processes where the relation between the parameters changes [16]. This can limit the potential of continuous monitoring and control in a dynamic research and development environment, where aspects such as feeding strategies, mode of operation, or inoculation density can still be subject to change. As a consequence, upon the need of monitoring related processes, the extensive data collection and model calibration procedure has to be repeated. Whereas the collection of in-line Raman spectra has been automated with recent technological advancements, the collection of high-quality measurements required as orthogonal reference data is material- and labor-intensive. Furthermore, high quality process data is not directly available when changing process parameters, meaning that Raman spectroscopy monitoring is not available until the data of several new process runs is collected. Next to that, the effects of a changed process parameter on the concentration ranges of compounds of interest are not always known beforehand. This limits the implementation speed and impact of real-time monitoring with Raman spectroscopy in dynamic environments, where live process data could be essential to early process understanding. Therefore, it is desired to efficiently develop robust models, to ensure model predictive performance remains unaffected by (minor) variations in manufacturing and environmental conditions.

To prevent the adverse effects of highly cross-correlated process data on the performance of PLS models across related processes, calibration datasets should be adjusted accordingly. Alternative approaches for data collection to supplement regular process data include target compound spiking (i.e. analyte spiking), generating synthetic spectra, and combining these approaches with design of experiments (DoE). For the first approach, spiking, the level of cross-correlation in a calibration dataset is evaluated and subsequently disrupted by supplementing additional samples. For example, a cross-correlation between substrate consumption and product formation can be disrupted by spiking the substrate in the bioreactor to generate conditions that break the cross-correlation observed in the original process. However, spiking compounds during an ongoing at-scale bioprocess is not efficient, as it may impair the continuation of the process. Therefore, measurements of spiked samples are often acquired in smaller scale reactors or shake flasks [15, 17]. The second approach involves the generation of synthetic samples outside of the bioreactor process to mimic process conditions. Samples can either be completely synthetic by externally preparing single compound mixtures of target compounds in cultivation media, or bioreactor samples can be utilized by altering the compound concentrations and performing measurements in miniaturized acquisition setups [18]. DoE techniques are often employed to design samples with uncorrelated compounds [19]. The combination of bioreactor and synthetic spectral data has proven to increase calibration model performance for several processes [15, 19]. However, depending on the complexity of the process and the number of compounds to be considered, these approaches can still lead to material- and labor-intensive calibration dataset preparation. As PLS models are typically calibrated to quantify only one process compound, it is essential that the model identifies and correlates spectral features specific to that single compound. Considering this concept, it may not be necessary to add data containing variation on other compounds, as these are already accounted for by process data in the base dataset. By supplementing the dataset with spectra of the target compound at varying concentrations, spectral features associated with the compound of interest are emphasized, which reduces cross-correlations within the dataset. Through this supplementation approach models may be improved and extended beyond the calibration ranges of the original dataset based on process data.

This work investigates the applicability of single compound spectra supplementation as an efficient and simple alternative for calibration dataset adjustment to improve PLS model performance across related processes. The yeast *Saccharomyces cerevisiae* was used as a model system to generate batch process data for the calibration of base models for the quantification of glucose, ethanol, and biomass. These base models are subsequently transferred to a fed-batch cultivation where the models have to extrapolate and deal with new process conditions. We demonstrate how standard quantitative model validation is not always sufficient to assess model quality and could lead to poor performance when process parameters change. After a qualitative assessment of the models, the calibration datasets were supplemented with single compound spectra to improve model transferability by increasing the calibration range and target specificity. This work shows the impact of qualitative model assessment and how to efficiently adjust calibration datasets to improve model transferability. This concept of a base model in combination with data supplementation can be used to extend the applicability of Raman spectroscopy models across related processes, without the need for collecting new process data. Methods like these could allow more efficient and flexible model transfer within dynamic process development environments.

2.2 Materials & Methods

2.2.1 Fermentation

The yeast strain *Saccharomyces cerevisiae* CEN.PK113-7D was used for all experiments [20]. All cultures were grown on synthetic medium containing 5 g/L (NH4)2SO4, 3 g/L KH2PO4, and 0.5 g/L MgSO4.7H2O adjusted to pH 6.0 with 2M KOH [21]. The pH was measured with an offline pH probe (Consort, Turnhout, Belgium). Glucose was used as carbon source, and an initial concentration of 20 g/L was reached by the addition of sterilized 50% glucose (J.T. Baker, Philipsburg, NJ) solution (in-house). Vitamin solution was sterilized through 0.2 µm syringe filters (Whatman, Maidstone, UK) and added after sterilization of the medium. Bioreactor medium was supplemented with 0.2 g/L sterile Antifoam-C (BASF, Ludwigshafen, Germany) after autoclaving.

Pre-cultures were grown aerobically in 500mL shake flasks with a 100mL working volume and were incubated at 30°C and 150 rpm in an orbital shaker (Sartorius, Göttingen, Germany). Batch cultures were grown in 2L stirred tank reactors (Applikon, Delft, the Netherlands) using a working volume of 1L. The cultures were aerated with 0.5 L/min air while stirred at 800 rpm and maintained at a temperature of 30 °C. A pH of 6.0 was maintained by the automatic addition of 2M KOH. Batch cultures were inoculated at an initial OD660 of 0.3. The fed-batch culture started out as a batch process described above, but was spiked with sterile 50% glucose solution when substrate depletion was detected through a decrease in CO_2 production. CO_2 production was measured with off-gas analysis with a ServoPRO 4900 (Servomex, Crowborough, UK). The fed-batch culture was fed with 50% glucose solution three times, thereby extending the process duration and increasing the final concentrations of ethanol and biomass. An overview of glucose, ethanol, and biomass concentrations during the performed batch and fed-batch fermentations is provided in Supplementary Figure 2.5.1.1.

2.2.2 Single compound solutions

Single compound spectra were acquired in the same 2L bioreactor as described in Section 2.1 to maintain similar acquisition conditions, such as aeration, stirring, and temperature. For each analyte, the bioreactor was filled with 1L of synthetic media (as described in section 2.1) with a constant temperature of 30 °C, stirring rate of 800rpm, and aeration with 0.5 L/min of air. Glucose and ethanol concentration ranges were generated by the stepwise addition of 10 mL prepared glucose or ethanol solution. After acquiring Raman spectra at a concentration, a 10 mL sample was taken for reference measurements by HPLC, and the subsequent concentration was achieved by adding the next 10 mL prepared solution. The 10mL additions were prepared beforehand according to predefined concentration distributions. The glucose concentrations were increased in an exponentially increasing manner to generate more low concentration conditions. The ethanol concentration was increased with steps of 20mM. To generate biomass single compound spectra, a batch process was operated until glucose depletion occurred, after which the full bioreactor volume was harvested. The cell suspension was centrifuged and washed twice in fresh synthetic media to remove any remaining glucose, ethanol, and other compounds. Cells were resuspended in synthetic media to a concentration of 14.3 g/L and used to increase cell density in a clean bioreactor by adding the suspension in a stepwise manner. Concentrations were measured in the ranges: 0-247 mM for glucose, 0-500 mM for ethanol, and 0-4.9 g/L for biomass with HPLC and dry-weight determination (Supplementary Table 2.5.2.1).

2.2.3 Reference data

2.2.3.1 Biomass

The batch fermentations were sampled every hour and cell growth was determined by optical density at a wavelength of 660nm (OD660) using a Libra S11 spectrophotometer (Biochrom, Cambridge, United Kingdom). Dry-weight determination was performed by loading 10mL of culture broth on nitrocellulose membrane filters (pore size, 0.45 µm; Gelman Laboratory, Ann Arbor, MI), drying the filters in a microwave and subsequently weighing the dry biomass (Mettler Toledo, Columbus, USA).

2.2.3.2 HPLC

Sample supernatants of the batch, fed-batch, and single compound samples were analysed for the corresponding ethanol and glucose concentrations using an Agilent 1260 infinity HPLC (Agilent Technologies, Santa Clara, CA). A BIO-RAD Aminex HPX-87H (300 x 7.8 mm) cation-exchange column (Bio-Rad, Hercules, CA) operated at a temperature of 60°C and 0.5 g/l H2SO4 was used as eluent with a 0.6 ml/min flow rate. The injection volume was 5 µm and an Agilent 1260 refractive-index and VWD detector at 214 nm was used for characterization.

2.2.4 Raman spectroscopy

2.2.4.1 Signal acquisition

A Raman RXN2 analyzer (Kaiser Optical Systems Inc., Ann Arbor, MI) equipped with a 785 nm laser was connected to the bioreactor with a fibre optic cable and bIO-Optic immersion probe to collect spectra over the range of 100-3,400 cm^{-1}. The immersion probe was mounted through the head plate and sterilized with the bioreactor. Several acquisition settings were assessed to find the optimal acquisition settings which provided a good signal-to-noise ratio while maintaining a high monitoring resolution. An exposure time of 60-second resulted in a detector saturation between 30-58% over the full process range. The Raman spectrometer was set to continuously acquire individual 60-second spectra that could be combined into longer measurements after data collection. Datasets of 1, 2, 4, 6, 8, and 10 minutes matching the timepoint of reference sampling were generated according to a protocol similar to Andre et al. [22]. Initial PLS models were calibrated using all different acquisition lengths, and their prediction performance was evaluated according to the methods described in Section 2.4.2. Prediction performance remained similar for acquisitions times above 1 minute of acquisition time (Supplementary Figure 2.5.3.1). Therefore, all subsequent models were calibrated with 1-minute Raman spectra to maintain a high data resolution during process monitoring. The single compound Raman spectra were acquired by measuring ten individual spectra of 1 minute per concentration and the ten spectra were averaged to obtain noise-free spectra.

2.2.4.2 Signal pre-processing and model building

Spectral pre-processing and model building was performed in MATLAB R2020b (MathWorks, WA), using PLS_Toolbox (v 9.2, Eigenvector Research, WA). The first step of pre-processing was the selecting of the fingerprint region ranging from 450 cm^{-1} to 1800 cm^{-1}. It was observed that background fluorescence and scattering effects in the spectral dataset increased exponentially towards the lower wavenumber region. Therefore, a basic extended multiplicative scatter correction (EMSC) was chosen using a quadratic term and the average spectra as the regression reference to properly fit and eliminate scattering effects [23]. All datasets were mean-centered before the modeling steps. The PLS models were calibrated by loading the spectra as X-data and the reference measurements per target analyte as the Y-data (HPLC and biomass). Venetian Blinds cross-validation with 7 folds was applied and the number of latent variables were selected based on the elbow point of the root mean square error of calibration (RMSEC) and cross-validation (RMSECV) scores (Supplementary Figures 2.5.4.1, 2.5.4.2). The quantitative validation of the base and supplemented models was performed by applying the models to a validation dataset consisting of an unseen batch fermentation, and evaluating the performance based on the RMSEC, RMSECV, root mean square error of prediction (RMSEP). To compare the performance of the base and supplemented models, relative root mean square errors (rRMSE) were used. An rRMSE based on the interquartile range (IQR) of the calibration dataset (for rRMSEC and rRMSECV) or application dataset (rRMSEP) was chosen (Equation 1) to reduce the influence of skewed data distributions when assessing model performance:

$$rRMSE = \frac{RMSE}{Q3 - Q1} \times 100 \qquad (2.1)$$

where Q1 is the 25th percentile, and Q3 is the 75th percentile of the dataset. This approach ensures the relative error measure is robust to outliers and focuses on the variability within the central portion of the data. The regression coefficient vectors were investigated for qualitative validation of the models and compared to single compound spectra of the target analyte for each model.

2.3 Results and Discussion

The impact of combining Raman spectra obtained for batch fermentation processes and single compounds on model specificity and transferability to a related process is determined using a qualitative and quantitative assessment approach. Firstly, the standard modelling approach is performed to obtain a base model. Here, batch fermentation data is used for calibration and the model is validated using an unseen but similar batch fermentation dataset. This is included to highlight the importance of applying quantitative and qualitative model assessment, as well as for understanding the connection between model specificity and transferability to the fed-batch process. Subsequently, single compound data supplementation is showcased and the supplemented model performance on both unseen batch and fed-batch data is presented and discussed.

2.3.1 Quantitative analysis of base model performance for batch data

Partial Least Squares (PLS) models were calibrated using fermentation process data to monitor glucose, ethanol, and biomass concentrations during yeast fermentation with Raman spectroscopy. Three individual base models were built (glucose, ethanol, and biomass) using Raman spectra and reference measurements from three batch cultivations (38 samples, Supplementary Figure 2.5.1.1). These base models were subsequently validated on data of one unseen batch cultivation (13 samples). The resulting model statistics and performances are shown in Table 2.1.

Table 2.1: Overview of statistics for the quantitative assessment of the base models for glucose, ethanol, and biomass.

Parameter	Calibration range	RMSEC	RMSECV	RMSEP	rRMSEP	Latent Variables
Glucose	0 – 120.78 mM	1.67 mM	1.83 mM	1.46 mM	2.07 %	2
Ethanol	0 – 172.86 mM	3.67 mM	4.34 mM	4.36 mM	3.57 %	2
Biomass	0.10 – 3.23 g/L	0.07 g/L	0.08 g/L	0.11 g/L	5.38 %	2

The quantitative analysis for all three models displays relative root mean square error of prediction (rRMSEP) values below 5% for the glucose and ethanol model, and 5.38 % for the biomass model. The RMSEP and RMSECV values of each model are close together, suggesting over- or under-fitting of the calibration data does not occur, and that the models perform well on unseen data. This was expected as the unseen batch was operated under identical conditions as the batches used for calibration. This means that quantitative assessment of base model performance according to common practice in literature indicates that the models are well calibrated for quantification during batch fermentation [10].

2.3.2 Base model performance on fed-batch data

In this section, we simulate a model transfer case by transferring the validated base models to a yeast fed-batch fermentation. Through this transfer we evaluate the effectiveness of the base models on a related process containing the same process analytes, but with altered inter-compound ratios (all three analytes) extended concentration ranges (ethanol, biomass). The model performance in shown in the form of measured versus predicted concentration plots in Figure 2.1A-C, along with the statics for quantitative assessment of the base model performance on unseen fed-batch process data in Figure 2.1D.

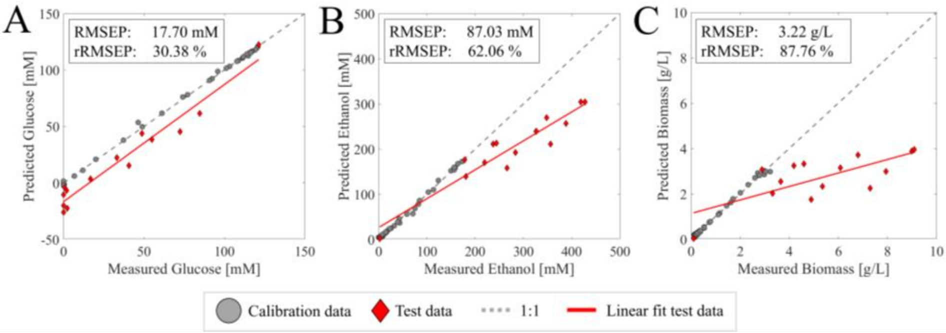


Figure 2.1: Measured (x-axis) versus predicted (y-axis) concentration plots of (A) glucose in mM, (B) ethanol in mM, and (C) biomass in g/L base models applied to spectra obtained during fed-batch cultivation. The fed-batch data is shown as red diamonds, the base calibration data as gray circles, the 1:1 fit as the grey dotted line, and the data fit as the red line. The RMSEP and rRMSEP of each model applied to the fed-batch are shown in the boxes.

Bolus feeding of glucose during the fed-batch process disrupted the ability of all three base models to accurately quantify their target analytes. This is represented by the increased RMSEP and rRMSEP values shown in Figure 2.1. Figure 2.1A shows that the glucose sample at 120 mM is predicted well by the base model, which was taken directly after inoculation and therefore considered highly similar to the batch data. However, new spectral variations in subsequent fed-batch samples caused an underestimation of the glucose concentrations. The RMSEP of the glucose model is 12 times higher when applied to the fed-batch when compared to the batch fermentation, despite the concentrations being in the same range. This highlights the lack of model specificity of glucose, and the model not being able to deal with the new ratios between the compounds. A comparable underestimation is seen for ethanol and biomass (Figure 2.1B and Figure 2.1C, respectively), where the base models was forced to extrapolate due to the extended concentration ranges. For the ethanol and biomass models, the three feeding moments are visible as individual groups of samples, where the predictions move further from the measured values with each glucose addition. Next to extrapolation, the decrease in prediction performance could result from the exponential process and hourly sampling interval of the batch calibration dataset. This led to an uneven sample distribution and a possible accuracy bias towards early-stage batch conditions, which contains high glucose, low ethanol, and low biomass concentrations, and therefore underestimating ethanol and biomass concentrations.

2.3.3 Qualitative assessment of base model

Extrapolation and skewed sample distributions may not be the sole reasons for the decreased performance. PLS models assume a linear relation between the signal response and analyte concentration, and a robust model can sometimes to extrapolate predictions outside of its calibration range with moderate accuracy, assuming that the relationships between the variables remain consistent [24]. As no new analytes were introduced in the fed-batch, the new spectral variations and high RMSEPs are most likely related to the different proportions between the parameters as a result of bolus feeding. Furthermore, the lack of extrapolation capabilities indicates that the obtained base models are not specific to their target analyte. Calibration with batch data led to cross-correlations in the base models, supported by the Pearson coefficients above 0.980 between all analytes (Supplementary Figure 2.5.5.1). A standard quantitative assessment statistic such as RMSEP does not indicate the specificity of each model to its target. The specificity of each base model can be visually inspected by comparing the regression coefficient values for each wavenumber with the single compound spectra of the targets [25]. The regression coefficient vector (RCV) of each base model is compared to single compound spectra of the target analyte in Figure 2.2.

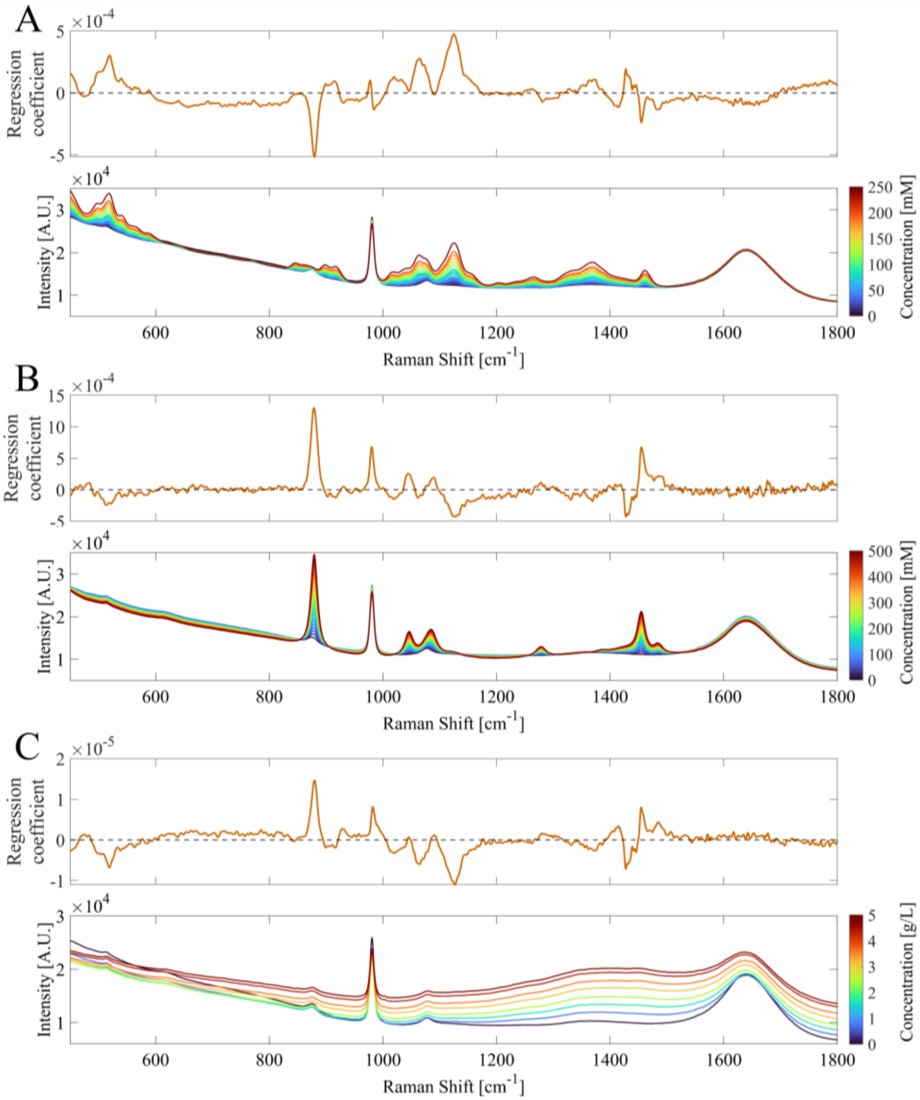


Figure 2.2: The regression coefficient vectors (top) of the (A) glucose in mM, (B) ethanol in mM, and (C) biomass in g/L base models. Each regression vector coefficient is juxtaposed with a concentration range of unprocessed single compound spectra of the models target parameter (bottom). The heatmap of the concentration range shows a low analyte concentration in blue and a high analyte concentration in red.

The regression coefficient vector of the glucose base model (Figure 2.2A) displays a positive correlation to known spectral markers for glucose, such as the C_2-C_1-O_1 bending at 517 cm^{-1} and the C-O-H bending at 1125 cm^{-1} [26]. However, negative correlations to ethanol peaks are also observed, such as the C-C stretching peak at 879 cm^{-1} [27]. It should be noted that negative regression coefficients do not necessarily represent negative correlations, as negative regression coefficients could result from the mathematical constraints of PLS when peaks overlap [25]. Based on single compound spectra of the three main analytes, the 879 cm^{-1} peak is mostly free from glucose and biomass peaks. Thus, it can be concluded that glucose quantification is coupled to the 879 cm^{-1} ethanol peak, causing underprediction of glucose concentrations for the higher ethanol concentrations during the fed-batch fermentation. The regression coefficient of the ethanol model (Figure 2.2B) is positively correlated to ethanol characteristic spectral markers, such as C-C stretching, C-O stretching, and CH_3 rocking at 879 cm^{-1}, 1046 cm^{-1}, and 1084 cm^{-1}, respectively [27]. However, it also includes negative correlations to known glucose peaks (C_2-C_1-O_1 bending at 517 cm^{-1}, C-O-H bending at 1125 cm^{-1}). Similarly to the glucose base model, the correlations to glucose-specific peaks cause the ethanol base model to underpredict ethanol concentrations. Each time the glucose concentration was increased through bolus feeding, the underprediction of ethanol concentration increased (Figure 2.1B). The regression coefficient of the biomass base model (Figure 2.2C) is highly similar in structure to the ethanol base model (Figure 2.2B), indicating that biomass prediction is based on the decrease of glucose and increase of ethanol concentration. This is a result of the high cross-correlation between biomass and both ethanol and glucose in the calibration dataset, reflected by a R^2 of 0.992 and 0.988, respectively (Supplementary Figure 2.5.5.1).

The base models were quantitatively validated by getting low RMSEP values when applied to the unseen batch data (Table 2.1). However, qualitative assessment showed that the base models are heavily dependent on variations not related to the target analyte, but rather reflect batch process evolution, which is a result of maximizing the covariance between X and Y with an implicit modelling technique. Every biological process has inherent cross-correlations, and if calibration datasets are not constructed appropriately, these cross-correlations are integrated into the analyte quantification model. To extend the applicability of process data beyond the original process, target specificity needs to be assessed before model transfer. The lack of qualitative assessment may not only pose a risk when moving to different modes of operation, model performance may also be compromised when there is a deviation in one of the correlated analytes while running a similar process. Events such as a deviation in inoculation cell density, or ethanol carryover from the preculture to a bioreactor can all introduce errors, as a change in one analyte concentration directly affects the prediction of the other two compounds. Although, the challenges shown with the models in this work are of a specific case where only exponential growth phase data was used for calibration, the issues related to non-specificity in Raman model development for upstream bioprocesses have been reported on before [15, 18].

2.3.4 Impact of single compound spectra data supplementation

The decrease in base model performance when transferred from batch data to fed-batch data was due to (1) redistribution of ratios between the target compounds and (2) extrapolation from calibration ranges. This indicated a lack of target analyte specificity in the base model and prediction performance on the fed-batch process should increase when this limitation is overcome. The standard approach is to run one or more fed-batch fermentations and collect new in-line and reference data, that can be either used to train a specific fed-batch fermentation model or combined with the existing batch fermentation data. However, additional fermentation runs would require significant time and effort, thus leading to a delayed ability to monitor a new but related process. As a faster and less labor-intensive alternative, we propose model transfer from a batch fermentation to a fed-batch fermentation by including solely single compound spectra to the batch fermentation calibration dataset. This means that the resulting calibration dataset is a combination of process data and single compound spectra. The single compound data supplementation aims to extend the calibration ranges for the ethanol and biomass models, and to improve the specificity of all three models towards its analytical target. An overview of sample distributions of the different datasets is shown in Figure 2.3.

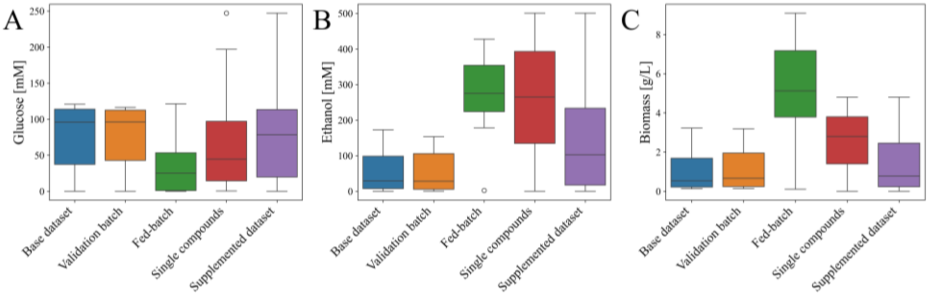


Figure 2.3: Sample distributions for (A) glucose in mM, (B) ethanol in mM, and (C) biomass in g/L for each of the datasets: base dataset, validation batch, fed-batch, single compounds, and the supplemented dataset.

The supplementary spectra were collected using the same 60-second acquisition time as during the fermentations. However, constant concentration conditions enabled the acquisition of multiple spectra per concentration point. For each point, ten 60-second measurements were averaged into one spectrum, resulting in high-quality smooth spectra with detector saturations identical to those observed during the fermentations. The new glucose calibration range was extended from 0-120.78 mM to 0-247.08 mM, with an increased sample density in the low concentration range (Figure 2.3A). The calibration range of ethanol was extended from 0-172.86 mM to 0-500.68 mM to remove the need of extrapolation outside of the calibration data (Figure 2.3B). Acquisition of single biomass spectra resulted in an extended calibration range from 3.23 g/L up to 4.81 g/L (Figure 3C). Due to technical difficulties not all fed-batch biomass concentrations were reached. Figure 2.3 shows an increase in sample density in the low glucose and high ethanol concentration ranges, but the data supplementation did not shift the mean concentrations to center of the calibration range. This is important as a mean near the sample distribution center indicates that the samples are properly distributed to prevent accuracy biases to specific concentration regions. The three base models were re-calibrated with the supplemented datasets, referred to as supplemented models. The performance of the supplemented models was evaluated with the fed-batch dataset, shown in Table 2.2, as well as the unseen batch data (Supplementary Table 2.5.6.1) to show maintained prediction accuracy for the original process.

Table 2.2: Overview of statistics for the supplemented models for glucose, ethanol, and biomass when applied to the fed-batch data.. The last column shows the improvement in rRMSEP of the supplemented model over the base model for the fed-batch data.

Model target	Calibration range	RMSEC	RMSECV	RMSEP	rRMSEP	Latent Variables	rRMSEP Base model	rRMSEP Improvement
Glucose	0 – 247.08 mM	3.05 mM	3.39 mM	3.06 mM	5.25 %	2	30.38 %	82.72%
Ethanol	0 – 500.68mM	8.02 mM	8.14 mM	8.65 mM	6.17 %	2	62.06 %	90.05%
Biomass	0.10 – 4.81 g/L	0.18 g/L	0.24 g/L	0.99 g/L	26.98 %	3	87.76 %	69.26%

The number of latent variables for the biomass supplemented model increased to 3, based on the RMSEC vs RMSECV graphs (Supplementary Figure 2.5.4.2). The predictive performance of all supplemented models on the fed-batch data increased compared to the base models, reflected by an 82.72%, 90.05%, and 69.26% rRMSEP decrease for glucose, ethanol, and biomass, respectively. The performance of the glucose and ethanol supplemented models is sufficient for accurate monitoring, as the rRMSEP values were close to 5%. The rRMSEP of the biomass supplemented model was 26.98%, and only gives an approximation of the biomass concentration. The biomass supplemented model showed increased values for rRMSEC (from 4.45% to 7.85%), rRMSECV (from 6.99% to 10.47%), and rRMSEP (from 5.38% to 8.32%) when applied to the batch validation dataset, meaning that supplementation and improved fed-batch performance came at the cost of batch prediction accuracy, and led to a more complex model (from 2 to 3 latent variables) (Supplementary Table 2.5.6.2).

For glucose and ethanol, data supplementation resulted in improved performance on the fed-batch cultivation while maintaining similar rRMSEC and rRMSECV values and performance on the validation batch. The decreased RMSEP, improved sample distribution for glucose, and the broader calibration range for ethanol successfully extended the applicability of the models, without compromising on the prediction performance on the original batch validation set. In addition to quantitative validation of the supplemented models, a qualitative assessment using regression vectors was performed to assess the impact on model specificity toward the target analytes (Figure 2.4A, Figure 2.4C, Figure 2.4E). The corresponding measured versus predicted plots of the supplemented models applied to the fed-batch data are shown in Figure 2.4B, Figure 2.4D, and Figure 2.4F.

The noise in all regression coefficients was reduced by data supplementation with single compound spectra. The coefficients of the glucose supplemented model (Figure 2.4A) show decreased dependency on the major ethanol peak at 879 cm^{-1}, but the correlation was not completely removed. In addition, the glucose supplemented model showed a decrease in the magnitude of the negative regression coefficients around 1452 cm^{-1}. Ethanol has a strong peak at 1455 cm^{-1} belonging to the asymmetric deformation of CH_3, which overlaps with glucose peaks in the same region. The glucose supplemented model corrects for this overlap by assigning negative regression coefficients to the ethanol peak to prevent an additive effect of these peaks and overestimation of the concentration [25]. For the glucose base model these negative regression coefficients were stronger due to the cross-correlation to the ethanol peak at 1455 cm^{-1}. Figure 2.4C shows that single compound samples had high leverage on the ethanol supplemented model as the regression coefficient vector closely represents the ethanol single compound spectra (Figure 2.2B). Although specificity of the model has increased, the current regression coefficient vector does not extensively compensate for overlapping glucose peaks. This could lead to possible overestimation of ethanol concentrations in situations where the concentrations of both ethanol and glucose are high, due to overlapping peaks in the 1000-1150 cm^{-1} and 1400-1500 cm^{-1} regions.

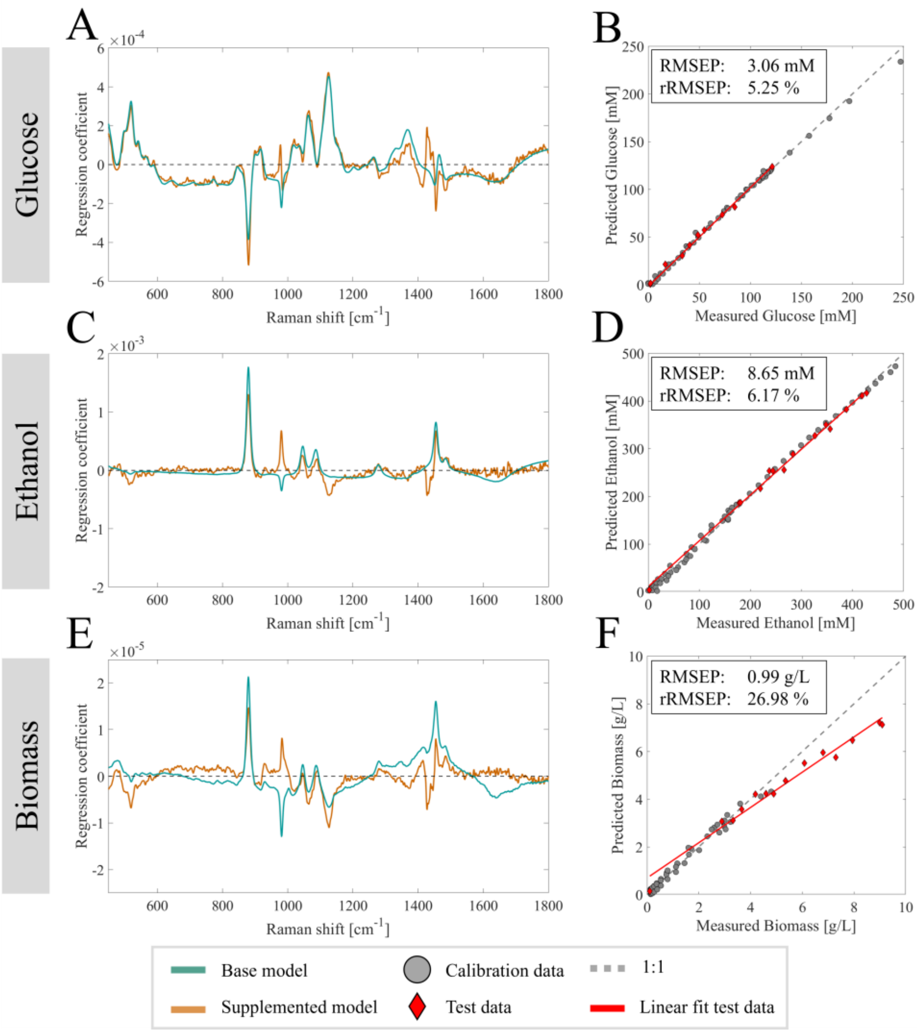


Figure 2.4: Regression coefficient vectors of the base model (cyan), supplemented model (orange), and the measured versus predicted plots of the supplemented models on the fed-batch dataset with the calibration data (grey) and fed-batch data (red) for glucose (A & B), ethanol (C & D), and biomass (E & F).

The biomass supplemented model underpredicts the actual biomass concentrations, suggesting there still is a dependency on non-target related peaks. The single compound spectra for biomass leveraged the model to assign more weights to the 1300-1500 cm^{-1} range where the single compound spectra for biomass displayed a baseline increase in the fingerprint region located between 1200 and 1600 cm^{-1} for higher biomass concentrations (Figure 2.2C). However, the regression coefficients are not specific enough for predictions independent of glucose and ethanol peaks. During the batch fermentations a non-linear signal extinction over the fingerprint region was observed, contradictory to the ascending baseline for single compound biomass spectra (Figure 2.2C). Iversen observed a similar non-uniform signal extinction with increasing yeast biomass during fermentation, and explained the effect to be caused by Lorenz-Mie scattering from the cell as particles [28]. It is unknown what caused the baseline increase in the single compound biomass spectra. At this point in time, we cannot fully exclude the impact of increased background fluorescence by media compounds which was used as measurement matrix, (decreased) cell culture viability, or possible leakage of cell contents. The yeast biomass used for these measurements was harvested after glucose depletion during batch fermentation, then washed and resuspended in clean synthetic media devoid of substrates and metabolites. Supernatant analysis with HPLC of samples taken during the biomass measurements showed no presence of glucose, ethanol, or other metabolites. While the consistent use of synthetic media maintained osmotic consistency, the extended depletion of substrates over the ~2 hour harvest and measurement procedure may have induced stress in the cells. More measurements of single yeast suspensions are required to gain knowledge on the spectral contribution of cells.

After evaluating the performance of the base models on the fed-batch data, it was observed that the decrease in performance was due to the redistribution of ratios between target compounds and the extrapolation beyond the original calibration ranges. Since PLS models assume a linear relationship between compound concentration and signal intensity, moderate prediction accuracy can be expected when the model extrapolates. It is therefore crucial that the models are first made robust against cross-correlations between process compounds before considering any calibration range extensions. Had the original base models been calibrated using batch process data with higher concentration ranges for all compounds, the fed-batch concentrations would have fallen within those ranges. However, the batch data would still contain strong cross-correlations between compounds, and application to the fed-batch where the ratios between these compounds change would still be problematic. Thus, addressing cross-correlation issues is key to developing more robust models, which provide a stronger foundation for extending the calibration range.

Qualitative assessment show that PLS models calibrated with process specific Raman spectra can be improved in terms of target specificity and quantification range by supplementing calibration datasets with single compound spectra. Supplementation with single compound spectra obtained in bioreactors offers a simple method for calibration dataset improvement without the need for extensive experimental designs and sample preparation, while maintaining process conditions such as temperature, stirring, and sparging. Moreover, increased model specificity results in the ability to transfer a base model beyond the original process it was trained for. Although the single compound supplementation successfully reduced cross-correlations in the calibration datasets, a next step could be to investigate solutions to efficiently generate mixture spectra which could disrupt the cross-correlations more efficiently. The observed discrepancy between single biomass spectra and process data, and the lack of knowledge on clear spectral markers for yeast biomass in literature, indicate that extensive studies on the individual effect of biomass on Raman spectra should be performed. More research on the spectral signals of cell density and viability will aid in developing specific biomass quantification models independent of substrate and product peaks.

2.4 Conclusion

Partial Least Squares (PLS) quantification models for Raman spectroscopy-based real-time fermentation monitoring are commonly calibrated with process data. However, fermentation processes have inherent cross-correlations leading to process specific models which do not transfer to related processes. Model calibration with highly cross-correlated data leads to prediction co-dependencies and standardly reported quantitative model validation, using metrics such as (r)RMSEP, does not guarantee model quality. Model specificity and spectral selectivity is essential for model robustness and should be evaluated during model calibration by investigating model statistics, such as the regression coefficients. Our approach of data supplementation with single compound spectra of the quantification target can expand calibration ranges, re-distribute a model's weights, and improve specificity to the relevant spectral markers. This extends the applicability of existing models to related processes, without the need of collecting new process data. We demonstrated this by adjusting base models calibrated on batch fermentation data, allowing transfer to a fed-batch mode of operation. This is represented by an rRMSEP improvement of 82.72%, 90.05%, and 69.26% for glucose, ethanol, and biomass respectively, leading to the absolute RMSEPs of 3.06 mM, 8.65 mM, and 0.99 g/L. Using single compound data spectra supplementation offers a fast and simple alternative to full model re-calibration, spiked samples integration, or extensive DoE approaches. Approaches like these can speed up the implementation and application of real-time monitoring with Raman spectroscopy, and thereby aid to early process monitoring and efficient process development.

Acknowledgements

We would like to thank Christiaan Mooiman for the technical support. This project is funded by the Department of Biotechnology at Delft University of Technology.

References

1. O'Mara, P., et al., *Staying alive! Sensors used for monitoring cell health in bioreactors.* Talanta, 2018. **176**: p. 130-139.
2. FDA, *Guidance for industry: PAT—A framework for innovative pharmaceutical development, manufacturing, and quality assurance.* Food and Drug Administration, Rockville, MD, 2004.
3. Vasilescu, A., et al., *Progress in electrochemical (bio) sensors for monitoring wine production.* Chemosensors, 2019. **7**(4): p. 66.
4. Bergin, A., J. Carvell, and M. Butler, *Applications of bio-capacitance to cell culture manufacturing.* Biotechnology advances, 2022: p. 108048.
5. Wasalathanthri, D.P., et al., *Technology outlook for real-time quality attribute and process parameter monitoring in biopharmaceutical development—A review.* Biotechnology and Bioengineering, 2020. **117**(10): p. 3182-3198.
6. Rathore, A., R. Bhambure, and V. Ghare, *Process analytical technology (PAT) for biopharmaceutical products.* Analytical and bioanalytical chemistry, 2010. **398**(1): p. 137-154.
7. Lourenço, N., et al., *Bioreactor monitoring with spectroscopy and chemometrics: a review.* Analytical and bioanalytical chemistry, 2012. **404**(4): p. 1211-1237.
8. Hirsch, E., et al., *Inline noninvasive Raman monitoring and feedback control of glucose concentration during ethanol fermentation.* Biotechnology Progress, 2019. **35**(5): p. e2848.
9. Webster, T.A., et al., *Development of generic raman models for a GS-KOTM CHO platform process.* Biotechnology Progress, 2018. **34**(3): p. 730-737.
10. Esmonde-White, K.A., M. Cuellar, and I.R. Lewis, *The role of Raman spectroscopy in biopharmaceuticals from development to manufacturing.* Analytical and Bioanalytical Chemistry, 2021: p. 1-23.
11. Zavala-Ortiz, D.A., et al., *Comparison of partial least square, artificial neural network, and support vector regressions for real-time monitoring of CHO cell culture processes using in situ near-infrared spectroscopy.* Biotechnology and Bioengineering, 2022. **119**(2): p. 535-549.
12. Rafferty, C., et al., *Analysis of chemometric models applied to Raman spectroscopy for monitoring key metabolites of cell culture.* Biotechnology progress, 2020. **36**(4): p. e2977.
13. Kozma, B., A. Salgó, and S. Gergely, *Comparison of multivariate data analysis techniques to improve glucose concentration prediction in mammalian cell cultivations by Raman spectroscopy.* Journal of Pharmaceutical and Biomedical Analysis, 2018. **158**: p. 269-279.
14. Wold, S., M. Sjöström, and L. Eriksson, *PLS-regression: a basic tool of chemometrics.* Chemometrics and intelligent laboratory systems, 2001. **58**(2): p. 109-130.
15. Santos, R.M., et al., *Monitoring mAb cultivations with in-situ Raman spectroscopy: The influence of spectral selectivity on calibration models and industrial use as reliable PAT tool.* Biotechnology progress, 2018. **34**(3): p. 659-670.
16. André, S., et al., *Developing global regression models for metabolite concentration prediction regardless of cell line.* Biotechnology and bioengineering, 2017. **114**(11): p. 2550-2559.
17. Domján, J., et al., *Raman-based dynamic feeding strategies using real-time glucose concentration monitoring system during adalimumab producing CHO cell cultivation.* Biotechnology Progress, 2020. **36**(6): p. e3052.
18. Romann, P., et al., *Advancing Raman model calibration for perfusion bioprocesses using spiked harvest libraries.* Biotechnology Journal, 2022: p. 2200184.
19. Webster, T.A., et al., *Automated Raman feed-back control of multiple supplemental feeds to enable an intensified high inoculation density fed-batch platform process.* Bioprocess and Biosystems Engineering, 2023: p. 1-14.

20. Nijkamp, J.F., et al., *De novo sequencing, assembly and analysis of the genome of the laboratory strain Saccharomyces cerevisiae CEN. PK113-7D, a model for modern industrial biotechnology.* Microbial cell factories, 2012. **11**(1): p. 1-17.
21. Verduyn, C., et al., *Effect of benzoic acid on metabolic fluxes in yeasts: a continuous-culture study on the regulation of respiration and alcoholic fermentation.* Yeast, 1992. **8**(7): p. 501-517.
22. André, S., et al., *Mammalian cell culture monitoring using in situ spectroscopy: Is your method really optimised?* Biotechnology Progress, 2017. **33**(2): p. 308-316.
23. Martyna, A., et al., *Improving discrimination of Raman spectra by optimising preprocessing strategies on the basis of the ability to refine the relationship between variance components.* Chemometrics and Intelligent Laboratory Systems, 2020. **202**: p. 104029.
24. Balabin, R.M. and S.V. Smirnov, *Interpolation and extrapolation problems of multivariate regression in analytical chemistry: benchmarking the robustness on near-infrared (NIR) spectroscopy data.* Analyst, 2012. **137**(7): p. 1604-1610.
25. Seasholtz, M.B. and B.R. Kowalski, *Qualitative information from multivariate calibration models.* Applied spectroscopy, 1990. **44**(8): p. 1337-1348.
26. Dudek, M., et al., *Raman Optical Activity and Raman spectroscopy of carbohydrates in solution.* Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2019. **206**: p. 597-612.
27. Boyaci, I.H., et al., *A novel method for quantification of ethanol and methanol in distilled alcoholic beverages using Raman spectroscopy.* Journal of Raman Spectroscopy, 2012. **43**(8): p. 1171-1176.
28. Iversen, J.A., R.W. Berg, and B.K. Ahring, *Quantitative monitoring of yeast fermentation using Raman spectroscopy.* Analytical and bioanalytical chemistry, 2014. **406**(20): p. 4911-4919.

2.5 Supplementary

2.5.1 Fermentation time evolution graphs

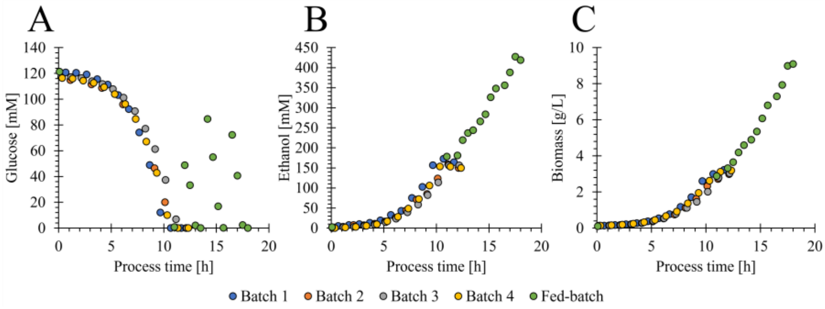


Figure 2.5.1.1: The concentrations of glucose (A), ethanol (B), and biomass (C) over process time for the calibration batches (Batch 1-3, blue, orange, and gray), validation batch (batch 4, yellow), and the fed-batch test dataset (green).

2.5.2 Overview of pure component measurement concentrations

Table 2.5.2.1: Concentrations of the supplementary measured samples for glucose, ethanol, and biomass.

Sample	Glucose [mM]	Ethanol [mM]	Biomass [g/L]
1	0.45	0.00	0.00
2	0.87	17.46	0.80
3	2.33	36.03	1.60
4	4.37	55.05	2.50
5	7.22	74.35	3.10
6	9.64	91.12	3.60
7	14.45	109.82	4.40
8	19.53	123.72	4.80
9	24.51	145.42	-
10	29.50	161.07	-
11	33.88	179.52	-
12	39.00	198.80	-
13	44.51	216.16	-
14	49.32	233.44	-
15	59.14	248.02	-
16	68.79	265.12	-
17	78.48	282.14	-
18	87.84	299.38	-
19	97.11	315.18	-
20	112.80	332.94	-
21	138.57	347.86	-
22	157.54	367.24	-
23	177.43	386.56	-
24	196.91	399.72	-
25	247.08	416.08	-
26	-	430.80	-
27	-	444.84	-
28	-	455.32	-
29	-	474.44	-
30	-	484.08	-
31	-	500.68	-

2.5.3 Influence of acquisition time on model accuracy

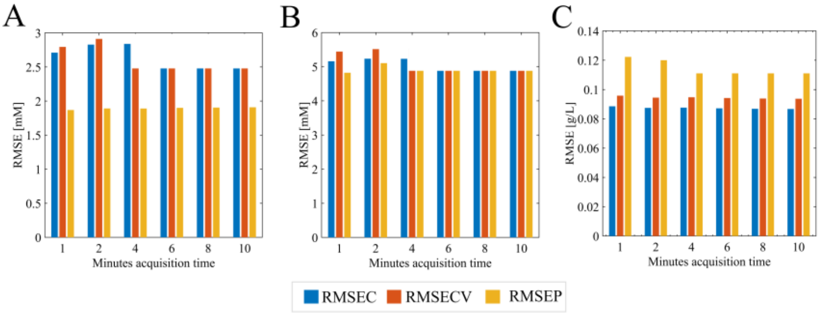


Figure 2.5.3.1: The performance for different acquisition times in minutes shown in root mean square errors of calibration (RMSEC), cross-validation (RMSECV), and prediction (RMSEP) for the base model calibration (Batch 1-3) and application to the batch validation dataset (Batch 4) of glucose (A), ethanol (B), and biomass (C).

2.5.4 RMSEC vs RMSECV plots for base and supplemented models

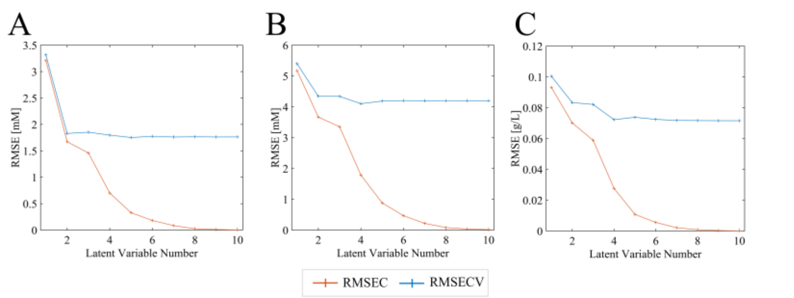


Figure 2.5.4.1: Elbow plots showing the root mean square error of calibration (RMSEC) and cross-validation (RMSECV) (x-axis) vs latent variables (y-axis) to select the number of latent variables for the base models of glucose (A), ethanol (B), and biomass (C).

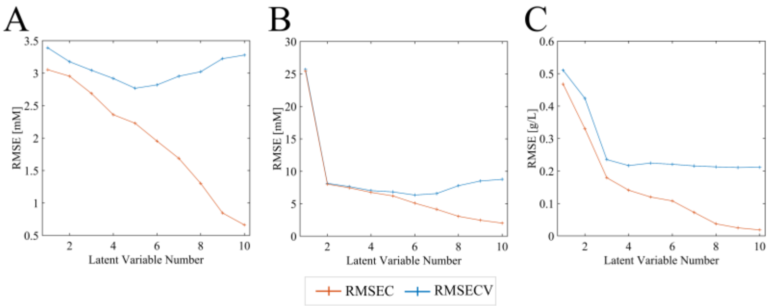


Figure 2.5.4.2: Elbow plots showing the root mean square error of calibration (RMSEC) and cross-validation (RMSECV) (x-axis) vs latent variables (y-axis) to select the number of latent variables for the supplemented models of glucose (A), ethanol (B), and biomass (C).

2.5.5 Cross-correlations between process compounds in the base and supplemented datasets

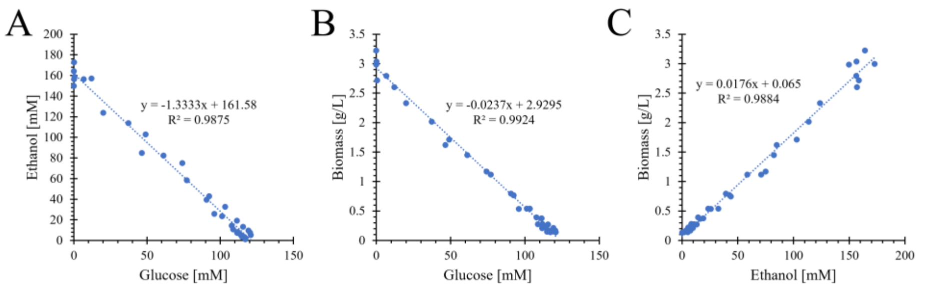


Figure 2.5.5.1: The cross-correlations in the base calibration dataset represented as glucose versus ethanol (A), glucose versus biomass (B), and ethanol versus biomass (C).

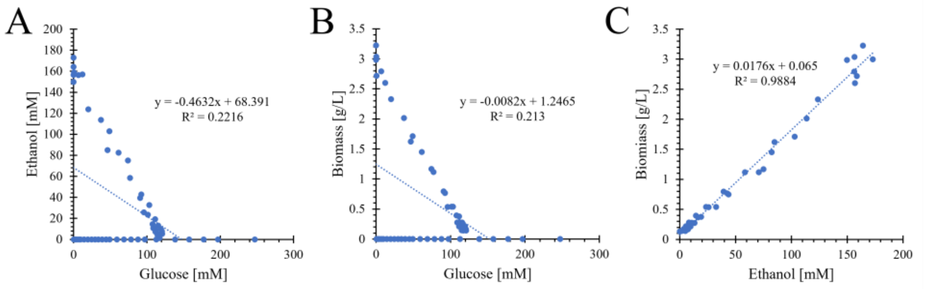


Figure 2.5.5.2: The cross-correlations in the glucose supplemented calibration dataset represented as glucose versus ethanol (A), glucose versus biomass (B), and ethanol versus biomass (C).

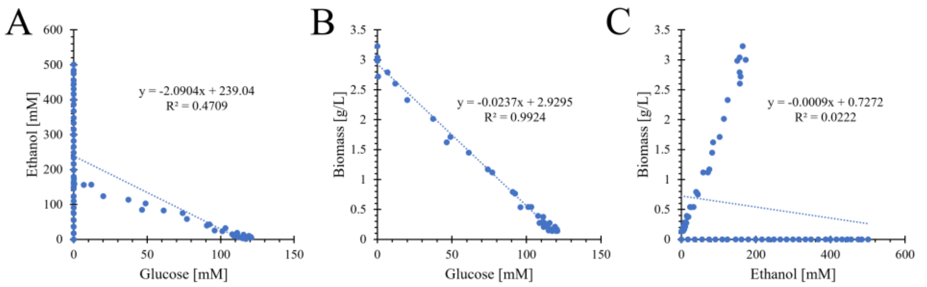


Figure 2.5.5.3: The cross-correlations in the ethanol supplemented calibration dataset represented as glucose versus ethanol (A), glucose versus biomass (B), and ethanol versus biomass (C).

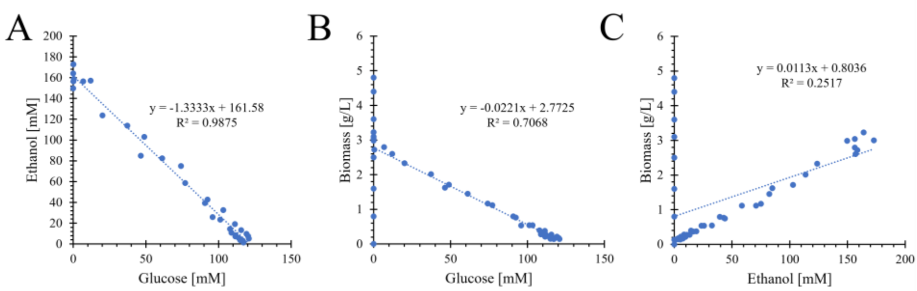


Figure 2.5.5.4: The cross-correlations in the biomass supplemented calibration dataset represented as glucose versus ethanol (A), glucose versus biomass (B), and ethanol versus biomass (C).

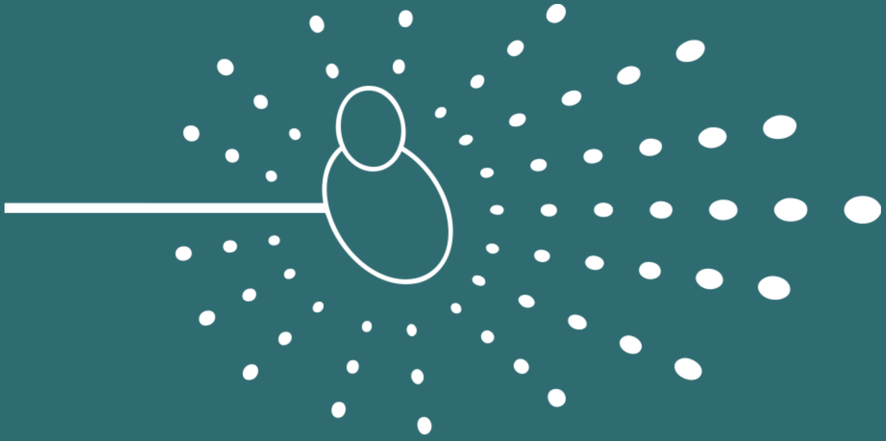
2.5.6 Performance of supplemented models on batch validation dataset

Table 2.5.6.1: Overview of statistics for the supplemented models applied to the validation batch and fed-batch datasets for glucose ethanol and biomass.

Model target	Calibration range	R2 Cal./CV/Pred. (batch)	RMSEC	RMSECV	RMSEP batch	RMSEP fed-batch	Latent Variables
Glucose	0 – 247.08 mM	0.997/0.996/0.999	3.05 mM	3.39 mM	1.71 mM	3.06 mM	2
Ethanol	0 – 513.28 mM	0.997/0.997/0.999	8.02 mM	8.14 mM	4.20 mM	8.65 mM	2
Biomass	0.10 – 4.81 g/L	0.981/0.967/0.995	0.18 g/L	0.24 g/L	0.17 g/L	0.99 g/L	3

Table 2.5.6.2: Comparison table of the relative root mean square errors of calibration (rRMSEC), cross-validation (rRMSECV) and prediction (rRMSEP) of the base and supplemented (suppl.) models. The errors were normalized to the interquartile ranges of the calibration dataset (for rRMSEC and rRMSECV) or the prediction dataset (for rRMSEP).

Model target	rRMSEC base	rRMSEC suppl.	rRMSECV base	rRMSECV suppl.	rRMSEP base	rRMSEP suppl.	rRMSEP base	rRSMEP suppl.
					Applied to batch		Applied to fed-batch	
Glucose	2.17 %	3.27 %	1.89 %	3.63 %	2.07 %	2.42 %	30.38 %	5.25 %
Ethanol	3.75 %	3.59 %	4.45 %	3.65 %	3.57 %	3.44 %	62.06 %	6.17 %
Biomass	4.45 %	7.85 %	6.99 %	10.47 %	5.38 %	8.32 %	87.76 %	26.98 %



Chapter 3

Impact of bioreactor process parameters and yeast biomass on Raman spectra

This chapter has been published as:

Klaverdijk, M., Nemati, M., Ottens, M., & Klijn, M. E. (2025). Impact of bioreactor process parameters and yeast biomass on Raman spectra. Biotechnology Progress, Article e70050 (https://doi.org/10.1002/btpr.70050)

Abstract

In-line Raman spectroscopy combined with chemometric modelling is a valuable process analytical technology (PAT) providing real-time quantitative information on cell culture compounds. Considering that compound quantification through chemometric models depends on pre-processing to maintain consistent changes in intensity at certain wavenumbers, all causes of signal distortion should be well understood to prevent quantification inaccuracies. This work investigated spectral distortion caused by the changing bioreactor parameters temperature, bubble quantity, and medium viscosity. In addition, the isolated spectral contribution of *Saccharomyces cerevisiae* cells in suspension was also determined. A temperature range from 20^0C to 40^0C resulted in peak shifts up to 0.8 cm^{-1} to lower wavenumbers, bubbles generated under standard bioreactor operation conditions led to signal attenuation of up to 7.93% reduction in peak intensity, and changes in liquid viscosity resulted in complex peak shift behavior. Isolated biomass concentrations reaching 5 g/L caused up to 44.6% reduction in distinct peak intensity, which was similar to spectra from batch process fermentations. Correcting for the attenuation revealed spectral features of biomass associated with proteins and lipids in the 1000-1500 cm^{-1} region. However, the spectral contribution of yeast biomass is dominated by signal extinction, which attenuates Raman spectra in a non-linear manner as biomass accumulates. The obtained knowledge on different sources of spectral distortion aids in the development of robust pre-processing and modeling strategies to obtain chemometric models applicable across experimental setups.

Keywords: *Raman spectroscopy, in-line measurements, Saccharomyces cerevisiae, spectral pre-processing, real-time monitoring, process analytical technology*

3.1 Introduction

Monitoring the progression of substrate, product, and biomass concentration in bioprocesses is traditionally measured off-line. This means that a physical sample is manually taken from the system and analysed by using an external standalone machine. Off-line measurements are labour-intensive and severely limit the real-time resolution of monitoring a process. Furthermore, manual sampling is invasive and increases the risk of altering or contaminating the process. The process analytical technology (PAT) framework published in 2004 [1] addressed these issues and aims to improve process understanding and control through real-time monitoring strategies. The PAT initiative and eased regulatory acceptance towards new analytical tools led to an increase in research and application of real-time monitoring strategies [2].

Optical spectroscopy PAT offers non-invasive and non-destructive measurement methods and can be placed in-line to the bioreactor allowing direct measurements, thereby lowering the risk of contamination [3]. Of the available spectroscopic tools, Raman spectroscopy is especially suitable for bioreactor environments due to the low signal interference from water. It also enables measurements through glass interfaces, facilitating easy integration with closed systems [4]. Raman spectroscopy is based on measuring inelastically scattered light coming from the interaction between monochromatic light and molecules in the sample. Each molecule provides a unique pattern of peaks based on the vibrational modes in the molecule's structure (Raman fingerprint), making Raman a powerful analytical tool for analysing specific targets of interest. Furthermore, there is a linear correlation between Raman signal intensity and molecule concentration, which allows for compound quantification after signal calibration [5]. Raman spectroscopy can be applied in-line, on-line, or at-line for the automated monitoring of bioreactor processes, using immersion probes, flow-cells, or measurements through glass windows [6, 7]. When applying in-line Raman spectroscopy in cell cultures through immersion probes, the obtained spectra contain contributions from all compounds in the medium. This allows for multiplexed monitoring with a single probe, but this also makes the signal highly complex. To address this complexity, multivariate modelling approaches based on dimensionality reduction and linear regression are commonly used to extract relevant spectral features for a target molecule. These models are typically calibrated with a dataset of Raman spectra capturing variations in the process, combined with quantitative reference measurements of the compound of interest. A calibrated model translates new spectral data into quantitative predictions for the compound of interest by applying learned weights to the spectral features associated with the compound. Through this approach, a single signal can be processed by multiple models to quantify various compounds simultaneously. Reported studies show that similar spectral data pre-processing strategies are often used to establish models for different compounds [8].

The aim of data pre-processing is to correct for signal distortion to ensure that model calibration and prediction is based solely on the variations linearly correlated with the concentration of the molecular compounds of interest [9]. However, data pre-processing steps should be applied with caution as excessive adjustments may introduce undesired spectral artifacts. It is therefore essential to have a good understanding of the multiple sources of signal distortion during bioreactor processes. Spectral distortion in Raman spectra can occur in three categories: baseline shifts, peak shifts over the wavenumber axis, and peak intensity changes not related to concentration changes. All these types of distortion can potentially decrease quantitative model performance by disrupting signal linearity and the alignment of model weights to the appropriate wavenumber variables. Raman spectra acquired in-line in bioreactor systems are subject to multiple sources of signal distortion and noise that can affect the spectra negatively. Strong background fluorescence from compounds in the sample can potentially mask signal of interest, making it challenging to detect and analyse subtle spectral changes [10]. In addition, particles such as cells, cell debris, and bubbles with sizes near the excitation laser wavelength can induce undesired optical scattering effects, leading to Raman signal attenuation or distortion [11]. These effects become more pronounced as a cell culture progresses, as increasing biomass concentrations and the accumulation of fluorescent compounds further amplifying scattering and fluorescence effects [12-14]. Whereas molecules and particles significantly smaller than the excitation wavelength ($\pi d/\lambda \ll 1$) scatter light isotropically according to Rayleigh scattering, larger particles ($\pi d/\lambda \sim 1, \pi d/\lambda \gg 1$) can scatter light asymmetrically, and predominantly in the forward direction [11]. The increased forward scattering and multiple scattering events by particles can lead to a reduction in the measured signal by in-line probes that use backscattering as measurement mode. While the heterogeneous composition of yeast biomass should give rise to spectral bands associated with proteins, lipids, and nucleic acids [15], in-line Raman spectra of yeast suspension cultures are often dominated by non-linear spectral extinction [16]. The yeast *Saccharomyces cerevisiae* has an average single cell diameter of 8 µm [17], significantly larger than the laser excitation wavelengths used by the Raman spectroscope (typically in between 532nm to 1064nm). This means that the cells can attenuate spectral features of other process compounds by causing anisotropic light scattering. The dominance of the resulting extinction effect complicates quantitative modelling approaches, and the spectral fingerprint related to the yeast cell's molecular composition has not yet been observed during in-line measurements. To remove extinction effects caused by biomass to ensure high quantification model performance for other molecular compounds in the bioreactor, stable peaks from medium sulphate and water have been used as internal normalization standards. This approach is based on the assumption that sulphate or water have a constant concentration and that the presence of biomass leads to a uniform signal extinction over the spectral fingerprint region [18, 19]. In other cases, fluorescence and scattering effects are corrected using spectral normalization steps in combination with Whittaker baseline corrections that fit a smooth baseline using penalized least squares, followed by baseline subtraction [4, 20]. Within bioreactor processes, optical scattering effects are not limited to cells, as bubbles formed in the bioreactor can reach sizes at which anisotropic light scattering occurs [21]. It is essential to understand spectral attenuation caused by bubbles as recent efforts in Raman spectroscopy model calibration involve miniaturization, flow-cell measurements, and high-throughput data collection. Some of these systems do not introduced bubbles in the matrix, which may lead to poor model transferability to bioreactor applications.

In addition to baseline distortions and signal extinction, changes in measurement conditions affecting molecular vibrations can cause unwanted peak shifts and intensity changes. An increase in system temperature can lead to the lengthening of molecular bonds by thermal expansion, lowering the vibrational frequency that causes spectral features to shift towards lower wavenumbers (red shift) [22]. Furthermore, an increase in temperature can decrease the population of molecules occupying the ground vibrational state, reducing the measured Stokes scattering intensity [23]. The level of hydrogen bonding in the medium can also affect vibrational frequencies, and several process parameters can dictate hydrogen bonding between water and medium compounds (e.g., temperature, ionic strength, pH). The spectral baseline caused by the cell culture medium predominantly consists of contributions from water, with features belonging to restricted translational and librational movements in the low wavenumber region (< 800 cm^{-1}), the water HOH-bending peak at 1641 cm^{-1}, and the OH-stretching modes in the high wavenumber region (> 3000 cm^{-1}) [24]. The stability of this water baseline is dependent on multiple medium factors, such as the temperature [25], presence of hydrogen bonding compounds, and salt concentration [26]. Conformational changes of these water bands can cause misalignment in spectral baselines, and influence the efficiency of common data pre-processing strategies.

The position of key spectral features along the wavenumber axis must remain stable, as misalignment between the wavenumber variables and weights assigned by the model can lead to impaired predictive performance [27]. Peak shifts are especially problematic for sharp spectral features that comprise of only a few wavenumbers, where a small wavenumber shift can lead to large prediction errors. Small shifts can be less problematic for broad spectral features as the prediction is based on a wider span of variables, providing more robustness against misalignments. However, when broad features overlap with signal of other compounds, a model has to decompose the peak and correct for the overlap [28], again increasing the sensitivity to wavenumber shifts. Misalignment can also affect model performance through data pre-processing. For example, internal standard normalizations can be sensitive to wavenumber shifts due to the high dependency on a singular spectral feature, where a misalignment will lead to a skewed correction of the entire spectra. Derivative-based pre-processing steps amplify a spectrum's inflection points causing sharp minima and maxima in the derivative spectrum, and peak shifts were reported to increase prediction errors [29]. The importance of minimizing the impact of spectral misalignment is reflected by the availability of peak alignment methods correcting wavenumber shifts caused by changes in measurement conditions or differences between analytical instruments [30].

We need to understand how spectra are influenced by measurement conditions to build robust and accurate prediction models for in-line Raman spectroscopy. In this work, we systematically explore the individual influence of temperature, bubble quantities, viscosity, and *Saccharomyces cerevisiae* biomass on Raman spectra. The resulting spectra are analysed for baseline shifts, wavenumber shifts, and changing peak intensities to determine the spectral contributions of each parameter as well as the potential impact on data pre-processing strategies and quantitative model performance. These findings contribute to the understanding of spectral distortion in bioreactor-based cell cultures. This knowledge will support the development of more robust and accurate quantification models, and facilitate the use of spectral data obtained from different experimental setups and processes. Moreover, improving knowledge on the spectral contribution of yeast biomass provides the means to critically evaluate standard quantitative modelling approaches and the development of innovative biomass quantification methods.

3.2 Materials & Methods

3.2.1 Measurement setup

3.2.1.1 Bioreactor and controller setup

All experiments were performed in an Applikon Bio 2L bioreactor systems (Getinge, Sweden). To prevent light contamination in the Raman measurements, the bioreactor was encapsulated with a custom PVC light cover. The temperature was controlled by a Biostat B-Plus controller (Sartorius Stedim, Germany) while the stirring speed was controlled by an Applikon ADI 1032 Stirrer Controller P100 (Getinge, Sweden).

3.2.1.2 Raman signal acquisition

A Raman RXN2 analyzer (Endress + Hauser Inc., Switzerland) equipped with a 400 mW 785 nm laser was connected to the bioreactor via an RXN-10 optical fiber and bIO-Optic immersion probe. The immersion probe was mounted through the bioreactor head plate and submerged into the liquid, and sterilized by autoclaving along with the bioreactor when necessary. Spectra were collected over the range of 100-3400 cm^{-1} with a resolution of 4 cm^{-1}, and a 60-second acquisition time resulted in detector saturations between 30-58%. The Raman spectroscope was set to continuously collect 60-second spectra, and at least 10 spectra were acquired per measurement condition. This resulted in low noise measurements.

3.2.2 Bioreactor parameter experiments

3.2.2.1 Media solutions

Four different types of media were used to study the bioreactor parameter effects: 1) Distilled water, 2) Synthetic media, 3) Synthetic media with low glucose concentration ($c_{glucose}$ = 24.4 g/L), 4) Synthetic media with high glucose concentration ($c_{glucose}$ = 47.6 g/L). The synthetic media was prepared with distilled water, and contained 5 g/L (NH4)2SO4 (Merck, Darmstadt, Germany), 3 g/L KH2PO4 (Merck, Darmstadt, Germany), and 0.5 g/L MgSO4.7H2O (Honeywell, Seelze, Germany) [31]. The media was adjusted to pH 6.0 with 2M KOH (Merck Sigma, Darmstadt, Germany), and 50% w/v glucose solution was prepared with glucose monohydrate (Merck, Darmstadt, Germany) and added to the desired concentration. The synthetic medium was completed with 0.2 g/L Antifoam-C (BASF, Ludwigshafen, Germany). An additional 15% vol/vol glycerol (Merck Sigma, Darmstadt, Germany) solution was prepared for studying the effect of viscosity on the spectra.

3.2.2.2 Temperature control

Temperature effects on the Raman spectra were studied in the range of 20-40^{0}C with step sizes of 5^0 C. The temperature values were reached using the bioreactor thermostat and control from the Biostat B-plus bioreactor controller. Media mixing was performed by continuously stirring at 830 rpm without sparging. A maximum deviation of 0.1^0 C from setpoint was allowed during the measurements.

3.2.2.3 Sparging and bubble size control

The bioreactor stirrer was operated at 80, 500, 675, and 830 rpm to form different quantity of bubbles and bubble sizes without the need of sparging. Bubble formation started at 500 rpm, and increased for higher stirring speeds (Figure 3.5.2.1). During these measurements the temperature was kept constant at 30^0 C using the same setup as described in 3.2.2.2.

3.2.2.4 Viscosity control

The 15% v/v glycerol solution was used in combination with active temperature control to generate different viscosity levels in the bioreactor. The temperature of the glycerol solution was operated in the range of 20-45^0 C with a step size of 5^0 C. A Lovis 2000 M/ME viscometer (Anton Paar, Austria) was used to measure the viscosity of this solution at each temperature. During measurements, the bioreactor was mixed by stirring at 830 rpm without sparging. The viscosity reference measurements are provided in Table 3.5.3.1.

3.2.3 Fermentations and biomass measurements

3.2.3.1 Batch fermentation

Two batch fermentations were conducted with an inoculation cell density of 0.015 g/L. The yeast strain *Saccharomyces cerevisiae* CEN.PK113-7D was used for all cell culture experiments [32]. All cultures were grown on sterile synthetic media, (see 2.2.1) supplemented with 0.2 g/L sterile Antifoam-C (BASF, Ludwigshafen, Germany), and filter sterilized vitamin solution (in-house) after autoclaving. Glucose was used as carbon source, and an initial concentration of 40 g/L was reached by the addition of sterilized 50% glucose solution (see 2.2.1). Medium aeration was performed by stirring at 830 rpm and sparging air at 0.5 L/min. A pH of 6.0 was maintained by the addition of 2 M KOH, and the temperature was kept constant at 30^0 C. The working volume of the reactors was 1L and the fermentations were operated until glucose depletion was detected by the drop in off-gas CO_2 measured with a ServoPRO 4900 off-gas analyzer (Servomex, UK).

3.2.3.2 Reference sampling

Samples of the supernatants of the batches were analyzed for their ethanol and glucose concentrations using an Agilent 1260 Infinity HPLC (Agilent Technologies, USA). A BIO-RAD Aminex HPX-87H (300 x 7.8 mm) cation-exchange column (Bio-Rad, USA) operated at 60^0 C was used with a 0.5 g/L H2SO4 eluent at a flow rate of 0.6 ml/min. The injection volume was 5 µm, and an Agilent 1260 refractive-index and variable wavelength detector were used for sample characterization.

The cell density of the fermentations was determined off-line by optical density at a wavelength of 660 nm (OD660) using a Libra S11 spectrophotometer (Biochrom, UK). The dry biomass weight was measured by loading 10 mL of cell suspension on nitrocellulose membrane filters (pore size, 0.45 µm; Gelman Laboratory, USA), drying the filters in a microwave, and subsequently weighing the dry biomass on a precision scale (Mettler Toledo, USA).

3.2.3.3 Biomass harvest and measurement

After glucose depletion occurred in the batch fermentations, the bioreactor broth was harvested and the biomass was washed twice by centrifuging and resuspending the cells. The suspension was centrifuged at 5000 rpm for 5 minutes in 400 mL bottles using an Avanti J-E centrifuge (Beckman Coulter, USA) to pellet the cells. After discarding the supernatant, the cell pellet was resuspended in fresh synthetic media. After the second washing step, the biomass was resuspended in 350 mL of synthetic media to a final concentration of 15 g/L. The washed cell suspension was added in steps of 25 mL to a bioreactor with 700 mL of clean synthetic media to measure concentration ranges of biomass. The Raman measurements were performed while maintaining 830 rpm of stirring, sparging with 0.5 L/min of air, and temperature control at 30^0 C. During the measurements samples were taken to determine the cell viability using a NucleoCounter NC-202 (Chemometec, Denmark).

3.2.4 Data analysis

3.2.4.1 Spectral processing

Spectral pre-processing was performed in PLS_Toolbox (v 9.3, Eigenvector Research, WA) running on MATLAB 2023a (MathWorks, WA), and Python 3.10 using libraries from scikit-learn (https://scikit-learn.org [33]) and Chemotools (https://paucablop.github.io/chemotools/ [34]). For each measurement condition 10 spectra of 1 minute were averaged to a single spectra by an in-house Python script to improve spectral quality and reduce noise.

3.2.4.2 Peak shift detection

Quantification of peak shifts was performed by determining the exact peak location. The location of spectral peaks was determined using the zero-crossing point after spectral derivatization, which coincides with the position of a peak. Peak position determination was performed using an in-house Python script and the workflow consisted of three steps. First, spectra of the bioreactor parameters and biomass measurements were reduced to the fingerprint region (350-1800 cm^{-1}) to minimize the leverage of the high spectral tails on the baseline correction. Second, AirPls baseline correction ($\lambda = 300$) was applied to remove spectral baselines, thereby reducing the influence of baselines on peak locations. Third, a second-order derivative (15-point window, 2nd order polynomial) was taken from the baseline-corrected spectra. Fourth, linear regression was applied between the datapoints before and after the zero-crossing of the second derivative spectra, and the exact crossing point was calculated to determine the peak position.

3.2.4.3 Signal extinction analysis

The decrease in peak intensity was measured to determine the level of signal extinction caused by bioreactor parameters and biomass. The spectra were reduced to the fingerprint region (350-1800 cm^{-1}) and pre-processed using the AirPls baseline correction ($\lambda = 300$), effectively removing the baseline while retaining the peak intensity in relation to the original baseline, as well as the relative differences between the individual measurements. After pre-processing the absolute peak intensity was used for analysis.

3.3 Results and discussion

Spectral distortion as a result of the bioreactor process parameters temperature, bubble quantity, and viscosity, as well as biomass concentration was investigated. The influence on the spectral baseline was determined by inspecting the full Raman spectrum, and changes in wavenumber shifts and peak intensities were quantitatively assessed using four pre-defined peaks: immersion probe window (406 cm-1), water HOH-bending (1641 cm^{-1}) [25], and when present in the mixture, the peaks of media sulphate (981 cm-1) and glucose COH-bending (1125 cm-1) [35]. Four measurement matrixes with known spectral features were used to perform these experiments: 1) distilled water, 2) synthetic media, 3) synthetic media with low glucose concentration (25 g/L), and 4) synthetic media with high glucose concentration (50 g/L). To study the effects of viscosity, a mixture of 15% (v/v) glycerol solution in water was measured under different temperatures to change the liquid viscosity without adjusting the absolute glycerol concentration.

3.3.1 Spectral effects of temperature

Changes in temperature can occur during process development or when spectral data is combined from different (miniaturized) experimental setups. Temperature influences the vibrational state populations of molecules and alters hydrogen bonding dynamics, both between water molecules and between water and other compounds in the medium. These changes can alter peak positions and intensities in a Raman spectrum. The spectral baseline of cell culture media is primarily influenced by contributions from water, and the resulting spectral changes in the water matrix due to temperature increases from 20⁰C to 40⁰C are shown in Figure 3.1. The baseline effects of temperature on the other matrices are shown in Supplementary Figure 3.5.1.1.

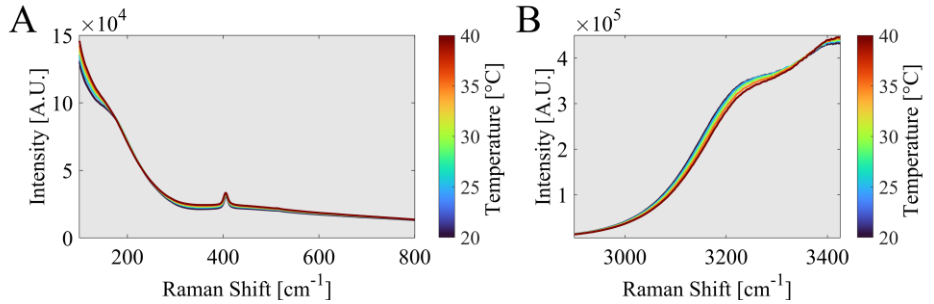
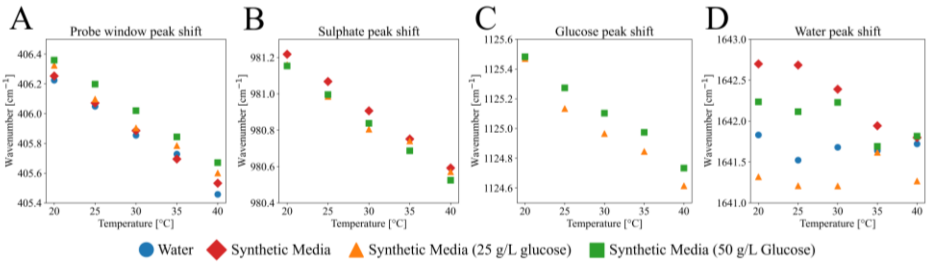


Figure 3.1: The low (A) and high (B) wavenumber region of water matrix for a temperature range from 20 to 40⁰C (blue to red in the colour bar). The displayed spectra are the average of 10 individual 60-second spectra acquired at each temperature.

The largest spectral baseline changes occurred in the broad bands related to the vibrational modes of water in the low (< 800 cm^{-1}) and high (> 3000 cm^{-1}) wavenumber regions. The baseline below 200 cm^{-1} (Figure 3.1A) showed an increase in intensity at higher temperatures as a result of the increasing band at 60 cm^{-1} just outside of the spectroscopes range (<100 cm^{-1}), which belongs to restricted translational movements of water molecules [36, 37]. Furthermore, the baseline from approximately 300 cm^{-1} to 800 cm^{-1} also increases with higher temperatures as the underlying broad features associated with the librational modes of water become more prominent due to the reduced hydrogen bonding and increased molecular motion [24]. The observed baseline changes in the low wavenumber region were similar in the synthetic media matrices (Supplementary Figure 3.5.1.1). The bands in the high wavenumber region (Figure 3.1B) originate from OH-stretching vibrations of water, and although a large section of these peaks is outside of this spectroscope's measurement range (>3425 cm^{-1}), some of the conformational changes are still visible between 3000-3425 cm-1. For example, higher temperatures weaken hydrogen bonds between water molecules, reducing the formation of large water molecule clusters, which typically decreases the intensity of the strong hydrogen bonding band at 3200 cm-1 and increases or broadens the intensity of the weaker hydrogen bonding band around 3400-3600 cm-1 [24, 38]. For all four matrixes, a decrease in the band at 3200 cm-1 was observed (Supplementary Figure 3.5.1.1), while the visible part of the band at 3400 cm-1 increased in water and decreased in the synthetic media matrices. This difference is most likely caused by the dissolved salts in the synthetic medium, which can disrupt the hydrogen bonding behaviour of water molecules [39]. The conformational change between these two OH-stretching bands in water matches examples in the field of Raman thermometry, where the ratio between these bands is used as an indicator for system temperature [40]. However, as the majority of this second band was outside of the detectable range for this spectroscope, further analysis was not performed on this region.



E

Measurement Matrix	**Peak intensity decrease over temperature range**			
	Probe window	**Sulphate**	**Glucose**	**Water**
Water	1.89%			5.38%
Synthetic Media	2.61%	6.57%		1.28%
Synthetic Media (25 g/L glucose)	2.54%	4.57%	7.07%	1.53%
Synthetic Media (50 g/L glucose)	2.29%	2.23%	2.23%	2.03%
Average	2.33%	4.46%	4.65%	2.56%

Figure 3.2: The effect of temperature on the peak locations of probe window (A), media sulphate (B), glucose (C), water HOH-bending (D), and the total intensity decrease per peak and measurement matrix (E). The peak location was derived by pre-processing the spectra with AirPls baseline correction (lambda = 300), taking a Savitzky Golay derivative (1st order, 15-point window length), and interpolating the zero-crossing point of the x-axis.

Four peaks were quantitatively analysed for wavenumber shifts caused by temperature and the results are displayed in Figure 3.2. An average peak shift across the four different mixtures of 0.73 cm-1, 0.61 cm-1, and 0.80 cm-1 was found for the immersion probe glass, sulphate, and glucose peak, respectively, while the water HOH-bending peak at approximately 1641 cm-1 showed a less consistent shift pattern. A peak shift to lower wavenumbers can occur due to thermal expansion at higher temperatures, causing increasing bond lengths and the disruption of hydrogen bonds. Both of these factors can reduce the energy difference between the ground and elevated vibrational states, thereby affecting the peak position. In AirPls baseline corrected spectra, the temperature increase led to an average relative intensity decrease of 2.33%, 4.46%, 4.65%, and 2.56% for the probe window, sulphate, glucose, and water peak across all measurement matrixes (Figure 3.2E).

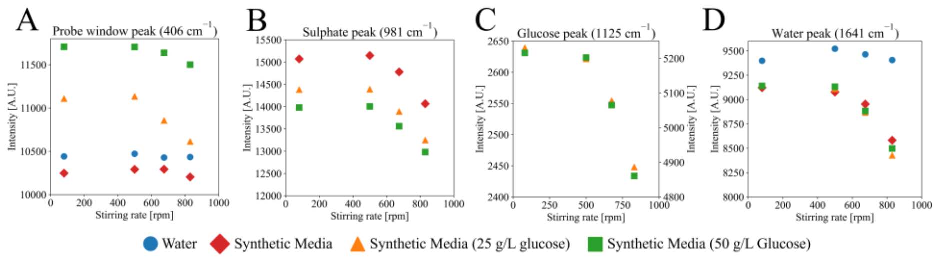
Figure 3.2A shows that the linear trajectory of the probe window peak shifts at 406 cm^{-1} is similar for each matrix, where the mean peak shift per 5^0C step was 0.18 cm^{-1} with a standard deviation of 0.03 cm^{-1} across all measurement matrixes. Peak shifts when measuring crystal structures are commonly studied over wider temperature ranges, and the observations on the probe window glass match in order of magnitude with examples of Raman sapphire crystal studies in literature (~0.41 cm^{-1} per 20 °C) [41]. For the two matrices containing glucose it was observed that the probe window peak is identified at slightly higher wavenumbers compared to the matrices without glucose. Glucose has a broad CCC-bending peak at 432 cm^{-1} [35] that partly overlaps with the 406 cm^{-1} probe window peak, thereby slightly affecting the peak position in the 25 g/L and 50 g/L glucose synthetic media samples. The average peak shift for sulphate (0.61 cm^{-1} at 981 cm^{-1}) is comparable to observations under similar conditions in literature (~0.30-0.70 cm^{-1}), where it was also found to be dependent on the salt concentration [42, 43]. The average glucose peak shift (0.80 cm^{-1} at 1125 cm^{-1}) could not be directly verified with existing literature under similar measurement conditions. The changing peak positions of sulphate and glucose can be the result of changes in hydrogen bonding, while for sulphate, ionization effects in the media can also play a role when temperature changes. The water peak at approximately 1641 cm^{-1} shown in Figure 3.2D belongs to the HOH-bending vibrational mode. The HOH-bending is more confined to individual water molecules and it should therefore be less affected by hydrogen bonding [25]. The results do not show a consistent pattern in the wavenumber shift and the detected decreases in peak intensities were relatively small (2.56% on average). A possible explanation for the decrease in measured peak intensities relative to the baseline is the increase of molecules in higher vibrational states when temperature increases. As the Raman spectroscope used in this work measures Stokes scattering, which primarily occurs when molecules are in the ground vibrational state, a shift of molecules to higher vibrational states could decrease the measured signal intensity.

The experimental data showed baseline and conformational changes mainly in the low and high wavenumber regions as a result of temperature changes. For the application of Raman spectra for quantitative real-time monitoring, spectra are typically reduced to the fingerprint region (around 350 to 1800 cm^{-1}) as a part of data pre-processing, sometimes including the region in-between 2800 to 3000 cm^{-1}, as the excluded regions contain little chemical information on biological molecules. Therefore, the expectation is that temperature induced changes observed at the tails of the spectra (<300 cm^{-1} and >3000 cm^{-1}) will not influence model building unless the full spectra are used for analysis. Thus, if the full spectrum is subjected to pre-processing involving polynomial fitting, the intensity differences in the spectral tails can strongly influence baseline and scatter corrections by exerting high leverage on the polynomial fit. As a consequence, improper baseline alignment may occur in key spectral areas such as the fingerprint region, and thereby impact the accuracy of compound quantification using chemometric models.

The observed peak shifts and intensity changes as a result of changing temperature pose a far larger challenge for model building, as these phenomena can directly affect the linear correlation between signal intensity and molecule concentration, and can lead to spectral misalignments between peaks and model coefficients. In general there are few bioreactor or cell culture processes that see a temperature shift of more than 5^0 C (leading to an average shift of 0.2 cm^{-1} in the glucose peak). However, peak shifts can be problematic in model calibration approaches where data from different experimental sources with varying levels of temperature control are combined (e.g., bioreactor, glassware, flow-cells, and alternative miniaturized setups) or during process development. Calibrating a model with samples at room temperature and subsequently applying it to cell culture conditions may lead to poor predictive performance as a result of spectral feature misalignment and peak intensity differences. A potential mitigation strategy is the use of spectral alignment techniques to correct feature positions between different measurement conditions [30]. In applications where multiple spectrometers are used across different facilities, these methods are already implemented to correct for variations between instruments [44]. Peak alignment approaches, combined with methods to account for normalizing the changes in peak intensities, could effectively standardize the spectra and facilitate the fusion of data obtained from various sources and at different process conditions.

3.3.2 Spectral effects of bubbles

Both bubbles and cells attenuate Raman spectra by acting as light scattering particles [16, 21]. To isolate the spectral contribution of biomass that can be used as input for biomass quantification models, it is key to understand the light attenuating effects of bubbles. To this end, bioreactors were operated at different impeller speeds (80, 500, 675, and 830 rpm) to generate bubbles of different sizes and quantities within the four matrices. We observed bubble formation around the baffles starting at 500 rpm, which was therefore chosen as the second step after the 80 rpm baseline condition. Bubble formation increased with higher rpms regardless of the mixture composition (Supplementary Figure 3.5.2.1). This section focuses on peak intensity changes to assess the impact of bubbles, as a peak shift analysis did not result in peak shifts larger than 0.11 cm^{-1} across different impeller speeds (Supplementary Figure 3.5.2.2). Although the water HOH-bending peak showed peak shifts of up to 0.67 cm^{-1}, the inconsistent patterns were attributed to noise and therefore considered to be independent of the bubbles. Figure 3.3 shows the impact of bubble formation on the absolute peak intensity for the probe window, sulphate, water HOH-bending, and glucose peak. The intensity decreases per peak and measurement matrix for increasing stirring rates are summarized in Figure 3.3E.



E

Measurement Matrix	Intensity decrease over RPM range			
	Probe window 406 cm^{-1}	Sulphate 981 cm^{-1}	Glucose 1125 cm^{-1}	Water 1641 cm^{-1}
Water	0.42%			1.30%
Synthetic Media	0.87%	7.14%		5.91%
Synthetic Media (25 g/L glucose)	4.68%	7.93%	7.25%	7.81%
Synthetic Media (50 g/L glucose)	1.77%	7.31%	7.31%	7.02%

Figure 3.3: The effect of bubble quantity caused by increasing stirring speeds on the peak intensities for the probe window (A), media sulphate (B), glucose with low concentration on the left y-axis and high concentration on the right y-axis (C), water HOH bending (D), and the total intensity decrease per peak and measurement matrix due to bubble quantity increase (E). The datapoints were acquired by taking the absolute peak intensities at the indicated wavenumbers after pre-processing spectra with AirPls baseline correction (lambda = 300).

Interestingly, extinction of 0.87% to 4.68% for the peak of the immersion probe (406 cm-1) was observed when measuring the synthetic media matrixes. This peak should not be affected by bubbles behind the window, as seen for water and synthetic media without glucose, because it arises from the immersion probe itself. For the synthetic media matrices with glucose, this observed intensity change might be due to extinction of the broad glucose CCC-bending peak at 432 cm^{-1} by the bubbles, which partly overlaps with the probe window peak at 406 cm^{-1} [35], similar to the results in the temperature change experiment. The extinction of the sulphate, glucose, and water peaks was 6-8% in the synthetic media measurements, while the concentration of these compounds remained constant. For the water matrix, the probe window and water HOH-bending peaks only showed a 0.42% and 1.30% decrease in peak intensity, respectively. This aligns with the visual inspection of the overall spectral changes in water, which are small compared to those in synthetic media (Supplementary Figure 3.5.2.3). The synthetic media matrices contained high salt concentrations and antifoam-C that can impact the bubble formation while stirring. High salt concentrations increase surface tension and can reduce bubble coalescence, and thereby increase the stability of small bubbles when compared to water [45]. Antifoam-C works as a defoaming agent by lowering surface tension and is added to cell culture medium to displace bubble stabilizing surfactants. Despite the presence of antifoam, it was visually observed that bubble formation increased drastically in the synthetic media matrices when compared to water, leading to a higher bubble density during the Raman measurements and thus greater signal attenuation. The bubble size and quantity was not measured during these experiments, and there is only visual proof of the increasing bubble quantity and decreasing bubble size (Supplementary Figure 3.5.2.1). The exact bubble and particle size distribution during bioreactor processes are challenging to estimate and measure, and is an active field of research [46].

The reported measurements here show that the spectral extinction caused by increasing bubble quantities and decreasing bubble size are significant (up to 8%), especially considering the constant concentration of compounds in the samples during the measurements. This means that the observed peak extinction can disrupt the proportionality between signal intensity and molecule concentration, and that bubble size and quantity should be taken into account as signal attenuating particles during model development. Furthermore, the production of fermentation compounds, such as ethanol, glycerol, or extracellular proteins, can also affect the surface tension and thereby bubble coalescence [45]. The combination of these changing factors over the course of a fermentation make bubble formation dynamic factor during a single process. In terms of high-throughput model calibration for quantitative monitoring, differences in signal intensity between spectra obtained in bioreactor settings and (miniaturized) high-throughput setups may arise due to the absence of bubbles (e.g., as seen for flow-cells and sample chambers). For example, the sulphate peak that is often used as a normalization reference [18] seems to suffers from similar levels of signal extinction compared to the glucose and water peak. As the peak shift was minimal for these spectra and the signal extinction appears uniformly distributed over the fingerprint peaks, sulphate normalization would lead to appropriate correction of signal extinction caused by bubbles. In cases where the extinction is not uniformly distributed, multiplicative scatter corrections such as extended multiplicative scatter correction (EMSC) would be more appropriate. In both cases, the normalization step should be critically evaluated to ensure that spectra from different experimental setups are adjusted equally to make signal intensities representative of the environment in which the models will be applied.

3.3.3 Effects of viscosity

Cell culture media viscosity is influenced by changing concentrations of substrates and products, which interact with water through hydrogen bonding, as well as the accumulation of biomass that increases viscoelastic properties of the medium. To study the spectral changes caused by viscosity, a 15% v/v glycerol solution was measured under temperatures ranging from 20 to 45^{0}C with increments of 5^{0}C. Increasing the temperature leads to the dissociation of hydrogen bonds between glycerol and water, and allowed a viscosity range of 1.55 to 1.03 mPa.s, thereby simulating a yeast fermentation of up to 55 g/L of yeast biomass [47] (the measured viscosities are shown in Supplementary Table 3.5.3.1). This approach was selected to allow for viscosity adjustments without changing the compound (glycerol) concentrations. Several peaks belonging to glycerol vibrational modes were investigated for wavenumber shifts and intensity changes. The fingerprint regions of the acquired spectra are shown in Figure 3.4 (peak shifts are shown in Supplementary Figure 3.5.3.1).

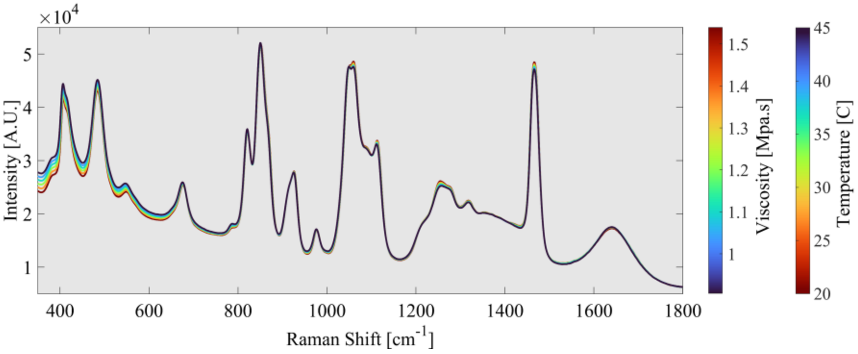


Figure 3.4: The Raman spectra fingerprint region of 15% v/v glycerol solution measured at a viscosity ranging from 1.55 to 1.03 mPa.s (second colourbar) obtained by increasing the temperature from 20 to 45^{0}C in increments of 5^{0}C (first colourbar). The displayed spectra are the average of 10 individual 60-second spectra acquired at each temperature step.

As a changing temperature setpoint was required to control the viscosity of the 15% v/v glycerol solution, the data does not allow to evaluate the effects of temperature and viscosity independently. In the previous temperature measurements (section 3.1) an increasing temperature generally led to an increasing baseline intensity in the low wavenumber range (<800 cm^{-1}) and peak shifts towards lower wavenumbers (up to 0.8 cm^{-1}). Similar spectral changes were observed in the 15% v/v glycerol mixture, where the baseline in the low wavenumber region shifted up and the glycerol peaks shifted to lower wavenumbers. Glycerol fingerprint features, such as the two CC-stretch (821 and 851 cm^{-1}) and two CH_2-rock (925 and 977 cm^{-2}) peaks [48], moved to lower wavenumbers by 0.63 cm^{-1}, 0.95 cm^{-1}, 0.83 cm^{-1}, and 0.84 cm^{-1}, respectively. These peak shifts were comparable in magnitude to the shifts observed for other peaks analysed during the temperature experiments. The overlap of glycerol peaks in the 1000 cm^{-1} to 1350 cm^{-1} region complicate peak position determination due to influences of shouldering peaks. The peak shift determination method calculated a shift to lower wavenumbers by 2.48 cm^{-1} for the observed spectral features at 1050 cm^{-1} and 1060 cm^{-1} (Supplementary Figure 3.5.3.1). Additionally, a 1.50 cm^{-1} shift was calculated for the peak at 1111 cm^{-1} attributed to the CO-stretch vibrational mode involved in hydrogen bonding. However, the large overlap with the shoulder at 1090 cm^{-1} coming from a CH_2 rocking mode affects the accuracy of peak shift determination. The overall observed effects (hydrogen bond strength) match those seen during the changing temperature experiments and no additional effects of viscosity could be identified. It should be noted that viscosity is influenced by multiple factors beyond hydrogen bonding during fermentation such as ionic strength and cell density, where various chemical and physical properties of the medium may affect Raman spectral features differently.

3.3.4 Effects of yeast biomass

Yeast cells constitute the majority of the particulate matter during fermentation in a bioreactor. The *S. cerevisiae* cells used in this work have an average diameter of ~8 μm in their single cell form with an ellipsoidal shape [17], and at this size the cells can cause anisotropic light scattering ($\pi d/\lambda \gg 1$) when considering the excitation laser wavelength of 785nm. To determine the impact of anisotropic light scattering, two batch fermentations were operated to generate yeast cells for the measurement of isolated biomass, using a concentration range 0 to 5 g/L. The yeast cell viability was above 96% during all the washing steps and measurements, minimizing the contribution of different scattering effects and fluorescence as a result of dead cells and cell debris. The experiment was performed twice with biomass from two individual batch cultures to test for reproducibility (Supplementary Figure 3.5.4.1) and all resulting spectra of the isolated biomass are shown in Figure 3.5. Two high concentration biomass spectra were poorly corrected by the baseline correction (Figure 3.5C) and therefore removed for further analysis.

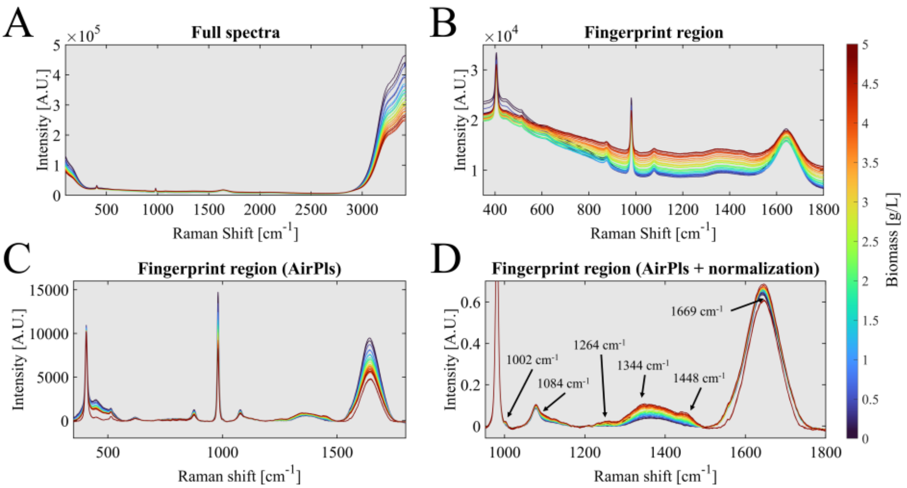


Figure 3.5: The full (A), fingerprint region (B), AirPls (lambda = 300) pre-processed fingerprint region (C), and AirPls (lambda = 300) followed by sulphate peak normalization pre-processed fingerprint region (D) of spectra acquired for 0 to 5 g/L isolated biomass (blue to red in the colourbar). Each displayed spectra is the average of 10 individual 60-second spectra acquired at a single concentration step.

Peak shift analysis resulted in an average wavenumber shift of 0.36 cm^{-1}, 0.29 cm^{-1}, and 1.90 cm^{-1} for the probe window, sulphate, and water HOH-bending peaks, respectively (Supplementary Figure 3.5.4.2). Increasing the biomass concentration was not expected to affect vibrational modes and cause peak shifts, and this is most likely due to the addition of underlying spectral features. The overall spectral changes caused by biomass did not directly show distinct peaks or spectral markers, as is typically observed for molecular compounds. The major observed spectral effects over the full spectral range were a strong intensity decrease in the low (<800 cm^{-1}) and high (>3000 cm^{-1}) wavenumber regions (Figure 3.5A), and a slight baseline increase caused by background fluorescence in the fingerprint region (350 cm^{-1} – 1800 cm^{-1}, Figure 3.5B). After correcting for this fluorescence baseline increase in the fingerprint region with an AirPls baseline correction, the impact of increasing biomass becomes visible. Figure 3.5C shows the extinction of other spectral features (sulphate peak 981 cm^{-1}, water peak 1641 cm^{-1}) relative to the baseline for increasing biomass concentration. Subsequent normalization to correct for the signal extinction using the sulphate peak [4] reveals and increase in intensity in the 1000-1500 cm^{-1} range for increasing biomass (Figure 3.5D). The intensity changes in this range linearly correlate with the biomass concentration (Supplementary Figure 3.5.4.3) and can be associated with the composition of *S. cerevisiae*, which is mainly characterized by a heterogeneous distribution of protein and lipid structures [49]. These compounds typically display broad bands, where lipids and proteins have strong features in the 1400-1500 cm^{-1} range [50], while proteins also display broad features in the 1200-1400 cm^{-1} range [51]. Moreover, the observed signal in these regions aligns with Raman microscopy measurements of *S. cerevisiae* strains in literature on strain discrimination [15]. The spectral features indicated in Figure 3.5D could be linked to phenylalanine (1002 cm^{-1}) [52], phospholipids (1084 cm^{-1}) [53], broad amide III bands (1246 cm^{-1}) [51], protein CH deformation (1344 cm^{-1}), CH$_2$-deformation of proteins and lipids (1448 cm^{-1}), and amide I stretching (1669 cm^{-1}) [51], although the low intensity of these features make a more exact identification of each band challenging. Nevertheless, these measurements indicate that the molecular composition of *S. cerevisiae* can be measured using in-line Raman spectroscopy, even though literature suggests that no identifiable Raman spectroscopy features are detectable with in-line measurements [16, 54]. This is most likely due to the difference in pre-processing strategy, where we correct for the otherwise dominant extinction effects. The measured shift of the water HOH-bending peak to higher wavenumbers by 1.90 cm^{-1} seems to be caused by the increase of the underlying spectral feature at 1669 cm^{-1} (amide I stretching). Similarly, the increase of the spectral feature at 1002 cm^{-1} (phenylalanine) most likely caused the shift to higher wavenumbers of the sulphate peak.

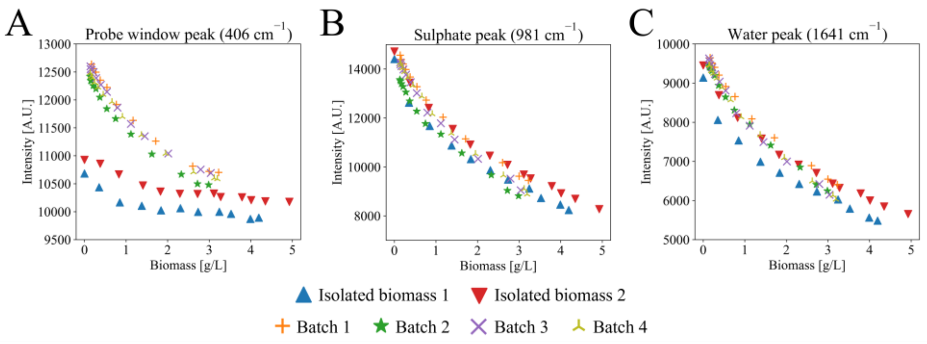


Figure 3.6: The effect of biomass on the peak intensities for the probe window (A), media sulphate (B), and water HOH-bending (C) peak. The datapoints were acquired by taking the absolute peak intensities at the indicated wavenumbers after pre-processing spectra with AirPls baseline correction (lambda = 300).

The intensity of the specific peaks associated with the composition of *S. cerevisiae* is considerably lower compared to the extinction effects induced by the cells as particles. To determine the degree of signal extinction caused by the yeast cells as particles, the peak intensity of the probe window (406 cm^{-1}), sulphate (981 cm^{-1}), and water HOH-bending (1641 cm^{-1}) was determined from the AirPls corrected biomass spectra. To investigate if these peak extinctions translate to real fermentation data, a comparison was made to spectra acquired during yeast batch cultivations. Raman spectra from four individual batch cultivations with biomass concentrations reaching from 0.1 to 3 g/L were pre-processed by the same AirPls baseline correction, and the measured signal decreases are shown in Figure 3.6.

The pattern of signal extinction for all peaks was similar between the two isolated biomass experiments and resulted only in a slight intensity offset. The average intensity decrease of the probe window peak at 406 cm^{-1} (Figure 3.6A) was 4.8% and 16.1% for the isolated yeast measurements and batch fermentations, respectively, and a large offset in signal intensity between the isolated yeast measurements and batch data was observed. The difference in intensity decrease stems from a different fit of the AirPls baseline correction caused by overlapping spectral features of glucose and other compounds during the fermentations with the probe window peak. Without this overlap, it is clear that the probe window peak experiences little intensity decrease for the isolated biomass measurements compared to the fermentations. In contrast, signal extinction was comparable for the sulphate and water peaks between the isolated biomass measurements and the batch cultivations, which both show a non-linear decrease in peak intensities when the biomass concentration increases (Figure 3.6B and 3.6C). Increasing the yeast concentration from 0 to 5 g/L led to average signal extinctions of 44.7 % and 44.6 % for the sulphate (981 cm^{-1}) and water (1641 cm^{-1}) peaks, respectively. The batch cultivations only reached up to 3 g/L of biomass and resulted in an average extinction of 36.6% and 33.3% for the sulphate and water peak. During both experiments, the sulphate and water peak were mostly free from other overlapping molecular signals, and their concentration is assumed to be constant, which is why these peaks are utilized as internal standards in other reported work [4, 16, 18]. The observed non-linear pattern for increasing the biomass concentration is similar to the observations of Yang et al. (2024) who measured the extinction of glucose and ethanol peaks [54]. Although the degree of extinction between the sulphate and water peak appears to be similar in this work, this does not guarantee uniform extinction across the fingerprint region in the raw spectra. The AirPls baseline correction performed optimally with a lambda value of 300. Slight changes to the lambda value affected the fit of both broad and narrow peaks, which could potentially influence the measured signal extinction. Previous literature on the extinction effect of yeast cells as particles also highlighted that the extinction effect is not uniformly distributed across the fingerprint region [16, 54]. Therefore, the use of internal standards for normalizing extinction effects should be carefully assessed, as this does not guarantee a proper correction for features far from the internal standard.

The biomass experiment revealed weak bands associated with the molecular composition of yeast proteins and lipids in the 1000-1500 cm^{-1} region and showed a strong non-linear signal extinction for increasing concentrations. However, the detected spectral features for biomass spanned a broad range that overlaps with the strong and sharp spectral features of commonly found molecular compounds in fermentations, namely glucose and ethanol. As the increase in biomass and product concentration typically correlate strongly during fermentation, it may be unlikely for multivariate linear regression models to extract the right spectral variation for biomass from the signals of ethanol. Broad features are also more sensitive to removal by baseline and scattering corrections as these can be perceived as spectral baseline effects.

The findings in this work highlight how multivariate linear regression models can be challenged by Raman spectra from yeast cultivations, as the strong signal extinction by cells as particles disrupts the proportionality between signal intensity and molecule concentration. When looking at bioreactor applications of Raman spectroscopy over the last decades, the majority was for mammalian cell cultures such as Chinese hamster ovary (CHO) cells [6]. The typical effective pre-processing strategies for CHO cell processes include a baseline correction or derivative step followed by a standard normal variate (SNV) scatter correction [20, 55], and strong signal extinction effects as observed for yeast cells are not reported. Yeast cells are smaller and have a high optical density, which results in stronger light scattering. The extremity of signal extinction is seen even at the low yeast concentrations in this work, and this supports approaches that employ the signal extinction itself to quantify yeast biomass. For example, Yang et al. (2024) used the intensity decrease of the water HOH-bending peak to quantify biomass concentrations [54]. However, such approaches might only be viable when yeast cells are the only light scattering particles in the medium. As presented in section 3.2, operating the bioreactor at a typical stirring rate of 830 rpm generated bubbles leading up to a 7.91% reduction of distinct peak intensity, while biomass of up to 5 g/L led to signal extinctions of 44.6%. Since the biomass can be expected to have a relatively uniform size, further research is required to determine the precise influence of bubble size and size distributions on signal attenuation. Despite the spectral distortion caused by biomass or bubbles, there have been many successful cases of monitoring compounds such as glucose, ethanol, and other products during yeast fermentation with Raman spectroscopy [4, 56]. This indicates that the extinction effects can be removed from spectra with the appropriate scatter corrections and normalization steps.

3.4 Conclusion

Raman spectroscopy is becoming an established analytical tool for cell culture monitoring and has been successfully applied to quantify small molecular compounds in real-time. However, without understanding the impact of different process parameters, quantitative model performance cannot be guaranteed when leveraging data from different scales, processes, or measurement conditions. This study investigated the spectral impact of temperature, bubble quantity, viscosity, and yeast biomass. Changing the temperature in the bioreactor from 20 to 40^{0}C resulted in a maximal peak shift of up to 0.80 cm^{-1} towards lower wavenumbers, while a decrease in peak intensity was observed of up to 4.46% (relative to the spectral baseline). An increasing number of bubbles generated through high stirring speeds resulted in signal extinction, reflected by up to 7.93% lower peak intensities in synthetic media samples. The spectral effects of liquid viscosity as a function of temperature in a 15% v/v glycerol water mixture led to complex peak shift behaviour with magnitudes similar to those observed for temperature alone, and we were therefore not able to truly isolate the effects of viscosity. Increasing biomass concentrations led to strong signal extinctions of up to 44.7% in the fingerprint region, causing a significant reduction in the measured signal of peaks from compounds with constant concentrations. In addition to signal attenuation, we were able to identify weak spectral bands associated with proteins and lipids in the 1000-1500 cm^{-1} spectral region after normalizing for scattering effects. This demonstrated that spectral features related to the molecular composition of yeast cells can be detected with in-line Raman spectroscopy, although the signal extinction caused by the cells as particles remains the dominant effect.

Overall, this work shows that Raman spectra are sensitive to wavenumber shifts and peak intensity changes for different operational conditions, and that particles such as bubbles and cells can cause significant signal extinction during in-line measurements. These effects can complicate model calibration and subsequent application, as mismatches between the affected spectral features might reduce quantification accuracy. Moreover, predictive models may face difficulties in generalizing between spectra acquired from different measurement conditions, especially considering sharp spectral features spanning few wavenumbers. A better understanding of these sources of spectral noise aids in the design of appropriate data pre-processing steps, thereby removing baseline shifts, restoring linearity through normalization, and reestablishing variable alignment. Furthermore, data pre-processing and modelling strategies should be tailored to each compound of interest. For example, the spectral contribution from yeast biomass is dominated by signal extinction rather than by its molecular spectral features. The insights obtained through this work contribute to a better understand of signal distortion in bioprocessing, thereby serving the development of robust quantification models. This work will support rapid model calibration with the use of (miniaturized) alternative setups to minimize time- and labour-intensive (re)calibration activities, thereby making real-time monitoring with Raman spectroscopy more accessible.

Acknowledgements

We would like to thank Christiaan Mooiman and Jeroen Schmidt for their technical support while setting up the lab and during the fermentations.

References

1. FDA, *Guidance for industry: PAT—A framework for innovative pharmaceutical development, manufacturing, and quality assurance.* Food and Drug Administration, Rockville, MD, 2004.
2. Esmonde-White, K.A., et al., *Raman spectroscopy as a process analytical technology for pharmaceutical manufacturing and bioprocessing.* Analytical and bioanalytical chemistry, 2017. **409**(3): p. 637-649.
3. Rathore, A., R. Bhambure, and V. Ghare, *Process analytical technology (PAT) for biopharmaceutical products.* Analytical and bioanalytical chemistry, 2010. **398**(1): p. 137-154.
4. Hirsch, E., et al., *Inline noninvasive Raman monitoring and feedback control of glucose concentration during ethanol fermentation.* Biotechnology Progress, 2019. **35**(5): p. e2848.
5. Pelletier, M.J., *Quantitative analysis using Raman spectrometry.* Applied spectroscopy, 2003. **57**(1): p. 20A-42A.
6. Esmonde-White, K.A., M. Cuellar, and I.R. Lewis, *The role of Raman spectroscopy in biopharmaceuticals from development to manufacturing.* Analytical and Bioanalytical Chemistry, 2021: p. 1-23.
7. Metcalfe, G.D., T.W. Smith, and M. Hippler, *On-line analysis and in situ pH monitoring of mixed acid fermentation by Escherichia coli using combined FTIR and Raman techniques.* Analytical and bioanalytical chemistry, 2020. **412**: p. 7307-7319.
8. Ryabchykov, O., S. Guo, and T. Bocklitz, *Analyzing Raman spectroscopic data.* Physical Sciences Reviews, 2019. **4**(2): p. 20170043.
9. Bocklitz, T., et al., *How to pre-process Raman spectra for reliable and stable models?* Analytica chimica acta, 2011. **704**(1-2): p. 47-56.
10. Lieber, C.A. and A. Mahadevan-Jansen, *Automated method for subtraction of fluorescence from biological Raman spectra.* Applied spectroscopy, 2003. **57**(11): p. 1363-1367.
11. Sinfield, J.V. and C.K. Monwuba, *Assessment and correction of turbidity effects on Raman observations of chemicals in aqueous solutions.* Applied spectroscopy, 2014. **68**(12): p. 1381-1392.
12. Shaw, A.D., et al., *Noninvasive, on-line monitoring of the biotransformation by yeast of glucose to ethanol using dispersive Raman spectroscopy and chemometrics.* Applied spectroscopy, 1999. **53**(11): p. 1419-1428.
13. Müller, D.H., et al., *Bioprocess in-line monitoring using Raman spectroscopy and Indirect Hard Modeling (IHM): A simple calibration yields a robust model.* Biotechnology and Bioengineering, 2023.
14. Jiang, H., et al., *Quantitative analysis of yeast fermentation process using Raman spectroscopy: Comparison of CARS and VCPA for variable selection.* Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2020. **228**: p. 117781.
15. Wang, K., et al., *Species identification and strain discrimination of fermentation yeasts Saccharomyces cerevisiae and Saccharomyces uvarum using Raman spectroscopy and convolutional neural networks.* Applied and Environmental Microbiology, 2023. **89**(12): p. e01673-23.
16. Iversen, J.A., R.W. Berg, and B.K. Ahring, *Quantitative monitoring of yeast fermentation using Raman spectroscopy.* Analytical and bioanalytical chemistry, 2014. **406**(20): p. 4911-4919.
17. Zakhartsev, M. and M. Reuss, *Cell size and morphological properties of yeast Saccharomyces cerevisiae in relation to growth temperature.* FEMS yeast research, 2018. **18**(6): p. foy052.

18. Picard, A., et al., *In situ monitoring by quantitative Raman spectroscopy of alcoholic fermentation by Saccharomyces cerevisiae under high pressure.* Extremophiles, 2007. **11**(3): p. 445-452.
19. Iversen, J.A. and B.K. Ahring, *Monitoring lignocellulosic bioethanol production processes using Raman spectroscopy.* Bioresource technology, 2014. **172**: p. 112-120.
20. Domján, J., et al., *Raman-based dynamic feeding strategies using real-time glucose concentration monitoring system during adalimumab producing CHO cell cultivation.* Biotechnology Progress, 2020. **36**(6): p. e3052.
21. Lee, H.L., et al., *In situ bioprocess monitoring of Escherichia coli bioreactions using Raman spectroscopy.* Vibrational Spectroscopy, 2004. **35**(1-2): p. 131-137.
22. Zhang, S., et al., *Raman spectroscopy study of acetonitrile at low temperature.* Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2021. **246**: p. 119065.
23. Yoshikawa, Y. and S. Shigeto, *A simple calibration method of anti-Stokes–Stokes Raman intensity ratios using the water spectrum for intracellular temperature measurements.* Applied Spectroscopy, 2020. **74**(10): p. 1295-1296.
24. Carey, D.M. and G.M. Korenowski, *Measurement of the Raman spectrum of liquid water.* The Journal of chemical physics, 1998. **108**(7): p. 2669-2675.
25. Seki, T., et al., *The bending mode of water: A powerful probe for hydrogen bond structure of aqueous systems.* The journal of physical chemistry letters, 2020. **11**(19): p. 8459-8469.
26. Đuričković, I., et al., *Experimental study of NaCl aqueous solutions by Raman spectroscopy: Towards a new optical sensor.* Applied spectroscopy, 2010. **64**(8): p. 853-857.
27. Vogt, F. and K. Booksh, *Influence of wavelength-shifted calibration spectra on multivariate calibration models.* Applied spectroscopy, 2004. **58**(5): p. 624-635.
28. Seasholtz, M.B. and B.R. Kowalski, *Qualitative information from multivariate calibration models.* Applied spectroscopy, 1990. **44**(8): p. 1337-1348.
29. Wise, B.M. and R.T. Roginski, *A calibration model maintenance roadmap.* IFAC-PapersOnLine, 2015. **48**(8): p. 260-265.
30. Witjes, H., et al., *Automatic correction of peak shifts in Raman spectra before PLS regression.* Chemometrics and Intelligent Laboratory Systems, 2000. **52**(1): p. 105-116.
31. Verduyn, C., et al., *Effect of benzoic acid on metabolic fluxes in yeasts: a continuous-culture study on the regulation of respiration and alcoholic fermentation.* Yeast, 1992. **8**(7): p. 501-517.
32. Nijkamp, J.F., et al., *De novo sequencing, assembly and analysis of the genome of the laboratory strain Saccharomyces cerevisiae CEN. PK113-7D, a model for modern industrial biotechnology.* Microbial cell factories, 2012. **11**(1): p. 1-17.
33. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* the Journal of machine Learning research, 2011. **12**: p. 2825-2830.
34. Lopez, P.C., *chemotools: A Python Package that Integrates Chemometrics and scikit-learn.* Journal of Open Source Software, 2024. **9**(100): p. 6802.
35. Dudek, M., et al., *Raman Optical Activity and Raman spectroscopy of carbohydrates in solution.* Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2019. **206**: p. 597-612.
36. Padro, J.A. and J. Martı́, *An interpretation of the low-frequency spectrum of liquid water.* The Journal of chemical physics, 2003. **118**(1): p. 452-453.
37. Walrafen, G., et al., *Temperature dependence of the low-and high-frequency Raman scattering from liquid water.* The Journal of chemical physics, 1986. **85**(12): p. 6970-6982.
38. Sun, Q., *The Raman OH stretching bands of liquid water.* Vibrational Spectroscopy, 2009. **51**(2): p. 213-217.

39. Ahmed, M., et al., *How ions affect the structure of water: a combined Raman spectroscopy and multivariate curve resolution study.* The Journal of Physical Chemistry B, 2013. **117**(51): p. 16479-16485.
40. Lednev, V.N., et al., *Quantifying Raman OH-band spectra for remote water temperature measurements.* Optics Letters, 2016. **41**(20): p. 4625-4628.
41. Thapa, J., et al., *Raman scattering in single-crystal sapphire at elevated temperatures.* Applied optics, 2017. **56**(31): p. 8598-8606.
42. Matsumoto, Y., et al., *Raman spectroscopic study of aqueous alkali sulfate solutions at high temperature and pressure to yield precipitation.* The Journal of Supercritical Fluids, 2009. **49**(3): p. 303-309.
43. Rull, F., F. Sobron, and O. Nielsen, *Dependence on concentration and temperature of the dynamics of SO in Li2SO4, Na2SO4 and K2SO4 aqueous solutions studied by Raman spectroscopy.* Journal of Raman Spectroscopy, 1995. **26**(8-9): p. 663-668.
44. Workman Jr, J.J., *A review of calibration transfer practices and instrument differences in spectroscopy.* Applied spectroscopy, 2018. **72**(3): p. 340-365.
45. Volger, R., L. Puiman, and C. Haringa, *Bubbles and Broth: A review on the impact of broth composition on bubble column bioreactor hydrodynamics.* Biochemical Engineering Journal, 2024. **201**: p. 109124.
46. Emmerich, J., et al., *Optical inline analysis and monitoring of particle size and shape distributions for multiple applications: Scientific and industrial relevance.* Chinese Journal of Chemical Engineering, 2019. **27**(2): p. 257-277.
47. Rudiš, M., V. Jezdinský, and Z. Štěrbáček, *Physical properties of microbial suspensions: II. Properties of microbial suspensions and their supernatants during fermentation conditions.* Folia Microbiologica, 1977. **22**: p. 128-133.
48. Mendelovici, E., R.L. Frost, and T. Kloprogge, *Cryogenic Raman spectroscopy of glycerol.* Journal of Raman Spectroscopy, 2000. **31**(12): p. 1121-1126.
49. Lange, H. and J. Heijnen, *Statistical reconciliation of the elemental and molecular biomass composition of Saccharomyces cerevisiae.* Biotechnology and bioengineering, 2001. **75**(3): p. 334-344.
50. Czamara, K., et al., *Raman spectroscopy of lipids: a review.* Journal of Raman spectroscopy, 2015. **46**(1): p. 4-20.
51. Rygula, A., et al., *Raman spectroscopy of proteins: a review.* Journal of Raman Spectroscopy, 2013. **44**(8): p. 1061-1076.
52. Hernández, B., et al., *Characteristic Raman lines of phenylalanine analyzed by a multiconformational approach.* Journal of Raman Spectroscopy, 2013. **44**(6): p. 827-833.
53. Kochan, K., et al., *Single cell assessment of yeast metabolic engineering for enhanced lipid production using Raman and AFM-IR imaging.* Biotechnology for biofuels, 2018. **11**: p. 1-15.
54. Yang, N., et al., *In-line monitoring of Bioreactor by Raman Spectroscopy: direct use of a standard--based model through cell--scattering correction.* Journal of Biotechnology, 2024.
55. Webster, T.A., et al., *Development of generic raman models for a GS-KOTM CHO platform process.* Biotechnology Progress, 2018. **34**(3): p. 730-737.
56. Wieland, K., et al., *Non-invasive Raman spectroscopy for time-resolved in-line lipidomics.* RSC advances, 2021. **11**(46): p. 28565-28572.

3.5 Supplementary

3.5.1 Temperature effects

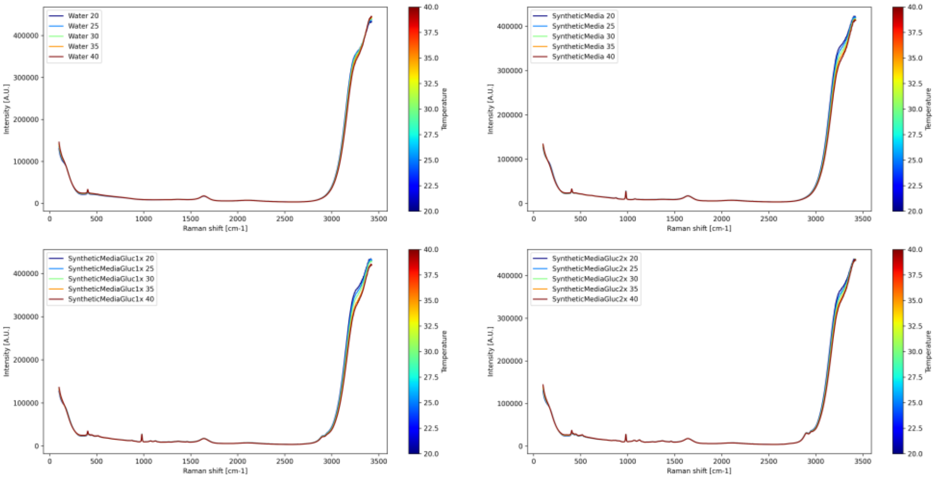


Figure 3.5.1.1: The full spectra of water (A), synthetic media (B), synthetic media with low glucose concentration (25 g/L) (C), and synthetic media with high glucose concentration (50 g/L) (D). The displayed spectra are the average of 10 individual 60-second spectra acquired at each temperature.

1.1.1. Bubble effects

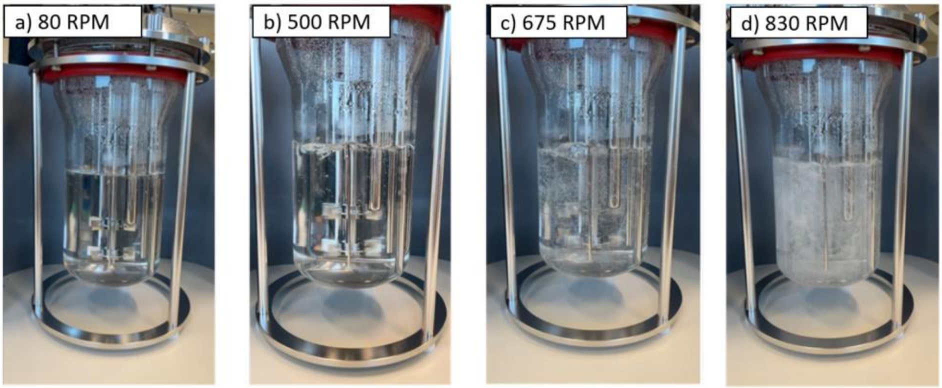


Figure 3.5.2.1: The bubble formation by increasing impeller speed without sparging in a 2L Applikon bioreactor with 1L of synthetic media at 80 (A), 500 (B), 675 (C), and 830 (D) rpm. The temperature was kept constant at 30^{0}C.

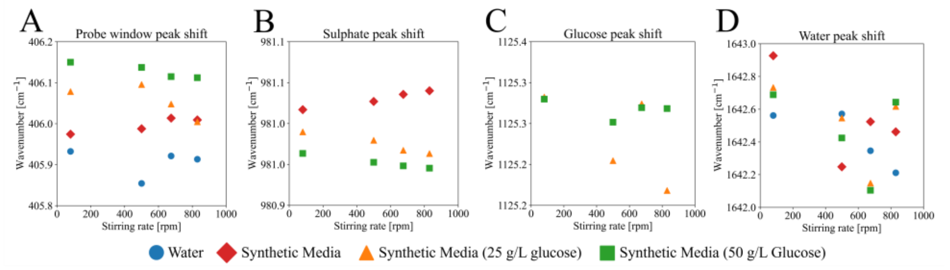


Figure 3.5.2.2: The effect of bubbles on the peak locations of the probe window (A), media sulphate (B), glucose (C), and water HOH-bending (D) peaks. The peak location was derived by pre-processing the spectra with AirPls baseline correction (lambda = 300), taking a Savitzky Golay derivative (1st order, 15-point window length), and interpolating the zero-crossing point of the x-axis.

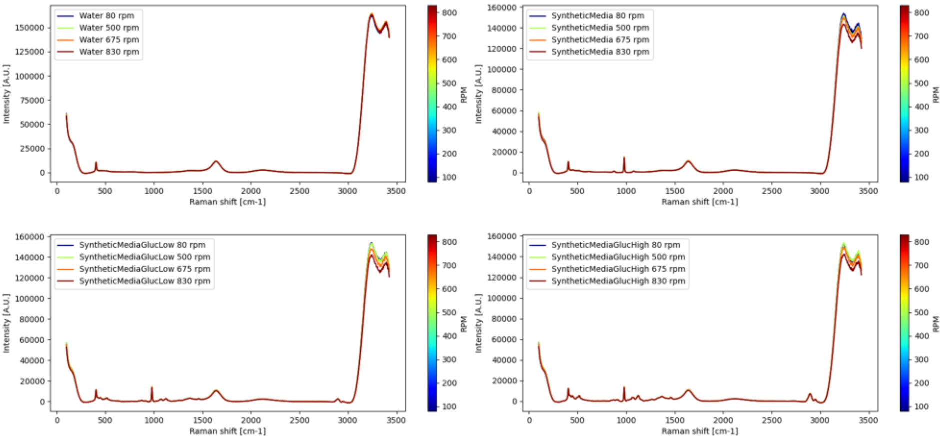


Figure 3.5.2.3: The spectral effects of bubble formation on spectra of water (A), synthetic media (B), synthetic media with 25 g/L glucose, and synthetic media with 50 g/L glucose. The colourbar indicates the stirring rate at which the spectra were acquired.

3.5.2 Viscosity effects

Table 3.5.3.1: The measured viscosity values (in duplicate) of the 15% glycerol v/v solution measured at different temperatures in the Lovis 2000 M/ME viscometer (Anton Paar, Austria).

Temperature (^{0}C)	µ (mPa.s)	ϱ (kg/m3)	µ (mPa.s) - Average	ϱ (kg/m3) - Average
20	1.5549	1036.806	1.5503	1036.81
	1.5458	1036.807		
25	1.3568	1035.217	1.3567	1035.22
	1.3567	1035.214		
30	1.2018	1033.423	1.2017	1033.41
	1.2017	1033.404		
35	1.0958	1031.377	1.0993	1031.33
	1.1028	1031.277		
40	1.0258	1028.601	1.02855	1028.58
	1.0313	1028.558		
45	0.8861	1026.902	0.8948	1026.87
	0.9036	1026.842		

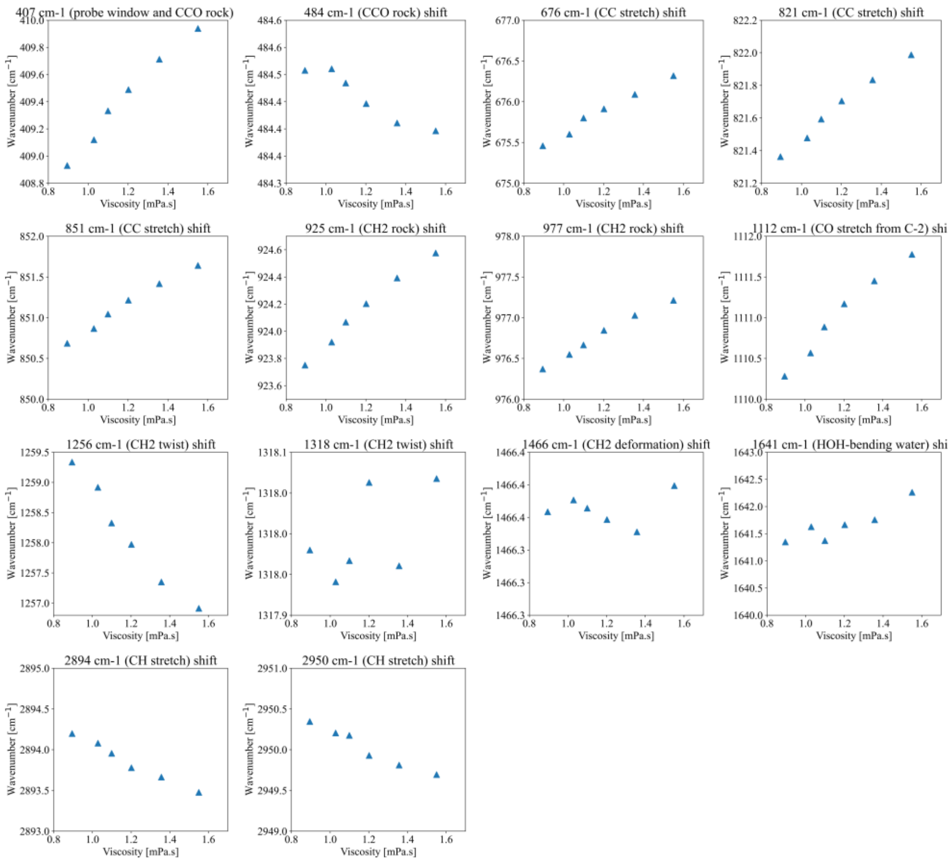


Figure 3.5.3.1: An overview of all glycerol Raman spectroscopy peak shifts analyzed of a 15% v/v glycerol water solution. The viscosity was adjusted by changing the temperature of the mixture.

3.5.3 Biomass effects

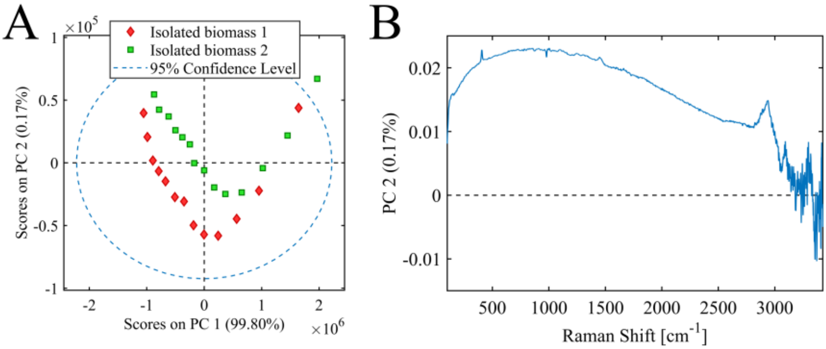


Figure 3.5.4.1: A Principal Component Analysis (PCA) performed on the biomass spectra obtained from resuspending biomass from two individual batch fermentations in fresh synthetic media. The full spectra were pre-processed with mean centering, resulting in a model with 2 principal components. Most variation is captured in PC1, and the spectra of the two experiments deviated slightly on baseline position in PC2.

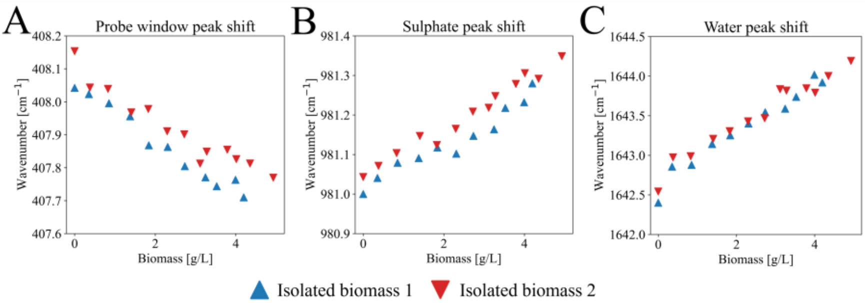


Figure 3.5.4.2: The effect of biomass on the peak locations of probe window (A), media sulphate (B), glucose (C), and water HOH-bending (D) peaks. The peak location was derived by pre-processing the spectra with AirPls baseline correction (lambda = 300), taking a Savitzky Golay derivative (1st order, 15-point window length), and interpolating the zero-crossing point of the x-axis.

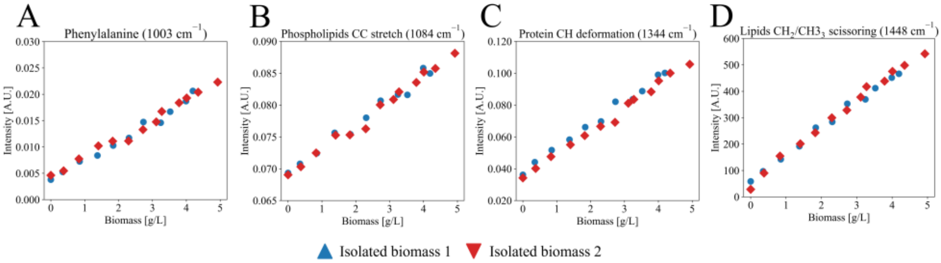
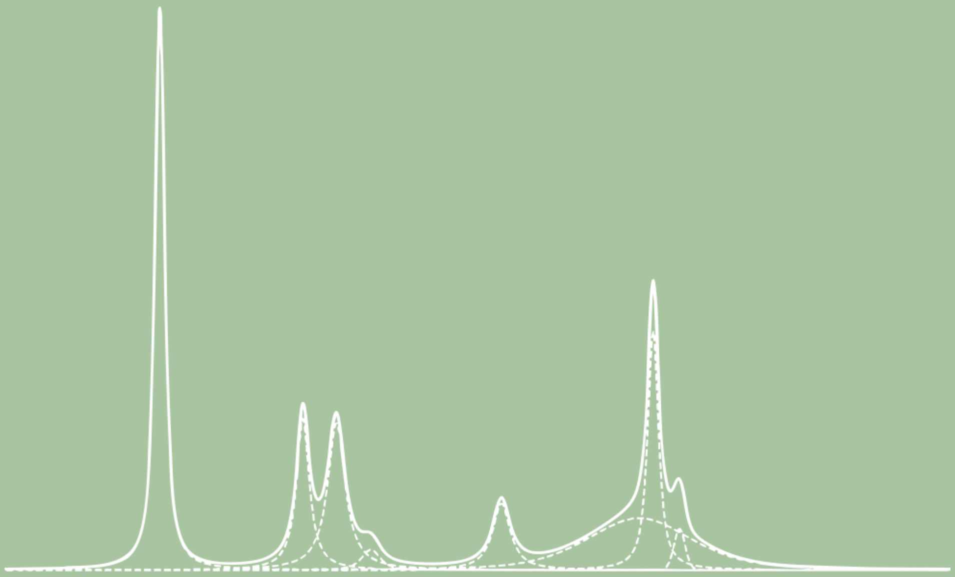


Figure 3.5.4.3: The linear correlation between biomass concentration and the intensity of the peaks related to phenylalanine (A), phospholipids CC stretching (B), protein CH deformation (C), and lipids CH_2/CH_3 scissoring (D). The spectra were pre-processed with AirPls baseline correction (lambda = 300), and sulphate peak normalization



Chapter 4

Towards rapid calibration of bioprocess quantification models using single compound Raman spectra: a comparison of four approaches

---

This article has been published as:

Klaverdijk, M., Smulders, L. A., Ottens, M., & Klijn, M. E. (2025). Towards Rapid Calibration of Bioprocess Quantification Models Using Single Compound Raman Spectra: A Comparison of Four Approaches. *Biotechnology and Bioengineering*.

Abstract

In-line Raman spectroscopy combined with accurate quantification models can offer detailed real-time insights into a bioprocess by monitoring key process parameters. However, traditional approaches for model calibration require extensive data collection from multiple bioreactor runs, resulting in process-specific models that are sensitive to operational changes. These challenges can be tackled by simplifying experimental data generation or implementation of computational methods to obtain synthetic and augmented Raman spectra. In this work, we utilized a small experimental dataset of 16 single compound spectra to calibrate quantification models by using Partial Least Squares (PLS) and Indirect Hard Modelling (IHM), leading to comparable rRMSEP values for glucose (4.8% and 4.2%), ethanol (11.6% and 6.3%), and biomass (16.2% and 10.0%) when applied to yeast batch and fed-batch bioprocesses. Subsequently, isolated spectral features extracted during IHM were used to generate fully synthetic spectral datasets for PLS model calibration, resulting in rRMSEPs of 3.2% and 14.5% for glucose and ethanol, respectively. Finally, spectra from a single batch process were augmented with the same isolated spectral features, and calibration with these augmented spectra reduced rRMSEP by 18.6 percent point (glucose) and 4.3 percent point (ethanol) compared to process-only calibrated models. This work demonstrates how different approaches may support robust development and rapid implementation of Raman spectroscopy-based models while minimizing experimental efforts, where even complete independence of process data can be achieved.

Keywords: *Raman spectroscopy, in-line measurements, data augmentation, Saccharomyces cerevisiae, process analytical technology, bioprocessing*

4.1 Introduction

Monitoring the concentration of metabolites, products, and biomass during a bioreactor process is often based on labor-intensive manual sampling and off-line sample analysis. The development and implementation of novel Process Analytical Technology (PAT) aims to automate the collection of quantitative data on these process parameters to achieve hands-free real-time monitoring. In recent decades, optical PAT has seen a rise in popularity as it allows for in-line measurements that can be combined with automated data analysis [1]. Specifically Raman spectroscopy is highly suitable for bioreactor processes due to its low signal interference from water and the ability to provide specific fingerprint signals for many compounds. Raman spectroscopy is successfully implemented across a wide range of bioreactor processes, from microbial to animal cell cultures, with the goal to quantify both simple and complex target compounds [2]. However, as the complexity of the measured systems increases, spectral features of all compounds in the system overlap in the singular spectrum, hindering direct interpretation of the raw signal. Therefore, multivariate modelling techniques are employed to translate the complex spectral signal into quantitative data.

The most popular multivariate technique is Partial Least Squares (PLS) regression and it is extensively used for a wide range of processes [3]. During PLS model calibration, spectral and reference value datasets are provided, and the model defines latent variables (LVs) that capture the most relevant spectral variations correlated to the target compound identity and abundance [4]. This makes PLS a powerful technique for Raman spectral decomposition to predict the target compound concentration, while little spectral knowledge of the system is required. These models are traditionally calibrated with extensive datasets for which multiple bioreactor runs have to be performed, leading to labor-intensive data collection procedure. Unfortunately, the required time and material investment delays the adoption of Raman spectroscopy as PAT to generate valuable process insights in early-stage development of new processes or in R&D-based environments. Moreover, repeating the same bioreactor process to collect calibration data leads to a limited design space, as the relationship between process compounds remains similar for every process run. As a consequence, the PLS model will learn to predict compound concentrations only under the circumstances occurring in that specific process. This means models can be trained to identify and predict abundance based on unspecific spectral features when compounds have strong cross-correlations (e.g., correlations between substrates, inverse correlations between substrates and product or biomass). Such models will perform poorly when process conditions or process operation disturbs this cross-correlation. To calibrate robust PLS models that can deal with process variations and transfers to related processes, the calibration dataset should include variation outside of the standard process evolution. However, collecting datapoints from the process itself by repeating bioreactor runs with different concentration settings is inefficient, especially when it is solely for the improvement of a Raman spectroscopy-based PLS model. Experimental methods for introducing these variations in the dataset, such as spiking the process with the compound of interest or creating custom samples in a cell culture matrix, can enhance a model's specificity for target compounds [5, 6]. However, these approaches are labor-intensive, requiring careful experimental design and sample preparation.

Computational approaches can also offer a solution to build spectral datasets for robust model calibration. When extensive process knowledge is available and the spectral composition of most individual process compounds is obtainable, methods such as Indirect Hard Modelling (IHM) can be used. IHM is a physics-based approach that incorporates known spectral properties to extract chemical information from a process spectra [7]. It differs from implicit modelling techniques, such as PLS, by explicitly modeling known spectral features of a compound with individual peak functions. Spectra of pure or dissolved compounds are deconvoluted by fitting Pseudo-Voigt peak profiles to the spectra until the residuals between the fitted model and experimental spectra are minimized. For mixtures containing a single unknown compound, the unknown spectral variation can be characterized through complemental hard modeling, allowing the spectral composition of the unknown compound to be extracted [8]. The defined individual models can be combined into a mixture model, which is calibrated on training spectra by weighing the intensity of each compound model to minimize the fit residuals. Complemental hard modelling was successfully applied to chemical processes and yeast bioprocesses [7, 9, 10]. When unexplained spectral residuals remain after optimization, the model can be expanded by extracting the unknown contribution, use it to develop a new hard model, and include it in the mixture model [10]. These applications demonstrate that the IHM approach offers a flexible and low calibration effort approach for quantification from spectral data. Nevertheless, IHM requires a high level of spectral knowledge on the process and the availability of isolated spectral measurements of the major process compounds.

As effective calibration data is labor-intensive to collect, alternative methods by which spectral data can be obtained computationally are highly desired. Methods to artificially generate Raman spectra or modify existing spectral data can alleviate current limitations, such as data scarcity and low variability of the compounds of interest. Several automated methods for generating synthetic spectra are developed for classification problems, such as Synthetic Minority Over-Sampling Technology (SMOTE) and Generative Adversarial Networks (GANs) algorithms [11]. The SMOTE algorithm interpolates between existing spectra of a minority-class to reduce class imbalance and to increase the diversity of a dataset [12]. The GAN approach consists of a generator and discriminator model that go through adversarial training, where the generator model learns to generate realistic spectra while the discriminator attempts to recognize synthetic data [11, 13]. Both methods can be used to expand small or imbalanced calibration datasets to improve the performance of classification models [14]. However, the use of these algorithms to generate spectra for quantification model calibration is limited, as a physically accurate relationships between spectral intensities and compound concentrations is not guaranteed. Examples for the generation of synthetic spectra to enhance quantification models are rare. Goldrick et al. generated synthetic Raman spectra simulating penicillin fermentation by combining empirical baseline spectra with simulated characteristic compound peaks in the form of Gaussian shapes [15]. Sulub & Small employed a similar method to simulate near-infrared spectra to calibrate a PLS model for the prediction of glucose in mixture measurements [16]. While these studies highlight the potential of augmenting spectral data for quantification problems, the application of synthetic Raman spectra remains largely unexplored.

This work compares four approaches for utilizing single compound spectra to calibrate Raman spectroscopy quantification models, applicable for bioprocess monitoring or control purposes. These approaches simulate scenarios where no or little process data is available prior to operating a bioreactor process, aiming to build robust models and enable availability of quantification models before a new process begins. The first approach uses a small experimental spectral dataset (16 spectra) containing single compound measurements of glucose, ethanol, and *Saccharomyces cerevisiae* biomass acquired under bioreactor conditions to calibrate PLS models directly. Secondly, a calibrated mixture model is obtained through IHM, calibrated with the same small experimental spectral dataset. Both models are validated on a bioprocess dataset of 4 batches and a single fed-batch (65 spectra total) to assess model performance. In the third approach, isolated spectral features extracted during the IHM approach are used to generate synthetic spectra that simulate bioprocess conditions. This yielded a full factorial dataset of 125 synthetic spectra to maximize spectral variation, with which PLS models were calibrated. In the fourth approach, isolated spectral features of glucose and ethanol from IHM were employed to augment a small spectral process batch bioprocess dataset (12 spectra). It was aimed to improve model specificity towards these targets by artificially increasing the spectral features variability, providing a data augmentation method for situations where process data is limited. To conclude, four approaches to leverage little to no process data for Raman spectroscopy-based quantification model calibration are compared in terms of prediction accuracy, calibration effort, and flexibility towards new compounds. By investigating these calibration approaches we contribute to rapid development of flexible and robust quantification models that can be obtained without running (additional) bioprocesses.

4.2 Materials & Methods

4.2.1 Experimental methods

4.2.1.1 Fermentation settings and reference sampling

The *Saccharomyces cerevisiae* strain CEN.PK113-7D was used for all bioprocesses [17], and cultures were grown on defined medium containing 5 g/L $(NH_4)_2SO_4$, 3 g/L KH_2PO_4, and 0.5 g/L $MgSO_4.7H_2O$ corrected to a pH of 6.0 with 2M KOH [18]. After medium sterilization 50% glucose (J.T. Baker, Philipsburg, NJ) solution (in-house) was added until 20 g/L, and vitamins and trace elements were added through 0.2 µm syringe filters (Whatman, Maidstone, UK). The medium was completed by adding 0.2 g/L sterile Antifoam-C (BASF, Ludwigshafen, Germany). Bioprocess data was collected by operating 4 batches and a single fed-batch in a 2L bioreactor system (Applikon, Delft, the Netherlands) using a 1L working volume. The cultures were maintained at 30^0C, stirred at 800 rpm, and aerated with 0.5 L/min of air by a Biostat B bioreactor controller (Sartorius, Göttingen, Germany). The pH setpoint of 6.0 was maintained by the automatic addition of 2M KOH. The batch bioprocesses were inoculated at an OD660 of 0.3 and sampled until glucose depletion. The fed-batch started as a batch culture operated at identical settings, and was bolus fed with 50% glucose solution three times whenever glucose depleted to extend the process. The bioprocesses were sampled every hour, and sample supernatants were analyzed for their glucose and ethanol concentrations with an Agilent 1260 infinity HPLC (Agilent Technologies, CA) equipped with a Bio-RAD Aminex HPX-87H (300 x 7.8 mm) cation-exchange column (Bio-Rad, Hercules, CA). The biomass concentration of each sample was determined by measuring the OD660 values using a Libra S11 spectrophotometer (Biochrom, UK), and dry-weight determination was performed by loading and drying 10mL of culture broth on nitrocellulose membrane filters (pore size: 0.45 µm); Gelman Laboratory, MI).

An overview of the reference measurements for each bioprocess is shown in Supplementary Figure 4.5.1.2.

4.2.1.2 Single compound measurements

The glucose, ethanol, and biomass single compound spectra were acquired in the same 2L bioreactor system operated under identical temperature and aeration settings as the fermentation (section 2.1.1). For each compound, the bioreactor was filled with 1L of defined media, and 5 concentrations values were achieved by adding 50% glucose solution (described in section 2.1.1), 96% ethanol, and biomass harvested from a batch fermentation and washed in defined media. The final spectral dataset with a total of 16 spectra consisted of a single defined media spectra followed by 5 glucose (50-250 mM), 5 ethanol (50-250 mM), and 5 biomass (0.8-5 g/L) spectra (Supplementary Figure 4.5.1.1A). The concentrations of each step were verified with HPLC and dry-weight determination as described in section 4.2.1.1.

4.2.1.3 Raman spectral acquisition

The glucose, ethanol, and biomass single compound spectra were acquired in the same 2L bioreactor system operated under identical temperature and aeration settings as the bioprocess (section 2.1.1). For each compound, the bioreactor was filled with 1L of defined media, and 5 concentrations values were achieved by adding 50% glucose solution (described in section 2.1.1), 96% ethanol, and biomass obtained from a batch bioprocess and subsequently washed in defined media. The final spectral dataset with a total of 16 spectra consisted of a single defined media spectra followed by 5 glucose (50-250 mM), 5 ethanol (50-250 mM), and 5 biomass (0.8-5 g/L) spectra (Supplementary Figure 4.5.1.1). The concentrations of each step were verified with HPLC and dry-weight determination as described in section 4.2.1.1.

4.2.2 Computational methods

4.2.2.1 Partial Least Squares model calibration

All Partial Least Squares (PLS) models were developing in PLS_Toolbox version 9.3.8 (Eigenvector Research Inc., WA) running on Matlab R2023a (MathWorks, WA). All spectra were pre-processed by reducing variables to the fingerprint region of 700-1800 cm^{-1}, Automatic Whittaker filter baseline correction ($\lambda=10000$, $\alpha=0.001$), sulphate peak normalization, and mean centering. The reference values for glucose, ethanol, and biomass were mean centered before calibration. An individual model was generated for each compound of interest, and Venetian blinds cross-validation was used. The number of latent variables for each model was selected based on the elbow point of the root mean square error of calibration (RMSEC) and cross-validation (RMSECV) plots, and by inspecting the loadings of each latent variable to prevent the inclusion of spectral noise. Model performance across calibration datasets was compared by using the relative root mean square error of prediction (rRMSEP) based on the interquartile range (IQR) shown in Equation 4.1:

$$rRMSEP = \frac{RMSEP}{Q3-Q1} \times 100 \qquad (4.1)$$

where Q1 and Q3 represent the first and third quartiles, respectively.

4.2.2.2 Indirect Hard Modelling

The single compound hard models and the mixture models used for quantification were generated in the PEAXACT version 5.9 (Aachen, Germany) spectroscopy software. The single compound spectra were reduced to the fingerprint region (700 to 1800 cm^{-1}), and corrected by sulphate peak normalization. The Complemental Hard Modeling (CHM) [8] approach was utilized to generate a hard model consisting of 7 peaks for defined media (Supplementary Figure 4.5.2.1). The defined media hard model was fitted into the highest concentration measurements of glucose, ethanol, and biomass, and Pseudo-Voigt profiles were fitted sequentially at the location with the highest residual error. This procedure was continued until the newly fitted peaks could not be verified with literature references of their Raman spectra. This resulted in models with 20 peaks for glucose, 8 peaks for ethanol, and 9 peaks for biomass. The four generated models were combined in a single bioprocess mixture model that was subsequently calibrated on the 16 single compound measurements by fitting each component to minimize the spectral residuals. For the reference concentrations, the weight of each component was balanced according to Equation 2:

$$1 = \omega_{DefinedMedia} + \omega_{Glucose} + \omega_{Ethanol} + \omega_{Biomass} \qquad (4.2)$$

The calibration procedure generated linear correlations between component weight and concentration (Supplementary Figure 4.5.2.3). During calibration and application the model was only allowed to change the weights of each hard model, without accounting for peak shifts and shape changes. The performance of the IHM model was expressed in RMSEP and rRMSEP values (Equation 1) to allow for comparison with the PLS models.

4.2.2.3 Synthetic spectra generation and augmentation

Synthetic Raman spectra simulating bioprocess conditions were generated using the Pseudo-Voigt profiles obtained during the IHM steps (section 2.2.2). Individual peak parameters and linear correlations between peak intensity and concentration were extracted and re-combined into bioprocess spectra using an in-house Python script. Concentration ratios between glucose, ethanol, and biomass were designed according to a full factorial design of experiments (DoE) approach with five concentrations per compound, leading to a total of 125 combinations (Supplementary Figure 4.5.3.1). The concentration ratios were inserted in Equation 2 to extract the weight of defined media, and the synthetic spectra were generated by multiplying the Pseudo-Voigt features with the weights corresponding to the desired concentration according to the calibration lines (Supplementary Figure 4.5.2.3). A detailed workflow of all steps is presented in Supplementary Figure 4.5.3.2.

4.2.2.4 Spectral augmentation of batch bioprocess data

The augmentation of Raman spectra from a single batch bioprocess was performed using the same spectral features and weight versus concentration calibrations as used during the generation of synthetic spectra (section 4.2.2.3). Two augmented datasets were generated by adjusting the concentrations of (1) glucose and (2) ethanol, where ± 10 and ± 20 mM around the original values was generated. This was done by adding and subtracting the Pseudo-Voigt profiles, with a boundary at a concentration of 0 mM. This resulted in two datasets consisting of 60 spectra (12 original batch process spectra and 48 spectra augmented with Pseudo-Voigt profiles), see Supplementary Figure 4.5.4.2. The detailed workflow of these steps is presented in Supplementary Figure 4.5.4.1.

4.3 Results and Discussion

In this work we compare four approaches using simple measurements to calibrate Raman spectroscopy quantification models for monitoring key compounds during bioprocessing (glucose, ethanol, and biomass) in scenarios where limited or no bioprocess data is available. A bioprocess setup with a simple broth composition was selected as the target process for quantification, with the main process components being: defined media, glucose, ethanol, biomass, and low amounts of glycerol and acetate (abundance for glycerol and acetate was considered insignificant for modelling). A small dataset of 16 single compound spectra was generated, consisting of one defined media spectrum and five concentrations of glucose, ethanol, and biomass each, plus their reference measurements (Supplementary Figure 4.5.1.1).

For the first approach, the dataset of 16 single compound measurements was used to directly calibrate three PLS models for the quantification of glucose, ethanol, and biomass (Figure 4.1.1, section 4.3.1). In the second approach, single compound spectra were used to generate compound hard models (HMs) for defined media, glucose, ethanol, and biomass. The HMs were combined in a single mixture model with the IHM method (Figure 4.1.2, section 4.3.2). In the third approach, the HMs were used to generate de novo synthetic mixture spectra of custom concentration ratios (Figure 4.1.3, section 4.3.3). This approach allowed for the simulation of bioprocess conditions across the entire design space defined by the concentration ranges of the single compound measurements. In the fourth approach, isolated spectral features of glucose and ethanol were used to augment a small dataset of a single batch process (12 spectra), by which the spectral variability for these compounds could be increased (Figure 4.1.4, section 4.3.4). The performance of these four modelling approaches was validated using a bioprocess dataset consisting of multiple batch bioprocesses and a single fed-batch bioprocess to investigate quantitative accuracy on process data. An overview of all experimental datasets and the four modelling approaches is shown in Figure 4.1.

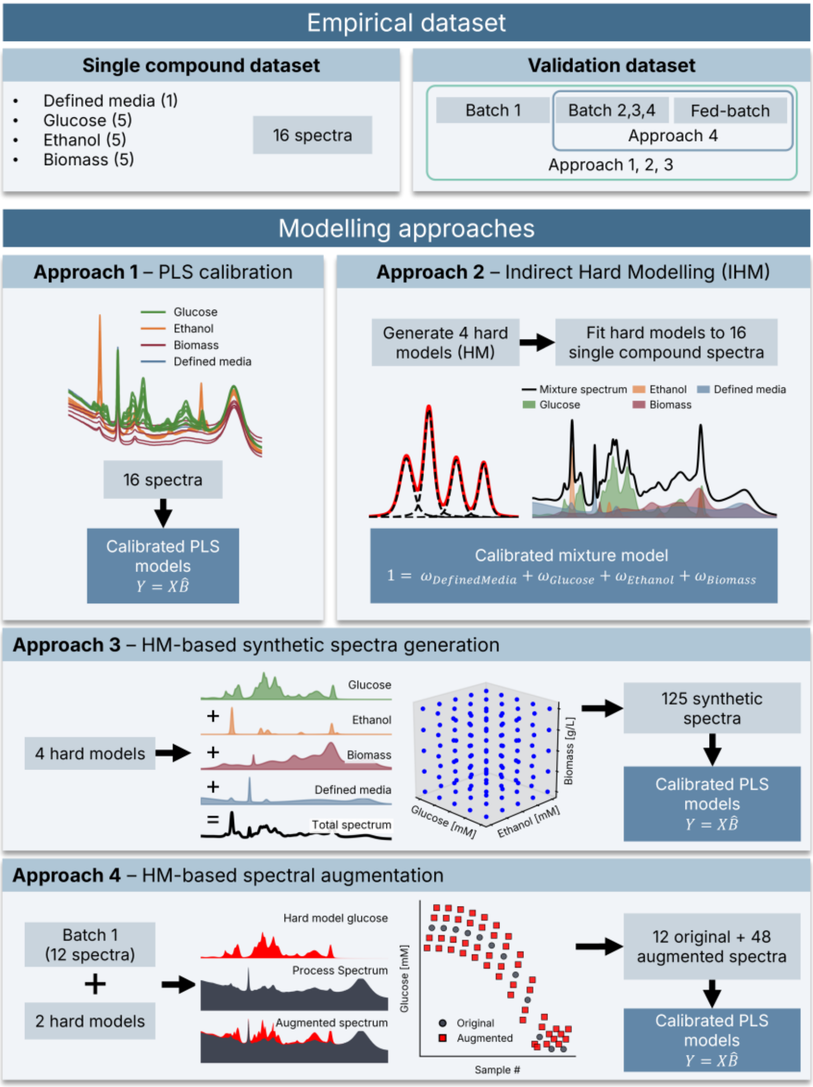


Figure 4.1: An overview of the experimental datasets (top row) and the four modelling approaches performed in this work. For Approach 1: 16 single compound spectra were used to calibrate Partial Least Squares (PLS) regression models for glucose, ethanol, and biomass directly. Approach 2: the single compound spectrum of defined media, and the highest concentration spectra of glucose, ethanol, and biomass were used to generate four hard models (HM) by fitting Pseudo-Voigt features. These HMs were combined in a mixture model that was calibrated on the full single compound dataset to establish relations between signal intensity and compound concentration. Approach 3: HMs of glucose, ethanol, biomass, and defined media and their respective intensity/concentration calibrations were used to generate synthetic Raman spectra with custom concentrations according to a full factorial Design of Experiments with 5 concentrations per quantification target. The resulting dataset of 125 synthetic spectra was used to calibrate PLS models for glucose, ethanol, and biomass. Approach 4: a single batch spectral dataset (Batch 1) was augmented with the HMs of glucose and ethanol to increase the spectral variability of each quantification target, resulting in two augmented datasets of 60 spectra used to calibrate PLS models for glucose and ethanol. All PLS models and the mixture model obtained with IHM were validated on a dataset consisting of multiple batch bioprocesses and a single fed-batch bioprocess.

4.3.1 Approach 1: PLS model calibration with single compound spectra

The performance of PLS models calibrated directly using single compound spectra (16 samples) obtained under standard bioreactor conditions was investigated for three targets: glucose, ethanol, and biomass. These models were assessed using the validation dataset of four batch processes and a single fed-batch bioprocess (65 samples), and the corresponding model performances, regression coefficient vectors (RCVs), and high concentration single compound spectra are shown in Figure 4.2.

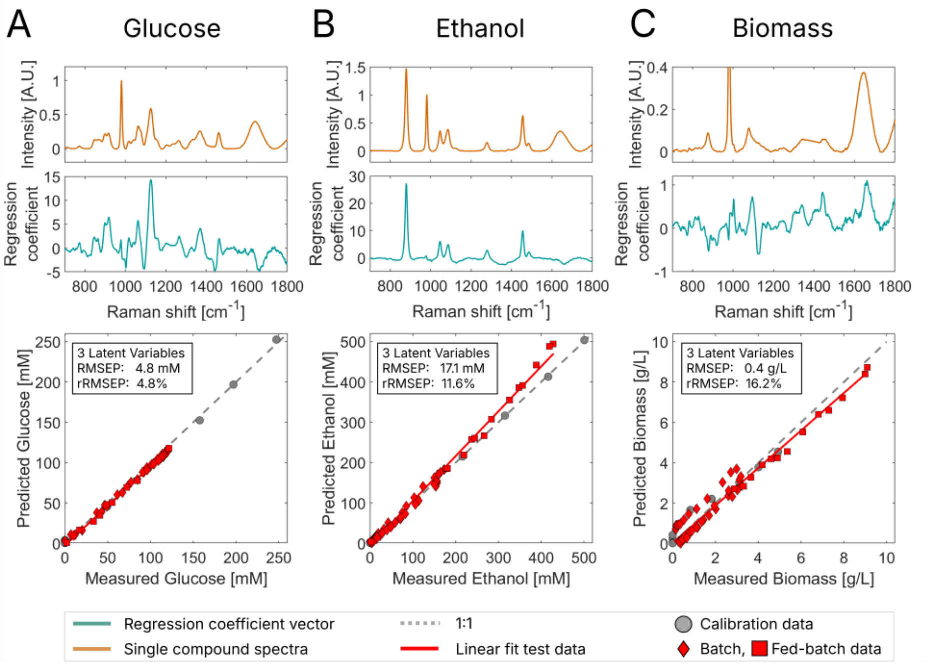


Figure 4.2: Pre-processed spectra (before mean-centering, top), regression coefficient vector (middle), and measured versus predicted plots (bottom) of the glucose (A), ethanol (B), and biomass (C) Partial Least Squares (PLS) models, respectively.

Quantitative analysis of model performance resulted in rRMSEP values of 4.8%, 11.6%, and 16.2% for glucose, ethanol, and biomass, respectively. Three latent variables were selected for each model, resulting in low RMSEC and RMSECV values by capturing the variation of glucose, ethanol, and biomass in separate components (Supplementary Figure 4.5.1.3). Glucose concentrations were accurately quantified across the batches and the fed-batch bioprocess. Qualitative model assessment indicates that the RCV of each model contains the key spectral features of their compound of interest, while correcting for overlapping spectral features. Glucose model specificity is reflected by the strong representation of the peaks for COH-bending (918 cm^{-1} and 1125 cm^{-1}), CO-stretching (1066 cm^{-1}), and CH-bending (1368 cm^{-1}) [19] in the RCV, and overlapping peaks of ethanol are corrected (e.g., 879 cm^{-1} and 1085 cm^{-1}). The ethanol model showed good prediction accuracy on the batch bioprocesses (rRMSEP of 6.74%), but the predictions on the fed-batch data deviated from the 1:1 line, leading to the overall rRMSEP of 11.6%. The ethanol model RCV closely resembles the single compound spectrum of ethanol, indicated by high coefficients for the CC-stretching (879 cm^{-1}), CO-stretching (1046 cm^{-1}), CH_3-rocking (1085 cm^{-1}), CH_2-twisting (1277 cm^{-1}), and CH_3-deformation (1456 cm^{-1}) peaks [20, 21]. An inspection of the residuals on the fed-batch samples revealed multiple regions where true and fitted spectra deviated (1250-1480 cm^{-1} and 1560-1660 cm^{-1}), but the pattern could not be directly related to a known compound.

Our previous work showed that spectral features associated with the molecular composition of biomass can be detected with in-line Raman spectroscopy after correcting for the extinction effect with a normalization step [22]. This is reflected by the biomass model RCV that contains spectral features matching with Raman spectroscopy studies of *S. cerevisiae*, displaying positive coefficients for bands related to phenylalanine (1002 cm^{-1}), phospholipids (1084 cm^{-1}), CH-deformation of proteins (1344 cm^{-1}), CH_2-deformation of lipids and proteins (1448 cm^{-1}), and amide I stretching (1669 cm^{-1}) [23]. In the work of Yang et al. (2024), PLS models for monitoring yeast bioprocesses were calibrated with single compound spectra of glucose, ethanol, peptone, yeast extract, and biomass, but no specific signal was found for the yeast cells [24]. Instead, signal extinction caused by biomass was modelled by measuring mixtures of glucose and ethanol at varying concentrations of biomass, and the non-linear relation was used to quantify biomass and correct predictions of the PLS models during bioprocessing. Other literature on monitoring of yeast with Raman spectroscopy mainly highlight the attenuation of spectral features with increasing biomass concentration, and there is little data on the detection of its protein and lipid signal during in-line measurements.

The results in Figure 4.2 show an accurate prediction of the biomass concentration in three out of four batches, while the predictions for one batch and the fed-batch bioprocess deviated from the 1:1 line. Further inspection of the residuals of the deviating batch dataset highlighted large differences in the water peak at 1640 cm^{-1}, but the cause of these differences could not be determined. For the fed-batch bioprocess, the biomass model had to extrapolate, as the single biomass spectra only reached 5 g/L while the fed-batch process went up to 9.2 g/L. The two lowest concentration biomass spectra from the calibration dataset were overpredicted during model development, and the need for extrapolation on the fed-batch data could have propagated this effect outside of the calibration concentration range.

The latent variable loadings of each model show that the third latent variable did not contain more than 0.45% of the spectral variation (Supplementary Figure 4.5.1.3). The vast majority of the spectral variation belonging to glucose and ethanol is explained in latent variables one and two (loadings of 28.6-70.9%), which may result from the difference in signal strength between biomass and the metabolites. To investigate the impact of including biomass spectra during calibration on overall model performance, models calibrated with only glucose and ethanol spectra were tested and applied to the same validation dataset (Supplementary Figure 4.5.1.4). This resulted in models with two latent variables, where the rRMSEP of the glucose model without biomass in the calibration dataset increased to 13.4% (from 4.8%) and the rRMSEP of the ethanol model increased to 14.8% (from 11.6%). The RCVs show that including biomass spectra in the calibration dataset allows the model to correct for overlapping spectral features (e.g., 1448 cm^{-1} and 1669 cm^{-1}). Moreover, the broad features of both glucose and biomass overlap over a large section of the spectra. Although these effects are less visible in the RCVs of the ethanol models due to narrow peaks, the inclusion of biomass spectra and the subsequent selection of an additional latent variable led to a higher predictive performance. Thus, despite the small magnitude of the biomass spectral signal, including biomass spectra improved spectral decomposition of the molecular features of yeast and increased prediction performance on bioprocess data.

The performance decrease seen for the fed-batch samples to quantify ethanol and biomass could also be related to the pre-processing strategy. Single compound spectra were acquired in individual bioreactor setups, which resulted in baseline offsets between the experiments (Supplementary Figure 4.5.1.1). An Automatic Whittaker baseline correction ($\lambda = 10000$, $\alpha = 0.001$) was utilized to achieve baseline alignment. After baseline correction, spectra needed to be normalized for intensity and a normalization based on the sulphate peak of the synthetic medium as an internal standard yielded the best results. However, the fed-batch was bolus fed with 50% sterile glucose solution three times, slightly diluting the sulphate peak, thereby compromising the intensity correction. This underlines the downside of utilizing internal standards for intensity normalization, as process adjustments can directly influence pre-processing accuracy. Despite these challenges, the sulphate peak normalization provided the most accurate models, and other normalization methods (e.g., standard normal variate) resulted in high prediction errors.

4.3.2 Approach 2: Indirect Hard Model (IHM) calibrated with single compound spectra

The 16 single compound spectra were used to generate HMs, where individual spectral features of each compound are modeled as Pseudo-Voigt profiles. To prevent the inclusion of noise into the HMs, fitted peaks were cross-checked with literature, resulting in 21 peaks for glucose [19], 8 peaks for ethanol [20, 21], and 9 peaks for biomass [23]. The fitted single compound models showed a high similarity to other work in literature using the IHM approach [10]. The individual HMs were combined to form a mixture model, which was calibrated on the single compound dataset. The calibrated mixture model was subsequently applied to evaluate performance with the bioprocess validation dataset (Figure 4.3). A detailed overview of the workflow is provided in Supplementary Figure 4.5.2.2.

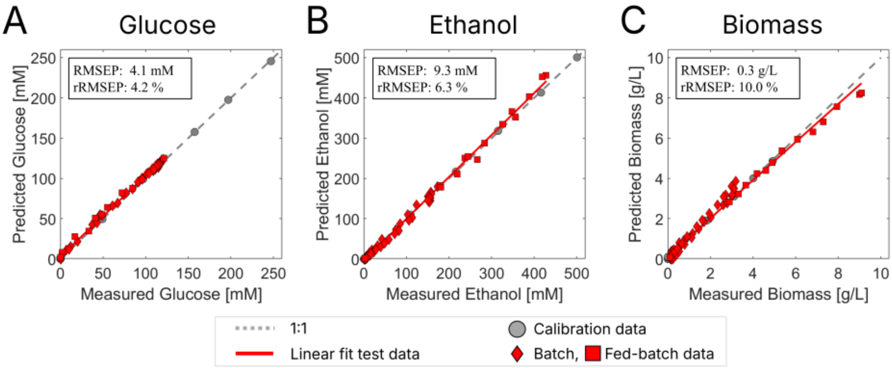


Figure 4.3: The measured (x-axis) versus predicted (y-axis) plots of an Indirect Hard Modelling (IHM) model applied to a bioprocess dataset consisting of 4 batches and a single fed-batch. The model quantified glucose (A), ethanol (B), and biomass (C).

Quantitative model assessment resulted in rRMSEP values of 4.2%, 6.3%, and 10.0% for glucose, ethanol, and biomass, respectively. The predictions showed high linearity for glucose and ethanol, with a slight overprediction of glucose. The prediction accuracy for biomass was considered decent for the batch bioprocesses, but the predictions for the late fed-batch samples deviated from the 1:1 line, where the model had to extrapolate past the 5 g/L upper limit of the calibration data. Another factor leading to the decrease in prediction accuracy for biomass in the late fed-batch samples may be the broad spectral features with low specificity obtained using the complemental hard modeling approach, even though the absolute position of each fitted peak closely matched the features associated with *S. cerevisiae* reported in literature [23].

The IHM approach was successfully applied by Muller *et al.* to monitor glucose and ethanol concentrations during yeast bioprocessing in a 20 mL cuvette setup, where Raman spectra were acquired with a Raman microscope through the bottom of the glass cuvette [10, 25]. A total of 11 mixture spectra plus a single measurement of yeast suspension were used to calibrate their model, and glucose and ethanol concentrations of around 100 g/L and 50 g/L, respectively were successfully quantified. Our model was calibrated without the need for mixture spectra, and glucose and ethanol concentrations only reached 21 g/L and 23 g/L, respectively, thus resulting in a weaker Raman signal. Despite the inherent differences between experimental setups and lower concentration ranges, our mixture model accuracy is in the same order of magnitude, as our RMSEPs for glucose and ethanol were 0.74 mg/g and 0.43 mg/g versus their 3.68 mg/g and 1.70 mg/g. It should be noted that the use of an immersion probe inside the bioreactor led to high signal extinction by biomass, and despite the signal attenuation our model provided accurate predictions after a simple normalization step.

4.3.3 Approach 3: PLS model calibration with synthetic spectra

In this section, we evaluate PLS model performance when calibrated with synthetic spectra. The HMs obtained in the previous section were extracted and utilized to generate de novo Raman spectra of custom concentration ratios. A total of 125 synthetic spectra were generated according to a full factorial design, with ranges of 0-200 mM for glucose, 0-500 mM for ethanol, and 0-5 g/L for biomass, including 5 concentration steps for each compound (Supplementary Figure 4.5.3.1). The synthetic dataset was subsequently used to calibrate PLS models for the quantification of glucose, ethanol, and biomass, and applied to predict four batch and one fed-batch bioprocess datasets (Figure 4.4).

PLS models calibrated on the dataset of 125 synthetic spectra resulted in rRMSEP values of 3.2%, 14.5%, and 256.0% for glucose, ethanol, and biomass, respectively. The synthetic spectra managed to simulate spectral variation of glucose closely, resulting in a more accurate prediction than direct calibration with the 16 experimental single compound spectra (1.6 percent point). The RCVs of the glucose models calibrated on experimental and synthetic data were similar, with the synthetic model containing less noise due to the smooth nature of the Pseudo-Voigt profiles. The ethanol model calibrated on synthetic spectra had a slightly higher (2.9 percent point) prediction error than the model calibrated on the experimental single compound spectra, but the RCVs were still considered highly similar between the models. The RCV similarity seen for ethanol models also resulted in a comparable deviation for the prediction of ethanol concentrations in the fed-batch dataset.

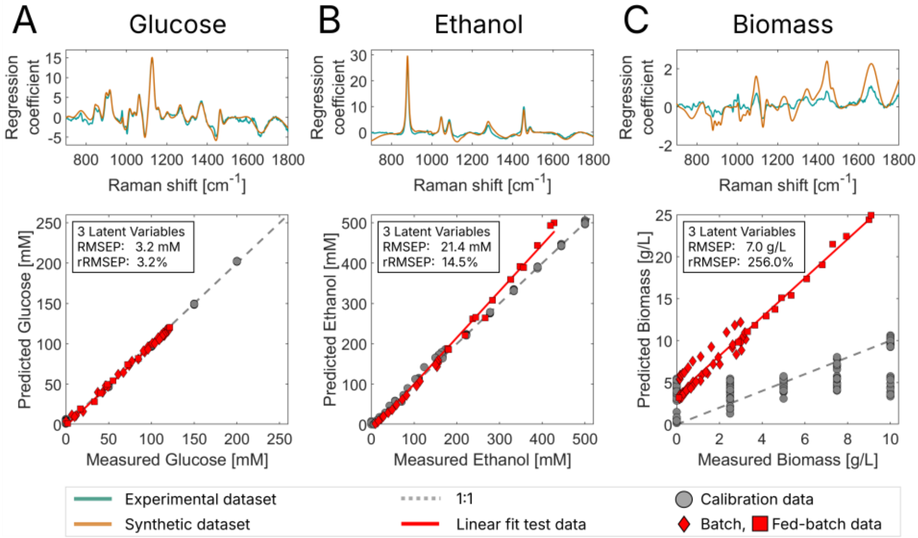


Figure 4.4: The regression coefficient vectors (RCV, top) of the Partial Least Squares (PLS) model calibrated with experimental data (orange), calibrated with synthetic spectra (blue), and the measured versus predicted plots (bottom) of the PLS models calibrated on synthetic spectra for glucose (A), ethanol (B), and biomass (C).

The poor performance of the biomass model clearly indicates that synthetic spectra cannot properly replicate the spectral variation caused by biomass. The broad spectral features of biomass extracted from the complemental hard modeling method did not accurately represent the true spectral contribution of *S. cerevisiae*, and this propagated to the synthetic spectra. Furthermore, spectral variation caused by biomass was a small percentage of the total spectral variation, as seen in the PLS models calibrated with experimental single compound spectra directly (section 3.1). The small magnitude of the biomass signal possibly caused a high sensitivity to small deviations in intensity, leading to difficulties of recreating the proper signal proportions. In addition, the broad shapes of the simulated biomass signal is sensitive to intensity changes by baseline correction steps.

This approach shows synthetic spectra allow for setting custom ratios between glucose and ethanol, by which the design space could be covered entirely without additional experimental effort. However, calibration with synthetic spectra did not lead to improved model performance for ethanol and biomass compared to calibration with 16 experimental single compound spectra. Furthermore, the use of synthetic spectra only led to very minor differences in the model RCVs for glucose and ethanol, supporting the lack of added benefit in terms of model performance or specificity. Nevertheless, this approach demonstrated the ability to expand a dataset with spectra highly similar to the process conditions, while maintaining the linear correlation between signal intensity and compound concentration. If this method can be expanded with HMs of additional compounds, it can generate high variability datasets without the need of collecting process data.

4.3.4 Approach 4: PLS model calibration with augmented process spectra

Calibrating quantification models with (repeated) process data limits the design space, which may lead to incorporation of cross-correlations and hinders model robustness [26]. However, generating process spectra de novo as discussed in section 3.3 is limited by the availability of HMs for all process compounds. In many applications, process knowledge is minimal and single compound spectra can only be acquired for a few compounds. This section investigates the augmentation of a small dataset of process spectra with spectral features of the compound of interest obtained from HMs. This approach utilizes the standard spectral variation of a small process dataset (a single batch) while attempting to improve the specificity of models towards a compound of interest, without needing to define other process compounds. A batch dataset of 12 spectra was expanded by synthetically modifying the concentration of either glucose or ethanol, up to a total of 60 spectra (Supplementary Figure 4.5.4.2). Augmenting spectra with the isolated biomass features was not considered because these features exhibited low specificity in section 3.3. PLS models were calibrated with the standard (12 spectra) and augmented (60 spectra) dataset, and applied to the reduced validation dataset (3 batches, 1 fed-batch). The prediction performance of these models on bioprocess data is shown in Figure 4.5.

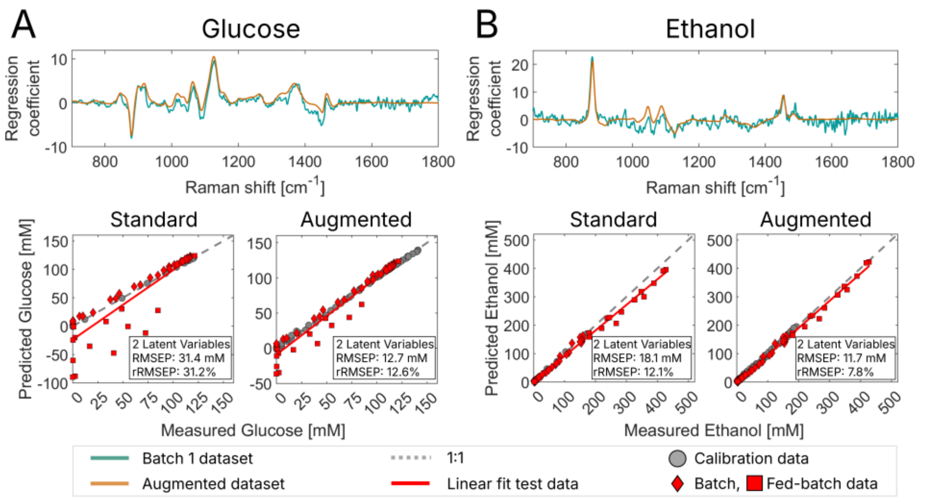


Figure 4.5: The performance of partial least squares (PLS) models quantifying glucose (A) and ethanol (B) while calibrated on a standard batch dataset (12 process spectra) and augmented batch spectra (12 process spectra and 48 augmented spectra). The calibration datasets were expanded by augmenting the spectral features of the quantification target by ±10 and ±20mM.

Despite the improvements in model performance and increased specificity of the RCVs, the loadings of latent variable 1 did not show large differences from those of the standard models and mainly captured batch evolution (Supplementary Figure 4.5.4.3). This is expected, as the augmented datasets followed an identical concentration trend to the standard batch bioprocess and the spectra were only slightly adjusted in concentration (Supplementary Figure 4.5.4.1). Modifying process spectra with concentrations at the extremes of the process design space to break cross-correlations (e.g., adjusting a true concentration of 20 mM to 140 mM) led to large prediction errors and non-linear effects. The errors in these cases could be attributed to spectral normalization, where we did not manage to equally normalize process spectra and synthetic spectral features before combining the two components in augmented spectra. As a result, biomass extinction effects in the batch spectra propagate throughout the augmentation process, underlining how augmentation approaches are highly sensitive to small intensity changes not related to concentration changes, as these disrupt signal linearity.

This section showcased how Pseudo-Voigt profiles isolated from single compound spectra can be utilized to customize the concentration of specific compounds in process spectra. This allows for expansion of a dataset's design space while minimizing the impact on other spectral features in the data. The augmentation approach can be used to artificially spike concentrations of compounds of interest without the need for extensive experimental setups where cell cultures cannot be recovered after spiking. However, to optimize augmentation methods, adaptive normalization techniques are necessary that transfer across spectra without the need for internal standards.

4.3.5 Discussion on modelling approaches

This section discusses the four modelling approaches demonstrated in this work. A comparison of the quantitative performance of each model, the calibration data used, and the complexity of each modelling approach is shown in Table 1. Time-evolution plots of the model predictions from all four approaches are provided in Supplementary Figures 4.5.5.1–4.5.5.4.

Table 4.1: Comparison of the four modelling approaches discussed in this work. The relative Root Mean Square Error of Prediction (rRMSEP) on the validation dataset is provided for each model and compound to directly compare quantitative performance between the methods.

Approach	Model type	Calibration dataset		Experimental time	Computational time	Model flexibility	rRMSEP		
		Total spectra	Spectral source				Glucose	Ethanol	Biomass
1	PLS	16	Experimental	Low	Low	Medium	4.8%	11.6%	16.2%
2	IHM	16	Experimental	Low	Medium	High	4.2%	6.3%	10.0%
3	PLS	125	Synthetic	Low	High	Medium	3.2%	14.5%	256.0%
4	PLS	12 + 48	Process + Augmented	High	High	Low	12.6%	7.8%	-

Approach 1 showed that PLS models calibrated using 16 experimentally obtained single compound spectra can achieve decent prediction performance without the need of collecting mixture (process) spectra. This approach resulted in compound specific models, reflected by distinct peaks in the model RCVs. Moreover, models obtained with Approach 1 outperformed models calibrated on batch process data supplemented with single compound spectra for biomass prediction (rRMSEP of 27.0%), as demonstrated in previous work [26]. The observed improvement for biomass quantification resulted from a higher quality of single-compound biomass measurements, which were consistent with spectral features reported in literature [22, 23]. Despite these improvements, the implicit PLS models are not expected to perform well outside of their calibration design space, as the performance relies on empirical relationships learned from training data and the models do not contain physical understanding of the monitored process. From this perspective, the semi-explicit IHM used in Approach 2 has some inherent advantages over implicit modeling techniques. Once individual HMs of the main process compounds are available, a mixture model can perform predictions based on chemical principles. Moreover, the baseline itself is defined as a process component (consisting of mainly water), thereby reducing the dependency on spectral pre-processing. In addition, IHM also offers more flexibility in situations where a novel or unknown compound is present. For example, spectral variation of glycerol and acetate was assumed negligible for this application, but when such a compound becomes more abundant in the process, its single compound spectra could be used to generate a hard model which is subsequently added to the mixture model. An alternative route would be fitting Pseudo-Voigt features to the residuals between the old mixture model and new process spectra to generate the model of an additional component [10]. Approach 2 is therefore considered highly flexible, as a database of hard models can be easily expanded with new compounds and can be calibrated based on simple measurements. In addition, where PLS models rely on learned weights at specific wavenumbers, the IHM approach can be tuned to allow peak shifts and shape changes during component fitting, leading to higher model robustness for changing measurement conditions or compound interactions. This tunability could also make quantification models perform better under extrapolation conditions as long as the number of process compounds does not change. However, these advantages come with increased modelling complexity, as HMs must be generated for all major process compounds, and the final mixture model must fit each HM to the process spectra for each prediction. This means that Approach 2 can become challenging for complex process mixtures where process knowledge and access to single compound spectra is limited. It is therefore important to note that the strength of PLS models lies in their simplicity and their ability to extract key spectral features of the target compound from complex spectra, thereby reducing the need for extensive process knowledge. This is considered beneficial for cell cultures with more complex media (e.g., for Chinese Hamster Ovary cells), where the number of relevant process compounds increases rapidly, and gaining complete spectral knowledge is challenging.

Calibrating PLS models with 125 synthetic mixture spectra during Approach 3 did not result in model improvements when compared to calibration with only the 16 experimental single compound spectra. Biomass prediction accuracy was particularly poor, likely due to the low specificity of the biomass HM. In addition, PLS models can benefit from calibration with experimental mixture spectra as interactions between compounds could influence the position and shape of their spectral features. The synthetic spectra generated in this work did not include these interactions, as each compound was modeled from single compound spectra.

However, when compared to automated methods for spectra generation such as GANs or SMOTE, our approach can maintain physically accurate linear relationships between signal and concentration, provided that two key assumptions are met. First, signal intensity must change linearly with compound concentration withing the calibration range, supported by our calibration lines based on five single compound spectra. Second, scaled HMs for individual compounds must combine additively to represent mixture spectra (Equation 2). Under these conditions, our method allows synthetic spectra generation for any concentration ratio within the single compound spectra measured range.

Another aspect that was considered challenging was matching the intensity between synthetic, augmented, and process spectra. We exclusively utilized the sulphate peak at 981 cm^{-1} as an internal standard for normalization, which is also reported for the application of Raman spectroscopy for other yeast bioprocesses [27, 28]. However, using internal standards for normalization should be done with great caution as they are dependent on a single variable, and therefore sensitive to changes in measurement conditions [24]. Furthermore, internal standards are not available in every measurement matrix, and alternative normalization methods should be explored when generating synthetic data. In this work, a robust method for normalizing the intensity of individual spectral components during spectra synthesis and augmentation was not found. Moreover, disruptions in signal linearity might occur at every spectral modification step, including normalization, and errors in intensity can propagate to the final spectrum.

The concept of Approach 4, which allows the artificial adjustment of spectral features related to target compounds, holds potential for complex processes where both process data availability and knowledge are limited. Since single compound spectra for common quantification targets (e.g., metabolites and products) can be easily measured, their key spectral features can be extracted to build HMs. The HMs can subsequently be used to enhance the spectral variation of this target compound within complex mixtures, even in situations where detailed process knowledge is missing. However, maintaining linearity between signal intensity and compound concentration is essential for linear regression techniques like PLS, but this is often disrupted by noise and scattering effects present in bioreactor spectra [22].

Despite the challenges highlighted in this work, the ability to generate synthetic and augmented spectra that accurately simulate process conditions can be valuable for calibrating quantification models for Raman spectroscopy. One of the largest hurdles for calibrating robust quantification models is the need for extensive data collection, especially capturing process states outside typical operational patterns, which can be crucial for improving model accuracy. The operation of bioreactor processes at different compound concentrations could provide valuable spectral information, but requires substantial time and material if solely performed for improving Raman spectroscopy quantification models. In addition, literature reports studies that investigated the effectiveness of compound spiking to generate this valuable data, but this typically leads to the loss of a cell culture [6]. The options to generate these valuable conditions synthetically or to augment existing process spectra towards the edges of the desired design space could provide efficient and low-effort alternatives.

4.4 Conclusion

Raman spectroscopy coupled with accurate quantification models serves as a powerful tool for monitoring bioreactor processes. Nevertheless, quantification model calibration is often labor-intensive and requires extensive experimental efforts. Furthermore, collecting large process datasets to calibrate these models often results in process-specific models with a narrow design space, highlighting the need for flexible methods to collect data and expand small process datasets. This study investigated four approaches by which a small dataset of 16 single compound measurements could be utilized to calibrate quantification models for glucose, ethanol, and biomass during *S. cerevisiae* bioprocesses.

The single compound dataset was used to calibrate quantification models using Partial Least Squares (PLS, Approach 1) and Indirect Hard Modelling (IHM, Approach 2). Both modelling approaches showed similar performance when comparing the rRMSEP values for glucose (4.8% and 4.2%), ethanol (11.6% and 6.3%), and biomass (16.2% and 10.0%). The PLS approach demonstrated how isolated biomass measurements incorporate spectral features associated with the molecular composition of yeast, while the IHM approach proved to be a robust and flexible method that can be easily extended to accommodate new process compounds or conditions.

The compound hard models were also applied to synthetically generate Raman spectra to directly calibrate PLS models (Approach 3) and to augment experimental process data to increase model specificity (Approach 4). Direct calibration with synthetic spectra proved effective for glucose and ethanol quantification PLS models, with rRMSEP values of 3.2% and 14.5%, respectively. Due to difficulties in isolating sharp spectral features for biomass, calibration of PLS models with synthetic spectra did not result in accurate biomass quantification. Spectral augmentation of a single batch bioprocess dataset led to rRMSEPs of 12.6% and 7.8% for glucose and ethanol, respectively, compared to 31.2% and 12.1% when calibrated solely on the standard batch data. The synthetic generation and augmentation of Raman spectra showed potential for the enhanced calibration of PLS models, but robust normalization steps are required to maintain signal integrity during these processes.

Overall, this work showcased multiple approaches by which simple spectral measurements can be applied to calibrate quantification models for bioprocesses, without the need for (additional) process data. This means that quantification models for yeast bioprocesses can be developed even before running the actual process, and models can be easily adapted to changes in process conditions or when transferring between processes. Furthermore, the possibility to augment existing spectra of complex processes enables model calibration improvement without extensive spectral knowledge of the system. The use of single compound, synthetic, or augmented Raman spectra supports efficient quantification model calibration, thereby simplifying the implementation of Raman spectroscopy for bioreactor monitoring.

Acknowledgements We would like to thank Christiaan Mooiman and Jeroen Schmidt for their technical support while setting up the lab and during the fermentations.

References

1. Esmonde-White, K.A., M. Cuellar, and I.R. Lewis, *The role of Raman spectroscopy in biopharmaceuticals from development to manufacturing.* Analytical and Bioanalytical Chemistry, 2021: p. 1-23.
2. Tanemura, H., et al., *Comprehensive modeling of cell culture profile using Raman spectroscopy and machine learning.* Scientific Reports, 2023. **13**(1): p. 21805.
3. Zavala‐Ortiz, D.A., et al., *Comparison of partial least square, artificial neural network, and support vector regressions for real‐time monitoring of CHO cell culture processes using in situ near‐infrared spectroscopy.* Biotechnology and Bioengineering, 2022. **119**(2): p. 535-549.
4. Wold, S., M. Sjöström, and L. Eriksson, *PLS-regression: a basic tool of chemometrics.* Chemometrics and intelligent laboratory systems, 2001. **58**(2): p. 109-130.
5. Romann, P., et al., *Advancing Raman model calibration for perfusion bioprocesses using spiked harvest libraries.* Biotechnology Journal, 2022: p. 2200184.
6. Santos, R.M., et al., *Monitoring mAb cultivations with in‐situ Raman spectroscopy: The influence of spectral selectivity on calibration models and industrial use as reliable PAT tool.* Biotechnology progress, 2018. **34**(3): p. 659-670.
7. Alsmeyer, F., H.-J. Koß, and W. Marquardt, *Indirect spectral hard modeling for the analysis of reactive and interacting mixtures.* Applied spectroscopy, 2004. **58**(8): p. 975-985.
8. Kriesten, E., et al., *Identification of unknown pure component spectra by indirect hard modeling.* Chemometrics and Intelligent Laboratory Systems, 2008. **93**(2): p. 108-119.
9. Echtermeyer, A., et al., *Inline Raman spectroscopy and indirect hard modeling for concentration monitoring of dissociated acid species.* Applied spectroscopy, 2021. **75**(5): p. 506-519.
10. Müller, D.H., et al., *Bioprocess in‐line monitoring using Raman spectroscopy and Indirect Hard Modeling (IHM): A simple calibration yields a robust model.* Biotechnology and Bioengineering, 2023.
11. Hao, Y., X. Li, and C. Zhang, *Improving prediction model robustness with virtual sample construction for near-infrared spectra analysis.* Analytica Chimica Acta, 2023. **1279**: p. 341763.
12. Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique.* Journal of artificial intelligence research, 2002. **16**: p. 321-357.
13. Goodfellow, I., et al., *Generative adversarial networks.* Communications of the ACM, 2020. **63**(11): p. 139-144.
14. Wu, M., et al., *Deep learning data augmentation for Raman spectroscopy cancer tissue classification.* Scientific reports, 2021. **11**(1): p. 23842.
15. Goldrick, S., et al., *Modern day monitoring and control challenges outlined on an industrial-scale benchmark fermentation process.* Computers & Chemical Engineering, 2019. **130**: p. 106471.
16. Sulub, Y. and G.W. Small, *Spectral simulation methodology for calibration transfer of near-infrared spectra.* Applied spectroscopy, 2007. **61**(4): p. 406-413.
17. Nijkamp, J.F., et al., *De novo sequencing, assembly and analysis of the genome of the laboratory strain Saccharomyces cerevisiae CEN. PK113-7D, a model for modern industrial biotechnology.* Microbial cell factories, 2012. **11**(1): p. 1-17.
18. Verduyn, C., et al., *Effect of benzoic acid on metabolic fluxes in yeasts: a continuous‐culture study on the regulation of respiration and alcoholic fermentation.* Yeast, 1992. **8**(7): p. 501-517.
19. Dudek, M., et al., *Raman Optical Activity and Raman spectroscopy of carbohydrates in solution.* Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2019. **206**: p. 597-612.

20. Boyaci, I.H., et al., *A novel method for quantification of ethanol and methanol in distilled alcoholic beverages using Raman spectroscopy.* Journal of Raman Spectroscopy, 2012. **43**(8): p. 1171-1176.
21. Pappas, C., et al., *Evaluation of a Raman spectroscopic method for the determination of alcohol content in Greek spirit Tsipouro.* Current Research in Nutrition and Food Science Journal, 2016. **4**(Special Issue Nutrition in Conference October 2016): p. 01-09.
22. Klaverdijk, M., et al., *Impact of bioreactor process parameters and yeast biomass on Raman spectra.* Biotechnology Progress, 2025: p. e70050.
23. Wang, K., et al., *Species identification and strain discrimination of fermentation yeasts Saccharomyces cerevisiae and Saccharomyces uvarum using Raman spectroscopy and convolutional neural networks.* Applied and Environmental Microbiology, 2023. **89**(12): p. e01673-23.
24. Yang, N., et al., *In-line monitoring of Bioreactor by Raman Spectroscopy: direct use of a standard--based model through cell--scattering correction.* Journal of Biotechnology, 2024.
25. Müller, D.H., et al., *Bioprocess in-line monitoring and control using Raman spectroscopy and Indirect Hard Modeling (IHM).* Biotechnology and Bioengineering, 2024. **121**(7): p. 2225-2233.
26. Klaverdijk, M., M. Ottens, and M.E. Klijn, *Single compound data supplementation to enhance transferability of fermentation specific Raman spectroscopy models.* Analytical and Bioanalytical Chemistry, 2025: p. 1-12.
27. Picard, A., et al., *In situ monitoring by quantitative Raman spectroscopy of alcoholic fermentation by Saccharomyces cerevisiae under high pressure.* Extremophiles, 2007. **11**(3): p. 445-452.
28. Hirsch, E., et al., *Inline noninvasive Raman monitoring and feedback control of glucose concentration during ethanol fermentation.* Biotechnology Progress, 2019. **35**(5): p. e2848.

4.5 Supplementary

4.5.1 Approach 1: PLS model calibration with single compound spectra

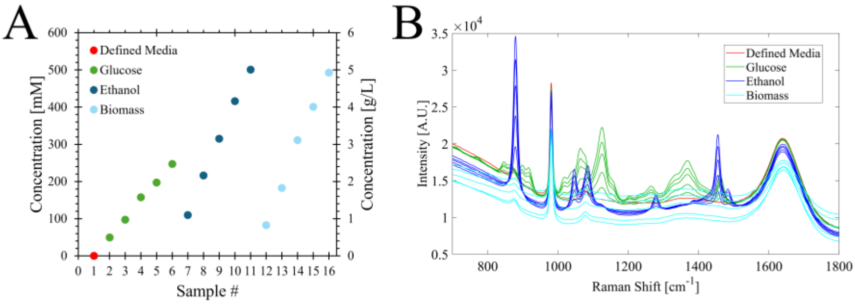


Figure 4.5.1.1: An overview of the concentrations (A) and spectra (B) of the single compound spectral dataset.

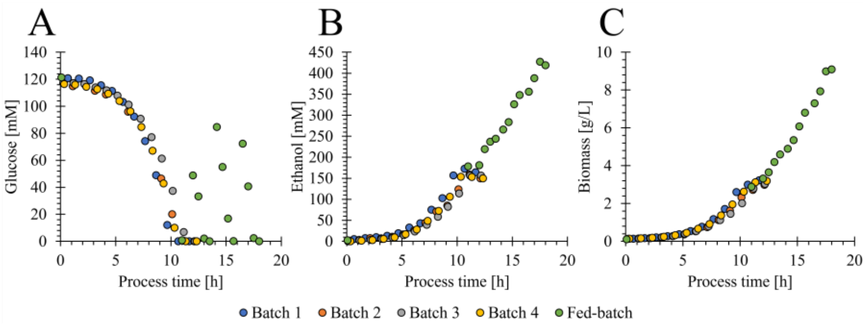


Figure 4.5.1.2: An overview of the concentrations of glucose (A) ethanol (B), and biomass of the fermentation validation dataset containing 4 batch and a single fed-batch process.

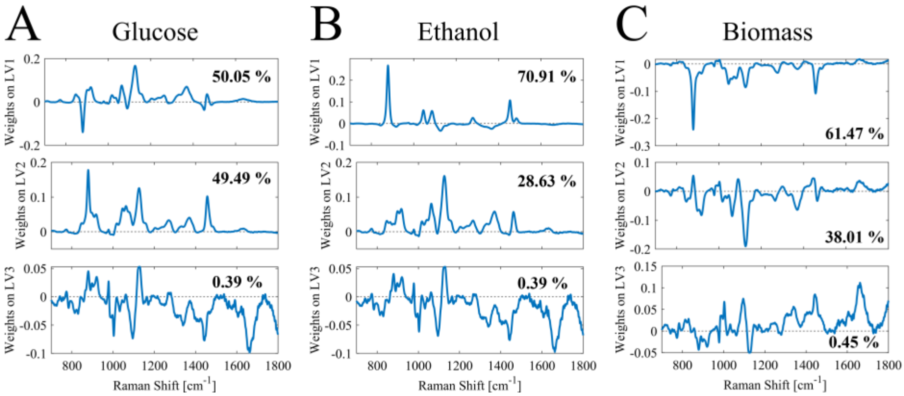


Figure 4.5.1.3: The latent variable loadings of the partial least squares (PLS) models calibrated on single compound spectra. The figures display the loadings and the percentage of captured variance of the glucose (A), ethanol (B), and biomass (C) models.

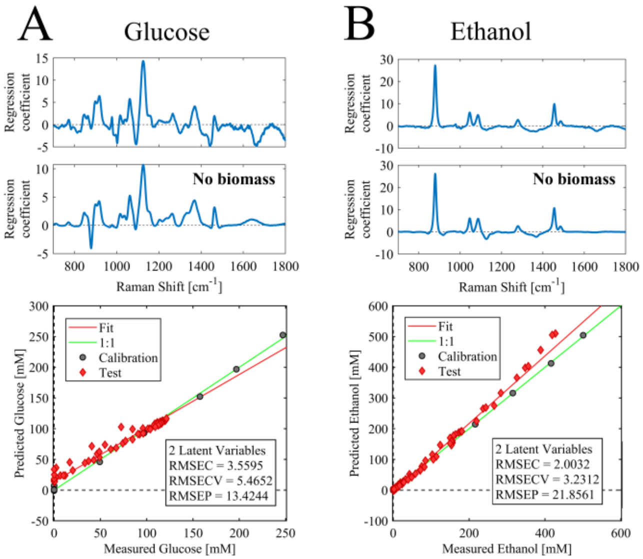
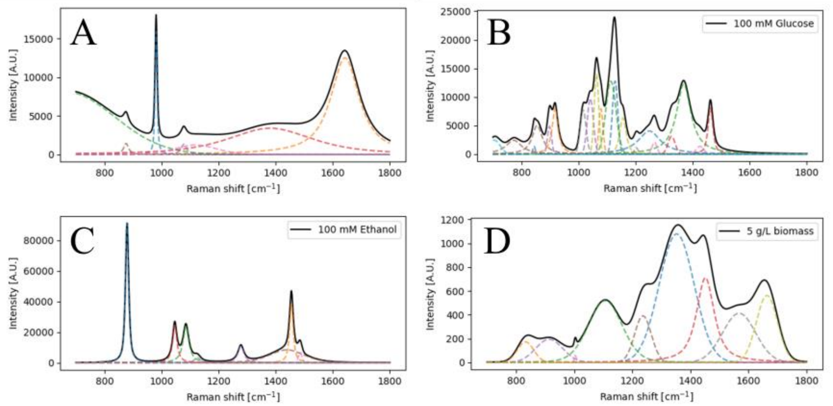


Figure 4.5.1.4: The performance and regression coefficient vectors (RCV) of partial least squares (PLS) models calibrated on the full single compound dataset (top) and only the glucose and ethanol spectra (bottom), plus the measured versus predicted plots of the models calibrated without biomass spectra. The models for glucose (A) and ethanol (B) are shown.

4.5.2 Approach 2: Indirect Hard Model (IHM) calibrated with single compound spectra



Defined media					Glucose					Ethanol					Biomass				
Peak	Position	Max	HWHM	Gaussian	Peak	Position	Max	HWHM	Gaussian	Peak	Position	Max	HWHM	Gaussian	Peak	Position	Max	HWHM	Gaussian
1	980.9	15849.7	4.4	0.3	1	1127.9	9612.6	10.9	1.0	1	879.2	17984.3	6.8	0.3	1	1352.5	16824.7	71.1	1.0
2	1642.6	12460.4	59.9	0.1	2	1078.3	5982.2	8.6	1.0	2	1454.9	7653.8	7.6	0.2	2	832.0	2700.2	33.5	1.0
3	664.7	8078.1	200.0	1.0	3	1368.4	9193.9	30.8	0.0	3	1085.3	4715.8	12.4	0.0	3	1106.6	8100.2	68.5	0.8
4	1380.7	3391.3	176.3	0.3	4	1462.2	6078.4	10.2	0.3	4	1046.2	4858.0	9.0	0.0	4	1451.5	11092.9	36.5	0.2
5	1077.8	1325.9	11.8	0.0	5	898.1	3659.2	8.5	1.0	5	1277.9	2132.5	12.5	0.0	5	913.6	3040.5	54.7	1.0
6	875.7	1419.3	10.7	0.1	6	855.0	3787.8	21.3	0.0	6	1438.6	1669.1	72.2	0.6	6	1236.5	6102.5	33.8	1.0
7	1119.1	1224.8	86.5	0.8	7	1265.9	1581.9	10.2	0.9	7	1485.7	1326.2	7.6	0.9	7	1003.5	949.6	4.3	1.0
					8	1016.1	5811.4	12.2	1.0	8	1125.6	657.9	12.9	0.3	8	1567.6	6440.9	66.8	1.0
					9	1155.5	4771.8	13.4	0.4						9	1663.4	8772.0	43.0	1.0
					11	1247.9	3060.6	47.3	0.0										
					12	917.8	6080.5	12.2	0.0										
					13	1113.2	9598.0	21.0	1.0										
					14	1325.8	2341.1	14.7	1.0										
					15	1041.2	7316.2	12.5	1.0										
					16	772.6	1831.5	33.2	0.0										
					17	1425.5	1043.9	15.5	0.9										
					18	1202.0	853.3	8.1	1.0										
					19	1062.7	10522.5	9.4	1.0										
					20	704.7	1848.7	24.2	1.0										
					21	844.1	1048.2	5.2	0.7										

Figure 4.5.2.1: The isolated peaks and total compound spectrum for defined media (A), glucose (B), ethanol (C), and biomass (D) extracted using complemental modeling.

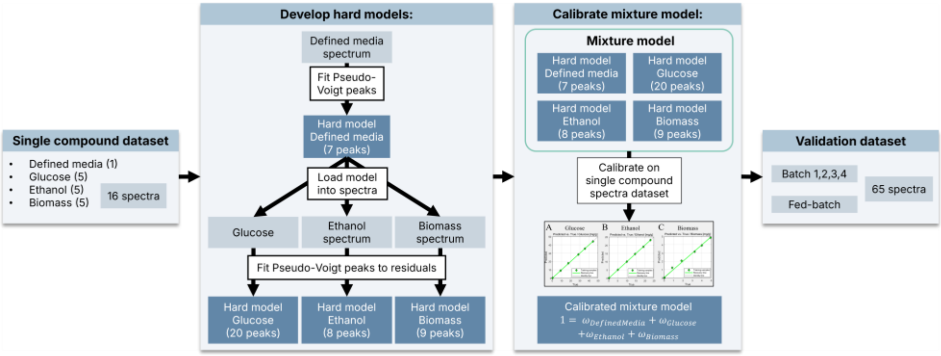

Single compound dataset
Defined media (1)
Glucose (5)
Ethanol (5)
Biomass (5)
16 spectra
Develop hard models:
Defined media spectrum
Fit Pseudo-Voigt peaks
Hard model Defined media (7 peaks)
Load model into spectra
Glucose
Ethanol spectrum
Biomass spectrum
Fit Pseudo-Voigt peaks to residuals
Hard model Glucose (20 peaks)
Hard model Ethanol (8 peaks)
Hard model Biomass (9 peaks)
Calibrate mixture model:
Mixture model
Hard model Defined media (7 peaks)
Hard model Glucose (20 peaks)
Hard model Ethanol (8 peaks)
Hard model Biomass (9 peaks)
Calibrate on single compound spectra dataset
A Glucose
B Ethanol
C Biomass
Calibrated mixture model
$1 = \omega_{DefinedMedia} + \omega_{Glucose} + \omega_{Ethanol} + \omega_{Biomass}$
Validation dataset
Batch 1,2,3,4
Fed-batch
65 spectra

Figure 4.5.2.2: Workflow used to generate the Indirect Hard Model (IHM). The single compound dataset was used to generate compound hard models. These were loaded into a mixture model, which was calibrated on the entire single compound dataset of 16 spectra, leading to linear correlations between model weight and concentration.

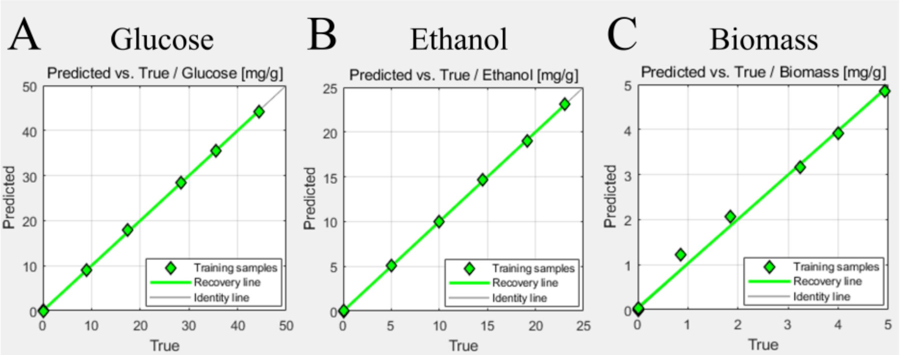


Figure 4.5.2.3: The calibration lines for glucose (A), ethanol (B), and biomass (C) of the fermentation indirect hard model.

4.5.3 Approach 3: PLS model calibration with synthetic spectra

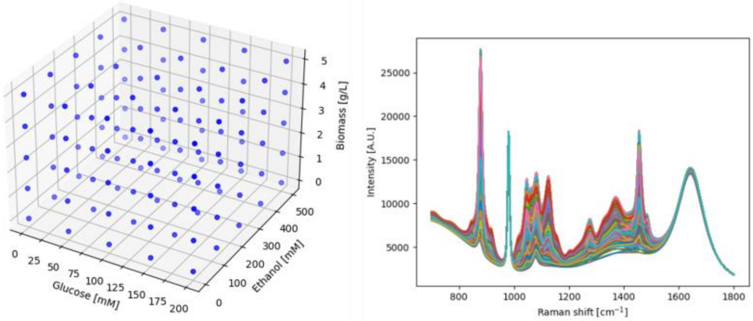


Figure 4.5.3.1: Scatterplot of the full factorial design used to generate synthetic Raman spectra (left), and the resulting synthetic spectra generated according to the concentrations (right). The synthetic spectra were generated according to this design by multiplying the compound hard model peak profiles with the appropriate weight.

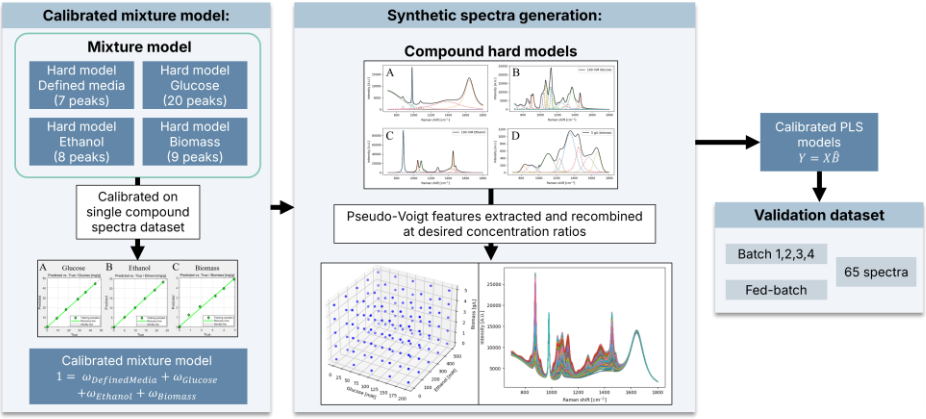

Calibrated mixture model:
Mixture model
Hard model Defined media (7 peaks)
Hard model Glucose (20 peaks)
Hard model Ethanol (8 peaks)
Hard model Biomass (9 peaks)
Calibrated on single compound spectra dataset
Calibrated mixture model
$1 = \omega_{DefinedMedia} + \omega_{Glucose} + \omega_{Ethanol} + \omega_{Biomass}$
Synthetic spectra generation:
Compound hard models
A
B
C
D
Pseudo-Voigt features extracted and recombined at desired concentration ratios
Calibrated PLS models
$Y = X\hat{B}$
Validation dataset
Batch 1,2,3,4
Fed-batch
65 spectra

Figure 4.5.3.2 The workflow used to generate synthetic spectra from the dataset of 16 single compound spectra. The Pseudo-Voigt profiles for each compound were extracted from the calibrated IHM model. These peak profiles were recombined in any desired ratio according to the mass balance, thereby generating synthetic spectra.

4.5.4 Approach 4: process spectra augmentation

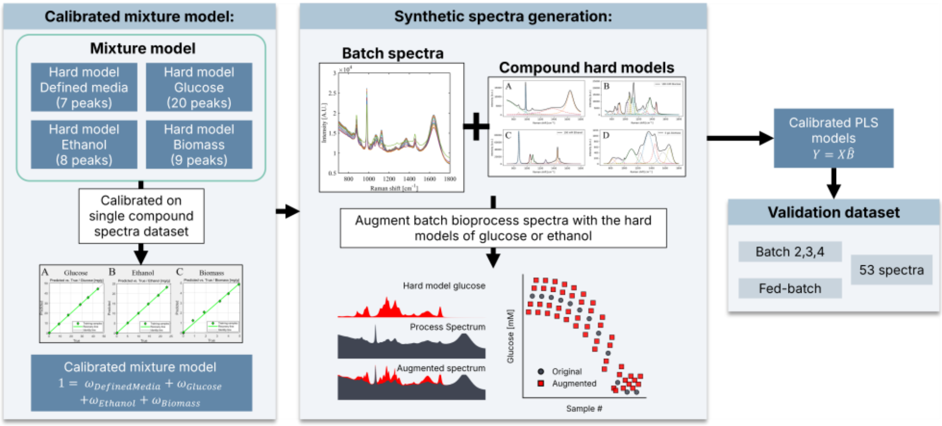

Calibrated mixture model:
Mixture model
Hard model Defined media (7 peaks)
Hard model Glucose (20 peaks)
Hard model Ethanol (8 peaks)
Hard model Biomass (9 peaks)
Calibrated on single compound spectra dataset
A Glucose
B Ethanol
C Biomass
Calibrated mixture model
$1 = \omega_{DefinedMedia} + \omega_{Glucose} + \omega_{Ethanol} + \omega_{Biomass}$
Synthetic spectra generation:
Batch spectra
Raman shift [cm⁻¹]
Intensity [A.U.]
Compound hard models
A
B
C
D
Augment batch bioprocess spectra with the hard models of glucose or ethanol
Hard model glucose
Process Spectrum
Augmented spectrum
Glucose [mM]
Original
Augmented
Sample #
Calibrated PLS models
$Y = X\hat{B}$
Validation dataset
Batch 2,3,4
Fed-batch
53 spectra

Figure 4.5.4.1: Workflow used to augment Raman spectra of a batch fermentation with the isolated spectral features of glucose and ethanol.

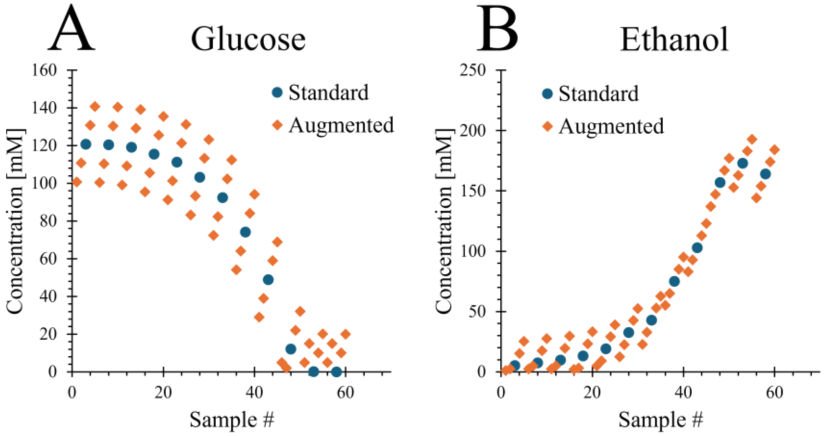


Figure 4.5.4.2: Scatterplots of the augmented dataset concentrations for glucose (A) and ethanol (B).The dataset was expanded by generating augmented spectra (orange diamonds) at ±10 and ±20 mM around the standard batch samples (blue circles).

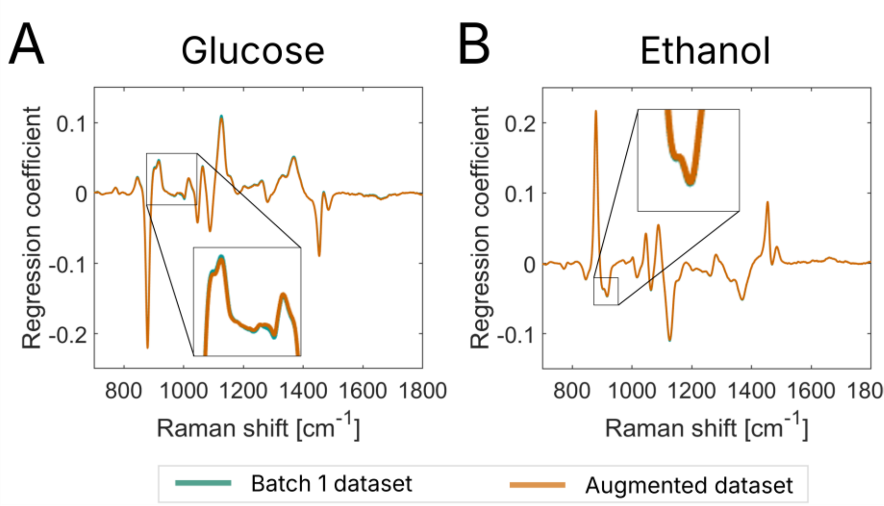


Figure 4.5.4.3: The regression coefficient vectors of models calibrated with the Batch 1 dataset (blue) versus models calibrated with the augmented dataset (orange).

4.5.5 Discussion on modelling approaches

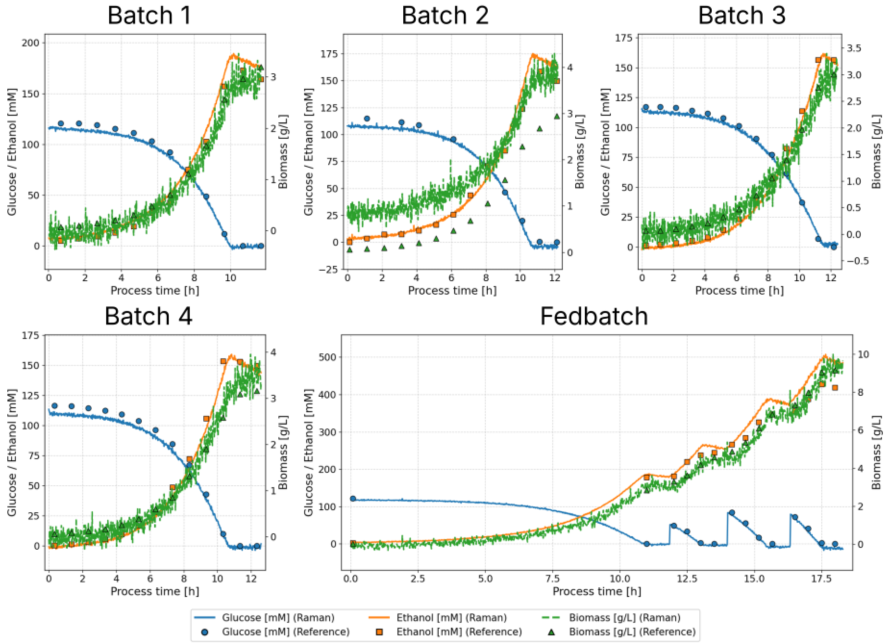


Figure 4.5.5.1: Time-evolution plots showing the quantitative predictions of the Raman models from Approach 1 for all spectra in the validation bioprocess dataset (lines), plotted alongside the corresponding reference measurements for those processes (markers).

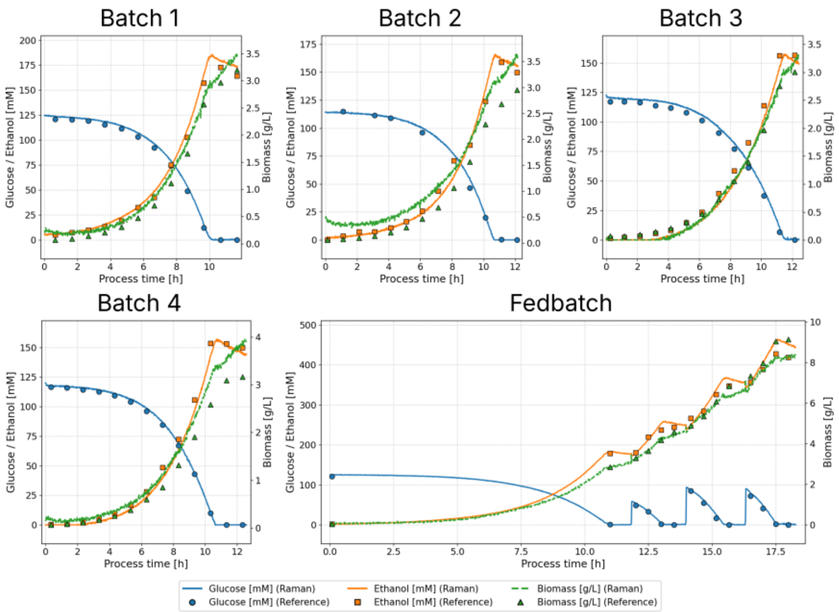


Figure 4.5.5.2: Time-evolution plots showing the quantitative predictions of the Raman models from Approach 2 for all spectra in the validation bioprocess dataset (lines), plotted alongside the corresponding reference measurements for those processes (markers).

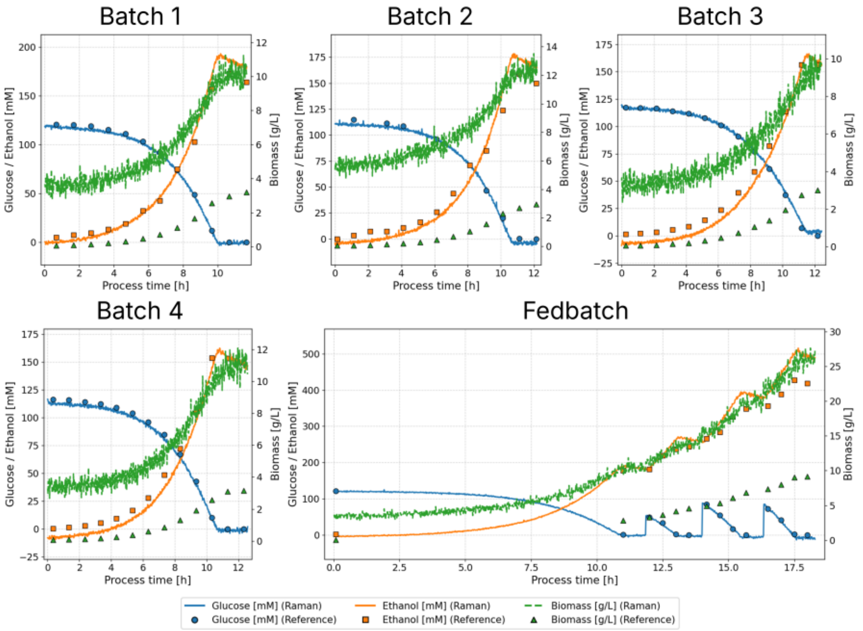


Figure 4.5.5.3: Time-evolution plots showing the quantitative predictions of the Raman models from Approach 3 for all spectra in the validation bioprocess dataset (lines), plotted alongside the corresponding reference measurements for those processes (markers).

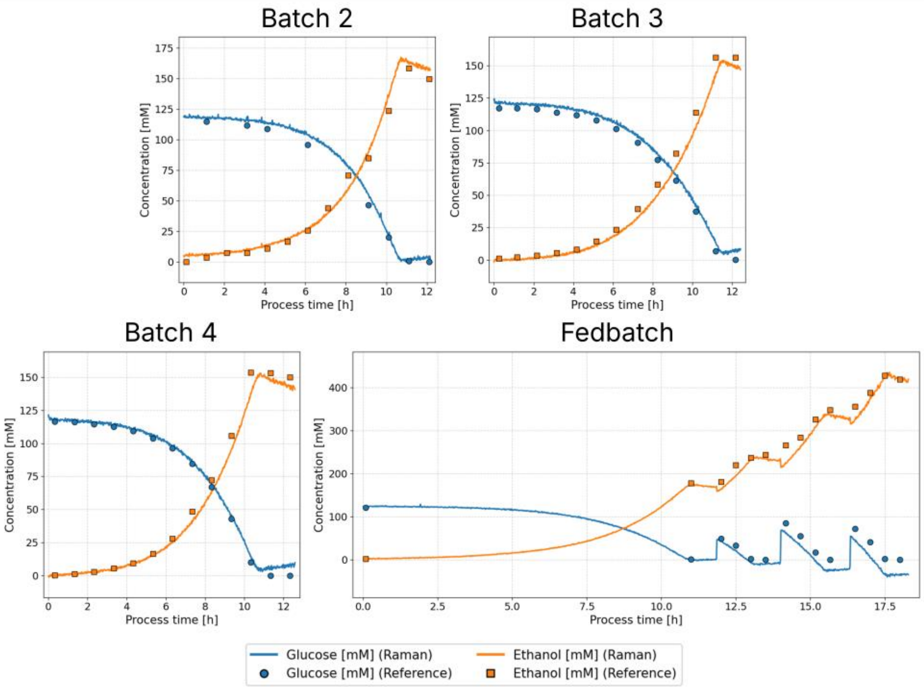


Figure 4.5.5.4: Time-evolution plots showing the quantitative predictions of the Raman models from Approach 4 for all spectra in the validation bioprocess dataset (lines), plotted alongside the corresponding reference measurements for those processes (markers)

Chapter 5

Conclusions & Outlook

5.1 Conclusions

The biotechnology industry is actively looking for Process Analytical Technology (PAT) tools capable of performing real-time measurements during bioreactor processes. In-line Raman spectroscopy has emerged as a powerful solution, offering non-invasive, multiplexed monitoring of key process parameters through a single optical probe. However, extracting a quantitative readout from spectral data depends on chemometric models calibrated with extensive bioprocess datasets, making model development both labor- and resource-intensive. Due to the narrow design spaces, these models become process-specific and are not robust to operational variations which hinders transferability to similar processes. This thesis identified three main challenges for the implementation of Raman spectroscopy as (1) the necessity of collecting process calibration data, (2) the process-specificity of models calibrated with process data, and (3) the reliance on data-driven models. The preceding chapters addressed these challenges by investigating alternative approaches to data collection and model calibration, and by studying the influence of changing measurement conditions on spectral integrity. This final chapter summarizes key findings, and highlights how they contribute to the rapid implementation of Raman spectroscopy for bioreactor processes.

The work in **Chapter 2** demonstrated how PLS models for glucose, ethanol, and biomass calibrated with *Saccharomyces cerevisiae* batch process data leads to process-specific models that predict batch evolution. This often goes unnoticed when models are used to predict a similar process, as standard quantitative assessments do not evaluate target specificity. Qualitative assessment indicated the integration of strong cross-correlations originating from the calibration dataset. This limited the applicability of the models trained on batch data to a related fed-batch mode of operation. A simple dataset supplementation with single compound spectra improved model specificity towards quantification targets and improved performance on the fed-batch process (decrease in rRMSEP of 82.7% for glucose, 90.1% for ethanol, and 69.3% for biomass). This highlights the importance of analyte variation in the calibration dataset for data-driven models, and demonstrates how solely incorporating process data does not result in robust prediction models. Designing an effective calibration dataset requires careful consideration through experimental planning, minimizing experimental workload and accelerating automated monitoring development.

Despite prediction performance improvement for glucose and ethanol in **Chapter 2**, model specificity towards *S. cerevisiae* biomass was limited. In reported literature, yeast biomass spectra are primarily characterized by strong scattering and signal extinction, and cell protein and lipid signals are only observed via Raman microscopy [1-3]. As a result, spectral pre-processing required to correct for scattering may inadvertently remove biomass specific signals. A deeper understanding of the influence of measurement conditions and light scattering particles on Raman spectra can aid in improved pre-processing strategies, distinguishing molecular signals from noise. **Chapter 3** investigated the influence of temperature, bubble quantity, viscosity, and *S.* cerevisiae biomass concentration on the peak position, peak intensity, and baseline shifts in Raman spectra obtained in a bioreactor setting. Despite the strong signal extinction caused by yeast cells as particles (44.6% at 5 g/L), the spectral contribution of cell proteins and lipids could be identified after correcting for signal extinction. The presence of bubbles in the bioreactor caused similar extinction effects (up to 7.9%) as cells, again highlighting the importance of intensity normalization for particle scattering effects. Changing the temperature from 20^{0}C to 40^{0}C caused peak shifts of up to 0.81 cm^{-1}, which could adversely affect model prediction performance. These findings are especially relevant to the developments in high-throughput Raman spectroscopy systems, where spectra acquired at different measurement conditions (e.g. immersion probe, flow-cells, or sample chambers) are combined. When it is not possible to maintain consistent measurement conditions through control steps, more extensive pre-processing is required to normalize spectral intensities and align spectral variables.

The detection of *S. cerevisiae* cell proteins and lipids in pure biomass measurements performed in **Chapter 3** allowed for the construction of a targeted calibration dataset acquired through simple measurements. **Chapter 4** focused on calibrating quantification models using a small dataset of 16 single compound spectra of defined media, glucose, ethanol, and biomass, aiming to calibrate models without process data. Models were calibrated through both PLS and IHM approaches, allowing for the comparison between a data-driven and model-driven quantification strategy. The accuracy and analyte specificity of these models demonstrated the effectiveness of utilizing process knowledge for the targeted collection of calibration data. The compound hard models allowed for the extraction of peak features and the generation of de novo synthetic spectra, resulting in simulation of custom process states. Models calibrated with synthetic spectra led to high analyte specificity (improved alignment of regression coefficients with known peaks), and showcased the potential for synthetic spectra generation when extensive spectral knowledge is available. Augmenting process spectra with the same peak features through the synthetic adjustment of target analytes demonstrated how spiking studies could be simulated without the need for complex experimental setups. These approaches indicate the potential of computational spectra generation and augmentation for quantification problems, leading to model calibration independent of process spectra.

Overall, this thesis describes several approaches by which simple spectral measurements can be applied to either improve, or fully calibrate Raman spectroscopy quantification models for yeast fermentation. Additionally, the improved understanding of how spectra are affected by bioreactor measurement conditions and yeast biomass enables targeted and effective pre-processing strategies. The approaches described in this thesis could lead to the development of quantification models using small, targeted datasets to generate models before running the actual process. Such modelling approaches are more flexible to process variations due to increased analyte specificity, and the calibration datasets based on single compound measurements can be easily expanded towards new analytes. This is particularly valuable in research environments where process parameters and conditions change frequently, making full model re-calibration using process data-driven models impractical. Furthermore, augmenting existing spectra of complex processes supports quantification model improvement, even when limited process data is available.

5.2 Outlook

The work in this thesis offers methods and guidelines for rapid development of Raman spectroscopy quantification models, nevertheless, more research is required to extend their applicability. The following three sections describe outlooks on the topics of leveraging rapid model development, Raman spectroscopy miniaturization, and the democratization of Raman spectroscopy in the field of bioprocessing. These sections are based on research directions that were explored during this thesis and showcase preliminary results.

5.2.1 Leveraging rapid model development

Chapter 4 demonstrates the potential of model calibration through single compound measurements, specifically by using the IHM approach. Despite the effectiveness of the IHM method for processes with few molecular species and high process knowledge, few studies have reported on this approach in literature [4-7]. For this approach to be effective, a high level of process knowledge and access to single compound spectra is required to generate compound hard models. In contrast to biopharmaceutical bioprocesses, where many studies have reported on the use of

Raman spectroscopy, a wide range of bioprocesses exist in industrial microbiology and biocatalysis for which extensive molecular knowledge of the system is readily available. The combination of extensive process knowledge and a limited number of process compounds could enable the use of rapid and flexible model calibration strategies, thereby reducing the experimental effort required for implementing Raman spectroscopy. To this end, an exploratory study was performed to investigate approaches for rapid Raman spectroscopy model development in biocatalysis process monitoring, an enzymatic reaction with low mixture complexity was analyzed. In a collaborative project with Prof. Frank Hollman and MSc Luc Zuhse the enzymatic oxyfunctionalization of ethylbenzene was monitored by in-line Raman spectroscopy (Figure 5.1A). This reaction takes place in 50% acetone in KPI buffer, and the substrate ethylbenzene is converted into 1-phenylethanol, which can be overoxidized into acetophenone. Despite not knowing the concentration trends of this reaction before running the process, the compounds present in the reaction mixture are fully defined providing an interesting proof-of-concept study for testing of rapid model development strategies, applying both PLS and IHM methods. The first reaction was operated using an ethylbenzene starting concentration of 100 mM, and samples were taken every 30 minutes while continuously acquiring Raman spectra with an acquisition time of 30 seconds. This resulted in a dataset with 18 reference concentrations of the reactants, following the progress of the reaction in Figure 5.1B.

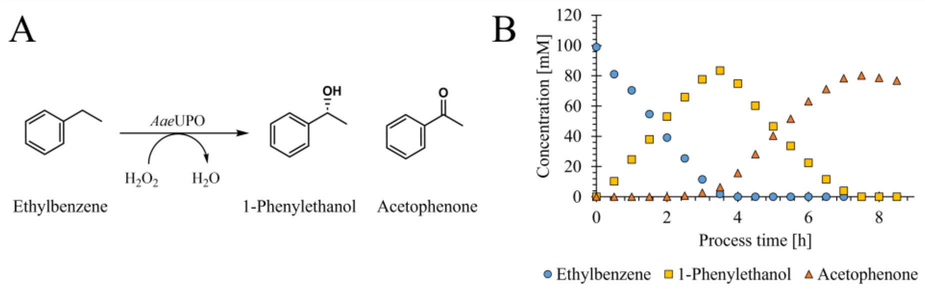


Figure 5.1: The reaction schematic (A) and the reaction progress through gas chromatography reference samples (B) of the oxyfunctionalization of Ethylbenzene by the *Aae*UPO enzyme. The process was initiated at an Ethylbenzene concentration of 100 mM.

As shown in Figure 5.1B, the concentration trends of the individual compounds lead to a decoupling of cross-correlations between the reactants. To study the applicability of Raman spectroscopy monitoring for this enzymatic reaction, the first run was used to calibrate three PLS quantification models, namely for ethylbenzene, 1-phenylethanol, and acetophenone. These three models were subsequently applied to predict concentrations during a second enzymatic reaction operated at identical settings. The resulting predictions are shown in Figure 5.2.

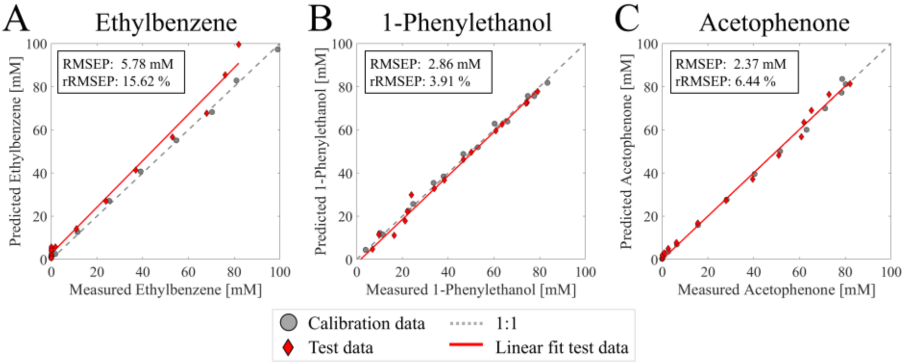


Figure 5.2: Measured versus predicted plots of PLS models for the prediction of ethylbenzene (A), 1-phenylethanol (B), and acetophenone (C). The models were calibrated on 17 spectra and reference measurements of one enzymatic reaction, and subsequently used to predict concentrations from a second enzymatic reaction operated at identical settings. The spectra were pre-processed by selecting the region between 700-1800 cm^{-1}, Automatic Whittaker Filter baseline correction (lambda = 10000, alpha = 0.001), and mean centering.

Model evaluation showed a relative Root Mean Square Errors of Prediction (rRMSEP, based on interquartile range) of 16.62%, 3.91%, and 6.44%, for ethylbenzene, 1-phenylethanol, and acetophenone, respectively. With the exception of two high concentration ethylbenzene samples, the calibration of PLS models with data from one operation resulted in accurate models specific to their analytical target (strong alignment of regression coefficients with known peaks). The use of a free AaeUPO enzyme solution in 50% acetone provided a particle free reaction mixture and minimal scattering effects were observed. The simplicity of the mixture, low water content, and sharp peaks from the reaction compounds resulted in spectra experiencing minimal spectral noise. A simple baseline correction step (Automatic Whittaker Filter, lambda = 10000, alpha = 0.001) resulted in clean spectra for the subsequent modelling and prediction steps. These experiments demonstrate the applicability of Raman spectroscopy for monitoring reactant concentrations during biocatalytic processes. Although only a single process run yielded accurate quantification models for the reactants, process data was still required for calibration. As a next step in the development of this enzymatic reaction process, the ethylbenzene concentration was increased to 1 M. Despite the accuracy of the PLS models when applied to the same 100 mM ethylbenzene process range, the prediction of reactants during the 1 M run was not feasible as the PLS models cannot extrapolate this far beyond the calibration range. To achieve similar model performance, the 1 M substrate reaction would have to be operated at least once to collect calibration data from the process itself, hindering direct monitoring of the new process.

In an attempt to develop monitoring models for the 1 M ethylbenzene reaction without the use of process data, custom mixtures of the reactants were analyzed using a miniaturized setup in 50 mL bottles. Raman spectra were acquired from single compound spectra as well as from 11 custom mixtures with concentration expected from the 1 M run, but following the concentration trends observed in the 100 mM ethylbenzene reaction. The single compound measurements were used to generate hard models for ethylbenzene, 1-phenylethanol, acetophenone, acetone, and KPI buffer by fitting Pseudo-Voigt functions. These hard models were combined in a mixture model that was calibrated on the 11 mixture spectra. This yielded a prediction model based on physical knowledge on the process compounds which adjusts the weight of each compounds spectra to minimize residuals in new process spectra. The 1 M ethylbenzene was monitored with Raman spectroscopy, and the mixture model was applied to the process spectra (Figure 5.3).

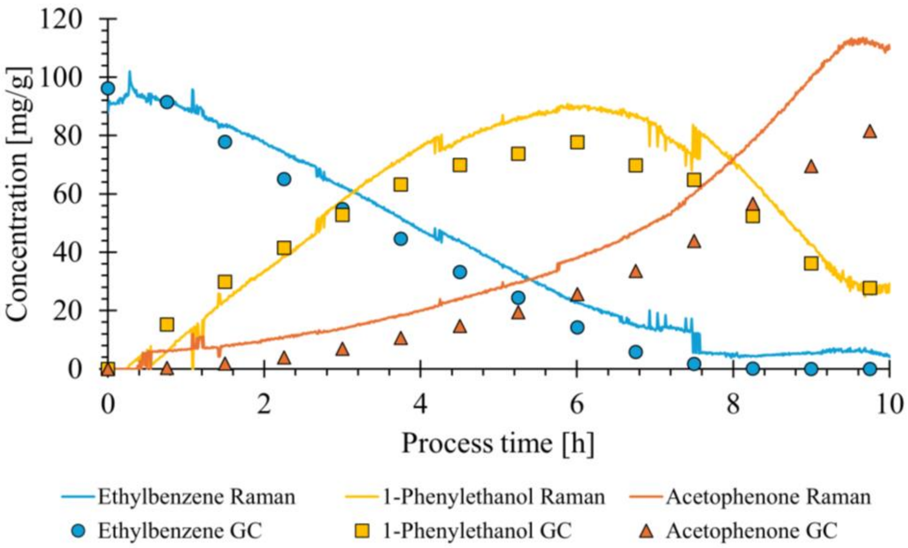


Figure 5.3: Comparison of predicted concentrations for ethylbenzene, 1-phenylethanol, and acetophenone using the Raman spectroscope in combination with the Indirect Hard Model (lines) versus reference measurements by gas chromatography (markers). The concentrations are expressed in mg/g due to the ratiometric approach used during IHM modelling. The oxyfunctionalization of ethylbenzene by the AaeUPO enzyme was initiated at an ethylbenzene concentration of 1M.

The IHM mixture model achieved rRMSEP values of 11.67%, 26.52%, and 42.70% versus the 16.62%, 3.91%, and 6.44% of the PLS models for ethylbenzene, 1-phenylethanol, and acetophenone, respectively. Although the IHM mixture model did not achieve the same level of prediction accuracy as the PLS models, the concentration trends were effectively captured, making it applicable to early-stage process monitoring. This showcases how single compound measurements accompanied by simple mixtures allow for the generation of a physically informed model that bases predictions based on spectral knowledge. Using flexible modeling methods based on existing pure compound models is highly applicable for dynamic research environments in which process parameters change rapidly.

This approach allows for an expendable compound library from which mixture models can be generated for new processes. The calibration of these mixture models solely serves to establish correlations between peak weights and compound concentrations. This minimizes calibration effort compared to data-driven methods like PLS, as the spectral characteristics of each compound are pre-established and do not need to be extracted for each new calibration. Additionally, the effort required to measure several single compound or synthetic mixture spectra is minimal (30 minutes) compared to performing the entire process (10 hours), and is significantly less resource intensive. Once a database of compounds is generated the spectral knowledge can be transferred across processes and departments, provided that spectroscope equipment and measurement conditions are standardized. This modular approach to model calibration allows for flexibility towards a wide range of bioprocesses and promotes collaboration across biotechnology disciplines. Leveraging Raman spectroscopy across different bioprocess types enables the compounding of knowledge, streamlining its implementation into new applications.

5.2.2 Raman miniaturization for accelerated implementation

Since the 1980s, advancements in Raman spectroscopy equipment have improved measurement stability and quality. The development of immersion probes connected to long optical fibers enabled in-line measurements within bioreactor processes while the spectrometer itself can be at distance from the process [8]. In addition, developments in sample chambers, flow-cells, and microscope-based setups allowed for the application of Raman spectroscopy through on-line and at-line measurement modes, allowing flexibility across experimental setups and scales. Despite the flexibility in application, combining spectral data from these different measurement modes is challenging due to inherent physical and optical differences between equipment types. Furthermore, the work in **Chapter 3** highlighted how changes in measurement conditions can affect the integrity of Raman spectra, stressing the importance of consistency across experiments. While in-line Raman spectroscopy using in-line immersion probes is widely applied for process monitoring, the ability to miniaturize measurement setups while preserving optical conditions offers promising opportunities for rapid model development.

A notable example of this flexibility is the integration of at-line Raman spectroscopes with automated mini-bioreactor systems. When combined with liquid handling stations and at-line reference analytics, this setup allows automated sampling and Raman spectral acquisition across a wide range of bioreactor conditions. In addition, the liquid handling station allows for the modification of bioreactor samples by spiking with the compound of interest (as simulated in **Chapter 4**) to enhance model specificity for this target [9]. Thus, this results in acquisition of calibration data covering a wide process design space without the need for full-scale bioreactor runs. Despite the availability of these advanced equipment setups, studies reporting on the application are limited. This may be related to the high equipment costs, reliance on single-use vessels, and the need for skilled technical personnel for operation and maintenance of the setup. This severely limits the applicability of these systems for small companies and research groups.

An alternative approach to miniaturizing the Raman measurement setup, while maintaining optical consistency, is the use of immersion probes in flow-cell assemblies [10, 11]. This enables one to maintain the optical configuration and minimize sample volume. The reduced sample volume is especially beneficial for the analysis of novel material and early-stage processes where available sample volumes are limited. During this thesis a flow-cell assembly was designed to fit the bIO-Optic immersion probe normally used in-line in bioreactor systems during fermentation. The design requirements for this component were to have a low internal volume (preferably <1mL), temperature control, and in-place cleaning in-between sample measurements (Figure 5.4).

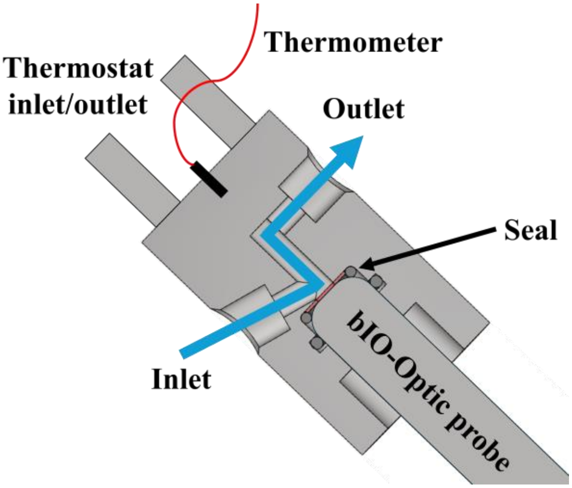


Figure 5.4: A schematic representation of the custom flow-cell for the bIO-Optic Raman immersion probe. The assembly is mounted at an upward 45^0 angle by which sample can be loaded from the bottom which fully submerges the probe window and pushes out air towards the outlet on the top. A water bath can be connected to the thermostat inlet and outlet by which the temperature of the system can be controlled, measured by a thermometer fitted on the end of the flow-cell.

The final design of the flow-cell features an internal measurement chamber of 15mm deep holding a volume of 200µL, meeting the requirement of sample volumes under 1mL (including the volume required for the inlet and outlet ports). With this design, samples could be loaded with a 1mL syringe. To prevent air bubbles in the sample chamber and on the probe window from interfering with the measurements, the flow-cell is mounted at a 45^0 upward angle. In this orientation samples are loaded from the bottom, thereby submerging the probe window and pushing out bubbles from the sample chamber. To establish temperature control a loop was built in around the sample chamber to circulate water from an external water bath. A thermometer was mounted on the side of the flow-cell to monitor the system temperature. In summary, this flow-cell design enables the measurement of low volume samples (<1mL) with the bIO-Optic probe in a temperature-controlled setup.

In the PhD project of Brenda Juarez hematopoietic stem cells were produced in 250 mL mini-bioreactors (Getinge) growing on APEL serum-free media. The size of the headplate does not permit the use of a 12mm bIO-Optic immersion probe, and the process was therefore explored with Raman spectroscopy by collecting weekly spent medium samples right before medium refreshment with the flow-cell assembly. Stored samples were thawed and measured in duplicates through a randomized order, with the flow-cell maintained at 37^0C to mimic the conditions of the bioreactor process. Since reference data for these samples was limited to glucose and lactate concentrations, the study focused exclusively on examining the changes in these spectra over the course of the bioreactor run. As the APEL media is a complex mixture with many proteins and growth factors at low concentrations, the major spectral changes were expected to originate from changes in glucose and lactate concentrations. Principal Component Analysis (PCA) was used to analyze spectra from the bioreactor runs and identify the most significant spectral changes over time (Figure 5.5).

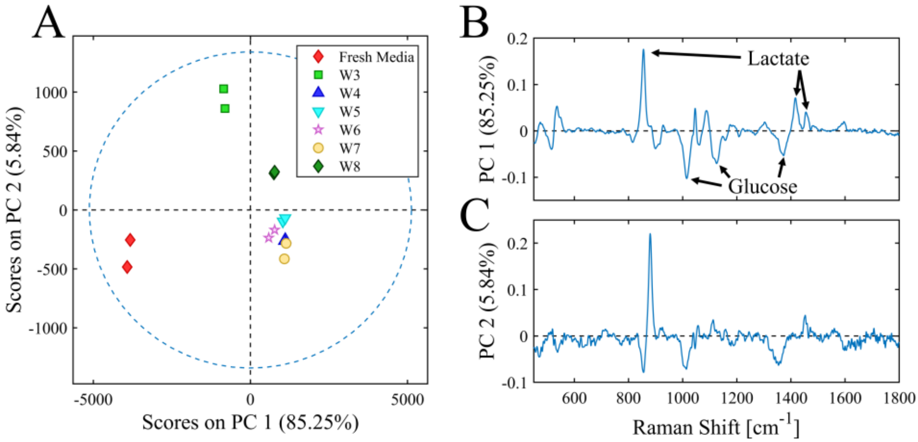


Figure 5.5: The sample scores of the media samples on Principal Component (PC) 1 and PC2 (A), loadings of PC1 (B), and the loadings of PC2 (C). The spectra were pre-processes with Automatic Whittaker baseline correction (lambda = 10000, alpha = 0.001) followed by mean centering. The spectral features associated with glucose and lactate are indicated on PC1, while the spectral feature likely associated with glutamine is indicated on PC2.

The majority of spectral variation (91.09%) could be explained with two Principal Components (PCs) (Figure 5.5A). Most sample duplicates group close together, and variations between the duplicates mostly occur along PC2. The variation along PC1 is mostly related to changes in the glucose and lactate concentration of the media, highlighted by positive loadings for lactate features and negative loadings for glucose features (Figure 5.5B). This was to be expected, as glucose is consumed and lactate is produced by the developing cell culture, causing fresh media samples to score low on PC1. The samples from week 3 and 8 vary from the other samples along PC2, and both these samples experienced lowered cell viability (data not shown). The peak at 880 cm^{-1} on PC2 could not be directly linked to a compound. The samples from week 3 show the greatest deviation from other samples along PC2, which may be explained by the accumulation of a compound due to the media not being refreshed up until this timepoint. If this compound is associated with reduced cell viability it could also account for the elevated scores on PC2 for the week 8 samples.

The use of miniaturized setups as shown above enable the efficient exploration of processes using existing sample banks and offer valuable insights into the molecular composition. This approach supplements the traditional reference measurements with additional process understanding, and helps researchers to anticipate monitoring challenges. Even with limited reference data, miniaturized systems allow for testing Raman spectroscopy during early-stage process development without the need of directly using an immersion probe. These exploratory measurements help to identify the key spectral contributors of the system and can highlight potential challenges such as fluorescence interference. In summary, these miniaturized setups are a low-risk starting point for integrating Raman spectroscopy during bioprocess development, and allow rapid screening of process conditions.

5.2.3 Democratization of Raman spectroscopy

Despite the technological developments in commercial Raman spectroscopes and the increasing utilization as PAT tool for process monitoring and control, its adoption is not widespread. The initial investment to obtain Raman spectroscopy equipment is steep and it is currently not a "plug-and-play" technology. The majority of industrial case studies reporting Raman spectroscopy implementation into production processes are performed by organizations with extensive R&D teams and funding [12]. However, even in these environments, the dependency on data-driven modelling approaches limits transferability [13, 14]. Overall, extensive calibration processes, lack of standardization and transferability of calibrated models, and most importantly, the lack of knowledge on efficiently implementing Raman spectroscopy hinders its widespread adoption.

To greatly reduce experimental efforts required to implement Raman spectroscopy into a new process, and to improve the transferability of existing models to related processes without requiring full re-calibration, we should democratize Raman spectroscopy models. This can be done with open-source standardized or generic prediction models. While several generic models for the prediction of metabolites and cell density during cell cultures have been reported in literature, these models are not publicly available [14, 15]. Concerns over confidentiality make companies reluctant to share process spectra and data, posing significant challenges for open-source modelling efforts through sharing process data.

When considering typical cell factory platforms in the biotechnology industry (e.g. *Escherichia coli*, *S. cerevisiae*, Chinese Hamster Ovary cells) the fundamental components required for cell growth are largely standardized per platform. While media compositions can vary to accommodate different nutrient requirements or different substrates for specific products, generating a general overview of the medium should be feasible. Moreover, compound concentrations close to the detection limit of the Raman spectroscope might not be essential to consider when the main goal is to monitor major compound concentrations. As Raman spectroscopy measures the molecular composition of a mixture, cell factory platforms can be represented as a list of major process compounds accompanied by their probable concentration ranges. Once this list of compounds is defined, single compound spectra can be acquired. Approaching model calibration with extended process and spectral knowledge allows for the calibration of physics-based models such as the IHM approach developed by Alsmeyer, Koß, and Marquardt in 2004 [4]. The fundamental difference between this technique and data-driven approaches is that spectral variation does not have to be modeled repeatedly for each new process. These new model types can incorporate prior knowledge from individual compound models and their probable concentration ranges, enabling them to perform more informed predictions.

While data collection presents a complex challenge for bioprocessing organizations to address independently, open-source modelling efforts offer a potential solution as the direct sharing of process data is not required. By building public libraries of single compound models and performing studies on general media compositions per cell factory platform, a knowledge base for chemometric modelling can be developed. While this data is not yet publicly available, instrumentation manufacturers have started to work towards plug-and-play Raman spectroscopy solutions [16]. By distributing Raman spectroscopes with built-in prediction models targeted at specific process ranges, the uncertainty in terms of implementation investments are greatly reduced. The spectroscope can be implemented immediately for monitoring basic process parameters, while process-specific data can be simultaneously collected for the development of models for additional compounds of interest. Yet, the commercialization of standardized models does not support small research organizations or academia, as that is not open source.

References

1. Iversen, J.A., R.W. Berg, and B.K. Ahring, *Quantitative monitoring of yeast fermentation using Raman spectroscopy.* Analytical and bioanalytical chemistry, 2014. **406**(20): p. 4911-4919.
2. Yang, N., et al., *In-line monitoring of Bioreactor by Raman Spectroscopy: direct use of a standard--based model through cell--scattering correction.* Journal of Biotechnology, 2024.
3. Wang, K., et al., *Species identification and strain discrimination of fermentation yeasts Saccharomyces cerevisiae and Saccharomyces uvarum using Raman spectroscopy and convolutional neural networks.* Applied and Environmental Microbiology, 2023. **89**(12): p. e01673-23.
4. Alsmeyer, F., H.-J. Koß, and W. Marquardt, *Indirect spectral hard modeling for the analysis of reactive and interacting mixtures.* Applied spectroscopy, 2004. **58**(8): p. 975-985.
5. Müller, D.H., et al., *Bioprocess in‐line monitoring using Raman spectroscopy and Indirect Hard Modeling (IHM): A simple calibration yields a robust model.* Biotechnology and Bioengineering, 2023.
6. Echtermeyer, A., et al., *Inline Raman spectroscopy and indirect hard modeling for concentration monitoring of dissociated acid species.* Applied spectroscopy, 2021. **75**(5): p. 506-519.
7. Kriesten, E., et al., *Identification of unknown pure component spectra by indirect hard modeling.* Chemometrics and Intelligent Laboratory Systems, 2008. **93**(2): p. 108-119.
8. Esmonde-White, K.A., et al., *Raman spectroscopy as a process analytical technology for pharmaceutical manufacturing and bioprocessing.* Analytical and bioanalytical chemistry, 2017. **409**(3): p. 637-649.
9. Sibley, M., et al., *Novel integrated Raman spectroscopy Technology for Minibioreactors.* BioProcess Int, 2020. **18**: p. 9.
10. Romann, P., et al., *Advancing Raman model calibration for perfusion bioprocesses using spiked harvest libraries.* Biotechnology Journal, 2022: p. 2200184.
11. Romann, P., et al., *Raman‐controlled pyruvate feeding to control metabolic activity and product quality in continuous biomanufacturing.* Biotechnology Journal, 2024. **19**(1): p. 2300318.
12. Esmonde-White, K.A., M. Cuellar, and I.R. Lewis, *The role of Raman spectroscopy in biopharmaceuticals from development to manufacturing.* Analytical and Bioanalytical Chemistry, 2021: p. 1-23.
13. Yousefi-Darani, A., et al., *Generic Chemometric Models for Metabolite Concentration Prediction Based on Raman Spectra.* Sensors, 2022. **22**(15): p. 5581.
14. Webster, T.A., et al., *Development of generic raman models for a GS‐KOTM CHO platform process.* Biotechnology Progress, 2018. **34**(3): p. 730-737.
15. Tulsyan, A., et al., *Automatic real‐time calibration, assessment, and maintenance of generic Raman models for online monitoring of cell culture processes.* Biotechnology and Bioengineering, 2020. **117**(2): p. 406-416.
16. 908 Devices. *MAVERICK: Process analytical technology for in-line bioprocess analysis.* n.d.; Available from: https://908devices.com/products/maverick/.

Acknowledgements

The contents of this thesis represent only a small part of the four-year journey of doing a PhD. Doing research is not possible without the help of the people around you. This holds for everyone at TU Delft, but also largely for many others in my life. I would like to personally thank all of you.

First of all, I would like to thank **Marieke** and **Marcel** for their excellent supervision over the years.

Marieke, I could not have wished for a better supervisor. When I started, being 50% of your group meant that I had the luxury of your almost undivided attention. It was great to observe how actively you worked on becoming the best supervisor possible, attending countless courses on supervision and management. I enjoyed giving feedback for those courses, though there was rarely anything I could add. While we barely had any disagreements, the combination of your very structured personality and my somewhat more chaotic approach presented a few challenges, but I feel we both learned how to make it work. I've seen your two person team grow into a full research group, and I am proud of what you have built over the four years I was there. You have had to face some challenges already, but I can truly say that you are doing great and your group appreciates how hard you work for them. I will miss our fun discussions about our creative projects at home, the warm MK-MK and MK group meetings, and the constant confusion over why my own initials kept appearing everywhere on Outlook and Teams. I am sure our paths will continue to cross, and I would love to collaborate again in the future.

Marcel, Thank you for your guidance and strong advice over these past four years. Our discussions provided the occasional but much-needed reality check and helped me refocus on the bigger picture and main goals of completing the PhD. Whenever a scientific discussion ended with a confident "goed, goed" I knew I was on the right track.

To **Astrid**, **Mehrab**, and **Lisa**, my BEP & MEP students.

Astrid, I was lucky to have such an independent and determined personality as my first student. Despite the many setbacks with the HPLC setup, you stayed persistent and deliver an excellent thesis. Your piano sessions in the atrium were a real highlight, it was the best music I ever heard in the building. **Mehrab**, your meticulous and precise lab work is the foundation of Chapter 3 of this thesis, and your lab journal is a work of art, it was a breeze to work with. You tackled your challenges with writing well, and I am proud to see how much you improved during your project. I liked our fun discussions about science and future career steps during the many breaks and our impromptu pipetting sessions. I am happy to see you enjoying your new position and I am sure that we will keep on running into each other. **Lisa**, working with you was a real pleasure. Your clear communication, efficient meetings, and structured approach led to valuable results. We only spent a total of 2 days in the lab together, yet combined with existing datasets, that was enough for a full MEP through your impressive use of spectral synthesis. Your work was the exploratory study to what eventually developed into Chapter 4 of this thesis. It is great to see you doing well in your traineeship at J&J.

To **Tim** and **Meryl,** my lovely paranymphs.

Tim, my PhD buddy from the first hour. Thank you for the first few years of dealing with my endless questions on Python and data structures, most of which I could have easily googled myself (I still never used classes in Python, sorry). The extra chair in my office was unofficially reserved for you, and I loved all those afternoons when you'd appear out of nowhere to survive the final hour of the day. I will never forget (and probably keep meeting) storytelling Tim, who appears after a few beers and speaks an octave higher, fiercely sharing opinions on the smallest of topics. Your passion for coffee is unmatched, and I have learned to limit my caffeine intake at your place to avoid intense shakiness. One day, for some reason, you decided to start running, and you just never stopped. I truly envy what your shins can endure, and will keep trying to reach even a fraction of your level of fitness. As for karting, I still have not managed to beat you, but I'm determined to keep trying, and maybe I will dethrone the karting king of BPE in our next career phase. I loved our adventure in the USA, hiking through Yosemite park, ending up in the wrong alleyways in San Francisco, dodging hurricanes in Santa Barbara, eating nutritious microwave chicken, gambling away our change in Las Vegas, and hugging trees in Sequoia park. What will our next adventure be?

Meryl, my hard-working senior-PhD office neighbor suddenly got replaced by a whirlwind of energy: high pitched yelling, random bouts of violence against my left shoulder, and endless "what are you doiiing" questions, and I loved every minute of it. A party in a concrete forest filled with glowing paprika's was where it instantly clicked, both as colleagues and beyond. My office soon transformed, new pictures on the walls, a growing hall of fame, Christmas lights arrived (and never left), and of course, many discussions about things like B&B vol liefde. Thank you for giving me the culture shock that is Vastenavend, with its unique dress code that apparently involves leather jackets and old curtains. I also finally had someone to endlessly ramble to about Raman spectroscopy in the office, hopefully it wasn't too much and actually helped you kick-start your own PhD project. I had a very gezellie time with you sitting next to me, and even though the last few months of finishing my thesis were a bit spannie, you made sure I stayed productive, at least in between the coffee breaks. Let's keep having great talks about life over long island ice teas or esma's.

To my warm and welcoming original C0.260 crew. **Tiago**, I want to sincerely thank you for all the honest advice, motivation, and support throughout the years we worked together, you knew when to taunt me, give a reality check, or just a good laugh. You had the privilege of observing me from the side, my neanderthal brow in full view, and I am sorry if I might have looked angry. It was a joy to look to my left, see you peek over the monitors, stretch, drum on your keyboard, or working hard on your second monitor. I loved the time we settled our differences with a bout of BJJ, if we are ever the same weight class we should try again. **Mariana**, kudos for dealing with my barrage of questions on what to do during the first months of my PhD. Your answers short, sharp, and direct, "Maarten stop… just do this instead". It was an honour to sit next to Mariana the queen of BPE, and the one time I accidentally called you Marina was frightening to say the least. If I ever need help getting tickets on Ticketmaster you are the first one I will call. **Zulhaj**, thank you for the great food, good conversations, and many walks around the building. If everyone had your energy level we could all afford standing desks by just giving up on office chairs.

To the "old" generation of PhDs. **Marijn** & **Oriol**, ever since I started at BPE you guys have been inseparable, and it has been greet to keep seeing you guys as a couple after the PhD. **Marijn**, it took me a few months to find out that the mysterious Olger I heard about was you. You are someone people can count on, and I remember the time you stepped in at the last minute to help me move in Delft. **Oriol**, your beautiful drawings and post-it art decorated our offices, your portrait of the queen stayed on our white board for two years. I like yoghurt too, but I've never seen anyone take it as seriously as you do. **Marina**, for someone half my size you somehow have twice my energy. From organizing activities, leading lively discussions, and making sure everyone in the group is doing well. I will never forget how warmly you welcomed me in the group and how much your advice about the PhD and life has meant to me. **Lars**, the true driving force behind many BPE events and Friday (or any other day of the week) drinks. Your discussions on politics and the royal family brought life to all the coffee and lunch breaks. Only you could come up with the idea of turning an old refrigerator into an automated sausage dryer, thanks again for the spare parts you could not take to Australia. I liked our debates on industry 4.0 and digitalization, did the digital twin thesis student ever work out in Brisbane? **Joan**, your art and video editing skills are impeccable, and that same attention to detail shone through in the protocols you created.

To my new and the future C0.260 crew. **Marika**, for some reason you just kept coming back to BPE? It must be a really fun group then. You joined BPE three times, and off-course the third time was best, forming the "Dutch office" and starting the Hall, later turned Wall of Fame (and sometimes shame) with Meryl. I will miss the slamming door and high pitched "Maarteeeeeen" whenever you came back from the lab. Keep doing things your own way, and don't let anyone throw you off balance. **Dimitri**, joining an office full of screaming and heavy sarcasm must have been quite the shock. I admire your dedication, balancing family life with a faraway PhD and the demanding lab work is no small feat. **Lorin**, I hope you like the way we decorated your corner desk, never forget that you are working in the best office of BPE. **Gianmarco**, I am sure you will have a great time in C0.260. I do wish you strength with the Dutch cuisine, hopefully your nonna won't have to send too many care packages from Italy.

To all members of the far away corner office. Every break was more fun when you were around **Roxana**, either because of your infectious good mood or the tasty bread you spoiled us with. I really enjoyed our trip to Yosemite park, thanks for making sure I did not become snake or bear food. **Daphne**, your sharp humour and perfect dose of sarcasm made work a lot more fun. I will never forget the how you and Roxana opened the dance floor in San Francisco, and made sure everyone joined in. **Tim Nijssen**, your quick wit and sharp remarks were impossible to match. **Ramon**, thank you for the many lively (and occasionally tipsy) discussions we've had, from your spontaneous rambling about science to my questionable experiments with airfryers (credits to Blokker quality, it still works). I loved our meetup in Turin, your speed on a Lime scooter through Italian traffic still baffles me. Our conversations about Ramon spectroscopy, CFD, industry adoption, and ideas for startups will be an inspiration for the rest of my career. Please contact me if you ever want to do some Raman modelling work. **Ben**, it's been great working with you during my final year, and after seeing your enthusiastic approach to your project, I might just give your post-it method a try in my next positions. I really enjoyed the party in Amsterdam, so when are we going to the Kurk? I am also slightly jealous of your private bioreactor collection, I will soon drop by for a proper Kombucha starter. **Jelle**, I appreciated your sharp sense of humour and thoughtful questions during my presentations. Your initiatives for activities and lekkere broodjes on Fridays are admirable and show how much you care about group spirit. **Rob**, thank you for your valuable advice, your sharp questions during the coffee breaks probably saved me from several mistakes.

To the lovely next-door office. Where there are bubbles, there's **Rik**. Apologies for all the random surfactant questions! You are a truly multidisciplinary scientist, spanning from space biology to X-ray tomography, and it was great having a fellow DnB fan in the department. **Hector**, Lars his spirit of organizing drinks definitely lives on in you, you are a one man machine when it comes to organizing BBQs complete with carbonated drinks and proper tequila. But please, refrain from cycling too fast after a long BBQ session. And I have to admit, I envy your confidence on the dance floor. **Tamara**, not to be overly dramatic, but you are a scientific force to be reckoned with. I am in awe of your productivity and efficiency, but despite your incredible work ethic I often ran into you at the coffee machine or printer, where you always took the time to ask how I was doing. I greatly appreciate the nice conversations we had about our projects and future plans. **Pieter**, I was initially surprised by your decision to leave industry for a PhD, but you are clearly flourishing in your new role. You found a perfect balance between productivity, making work fun, and having a party every now and then. You are a man with a plan and know exactly what you want. **Ivo**, it was great running into you again in Delft during the EngD, and I am sure that you will like the PhD just as much. Let's keep hitting Basic Fit together ,but please let's go slow on the Bulgarian split squats.

To the lovely people from the next office over. **Eduardo**, **Mona**, you have the rare ability to turn the craziest or most outlandish topics into genuine and thoughtful discussions. **Mun**, my apologies for the culture shock you experienced during those first lunch conversations in your first week. I have great respect for your decision to develop a career in a field you truly believe in, and one you envision will make the world a better place. Please keep that sense of purpose, it inspires others to do the same. **Moumita**, thank you for always walking around with a smile on your face and for the cheerful chats in the hallways and during the coffee breaks. I appreciated your honesty and the genuine questions about my wellbeing from time to time, it meant a lot. **Nicole**, a true bolt of energy, enthusiastic, and clearly taking the right number of holidays (as everyone should). I've enjoyed your challenging questions on Raman, you rarely settle for simple answers and always want to know more. I might think sailing to England is quite an adventure, but for you this is a regular weekend. **Joana**, your work ethic in the lab is impressive, and it's great to see BPE proudly maintaining its Portuguese spirit.

To the wonderful group in the upstairs office. **Brenda**, you were the most motivated candidate I've ever encountered during PhD interviews. You had it all planned out, from getting a PhD to starting your own company and becoming a young successful CEO. I think you're well on your way. I really enjoyed having you in the group, from the delicious Mexican food and candy to occasionally hearing "ai nooooo Marteeeen" whenever I made a bad joke. **Mariana the 2nd**, you always greet everyone with a smile and a kind word, and our chats during breaks were a welcome moment of calm in the day. It is always nice catching up with you (thanks for showing me the FTIR!) and I hope we'll keep running into each other to talk about spectroscopy and the industry, you are doing fantastic! I would like to thank and remember **Miki** for her kindness, enthusiasm, and the genuine care she showed for me and everyone around her. I am grateful for the years we worked together, and the conversations we had about our work but especially about life will remain close to me.

To the great PI's of BPE. **Adrie**, you are a steady factor during the coffee breaks and it is good to see how you make new students feel welcome in the group. **Ludo**, the lunch discussions with you were always highly entertaining, talking about anything from sports to French politics. Thank you for sharing your extensive cognac collection with me during the BPE BBQ at your place. **Cees**, it's been great to see the CASE group grow over time, and I'm impressed with how engaged you are with your students, both at and outside of work. I wish you all the happiness that comes with this new chapter of being a father. **Josh**, it was awesome to see how quickly you got your new labs up and running. Your enthusiasm, honesty, and work-hard-play-hard mindset are a great example to others. **Michel**, your enthusiasm during your lectures in Wageningen was contagious, and I am happy that you brought the same energy to motivate people at BPE in Delft.

To the unwaveringly supportive and helpful staff of BPE. **Christiaan**, my first fermentation experiments were so much better because we performed them together. Sampling until midnight becomes way more enjoyable with breaks spent talking about games and eating burgers. Although your tasks varied wildly over the years, you were always ready to help. I am happy that we still see each other outside of work for nice dinners and bowling competitions. We might have to add karting to the agenda to finally kick Tim of his throne. And **Carina**, Lisanne still wants to beat you at a few of the new games we just got! **Song**, it is always great to see you, and I greatly appreciated your encouragements and motivation during my project. Beyond all your help in the lab you made days at work more enjoyable with your positive energy and endless kindness. And not to forget, your tech knowledge and advice have saved me from more than a few unwise purchases. **Stef**, thank you for being such a great support to all the PhDs and for making the coffee breaks so much fun. Your determination to help me with my project, whether it meant digging through piles of old equipment or tracking down long forgotten parts, was deeply appreciated. Thank you **Max** for making my first year of teaching so enjoyable, and for not taking everything too seriously. **Jeroen**, keep motivating everyone to do mad science! And thank you for getting me out of the office whenever I tried to skip coffee breaks. **Simon**, it has been great getting to know you over the last year. Will you keep teaching us the ins and outs of German culture on the BPE Oktoberfests? **Kawieta**, you are the true MVP of BPE, the engine that keeps the entire department running. A secretary, travel agent, and problem-solver all in one. Whether I needed help with Finance, travel arrangements for conferences, or an HR issue, you would be in my office within minutes on our phone to get it done. Thank you so much for making the administrative side of my PhD so efficient. And special thanks to **Marcel Langeveld** for the creative solutions you came up with in the workshop. In a world where many rely on costly equipment, your ability to build what is needed is deserves far more recognition within BT.

Jort, thank you for the fun time I had during my MSc thesis on your project. It was the perfect stepping stone toward this PhD and most likely the rest of my career. I know I can always reach out to you for career advice, honest opinions about our field, or simply for lively discussions about every other topic imaginable. It is nice to keep seeing you at most of the events I go to, and I'm sure we will keep in touch.

Berry, **Anton**, en **Tim**, bedankt dat jullie me hebben verwelkomd op de Kwikka, jullie half studenten half professionele huishouden. Bij jullie wonen was de perfecte balans van hard werken, relaxen, en in het weekend compleet brak op de bank liggen. De heerlijke maaltijden, talloze flessen wijn, eindeloze goede series en de potjes call of duty en guitar hero waren de perfecte manier om weer op te laden.

Longbois United, your club is terrified. We kunnen niet dieper zinken dan ons onderzeese avontuur in 2017, en sinds toen zijn we alleen nog bergopwaarts gegaan.

Ik ben ontzettend blij dat we elkaar bijna tien jaar nadat we voor het eerst samen in een boot stapten nog zo vaak zien. Dus Dr. Bos en Klitschko, de club groeit, wat gaat ons businessplan worden? **Joris**, grote heit in het noorden, ik had nooit gedacht dat iemand vrijwillig een PhD zou beginnen in Leeuwarden, maar na mijn recente bezoek lijkt Wetsus me een geweldige plek om te werken. Ik kan geen genoeg krijgen van je dagelijkse content en ik kijk uit naar je defense! **Peter**, je zit nog een heel stuk noordelijker als Joris, hoe is het leven daar als Dr.? Je bent een geweldige gastheer en een fantastische kok, en ik weet zeker dat Finland er op vooruit gaat met jou erbij. Ik kan niet wachten om Turku te bezoeken en naar Turku carnaval te luisteren. **Arthur**, hoe je het voor elkaar krijgt om meerdere bedrijven te runnen naast een fulltime baan, en tóch geen feestje of activiteit met ons mist blijft mij een bewonderenswaardig raadsel. Ik kijk vol enthousiasme naar wat je allemaal doet en ben blij om na zo lang nog zulke goede vrienden te zijn. Ik kom graag weer eens door de uiterwaarden wandelen. **Daan**, ik kom altijd met veel plezier langs bij jou en Vera, en waardeer de openheid van onze gesprekken enorm. We zijn allebei wat zoekend naar wat we willen in de komende jaren, en ik vind het fijn om samen te praten over de grote keuzes die daar bij komen kijken. En mocht je over een paar jaar op zoek zijn naar een nieuw doel, dan kunnen we altijd nog een ploegbedrijf beginnen. **Jeroen**, hoewel ik wat ver weg zit kom ik nog steeds ontzettend graag bij je bankhangen met een bak koffie. Misschien het allerleukst vond ik toen we samen op de kop in de auto hingen om die radio aan te sluiten, zullen we binnenkort een gezamenlijk project beginnen? **Reinder**, ongekend hoeveel jij nog sport, maar ik klaag niet want dit houd soepkom in goede staat. Ik heb stiekem geoefend met Fifa, wanneer doen we een rematch? Dan kom ik graag een keer logeren in je nieuwe stekkie. **Jens**, menig bouwbedrijf mag even bij jou op les, en met jouw tempo van grond verleggen kunnen ze bij Rijkswaterstaat nog wat van je opsteken. Wanneer is de housewarming van je landgoed? **Kim**, je weet me altijd aan het denken te zetten, soms net iets verder dan ik zelf normaal doe. Ik waardeer de diepgaande gesprekken, ookal heb ik soms even tijd nodig om in mijn hoofd bij te benen. **Amarens**, de commando's "BENEN" en "laatste 10 halen" echoën nog door mijn hoofd, en komen soms zelfs van pas. Soms moet je jezelf een beetje voorliegen om de eindstreep te halen toch? Toch gaat het vaak makkelijker met wat externe sturing, en sta jij voor ons klaar als dit zooitje ongeregeld de verkeerde kant op drijft.

Martin, de eerste dag op de middelbare school dacht ik eerlijk gezegd dat je zo'n typische scoutingnerd was, maar dat beeld veranderde snel toen we onze gezamenlijke obsessie voor Lego Technic ontdekten (ik was ook een nerd). Vanaf dat moment raakten we onafscheidelijk: in de klas, na school, bij de Plus of KFC, op feestjes (best handig zo'n scoutinggebouw) en op vakanties. En alsof dat nog niet genoeg was kwam je me ook nog achterna naar Wageningen om ook Biotechnologie te studeren, en werden we zelfs een tijdje huisgenoten. Na ook dezelfde master en zelfs een stage bij hetzelfde bedrijf ben je nu toch een andere richting in geslagen, en dat gaat je heel goed af. Je bent een van mijn beste vrienden, en ik hoop dat we elkaar in de komende zeventien jaar net zo vaak zien als in de afgelopen zeventien.

Aan **Cun**, **Max**, **Henkjan**, **Thom**, en **Luuk**, ook al zien we elkaar soms lange tijd niet, het is er niet minder gezellig om. Met kort bijpraten zitten we snel weer op onze oude lijn en is het even gezellig als 10 jaar geleden.

Aan mijn ouders, zus en broer en schoonfamilie. Nou **Maike**, het is me gelukt! Na al die jaren zwoegend naar de mentorgesprekken op de middelbare school hadden we dit misschien niet meer durven dromen. We wisten allebei dondersgoed dat ik het prima kon, maar ik koos zelden de makkelijkste manier. Toch heb ik tijdens het studeren mijn interesse gevonden en ging alles toen een stuk soepeler. Bedankt voor oneindige steun, motivatie, grote knuffels en lieve woorden, zonder deze dingen was dit niet gelukt. **Harm**, bedankt voor de altijd kritische vragen en voor het steeds weer doorvragen waarom ik iets deed en waarom het belangrijk was. Jouw technische achtergrond is duidelijk een beetje op mij overgeslagen, misschien via je oude lego, en ik heb daar een praktische instelling aan overgehouden. Ook bij toekomstige banen en uitdagingen blijf ik graag sparren over wat belangrijk is en hoe dingen opgelost kunnen worden. **Elke**, soms denk ik dat onze interesses ver uit elkaar liggen, totdat er plots een oude wasmachine voor onze neus staat. Binnen een uur moest dat ding volledig uit elkaar puur uit interesse hoe zon ding werkt en om onderdelen voor **Floris** te scoren. Ik heb enorm veel bewondering voor je creativiteit en alle kunstwerken die je maakt. **Thomas**, klein broertje, ik vind het geweldig om te zien hoe erg je opleeft in Wageningen en het plezier wat je haalt uit je onderzoeksprojecten. Ik ben heel benieuwd of je ook een PhD gaat doen, ik denk in ieder geval dat het bij je past. Familie **Wisman**, bedankt voor de leuke en warme weekenden, gezellige vakanties, en het vele eten wat ik altijd van jullie krijg.

Dan als laatste mijn dank aan het thuisfront. **Ollie** & **Mees**, bedankt voor het tien keer gezelliger maken van het thuis werken de afgelopen 2 jaar, ook al ging dat gepaard met de nodige spelfouten doordat jullie het liefst over mijn handen liggen. Jullie waren duidelijk geïnteresseerd in mijn werk, vooral het gedeelte over lasers. **Ollie**, jouw oneindige honger in de ochtend was de perfecte wekker wanneer wij waren vergeten er een te zetten. **Mees**, laten we vooral geluiden naar elkaar blijven maken, ook al begrijpen we geen bal van elkaar, het is in ieder geval heel gezellig.

En dan als laatste, **Lisanne**. Hey smurf, zonder jou was dit nooit gelukt. Ergens anno 2017 zagen we elkaar in een boot, en vanaf daar ging alles in een stroomversnelling. Naast dat ik je onmiddellijk een geweldig leuk en mooi persoon vond, was je ook precies de schop onder mijn kont die ik nodig had om eens iets harder te gaan studeren. Je bent me sinds toen altijd onvoorwaardelijk blijven steunen, en motiveerde mij om zoveel mogelijk met twee handen aan te pakken in plaats van eindeloos blijven treuzelen. Waar ik vaak bleef hangen in gedachten was jij degene die altijd meteen in beweging kwam. Toen ik deze PhD begon ben je me vrolijk achterna gekomen om ook in Delft te komen wonen. Bij momenten van frustratie of stress was een knuffel en wat bemoedigende woorden al genoeg, en stond je meteen klaar om me te helpen het op te lossen. Ik ben ontzettend trots op hoe ver we samen zijn gekomen, en ik ben benieuwd naar wat ons volgende avontuur gaat brengen.

Curriculum Vitae

Maarten Klaverdijk was born on August 28th, 1996, in Arnhem, The Netherlands. After graduating high school he started his BSc in Biotechnology at Wageningen University & Research, attracted by the study's combination of biology, chemistry, and engineering. Motivated by an example of biopharmaceutical production in plants during a minor in plant biotechnology, he continued with the Master's programme Biotechnology specializing in medical biotechnology and biopharmaceutical production.

For his MSc thesis he joined the Bioprocess Engineering (BPE) department, where he explored process monitoring of the baculovirus expression system for vaccine production using on-line holographic microscopy. This project sparked his interest in bioprocess monitoring, which was further strengthened during an internship at Byondis where he worked on optimizing a CHO cell perfusion process, and experienced firsthand the challenges of process monitoring and sampling in biopharmaceutical manufacturing. Presenting a poster on his MSc thesis at the online COVID-19 edition of the Netherlands Biotechnology Congress (NBC) gave him the opportunity to share his research, but also led to his first contact with Marieke Klijn.

After obtaining his MSc degree, Maarten moved to Delft to start his PhD in the group of Marieke Klijn, within the Bioprocess Engineering department at Delft University of Technology. His research initially focused on real-time bioreactor monitoring using a combination of in-line microscopy and Raman spectroscopy. Following some challenges with the microscopy setup in the first year, his work during the following three years focused on streamlining the implementation of Raman spectroscopy for bioreactor monitoring, with an emphasis on novel strategies for data collection and chemometric model calibration.

The work in this thesis reflects his interest in combining biotechnology, data analysis, and process monitoring.

List of Publications

Journal articles

- **Klaverdijk, M.**, Smulders, L. A., Ottens, M., & Klijn, M. E. (2025). Towards rapid calibration of bioprocess quantification models using single compound Raman spectra: A comparison of four approaches. *Biotechnology and Bioengineering*. Advance online publication. https://doi.org/10.1002/bit.70092

- **Klaverdijk, M.**, Nemati, M., Ottens, M., & Klijn, M. E. (2025). Impact of bioreactor process parameters and yeast biomass on Raman spectra. *Biotechnology Progress*, Article e70050. https://doi.org/10.1002/btpr.70050

- **Klaverdijk, M.**, Ottens, M., & Klijn, M. E. (2025). Single compound data supplementation to enhance transferability of fermentation specific Raman spectroscopy models. *Analytical and Bioanalytical Chemistry*, 1-12.

- Altenburg, J. J., **Klaverdijk, M.**, Cabosart, D., Desmecht, L., Brunekreeft-Terlouw, S. S., Both, J., ... & Martens, D. E. (2023). Real-time online monitoring of insect cell proliferation and baculovirus infection using digital differential holographic microscopy and machine learning. *Biotechnology Progress*, *39*(2).

Conference contributions

- **Klaverdijk**, **M**, Ottens, M., Klijn, M.E., Setting Up Raman Spectroscopy for Effective Upstream Bioprocess Monitoring, Nederlandse Biotechnologie Vereniging (NBV) Biopharma Event 2024, Leiden, November 2024, Oral Presentation

- **Klaverdijk**, **M**, Ottens, M., Klijn, M.E., Raman Spectroscopy Models for Upstream Process Development: Improving Transferability Across Processes, American Chemical Society (ACS) Fall meeting, San Francisco, United States, August 2023, Oral Presentation

- **Klaverdijk**, **M**, Ottens, M., Klijn, M.E., Combination of Raman Spectroscopy and In-line Microscopy Monitoring for Yeast Fermentation, Netherlands Biotechnology Conference (NBC) 2022**,** Leiden, The Netherlands, October 2022, Oral Presentation

- **Klaverdijk**, **M**, Ottens, M., Klijn, M.E., Combination of Raman Spectroscopy and In-line Microscopy Monitoring for Yeast Fermentation, 13th ESBES Symposium, Aachen, Germany, September 2022, Oral Presentation