
THE UNIVERSITY
*of* EDINBURGH

# The Use of On-line Metabolomics for Bioprocess Monitoring

*Joan Cortada García*


*Doctor of Philosophy*

THE UNIVERSITY OF EDINBURGH

2022

***Few will have the greatness to bend history; but each of us can work to change a small portion of the events, and in the total of all these acts will be written the history of this generation.***

*Senator Robert F. Kennedy's address to students at Day of Affirmation ceremonies at the University of Capetown, Capetown, South Africa, 6th June 1966.*

# Abstract

Biotechnology has the potential to play a key role in the transition from an economy based on non-renewable fossil fuels to a greener economy in which renewable feedstocks can be used to generate a wide range of products. Fermentation monitoring in real-time can provide detailed biological and engineering information about the bioprocess, which is a very valuable asset for process development and is an instrumental tool for product manufacturing, allowing to quickly detect and react to deviations from desired process specifications, potentially preventing bioprocess failure and the loss of capital and resources.

While many different parameters can be monitored in a bioprocess, the monitoring of metabolites is particularly useful, for these provide the best representation of the phenotypic state of a living organism at any specific time. The aim of this thesis is to explore the use of metabolomics to monitor the metabolites present in the liquid phase of a bioreactor in real-time. Metabolomics offers the possibility to simultaneously detect hundreds of compounds without the need to develop complex and laborious chemometric models. To study this technology, a fluidics system was developed coupling a bench-top bioreactor to a mass spectrometer, allowing automatic on-line metabolomics analysis every five minutes.

The on-line metabolomics analysis system was implemented with an *Escherichia coli* (*E. coli*) fermentation process of succinate production and 886 different ions were monitored in an untargeted fashion using an Exactive Orbitrap mass spectrometer. This bioprocess was also analysed by liquid chromatography - mass spectrometry (LC-MS) in order to better characterise the signals monitored by untargeted on-line metabolomics, find process biomarkers and identify potential strain engineering targets. Being able to monitor hundreds of metabolites with a high time resolution can be a very powerful tool for process development, allowing the comparison of different strains and cell lines, as well as identifying by-products and bottlenecks without requiring extensive analytical calibration.

Based on the on-line metabolomics results and the LC-MS analysis, metabolites of interest for the bioprocess were identified and this information was used to develop a targeted on-line metabolomics method. Forty-one metabolites were monitored on-line in a targeted fashion for three *E. coli* succinate fermentation replicates and univariate linear regression models were built to correlate the mass spectrometry signal to the concentration of 12 metabolites and the biomass in the bioreactor, demonstrating how this technology can be used to monitor the concentration of multiple bioprocess metabolites of interest on-line. These regression models were built by splitting the triplicate fermentation data into a training set and a validation

set, and model performance was statistically assessed. Importantly, regression models for the 41 metabolites monitored could have been developed with further analytical calibration. Therefore, metabolomics allows the monitoring of a significantly larger amount of compounds compared to alternative technologies, such as vibrational spectroscopy.

Finally, a kinetic model was built to describe the aerobic batch phase of the succinate bioprocess, allowing improved process understanding. Furthermore, the model was employed in combination with targeted on-line metabolomics to forecast the future evolution of key process compounds at different stages of the process, showing how on-line metabolomics can be used as a forecasting tool during product manufacturing.

In summary, this thesis demonstrates the use of on-line metabolomics for bioprocess improvement in two main ways. Firstly, on-line metabolomics can be used to monitor hundreds of metabolites at a very high time resolution without requiring prior calibration or process knowledge (untargeted method). This is a very powerful tool for bioprocess development, allowing the comparison of different strains, cell lines and process conditions at a much faster throughput than alternative technologies, particularly when LC-MS is used to complement the on-line metabolomics data. Secondly, on-line metabolomics can be used to monitor the concentration of specific metabolites (targeted method), which has great potential in improving the performance of product manufacturing in the biotechnology industry, for example by using predictive kinetic models in real-time.

# Lay Summary

Biological systems such as bacteria, yeast, filamentous fungi, microalgae and mammalian cells can be used to produce medicines, food, enzymes, fuel and materials, among other products for human benefit. The great advantage of these so-called bioprocesses is that they can be performed using renewable feedstocks as the main substrates - such as sugars and starch from plant materials - rather than using non-renewable fossil fuels. This makes biotechnology instrumental in the transition from an economy based on non-renewable fossil fuels to a greener economy.

Nevertheless, the successful development of industrial bioprocesses depends on them being economically competitive with the equivalent petrochemical processes, which is not trivial. For this reason, it is important to develop better tools and technologies that can help optimise and increase the success rate of bioprocesses, as well as ensuring that the end product quality meets the required market standards. One of these technologies is mass spectrometry, which allows the simultaneous measurement of hundreds of compounds, even at very low concentrations. When the compounds analysed are small molecules - such as glucose, citric acid, ethanol or penicillin - this global analysis is called metabolomics.

Measuring how these compounds are produced and consumed during the bioprocess can provide very useful information regarding the health of the biological system used (*e.g.* the bacterial cells), product formation, substrate consumption or the production of unwanted by-products and process inhibitors that can compromise the bioprocess. Knowing this information in real-time is key for the proper operation and optimisation of bioprocesses.

In this thesis, the use of real-time metabolomics was studied for the analysis and improvement of a bioprocess of succinate production using the bacterium *Escherichia coli* (*E. coli*), a common production organism in biotechnology. Succinate is a compound that can be used as an intermediate in the manufacturing of many different products such as personal care items, pharmaceutical intermediates, food and drink additives, polymers, resins, coatings, lubricants and textiles. The use of real-time metabolomics with this bioprocess allowed the quantitative and semi-quantitative analysis of a wide range of key process compounds (a larger number of compounds than in alternative techniques, such as infrared and Raman spectroscopy) and forecasting the evolution of some of these compounds at various points of the process using mathematical models. This technology also allowed the identification of potential ways to improve the *E. coli* bacterium with genetic engineering in order to increase succinate production. All this shows how real-time metabolomics has great potential for bioprocess analysis, both during the manufacturing of biotechnology products and for bioprocess development.

# Resumen para público no especializado

Sistemas biológicos como bacterias, levaduras, hongos filamentosos, microalgas o células de mamífero se pueden usar para producir medicinas, alimentos, encimas, combustibles y materiales, entre otros productos de interés humano. La gran ventaja de estos bioprocesos es que se pueden llevar a cabo usando materiales renovables como substratos (como azúcares y polisacáridos vegetales) en vez de usar combustibles fósiles no renovables. Por este motivo, la biotecnología puede desempeñar un papel fundamental en la transición de una economía basada en combustibles fósiles no renovables hacia una economía más sostenible.

Sin embargo, el creciente desarrollo de bioprocesos industriales depende de que éstos sean económicamente competitivos con sus respectivos procesos petroquímicos, lo cual no es baladí. Por eso, es de gran importancia desarrollar herramientas y tecnologías que faciliten la optimización y aumenten el porcentaje de éxito de los bioprocesos, a la vez que permitan garantizar que el producto del bioproceso cumpla con los requisitos de calidad del mercado. Una de estas tecnologías es la espectrometría de masas, la cual permite el análisis simultáneo de cientos de compuestos, incluso a concentraciones muy bajas. Cuando las moléculas analizadas son pequeñas (como la glucosa, el ácido cítrico, el etanol o la penicilina), a este análisis global se le llama metabolómica.

Medir cómo estos compuestos se producen y consumen durante el bioproceso puede aportar gran cantidad de información acerca de la salud del sistema biológico usado (por ejemplo las células bacterianas), la formación de producto, el consumo del substrato o la producción de subproductos e inhibidores que puedan poner en riesgo la ejecución del bioproceso. Conocer esta información en tiempo real es de gran importancia para la correcta operación y para permitir la optimización del bioproceso.

En esta tesis se ha estudiado el uso de la metabolómica en tiempo real para el análisis y la mejora de un bioproceso de producción de succinato usando la bacteria *Escherichia coli* (*E. coli*), uno de los organismos más usados en biotecnología. El succinato es un compuesto que se puede usar como paso intermedio en la fabricación de muchos productos, como productos de belleza, productos farmacéuticos, aditivos alimentarios, polímeros, resinas, materiales de revestimiento, lubricantes y materiales textiles. El uso de la metabolómica en tiempo real con este bioproceso permitió analizar de forma cuantitativa y semicuantitativa una gran cantidad de compuestos clave del bioproceso (un mayor número de compuestos que en técnicas alternativas como la espectroscopía infrarroja o la espectroscopía Raman) y anticipar la evolución de algunos de estos compuestos en distintos puntos del proceso usando modelos matemáticos. Esta tecnología también permitió identificar posibles maneras de mejorar la

bacteria *E. coli* con ingeniería genética para augmentar la producción de succinato. Todo esto demuestra cómo la metabolómica en tiempo real tiene un gran potencial para el análisis de bioprocesos tanto durante la manufacturación de productos como en el desarrollo y mejora del proceso.

# Acknowledgements

I would like to thank my supervisors Karl Burgess, Rónán Daly and Ali Arnold for providing guidance throughout the four years of this PhD in the fields of metabolomics, data analysis and fermentation, and for giving me the freedom to explore the avenues that I found interesting, important and worth researching. I especially extend my gratitude to Karl for his fast yet thorough feedback on my scientific output.

This PhD project has been made possible by the Industrial Biotechnology Innovation Centre (IBioIC) and the UK Research and Innovation Biotechnology and Biological Sciences Research Council (UKRI BBSRC), who provided the funding, training and access to a network of industrial biotechnology partners.

I would like to express my gratitude to Glasgow Polyomics, where I spent the first 18 months of the PhD, particularly to Jeni Haggarty for sharing long days in the lab developing the on-line metabolomics monitoring system hand in hand. I would also like to thank all the other people that made my stay in Glasgow unforgettable, especially Gabri for filling my memories of my time there with jazz tunes.

The move from Glasgow to Edinburgh brought a lot of new colleagues and friends. I would like to thank Karl Burgess' group and EdinOmics for the moments spent inside and outside of university, especially Martina and Haig for all the trips to Craigmillar Castle and Prestonfield House during the difficult times of lockdown due to the COVID-19 pandemic. I would like to extend my thanks to all the friends that I have met in Edinburgh. To Hazel, Yu and Ricardo for all the great times in the tennis court and around the dining table. To Hanna, Robert, Nahuel and Nadine for all the trips in the stunning Scottish countryside. To Alán and Ricardo for being great opponents at chess. And to Martina and Michael for their contagious cheerfulness. I would like to give a special thanks to Marcos for being a great friend from the moment I first landed in Scotland.

In 2020, I spent nine months on industrial placement at Ingenza Ltd. I would like to thank the whole Ingenza team, particularly the fermentation department, for giving me the chance to collaborate on a wide range of projects and learn from exposure to many different fermentation processes. I would like to express my sincere gratitude to Ali Arnold for being a great boss, colleague and friend.

My heartfelt thanks to my friend Pau for being a source of wisdom and inspiration, both professionally in the field of fermentation analysis and personally. Thank you, Marc, for always being so easy to talk to and willing to discuss absolutely any topic. To Joan for your huge altruistic heart. And in general to my friends from Barcelona - Maria, Sergio, Laia, Mar and Miriam - for sharing so many great moments with me.

I would also like to thank my parents for giving me both roots with strong values and wings to fly to where I am today. Thanks also to my brothers for your wisdom and meaningful conversations.

Finally, to Rebekah, for brightening my days and filling them with kindness and happiness.

Joan Cortada García

# Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

**Joan Cortada García**

# Contents

# Figures and Tables

## Figures

---

## Tables

# Nomenclature

| | |
|---|---|
| $\alpha$ | Luedeking-Piret factor proportional to the biomass growth rate [mol P/C-mol X] |
| $\beta$ | Luedeking-Piret factor not proportional to the biomass growth rate [mol P/(C-mol X · h)] |
| $\mu$ | Biomass-specific formation rate of biomass [$h^{-1}$] |
| $\mu_{Ace}$ | Biomass-specific formation rate of biomass due to acetate consumption [$h^{-1}$] |
| $\mu_{Glu}$ | Biomass-specific formation rate of biomass due to glucose consumption [$h^{-1}$] |
| $\mu_{max}$ | Maximum biomass-specific formation rate of biomass [$h^{-1}$] |
| $\hat{y}_i$ | Estimated value by the model for the *i-th* sample |
| $y_i$ | Measured value for the *i-th* sample |
| Ace | Acetate concentration [mol Ace/L] |
| C-mol X | Moles of biomass normalised to one atom of carbon |
| Glu | Glucose concentration [mol Glu/L] |
| k | most meaningful principal components |
| $k_d$ | Biomass decay rate [$h^{-1}$] |
| $K_i$ | Half-saturation constant of substrate *i* consumption [mol i/L] |
| m | number of variables |
| n | number of samples |
| $N_i$ | Amount of moles of compound *i* [mol i] |
| $q_{Ace_{cons}}$ | Biomass-specific acetate consumption rate [mol Ace/(C-mol X · h)] |
| $q_{Ace_{prod}}$ | Biomass-specific acetate production rate [mol Ace/(C-mol X · h)] |
| $q_{i_{max}}$ | Maximum biomass-specific consumption or production rate [mol i/(C-mol X · h)] |
| $q_i$ | Biomass-specific consumption or production rate [mol i/(C-mol X · h)] |
| S | Substrate concentration [mol S/L] |
| X | Biomass concentration [C-mol X/L] |
| $Y_{Ace_{i/s}}$ | Yield of production or consumption of compound *i* on acetate [mol i/mol Ace] |
| $Y_{i/s}$ | Yield of production or consumption of compound *i* on glucose [mol i/mol Glu] |

Across the text of this tests, scalars are written in lowercase italics ($x$), vectors are written in lowercase bold italics ($\boldsymbol{x}$) and matrices in uppercase bold ($\mathbf{X}$).

## List of Acronyms

**AA** acetic acid

**AC** ammonium carbonate

**ACN** acetonitrile

**APCI** atmospheric pressure chemical ionisation

**APPI** atmospheric pressure photo ionisation

**BCAA** branched-chain amino acid

**CHO** chinese hamster ovary

**CID** collision induced dissociation

**C:M:W** chloroform:methanol:water

**DCW** dry cell weight

**DO** dissolved oxygen

**DoE** design of experiments

**EMP** Embden-Meyerhof-Parnas

**ESI** electrospray ionisation

**FA** formic acid

**FDA** Food and Drug Administration

**FIR** far-infrared

**FWHM** full width at half maximum

**GC-MS** gas chromatography-mass spectrometry

**HESI** heated electrospray ionisation

**HILIC** hydrophilic interaction liquid chromatography

**HPLC** high-performance liquid chromatography

**IM-Q-TOF** ion mobility quadrupole time-of-flight

**IPA** isopropanol

**IR** infrared

**LC-MS** liquid chromatography-mass spectrometry

**LOD** limit of detection

**LOESS** locally estimated scatterplot smoothing

**LOOCV** leave-one-out cross-validation

**m/z** mass-to-charge ratio

**MeOH** methanol

**MIR** mid-infrared

**MRM** multiple reaction monitoring

**NIR** near-infrared

**NMR** nuclear magnetic resonance

**$OD_{600}$** optical density at 600 nm

**PALS** pathway activity level scoring

**PBS** phosphate buffer solution

**PC** principal component

**PCA** principal component analysis

**PEEK** polyetheretherketone

**PEP** phosphoenolpyruvate

**PLS** partial least squares

**PLS-DA** partial least squares – discriminant analysis

**ppm** parts per million

**PPP** pentose phosphate pathway

**PiMP** Polyomics integrated Metabolomics Pipeline

**PRESS** predictive residual sum of squares

**QQQ** triple quadrupole

**RMSE** root mean squared error

**RMSECV** root mean squared error of cross-validation

**RMSEP** root mean squared error of prediction

**RMSRE** root mean squared relative error

**RMSRECV** root mean squared relative error of cross-validation

**RMSREP** root mean squared relative error of prediction

**S/N** signal to noise ratio

**SRM** selected reaction monitoring

**SSR** residual sum of squares

**SST** total sum of squares

**TCA** tricarboxylic acid

**TIC** total ion chromatogram

**TOF** time-of-flight

**VIP** variable importance in projection

**WCW** wet cell weight

# PART I

# Introduction and Methods

Chapter 1

# Introduction to fermentation analysis and metabolomics


## Abstract

Industrial biotechnology has the potential to play a key role in the transition from an economy totally dependent on fossil fuels to a more sustainable economic model based on the use of renewable sources to produce energy and goods. This chapter is divided into two parts. The first part of this chapter provides an overview of biotechnology and its industrial applications. The concept of bioreactors, fermenters and fermentation monitoring are introduced. The latter is instrumental for the successful scale-up and performance of industrial biotechnological processes. Different methods of fermentation monitoring are reviewed with a focus on biomass and metabolites in the liquid phase. One of these methods is metabolomics, which will be introduced in the second part of this chapter together with mass spectrometry. The different parts of a mass spectrometer are explained, paying special attention to Orbitrap and triple quadrupole mass analysers, which will be used in the different experiments in the thesis. Different important concepts of metabolomics are introduced, such as the difference between targeted and untargeted metabolomics as well as metabolite annotation. The most common statistical tools of metabolomics data analysis are also explained, namely multivariate analysis and pathway analysis. Finally, the main goals of the thesis are detailed at the end of the chapter.


## 1.1 Industrial biotechnology and fermentation processes

Biotechnology is the exploitation of biological systems for industrial and environmental applications. These biological systems typically include microorganisms - such as bacteria, yeasts, filamentous fungi or microalgae - mammalian cell cultures and plants. Common industrial biotech products include fermented food products and food additives, biomass, pharmaceutical and nutraceutical products such as antibiotics, vaccines, hormones, antibodies, antioxidants and vitamins (Silva *et al.*, 2020), enzymes for a vast range of applications (such as the

production of detergents and the processing of food and materials) (May, 2019), fuels such as ethanol, isobutanol and diesel (Peralta-Yahya *et al.*, 2012), and a wide range of materials such as polymers and fibres (or their monomers) (S. Y. Lee *et al.*, 2019; Suitor *et al.*, 2020; Tamis *et al.*, 2018). Waste-water treatment and bioremediation of pollutants from the environment are also industrial and environmental applications of biotechnology (Rittmann and McCarty, 2020).

Biotechnology has great potential to play a key role in the transition from an industry totally dependent on fossil fuels to a more sustainable economic model based on the use of renewable sources. This is because the biological systems used in the bioprocesses mentioned above have the ability to use renewable feedstocks as substrates - such as sugars and starch from plant materials - and turn them into products and energy.

### 1.1.1 A brief history of industrial biotechnology

The history of biotechnology dates back long before humans knew of the existence of microorganisms. Some of the oldest civilisations in Mesopotamia and China already knew how to produce fermented beverages such as beer and wine by 7000 BC, ancient Egyptians fermented bread as early as 4000 BC and people in China fermented soya by 1500 BC (K. Buchholz and Collins, 2013; Demain *et al.*, 2016; P. E. McGovern *et al.*, 2004). However, it was not until the 17$^{th}$ century, with the creation of the first microscopes, that microorganisms were observed for the first time by Antonie van Leeuwenhoek (Schürrle, 2018). Despite this, their significance for human society remained unknown until the 19$^{th}$ century, when a series of discoveries by scientists such as Theodor Schwann, Charles Cagniard de la Tour and Louis Pasteur led to the understanding that ethanol fermentation was caused by yeast (Barnett, 2003; Pasteur, 1995; Schwann *et al.*, 1847).

Louis Pasteur revolutionised fermentation practices during the second half of the 19$^{th}$ century, emphasising the importance of inoculation and sterility to avoid contaminated batches. Several enzymes were discovered during this period (such as invertase, lipase and emulsin) and the first industrial plants for lactic acid production started to appear (Avery, 1881; K. Buchholz and Collins, 2013; Demain *et al.*, 2016). Towards the end of the 19$^{th}$ century, more knowledge was gained about the specificity of enzymes and the fact that biological reactions are (mostly) performed by enzymes (Buchner, 1897; Kohler, 1971). These discoveries represented a change in paradigm, understanding that biological reactions are regulated by the rules of chemistry, disregarding the coetaneous commonly accepted theory of *vis vitalis*, which postulated that biology was ruled by special forces that cannot be explained by physics or chemistry.

At the beginning of the 20$^{th}$ century, the characterisation of the acetone-butanol-ethanol (ABE) fermentation process using the bacterium *Clostridium acetobutylicum* made it possible to find biotechnological solutions for the large-scale production of synthetic rubber (using butanol as an intermediate) and acetone for British weaponry production for WWI. Glycerol produced by fermentation was also used by Germany to make explosives in WWI (Demain *et al.*, 2016; Schürrle, 2018). On a brighter side of history, other enzymes were commercialised at this time (such as pancreatic trypsin and pectinase) and the first industrial plants of microbially-produced citric acid started to appear (Lombardino, 2000).

The serendipitous discovery in the 1920s by Alexander Fleming that the filamentous fungus *Penicillium rubens* (initially misidentified as *Penicillium rubrum* (Houbraken *et al.*, 2011)) produced penicillin (Fleming, 1929) and the successful large-scale production of the antibiotic in the 1940s after strenuous efforts by multiple academic and industrial partners represented another tipping point in the field. Especially, it demonstrated the potential of strain screening for increasing bioprocess performance and it promoted technical advances in fermentation (such as submerged fermentation of filamentous fungi) and downstream processing. The successful commercialisation of penicillin also motivated in the coming decades the discovery and industrial production of other antibiotics, such as actinomycin, streptomycin, cephalosporin C, tetracycline and erythromycin. During these decades, biotechnology was also used for the production of vitamins B2 and B12, and L-sorbose, an intermediate of the chemical synthesis of vitamin C (Bremus *et al.*, 2006; Revuelta *et al.*, 2017).

The decade of the 1950s saw more revolutionary biological breakthroughs, such as the sequencing of insulin and the discovery of the molecular structure of DNA (Stretton, 2002; Watson and Crick, 1953), which set up the basis for the development of the first molecular cloning tools in the coming decades, especially after the development of the PCR in the 1980s by Kary B. Mullis (Mullis *et al.*, 1986). Large-scale production of glutamic acid in the 1950s using fermentation with the bacterium *Corynebacterium glutamicum* also served as a precedent in the following decades for the industrial production of other amino acids using biotechnology, such as lysine, threonine, phenalanine and tryptophan (Hashimoto, 2017).

DNA cloning allowed the production of recombinant products, with the first successful example being the production and approval for clinical use of human insulin expressed in *Escherichia coli* in the late 1970s / early 1980s (Baeshen *et al.*, 2014). Other recombinant products started appearing on the market in the following years, including, among others, human growth hormone, $\beta$-interferon and human tissue plasminogen activator, which was the first approved recombinant protein produced using mammalian cells (Collen and Lijnen, 2004).

Table 1.1 shows an overview of some of the most relevant events in the history of industrial biotechnology up to the present. Currently, in the 21$^{st}$ century, the world highly depends on fossil fuels for the vast majority of the global production of energy and goods. For example, in 2019, the 84 % of the world energy consumption was obtained from fossil fuels: 33 % oil,

24 % natural gas and 27 % coal (Bp, 2021). However, fossil fuels are a limited non-renewable resource and the emissions generated from their consumption are a dominant contributor to climate change, one of the major threats for human society at present (in 2010, 65 % of the greenhouse gas emissions were attributed to fossil fuel combustion and industrial processes (IPCC, 2014)). The different successful stories of biotechnology, especially since the 20$^{th}$ century, have showcased how it can be used to fully or partially replace chemical processes, which typically use fossil fuels as the main feedstock. However, in order to be competitive with chemical processes, proper monitoring and control tools need to be implemented in industry for the successful scale-up and performance of biotechnological processes, especially for commodity chemicals, which are produced in large quantities at a low price per volume.

In this thesis, the use of metabolomics as a fermentation monitoring tool will be explored. An *E. coli* fermentation process of succinate production will be used as a proof of concept for developing said metabolomics monitoring platform. Succinic acid is an example of a molecule with many industrial applications and which can be produced by fermentation. The next sections will provide some background on succinic acid and fermentation analysis.

**Table 1.1:** Most relevant events in the history of industrial biotechnology.

| Time period | Event | Reference |
|---|---|---|
| 7000 BC – 1500 BC | Ancient civilisations ferment foods and beverages such as beer, wine, bread and soya | P. E. McGovern (2019) and P. E. McGovern *et al.* (2004) |
| 1680 | Antonie van Leeuwenhoek observes microorganisms in liquid samples under the microscope | Schürrle (2018) |
| 1830s-1840s | Early industrial enzymatic processes using amylases (formerly named diastase) to produce dextrin from starch<br>Theodor Schwann and Charles Cagniard de la Tour attribute alcohol fermentation to yeast | Armstrong (1933)<br>Barnett (2003) and Schwann *et al.* (1847) |
| 1850s-1880s | Louis Pasteur consolidates the idea that fermentation is caused by living organisms, develops the pasteurisation process and highlights the importance of inoculation for reliable beer production<br>First lactic acid fermentation plants start to appear<br>Discovery of several enzymes (invertase, lipase, emulsin, *etc.*) | Pasteur (1995)<br>Avery (1881)<br>K. Buchholz and Collins (2013) |
| 1890s | Eduard Buchner performs alcohol fermentation from sucrose with only enzymes (pressed yeast extract) | Buchner (1897) and Kohler (1971) |
| 1900s-1910s | Butanol fermentation with *Clostridium acetobutylicum* for synthetic rubber production<br>Industrial-scale production of acetone (Britain) and glycerol (Germany) via fermentation for WWI weaponry production<br>Leonor Michaelis and Maud Menten bring insight into enzyme kinetics<br>Expansion of the industrial uses and production of enzymes | Schürrle (2018)<br>Demain *et al.* (2016)<br>Michaelis and Menten (2013)<br>Demain *et al.* (2016) |
| 1920s-1940s | Industrial production of citric acid by microbial fermentation by Pfizer<br>Alexander Fleming discovers penicillin, produced by the filamentous fungus *Penicillium rubens*<br>Large-scale production of penicillin marks the beginning of the era of biological antibiotics<br>Microbial oxidation of sorbitol into L-sorbose, an intermediate for vitamin C production using *Gluconobacter oxydans*<br>Industrial production of vitamin B2 and B12 by fermentation | Lombardino (2000)<br>Fleming (1929)<br>Demain *et al.* (2016)<br>Bremus *et al.* (2006)<br>Revuelta *et al.* (2017) |

Continued on next page

**Table 1.1 – continued from previous page**

| Time period | Event | Reference |
|---|---|---|
| 1950s | Frederick Sanger publishes the amino acid sequence of insulin | Stretton (2002) |
| | James Watson and Francis Crick publish their model of the three-dimensional structure of DNA based on X-ray diffraction results from Rosalind Franklin and Maurice Wilkins | Watson and Crick (1953) |
| | Industrial production of L-glutamic acid with *Corynebacterium glutamicum* | Hashimoto (2017) |
| 1960s-1980s | Early development of molecular cloning tools. Recombinant DNA is cloned into a plasmid for the first time. | Cohen *et al.* (1973) |
| | Increased use of bio-ethanol as a fuel, particularly in the US and Brazil | Solomon *et al.* (2007) |
| | Georges Köhler and César Milstein presented hybridoma technology for monoclonal antibody production | Köhler and Milstein (1975) |
| | Recombinant production of human insulin in *E.coli* by Eli Lilly and Genentech, being the first recombinant pharmaceutical approved for clinical use | Baeshen *et al.* (2014) |
| | Kary B. Mullis develops the PCR | Mullis *et al.* (1986) |
| | More recombinant products appear on the market, including human tissue plasminogen activator, the first recombinant therapeutic protein produced by mammalian cells to get market approval | Collen and Lijnen (2004) |
| 1990s-present | Flavr Savr tomatoes are the first genetically modified crop approved in the market | Redenbaugh *et al.* (1994) |
| | First complete bacterial genome (*Haemophilus influenzae*) | Fleischmann *et al.* (1995) |
| | Adalimumab, for treating rheumatoid arthritis, becomes the first fully humanised mAb approved in the market | R. M. Lu *et al.* (2020) |
| | Human genome completed | Lander (2011) |

### 1.1.2 Succinic acid

Succinic acid is a dicarboxylic acid with many industrial applications (Figure 1.1). At the beginning of the 21st century, it was regarded by the US Department of Energy as one of the top twelve building block chemicals that can be produced from renewable sugars (Werpy and Petersen, 2004). The importance of succinic acid in the bioeconomy has also been acknowledged by the European Commission, consolidating the driving force in developing large scale bioprocesses for succinic acid production (E4tech *et al.*, 2015).

Succinic acid is used as an intermediate in the manufacturing of high-value consumer products such as personal care items, pharmaceutical intermediates and food and drink additives, as well as a commodity in the manufacturing of high production-volume products such as polybutylene succinate (PBS), polybutylene succinate adipate (PBSA), resins, coatings, lubricants and polyurethanes and it is also used as plasticiser for polymer modification. Furthermore, succinic acid can also be derivatised into other platform chemicals such as 1,4 butanediol (1,4 BDO), tetrahydrofuran (THF), gamma-butyrolactone (GBL), maleic anhydride and succinimide (E4tech *et al.*, 2015; Matano *et al.*, 2014; Nghiem *et al.*, 2017; Saxena *et al.*, 2017; Thakker *et al.*, 2012; Weastra, 2012), all of which have significant market applications, such as the production of elastic fibres, plastics and polyurethanes.

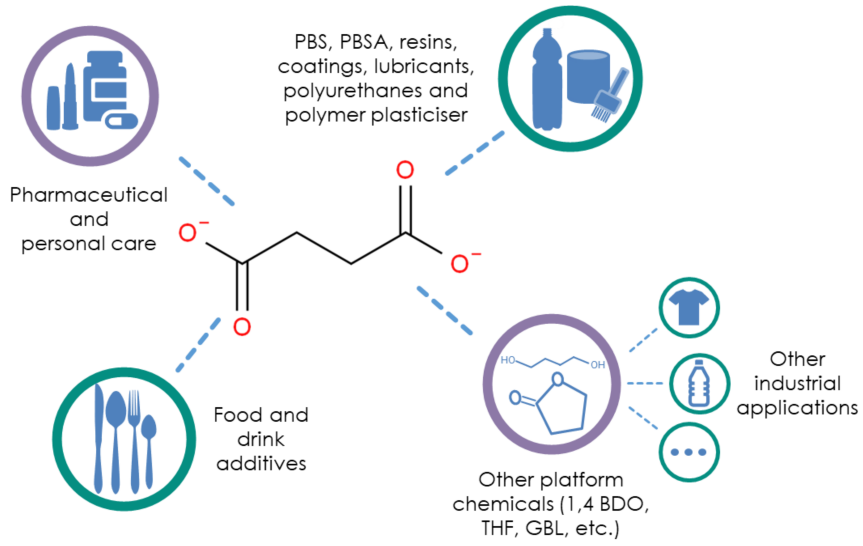



**Figure 1.1:** Industrial applications of succinic acid. 1,4 BDO: 1,4 butanediol; GBL: gamma-butyrolactone; PBS: polybutylene succinate; PBSA: polybutylene succinate adipate; THF: tetrahydrofuran.

In 2015, there were four major suppliers of succinic acid derived from bio-renewable sources based in Europe and North America: BioAmber, Myriant, Reverdia and Succinity, collectively producing 65,000 tonnes/year globally (Jiang *et al.*, 2017; Spekreijse *et al.*, 2019). However, low oil prices since late 2014 have challenged the economic competitiveness of bio-succinate, causing a halt to most of these biobased production plants and a shift to petroleum-based

production facilities in the Asia-Pacific region, especially in Japan and China (MarketsandMarkets™, 2018; McCoy, 2019). Due to its large number of commercial applications and its large market size, improving the yield of biobased succinic acid production could have a substantial impact on its competitiveness with petroleum-derived succinic acid with large economic and environmental implications worldwide.

Succinate is a compound of central metabolism: an intermediate of the tricarboxylic acid (TCA) cycle that can be produced by both oxidative and reductive pathways under aerobic and anaerobic conditions respectively (Figure 1.2). This makes several microbial hosts available for industrial succinate bio-production, with some of the best performing ones being *Actinobacillus succinogenes*, *Anaerobiospirillum succiniciproducens*, *Corynebacterium glutamicum, Escherichia coli, Mannheimia succiniciproducens, Saccharomyces cerevisiae and Yarrowia lipolytica* (S. Y. Lee *et al.*, 2019). High performance values of yield, productivity and titre have been reported for all these, reaching yields close to 0.8 – 0.9 g/g, productivities of 1.75 g/L/h and titres of 50 – 100 g/L (Jiang *et al.*, 2017). This thesis will focus on an *E. coli* succinate process, as detailed below.

### The *E. coli* succinate bioprocess

In this thesis, a dual-phase *E. coli* succinate fermentation process will be studied. The process consists of a batch aerobic growth phase followed by an anaerobic succinate production phase. The industrial strain used in this work was similar to the strain NZN111 reported by Chatterjee *et al.* (2001), with which Y. Liu *et al.* (2011) and H. Wu *et al.* (2007) reported succinate titres of 10 – 20 g/L after 24 h of anaerobic production in an aerobic/anaerobic dual-phase fed-batch processes. Although good performance indicators of productivity, yield and titre have been reported using genetically modified *E. coli* strains, the choice of organism and fermentation method was not strictly dictated by the aim of optimising a specific process, but instead having a relatively simple process to use as a proof of concept for developing a real-time metabolomics monitoring tool.

*E. coli* accumulates and secretes succinate anaerobically via the “anti-clockwise” reductive TCA cycle (Figure 1.2), beginning with the conversion of phosphoenolpyruvate (PEP) into oxaloacetate mainly by the action of the enzyme PEP carboxylase (Gokarn *et al.*, 2000; S. Lu *et al.*, 2009; Saxena *et al.*, 2017). In the absence of oxygen as electron acceptor, oxaloacetate and fumarate behave as electron acceptors, allowing $NAD^+$ regeneration and resulting in succinate production. Succinate excretion in *E. coli* is yet not fully characterised. Fukui *et al.* (2017) recently reported the *yjjP* and *yjjB* genes as candidate succinate transporters in *E. coli* under anaerobic conditions. However, the authors did not provide mechanistic details or energy requirements of the transport system. Other means of succinate secretion under

anaerobic conditions in *E. coli* have been reported, such as the DcuA-C fumarate-succinate antiport system (Six *et al.*, 1994; Zientz *et al.*, 1996) and the CitT citrate-succinate antiporter (Takahashi *et al.*, 2021), where the export of succinate is coupled with the import of fumarate or citrate, respectively.

It is also important to bear in mind that the major mechanism of glucose uptake in bacteria is via the PEP-dependent phosphotransferase system (PTS), which transfers the phosphate group from PEP to glucose, incorporating glucose-6P into the cell and generating pyruvate (Nuoffer *et al.*, 1988; Zhuang *et al.*, 1999). This has two implications for the anaerobic production of succinate: first, that some glucose consumption will need to be dedicated to produce PEP (for further glucose uptake), and secondly, that the anaerobic formation of succinate will be coupled to pyruvate production (as a by-product of glucose-6P formation during glucose uptake).

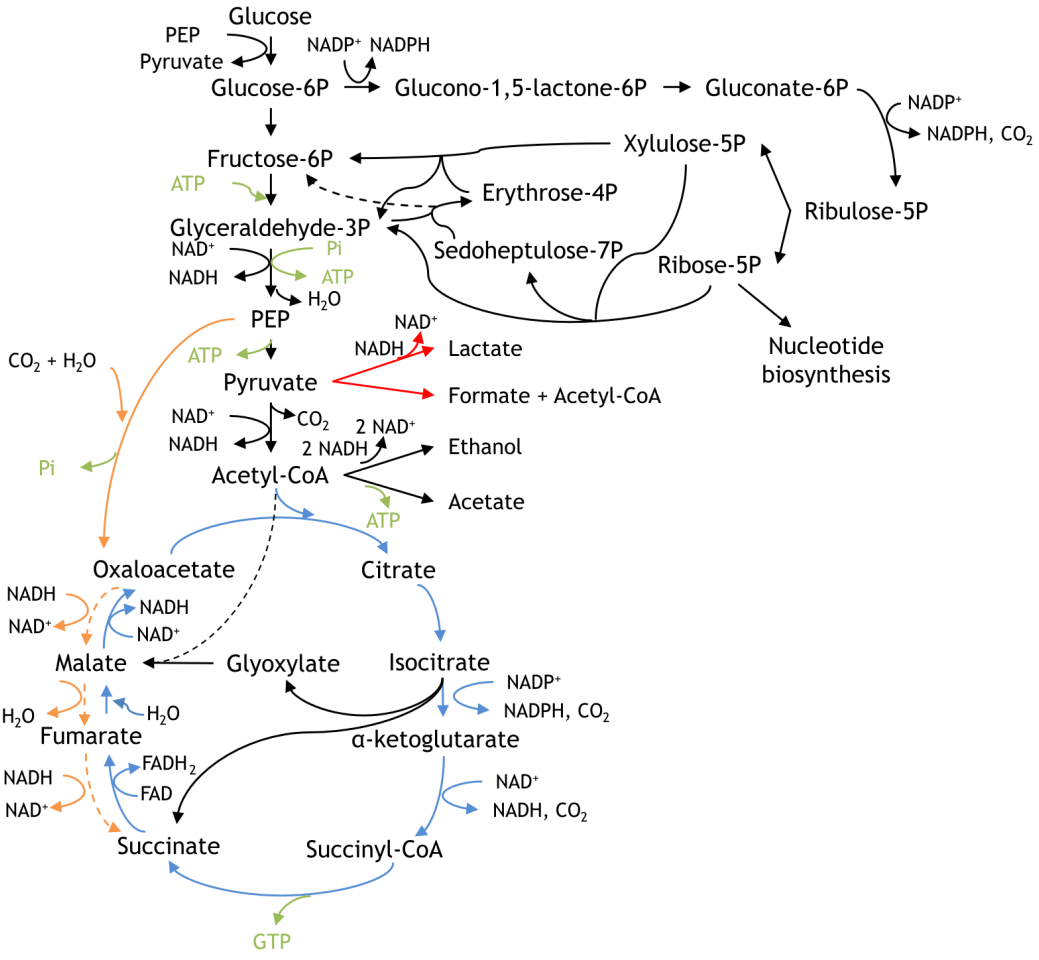



**Figure 1.2:** Main metabolic reactions involved in succinate production in *E. coli*. Blue lines indicate the oxidative TCA cycle under aerobic conditions. Orange lines indicate the reductive TCA cycle under anaerobic conditions. Red lines indicate deleted reactions in the industrial strain used. Arrows crossing other reactions are marked with black dashes. ATP, GTP and free phosphate (Pi) are indicated in green.

### 1.1.3 The bioreactor

Large-scale biotechnological processes are performed in bioreactors (also known as fermenters). These are vessels designed to provide a controlled and aseptic environment for the growth of microbial or mammalian cells and product formation. Aeration and mixing are provided with the use of impellers and/or air bubbles, which ensure that the required nutrients dissolved in the liquid medium reach (ideally) all the cells in the vessel. Cylindrical bioreactors are the most commonly used in industry, with the most popular configurations being the stirred tank reactor, the bubble column and the airlift (Doran, 2013). Stirred tanks provide mechanical agitation to the fermentation broth with impellers, which very effectively disperse the air bubbles sparged in the fermenter but at the same time require more mechanical power and cause more stress to the cells. Bubble columns and airlifts provide aeration and mixing without mechanical agitation. Instead, the same up-ward movement of the air bubbles causes the liquid to mix. Airlifts have additional physical separators inside the reactor, which directs the flow of bubbles and creates very defined up-flow and down-flow mixing patterns for the liquid broth (Figure 1.3).

Bioreactors typically also contain sensors and sample ports to monitor and control that the fermentation is progressing as expected. This is important for making sure that key process targets are met, such as product quality, yield, titre or productivity and that the sterility of the culture has not been compromised. Different methods of fermentation monitoring are detailed below.

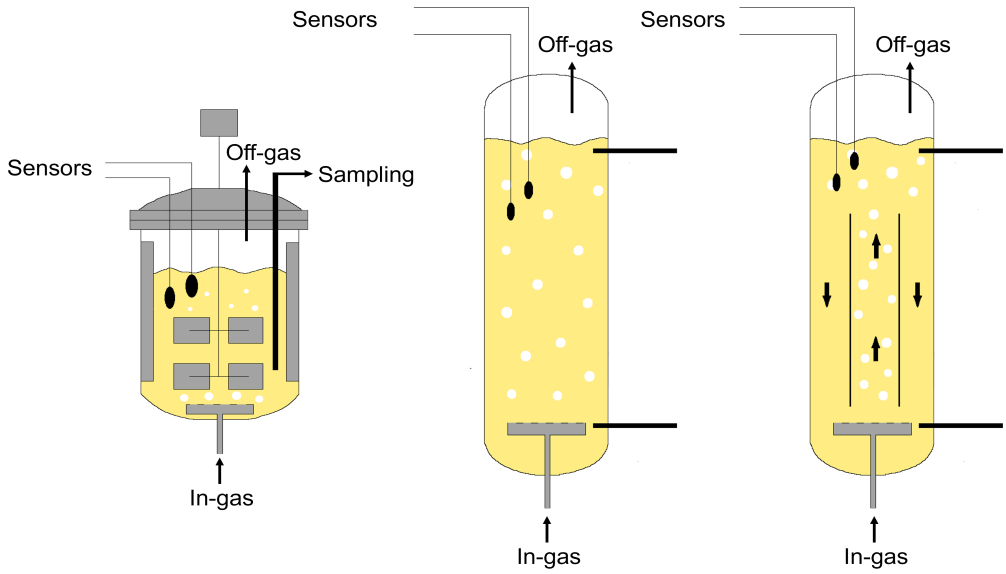



**Figure 1.3:** Most common bioreactor configurations: stirred tank reactor (**left**), bubble column (**middle**) and air lift (**right**). Liquid phase in yellow, gas phase in white.

### 1.1.4 Fermentation monitoring

Fermentation monitoring is a crucial step to understand and control the evolution of a bioprocess. Being able to assess the state of the process in real-time might allow to quickly detect and correct any possible deviations from desired process specifications (Zu *et al.*, 2017). Critical process parameters and critical quality attributes are monitored and controlled in order to guarantee the quality of the end product (Svendsen *et al.*, 2015). There is a wide range of sensors available to monitor these parameters, depending on their nature (Mandenius and Titchener-Hooker, 2013; Vogel and Todaro, 2014). Examples of the sensors most commonly found in fermentation technology are listed:

- Physical: Temperature, pressure, agitation speed, vessel weight, foam level and gas flow rate.
- Chemical: Dissolved $O_2$, dissolved $CO_2$, pH, redox potential and gaseous $O_2$ and $CO_2$ concentration.

In the last two decades the US Food and Drug Administration (FDA) has led a transformation in the way drugs are industrially manufactured. This initiative started with the publication of drug manufacturing guidelines (FDA, 2004) that promote the use of science and engineering to set and follow quality boundaries in the design, development and manufacturing steps of drugs (Clementschitsch and Bayer, 2006; Wechselberger *et al.*, 2010). These guidelines stress the importance of process analytical technologies (especially real-time monitoring), defining three types of measurements for process analysis, depending on their level of automation and integration within the process:

- In-line: the sample is not removed from the process, but instead is analysed *in situ* and the measurement can be invasive or non-invasive.
- On-line: the sample is diverted from the process for analysis and may be returned to the process stream.
- At-line: the sample is removed from the process, isolated and analysed in its proximity.

The three levels above can be understood in opposition to traditional off-line analytical methods, where the sample is removed from the process and analysed separately at a later time, hence not allowing for process control. Automated systems allow faster analysis and quicker response times, facilitating better process monitoring and control. However, automation is not always straightforward, often requiring specialised knowledge and more instrumental investment. For this reason, despite the advantages of real-time analysis, off-line analysis is still very common in industrial environments.

Typical fermentation processes consist of a liquid phase and a gas phase, as it was illustrated in Figure 1.3. The liquid phase contains the cells, medium nutrients and product released by the cells, whereas the gas phase contains the in-gas carrying the gas necessary for the desired metabolic behaviour of the cells, and the off-gas with the remaining in-gas not

consumed by the cells and product gases resulting from the metabolic activity of the cells (such as $CO_2$ or $CH_4$). Aerobic fermentations typically use compressed air as the in-gas, although in some cases where aeration is limiting, a combination of air and concentrated oxygen can be used. Mammalian cell cultures often use a combination of air, oxygen and carbon dioxide in different ratios to control the dissolved oxygen in the broth and its pH (due to the acidity of carbonic acid formed when the $CO_2$ is dissolved in the liquid). This is a milder way to acidify the medium, rather than using strong acids such as $H_2SO_4$ or HCl. Anaerobic processes often do not even have a sparger (Doran, 2013).

The cells are the biocatalyst of any fermentation process. For this reason, most reactions in a bioreactor take place in the liquid phase. Even when a gas compound is one of the major substrates of the process (such as $CH_4$ for methanotrophic bacteria, $CO_2$ for microalgae, or $O_2$ for aerobic organisms), this needs to dissolve in the liquid to become available to the cells. For this reason, a clear distinction is made between the monitoring of the gas phase and liquid phase.

#### Monitoring the gas phase

The most conventional instruments for gas phase monitoring are infrared and paramagnetic analysers for the analysis of $CO_2$ and $O_2$ (Vogel and Todaro, 2014). Nevertheless, measuring other compounds in the gas phase can be very informative, especially if any product of interest is volatile, such as ethanol or terpenoids. In these cases, mass spectrometry (see Section 1.2.1) might be the technology of choice, as illustrated in Table 1.2.

**Table 1.2:** Examples of mass spectrometry based on-line metabolomics for the analysis of the off-gas of bioprocesses.

| Technology | Variables monitored | Reference |
|---|---|---|
| Selected ion flow tube mass spectrometry (SIFT-MS) with a triple-quadrupole (QQQ) mass filter. Fermentations, using *Bacillus subtilis*, were performed in a 2 L bioreactor. | Thirteen volatile compounds, including acetaldehyde, acetoin, acetone, meso-2,3-butanediol, n-butanol, butanone, butyraldehyde, butyl acetate, diacetyl, dimethyl sulfide, ethanol, ethyl acetate, isoamyl alcohol, and isoprene. | Custer *et al.* (2003) |
| Mass spectrometer with a single quadrupole mass filter. Fermentations, using *B. subtilis* and *S. cerevisiae*, were performed in a 14 L bioreactor. | Acetoin and ethanol. | Oeggerlit and Heinzle (1994) |
| Flow injection secondary electrospray ionisation mass spectrometry (SESI-MS) with an orbitrap mass analyser. Fermentations, using *S. cerevisiae*, were performed in a 100 mL three-neck flask. | A total of 263, 636 and 43 m/z signals detected in three different experiments reported. Some of the metabolites annotated included acetic acid, ethanol, ethyl decanoate, farnesene and fatty acids of different chain lengths. | Tejero Rioseras *et al.* (2017) |
| Proton transfer reaction mass spectrometry (PTR-MS) with a single quadrupole mass filter. Fermentations, using *E. coli*, were performed in a 20 L bioreactor. | A total of 15 different m/z signals were monitored. Potential compounds identified were acetaldehyde, acetone, butanedione, ethanol, indole, isoprene, methanethiol and methanol. | Luchner *et al.* (2012) |
| PTR-MS with a time-of-flight mass analyser. Fermentations, using *S. cerevisiae*, *Metschnikowia pulcherrima* and *Torulaspora delbrueckii* yeast species, were performed in vials of unspecified volume. | At least 30-50 different m/z signals were monitored, although a clear list or potential annotations were not reported. | Berbegal *et al.* (2020) |
| Mass spectrometer with a magnetic sector analyser. Fermentations, using mammalian cells, were performed in 5, 15 and 50 L bioreactors. | Carbon dioxide and oxygen. | Behrendt *et al.* (1994), Goh *et al.* (2020) and Goudar *et al.* (2011) |

### Monitoring the liquid phase

As mentioned earlier, most reactions, metabolites and the biomass can be found in the liquid phase. For this reason, the monitoring of biomass and metabolites in the liquid broth has received a lot of attention in the last decades. Different technologies can be used to monitor biomass, including optical density (turbidity), dielectric spectroscopy, microscopy, flow cytometry, fluorescence spectroscopy, calorimetry and vibrational spectroscopy, whereas the monitoring of metabolites can be achieved using HPLC, vibrational spectroscopy, NMR and mass spectrometry (Marison *et al.*, 2012; Sonnleitner, 2012). Spectroscopy studies the absorption and emission of electromagnetic radiation by matter, whereas spectrometry is the measurement of a specific spectrum (a range of elements). Mass spectrometry, in particular, measures a range of ionised molecules from a sample and discriminates them according to their mass-to-charge ratio (m/z) (see Section 1.2.1).

Following, a brief description of each technology mentioned above is given, with special attention on those applied for the monitoring of metabolites.

*Optical density*

Optical density sensors are used to measure the turbidity in the fermentation medium caused by suspended matter that absorbs or scatters light (Grobbelaar, 2009; Kiviharju *et al.*, 2008). Light is emitted from a source and passes through the medium before reaching a detector. The higher the concentration of suspended matter, the less light will reach the detector. The light absorbed or scattered by the media is used to quantify the biomass. However, apart from biomass, other suspended matter and air bubbles can also contribute to the medium turbidity. For instance, media components such as macroscopic cellulosic substrates used for ethanol fermentation also cause light absorption making these turbidity sensors unpractical for biomass monitoring using this type of media (Cabaneros Lopez *et al.*, 2019).

*Dielectric spectroscopy*

Dielectric spectroscopy studies the response of a sample under an electric field of fixed or changing frequency. This response is described with two properties, the permittivity and conductivity, which respectively measure the capacity of a material to polarise (and therefore store energy) and let current flow through it (Deshmukh *et al.*, 2017). Cells have different layers (cell wall, cell membrane and cytoplasm) with very different biochemical compositions of proteins, lipids, polysaccharides, free ions, nucleic acids, small metabolites, *etc.* For this reason, these different layers have very different dielectric properties and make viable cells very polarisable when exposed to an alternating electrical field (Markx and Davey, 1999). Effectively, viable cells behave like electric capacitors when exposed to an alternating electrical field. Consequently, measuring the capacitance and conductivity of the fermentation medium

with in-line dielectric probes it is possible to monitor the concentration of viable cells (Harris *et al.*, 1987; Markx and Davey, 1999; Sonnleitner, 2012). Note that the capacitance is a function of the permittivity of the measured material and the dimensions of the capacitor (Harris *et al.*, 1987).

A strength of monitoring the biomass using dielectric spectroscopy is that this technology does not suffer interference from air bubbles and dead cells (which do not sustain an intact and polarisable cell membrane). This allows the monitoring of viable cells, which is of great use for bioprocess monitoring.

*Microscopy*

Counting cells under a microscope is one of the oldest methods to determine the concentration of biomass in a sample. Furthermore, staining techniques can be used to measure cell viability (Kamiloglu *et al.*, 2020). Modern developments in automation and data analysis have facilitated the coupling of microscopes to bioreactors for real-time fermentation monitoring. Bluma *et al.* (2010) described different in-line microscopy probes available for bioprocess monitoring. Image-processing software can be used to distinguish different particles and cell types.

Belini *et al.*, 2020 reported how in-line microscopy can be used to distinguish and monitor different morphological forms of *S. cerevisiae* (single cells, budding cells and pseudohyphae of at least three cells) during ethanol fermentation from sugarcane molasses. They could also detect morphological changes when the cells were heated from the process temperature (33 ℃) to 45 ℃ to simulate heat stress caused by inefficient cooling under industrial fermentation conditions.

*Flow cytometry*

Flow cytometry is a technology that can perform rapid single-cell analysis using lasers, detecting both fluorescent and light scattering signals. This information can be used to analyse and/or sort cells based on different properties. Light scattering can be used to analyse cell size (forward scattering) and internal complexity or granularity (side scattering). Furthermore, there is an extensive variety of dyes that can be used to analyse different parameters by fluorescence, such as cell viability, membrane potential or the concentration of nucleic acid, lipids or reactive oxygen species (da Silva *et al.*, 2012; McKinnon, 2018). For example, Kacmar and Srienc (2005) used flow cytometry with propidium iodide staining to measure the concentration and viability of chinese hamster ovary (CHO) cells in a bioreactor. R. Zhao *et al.* (1999) and Broger *et al.* (2011) respectively monitored population differences in *E. coli* and *Pichia pastoris* fermentation cultures during GFP production using on-line flow cytometry. In both cases, a relatively high percentage of the population (20 and 25 % respectively) was found

not to be GFP-producers due to plasmid or GFP gene loss at the end of the fermentation. Furthermore, Broger *et al.* (2011) demonstrated good correlation of the biomass concentration measurements obtained by on-line flow cytometry compared to alternative on-line and off-line methods.

An important consideration for using flow cytometry for on-line or at-line fermentation monitoring is that the instrumentation needs to incorporate a system for automatically diluting (to *ca.* $1x10^6$ cells/mL) and staining the cells to obtain accurate measurements of individual cells and achieve homogeneous staining. Different flow injection configurations have been reported in the literature (Abu-Absi *et al.*, 2003; Sonnleitner, 2012; R. Zhao *et al.*, 1999).

*Fluorescence spectroscopy*

Fluorescence is the phenomenon by which a material absorbs light at a high energy (short wavelength) and emits light of lower energy (longer wavelength). Several biomolecules are fluorescent, including tryptophan, NAD(P)H and flavin-containing molecules such as riboflavin, FAD or FMN. Furthermore, these molecules are closely linked to cellular growth, which makes it possible to use them as biomarkers for biomass concentration using in-line fluorescent probes during fermentation processes (Haack *et al.*, 2004; Hisiger and Jolicoeur, 2005; Surribas, Montesinos *et al.*, 2006). However, the levels of these molecules are closely dependent on the metabolic state of the cell and therefore these might change even if the concentration of cells does not, for instance at the end of biomass growth, or at the beginning of the product formation phase. NAD(P)H can be particularly sensitive to metabolic changes, as it can be rapidly oxidised to $NAD(P)^+$ depending on the redox state of the cell (Siano and Mutharasan, 1989; Surribas, Geissler *et al.*, 2006).

Nevertheless, fluorescence can be scanned over a wide range of excitation and emission wavelengths in a so-called 2D fluorescence spectroscopy (Marose *et al.*, 1998). This way, multiple fluorophores can be used to model biomass concentration. This approach also allows the use of multivariate analysis to exploit the whole range of excitation and emission wavelengths measured, not requiring a pre-selection of fluorophores. Surribas, Geissler *et al.* (2006) used this approach to monitor biomass concentration, substrate (glycerol) concentration and recombinant *Rhizopus oryzae* lipase (product) activity in *Pichia pastoris* batch fermentations. Substrate concentration and product formation were stoichiometrically well correlated to biomass growth, which made their prediction possible even though glycerol is not a fluorophore, and the lipase produced is not a fluorescent protein (although it will generate some fluorescent signal due to its aromatic amino acids). Bayer *et al.* (2020) also used a multivariate data analysis approach to build a model for biomass monitoring in an *E. coli* fed-batch process using fluorescent spectroscopy. Despite the added value that these multivariate

models could potentially bring over mechanistic models using specific fluorophores, Bayer *et al.* (2020) and Surribas, Geissler *et al.* (2006) reported that the chemical species with the highest contribution to their multivariate models to monitor biomass corresponded to the excitation and emission wavelengths of flavins, NAD(P)H and tryptophan.

*Calorimetry*

Microbial growth is stoichiometrically linked to heat formation (Maskow and Paufler, 2015). For aerobic processes, the metabolic heat production has experimentally and theoretically been found to be 440 $\pm$ 33 kJ for every mole $O_2$ consumed (Birou *et al.*, 1987). Thanks to these stoichiometric correlations, measuring heat flows around the bioreactor it is possible to correlate the measured heat production rate to the biomass formation rate (Müller *et al.*, 2018). Compared to other methods of biomass monitoring, calorimetry requires relatively simple instrumentation, namely temperature probes and flow meters.

The methods described above are used to monitor the biomass. The methods used to monitor metabolites will be described below.

*High-performance liquid chromatography*

High-performance liquid chromatography (HPLC) is one of the most widely used techniques in analytical chemistry. Using columns of different properties, it allows to separate compounds of a mixture based on their physicochemical characteristics (hydophobicity, charge, size, *etc.*). After eluting from the column, the separated compounds are analysed in a detector, typically an absorbance, fluorescence or refractive index detectors, absorbing wavelengths in the ultraviolet (190-400 nm), visible (400-700 nm) or infrared (800 nm - 1 mm) range, or a mass spectrometer (Bélanger *et al.*, 1997).

Examples of on-line HPLC analysis have been reported in the literature. Warth *et al.* (2010) used a filtration probe to send cell-free samples to a valve and then into an on-line HPLC system equipped with a refractive index detector to analyse extracellular organic acids, alcohols, and carbohydrates in an *E. coli* fermentation every 10 minutes (chromatography run time). Koliander *et al.* (1990) developed a similar flow injection system using a magnetically stirred filtration unit to remove the cells and used it for monitoring *S. cerevisiae*, *Streptomyces olivaceus* and *Arthrobacter sp.* fermentations every 10-30 minutes, depending on the analyte.

Koch *et al.* (2016) also used a biomass filtration probe to analyse by on-line HPLC Penicillin V and its precursor phenoxyacetic acid in a *Penicillium chrysogenum* fermentation process every 16 minutes. The flow injection system had a flow-through cell from which the filtered samples were sent to the HPLC. Tohmola *et al.* (2011) developed an on-line HPLC system to monitor glucose, glycerol, acetate, and ethanol in *S. cerevisiae* fermentations every five minutes. Their system contained a sample collector and a cross-flow filter to remove the cells

prior to injection. Furthermore, between samples they implemented a cleaning step for the tubing lines with 70 % ethanol and another one for the cross-flow filter using 0.9 % saline. Nevertheless, some data points were missing in their time-course data graphs, indicating potential system blockages.

Overall, on-line HPLC has been successfully used to monitor key fermentation compounds. However, typical time delays between samples due to separation in the column were generally close to 10 minutes. Furthermore, all these systems were implemented using biomass filtration units, only analysing extracellular metabolites. Finally, the use of both chromatography columns and filters makes this technology susceptible to blocking.

*Vibrational spectroscopy - Infrared spectroscopy*

Infrared (IR) spectroscopy studies the vibration of intramolecular bonds when irradiated with infrared light at a given wavelength. If the frequency of the IR photon is the same as the frequency of vibration of the chemical bond, this will absorb some of the photon energy and will vibrate (Larkin, 2011a; Stuart, 2005). The two main types of vibration measured are stretching and bending, and some examples of chemical bonds that vibrate under IR radiation are C−H, O−H, N−H, C=C, C=O, C≡C and C≡N (Stuart, 2005). Consequently, IR probes can be used in-line to detect a vast array of compounds.

IR spectroscopy is a fast, non-invasive and non-destructive technology (Sonnleitner, 2012). However, an important disadvantage for monitoring fermentations is that both gas bubbles and vibrations from the impeller can cause disturbance to the IR probe readings (Koch *et al.*, 2016). Another disadvantage is that biological mixtures are chemically very complex, leading to highly convoluted IR spectra containing a lot of overlapping peaks. Furthermore, compounds at higher concentrations often mask those more diluted ones. This requires the use of first- and second-order derivatives to deconvolute the different peaks of the spectra to create chemometric models for metabolite quantification (Marison *et al.*, 2012; Stuart, 2005). These models are time consuming to build and usually not transferable, *i.e.* they are only applicable to the configuration used to build them (bioreactor, medium, strain, *etc.*) (Marison *et al.*, 2012; Roggo *et al.*, 2007).

The IR region is split into three groups depending on their wavelength: the near-infrared (NIR) (800-2500 nm), mid-infrared (MIR) (2500-25,000 nm) and far-infrared (FIR) (25,000-100,000 nm) (Marison *et al.*, 2012). The high absorbance of water in the FIR region makes it impractical for fermentation monitoring, therefore, the focus is put on NIR and MIR. The primary (or fundamental) energy level of vibration of most of the bonds found in biomolecules are excited in the MIR electromagnetic range, producing sharper and stronger signals than in the NIR region (Sonnleitner, 2012; Stuart, 2005). The NIR electromagnetic region, on the other hand,

causes the excitation of chemical bonds to energy levels of vibration above the fundamental one (called overtones or combination bands). As a result, NIR absorption bands are broader, weaker and more overlapped, making NIR less sensitive to low concentration analytes and complicating peak assignment (Pu *et al.*, 2020; Svendsen *et al.*, 2015; Vann *et al.*, 2017).

Infrared and Raman spectroscopy are so far the most successful technologies employed for liquid-phase monitoring of metabolites in biotechnology. Some of these examples are gathered in Table 1.3.

*Vibrational spectroscopy - Raman spectroscopy*

Raman spectroscopy studies the vibrations of intramolecular bonds when irradiated with a laser (A. C. McGovern *et al.*, 2002). Instead of being a resonance effect (as in IR), in Raman spectroscopy a high-energy photon hits the molecule and loses part of its energy to make the molecule vibrate and polarise. The remaining energy is scattered as a lower-frequency photon (Colomban and Gouadec, 2008; Larkin, 2011b). Compared to IR, Raman spectroscopy is better at exciting non-polar chemical groups, while IR spectroscopy is better at exciting polar ones (Larkin, 2011a).

In common with IR spectroscopy, Raman spectroscopy is label-free, non-sample-destructive and can be employed using *in situ* sensors (Golabgir and Herwig, 2016; H. L. T. Lee *et al.*, 2004). Furthermore, Raman spectroscopy shows little water interference, which is an important advantage for bioprocess monitoring (Shalabaeva *et al.*, 2017). Although Raman peaks are narrower than with IR spectroscopy, they also show a lot of overlap, especially in biological mixtures, requiring the use of chemometric models to deconvolute the different peaks (Abu-Absi *et al.*, 2011; Iversen *et al.*, 2014; H. L. T. Lee *et al.*, 2004; Shalabaeva *et al.*, 2017). One of the biggest challenges of Raman spectroscopy is background noise caused by fluorescence (Jestel, 2010), which can be caused by several biomolecules. Low-frequency lasers can be used to reduce fluorescence background, but these lasers need long exposure times, which can lead to undesired localised heating effects (Marison *et al.*, 2012).

Some examples of the application of Raman spectroscopy for fermentation monitoring are gathered in Table 1.3.

**Table 1.3:** Examples of applications of infrared and Raman spectroscopy for monitoring fermentation processes.

| Technology | Variables monitored | Comments | Reference |
|---|---|---|---|
| On-line MIR spectroscopy | Biomass, ethanol, fructose, glucose, glycerol and sucrose. | *S. cerevisiae* process for ethanol production in a 2 L bioreactor. Five fermentation runs were used to calibrate and test a PLS[1] regression model for each item monitored using off-line HPLC data and weight measurements as reference concentration values for metabolites and biomass, respectively. MIR spectra were acquired every 10 minutes. | Rodrigues *et al.* (2018) |
| On-line MIR spectroscopy | Glucose, glycerol, ammonium and acetate | *E. coli* biomass formation batch process. A library of pure component spectra was used to build the model, which was tested with nine fermentations. Off-line concentrations for checking the model performance were measured by enzymatic assay. MIR spectra were acquired every 2 minutes. | Schenk *et al.* (2008) |
| In-line NIR spectroscopy | Biomass, ethanol, glucose and glycerine. | *S. cerevisiae* process for ethanol production in a 2 L bioreactor. Five fermentation runs were used to calibrate and test a PLS regression model for each item monitored using off-line HPLC data and UV absorption as reference concentration values for metabolites and biomass, respectively. Although spectra acquisition frequency was not clearly reported, it was stated that each NIR spectrum was the average of 64 scans and that spectrum can be acquired in less than 1 minute. | do Nascimento *et al.* (2017) |
| In-line NIR spectroscopy | Ammonia, glucose, glutamine and lactate | CHO process for the production of a human protein in a 2 L bioreactor. A regression model for each item monitored was developed using off-line enzymatic assay data as reference concentration values. Average spectra was acquired using 32, 64 and 128 scans. | Arnold *et al.* (2003) |
| In-line Raman spectroscopy | Ammonia, biomass (viable and total), glucose, glutamate, glutamine and lactate. | CHO process for the production of an antibody fusion protein in a 500 L bioreactor. Four bioreactor runs were used to calibrate and test a partial least squares (PLS) regression model for each item monitored. Reference values were determined using a Cedex cell counter and a BioProfile 400 Analyser for cell density and metabolite concentrations, respectively. Raman spectra were acquired every two hours. | Abu-Absi *et al.* (2011) |
| In-line Raman spectroscopy | Biomass, ethanol and glucose. | *S. cerevisiae* process for the production of ethanol in a 1 L bioreactor. Standard solutions were used to calibrate a linear regression model for each item monitored using off-line HPLC data and UV absorption as reference concentration values for metabolites and biomass, respectively. The water signal was used to determine the concentration of biomass, which in turn was used to correct signal suppression of the other analytes consequence of Mie scattering caused by the biomass . Raman spectra were acquired every 15 minutes. | Iversen *et al.* (2014) |
| In-line Raman spectroscopy | Monoclonal antibody glycosylation site occupancy. | CHO process for the production of Immunoglobulin G (IgG) antibody in a 2 L bioreactor. Six bioreactor runs were used to calibrate and test a PLS regression model for total and non-glycosylated monoclonal antibody. Reference values of total and non-glycosylated antibody were determined using a photometric analyser and by LC-MS, respectively. Raman spectra were acquired every 20 minutes. Additionally, viable cell density and cell viability were monitored at-line using a cell counter, and ammonium, glucose, glutamine and lactate were monitored at-line using a photometric analyser. | M.-Y. Li *et al.* (2018) |
| In-line Raman spectroscopy | Acetate, butanol, butyrate and glucose | *Clostridium acetobutylicum* process for the production of acetate, butanol and butyrate in a 2 L bioreactor. Four fermentation runs were used to calibrate a PLS regression model for each item monitored using off-line HPLC data as reference concentration values. Raman spectra were acquired every 1 hour. | Zu *et al.* (2017) |

[1] Partial least squares (PLS) will be discussed in Section 1.2.4

*Nuclear magnetic resonance*

Nuclear magnetic resonance (NMR) is a technology based on the excitation of atomic nuclei when irradiated with a magnetic field. There have been some examples in the literature using NMR for on-line fermentation monitoring at laboratory scale. The first demonstrations of this technology were the so-called in-magnet bioreactors, which consisted on miniaturising the fermenter into conventional or adapted NMR tubes, which were placed inside the NMR instrument for the whole fermentation run (Brecker *et al.*, 1999; Majors *et al.*, 2008). Kreyenschulte *et al.* (2015) reported a more scalable on-line approach consisting of a circulation loop drawing unfiltered samples from the fermentation broth and pumping them through a low-field NMR spectrometer before aseptically returning them into the fermenter. The logic behind using a low-field NMR rather than more precise high-field NMR instruments is the reduced size, complexity and price of the former, which makes them a more realistic choice for industrial purposes. Legner *et al.*, 2019 more recently reported a similar setup using NMR and Raman spectrometers in a flow cell to monitor on-line an ethanol fermentation with *S. cerevisiae* cells immobilised in alginate beads.

Similar to vibrational spectroscopy, a limitation of NMR for bioprocess monitoring is the presence of overlapping peaks, which limits the number of compounds that can be detected and quantified; usually less than ten (Brecker *et al.*, 1999; Kreyenschulte *et al.*, 2015; Majors *et al.*, 2008).

*Mass spectrometry*

The use of mass spectrometry for bioprocess monitoring is yet a largely unexplored field with a vast potential. Investigating the potential of this technology for fermentation monitoring is the purpose of this thesis. An introduction to mass spectrometry and metabolomics is covered in the next section.

## 1.2 Metabolomics

The word metabolome was first coined by Oliver *et al.* (1998) and it describes the global collection of small to medium size molecules (*i.e.* up to 1000 - 2000 Da) present in the metabolism of a given biological system. As metabolites are at the final step of biological regulation, metabolite concentrations and metabolic fluxes provide the best picture of cellular phenotypes (Farrell *et al.*, 2014; Fiehn, 2002). The analysis of the metabolome is studied with metabolomics, typically with mass spectrometry and NMR instruments (Idle and Gonzalez, 2007. Although NMR is more quantitative and better at elucidating molecular structures, mass spectrometry is a more sensitive technology with a wider detection capacity, faster analytical

throughput (unless it is coupled to chromatography) and requires a much smaller sample volume (10-100 $\mu$L compared to 100-500 $\mu$L) (Cajka and Fiehn, 2015; Coene *et al.*, 2018; Wishart, 2016). For these reasons, mass spectrometry will be the technique of choice for real-time metabolomics monitoring of fermentation processes studied in this thesis.

### 1.2.1 Fundamentals of mass spectrometry

Mass spectrometry is an analytical technique capable of detecting a vast array of chemical species, from small metabolites to entire proteins (Chi *et al.*, 2007; Ge *et al.*, 2002). Other key advantages of mass spectrometry include its analytical speed (Stanbury *et al.*, 2017; Vogel and Todaro, 2014), its ability to simultaneously detect hundreds of compounds and its high sensitivity - some compounds can be detected at concentrations as low as a few parts-per-billion (ppb) (Lindinger and Hansel, 1997; Luchner *et al.*, 2012). Often, mass spectrometry is coupled to chromatography for prior separation of the different molecules of a mixture before analysing them. These hyphenated technologies are liquid chromatography-mass spectrometry (LC-MS), for analysing compounds in the liquid phase, and gas chromatography-mass spectrometry (GC-MS), for molecules in the gas phase. The aim of this thesis is to study mass spectrometry for the analysis of fermentation metabolites in the liquid phase, therefore, the focus will be placed on mass spectrometry and LC-MS.

Mass spectrometers detect gas-phase ions, even when studying liquid mixtures. Consequently, mass spectrometers perform three distinct steps when analysing liquid mixtures; production, separation and detection of gas-phase ions. These are carried out in the three parts of a mass spectrometer described below and as indicated in Figure 1.4:

- Ion source: Generates gas-phase ions that will enter the mass spectrometer.
- Mass analyser: Filters/separates the different ions by employing electrical and/or magnetic fields.
- Ion detector: Detects the ions filtered by the mass analyser.

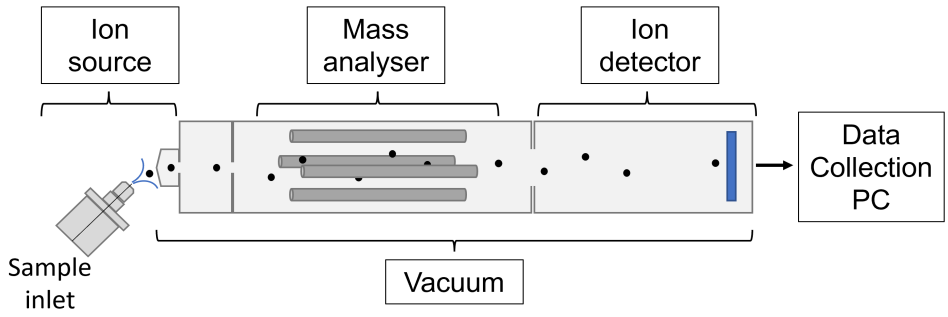



**Figure 1.4:** Schematic structure of a mass spectrometer with a quadrupole mass analyser. Gas-phase ions are represented as black circles.

### Ion Source

The main function of the ion source in an LC-MS system is to ionise the molecules in the sample and get them into the gas phase. Most conventional modern mass spectrometers use atmospheric pressure ionisation methods. These are also known as soft ionisation methods because they cause little to no molecule fragmentation during the ionisation step (Altuntaş *et al.*, 2013). The most common atmospheric pressure ionisation methods include:

- Electrospray ionisation (ESI)
- Atmospheric pressure chemical ionisation (APCI)
- Atmospheric pressure photo ionisation (APPI)

Out of the three, ESI is the most widely used in modern mass spectrometers (particularly for LC-MS) and it performs well for molecules of medium to high polarity (Figure 1.5). APCI is an alternative ionisation method for non-polar molecules such as lipids and steroids, and APPI is mostly used for aromatic hydrocarbons, which do not ionise easily. Due to its wider applicability and frequency of use, this thesis will focus on ESI.

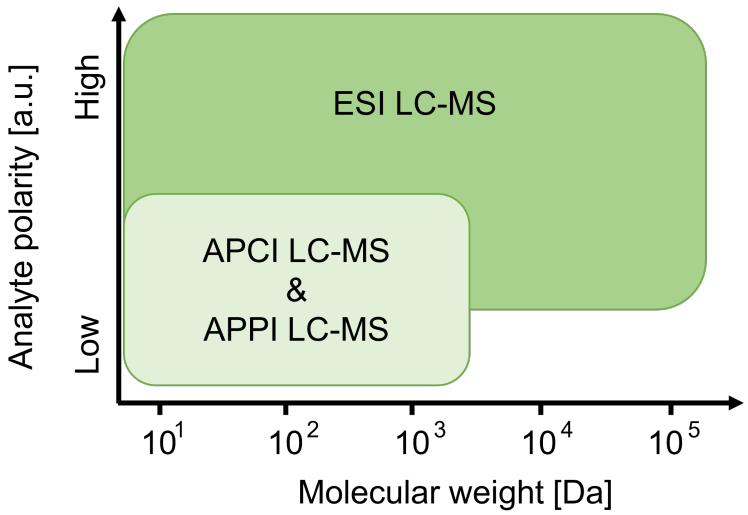



**Figure 1.5:** Use range of different mass spectrometry atmospheric pressure ionisation modes depending on analyte chemistry.

*Electrospray Ionisation*

Figure 1.6 shows a schematic representation of ESI. In short, a capillary tip continuously ejects sample droplets into the ionisation chamber. An applied potential difference between the capillary and the sampling cone of the mass spectrometer causes the liquid in the capillary to form a conical shape (Taylor cone), which facilitates droplet formation. The high voltage also causes ionisation of the analytes in the liquid phase and the selection of ions of a particular charge. Following this, the droplets containing the ionised analytes shrink due to solvent evaporation, thus increasing the charge density in the droplet. When the repulsive forces

inside the droplet become higher than the droplet surface tension (Rayleigh instability limit), the droplet explodes in a process known as Coulomb fission. This leads to the formation of smaller droplets that will continue evaporating and exploding. This process eventually favours the ionised molecules to move to the gas phase, either due to solvent evaporation to dryness - more predominant mechanism for bigger molecules - or due to the ion being evaporated of the droplet - more predominant mechanism for smaller molecules (Gross, 2017; Konermann *et al.*, 2013; Yamashita and Fenn, 2002). The mass spectrometer is under high vacuum, which together with the voltage potential causes these gas phase ions to enter the mass spectrometer.

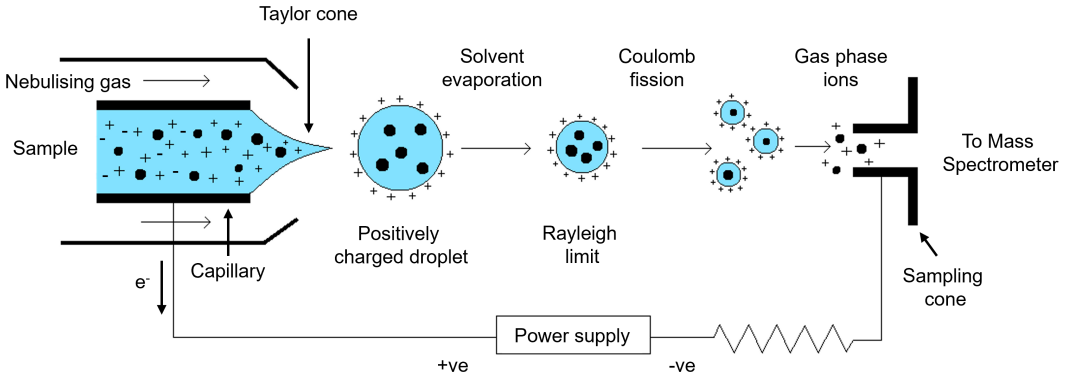



**Figure 1.6:** Electrospray ionisation diagram on positive ionisation mode. The analytes of interest are represented as black circles.

There are many different factors affecting the ionisation performance in ESI, the main ones being:

- Applied potential difference: Increasing it improves ionisation up to a certain voltage, when the signal starts becoming less reproducible and the noise signal increases.
- Eluent flow rate: Decreasing it helps solvent droplet reduction and therefore ion formation. However, slow flow rates lead to wider chromatographic peaks and a slower analysis.
- Choice of mobile phase: solvents with low surface tension (such as methanol, isopropanol or acetonitrile) are usually chosen as they allow a stable and reproducible electrospray.
- Nebulising gas flow rate and temperature: Using a nebulising gas in the ESI chamber can be beneficial to reduce solvent droplet size, which in turn favours gas-phase ion formation and analyte signal (Bruins *et al.*, 1987), especially at high nebulising gas flow rates and temperatures, up to a point where the spray becomes unstable. Furthermore, the use of a nebulising gas allows an increase in the eluent flow rate yet still retaining good ionisation.

- Capillary position: If properly adjusted, it can considerably decrease noise due to the removal of uncharged species. Off-axis to the sample plate is commonly the best performing capillary position (Bruins *et al.*, 1987).

*Adduct formation and in-source fragmentation*

Common ions analysed in metabolomics using ESI ionisation are protonated and deprotonated species, in positive and negative mode respectively. Given a molecule (M), these ions are notated as $[M+H]^+$ and $[M-H]^-$. However, this molecule M can also form different ions by interacting with species in the mobile phase, for example with $Cl^-$ or $Br^-$ anions in negative mode, or with $Na^+$ or $K^+$ cations in positive mode (Cech and Enke, 2001). It is also possible to form ions with multiple charges, for example if a molecule loses or gains three protons. Taking the example of glucose, which has a monoisotopic mass of 180.0634 Da, a few common adducts that may be observed are listed below, in ascending m/z order:

- $[M-3H]^{3-}$: 59.0139 m/z
- $[M+3H]^{3+}$: 61.0284 m/z
- $[M-2H]^{2-}$: 89.0244 m/z
- $[M-H2O-H]^-$: 161.0450 m/z
- $[M-H]^-$: 179.0562 m/z
- $[M+H]^+$: 181.0712 m/z
- $[M+Na]^+$: 203.0526 m/z
- $[M+Cl]^-$: 215.0328 m/z
- $[M+FA-H]^-$: 225.0616 m/z
- $[2M-H]^-$: 359.1195 m/z

The list above is only a short example. An extensive list of possible adducts has been reported by Huang *et al.* (1999). When looking for potential adducts, it is important to consider the sample matrix, *i.e.*, the chemical species in the sample that can lead to the formation of adducts, for instance, the presence of salts and the solvent used. The use of mobile phase additives such as weak acids and bases - which can work as proton donors or acceptors - can help to generate $[M-H]^-$ and $[M+H]^+$ ions. Typically, 0.1-1.0 % acetic acid or formic acid is used in positive mode, while 10-20 mM of an ammonium salt (typically ammonium acetate, ammonium carbonate or ammonium formate) are used in negative mode.

Apart from adducts, some fragmentation can also occur during the ionisation process due to the high voltages. This phenomenon is called in-source fragmentation and can lead to new product species (Y. .-. Xu *et al.*, 2015).

### Mass Analyser

Ions generated in the ionisation chamber are discriminated in the mass analyser according to their m/z. Mass analysers can be operated to either specifically analyse selected ions of a specific m/z value or to scan the sample over a range of m/z values, depending on how targeted the analytical method is. More details about targeted and untargeted mass spectrometry can be found on Section 1.2.2.

In order to discuss some of the main types of mass analysers, it is important to introduce the concepts of mass resolution and mass accuracy.

*Mass resolution*

In general terms, the resolution of a mass spectrometer is its capacity to distinguish ions that are similar in mass. Although there are different ways to formally define mass resolution, the most common way to measure it is using the full width at half maximum (FWHM) of the peak (Hoffmann and Stroobant, 2007; Mulvaney, 1993). That is, dividing the m/z value of a specific ion by the width of its peak at half of its maximum height, as shown in equation 1.1 and Figure 1.7. In essence, the sharper the peaks are, the higher the resolution. When reporting mass resolution and comparing different instruments, it is important to mention the mass of the ion used, as it will influence the value obtained.

$$Resolution = \frac{m/z_{measured}}{FWHM} \tag{1.1}$$

*Mass accuracy*

Mass accuracy is the difference between the theoretical and measured m/z value of a molecule, and it is often expressed in parts per million (ppm) units (Hoffmann and Stroobant, 2007), as shown in equation 1.2. An example is illustrated in Figure 1.7.

$$Accuracy = \frac{m/z_{theoretical} - m/z_{measured}}{m/z_{theoretical}} \cdot 10^6 \tag{1.2}$$

After introducing these concepts, the most common types of mass analysers are explained below. Special attention will be paid to quadrupoles and Orbitraps, as these are the analysers that will be used in this thesis.

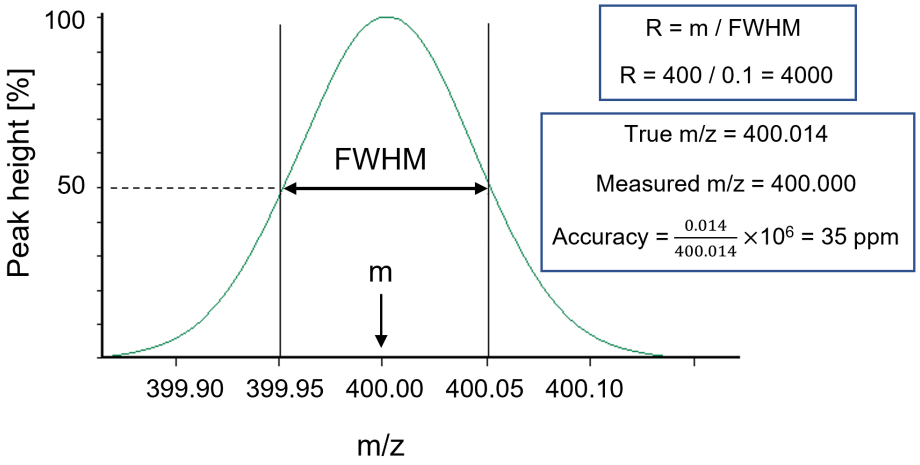



**Figure 1.7:** Mass resolution (R) and mass accuracy definitions.

*Quadrupole*

Quadrupole mass analysers comprise four parallel equidistant rods with an applied static potential (DC voltage) and an alternating potential (RF voltage) field. Opposing rods have the same alternating voltage and polarity at all times. The alternating voltage creates a fluctuating electric field between the four rods, which filters ions of a certain m/z allowing them to pass through the analyser, whereas non-selected ions acquire an unstable trajectory that makes them leave the quadrupole or collide with one of the four rods, thus not reaching the ion detector (Paul and Raether, 1955) (Figure 1.8).

Quadrupoles are reproducible and inexpensive but have low resolution. Nevertheless, quadrupoles are very useful when employed as mass filters and fragmentation cells in tandem mass spectrometers, as it will be discussed in Section 1.2.1.

*Quadrupolar Ion Trap*

An ion trap is an instrument that stores ions by applying an oscillating electric field in either two or three dimensions. 2D ion traps are like a normal quadrupole but ions are retained oscillating inside its axial dimension by the presence of repulsive lenses at the ends of the quadrupole. 3D ion traps are commonly described as a quadrupole bent on itself and closed on a loop, with the inner rod being reduced to the centre of the trap. Both 2D and 3D ion traps use an oscillating electric field to keep the ions of all masses oscillating together in the trap. An RF resonant frequency is applied to eject the ions of a given mass from the trap (Hoffmann and Stroobant, 2007).

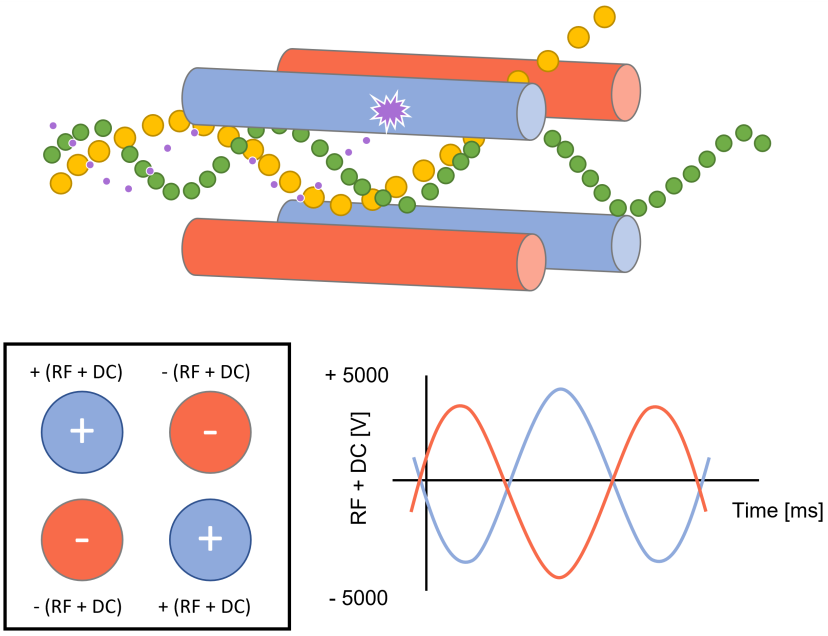



**Figure 1.8:** Diagram of a quadrupole mass analyser showing an example with three ions, where the green ones are selected while the yellow and purple are filtered out (**top**). Diagram of the oscillating voltage of the four rods with opposing rods sharing the same properties (**bottom**)

Ion traps have high sensitivity and moderate resolution (depending on the instrument) but limited ion capacity (easily saturated) due to charge repulsion of the contained ions. Furthermore, spectral quality decreases if the ions are stored in the trap for too long.

*Time-of-flight*

In time-of-flight (TOF) analysers, ions are extracted from the ion source in short bursts and subjected to an accelerating region (homogeneous electric field) followed by a drift (field-free) region before reaching the detector (Boesl, 2017). Ions with a higher m/z ratio will travel (or fly) slower, and therefore be detected later. This time difference to reach the detector is what allows to distinguish ions of different masses (Cameron and Eggers, 1948; Wolff and Stephens, 1953).

TOF analysers have the highest mass range of all mass spectrometry analysers and have high resolution and mass accuracy (Hoffmann and Stroobant, 2007). However, TOFs are very sensitive to temperature changes and their resolution decreases as the m/z value increases. The latter phenomenon is observed due to the linear dependency of the flying time to the square root of the m/z value (Boesl, 2017). Nevertheless, TOF resolution can easily be optimised by designing longer flying paths or by using a reflectron (ion mirror), which in turn also reduces peak width by homogenising the stochastic dispersion of ion velocities for ions of the same m/z produced in the accelerating region (Mamyrin *et al.*, 1973).

*Orbitrap*

An Orbitrap mass analyser consists of an outer barrel-like electrode and a central spindle-like electrode along the axial axis (Figure 1.9). Both electrodes have an independent voltage applied to them and are separated by vacuum. Ions are introduced in this space and keep oscillating between both inner and outer electrodes - similarly to an ion trap - at a speed that depends on their m/z (Makarov, 2000). Detector plates at the ends of the outer barrel measure the current generated by the oscillating ions, which generate an intensity signal as a function of time. Then, Fourier transforms are used to mathematically convert this signal function in the time domain into a new function in the frequency domain. This new frequency function will contain the different frequencies of all the corresponding m/z ions oscillating in the Orbitrap. Finally, these frequencies are mathematically converted into m/z values, which allows the generation of the mass spectrum (Lange *et al.*, 2014; Scigelova *et al.*, 2011). Consequently, Orbitraps combine ion separation and detection at the same time.

Orbitrap mass analysers show high mass resolving power (up to 1,000,000) and mass accuracy (up to <1ppm for internal calibration) (Thermo Fisher Scientific, 2017), as well as good sensitivity and dynamic range (>$10^3$) (Hu *et al.*, 2005). In this thesis, an Exactive$^{TM}$ Orbitrap and a Q Exactive$^{TM}$ Orbitrap (Thermo Scientific$^{TM}$) will be used for on-line metbolomics and off-line LC-MS. A diagram of these two instruments is shown below in Figure 1.10.

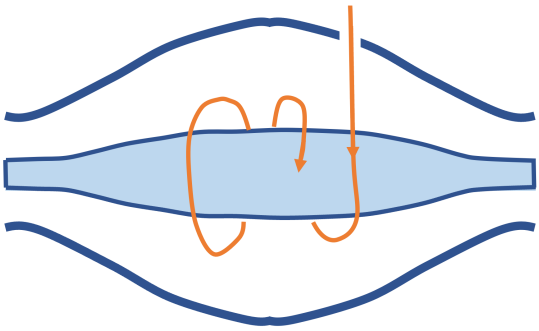


**Figure 1.9:** Diagram of an Orbitrap mass analyser. The ions introduced oscillate at different frequencies depending on their m/z.

The Orbitrap Exactive has an ion source, followed by ion focusing devices such as lenses and flatapoles (ion focusing metal electrodes), which make sure that neutral particles are removed before reaching the mass analyser. Then, the ions travel through the instrument via transmission multipoles (typically quadrupoles, hexapoles or octapoles, depending on the number of rods) until they reach a C-trap, which is a curved quadrupolar ion trap that collects the beam of ions and controls which ions are sent to the Orbitrap mass analyser. In the Orbitrap, ions oscillate at different frequencies depending on their m/z, which allows their differential detection. The Orbitrap Q Exactive is very similar but it contains a quadrupole mass

filter before the C-trap and a quadrupole for fragmentation (high-energy collision dissociation or HCD) after the C-trap. Fragmented ions can be sent back to the C-trap and from there to the Orbitrap. This allows the collection of fragmentation information, which can help identify metabolites (see Section 1.2.2).

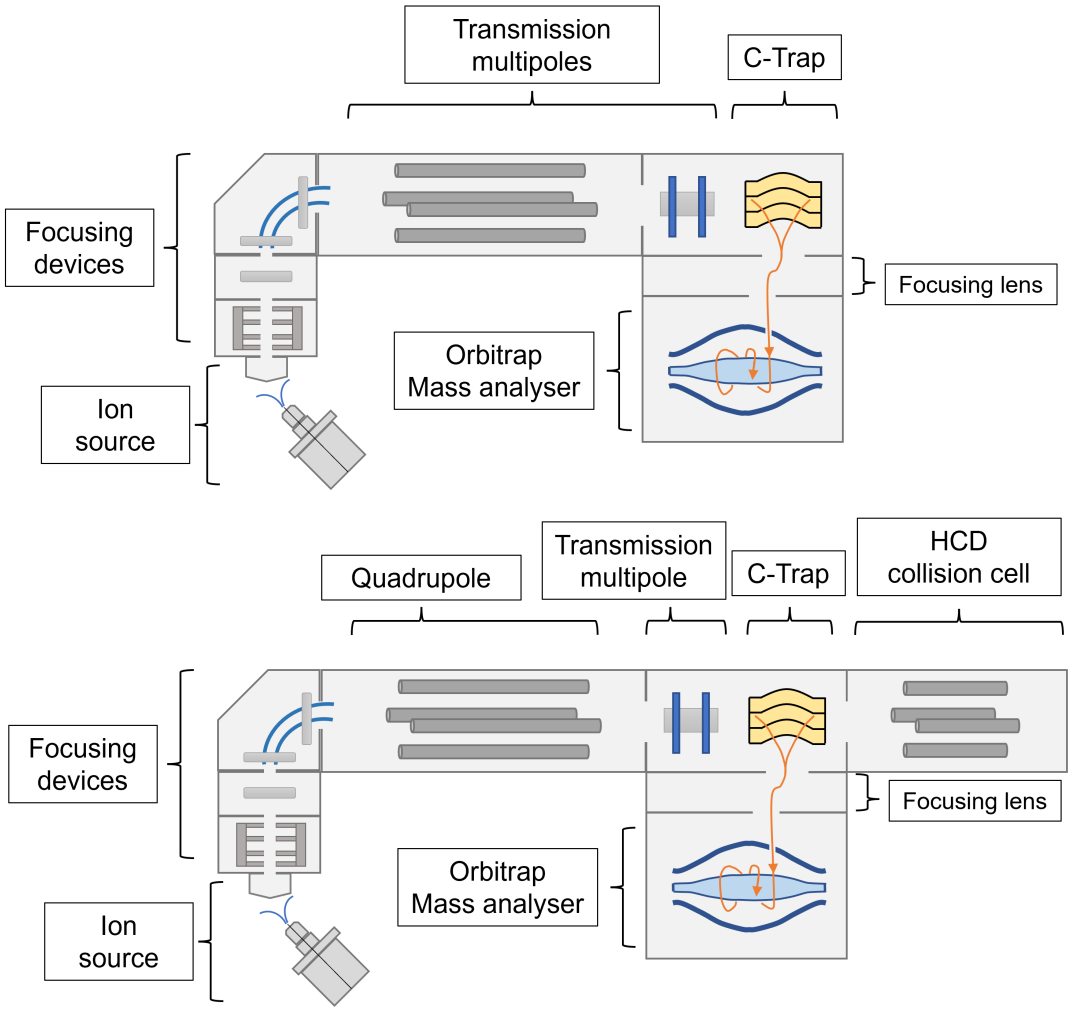



**Figure 1.10:** Diagram of the Orbitrap Exactive (**top**) and Q Exactive (**bottom**) mass spectrometers with their main components. Both instruments will be used in this thesis.

*Tandem mass spectrometry*

Multiple mass analysers (commonly two) can be combined in a single mass spectrometer in the so called tandem mass spectrometers. The first analyser is usually a quadrupole used as a mass filter, followed by a collision cell containing an inert gas (usually $N_2$ or Ar) which causes ion fragmentation and, finally, the second mass analyser - for example, another quadrupole, a TOF or an Orbitrap. Fragmentation allows for the gathering of structural information of the selected ions to better elucidate the nature of the original molecules.

Triple quadrupoles are an example of tandem mass spectrometry, where three quadrupoles are combined in a row. Usually, the first and third quadrupoles are configured as ion filters (as normal quadrupoles), while the middle one is configured as a collision cell. Triple quadrupole mass analysers are the most popular option for performing quantitative mass spectrometry analyses, mainly due to their high linear dynamic range and speed in transitioning between the different target metabolites to analyse (Gross, 2017). In this thesis, a TSQ Quantiva triple quadrupole mass spectrometer (Thermo Scientific$^{TM}$) (Figure 1.11) will be used for quantitative targeted metabolomics (see Section 1.2.2).

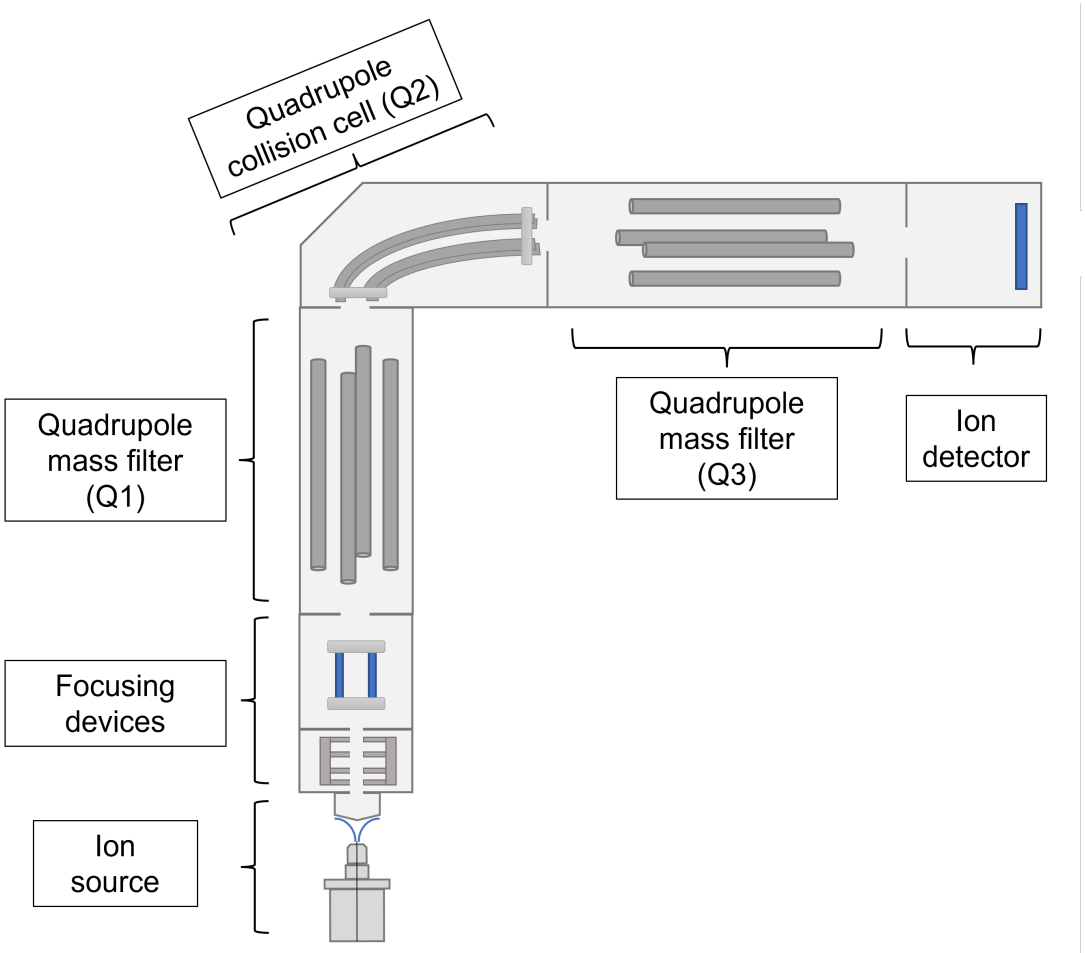



**Figure 1.11:** Diagram of the TSQ Quantiva triple quadripole mass spectrometer with its main components, which will be used in this thesis.

Figure 1.11 shows a diagram of the TSQ Quantiva. Briefly, the instrument contains ion focusing elements that remove some neutral species and allow the ions to reach the first quadrupole (Q1), where ions of a specific m/z can be filtered. Then, the selected ions reach the second quadrupole (Q2), which is typically used as a collision cell to generate fragment ions that will reach the third quadrupole (Q3). There, specific fragment ions can be filtered and be sent into the ion detector. Q1 and Q3 can be used to scan over a range of m/z values, or

to filter ions with a specific m/z. Combinations of these two operation modes provide a lot of versatility for developing different analytical methods. For instance, Q1 can be used to select ions of a specific m/z, which can be fragmented in Q2, and then, Q3 can scan over a range of m/z values to identify different fragment ions generated from the parent molecule filtered in Q1. More details on how triple quadrupoles can be used to monitor specific molecules are provided in Section 5.1.1.

#### Ion detector

The main functions of the detector are to count the ions that have passed the analyser and to amplify the signal. The most common detectors found in mass spectrometry are electron multipliers, but other types still exist such as faraday cups and electro-optical ion detectors (Hoffmann and Stroobant, 2007).

### 1.2.2 Untargeted and targeted metabolomics

Metabolomics is usually classified into two categories: untargeted and targeted metabolomics. Untargeted metabolomics is the global analysis of as many metabolites as possible in a given biological system. Although different extraction and analytical methods will favour specific compounds, an untargeted approach tries to analyse all measurable metabolites in a given sample. Targeted metabolomics is the analysis of a predefined specific set of compounds, which are identified by comparison to reference standards using information such as retention time, molecular mass, fragmentation patterns or collision cross section.

Confident metabolite annotation is more challenging in untargeted metabolomics, hence this method is often regarded as “discovery-based” or “hypothesis-generating”, while targeted metabolomics is considered “validation-based” or “hypothesis-driven” (Schrimpe-Rutledge *et al.*, 2016). This way, a common approach in metabolomics studies is to do an initial untargeted analysis to get a comprehensive overall idea of the results and identify potential metabolites of significant interest, then follow this up with a targeted analysis to validate these results.

The annotation of compounds plays a significant role in metabolomics analysis; this topic is further expanded below.

### 1.2.3 Metabolite annotation

Sumner *et al.*, 2007 described four different levels of compound identification for metabolomics results (Figure 1.12), and this has become the reference system to categorise annotation confidence in the field of metabolomics. According to this system, the highest degree of confidence is only achieved when the samples are compared to reference standards.

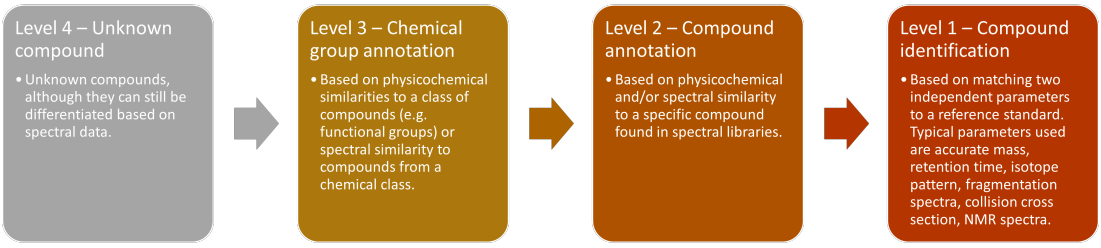



**Figure 1.12:** Metabolite identification levels, as described by Sumner *et al.* (2007).

Although untargeted metabolomics provides a lower degree of confidence in the annotations obtained, its strength lies in the large number of metabolites detected, which can provide a lot of information. Using the right tools, this information can generate a lot of biological context to understand and interpret results. Some of these tools are multivariate analysis and pathway analysis.

### 1.2.4 Multivariate analysis

Untargeted metabolomics typically results in hundreds or thousands of metabolite annotations for biological samples. This large number of features can easily become overwhelming to interpret. To deal with these large sets, different methods of data visualisation and dimensionality reduction are commonly used such as heatmaps, principal component analysis (PCA) and partial least squares – discriminant analysis (PLS-DA).

#### Heatmaps

Heatmaps are a common visualisation tool used in metabolomics, where the different metabolites and samples are displayed as the rows and columns of a matrix table and the metabolite intensity is displayed with a colour (see Figure 1.13). This way, many metabolites can be displayed at the same time. Furthermore, clustering algorithms allow to group together similar metabolic signals across the different samples, which enables easier data analysis and interpretation. One of the main disadvantages of heatmaps is that they can quickly become overcrowded with the hundreds or thousands of metabolites commonly detected in metabolomics studies. Therefore, it is useful to do some further analysis and only represent relevant metabolites, based on specific criteria, such as statistical significance or bioprocess relevance.

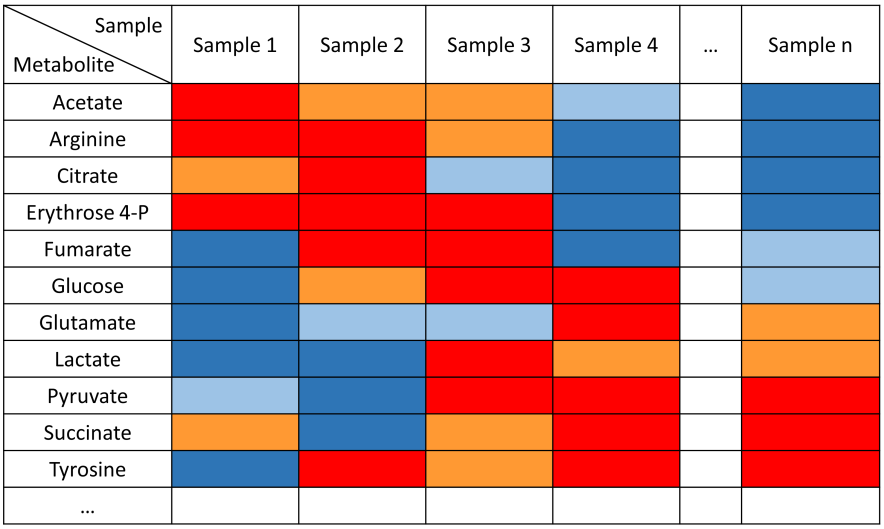



**Figure 1.13:** Example of a heatmap, where columns represent experimental samples and rows represent metabolites. Metabolite intensities range from high values (dark red) to low values (dark blue).

### Principal component analysis

Principal component analysis is an unsupervised mathematical technique that is used to transform a set of correlated variables (*m*) into a new set of uncorrelated variables called principal components (PCs). The first principal component (PC1) has the direction of highest variation of the original data. Mathematically, this is analogous to saying that the projection of the data points to this direction is minimum. The second principal component (PC2) describes a new direction orthogonal (*i.e.* uncorrelated) to PC1 that maximises the variation of the original data, and the same process is applied with the subsequent principal components. Orthogonality is used for convention and it makes it easier to work with and interpret the resulting PCA plot, where axes are independent from each other.

A simple graphical example with six measurements of two metabolites is represented in Figure 1.14. Each PC can be expressed as linear combinations of the *m* original variables, and each PC is sorted in decreasing order of variance of the data (Chatfield and Collins, 2018) so that PC1 describes the direction of highest variation of the original data, *etc.* In the simple example of Figure 1.14, using only two variables, if the slope of PC1 in the original coordinates is 0.25, the one-unit long vector of PC1 (called eigenvector or singular vector) would be PC1 = 0.97 x Succinate + 0.24 x Pyruvate, as illustrated in Figure 1.14B and D. In PCA nomenclature, the weighing factors of this linear combination (0.97 and 0.24 in the example) are called **loadings**, and the projection of the raw data samples into the new PC axes are called the **scores** of the samples. In the example, the scores of each sample on PC1 would be obtained by multiplying the sample succinate and pyruvate intensity values by the loadings of PC1 and adding them together.

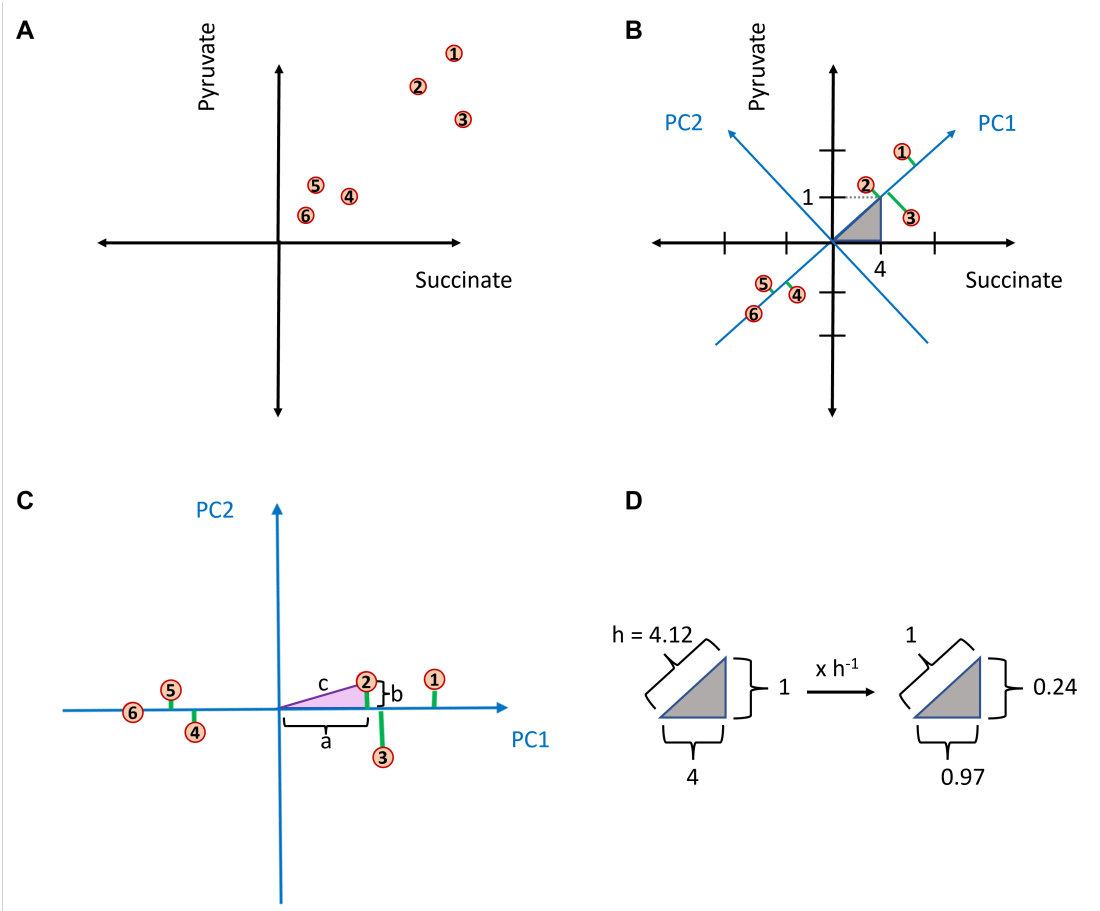



**Figure 1.14:** Graphical example of principal component analysis with six data points (in orange) and two metabolites in the original data (**A**) and mean-centered data (**B**). Principal components are represented in blue and the projection of the data points into PC1 are depicted in green. The PC plot (**C**) has the new axes defined by PC1 and PC2. The **score** of sample 2 on PC1 is represented by "a" and on PC2 it is represented by "b". Graphical example of the process of obtaining the **eigenvector of PC1**, which has a slope of 1/4 = 0.25 (**D**). The value of the hypotenuse (h) is calculated using the Pythagorean theorem. Dividing all sides of the triangle by h, a one-unit long vector is obtained. Then, the eigenvector of PC1 is: PC1 = 0.97 x Succinate + 0.24 x Pyruvate, and 0.97 and 0.24 are called weights or loadings. The grey triangle in subfigure **D** follows from subfigure **B**.

In matrix notation, PCA can be understood as the decomposition of the original data matrix **X** containing *n* samples and *m* variables into a loadings matrix **L** (transposed) and a scores matrix **S**. The mathematical development and final expressions are given in equations 1.3-1.5, where $\mathbf{L^{-1}}$ denotes the inverse of the loadings matrix, $\mathbf{L^{T}}$ denotes the transposed loading matrix and **E** is the errors matrix, which can be understood as the unexplained variance of the original data after reducing the dimensionality.

$$S_{n,m} = X_{n,m} \cdot L_{m,m} \tag{1.3}$$

$$\begin{aligned} S_{n,m} \cdot L^{-1}_{m,m} &= X_{n,m} \cdot L_{m,m} \cdot L^{-1}_{m,m} \\ &= X_{n,m} \end{aligned} \tag{1.4}$$

$$\begin{aligned} X_{n,m} &= S_{n,m} \cdot L^{T}_{m,m} \\ &= S_{n,k} \cdot L^{T}_{m,k} + E_{n,m} \end{aligned} \tag{1.5}$$

where *k*<*m*, according to the dimensionality reduction. That is, the *m* variables (or dimensions) of the original data has been reduced to the first *k* most meaningful principal components. Note that the transposed of the loadings matrix can be used instead of its inverse because **L** is an orthogonal matrix, *i.e.* each of the columns (or rows) are independent of the others. This results from the PCs being orthogonal to each other and using $\mathbf{L^{T}}$ instead of $\mathbf{L^{-1}}$ makes solving the task computationally much easier. An illustration of equation 1.3 is shown in Figure 1.15.

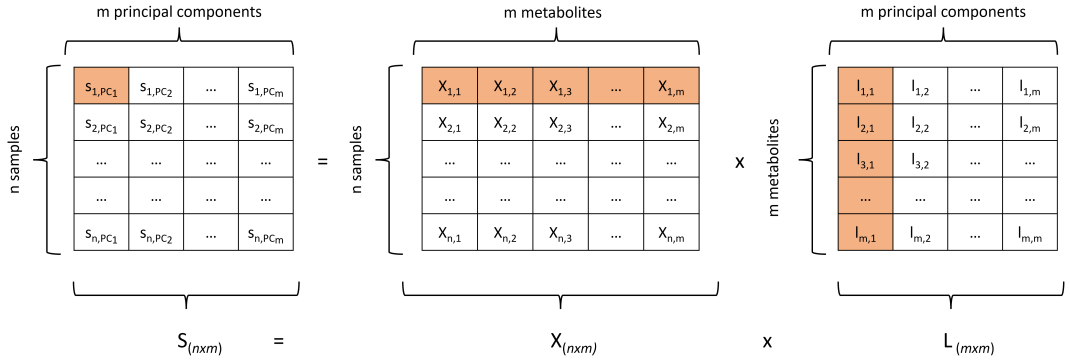



**Figure 1.15:** Illustration of the principle of PCA in matrix form before dimensionality reduction, where L, S and X are the loadings, scores and original data matrices. Note that, dimensionality reduction is achieved when the first "k" PCs are chosen, hence reducing the columns of S and L from "m" to "k". Highlighted in orange is the elements involved in obtaining the score for PC1 in the first sample.

If the first few PCs account for the majority of the variation of the original data, these can help identify which *m* variables contribute the most to the data variation. Furthermore, PCA can also help visualise which data samples cluster together, suggesting that these have more in common with each other than with samples from other clusters. Finally, it is worth mentioning that in cases when the number of variables is larger than the number of samples – as it is often the case for biological data – PCA can bring the dimensionality of the data as low as the number of samples without losing information (Ringnér, 2008).

### Partial least squares regression

Partial least squares regression is a supervised method that correlates a multivariate data matrix (*e.g.* metabolite intensities) to an observable y-variable (*e.g.* biomass concentration). In mathematical notation, this could be expressed following equation 1.6, where ***y*** is the *n*-length column vector containing the observable variable for each of the *n* samples, **X** is the *n*x*m* matrix containing metabolite intensities for *m* metabolites (columns) and *n* samples (rows), $\beta$ is the *m*-length column vector containing the regression coefficients for each metabolite, and $\varepsilon$ is the *n*-length column vector containing the residual error for each sample. A visual illustration of a linear regression problem with a metabolomics data matrix containing four samples trying to correlate the signal intensity of six different metabolites to the concentration of biomass for each sample is shown in Figure 1.16.

$$y_n = X_{n,m} \cdot \beta_m + \varepsilon_n \tag{1.6}$$

| Biomass [g/L] |
|---|
| 2.40 |
| 7.10 |
| 23.25 |
| 21.70 |

| Metabolite intensity / Sample | Arginine | Erythrose 4-P | Glucose | Glutamate | Pyruvate | Succinate |
|---|---|---|---|---|---|---|
| Sample 1 | 1.56E3 | 0.00 | 3.00E5 | 0.00 | 1.35E4 | 4.43E5 |
| Sample 2 | 5.80E3 | 0.00 | 2.52E5 | 2.00E4 | 3.24E4 | 3.13E5 |
| Sample 3 | 5.41E4 | 3.84E3 | 5.55E4 | 1.30E5 | 1.83E4 | 4.00E5 |
| Sample 4 | 2.00E4 | 1.14E5 | 2.32E5 | 2.12E4 | 2.32E6 | 1.14E7 |

| Coefficients |
|---|
| β_Arginine |
| β_Erythrose 4-P |
| β_Glucose |
| β_Glutamate |
| β_Pyruvate |
| β_Succinate |

| Error [g/L] |
|---|
| ε_1 |
| ε_2 |
| ε_3 |
| ε_4 |

$y_{(n)}$ = $X_{(nxm)}$ x $\beta_{(m)}$ + $\varepsilon_{(n)}$

**Figure 1.16:** Example of a regression problem using an $X_{(nxm)}$ data matrix containing four samples and six metabolites to try to predict the concentration of biomass y(n) for each sample.

Although the above example contained a ***y*** vector of one single observable variable, PLS models can be used to predict a matrix of multiple dependent **Y** variables at the same time. In the literature, this distinction is often made by respectively referring to these as PLS1 and PLS2 modelling (L. C. Lee *et al.*, 2018).

The aim of regression modelling is to find a set of $\beta$ coefficients that when multiplied by the predictor variables in **X** result in a good prediction of the response variables of interest ***y***, as described in equation 1.6. However, metabolomics data usually has a much larger number of metabolites than samples and many signals are correlated to each other (multicollinearity). As a consequence, classical multivariate linear regression is not suitable to obtain robust predictions and dimensional reduction methods such as principal component regression (PCR) and PLS are often used instead (L. C. Lee *et al.*, 2018; Pérez-Enciso and Tenenhaus, 2003). The idea of PCR is to use the first principal components of the original **X** predictor variables to predict ***y*** (or **Y**).

It was previously mentioned that PCA finds a new set of orthogonal variables that maximise the variance of the data (principal components). In the case of PLS, the **X** and **Y** matrices are decomposed into two respective "latent variables" similar to how PCA decomposes the **X** matrix (see equations 1.7 and 1.8), but in this case the model does not maximise the variation of **X** but instead it maximised the covariance between **X** and **Y** (Pérez-Enciso and Tenenhaus, 2003). By maximising the covariance, higher loadings are assigned to variables (or metabolites) that are highly correlated with the response variables (Miller and Miller, 2005).

To sum up, PLS generates latent variables using those variables in **X** that better predict **Y**, while reducing the dimensionality of the matrices, trying to achieve a robust regression even in ill-conditioned highly-colinear systems (Garthwaite, 1994).

$$X_{n,m} = T_{n,k} \cdot P^T_{m,k} + E_{n,m} \tag{1.7}$$

$$Y_{n,o} = U_{n,k} \cdot Q^T_{o,k} + F_{n,o} \tag{1.8}$$

where **T** and **U** are the scores matrices of **X** and **Y**, **P** and **Q** are the respective loading matrices and **E** and **F** the error matrices. The classical and more widespread PLS algorithm is called Non-linear Iterative Partial Least Squares (NIPALS) and it works iteratively across the different columns of **Y** by first calculating the weights (***w***) of **X** by using ordinary least squares regression on one single column of **Y** (***y***) (equation 1.9), and then calculate the scores of **X** (noted as ***t***) (equation 1.10). These ***t*** scores of **X** are then used to calculate the loadings of **X** (noted as ***p***) (equation 1.11) and the loadings of **Y** (noted as ***q***) (equation 1.12). Finally, the ***q*** loadings of **Y** are used to calculate the scores of **Y** (noted as ***u***) (equation 1.12). Now the process is repeated iteratively with the other columns of **Y** minimising the error of the covariance between **X** and **Y** as well as the error matrices **E** and **F**. Different ***w*** values are used until the minimisation criteria is met. After the first PLS component has been found, the process is repeated using the residuals of **X** and **Y** (equations 1.14 and 1.15) as the new input predictors (**X**) and response variables (**Y**) to find the second PLS component, and this is repeated successively to find the subsequent PLS components (L. C. Lee *et al.*, 2018).

$$y_n = X_{n,m} \cdot w_m \longrightarrow X_{n,m}^{-1} \cdot y_n = X_{n,m}^{-1} \cdot X_{n,m} \cdot w_m \longrightarrow w_m = X_{n,m}^{T} \cdot y_n \tag{1.9}$$

$$t_n = X_{n,m} \cdot w_m \tag{1.10}$$

$$t_n = X_{n,m} \cdot p_{1,m} \longrightarrow p_{1,m} = X_{n,m}^{T} \cdot t_n \tag{1.11}$$

$$t_n = Y_{n,o} \cdot q_{1,o} \longrightarrow q_{1,o} = Y_{n,o}^{T} \cdot t_n \tag{1.12}$$

$$u_n = Y_{n,o} \cdot q_{1,o} \tag{1.13}$$

$$res_X = X_{n,m} - t_n \cdot p_{1,m} \tag{1.14}$$

$$res_Y = Y_{n,o} - t_n \cdot q_{1,o} \tag{1.15}$$

For a more in depth description of the mathematics of PLS and the NIPALS algorithm, the reader is referred to Wold *et al.* (2001).

A common application of PLS in the field of metabolomics is as a classification tool. This technique is called partial least squares-discriminant analysis (PLS-DA) and consists of a PLS regression problem, where the observable **Y** variables are not numerical but categorical instead. For example, two categorical variables could be “successful fermentation run” and “unsuccessful fermentation run”. Mathematically, each categorical variable is assigned a value (usually -1 and 1, or 0 and 1) in order to compute the regression (L. C. Lee *et al.*, 2018; Szymańska *et al.*, 2012). Metabolites with high positive or negative loading values are the ones with a higher contribution to the prediction of the different data categories. Another important measure of metabolite importance in PLS-DA is called variable importance in projection (VIP). VIP adds the PLS-weight value (w) of a metabolite in each latent component multiplied by the relative y-variation explained in each latent component, as illustrated by equation 1.16 (Cocchi *et al.*, 2018; Wold *et al.*, 2001).

$$VIP_i^2 = \sum_{f=1}^{\phi} \frac{w_{if}^2 \cdot SSY_f \cdot m}{SSY_{total} \cdot \phi} \tag{1.16}$$

where $m$ is the number of metabolites in **X**, $i$ is the metabolite evaluated, $\phi$ is the number of latent variables, $w$ is the metabolite- and latent variable-specific PLS-weight value, and $SSY_f$ is the y-variance explained by each specific latent variable ($f$).

Adding up the sum squared VIP of each metabolite results in the number of metabolites (*m*). If all metabolites have the same contribution on the development of the PLS-DA model, they all will have a $VIP^2$ and VIP score of 1. For this reason, VIP values above 1 are often considered as significant in class discrimination (Chong and Jun, 2005). Nevertheless, there is some discrepancy in the literature regarding the appropriate value of this threshold and alternative values are sometimes used, such as the average VIP or a value of 0.67 (Cocchi *et al.*, 2018; Gorrochategui *et al.*, 2016).

*PLS model validity*

The most common criteria to evaluate the validity of PLS regression models are the parameters $R^2$ and $Q^2$ (also known as predictive or cross-validated $R^2$) (Triba *et al.*, 2015). On one hand, $R^2$ is the coefficient of determination, which measures the goodness-of-fit of the model, *i.e.* the proportion of variance of the dependent variable explained by the independent variable/s. On the other hand, $Q^2$ measures the predictability of the model, *i.e.* given a new data set, how accurately can the model predict the dependent variables based on the independent variable/s.

There are two typical ways of checking the prediction capacity of the model. The more robust one is splitting the data into a training set (used to build the model) and a test set (used to test the prediction capacity of the model on new data). Alternatively, if test data is not available, the whole data set can be used to build the model in a process known as cross-validation. A common cross-validation method is called k-fold cross-validation and it consists in splitting the data in k subsets and perform the model training with k-1 subsets and validation with the remaining subset. This is done for the k possible iterations and the prediction error is averaged, as schematically depicted in Figure 1.17. A specific case of k-fold cross-validation where the number of subsets equals the number of samples in the data (k=n) is called leave-one-out cross-validation (LOOCV). In this case, the regression model is iteratively trained with all the samples but one until all samples have been used (Nengsih *et al.*, 2019).

Maximising $Q^2$ is usually the criterion used to determine the optimum number of latent components of a PLS model. Unlike $R^2$, which will progressively increase towards a value of 1 with an increasing number of latent components, $Q^2$ will reach a maximum value, after which it will start decreasing. Although $Q^2$ value of 0.5 is generally considered acceptable, this highly depends on the nature of the specific data of study (Triba *et al.*, 2015). Equations 1.17 and 1.18 describe how to calculate $R^2$ and $Q^2$ (Szymańska *et al.*, 2012).

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1.17}$$

where the residual sum of squares (SSR) is the variance of the data unexplained by the model, the total sum of squares (SST) is the total variance of the data, $y_i$ and $\hat{y}_i$ are the measured and model-predicted dependent variables of the *i*-th sample and $\bar{y}$ is the mean value of the dependent variable across the *n* samples.

$$Q^2 = 1 - \frac{PRESS}{SST} = 1 - \frac{\sum_{i=1}^{n^{test}} (y_i^{test} - \hat{y}_i^{test})^2}{\sum_{i=1}^{n^{test}} (y_i^{test} - \bar{y}^{train})^2} \tag{1.18}$$

where the predictive residual sum of squares (PRESS) is the variance of the data unexplained by the model for the predicted data, $y_i^{test}$ and $\hat{y}_i^{test}$ are the measured and predicted dependent variable of the *i*-th sample in the test set and $\bar{y}^{\text{train}}$ is the mean value of the dependent variable across the samples in the training set. In the case of internal cross-validation (instead of external prediction), the "test" superscript refers to the internal cross-validation sets. Therefore, comparing $R^2$ and $Q^2$, the difference between SSR and PRESS is that SSR is calculated with the same data used to train the model, whereas PRESS is calculated from the test or cross-validation data.

In summary, VIP assesses, in every different latent variable, the importance of each metabolite by looking at the variation explained by each metabolite in proportion to the total variation of that specific latent variable, while $R^2$ and $Q^2$ look at the variation explained by the model (all metabolites included) compared to the total variation of the data. $R^2$ is calculated using the whole data, whereas $Q^2$ is calculated using test data or with cross-validation, which makes it a better indication of model robustness. A VIP value above 1 is often considered as significant for class discrimination and a $Q^2$ value of 0.5 is generally considered acceptable for PLS-DA, although this depends on the nature of the data.

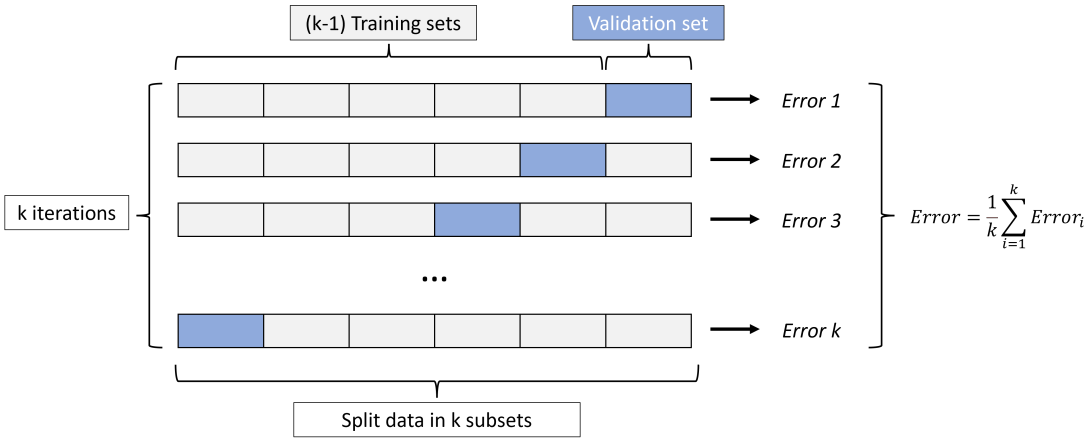



**Figure 1.17:** Schematic representation of the k-fold cross-validation process with the training (**grey**) and validation (**blue**) sets.

### 1.2.5 Pathway analysis

Apart from looking at individual metabolites using multivariate analysis, another way to understand biological data is to look at groups of metabolites that are biologically related and/or work together in metabolic pathways. Identifying whether different metabolic pathways are up- or down-regulated in different sample sets can be very useful to analyse the results. For example, a set of samples might show up-regulation of carbohydrate degradation and biosynthetic pathways, indicating good conditions for cellular growth, whereas another set of samples might show the up-regulation of secondary metabolism and catabolic pathways, indicating that the cells are no longer dividing, but instead are entering a phase of saturation or cell death. To perform this kind of analyses, it is necessary to look at a large number of metabolites in order to have a good coverage of the metabolic pathways. This is possible with metabolomics due to the vast detection capacity of mass spectrometers.

In the case of bioprocess optimisation, pathway analysis can help understand metabolic regulation, identify bottlenecks and point at specific targets for metabolic engineering. Two methods for analysing significant pathways in metabolomics experiments are enrichment analysis and pathway activity level scoring (PALS). Enrichment analysis algorithms – originally developed for genomics and transcriptomics data – is a method used to identify pathways with a large number of metabolites significantly different in different sample groups (Reimand *et al.*, 2019). PALS is a similar algorithm but it specifically rates small correlated changes across several compounds of a pathway as more significant than large changes in one or few compounds of a pathway (McLuskey *et al.*, 2021).

## 1.3 Hypothesis and aims

The hypothesis of this thesis is that monitoring the metabolites in the liquid phase of a fermentation processes in real-time can be used to improve the process.

The general approach to test this hypothesis will be to develop a method for monitoring metabolites present in the liquid phase of a bioreactor in real-time and explore different ways how this technology can be used to improve the process. Additionally, off-line LC-MS analysis will be used to complement and further exploit the real-time metabolomics monitoring data. In more detail, the main aims of the thesis are:

- To develop a platform for metabolomics monitoring of fermentation processes with direct sampling from the liquid phase and test it with an *E. coli* fermentation process of succinate production.

- To monitor the *E. coli* succinate fermentation process using metabolomics and analyse the data to find potential metabolites of interest. Typical compounds of interest for a bioprocess could be substrates, products, by-products, inhibitors or biomarkers of interest. This information can be used to develop genetic engineering strategies to improve the process (especially if inhibitors or by-products are found) or the monitoring itself (particularly if biomarkers are identified).
- To monitor the *E. coli* succinate fermentation process using metabolomics and develop mathematical models to correlate the metabolite mass spectrometry signal with metabolite concentration. These models have great potential in improving the performance of product manufacturing in the biotechnology industry.
- To implement a kinetic model describing the fermentation process and employ it in combination with on-line monitoring to forecast the evolution of different variables of interest in real-time.
- To perform an LC-MS analysis of the *E. coli* succinate fermentation process to complement the metabolomics monitoring data and obtain confident identification of metabolites of potential interest. Similarly, this data can be used to try to identify bioprocess biomarkers and potential targets for strain engineering.

# Chapter 2

# Materials and methods

This chapter describes the general materials and methods used throughout the thesis. Any experiment or analysis deviating from what is described in this chapter is specifically mentioned in the Materials and methods section of the respective chapter.

## 2.1 Reagents

All reagents used in the preparation of fermentation media and reference standards for targeted LC-MS and HPLC analysis were purchased from Sigma-Aldrich (Merck Life Science UK Limited, Gillingham, Dorset, United Kingdom) with the following exceptions. Acetic acid, glycine and all LC-MS grade reagents used to prepare mobile phase solutions were purchased from Fisher Scientific (Loughborough, Leicestershire, United Kingdom) except formic acid was, which was purchased from Merck KGaA (Darmstadt, Germany). Cis-aconitate, L-proline, L-threonine, L-tyrosine and L-valine were purchased from Alfa Aesar (Thermo Fisher Scientific, Heysham, United Kingdom). L-alanine and pyruvate were purchased from Acros organics (Thermo Fisher Scientific, Geel, Belgium). L-lysine, ribose-5P and xylulose-5P were purchased from Honeywell Fluka (Seelze, Germany). Fumarate was purchased from Tokyo Chemical Industry UK Ltd. (Oxford, United Kingdom). Oxythiamine was purchased from Bachem (Bubendorf, Switzerland). Polypropylene glycol 2000 was purchased from VWR Chemicals (Avantor, Fontenay-sous-Bois, France). Ammonium hydroxyde 28 % (v/v) and sulfuric acid 96 % (v/v) for pH control were purchased from Alfa Aesar (Thermo Fisher Scientific, Heysham, United Kingdom) and Chem-Lab NV (Zedelgem, Belgium), respectively.

## 2.2 Bacterial strain

All experiments described in the thesis were carried out using a proprietary industrial *E. coli* BW25113 strain [Δ*pflB::kanR,* Δ*ldhA::cmlR*] (Ingenza Ltd., UK), based on the NZN111 *E. coli* strain with the same two deletions described by Chatterjee *et al.* (2001). The strain was called IGZ006 and cannot grow anaerobically on glucose. The knock-out of the pyruvate-formate lyase (*pflB*) and lactate dehydrogenase (*ldhA*l) genes was done by the chromosomal insertion in their respective loci of the kanamycin and chloramphenicol resistance genes with the aim of reducing the formation of two of the major by-products in anaerobic fermentation in *E. coli* – formate and lactate (Singh *et al.*, 2011) – and increasing the production of succinate.

## 2.3 Growth media

All fermentation experiments described in the thesis were carried out with a batch phase for biomass formation using a defined minimal medium containing 11.90 g/L glucose as the sole carbon source, 2.00 mM $MgSO_4$, a mix of salts solution (2.00 g/L $(NH_4)_2SO_4$, 14.60 g/L $K_2HPO_4$, 3.60 g/L $NaH_2PO_4 \cdot 2H_2O$, 0.50 g/L $(NH_4)_2$H-citrate), a mix of trace elements (1.0 mg/L $CaCl_2 \cdot 2H_2O$, 20.06 mg/L $FeCl_3$, 0.36 mg/L $ZnSO_4 \cdot 7H_2O$, 0.32 mg/L $CuSO_4 \cdot 5H_2O$, 0.30 mg/L $MnSO_4 \cdot H_2O$, 0.36 mg/L $CoCl_2 \cdot 6H_2O$, 44.60 mg/L $Na_2EDTA \cdot 2H_2O$), antibiotics (100 mg/L kanamycin, 34 mg/L chloramphenicol) and antifoam (33.33 $\mu$L/L polypropylene glycol (PPG) P-2000). Shake flask overnight cultures were prepared using the same medium but with 10.00 g/L glucose and no antifoam.

## 2.4 Fermentation process conditions

All fermentation experiments were carried out in a 5 L Applikon stirred tank fermenter (ADI 1030 Bio Controller, 1035 Bio Console), and the process consisted of an initial aerobic batch phase where the minimal medium was primarily used for biomass formation, followed by a 24 h anaerobic succinate production phase (Figure 2.1), similar to the process described by Vemuri *et al.* (2002).

### 2.4.1 Inoculum

Fermentation inocula were prepared by inoculating 50 $\mu$L of cell bank into 100 mL of growth medium in a 500 mL baffled shake flask and incubated at 37 ℃ and 165 rpm for 17 – 17.5 h.

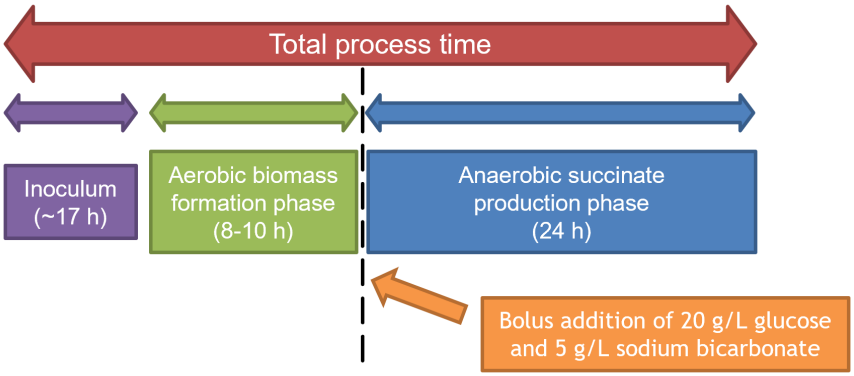



**Figure 2.1:** Schematic diagram of the fermentation process. The dashed black line splits both phases of the process.

### 2.4.2 Aerobic batch phase for biomass growth

The fermentation was started by inoculating 100 mL of overnight culture into 3 L of growth medium in the 5 L fermenter for a starting $OD_{600}$ of 0.21 +/- 0.025. During biomass growth the conditions were maintained at 37 ℃ temperature, 500-900 rpm agitation (controlled to keep the dissolved oxygen (DO) > 30 %), 4.00 L/min air (1.33 vvm) and pH 7.0 ± 0.1, controlled with 2.00 M $H_2SO_4$ and 28 % w/v $NH_4OH$.

### 2.4.3 Anaerobic succinate production phase

At the beginning of the production phase, glucose from a 500 g/L solution and sodium bicarbonate from a 100 g/L solution were added to the fermenter as a single bolus addition to a final concentration of 20 g/L and 5 g/L respectively in the vessel, as described by H. Wu *et al.* (2007). The sodium bicarbonate provides soluble $CO_2$, which is required for the conversion of PEP to oxaloacetate (Figure 1.2) (Thakker *et al.*, 2012). Once the glucose and sodium bicarbonate were added to the fermenter, the sparged air was replaced by pure (99.8 %) $CO_2$ at 0.50 L/min (0.17 vvm), agitation was set to 300 rpm, temperature at 37 ℃ and pH at 7.0 ± 0.1 controlled with 2.00 M $H_2SO_4$ and 28 % w/v $NH_4OH$.

## 2.5 Biomass measurement

Biomass levels were reported as optical density at 600 nm ($OD_{600}$), wet cell weight (WCW) and dry cell weight (DCW). The WCW was determined by spinning down 1 mL of sample for 5 min at 14,000 x g twice in a pre-weighed microtube, removing the supernatant and weighing the resulting pellet. The DCW was determined by drying the WCW pellet in an oven at 100 ℃ for 48 h. In both cases, the weight of the pellet in g/L was calculated from gravimetric difference (microtube with pellet - empty microtube).

## 2.6 On-line metabolomics method

This section details the on-line metabolomics method for both untargeted on-line metabolomics using the Exactive™ Orbitrap, and for targeted on-line metabolomics using the TSQ Quantiva triple quadrupole.

On-line metabolomics was conducted by connecting the fermenter to either an Exactive™ Orbitrap (Thermo Scientific™) mass spectrometer or a TSQ Quantiva (Thermo Scientific™) triple quadrupole mass spectrometer with a fluidics system similar to what had previously been described in the literature (Link *et al.*, 2015), but adapted to inject undiluted fermentation broth samples straight into the mass spectrometer. The modified fluidics system consisted of a peristaltic pump and two valves (Figure 2.2), and sample injections were alternated with washing steps, one-to-one. The peristaltic pump was a Masterflex™ L/S ® (Cole-Parmer) high-performance pump model 77252-72 and was operated at a high flow rate of 75-100 mL/min. The first valve was a six-port, two-position valve (Vici Valco®) and the second one was a ten-port, two-position valve (Dionex Corporation). Note, however, that no chromatography was used.

### 2.6.1 Sample injections

During sample injections, fermentation broth is continuously extracted from the fermenter with the peristaltic pump, injected into the six-port valve, circulated through a 50 $\mu$L loop and returned to the fermenter. Upon valve switching, the broth sample from the 50 $\mu$L loop of the six-port valve is carried by sterile water pumped at a 200 $\mu$L/min flow rate using an external piston pump into the ten-port valve, where it is collected in a 1 $\mu$L loop. The sample is finally injected into the mass spectrometer carried by a 70:30 ACN:IPA + 0.1 % FA mixture at a 400 $\mu$L/min flow rate. The duration of the injection method is 1 minute, and the total travelling time from the bioreactor to the mass spectrometer 30 seconds, with *ca.* 10 seconds to reach the six-port valve and 20 more seconds to reach the mass spectrometer.

### 2.6.2 Washing steps

Each sample injection was followed by a 4 min washing step to avoid system blockage and signal loss. The six-port valve was washed with 70:30 IPA:ACN + 0.1 % FA for 1 min and with sterile water for 3 min, both at a 400 $\mu$L/min flow rate. The ten-port valve was washed with sterile water for 2 minutes and then 70:30 ACN:IPA + 0.1 % FA for 2 min, both at a 600 $\mu$L/min flow rate (Figure 2.3). The washing solutions were sent to waste and did not enter either the fermenter or the mass spectrometer.

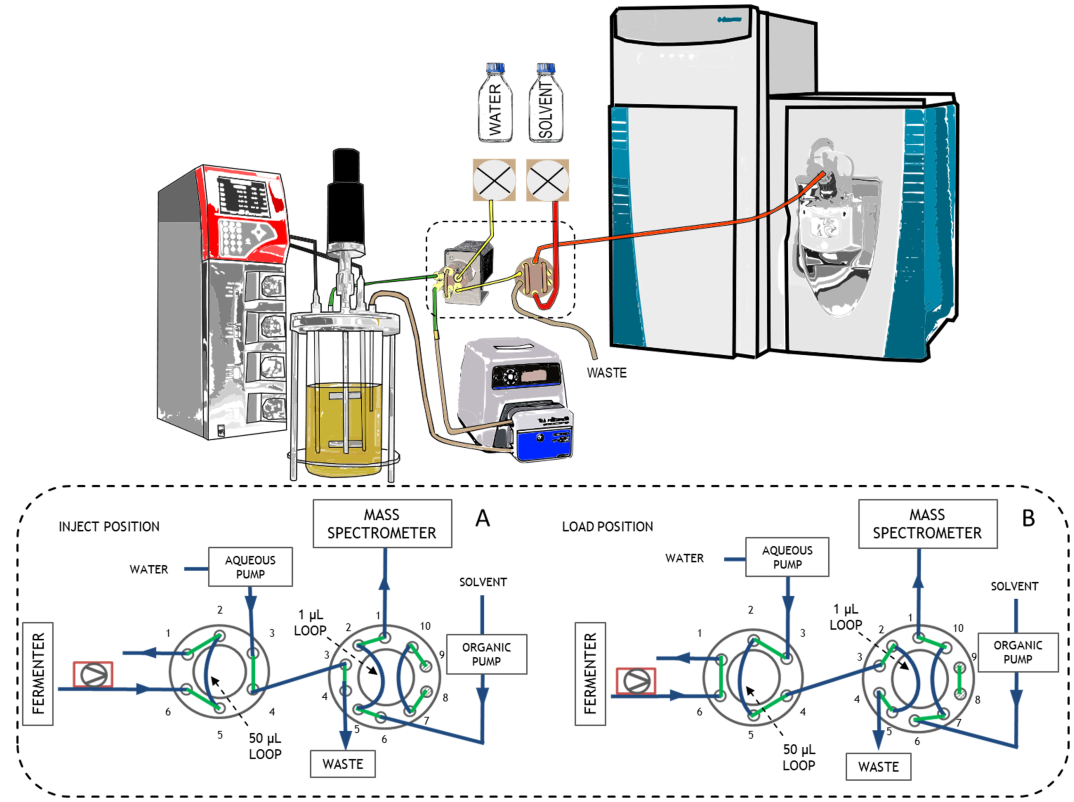



**Figure 2.2:** Experimental setup of the on-line metabolomics system (**top**) and a detailed diagram of the dashed area showing the two valve positions (**bottom**). During the **"inject position"**, fermentation broth sample is continuously circulated through the 50 $\mu$L of the left 6-port valve (sample n+1) and the fermentation broth sample from the 1 $\mu$L loop in the right 10-port valve (sample n) is pushed to the mass spectrometer. During the **"load position"**, the sample n+1 is pushed from the 50 $\mu$L loop of the 6-port valve into the 1 $\mu$L loop of the 10-port valve, ready for injection at the next "**inject position**".

### 2.6.3 On-line mass spectrometer parameters - Orbitrap Exactive

Gas-phase ions were generated using an Ion Max source (Thermo Scientific$^{TM}$) with an ESI probe. The mass spectrometer was operated at 50,000 resolution, mass range 50 – 1000 m/z in polarity switching mode with a spray voltage of ±3.5 kV. Capillary temperature was set to 350 ℃, sheath gas flow rate 40 a.u., automatic gain control (AGC) target 1x10$^6$ a.u. and the lock masses in positive and negative mode were 74.0964 m/z and 112.9856 m/z, respectively.

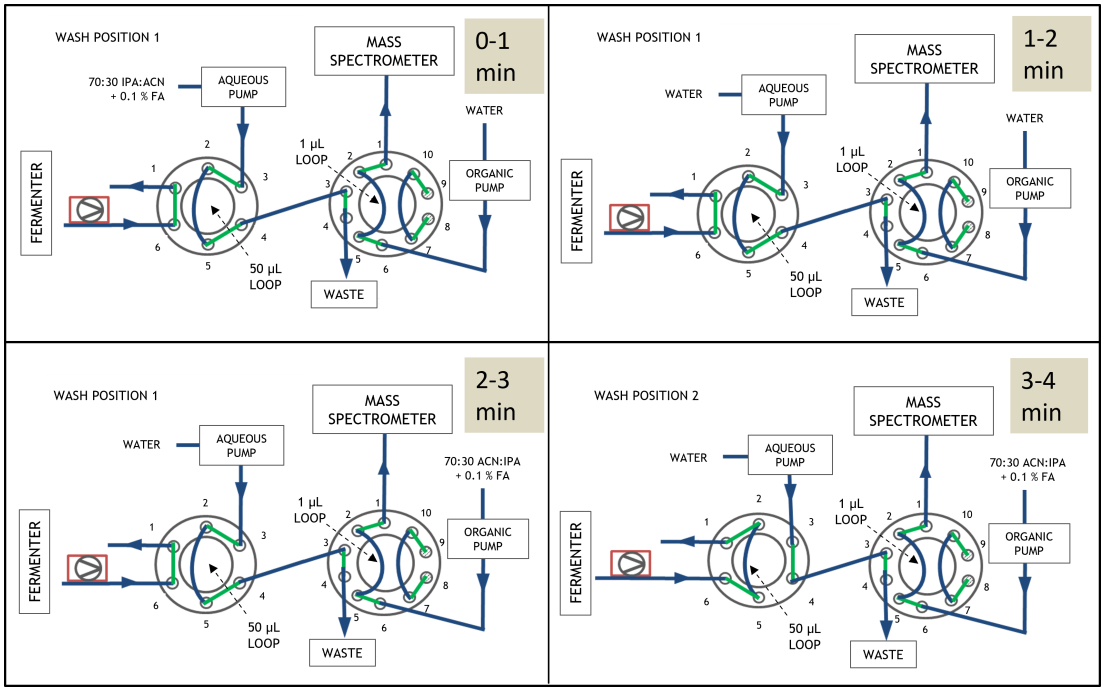



**Figure 2.3:** Diagram showing the positions of the two valves and liquid solutions used during the four-minute washing step between sample injections. The duration of each configuration is indicated on the top right of each diagram. The nomenclature of the pumps (aqueous and organic pump) is based on the solutions used during the injection step.

### 2.6.4 On-line mass spectrometer parameters - Quantiva triple quadrupole

Gas-phase ions were generated using an Ion Max NG source (Thermo Scientific$^{TM}$) with a heated electrospray ionisation (HESI) II probe. The mass spectrometer was operated using an MRM method in polarity switching mode (see Section 2.8) with a spray voltage of ±3.5 kV. Capillary temperature was set to 25 ℃, ion transfer tube (sheath gas) temperature 325 ℃, sheath gas flow rate 50 a.u., auxiliary gas 10 a.u. and sweep gas 1 a.u. Resolution for both Q1 and Q3 was set at a FWHM of 0.7 m/z, cycle time was 0.8 s and CID fragmentation in Q2 was done under 1.5 mTorr pressure.

### 2.6.5 On-line metabolomics data processing and analysis

Real-time metabolomics raw data was processed with Xcalibur$^{TM}$ software (Version 3.1.66.10 for the Orbitrap Exactive and Version 4.2.28.14 for the Quantiva triple quadrupole) using two different peak detection methods depending on the metabolite. Method selection was based on peak integration performance, which was assessed by visual inspection. The first method was the Genesis peak detection method, using the following parameters: the peak integration threshold was set to 0.5 signal to noise ratio (S/N) and peak detection was set to the highest peak within a 15 s retention window, with a minimum peak height threshold of 3 S/N. The second method was the ICIS method with the following parameters: one smoothing point,

a baseline window of 40, area noise factor of five, a peak noise factor of 10 and minimum peak height threshold of 3 S/N. All the Quantiva triple quadrupole data was processed using the ICIS peak detection method. After processing the raw data with the Xcalibur$^{TM}$ software, metabolite features were extracted as a .csv file which was further analysed using the ggplot2 package (Version 3.3.3; Wickham, 2016) in the statistical software environment R (Version 3.6.1).

## 2.7 Off-line LC-MS analysis

This section details the off-line metabolomics method for both the Q Exactive$^{TM}$ Orbitrap, and the TSQ Quantiva triple quadrupole.

### 2.7.1 Sample extraction for LC-MS analysis

At every time point, a 1 mL fermentation sample was taken into a pre-chilled microtube for off-line LC-MS analysis and quenched with 3 $\mu$L of 5 M $H_2SO_4$. The samples were then spun down at 4 ℃ and 13,000 x g for 10 min immediately after being removed from the bioreactor. The supernatant and cell pellet were collected as extracellular and intracellular fractions respectively and stored at -80 ℃ until further extraction for LC-MS analysis.

#### Extracellular fractions

Extracellular fraction extractions were prepared by diluting 10 $\mu$L of sample into 390 $\mu$L of 1:3:1 chloroform:methanol:water (C:M:W). The samples were then vortexed for 5 seconds and mixed in a chilled microtube mixer or a rotary shaker for 5-min at 4 ℃ and then centrifuged for 3 min at 13,000 x g and 4 ℃. At this point, 360 $\mu$L of supernatant were transferred into a new microtube and stored at -80 ℃ until LC-MS analysis. During handling, the 1:3:1 C:M:W extraction solvent and the samples were kept on ice.

#### Intracellular fractions

Prior to extraction, intracellular fractions were washed by resuspending the cell pellets in 1 mL of sterile phosphate buffer solution (PBS). The PBS was removed by spinning down the samples twice for 10 min at 13,000 x g and 4 ℃. For metabolite extraction, 200 $\mu$L of 1:3:1 C:M:W were added for every 5 mg of WCW pellet. Cell pellets were resuspended by pipetting, and then the samples were mixed in a chilled microtube mixer or a rotary shaker for 1 h at 4 ℃. Then, the samples were centrifuged for 3 min at 13,000 x g and 4 ℃. At this point, 200 $\mu$L of supernatant were transferred into a new microtube and stored at -80 ℃ until LC-MS analysis. During handling, the 1:3:1 C:M:W extraction solvent and the samples were kept on ice.

#### Pooled samples

Pooled samples for quality control for LC-MS analyses were prepared by combining 25 $\mu$L of each extracted sample to be analysed into one single vial.

### 2.7.2 Off-line LC-MS analysis - Orbitrap Q Exactive

Metabolite separation was performed with a Thermo Scientific$^{TM}$ UltiMate$^{TM}$ 3000 RSLCnano UHPLC system using a polymeric zwitterionic hydrophilic interaction liquid chromatography (ZIC®-pHILIC) column (Merck SeQuant®) (150 mm x 4.6 mm, 5 $\mu$m particle size) equipped with the corresponding guard column (20 mm x 2.1 mm, 5 um particle size) (Merck SeQuant®). A linear gradient was applied to the column, running from 80 to 20 % solvent B over 15 min, followed by a 2 min wash with 5 % solvent B, and 9 min re-equilibration with 80 % solvent B, where solvent B was acetonitrile and solvent A (the remaining percentage) was 20 mM ammonium carbonate in water (pH 9.16). The total flow rate was 300 $\mu$L/min, column temperature was maintained at 25 °C, sample injection volume was 10 $\mu$L, samples were maintained at 4 °C for the duration of the analysis and a HESI probe was used on the ion source.

Metabolite detection was done in a high-resolution Thermo Scientific$^{TM}$ Q Exactive$^{TM}$ Orbitrap mass spectrometer at 70,000 resolution, mass range 70 – 1050 m/z in polarity switching mode with a spray voltage of ±3.8 kV. Capillary temperature was set to 320 °C, sheath gas 40 a.u., AGC target 1x10$^6$ a.u. and the lock masses in positive and negative mode were 144.9822 m/z and 100.9856 m/z, respectively.

#### Off-line LC-MS data processing - Orbitrap Q Exactive

Raw mass spectrometry files were converted into .mzXML files in profile mode with the open-source software ProteoWizard (Version 3.0; Chambers *et al.*, 2012). Further data processing and analysis was performed using the Polyomics integrated Metabolomics Pipeline (PiMP) online platform (Gloaguen *et al.*, 2017; date of use: 08 Jul 2020). Peak detection and filtering were set to 3 ppm of the theoretical monoisotopic mass, minimum intensity to 5000, noise to 0.8, retention window to 0.05 and minimum number of detections to 3. Peak retention time was corrected using the Obiwarp algorithm (Prince and Marcotte, 2006; from xcms package Version 1.48.0). Metabolite features were extracted from PiMP as a .csv file and bar charts were created using the ggplot2 package (Version 3.3.3; Wickham, 2016) in the statistical software environment R (Version 3.6.1).

Fragmentation was performed on pooled samples by isolating ions in a 1.2 m/z window and fragmentation with stepped HCD collision energy of 24.8, 60.0 and 94.8 % for both polarities with 17,500 resolution and AGC target 1x10$^5$ a.u. Top 10 ions (intensity threshold 1.3x10$^5$) were selected for fragmentation and then added to a dynamic exclusion window for 15 s.

### 2.7.3 Off-line LC-MS analysis - Quantiva triple quadrupole

The LC-MS method was the same as the one used for the Orbitrap Q Exactive (see Section 2.7.2) but using a Thermo Scientific™ UltiMate™ 3000 UHPLC system, not using a guard column, and keeping the column temperature at 40 ℃.

Metabolite detection was done in a Thermo Scientific™ TSQ Quantiva triple quadrupole mass spectrometer using an MRM method in polarity switching mode (see Section 2.8) with a spray voltage of ±3.5 kV. Capillary temperature was set to 350 ℃, ion transfer tube temperature set to 325 ℃, sheath gas 50 a.u., auxiliary gas 10 a.u. and sweep gas 1 a.u. Resolution was set at a FWHM of 0.7 and 1.2 m/z respectively for Q1 and Q3, cycle time was 0.8 s and CID fragmentation in Q2 was done under 1.5 mTorr pressure.

#### Off-line LC-MS data processing - Quantiva triple quadrupole

Off-line metabolomics data generated with the Quantiva triple quadrupole was processed in the same way as the on-line metabolomics data generated with the same instrument (see Section 2.6.5).

## 2.8 Multiple reaction monitoring

The parameters for the multiple reaction monitoring (MRM) method used with the triple quadrupole mass spectrometer are shown in Table 2.1. Most metabolites have two product ions, but in some instances only a good product ion could be found. Acetate and the three compounds identified as possible biomass biomarkers (R-2,3-Dihydroxy-3-methylpentanoate, R-2,3-Dihydroxy-isovalerate and S-2-Aceto-2-hydroxybutanoate) were not fragmented (collision energy = 0 V), the former because no good product ion could be identified, and the three latter due to the lack of access to their reference standards.

**Table 2.1:** Multiple reaction monitoring parameters. The retention time was only applicable for LC-MS analysis; for on-line metabolomics, retention time was set at 0.4 $\pm$ 0.4 min for all compounds.

| Compound | Polarity | Precursor [m/z] | Product [m/z] | Collision energy [V] | Retention time [min] |
|---|---|---|---|---|---|
| 3-Phospho-D-glycerate | Negative | 184.99 | 123.11<br>166.92 | 17.01<br>9.04 | 12.7 $\pm$ 2.0 |
| 6-Phosphogluconic acid | Negative | 275.02 | 195.14<br>256.90 | 21.68<br>12.83 | 13.0 $\pm$ 13.0 |
| Acetate | Negative | 59.01 | 59.01 | 0 | 7.9 $\pm$ 2.0 |
| cis-Aconitate | Negative | 173.00 | 84.96<br>129.01 | 12.75<br>5.25 | 13.0 $\pm$ 2.0 |

Continued on next page

**Table 2.1 – continued from previous page**

| Compound | Polarity | Precursor [m/z] | Product [m/z] | Collision energy [V] | Retention time [min] |
|---|---|---|---|---|---|
| Citrate | Negative | 191.02 | 86.85 | 17.89 | 13.2 ± 2.0 |
| Erythrose-4P | Negative | 199.00 | 139.13 | 12.54 | 11.6 ± 2.0 |
| Fructose-6P | Negative | 259.02 | 138.93<br>168.71 | 16.50<br>8.71 | 11.7 ± 2.0 |
| Fumarate | Negative | 115.00 | 69.05<br>71.03 | 13.34<br>8.16 | 12.0 ± 2.0 |
| Glucose | Negative | 179.06 | 59.14<br>73.03<br>135.11 | 19.62<br>5.25<br>11.91 | 11.3 ± 2.0 |
| Glyceraldehyde-3P | Negative | 169.05 | 151.01 | 5.25 | 11.4 ± 2.0 |
| Glycine | Positive | 76.04 | 30.19<br>44.01 | 12.96<br>11.23 | 11.7 ± 2.0 |
| Isocitrate | Negative | 191.02 | 116.96<br>172.93 | 14.81<br>7.86 | 13.5 ± 2.0 |
| L-Alanine | Positive | 90.06 | 44.08<br>62.08 | 12.71<br>5.25 | 11.1 ± 2.0 |
| L-Arginine | Positive | 175.12 | 70.07<br>158.05 | 23.70<br>12.71 | 18.1 ± 2.0 |
| L-Asparagine | Positive | 133.06 | 73.86<br>87.07 | 17.47<br>11.07 | 11.3 ± 2.0 |
| L-Aspartate | Negative | 132.03 | 88.14<br>115.07 | 11.91<br>9.80 | 11.1 ± 2.0 |
| L-Glutamate | Negative | 146.05 | 73.91<br>128.14 | 19.74<br>9.76 | 14.0 ± 4.0 |
| L-Glutamine | Positive | 147.08 | 84.00<br>101.21<br>129.93 | 17.55<br>11.91<br>9.63 | 14.0 ± 4.0 |
| L-Histidine | Positive | 156.08 | 93.00<br>110.05 | 23.62<br>15.19 | 10.9 ± 2.0 |
| L-Isoleucine | Positive | 132.10 | 57.07<br>69.00 | 30.44<br>17.76 | 8.9 ± 2.0 |
| L-Lactate | Negative | 89.02 | 45.13<br>71.01 | 12.20<br>9.21 | 7.7 ± 2.0 |
| L-Leucine | Positive | 132.10 | 30.21<br>43.30 | 18.23<br>25.3 | 8.5 ± 2.0 |
| L-Lysine | Positive | 147.11 | 84.07<br>130.07 | 17.89<br>10.43 | 14.0 ± 4.0 |
| L-Methionine | Positive | 150.06 | 56.05<br>133.00 | 18.06<br>10.05 | 9.0 ± 2.0 |
| L-Phenylalanine | Positive | 166.09 | 77.00<br>120.05 | 39.84<br>14.06 | 8.2 ± 2.0 |
| L-Proline | Positive | 116.07 | 68.00<br>70.07 | 28.38<br>16.12 | 9.8 ± 2.0 |
| L-Serine | Positive | 106.05 | 60.00<br>88.04 | 12.08<br>9.38 | 11.7 ± 2.0 |
| L-Threonine | Positive | 120.07 | 74.00<br>103.00 | 11.40<br>18.44 | 9.0 ± 3.0 |
| L-Tryptophan | Positive | 205.10 | 146.07<br>188.07 | 18.18<br>10.35 | 9.4 ± 2.0 |

Continued on next page

**Table 2.1 – continued from previous page**

| Compound | Polarity | Precursor [m/z] | Product [m/z] | Collision energy [V] | Retention time [min] |
|---|---|---|---|---|---|
| L-Tyrosine | Positive | 182.08 | 136.10<br>165.07 | 13.93<br>10.27 | 10.2 ± 2.0 |
| L-Valine | Positive | 118.09 | 57.05<br>72.00 | 29.94<br>11.49 | 9.6 ± 2.0 |
| Pyruvate | Negative | 87.01 | 43.08<br>68.93 | 8.79<br>5.25 | 7.0 ± 2.0 |
| R-2,3-Dihydroxy-3--methylpentanoate | Negative | 147.07 | 147.07 | 0 | 13.0 ± 13.0 |
| R-2,3-Dihydroxy--isovalerate | Negative | 133.05 | 133.05 | 0 | 13.0 ± 13.0 |
| Ribose-5P | Negative | 229.01 | 138.97<br>169.04 | 12.58<br>11.57 | 11.6 ± 2.0 |
| Ribulose-5P | Negative | 229.01 | 139.01 | 16.67 | 12.5 ± 3.0 |
| S-2-Aceto-2--hydroxybutanoate | Negative | 145.05 | 145.05 | 0 | 13.0 ± 13.0 |
| S-Malate | Negative | 133.01 | 70.99<br>114.93 | 13.34<br>11.15 | 11.9 ± 2.0 |
| Sedoheptulose-7P | Negative | 289.03 | 198.93<br>228.89 | 11.19<br>11.32 | 11.8 ± 2.0 |
| Succinate | Negative | 117.02 | 73.11<br>98.83 | 11.74<br>7.86 | 11.3 ± 2.0 |
| Xylulose-5P | Negative | 229.01 | 139.09<br>169.11 | 13.64<br>10.86 | 12.5 ± 3.0 |

## 2.9 Off-line HPLC-UV/Vis-RI analysis

HPLC coupled to UV/Vis and refractive index detectors (HPLC-UV/Vis-RI) analysis was carried out using a Rezex$^{TM}$ ROA Organic Acid H+ ion-exclusion column (Phenomenex®) (300 mm x 7.8 mm) equipped with a Carbo-H4 guard column (SecurityGuardTM) (3.0 mm i.d.). An isocratic method was applied to the column, running a 5 mM $H_2SO_4$ mobile phase solution for 30 min. The total flow rate was 800 $\mu$L/min, column temperature was maintained at 65 ℃, sample injection volume was 10 $\mu$L and samples were maintained at 4 ℃ for the duration of the analysis. The HPLC-UV/Vis-RI data was extracted as a .csv file and was further analysed using the ggplot2 package (Version 3.3.3; Wickham, 2016) in the statistical software environment R (Version 3.6.1).

## 2.10 Limit of detection calculation

The limit of detection (LOD) for analytes was calculated following equation 2.1 (Irving *et al.*, 1978; Miller and Miller, 2005)

$$LOD = 3 \cdot \frac{SD_{intercept}}{m_{cc}} \tag{2.1}$$

where $SD_{intercept}$ and $m_{cc}$ are the standard deviation of the intercept and slope of the calibration curve, respectively. Calibration curves were built in the statistical software environment R (Version 4.1.2) and LODs were calculated using the chemical package (Version 0.2.2; Ranke, 2021).

## 2.11 Data smoothing

Data smoothing was carried out in the statistical software environment R using the ggplot2 package (Version 3.3.3; Wickham, 2016) with a locally estimated scatterplot smoothing (LOESS) method with a span between 0.15 and 0.5, depending on the metabolite.

## 2.12 Data normalisation

Data normalisation was done by applying equation 2.2, where $\boldsymbol{y}$ is any given time-course vector, such as the mass spectrometry intensity of a metabolite throughout the duration of the fermentation. After normalising, the normalised data will range from 0 to 1.

$$y_{norm} = \frac{y - min(y)}{max(y) - min(y)} \tag{2.2}$$

## 2.13 Model diagnostics

Different statistical tools are used throughout the thesis to assess the quality of mathematical models. These are described below.

### 2.13.1 Coefficient of determination ($R^2$)

The coefficient of determination ($R^2$) is calculated with equation 2.3

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.3}$$

where $y_i$ and $\hat{y}_i$ are the measured and model-predicted dependent variables of the *i*-th sample and $\bar{y}$ is the mean value of the dependent variable across the *n* samples.

The predictive (or cross-validated) coefficient of determination ($Q^2$) is calculated with equation 2.4

$$Q^2 = 1 - \frac{\sum_{i=1}^{n^{test}} (y_i^{test} - \hat{y}_i^{test})^2}{\sum_{i=1}^{n^{test}} (y_i^{test} - \bar{y}^{train})^2} \tag{2.4}$$

where $y_i^{test}$ and $\hat{y}_i^{test}$ are the measured and predicted dependent variable of the *i*-th sample in the test set and $\bar{y}^{\text{train}}$ is the mean value of the dependent variable across the samples in the training set (Schüürmann *et al.*, 2008). In the case of internal cross-validation (instead of external prediction), the "test" superscript refers to the internal cross-validation sets.

### 2.13.2 Root mean squared error

The root mean squared error (RMSE) of regression models were calculated following equation 2.5

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(c_i - \hat{c}_i)^2} \tag{2.5}$$

where $c_i$ and $\hat{c}_i$ are the measured and predicted metabolite concentration of the *i*-th sample and *n* is the total number of samples.

Similarly, root mean squared error of cross-validation (RMSECV) and root mean squared error of prediction (RMSEP) were calculated following equations 2.6 and 2.7

$$RMSECV = \sqrt{\frac{1}{n^{CV}}\sum_{i=1}^{n^{CV}}(c_i - \hat{c}_i)^2} \tag{2.6}$$

$$RMSEP = \sqrt{\frac{1}{n^{test}}\sum_{i=1}^{n^{test}}(c_i - \hat{c}_i)^2} \tag{2.7}$$

where $n^{CV}$ and $n^{test}$ are the samples in the cross-validation and test sets used to develop and test the model.

It is also possible to get an estimate of the magnitude of the RMSE as a percentage relative to the variable measured following equations 2.8, 2.9 and 2.10

$$RMSRE = 100 \cdot \frac{RMSE}{\bar{c}} \tag{2.8}$$

$$RMSRECV = 100 \cdot \frac{RMSECV}{\bar{c}^{CV}} \tag{2.9}$$

$$RMSREP = 100 \cdot \frac{RMSEP}{\bar{c}^{test}} \tag{2.10}$$

where $\bar{c}$, $\bar{c}^{\text{CV}}$ and $\bar{c}^{\text{test}}$ are the average concentration of the metabolite in the whole data set, cross-validation set and test set, respectively.

# PART II

# Untargeted on-line metabolomics analysis of an *Escherichia coli* succinate fermentation process and subsequent characterisation by LC-MS

Chapter 3

# Fermentation monitoring with on-line metabolomics

Part of the experimental work from this chapter was carried out in collaboration with Dr. Jennifer Haggarty at the University of Glasgow, specifically, the development of the on-line metabolomics monitoring system, the selection of the solvent system and the introduction of a washing step. The HPLC analysis of fermentation samples was carried out by Ingenza Ltd.

The contents of this chapter were submitted for publication in the journal *Biotechnology and Bioengineering* on 15 April 2022 under the title "On-line untargeted metabolomics monitoring of an *E. coli* succinate fermentation process".

## Abstract

Process monitoring is of crucial importance for the industrial production of biobased fermentative products. This chapter shows the development and application of a mass spectrometry-based on-line monitoring system capable of sampling undiluted fermentation broth samples automatically from a 5 L bioreactor. The different elements of the system, as well as the choice of solvent and the implementation of a washing step between sample injections will be discussed as important factors in the performance of the on-line monitoring system. Using mass spectrometry allows, on one hand, the direct monitoring of targeted key process compounds and, on the other hand, provides information of hundreds of additional untargeted compounds without requiring previous calibration data. An *Escherichia coli* succinate fermentation process was monitored with the on-line metabolomics system developed and the results indicated that significant carbon pools accumulating in the pentose phosphate pathway (PPP) during the anaerobic succinate production phase, highlighting the PPP as a possible target for carbon flux reshuffling to improve succinate production, which will be the subject of Chapter 4.

## 3.1 Introduction

Different variables are frequently monitored in fermentation processes, such as the pH, temperature and dissolved oxygen of the liquid phase, as well as the oxygen, carbon dioxide and pressure of the gas phase. Although there are many other parameters that can be monitored – such as turbidity, rheology, enzyme activity or metabolite concentration among others (Y. Harada *et al.*, 2014) – their use is less ubiquitous, especially for large scale fermentations.

Biological systems are complex, with different levels of regulation at the DNA, RNA, protein and metabolite level. However, as metabolites are at the final step of biological regulation, metabolite concentrations and metabolic fluxes provide the best picture of cellular phenotypes (Farrell *et al.*, 2014; Fiehn, 2002). For this reason, the monitoring of the chemical species in the fermentation broth using HPLC and, especially, vibrational spectroscopy has received more attention in the last couple of decades (see Section 1.1.4 and Table 1.3). Although these technologies have been successfully used for real-time monitoring purposes in bioreactors, these are usually limited to analysing a small number of target compounds (usually no more than six metabolites reported using vibrational spectroscopy). Furthermore, vibrational spectroscopy requires the laborious development of chemometric models such as PLS regression to distinguish the different compounds in the mixture.

Compared to these techniques, mass spectrometry offers a much wider detection capacity, making it an attractive alternative for bioprocess monitoring. Furthermore, metabolomics can be used in an untargeted fashion, making it possible to use this technology not only for the monitoring of compounds during manufacturing, but also as a powerful tool during earlier research and development phases, for instance in the identification of by-products, inhibitors or engineering targets, as it will be elaborated in this chapter.

Despite these advantages, to date, on-line metabolomics still remains unexploited for bioprocess monitoring. Link *et al.* (2015) reported the use of metabolomics to monitor different organisms directly from a cultivation flask. However, the cells grown in these experiments were cultured in fermentation media that was up to eight times diluted, which is an impractical imitation for bioprocess monitoring. Plum and Rehorek (2005) reported an on-line mass spectrometry system for analysing nine azo dyes in a waste water treatment process. However, this system contained a biomass filtration unit and was targeted to only nine compounds, thus limiting the vast detection capacity of mass spectrometry. A range of so-called “miniaturized” mass spectrometry analysers – such as the MiD (Microsaic) and the Rebel (908 devices) are also appearing in the market due to the increasing demand for metabolite monitoring tools (Hamilton *et al.*, 2014; Synoground *et al.*, 2021). However, these are still almost exclusively being used at-line or off-line, thus limiting their monitoring potential, and only targeting predefined set of compounds (typically glucose, lactate, ammonia, glutamine and glutamate).

More examples of the application of mass spectrometry for on-line bioprocess monitoring have been reported for gas-phase analysis (see Table 1.2). IONICON Analytik and Thermo Scientific™ are two companies commercialising mass spectrometers for on-line monitoring of the gas phase of fermentation processes. However, analysing the gas phase limits the high detection capacity of mass spectrometry to gases and volatile compounds. In microbial strains and cell lines that release few volatile compounds, the off-gas analysis might be limited to just $O_2$ and $CO_2$, as it is the case for the reported examples using mammalian cells in Table 1.2. As cells cultured in a bioreactor are dispersed in the liquid phase, the majority of metabolites are also present in the liquid. Therefore, employing metabolomics for on-line monitoring of the liquid phase would be applicable for a much larger list of compounds, especially using untargeted metabolomics. The development of a fermentation monitoring system using untargeted metabolomics with direct sampling from the liquid phase will be the focus of this chapter.

### 3.1.1 On-line metabolomics

The FDA drug manufacturing guidelines defined three types of measurements for process analysis: in-line, on-line and off-line (FDA, 2004) (see Section 1.1.4). For reasons of preserving the sterility of the process, in-line measurements are often preferred in industrial setups. However, mass spectrometry is a sample-destructive technology that has so far never been designed in the form of an in-line sensor. For this reason, the real-time metabolomics platform described in this work is designed to work as an on-line monitoring tool.

There are different important choices and parameters concerning an on-line metabolomics system. One of these choices dealing with liquid samples is whether to use a chromatographic separation step or not. The advantage of using chromatography is that it provides additional information to identify the compounds in the mixture, but the disadvantages is that it adds complexity to the system and increases the time of analysis. In order to reduce the chances of blocking the on-line system with salts and cells from the medium, it was decided not to use chromatography. Instead, the samples were analysed with flow injection mass spectrometry (Sarvin *et al.*, 2020). Without chromatography, there are a few other important considerations for the performance of an on-line metabolomics system. These are introduced below.

#### Important considerations for on-line metabolomics

The choice of solvent for the on-line metabolomics system has an effect over different factors that need to be considered. Properties such as the viscosity of the solvent have a direct impact on the resistance to the flow (back-pressure) experienced by the pumps of the system (Gross, 2017). Furthermore, the solvent polarity and pH will impact the accumulation of particulates and dirt in the lines, which will also affect the back-pressure of the pumps. Being able to maintain this pressure below the instrumental limits throughout the duration of the

experiment is one of the most important practical aspects for the development of an on-line monitoring system with automatic sampling. The polarity and pH of the solvent of choice will also affect ionisation and detection of metabolites in the mass spectrometer (Hoffmann and Stroobant, 2007). However, good ionisation conditions for a given group of compounds might be detrimental for another. For instance, a low pH will favour protonation of compounds and therefore improve metabolite signals in positive mode, and the opposite for solvents with a high pH. Another important consideration using a real-time metabolomics system that analyses liquid-phase samples of a bioreactor is the presence of biomass and a complex mixture of chemicals – some of which might be at concentrations as high as several grams per litre – which can be challenging to operate continuously without any blockage of the analytical instruments. Consequently, the use of a washing step between sample injections might need to be considered.

To summarise, the choice of solvent and the addition of a washing step will have an effect on several aspects of the on-line monitoring tool and therefore, different chemical solutions should be tested, and their impact assessed.

### 3.1.2 Aims and objectives

The main aim of this chapter is to establish an on-line metabolomics system for monitoring the liquid phase of fermentation processes and to study how such platform could be used to better understand and improve the fermentation. As a case study, an *E. coli* succinate fermentation process will be used to develop and demonstrate the on-line metabolomics monitoring platform. Special attention will be paid to the solvent system used to introduce the fermentation sample into the mass spectrometer, the use of an auxiliary washing step to prevent blockages and signal loss of the monitoring system, and to metabolite annotation.

## 3.2 Results

### 3.2.1 The fermentation process

Figure 3.1 shows the typical profile of the dual-phase *E. coli* succinate fermentation process. The initial biomass growth phase is characterised by an exponential increase in the cell concentration – represented as the biomass WCW – and a corresponding exponential decrease in dissolved oxygen (DO). The DO was set to be maintained above 30 % of oxygen saturation, controlled by the agitation of the impeller, which can lead to noisy DO profiles when this parameter is around 30 %. This is caused by the control system being turned on and off around the 30 % DO value. When the DO goes below 30 %, the impeller agitation increases, causing the DO to increase, thus causing the mentioned fluctuations.

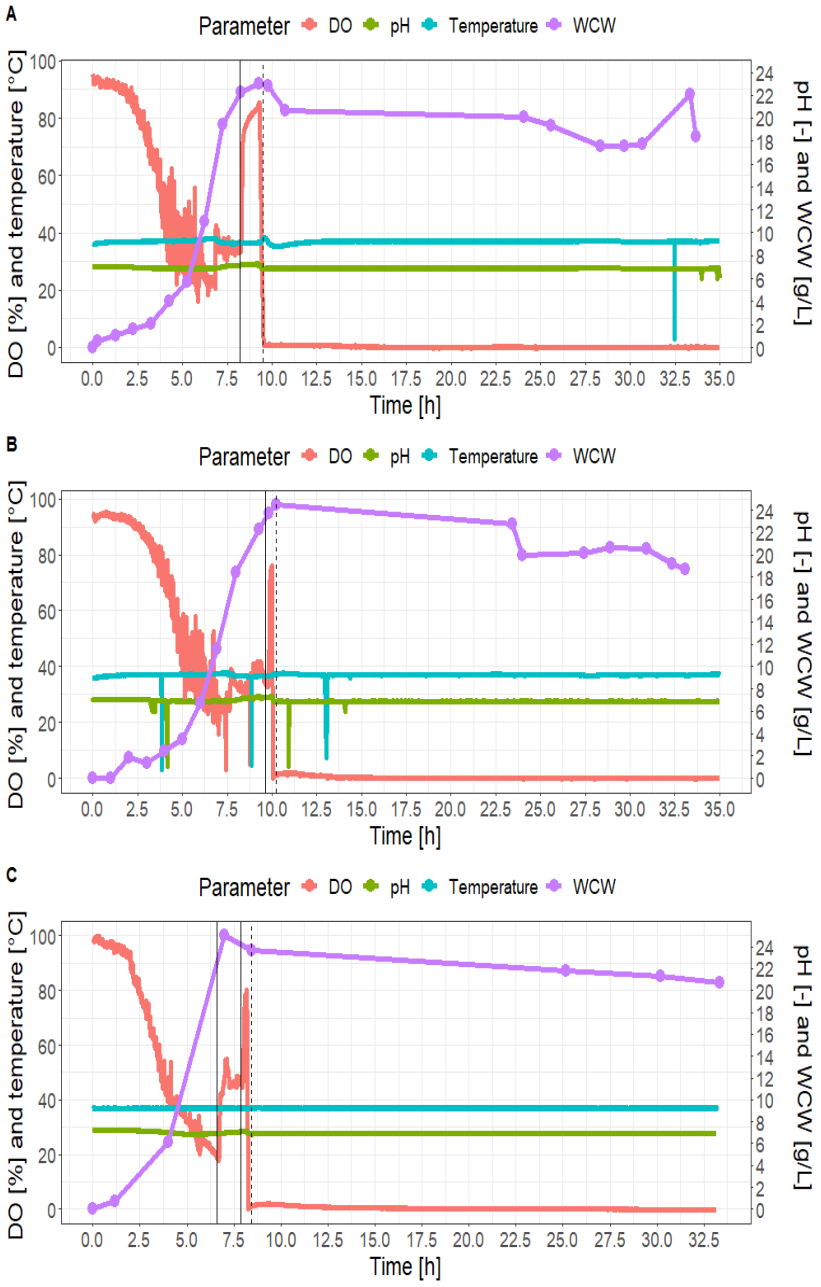



**Figure 3.1:** Three examples of the typical profile of the dissolved oxygen (DO), pH, temperature and biomass wet cell weight (WCW) of the *E. coli* succinate dual-phase fermentation. The vertical black solid line indicates a spike up in the DO, which is indicative of glucose starvation. The vertical black dashed line indicates the transition from the aerobic growth phase to the anaerobic succinate production phase.

Analysis of the off-gas during the biomass production phase (Figure 3.2) also shows a drop in oxygen percentage caused by the transfer and consumption of oxygen in the liquid phase, as well as the increase in carbon dioxide concentration in the gas exhaust, due to cell respiration. The exhaust $CO_2$ is often used to calculate the carbon dioxide evolution rate (CER), which is an indicator of activity of primary metabolism and biomass growth (Spéranido and Paul, 1997; L. Wu *et al.*, 2003). When glucose is depleted from the medium, the cells abruptly stop consuming oxygen, which is characterised by a spike up in the DO. Succinate production using the IGZ006 *E. coli* strain takes places anaerobically. Therefore, at the beginning of the succinate production phase, air is no longer supplied to the fermenter, and $CO_2$ is sparged through the medium instead, leading to close to 0 % DO during the whole production phase. The temperature and pH were controlled at 37 ℃ and 7, respectively for the duration of the whole fermentation process. The simplicity of this fermentation process makes it an excellent case study for the development of an on-line metabolomics platform, which will be described in detail in the next section.

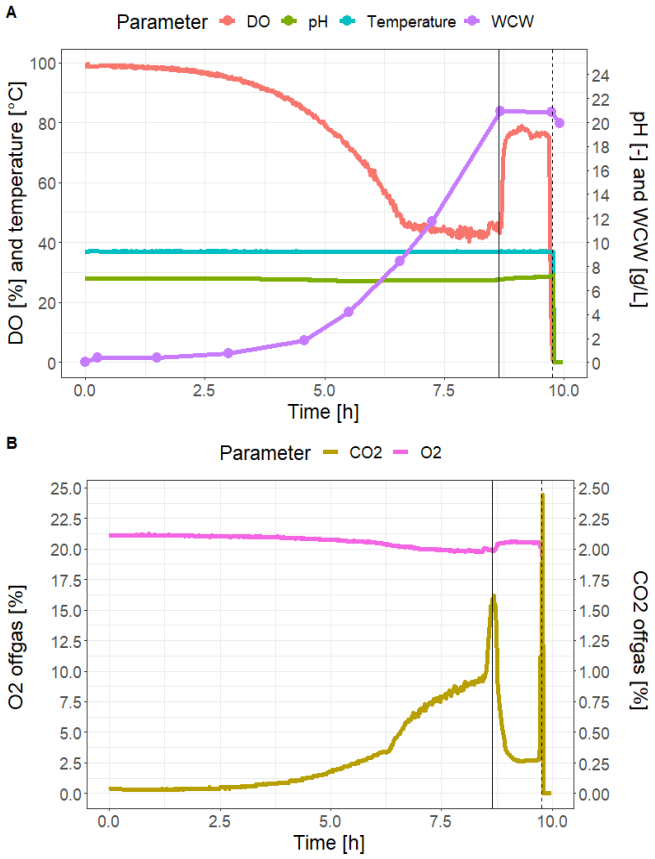



**Figure 3.2:** Fermentation profile (**top**) and off-gas analysis (**bottom**) of an *E. coli* succinate production fermentation during the aerobic biomass production phase. The vertical black solid line indicates the drop in off-gas $CO_2$, which is indicative of glucose starvation. The vertical black dashed line indicates the transition from the aerobic growth phase to the anaerobic succinate production phase.

### 3.2.2 Development of the on-line metabolomics monitoring system

Different aspects were found of critical importance for the development of the on-line metabolomics monitoring system, namely the sampling system, the solvent selection and the use of a washing step. These will be expanded below.

#### Sampling system

A schematic of the sampling system is provided in Figure 3.3. The system consists of a peristaltic pump, a 6-port valve with a 50 $\mu$L sampling loop, a 10-port valve with a 1 $\mu$L sampling loop, an aqueous pump and a solvent pump. Both valves have two positions: “injection” and “load” (Figure 3.3). In the "inject position", fermentation broth from the bioreactor is continuously circulated through a 50 $\mu$L loop in the first valve. When the system switches to the "load position", water pushes the sample into a 1 $\mu$L loop situated in the second valve. Then, the valves switch back to the "inject position", causing the solvent system to push the fermentation sample into the mass spectrometer. Using two valves aims to mitigate the solvent incompatibility at the two ends of the system. Namely, at one end, the bioreactor contains a water-based environment with living cells, and at the other end, ESI mass spectrometer works best with volatile organic solvents, which are more effective than water at generating gas-phase ions (Hoffmann and Stroobant, 2007). Organic solvents commonly have a detrimental effect on the health and behaviour of cells, which are adapted to live in aqueous environments, therefore the two solvents – organic and aqueous – need to be separated.

It is worth mentioning that the system described in Figure 3.3 does not contain a chromatographic step where molecules are separated on given physicochemical properties such as size, charge or hydrophobicity. This type of setup is called flow injection mass spectrometry (Sarvin *et al.*, 2020). In the absence of chromatography, the identification of metabolic features relies to a larger extent on accurate mass, fragmentation and the dynamic profile observed throughout the experiment, as it will be commented in Section 3.2.3. Different elements of the sampling system are detailed further below. For more details of the sampling method, see Section 2.6.

*Peristaltic pump*

Fermentation broth is continuously pumped out of the bioreactor through a peristaltic pump at a high flow rate (75-100 mL/ min) in order to avoid accumulation of biomass in the system and to minimise the traveling time of the sample from the fermenter into the mass spectrometer. It was observed that a high-pressure peristaltic pump was necessary in order to prevent the malfunctioning of the pump mid-way through a fermentation experiment. A Masterflex$^{TM}$ L/S ® (Cole-Parmer) high-performance pump model 77252-72 was chosen for this purpose.

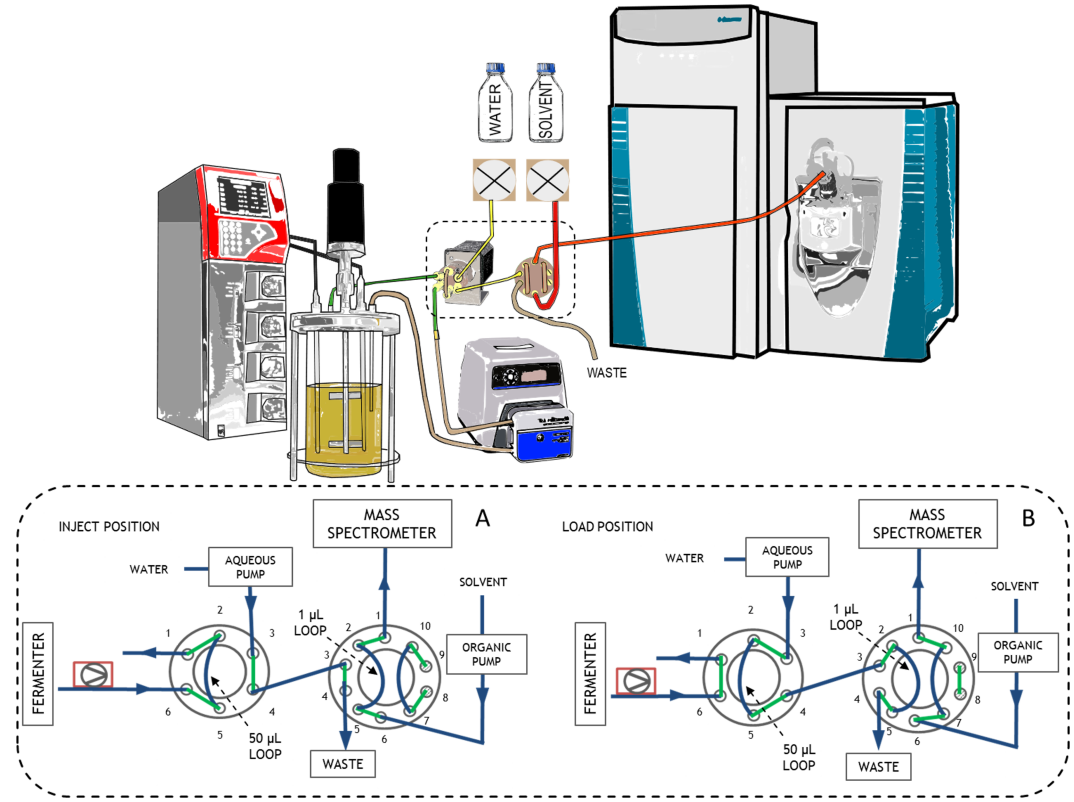



**Figure 3.3:** Experimental setup of the on-line metabolomics system (**top**) and a detailed diagram of the dashed area showing the two valve positions (**bottom**). During the **"inject position"**, fermentation broth sample is continuously circulated through the 50 $\mu$L of the left 6-port valve (sample n+1) and the fermentation broth sample from the 1 $\mu$L loop in the right 10-port valve (sample n) is pushed to the mass spectrometer. During the **"load position"**, the sample n+1 is pushed from the 50 $\mu$L loop of the 6-port valve into the 1 $\mu$L loop of the 10-port valve, ready for injection at the next **"inject position"**.

*Six-port valve*

The fermentation broth is pushed by the peristaltic pump into a 6-port valve, which has a 50 $\mu$L sampling loop. The broth is recirculated back into the fermenter for the majority of the time in the "inject position". When the valve position is changed to "load position", the 50 $\mu$L of the sampling loop is pushed by sterile water and travels to the 10-port valve, where 1 $\mu$L of sample gets collected in a 1 $\mu$L sampling loop. A VICI® Cheminert® 6 port 2-position valve is used in the on-line system. A wide-bore diameter valve (0.75 mm) was chosen to minimise the risk of blockages from the biomass. This restraints the valve pressure limit to 250 psi.

*Tubing adapter*

The sampling system contains tubing of different sizes. Namely, the peristaltic pump is operated with Masterflex$^{TM}$ flexible tubing, while the 6-port VICI® valve uses polyetheretherketone (PEEK) semirigid tubing. The difference in tubing diameter requires the use of tubing adapters, which might consist of different pieces. In this particular case, the adapter consisted of four pieces (Figure 3.4) and one of these was particularly problematic to purchase online. For this reason, knowing the dimensions of the other pieces, a 3D version of the missing piece was designed in the open access online tool Tinkercad (https://www.tinkercad.com. Accessed on 29$^{th}$ July 2019) (Figure 3.5) and printed on a Ultimaker3 3D printer with a PA6/66 nylon polymeric mixture, which is suitable for autoclave conditions.

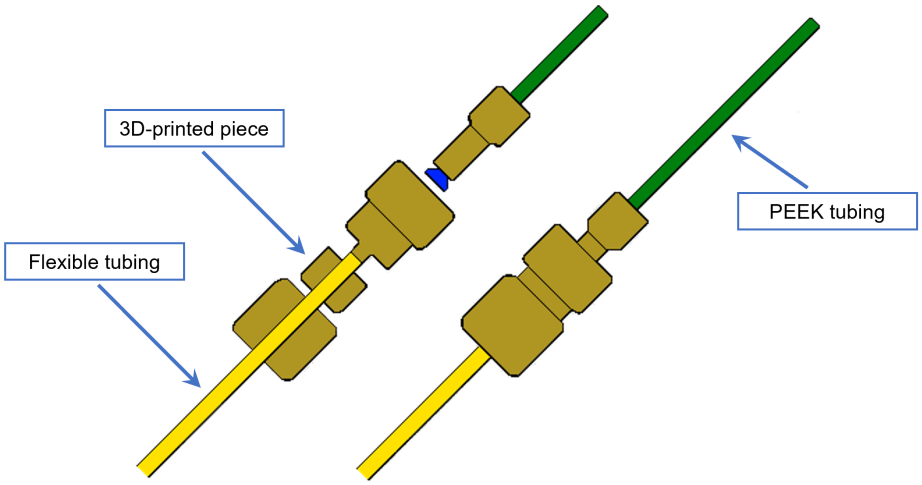



**Figure 3.4:** Flexible tubing to PEEK tubing adapter.

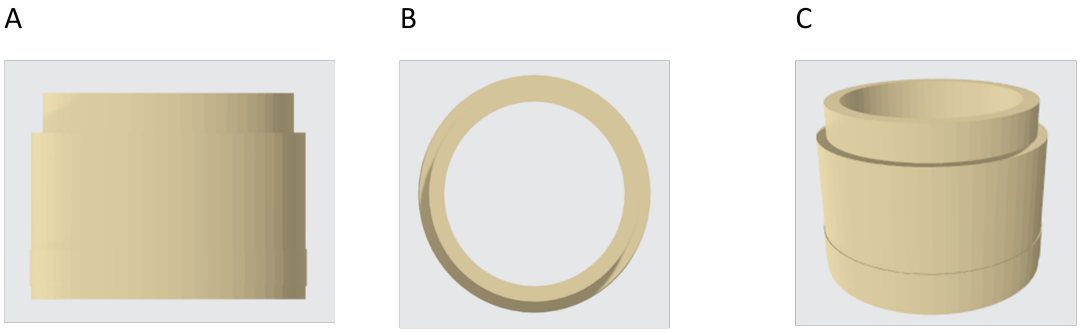



**Figure 3.5:** Tubing adapter designed on the open access online tool Tinkercad for 3D printing. Front view (**A**), top view (**B**) and 3D view (**C**).

*Ten-port valve*

The 50 $\mu$L collected in the sampling loop of the 6-port valve are pushed with sterile water to the 10-port valve using a piston pump from an HPLC instrument (aqueous pump). The fermentation sample gets pushed into a 1 $\mu$L sampling loop on the 10-port valve during the "load position", and when the valve switches to the "inject position", the 1 $\mu$L sample gets injected into the mass spectrometer, pushed with organic solvent using a second piston pump from the same HPLC instrument (organic pump).

### Solvent system selection

Two experiments were carried out to assess the effect that the solvent used to introduce the fermentation sample into the mass spectrometer has on the performance of the on-line metabolomics monitoring system. Thirteen different solvent mixtures with different levels of polarity, pH and physicochemical properties were tested with the *E. coli* succinate process. These were different combinations of acetonitrile (ACN), $H_2O$, methanol (MeOH) and isopropanol (IPA), with acetic acid (AA), ammonium carbonate (AC) and formic acid (FA) being used for pH adjustment.

The different parameters used for this assessment were:

- Positive and negative mode total ion chromatogram (TIC).
- Positive and negative mode total number of m/z signals detected.
- Signal intensity of glucose and succinate.
- Aqueous and solvent pump back-pressure.

These parameters were selected as general indicators of ionisation capacity and operational performance. The exception would be the signal of glucose and succinate, which were specifically analysed as they are the substrate and product of the *E. coli* succinate bioprocess studied. The first experiment was conducted in a 1 L Schott bottle, as a scale-down and less-complex version of the bioprocess. Then, the best performing solvent systems were tested in a second experiment using the 5 L bioreactor, in order to better reproduce the conditions of the bioprocess (see section 3.5.1). In both experiments, the aim was to test the different solvents in a culture that would change as little as possible in time, in order to avoid significant changes in the metabolic profile of the culture caused by events such as substrate starvation (either glucose or oxygen), or by differences in biomass levels at the beginning and the end of the experiment.

The first experiment was done on a 1 L Schott bottle at a biomass concentration of 1.23 ± 0.024 $OD_{600}$ (see section 3.5.1), and for technical reasons, solvent mixtures 1 and 11 (Figure 3.6) could not be tested in this initial test. The results are shown in Figure 3.6 and, as anticipated, different solvent systems perform better depending on the parameter of interest analysed. The most intense glucose signal was obtained using solvent mixture number 4,

whereas solvent mixture number 3 gave the best succinate signal and highest TIC in negative mode. These two parameters are closely related, with succinate representing approximately 25 % of the negative mode TIC on solvent mixture 3. Surprisingly, solvent mixture 3 has a relatively neutral pH (7.06), and therefore, basic solvent mixtures – such as 5 or 8, with a pH of 9.40 and 8.97, respectively – were expected to show a higher TIC on negative mode. In fact, it was even more surprising the fact that solvent solutions 9 and 10 – with a pH of 2.95 and 3.09 respectively – showed a higher succinate intensity than solvent mixtures 5 and 8.

The highest TIC signal on positive mode was observed with solvent mixtures 9 and 10 – with a pH of 2.95 and 3.09 respectively. Although a low pH should favour ionisation on positive mode, solvent s 9 and 10 – especially 10 – outperformed all the other solvent solutions, even though 2, 7, 12 and 13 were also very acidic (pH between 2.02 and 3.55).

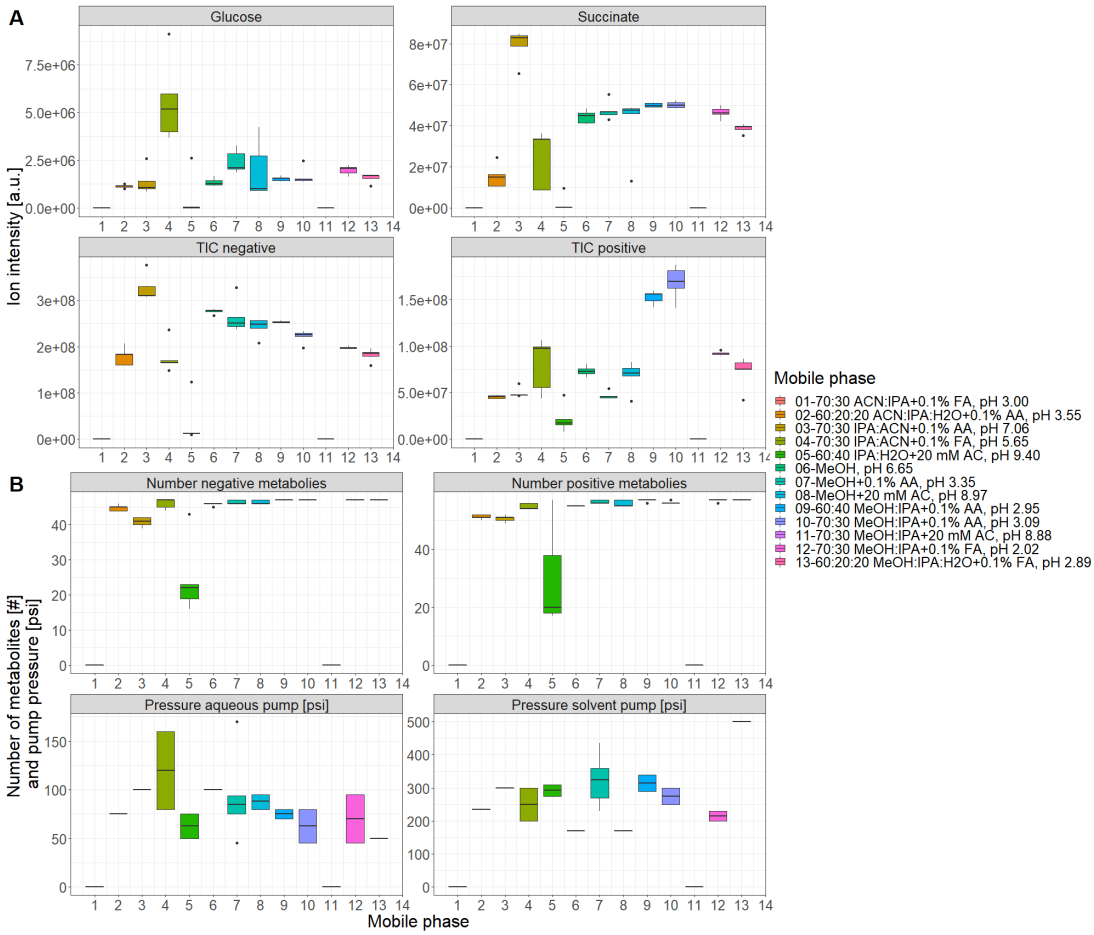



**Figure 3.6:** Variables of interest for solvent system selection. Experiment run with the on-line metabolomics system connected to a Schott bottle. **A**: glucose intensity, succinate intensity, TIC in negative mode, TIC in positive mode. **B**: number of metabolites detected in negative and positive mode, aqueous pump and solvent pump back-pressure. Solvent mixtures 1 and 11 were not included, hence a value of 0 was given to these in all categories. Abbreviations: AA: acetic acid; AC: ammonium carbonate; ACN: acetonitrile; FA: formic acid; IPA: isopronaol; TIC: total ion chromatogram.

Regarding the number of m/z features found, similar values were obtained across the different solvent systems tested. As far as the pumps back-pressure are concerned, on one hand, solvent system 4 resulted in considerably higher aqueous pump back-pressures than the rest of solvent mixtures, which operated at back-pressures no higher than 100 psi. On the other hand, only solvent mixtures 6 and 8 showed solvent pump back-pressures below 200 psi. As the pressure limit of the VICI valve was 250 psi, it was desired to operate the system below this pressure.

Solvent solution 5 performed particularly poorly in most of the parameters analysed, showing a low TIC as well as a low succinate and glucose intensity, and a low number of metabolites in both positive and negative mode. Finally, solvent solution 13 showed a very high solvent pump back-pressure of 500 psi. As discussed above, solvent solutions 3, 4, 6, 8 and 10 were the best performers in at least one of the categories assessed. Therefore, these were taken through to a second round of experimentation. Solvent solutions 6 and 8 performed very similarly in terms of solvent pump back-pressure. Therefore, only one of them was carried through in the second round. As solvent 6 was MeOH and solvent 8 was MeOH + 20 mM AC, solvent 6 was selected over 8 for a matter of simplicity.

These best performing solvent systems, together with solvent mixtures 1 and 11 – which could not be tested in the first experiment – were tested in a second experiment using a fermenter with an *E. coli* culture at a high concentration (OD600 ca. 8, see section 3.5.1). The results of this second experiment are shown in Figure 3.7.

Consistently with the results from Figure 3.6, solvent system 3 showed the highest TIC on negative mode and succinate intensity in Figure 3.7. Furthermore, solvent system 10 also scored the highest on positive mode TIC. In this case, however, solvent systems 1 and 11 outperformed solvent system 4 regarding glucose intensity. Once again, the number of metabolites was very similar across all the different solvent systems tested, in both positive and negative mode. Finally, as far as the pump back-pressure is concerned, on one hand, all six solvent systems showed a very similar aqueous pump back-pressure. On the other hand, only solvent systems 1, 6 and 11 showed a solvent pump back-pressure no higher than 200 psi (also consistently with the findings of the first experiment growing the cells in a Schott bottle). Furthermore, solvent mixture 1 showed by far the lowest solvent pump back-pressure (140 psi).

In both experiments it was observed that depending on the parameter selected, a different solvent mixture will be considered best. However, due to the low solvent pump back-pressure observed using solvent system 1 (70:30 ACN:IPA + 0.1 % FA), this mixture was considered to be the first option to try on subsequent on-line metabolomics experiments. For a further discussion, see Section 3.3.3.

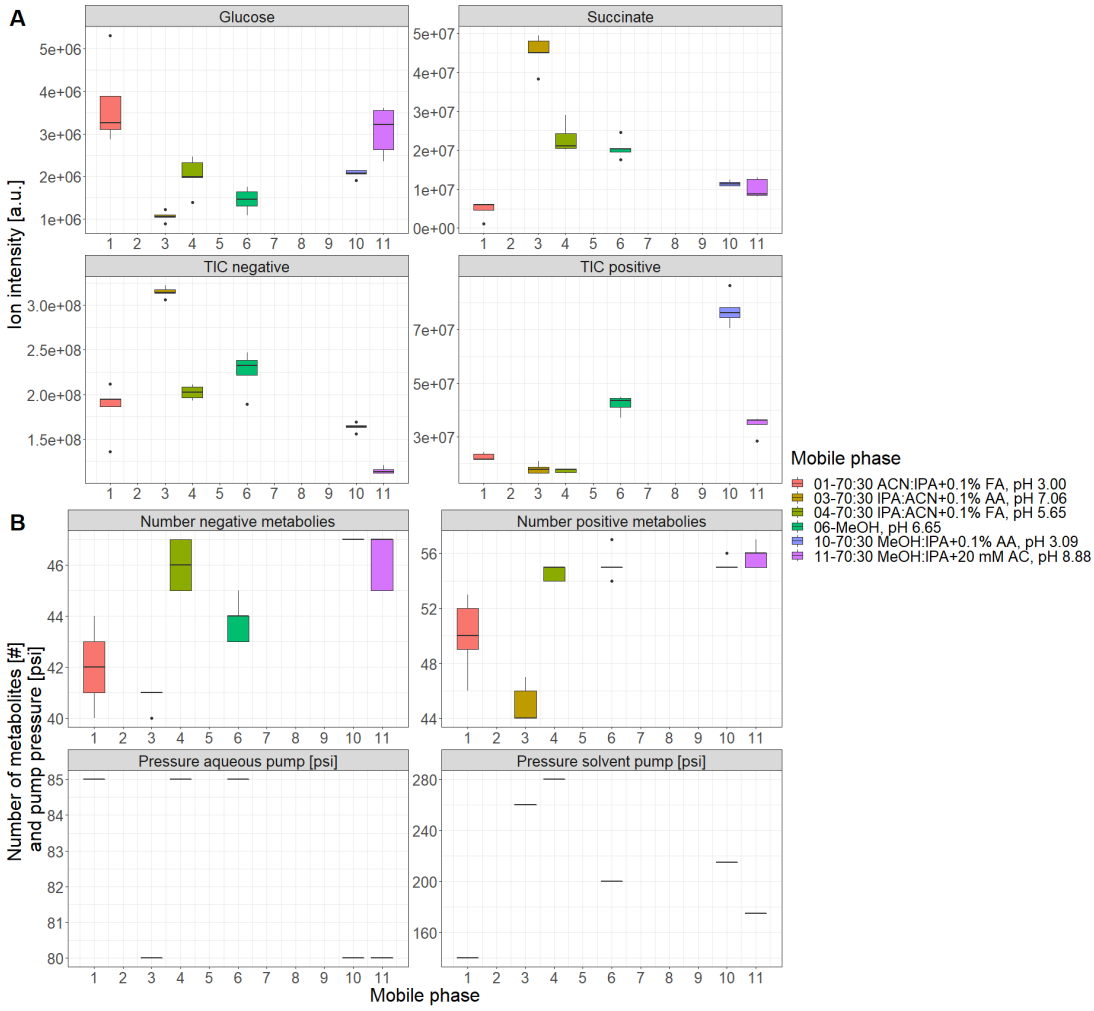



**Figure 3.7:** Variables of interest for solvent system selection. Experiment run with the on-line metabolomics system connected to a bioreactor. **A**: glucose intensity, succinate intensity, TIC in negative mode, TIC in positive mode. **B**: number of metabolites detected in negative and positive mode, aqueous pump and solvent pump back-pressure. Abbreviations: AA: acetic acid; AC: ammonium carbonate; ACN: acetonitrile; FA: formic acid; IPA: isopronaol; TIC: total ion chromatogram.

### Introducing a washing step between sample injections

The initial five fermentation runs using the on-line metabolomics setup for monitoring the *E. coli* succinate fermentation process were carried out sampling from the fermenter every minute with no intervention on the system. This led to blockages of the fluidics system in every experiment, sometimes during the first few hours of the fermentation. In order to solve this technical limitation, a washing step was introduced between sample injections using a washing solution of 70:30 IPA:ACN + 0.1 % FA, which is commonly used for cleaning ESI mass spectrometer needles. The implementation of this washing step was tested in a series of nine *E. coli* succinate fermentations, and the main variable parameters investigated were:

- Washing the 6-port valve or not.

- The flow rate of the washing solutions: 200, 400 and 600 $\mu$L/min.
- The total duration of the washing step and the duration that each specific mix was used for. It is worth mentioning that the longest section of tubing in the system was the PEEK tube connecting the 10-port valve and the mass spectrometer, and it was 250 $\mu$L in total tubing volume.
- The frequency of sampling: every 5 and 10 min.
- The sample:wash ratio: 12:1, 3:1, 1:1.
- Using a sequence of an aqueous solution and organic solvent to clean the mass spectrometer side of the system, or using a combination of two organic solvents instead.
- The solvent system carrying the sample into the mass spectrometer: MeOH, MeOH + 20 mM AC and 70:30 ACN: IPA + 0.1 % FA. The selection of solvent system (see Section 3.2.2) was carried out between the nine fermentations mentioned in this section. This resulted in three different solvent systems being tested, even though this factor was not initially designed to be a variable in the experiments.

The nine *E. coli* succinate fermentations compared are gathered in Table 3.1. Comparing experiment 1 and 2, it was observed that washing the 6-port valve of the real-time metabolomics system had no negative effect on the behaviour of the cells growing in the bioreactor, while it helped to prevent blockages. The main potential concern was that some solvent used during the wash of the 6-port valve could remain in the 50 $\mu$L sampling loop and lead to cell death during the “Inject position” of the sample following the wash (see Figure 3.3 and Figure 2.3). It must be admitted that comparing these two experiments, the flowrate of the washing step for the 10-port valve was also increased from 200 $\mu$L/min to 400 $\mu$L/min and, therefore, with only this information it is not possible to know the contribution of each of the two individual changes on extending the absence of blockages from 5.92 h to 25.33 h. Instead, all that can be concluded is that the combination of both changes had a positive effect.

Another important observation from Table 3.1 was that in experiment 4, which had a 3:1 sample:wash ratio, the TIC signal of consecutive injections consistently increased immediately after every washing step (Figure 3.8A). This indicated that the washing step helps to prevent not only system blockages, but also signal loss, potentially due to the removal of particulates and build-up molecules accumulated in the system during sample injection. When the washing step was used after every injection (experiments 5-9), the changes in the TIC did not follow any periodic pattern. An example with data from experiment 9 is shown in Figure 3.8B. In both cases there was a decreasing trend in the TIC during the first 15 injections, but this is probably not caused by signal loss, but by the consumption of glucose from the media by the cells for biomass formation. With these observations, it was deemed necessary to implement a washing step after every single injection.

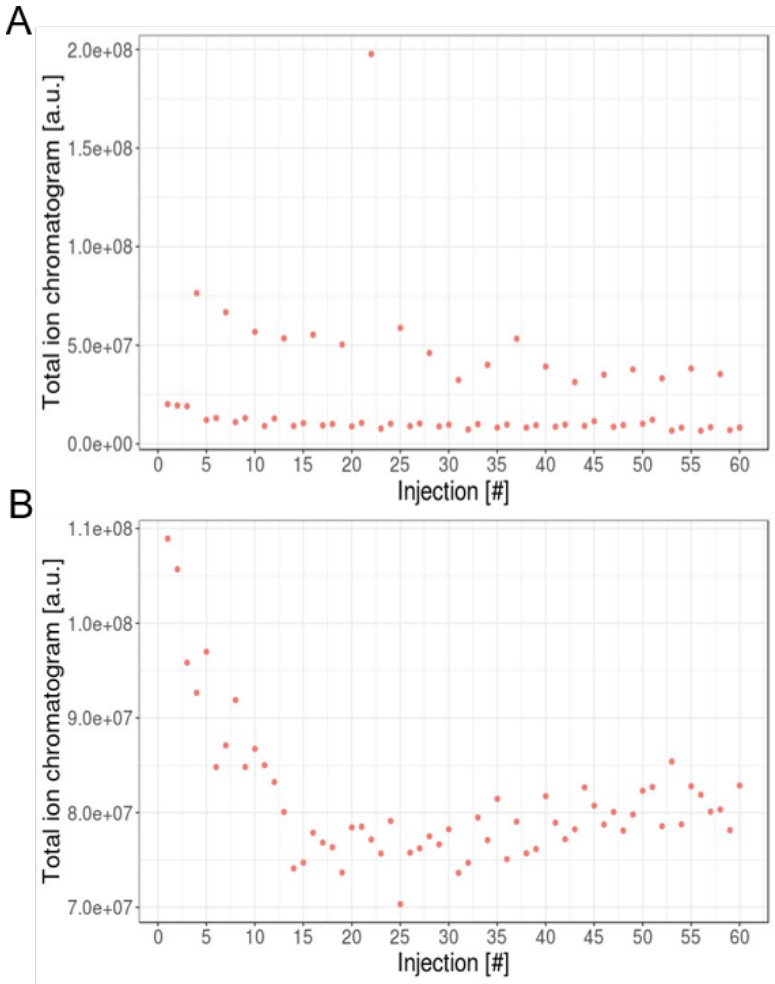



**Figure 3.8:** Total ion chromatogram (TIC) of the first 60 injections from a succinate production fermentation analysed by on-line metabolomics and with a sample:wash ratio of 3:1 (**A**) or 1:1 (**B**). Only the first 60 injections are shown for easier visualisation of the x-axis. In figure A it can be appreciated that after every third injection, the TIC increases, corresponding with injections immediately after a washing step.

Looking at experiments 5, 7 and 9, it can be observed that both injecting one sample every five and ten minutes it is possible to maintain the system running without blocking for more than 45 h, which is about 12 h longer than the duration of the fermentation process itself (approximately 33 h). Checking this is important, as the shorter the time between samples, the higher the monitoring time resolution. Increasing the sampling frequency to less than five minutes$^{-1}$ would mean that the washing step would need to be shorter than four minutes. Given that the longest section of tubing came after the 10-port valve and had 250 $\mu$L volume, washing it with a flow rate of 600 $\mu$L/min for two minutes with solvent mixtures A and C represented nearly five volume changes of the inside of the tubing with each solvent mix. Although it might be possible to shorten the washing step, three volume changes with each solvent mixture at 600 $\mu$L/min would take 1.25 min with each solvent (2.50 min total). Finally,

it is worth pointing out that experiments 6 and 8 blocked much earlier than 45 h, especially experiment 8. It was suspected that this could be caused by either the tubing or the mass spectrometer spray needle being partially dirty from the beginning of the experiment, although it is not possible to prove this retrospectively.

Comparing experiment 6 to 7 and 9, it was concluded that it helped to prevent blockages cleaning the 10-port valve first with water and then with organic solvent (mix A and C as labelled in Table 3.1), for two minutes each, rather than using two different organic solutions (mix B and C as labelled in Table 3.1), for one and three minutes respectively. The reasoning behind testing this was that using two solutions of different hydrophobicity probably helps to clean the lines of a broader range of molecules of different properties that could potentially cause blockages. However, because not only the type of washing mix was changed (A or B), but also the duration that these were used for before switching back to mix C (2 minutes or 1 minute), it is not possible to separate these two individual effects and be certain that using water first and then an organic solvent on the 10-port valve is the best option to prevent blockages of the on-line monitoring system.

Finally, in Table 3.1 it can also be observed that using high flow rates of 400 and 600 $\mu$L/min for the 6-port and 10-port valves respectively during the wash step did not lead to issues of overpressure of the system (experiments 3 to 9). Therefore, it was decided to keep these high flow rates because they should help clean the lines of particulates due to their higher pressure.

Out of all the experiments shown in Table 3.1, experiments 7 and 9 were considered to have the best conditions to be used for the following fermentation experiments, as the on-line monitoring worked successfully for a long time without blockages, had a washing step after every sample and a good monitoring resolution (one sample every five minutes). These conditions are gathered in Table 3.2.

**Table 3.1:** Conditions tested for the washing step. For all experiments, mix A was $H_2O$ and mix B was 70:30 IPA:ACN + 0.1 % FA. Abbreviations: AC, ammonium carbonate; ACN, acetonitrile; FA, formic acid; IPA, isopropanol; MeOH, methanol.

| Experiment number | Solvent system (mix C) | Six-port valve | Ten-port valve | Frequency of sampling | Sample:wash ratio | Comments |
|---|---|---|---|---|---|---|
| 1 (Batch 5) | MeOH | No wash | 3 min B at 200 $\mu$L/min + 7 min C at 200 $\mu$L/min | 10 $min^{-1}$ | 12:1 | System blocked after 5.92 h |
| 2 (MethDev 1) | MeOH | 2 min B at 200 $\mu$L/min + 7 min A at 200 $\mu$L/min | 3 min B at 400 $\mu$L/min + 6 min C at 400 $\mu$L/min | 10 $min^{-1}$ | 12:1 | System blocked after 25.33 h |
| 3 (MethDev 2) | MeOH | 4 min B at 400 $\mu$L/min + 5 min A at 400 $\mu$L/min | 4 min B at 600 $\mu$L/min + 5 min C at 600 $\mu$L/min | 10 $min^{-1}$ | 3:1 | Pressure started to look quite high after 20 h |
| 4 (Batch 6) | MeOH | 4 min B at 400 $\mu$L/min + 5 min A at 400 $\mu$L/min | 4 min B at 600 $\mu$L/min + 5 min C at 600 $\mu$L/min | 10 $min^{-1}$ | 3:1 | No blockages after 48.60 h but the back-pressure increased after 27.75 h. The TIC was much higher for the samples immediately after a wash step (Figure 3.8) |
| 5 (MethDev 4) | MeOH | 4 min B at 400 $\mu$L/min + 5 min A at 400 $\mu$L/min | 4 min B at 600 $\mu$L/min + 5 min C at 600 $\mu$L/min | 10 $min^{-1}$ | 1:1 | No blockages or overpressure after 45.83 h |
| 6 (MethDev 9) | 70:30 ACN: IPA + 0.1 % FA | 1 min B at 400 $\mu$L/min + 3 min A at 400 $\mu$L/min | 1 min B at 600 $\mu$L/min + 3 min C at 600 $\mu$L/min | 5 $min^{-1}$ | 1:1 | System blocked after 21.83 h |
| 7 (MethDev 10) | 70:30 ACN: IPA + 0.1 % FA | 1 min B at 400 $\mu$L/min + 3 min A at 400 $\mu$L/min | 2 min A at 600 $\mu$L/min + 2 min C at 600 $\mu$L/min | 5 $min^{-1}$ | 1:1 | System ran for 45.5 h non-stop without blockages |
| 8 (Batch 7) | MeOH + 20 mM AC | 1 min B at 400 $\mu$L/min + 3 min A at 400 $\mu$L/min | 2 min A at 600 $\mu$L/min + 2 min C at 600 $\mu$L/min | 5 $min^{-1}$ | 1:1 | System blocked several times after 5.92 h |
| 9 (Batch 9) | 70:30 ACN: IPA + 0.1 % FA | 1 min B at 400 $\mu$L/min + 3 min A at 400 $\mu$L/min | 2 min A at 600 $\mu$L/min + 2 min C at 600 $\mu$L/min | 5 $min^{-1}$ | 1:1 | Peristaltic pump temporarily stopped after 28.33 h, but suspectedly due to a power issue. No blockage issues in 47.07 h |

**Table 3.2:** Washing step conditions considered most appropriate to operate the real-time metabolomics on-line monitoring system without blocking. Abbreviations: ACN, acetonitrile; FA, formic acid; IPA, isopropanol.

| Mix A | Mix B | Solvent system (Mix C) | Six-port valve | Ten-port valve | Frequency of sampling | Sample:wash ratio |
|---|---|---|---|---|---|---|
| $H_2O$ | 70:30 IPA:ACN + 0.1 % FA | 70:30 ACN: IPA + 0.1 % FA | 1 min B at 400 $\mu$L/min + 3 min A at 400 $\mu$L/min | 2 min A at 600 $\mu$L/min + 2 min C at 600 $\mu$L/min | 5 $min^{-1}$ | 1:1 |

### 3.2.3 Metabolic profile of key fermentation compounds

Once the conditions in Table 3.2 were considered satisfactory for operating the on-line metabolomics system, the data from experiment 9 in Table 3.1 (which was run under the same conditions) was analysed. A total of 359 and 527 different m/z values were observed in negative and positive mode respectively (886 in total). Some of the different patters observed are represented in Figure 3.9. From these, 124 signals were annotated as different metabolites based on accurate mass within 5 ppm m/z error, and 67 of these were matched to annotations performed by off-line LC-MS analysis (see Chapter 4, Section 4.2.2). A few examples of these annotated metabolites are shown in Figure 3.10.

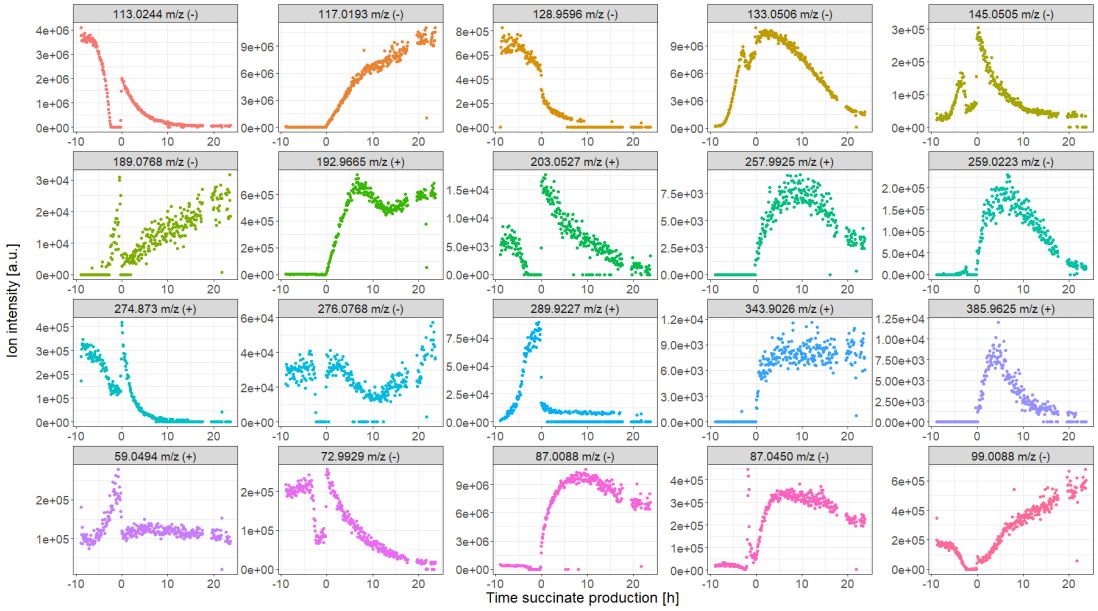



**Figure 3.9:** Time-course patterns observed with on-line metabolomics monitoring of a succinate production fermentation process in *E.coli*. Time is indicated with respect to the beginning of the succinate production phase.

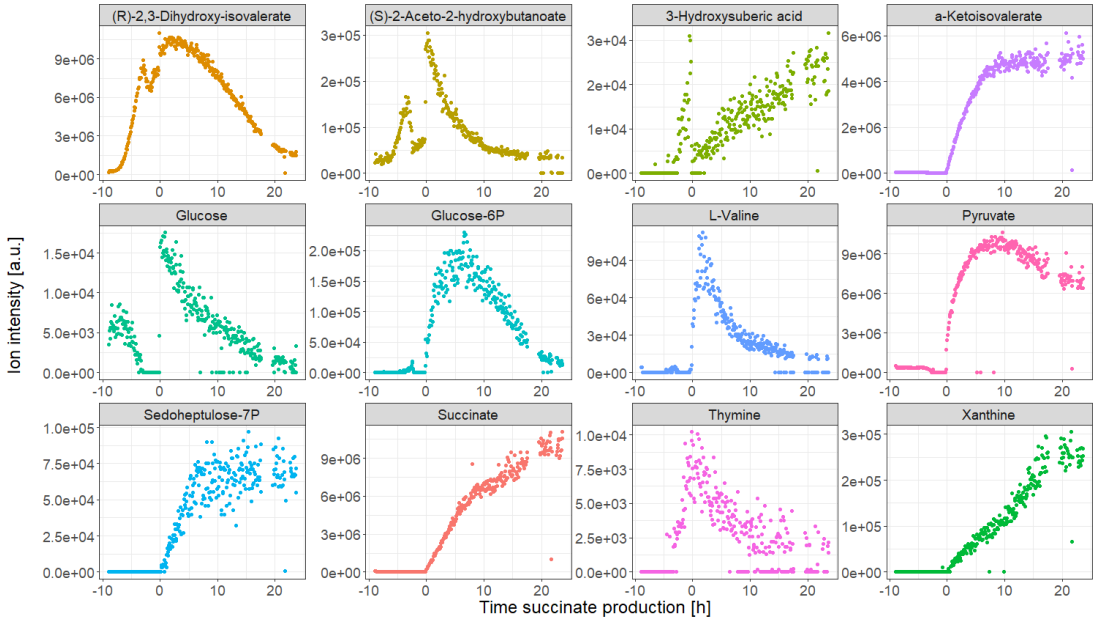



**Figure 3.10:** Example of annotated metabolites observed with on-line metabolomics monitoring of a succinate production fermentation process in *E.coli*. Time is indicated with respect to the beginning of the succinate production phase.

The data shows that some metabolites such as (R)-2,3-dihydroxy-isovalerate and (S)-2-aceto-2-hydroxybutanoate follow an exponential increase in intensity during the biomass formation phase, indicating that these could potentially be used as biomass biomarkers for on-line monitoring. Another observation is that sugar phosphates involved in the PPP, such as glucose-6P and sedoheptulose-7P, and some compounds involved in nucleotide metabolism, such thymine and xanthine, show a significant increase in intensity at the beginning of the sucinate production phase. Added to these examples, compounds of glycolysis, the TCA cycle, metabolites of anaerobic fermentation and amino acids are other examples of metabolites describing patterns worth monitoring in the bioprocess. Some of these are also shown in Figure 3.10.

Being able to see the dynamic evolution of a given m/z signal over the whole experiment can not only provide very meaningful information to increase the biological understanding of the bioprocess, but also avoid misannotations, particularly as there is no chromatography in flow injection mass spectrometry. Some of these misannotations could result from in-source fragmentation or adduct formation. An example of both cases is given below.

An example of a fragment ion is the signal 113.0244 m/z in Figure 3.9, which could be annotated as 3-oxo-4-pentenoic acid ($[C_5H_6O_3 - H]^-$), but seeing that it follows the same profile as glucose (Figure 3.12), it is more reasonable to annotate it as a glucose fragment with formula $C_5H_5O_3^-$ resulting from in-source fragmentation. To test this, a pure glucose standard was analysed in the Orbitrap mass spectrometer and a 113.0244 m/z fragment was

observed together with some other ions such as 91.0037 and 101.0244 m/z, which could be two other glucose in-source fragments with formulas $C_2H_3O_4^-$ and $C_4H_5O_3^-$, respectively. This first example shows how knowing the dynamic evolution of a given m/z signal can be used to improve its annotation.

Potential glucose adducts were found (within a mass error of 5 ppm) based on the rules reported by Huang *et al.* (1999) and plotted in Figure 3.11. Glucose is added at the beginning of the fermentation and consumed at an exponentially increasing rate - due to exponential biomass formation - until depletion at the end of the aerobic batch phase. Then, more glucose is added at the beginning of the anaerobic phase and it is consumed by the cells to make succinate (see Figure 3.12). Based on the knowledge of the dynamic evolution of glucose, signals 1, 2, 3, 5, 7 and 8 in Figure 3.11 could be annotated as true glucose adducts, whereas signals 4 and 6 follow a very different dynamic evolution. This second example shows, again, how the dynamic profile of a metabolite can be used to assess its annotation, especially with prior biological and bioprocess knowledge.

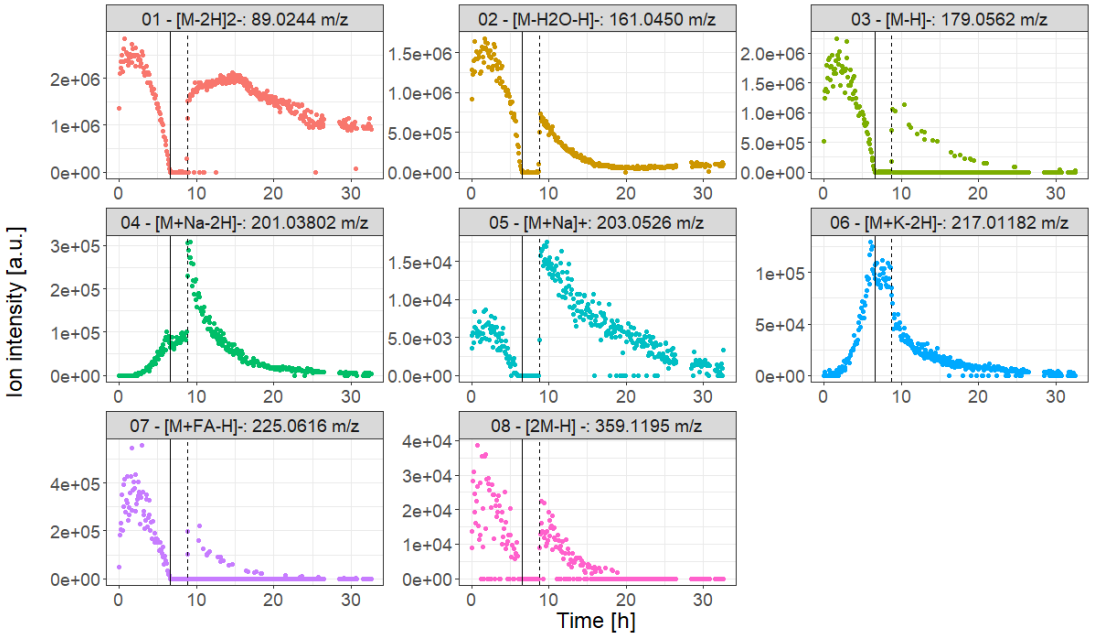



**Figure 3.11:** Potential annotated glucose adducts observed by on-line metabolomics analysis of a succinate production fermentation process using *E. coli* and a mass tolerance of 5 ppm. The vertical black solid line indicates glucose starvation during the aerobic growth phase. The vertical black dashed line indicates the transition from the aerobic growth phase to the anaerobic succinate production phase.

The fact that the six potential glucose adducts are not identical might be because these ionise better or worse at different stages of the experiment depending on what other compounds are present in the fermentation broth. For example, one must consider that during the succinate production phase, carboxylic acids such as pyruvate and succinate are produced in high

concentrations. These compounds ionise well in negative ESI mode and, therefore, might have an effect on the detection of other molecules due to ion suppression and saturation of the mass spectrometer detector, particularly in negative mode. Ions in positive mode (such as signal 5 in Figure 3.11) will be less affected by the presence of pyruvate and succinate.

### Comparing the on-line metabolomics data with off-line HPLC data

HPLC is the most widely used reference method for fermentation off-line analysis and it is commonly used for validation of monitoring methods (Cabaneros Lopez *et al.*, 2019; Legner *et al.*, 2019; Rodrigues *et al.*, 2018). Off-line samples from experiment 9 (Table 3.1) were analysed by HPLC coupled to UV/Vis and refractive index detectors (HPLC-UV/Vis-RI) and compared to the on-line metabolomics signals for the key process metabolites present in both data sets – glucose, succinate and pyruvate – (Figure 3.12). Comparing the off-line HPLC-UV/Vis-RI results with the on-line metabolomics data demonstrates the main advantages of real-time metabolomics. On one hand, the off-line data consists of 16 time points, taken on hourly basis and with time gaps of more than eight hours corresponding to overnight periods, whereas with on-line metabolomics, samples were automatically collected every 5 minutes resulting in 355 time points (a 22-fold increase in resolution compared to the off-line analysis). Furthermore, HPLC-UV/Vis-RI is run as a targeted analysis, leading to the analysis only a handful of compounds, whereas, as demonstrated in Figures 3.9 and 3.10, metabolomics can be used to monitor a much wider range of metabolites.

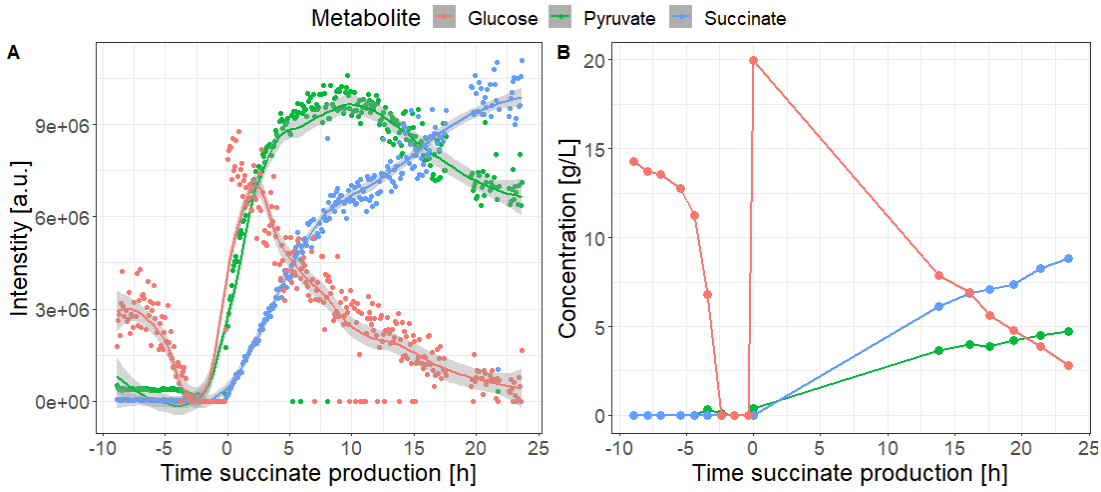



**Figure 3.12:** Glucose, pyruvate and succinate monitored by on-line metabolomics (**A**) and off-line HPLC (**B**). The mass spectrometry measurements are represented as dots and the corresponding smoothed signal is represented with lines and calculated with LOESS. Ions 203.0527, 87.0088 and 117.0193 m/z were respectively used for glucose, pyruvate and succinate. The HPLC data is represented as dots and the interpolated data is represented with lines.

Figure 3.13 shows that the Pearson correlation estimates between the two methods of analysis (on-line metabolomics and off-line HPLC) were 0.86, 1.00 and 0.96 for glucose, succinate and pyruvate respectively, showing a good positive correlation between the two methods. For on-line metabolomics, the signals $[glucose+Na]^+$, $[succinate-H]^-$ and $[pyruvate-H]^-$, corresponding to 203.0527, 117.0193 and 87.0088 m/z, respectively, were used. The glucose sodium adduct was used as it resembled more the signal observed by HPLC than any other glucose signals in negative mode found. The correlation between three glucose signals observed by on-line metabolomics and by HPLC was compared by Pearson correlation in Figure 3.14. Specifically, the glucose fragment of 71.0136 m/z and the $[glucose-H]^-$ adduct of 179.0561 m/z, both in negative mode, had Pearson correlation estimates of 0.81 and 0.75 to the glucose HPLC results, respectively. The correlation for these negative signals is particularly inaccurate for some data points of low real-time metabolomics intensity. This could be caused by ion suppression and detector saturation during the succinate production phase. The $[glucose+Na]^+$ positive mode adduct of 203.0527 m/z showed a Pearson correlation estimate of 0.86 to the glucose HPLC data, and the data points are scattered more randomly around the correlation line.

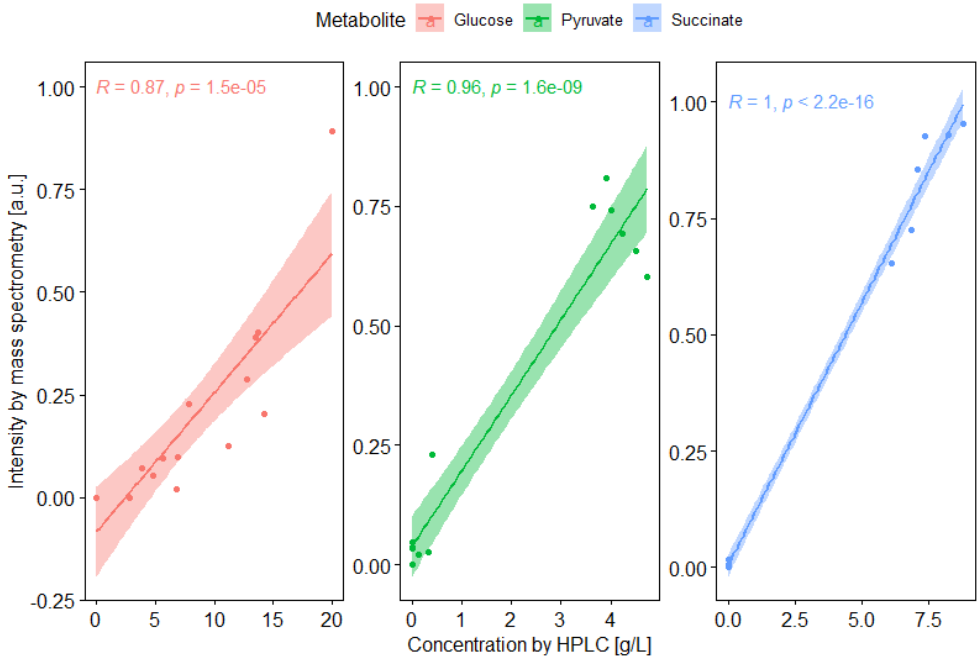



**Figure 3.13:** Pearson correlation between HPLC and on-line metabolomics data for glucose, pyruvate and succinate, where R is the Pearson correlation coefficient and p shows the p-value of the test.

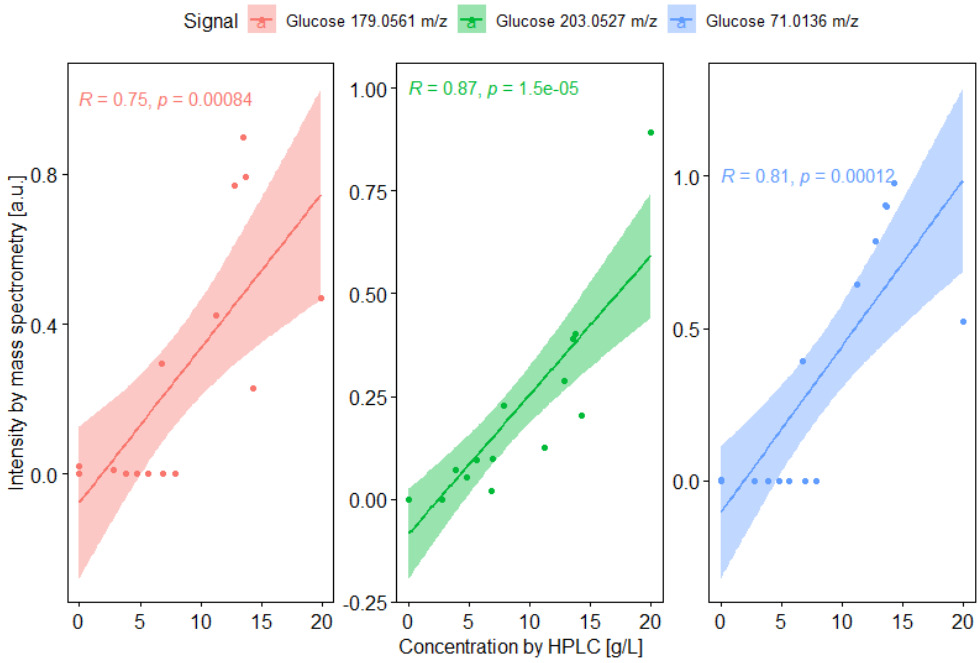



**Figure 3.14:** Pearson correlation between the glucose HPLC data and the on-line metabolomics data for three signals annotated as glucose: the $[glucose\text{-}H]^-$ adduct of 179.0561 m/z, the $[glucose+Na]^+$ adduct of 203.0527 m/z and the glucose fragment of 71.0136 m/z. R is the Pearson correlation coefficient and p shows the p-value of the test.

## 3.3 Discussion

The main aims of this chapter were to establish an on-line metabolomics system for monitoring the liquid phase of fermentation processes and to use the data generated with this system to obtain valuable information of the process. A system that could be applied for the entire duration of an *E. coli* succinate fermentation was developed and 886 m/z signals were monitored for 32 hours. The different elements of the on-line system were described (peristaltic pump, 6-port valve, 10-port valve and tubing adapter) and a washing step was implemented between sample injections in order to prevent accumulative signal loss and instrumental blockages. Finally, the impact of different solvent systems and metabolite annotation were also investigated.

### 3.3.1 On-line metabolomics for liquid-phase analysis

Previous examples of using mass spectrometry for on-line monitoring fermentation processes have been almost exclusively implemented for the gas-phase and, therefore, limited to gases and volatile species (Table 1.2). Although some biological systems produce a very wide range of volatile compounds (for instance, Tejero Rioseras *et al.* (2017) monitored 636 m/z signals in a *S. cerevisiae* fermentation), metabolic reactions occur in the liquid phase, and some

metabolites have very low volatility. This is the case for sugar phosphates (Rende *et al.*, 2019), which were detected in the liquid phase using the developed on-line metabolomics system during succinate production. These observations would have probably been missed using a system for monitoring the gas-phase.

A parameter commonly monitored in bioprocess analysis is the dissolved oxygen (DO) in the broth. The DO provides valuable information of the metabolic state of the cell, to the point that some fermentation processes use the DO as the main criteria for monitoring and control (Akesson *et al.*, 2001; Z. J. Wang *et al.*, 2010). However, as shown in Figure 3.1, DO readings can change rapidly if the aeration or agitation conditions change, potentially leading to misinterpretations of the state of the fermentation. For instance, Figure 3.1C showed the DO spike up at around 6.60 and 7.85 h. With only this DO reading, it is unclear if the cells had starved on glucose at the first DO spike. It is possible that this double spike could represent a switch from an exhausted primary substrate (glucose) to a secondary substrate such as acetate, which is accumulated under aerobic conditions with glucose excess due to overflow metabolism and the Crabtree effect (Akesson *et al.*, 2001; De Mey *et al.*, 2007; Wolfe, 2005; B. Xu *et al.*, 1999). With on-line metabolomics, it is possible to have a direct measurement of glucose and know exactly when it is depleted, as shown in Figure 3.12 (around 2 h before the beginning of the production phase).

The high time-resolution of the on-line metabolomics system developed allows accurate determination of the timings for specific metabolic shifts, which could easily be otherwise missed. For instance, Figure 3.9 shows rapid spikes or drops in specific metabolites, which could correspond to fast metabolic switches. Some of these fast changes can be observed at the time when glucose is depleted during the aerobic biomass formation phase, as well as when the anaerobic succinate production phase starts, including rapid increases in succinate and pyruvate abundance. Gathering this information in real-time can be instrumental for process monitoring and control. For instance, observing the depletion of the carbon source could trigger the start of a glucose feed in a fed-batch process, ensuring a good process productivity and avoiding the accumulation of undesired by-products resulting from glucose starvation.

Furthermore, being able to analyse the evolution of many different compounds can help discover potential metabolic bottlenecks and targets for genetic engineering. For instance, the intensity of sedoheptulose-7P and glucose-6P was observed to rapidly increase at the start of the succinate production phase. This kind of information could direct strategies of strain and process engineering, as it will be discussed in more detail in Chapters 4 and 7.

### 3.3.2 Metabolite annotation

Real-time monitoring systems are characterised by their speed of analysis: the fermentation is analysed in a short time window, leading to the possibility to react quickly to changes of the system. In the on-line monitoring system described in this chapter, speed comes with the removal of the chromatographic separation step. On top of reducing the measuring time, omitting the chromatography also simplifies the system. It removes an element that could lead to potential blockages – due to the high concentrations of biomass and salts in the samples – and that often needs optimisation. However, the caveat of not having a chromatographic step is that all compounds are analysed at the same time, thus relying on accurate mass and/or fragmentation information to distinguish the different metabolites. Even in these cases, isobaric and isomeric compounds might be indistinguishable. Other authors have acknowledged this limitation using flow injection mass spectrometry, but still claim the biological relevance and usefulness in the data generated, even when the samples analysed do not belong to time-course experiments (Fuhrer *et al.*, 2011; Fuhrer and Zamboni, 2015). For instance, the information obtained can be used to indicate specific metabolic pathways of potential interest for follow-up experiments. In time-course experiments – such as fermentation monitoring – the dynamic evolution throughout the course of the fermentation adds a new dimension of information for annotating any given m/z signal, increasing the confidence of the annotations.

### 3.3.3 Solvent system selection

Two experiments were performed to assess the effect of the solvent system used in the on-line metabolomics platform. Although the first experiment was run at a biomass concentration nearly 10 times lower than the maximum values obtained in the bioreactor, the mere presence of biomass and the profile of metabolites characteristic of the succinate process made it useful to assess the effect of different solvent mixtures on critical operational parameters of the on-line metabolomics system. The validity of this assertion can be appreciated by observing the similarity of these results with the ones obtained using a bioreactor (Figure 3.6 and Figure 3.7).

In both experiments, the aim was to use a culture that would change in time as little as possible, in order to test the different solvent mixtures under the same conditions. The ideal case scenario would have been to test these with a culture at steady state, for instance using a chemostat or a turbidostat. However, these setups were not readily available to the experimenters. Instead, it was decided to use a culture at anaerobic succinate production phase, mainly because the *E. coli* strain used in this bioprocess does not grow anaerobically on glucose, as observed experimentally (Figure 3.1). The lack of ability to grow anaerobically on glucose has also been reported in the literature for the analogous *E. coli* strain NZN111 strain by H. Wu *et al.* (2007) and Y. Liu *et al.* (2011) using a similar dual-phase succinate process.

The purpose of using a culture with non-growing biomass is to avoid significant changes in the metabolic profile of the culture caused by events such as substrate starvation (either glucose or oxygen), or by differences in biomass levels at the beginning and the end of the experiment. These changes could drastically change the parameters assessed in the experiment – especially the TIC and number of metabolites detected – and make more difficult the interpretation of the results. In the bioreactor, it is much easier to be confident about reaching anaerobic conditions, as it is possible to actively sparge $CO_2$, and monitor the dissolved oxygen. In a Schott bottle, however, it is harder to control an anaerobic environment. For this reason, the culture had been incubating for 18 h in the sealed Schott bottle before starting the test, in order to try to use all the oxygen present in the bottle. The presence of succinate at the start of both experiments shows that anaerobic conditions had been reached by that point.

A total of 13 different solvent mixtures were tested with the on-line monitoring system. These were combinations of ACN, IPA, MeOH and $H_2O$ in different proportions, and AA, AC and FA were added to some mixtures to adjust the pH and help ionisation of specific compounds. Other authors have used similar mixtures for untargeted flow injection mass spectrometry analysis. Fuhrer *et al.* (2011) used 60:40 IPA:H2O with 5 mM AC at pH 9 for negative mode analysis and 60:40 MeOH:$H_2O$ with 0.1 % FA at pH 3 for positive mode. Link *et al.* (2015) used 60:40 IPA:$H_2O$ with 1 mM ammonium fluoride at pH 9 for negative mode analysis. Beckmann *et al.* (2008) described that different MeOH:$H_2O$ mixtures ranging from 30-70 % MeOH have been used for flow injection mass spectrometry, recommending a 50:50 MeOH:$H_2O$ mixture for soft leaf tissue extracts analysis. As a last example, Vaidyanathan *et al.* (2002) used a 50:50 ACN:$H_2O$ mixture with 10 mM FA for a high-throughput bacterial identification method. All solvent mixtures described above are given in volume:volume ratios.

Despite the variety of solvents used for flow injection analysis in the works mentioned above, the authors did not justify their solvent mixture of choice or provide details of any solvent optimisation efforts. In fact, Beckmann *et al.* (2008) even reckoned that the solvent of choice does not directly affect ionisation of the sample in flow injection electrospray mass spectrometry, due to a plug flow effect (very low axial mixing across the injection tube). Instead, the authors claim that the largest effect of the solvent of choice is for preventing the build-up of contaminants in the lines. In this work, however, the different solvent mixtures used showed a significant effect in the ionisation of glucose, succinate and in the total ion chromatogram in positive and negative mode (Figure 3.6 and Figure 3.7). This effect might be caused by a higher degree of turbulence and mixing at higher flow rates, as described by the dimensionless Reynolds number (equation 3.1).

$$Re = \frac{\rho u W}{v} \tag{3.1}$$

where $\nu$ is the viscosity of the fluid, $\rho$ is the density of the fluid, u is the mean inflow velocity of the fluid and W is the diameter of the pipe. The Reynolds number is a common engineering parameter used to model the turbulence of fluids (Davidson and Harvie, 2007. Specifically, the on-line monitoring system used in this work was operated at significantly higher flow rates (600 $\mu$L/min) than what Beckmann *et al.* (2008) reported (100 $\mu$L/min).

As mentioned in Section 3.2.2, depending on the parameter selected, a different solvent mixture might be considered best. Therefore, with no solvent system clearly superior to the rest, a pragmatic approach was followed and solvent system 1 was chosen based on the low solvent pump working pressure observed. Alternatively, an objective function including all the eight different parameters analysed (glucose and succinate intensity, TIC on positive and negative mode, aqueous and solvent pump pressures and number of m/z features detected on positive and negative mode) could have been used to find the optimal solvent system. This approach can incorporate weighting factors to give more importance to specific parameters (for instance, the aqueous and solvent pump pressures). Van Poucke *et al.* (2005) used a similar approach to find the optimum mobile phase ratio of MeOH, $H_2O$ and FA in order to simultaneously analyse five antibacterial growth promoters (bacitracin, olaquindox, spiramycin, tylosin and virginiamycin) by LC-MS. The authors used a weighing factor to one of the five analytes to increase its relative contribution to the objective function. Although this method can result effective, the final results will depend on the value of the weighting, which often has an arbitrary value.

### 3.3.4 Introducing a washing step between sample injections

The study of the implementation of a washing step could have been done in a more systematic manner. Looking at Table 3.1, it can be appreciated that often two experiments are compared where at least two parameters were changed at the same time. This significantly limits the understanding of the main effects studied on the observable factor (*i.e.* the duration of use of the on-line monitoring tool without blockages). A formal design of experiments (DoE) approach might have resulted in better use of time and resources. However, it must be borne in mind that the economic and time requirements of performing each of these experiments limit the number of runs that can be feasibly performed.

A potentially better design is proposed in Table 3.3. Note, however, that a full factorial design with three factors at two levels each would result in $2^3$ = 8 experiments, which is already close to the nine experiments that were run. Furthermore, this would not include any replicates or mid-points and it considers many of the other important factors that were tested as constant factors, such as the sample:wash ratio, the duration of the wash and the implementation of the washing step in the 6-port valve. The addition of any of these as a variable factor would double the number of experiments in a full factorial design ($2^n$), therefore, a potential fractional factorial design using four variable factors ($2^{4-1}$ = 8 experiments) could be suggested. This

would lead to a level IV resolution ($2^{4-1}_{IV}$), hence being able to estimate the main effects without aliasing with other main effects or with two-factor interactions, although two-factor interaction effects would be aliased with each other (Montgomery, 2012). However, this final number of 8 experiments does not include any pilot experiments that are usually carried out before finalising the design in order to avoid unforeseen issues or limiting the high and low levels to sensible parameters. For instance, using a flow rate of 600 $\mu$L/min could have been excessive and led to overpressurising the system, or another good example could be using a sample:wash ratio of 1:1 to prevent signal loss, as it was empirically observed (see Figure 3.8). Failing to check these aspects beforehand could lead to the complete waste of the experimental design. Therefore, a few pilot experiments would need to be added to either the $2^3$ or $2^{4-1}$ designs, thus exceeding the nine experiments described in Table 3.1.

**Table 3.3:** Full factorial design with three variable factors at two levels each ($2^3$ design) proposed for more systematically investigating the parameters of a washing step affecting the duration of use of the on-line metabolomics system without blocking. A is $H_2O$, B is 70:30 IPA:ACN + 0.1 % FA and C is 70:30 ACN:IPA + 0.1 % FA. Four factors are left as constant and therefore do not have low and high levels.

| Factor | Type of factor | Low level | High level |
|---|---|---|---|
| Flow rate of 6-port and 10-port valves | Variable | 200 $\mu$L/min | 600 $\mu$L/min |
| Frequency of wash | Variable | 5 $min^{-1}$ | 30 $min^{-1}$ |
| Solvent combination to clean the 10-port valve and the mass spectrometer side of the system | Variable | Mix B + Mix C | Mix A + Mix C |
| Sample:wash ratio | Constant | 1:1 | |
| Duration of the washing step | Constant | 4 min | |
| Six-port valve wash | Constant | Yes | |
| Duration of each solvent mix for both 6-port and 10-port valves | Constant | 2 min + 2 min | |

As far as the prior work is concerned, similar on-line methods described in the literature tend not to describe the use of a washing step. Instead, the use of filtration systems are more extended. The only example found to use a washing step is the work by Tohmola *et al.* (2011), who described an on-line HPLC system for monitoring *S. cerevisiae* fermentations every five minutes. Their system contained a sample collector and a cross-flow filter to remove the cells prior to injection. Between samples, the tubing lines were cleaned with 70 % ethanol and the cross-flow filter was cleaned with 0.9 % saline.

As mentioned earlier, on-line methods solely based on sample filtration are more commonly reported. For instance, Y.-C. Liu *et al.* (2001) reported a method of at-line HPLC monitoring of a *Zymomonas mobilis* ethanol fermentation. The system employed a commercial robotic BenchMate II device for the automatic pipetting of fermentation broth, filtration of the biomass through a 0.5 $\mu$m filter, dilution of the filtrate and injection of the diluted sample into an HPLC.

Another example was reported by Plum and Rehorek (2005), who used a system for on-line monitoring of a combined anaerobic and aerobic wastewater treatment process by HPLC coupled to a diode array detector and tandem mass spectrometry. The system consisted of two 30 L membrane reactors, respectively operated anaerobically and aerobically, with sludge immobilised in polyurethane foamed carriers. The liquid was continuously pumped through an ultrafiltration membrane of 2500 Da molecular weight cut-off, and the cell-free filtrate was automatically degassed and analysed by HPLC. Although the authors did not report a washing step between samples, the immobilisation of biomass and the use of continuous ultrafiltration would clean the sample of biomass and particulates, thus reducing the chances of blockages. The sampling frequency was not reported by the authors, but HPLC methods between 15 and 40 minutes were suggested by the authors, and the results of a 22-hour fermentation show stages where the analytes concentrations were measured many times a day, and stages where no analytical data is reported for a couple of days in a row. Due to the lack of detail, it is unclear if the periods of lower sampling frequency are due to technical issues such as the blockage or maintenance of the sampling system. On an earlier version of the same analytical system, Rehorek *et al.* (2002) reported that for one particular bioreactor run the system operated for more than 100 h without technical issues such as pressure increases or clogging, suggesting that the system occasionally suffers from blockages, especially after many hours of operation.

Other examples using similar filtration strategies are found in the literature (Koliander *et al.*, 1990; Warth *et al.*, 2010) and, although these did not report the incorporation of a washing step, both used guard columns to remove salts and impurities.

Link *et al.*, 2015 used on-line mass spectrometry to monitor metabolic switches between starvation and growth in *E. coli* cultures without filtering the cells from the sample. Although the authors did not report the use of any washing step, they grew the cells in eight-times diluted cultivation medium and only used the system for 150 minutes. Diluting the cultivation medium greatly reduces the concentration of salts, which can induce the accumulation of particulates in the on-line system and cause ion suppression in the mass spectrometer, but is not a viable option for monitoring fermentation processes.

In conclusion, previous authors have mostly avoided the use of a washing step by incorporating a filtration step in the on-line sampling system, or by diluting the cultivation medium. Filtering the cells limits the analysis to extracellular metabolites and culturing the cells in diluted medium is not a viable strategy in manufacturing scenarios. The strategy of including a washing step between samples described in this chapter allowed the on-line monitoring of unfiltered and undiluted *E. coli* fermentation broth samples for more than 45 h without blocking the system.

## 3.4 Conclusions

In this chapter, the application of real-time metabolomics has been demonstrated for on-line monitoring of an *E. coli* succinate fermentation process and the different elements of the system were described. It has been shown how the choice of solvent system can have a substantial effect in the amount of m/z ions observed, the TIC intensity, and it will condition which molecular species will ionise better or worse. Last but not least, physicochemical characteristics of the solvent system, such as the polarity and viscosity, will greatly condition the back-pressure of the pumps that keep the on-line metabolomics system functioning, both directly (due to the liquid properties) and indirectly (by their ability to maintain the lines clean of particulates). Similarly, it was empirically observed how important it is to use a washing step between injections in order to help prevent blockages of the real-time metabolomics system when using unfiltered samples, as well as to prevent signal loss of the mass spectrometer due to accumulating of dirt in the instrument.

It has also been exemplified that on-line metabolomics allows the measurement of the dynamic evolution of many different signals. These can be used to identify potential biomarkers (such as (R)-2,3-dihydroxy-isovalerate and (S)-2-aceto-2-hydroxybutanoate, which were identified as potential biomarkers for biomass) and by-products, which could be potential targets for strain and process engineering. For example, sugar phosphates from the PPP and compounds of nucleotide metabolism were significantly produced at the beginning of the succinate production phase. This will be further studied in Chapter 4. Moreover, being able to trace the dynamic evolution of a molecule over the entire duration of the experiment can be very helpful to correctly annotate said molecule.

Based on their dynamic patterns, compounds of glycolysis, the TCA cycle, the PPP, nucleotide metabolism, metabolites of anaerobic fermentation, amino acids and biomass biomarkers have been identified as worth monitoring in the bioprocess. Chapter 5 will discuss the development of a targeted method for specifically monitoring some of these compounds during the fermentation process.

## 3.5 Materials and methods

The materials and methods are as described in Chapter 2. Any differences and particular methods used exclusively in this chapter are described below.

### 3.5.1 Solvent system selection

The effect of using different solvent systems for real-time metabolomics was tested in two consecutive experiments. In the first one, the monitoring system was connected to a Schott bottle, and in the second one, a 5 L fermenter was used. A total of 13 different solvent mixtures with different levels of polarity, pH and physicochemical properties were tested for a number of injections (n= 5 to 15) with a succinate producing *E. coli* strain. The 13 solvent mixtures were:

1. 70:30 ACN:IPA + 0.1 % FA, pH 3.00
2. 60:20:20 ACN:IPA:$H_2O$ + 0.1 % AA, pH 3.55
3. 70:30 IPA:ACN + 0.1 % AA, pH 7.06
4. 70:30 IPA:ACN 0.1 % FA, pH 5.65
5. 60:40 IPA:$H_2O$ + 20 mM AC, pH 9.40
6. MeOH, pH 6.65
7. MeOH + 0.1 % AA, pH 3.35
8. MeOH + 20 mM AC, pH 8.97
9. 60:40 MeOH:IPA + 0.1 % AA, pH 2.95
10. 70:30 MeOH:IPA + 0.1 % AA, pH 3.09
11. 70:30 MeOH:IPA + 20 mM AC, pH 8.88
12. 70:30 MeOH:IPA + 0.1 % FA, pH 2.02
13. 60:20:20 MeOH:IPA:$H_2O$ + 0.1 % FA, pH 2.89

#### Schott bottle test

In the first experiment, a 1 L Schott botte containing 600 mL of growth media was inoculated with 10.6 mL of an IGZ006 *E. coli* overnight culture for an $OD_{600}$ of 0.088. This was incubated with the lid closed at 37 ℃ and 165 rpm for 18 h to an $OD_{600}$ of 1.23 +/- 0.024 CI 95 %. The Schott bottle was connected to the real-time metabolomics system and automatic samples were taken into an Exactive$^{TM}$ Orbitrap mass spectrometer using the solvent systems listed above, with the exception of number 1 and 11, which were excluded for technical reasons. Every time that a solvent mixture had to be changed, the lines of the real-time metabolomics system were cleaned with a 9-minute washing step as described for experiment 5 (MethDev 4) in Table 3.1, with the difference that in this case mix C was the specific solvent system tested rather than methanol. Furthermore, the line carrying the solvent was purged for 5 minutes at every change of solvent mixture.

### Fermentation test

In the second experiment, a 5 L fermenter containing 3 L of growth medium was inoculated with 100 mL of IGZ006 *E. coli* overnight culture to an $OD_{600}$ of 0.0071 in the fermenter. The cells were grown under aerobic conditions with 4 L/min air (1.33 vvm), 37 ºC and 600 rpm. After 23.75 h, the $OD_{600}$ had reached 8.47 +/- 0.096 CI 95 %, close to the maximum levels of biomass seen in the succinate production process studied. Once the culture had reached an age of 26 h, 50 mL of 500 g/L glucose were added to the fermenter as a bolus addition for an estimated concentration of 10 g/L in the fermenter, the air was turned off and the agitation was turned down to 150 rpm to start the anaerobic succinate production phase. After 14.92 h of anaerobic succinate production phase, the fermenter agitation was set to 500 rpm and it was connected to the real-time metabolomics system. Automatic samples were taken into an Exactive$^{TM}$ Orbitrap mass spectrometer using solvent mixtures 1, 3, 4, 6, 10 and 11, based on the results from the Schott bottle test. Every time that a solvent mixture had to be changed, the lines of the real-time metabolomics system were cleaned with a 9-minute washing step as described for experiment 5 (MethDev 4) in Table 3.1, with the difference that in this case mix C was the specific solvent mixture tested rather than methanol. Furthermore, the line carrying the solvent mixture was purged for 5 minutes at every change of solvent mixture.

### On-line metabolomics data processing for solvent system selection

The results from the solvent system selection experiment were not analysed using a pre-defined processing method using the Xcalibur$^{TM}$ software, but using the xcms package (Version 3.0.2) in the statistical software environment R (Version 3.4.4). Peak finding was done using the xcmsSet function using the “MSW”method to deal with non-chromatographic MS data, as is the case of the on-line metabolomics monitoring system. However, this method requires one single spectrum and therefore only two scans were analysed for each sample: one for positive and one for negative mode. It was attempted to choose the two scans in the middle of the injection peak (scan 11 for both negative and positive mode in the example illustrated in Figure 3.15). The peak threshold was set to 8000 a.u., amplitude threshold to 1E-04 and S/N set to 3. Peak grouping was done using the “mzClust” method with a relative m/z error of 5 ppm. Peak area integration was done using the “MSW” method. The total ion chromatogram (TIC) for each sample was calculated by adding the intensity of all m/z features detected in the sample. The extracted mass spectrometry data was analysed with box plots using the ggplot2 package (Version 3.3.3; Wickham, 2016) in the statistical software environment R (Version 3.4.4).

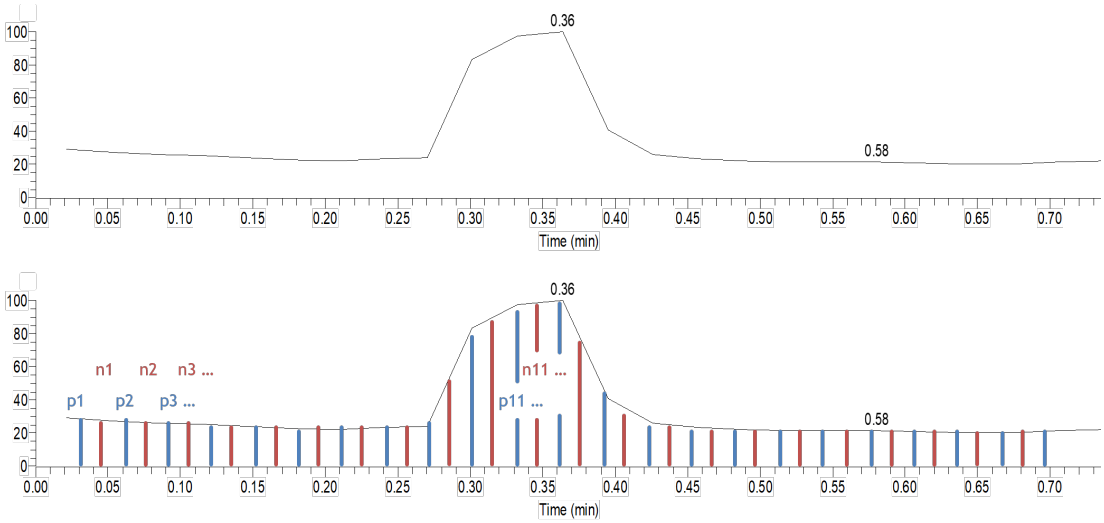



**Figure 3.15:** Example of the signal for a sample analysed by flow injection mass spectrometry (**top**). Positive (p) and negative (n) scans are alternated and represented in blue and red, respectively (**bottom**). Scan 11 is in the middle of the injection peak for both positive and negative mode.

### 3.5.2 Potential adduct annotation

A list of potential adducts was generated by adding isopropanol and water adducts to the list of adducts reported by Huang *et al.* (1999). This list of 42 adduct formulas was applied to 173 metabolites from the main metabolic pathways, generating a 42 x 173 matrix ($M_{42,173}$). Monoisotopic masses were used for the calculation of the theoretical adducts. Adducts were programmatically assigned in the statistical software environment R (Version 4.0.4) by comparing any given list of m/z values against the $M_{42,173}$ matrix and selecting only the matches within 5 ppm mass error.

### 3.5.3 Correlation analysis between HPLC data and on-line metabolomics data

A Pearson test was used to analyse the correlation between the HPLC and real-time metabolomics data of different metabolites. This test was performed using the ggpubr package (Version 0.4.0; Kassambara, 2020) in the statistical software environment R (Version 4.0.4).

Chapter 4

# Liquid chromatography – mass spectrometry analysis of succinate fermentation in *Escherichia coli*

The samples for LC-MS analysis of the three fermentation replicates were analysed on the Q Exactive Orbitrap by Glasgow Polyomics. The computational analysis of the results was carried out by Joan Cortada García.

## Abstract

Chapter 3 described the development of an on-line metabolomics tool and how it can be used for monitoring and enhancing the understanding of fermentation processes. This tool was used to monitor an *Escherichia coli* dual-phase succinate production fermentation process, and metabolites following patterns of potential interest for process monitoring and optimisation were identified. In order to characterise these metabolic signals further, the fermentation process was analysed by off-line liquid chromatography – mass spectrometry (LC-MS). Compared to the on-line metabolomics system developed, LC-MS contains a chromatographic step that separates the compounds prior to analysis. This separation step provides retention time information, which allows an increase in the confidence of compound annotations when these are compared to reference standards.

The analysis by off-line LC-MS allowed more confident annotation of the results obtained with on-line metabolomics, by comparing the time evolution of signals with the same mass-to-charge ratio measured with both methods. This novel approach is only possible with time-course data when enough time resolution is available, highlighting the advantage of on-line metabolomics. Furthermore, the LC-MS results were used to perform multivariate statistical analysis and pathway analysis to identify potential biomarkers of carbon starvation and pathways of interest for strain or process optimisation. Pathway analysis pointed at three highly-expressed pathways, namely the pentose phosphate pathway, ascorbate and aldarate

metabolism and pantothenate and coenzyme A biosynthesis. Of these three, the pentose phosphate pathway was investigated further and compounds belonging to this pathway were observed to rapidly increased at the beginning of the anaerobic succinate production phase, suggesting that it could be a potential route to try to improve succinate production.

## 4.1 Introduction

Liquid chromatography – mass spectrometry (LC-MS) has become the most common technique to generate metabolomics data (Gloaguen *et al.*, 2017). Different types of chromatography columns can be used depending on the nature of the molecules to be analysed. The two types of column most widely used for LC-MS are hydrophilic interaction liquid chromatography (HILIC) and reversed phase liquid chromatography (RP-LC) (Buszewski and Noga, 2012). HILIC is good at separating hydrophilic and amphiphilic compounds, and it uses a hydrophilic stationary phase and an aqueous-organic mobile phase of increasing water content (Creek *et al.*, 2011). Organic solvents that are highly soluble in water are commonly used, such as acetonitrile. RP-LC is used to separate hydrophobic compounds, and it uses a hydrophobic stationary phase and an aqueous-organic mobile phase of increasing organic content (if a gradient is used).

In this chapter, the *E. coli* fermentation process was analysed using LC-MS with a HILIC column, which is a suitable option for a global metabolomics analysis (Creek *et al.*, 2011) and, particularly, for some of the main compounds of the bioprocess, such as compounds from glycolysis, the TCA cycle and anaerobic fermentation.

### 4.1.1 Aims and objectives

In Chapter 3, an *E. coli* succinate production fermentation process was analysed with on-line metabolomics and nearly 900 m/z signals were monitored every 5 minutes. This resulted in a large variety of dynamic metabolic patterns, some of which were indicative of potential targets for bioprocess improvement. The main aim of this chapter is to determine whether LC-MS analysis can provide useful information for bioprocess improvement, especially by complementing the on-line metabolomics results. To do this, three main objectives were set:

- To do a combined targeted and untargeted LC-MS analysis of the succinate fermentation process
- To use the LC-MS results to complement the annotations of the on-line metabolomics experiments from the previous chapter.
- To analyse the LC-MS data with multivariate and pathway analysis to find useful bioprocess biomarkers and potential candidates for bioprocess optimisation.

The LC-MS analysis increases the confidence in annotation compared to the on-line metabolomics results by having a separation step using chromatography. The large amount of data and additional confidence in annotation obtained by LC-MS are instrumental in performing useful multivariate and pathway analysis.

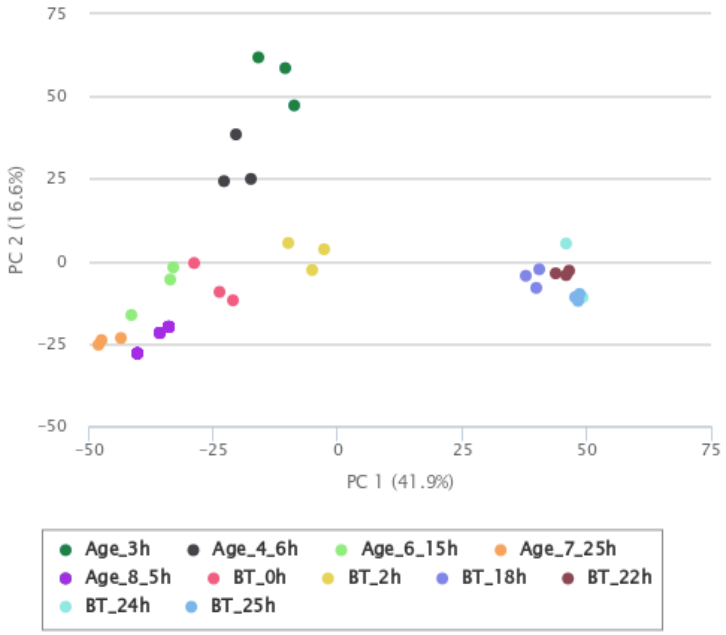
## 4.2 Results

### 4.2.1 *E. coli* succinate fermentation process for LC-MS analysis

The first aim of this chapter was to perform a combined targeted and untargeted LC-MS analysis of the succinate fermentation process. Three *E. coli* dual-phase succinate fermentation replicates were performed, samples were taken throughout the course of the fermentation and were processed and extracted under cold conditions for metabolomics analysis with a mixture of 1:3:1 C:M:W, as described in Section 2.7.1. Extracted samples were analysed by LC-MS in a high-resolution Q Exactive$^{TM}$ Orbitrap mass spectrometer, providing accurate mass to 3 ppm error, chromatographic retention time and fragmentation data.

It was empirically tested that the use of glass beads was not necessary to break the *E. coli* cells and extract intracellular metabolites. This was checked by extracting one sample from the triplicate fermentation experiment using two different extraction protocols – only differing in the use or not of glass beads during the one-hour C:M:W extraction incubation – and looking at the presence, intensity and peak shape of the compounds identified in the sample. The similarity of the results indicated that the use of glass beads was not necessary to extract intracellular metabolites from *E. coli*.

A total of 13 samples from each replicate fermentation were analysed by LC-MS for both intracellular and extracellular fractions. Figure 4.1 shows a PCA plot of the samples for both intra- and extracellular analysis. The triplicate samples cluster very well, being a good quality control indicator of the analysis. Table 4.1 summarises the number of features detected and the corresponding metabolites annotated and identified for both intra- and extracellular fractions. A total of 7334 features were found with unique m/z and retention time (RT) values. To avoid repetitions, m/z values were rounded to 4 decimal points and retention times to the closest 10$^{th}$ second. Of these features, 3549 were found in the intracellular fraction and 4538 in the extracellular fraction, with some peaks showing identical m/z and retention times in both data sets. In the intracellular fraction, 4372 compounds were annotated based on accurate mass and/or fragmentation data and 81 were identified based on matches to reference standards. The number of total compounds assigned might be larger than the number of peaks detected because one single m/z value can be assigned to multiple isobaric compounds, isomers, adducts and fragments, especially when there is no RT information from reference standards.



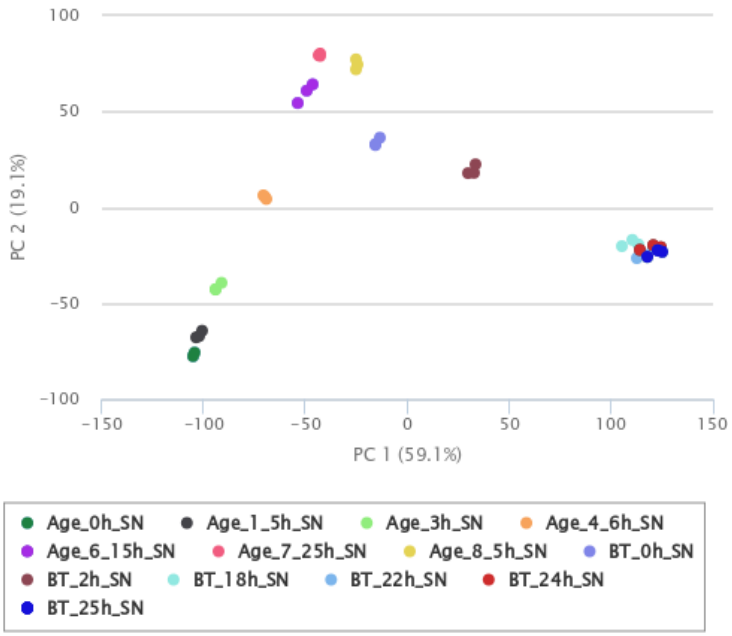




**Figure 4.1:** PCA plot of the triplicate *E. coli* fermentation samples for both intracellular (**top**) and extracellular (**bottom**) fractions. Note that the first two time points for the intracellular fraction were not analysed because the low amount of biomass in the sample made the metabolite extraction impractical. Age denotes the fermentation time during the biomass formation phase, and BT denotes the succinate production phase. Abbreviations: BT: biotransformation; SN: supernatant.

In the extracellular fraction, 71 compounds were identified and 4801 were annotated. The number of identifications is significantly lower than the annotations, as the former require matching results to reference standards and are therefore limited to how many compounds are prepared in the standards mix for targeted analysis.

**Table 4.1:** Summary of the LC-MS results of a triplicate *E. coli* succinate production fermentation process analysed on PiMP. Annotated compounds are based on accurate mass and/or fragmentation data. Identified compounds are matched against reference standards. The union of both intra- and extracellular fractions contains all elements in either fraction without repetitions.

| Parameter | Intracellular fraction | Extracellular fraction | Union |
|---|---|---|---|
| Peaks detected | 3549 | 4538 | 7334 |
| Total compounds | 4453 | 4872 | 6433 |
| Annotated compounds | 4372 | 4801 | 6341 |
| Identified compounds | 81 | 71 | 92 |

### 4.2.2 Complementing on-line metabolomics data with LC-MS

The first reason to do an LC-MS analysis of the fermentation process was to complement the annotations of the untargeted on-line metabolomics monitoring experiments from Chapter 3. To do that, the annotations from both data sets were compared.

On one hand, the LC-MS annotations are based on accurate mass allowing for a 3 ppm mass error window. Additionally, some compounds also have supportive fragmentation data and/or chromatographic retention time matched to a reference standard, increasing the confidence in annotation. On the other hand, the on-line metabolomics annotations are only based in either accurate mass (to 5 ppm error) or fragmentation (for Orbitrap and QQQ experiments, respectively) as well as the dynamic time evolution of each signal to be annotated. However, the latter can only be used when the profile of the metabolite is known beforehand with prior knowledge of the fermentation process, limiting it to key metabolites such as succinate, glucose or pyruvate.

To try to combine the information from both data sets, the 886 different m/z signals in positive and negative mode from untargeted on-line metabolomics monitoring were compared against the 6433 metabolites annotated in PiMP (both intracellular and extracellular fractions data) and matches within 5 ppm m/z error were used as potential metabolite annotations. The time evolution pattern obtained from both platforms (on-line monitoring and LC-MS) was also used as a criterion to evaluate the annotations by visual inspection. That is, m/z matches with completely different time evolution patterns were not considered as potential annotations. Finally, when different possible annotations were possible, priority was given to metabolites belonging to the *E. coli* K-12 MG1655 KEGG library. Although *E. coli* BW25113, was not available in the KEGG library, K-12 MG1655 is phylogenetically closely related to it (Baba *et al.*, 2006).

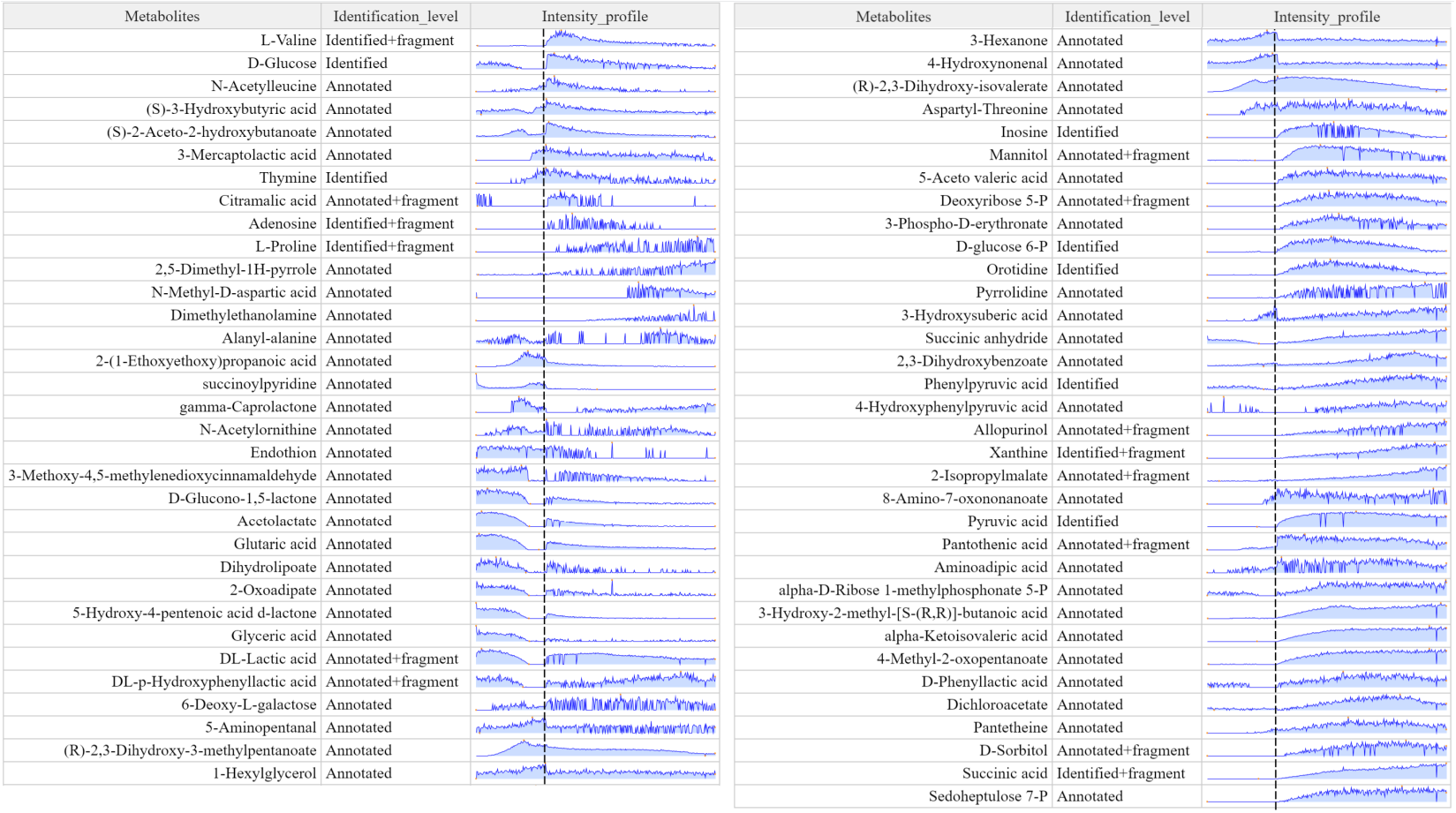
A total of 67 different metabolites were annotated following the procedure above and they are depicted in Figure 4.2. For an easier visualisation, the results were clustered according to similarity of their dynamic profile. Most of the noisier signals are between 2000 and 6000 units of ion intensity, close to the detection threshold commonly used in mass spectrometry, which can sometimes be set as high as 10,000 (Creek *et al.*, 2011). Therefore, these low signals might not be consistently captured in the detection and processing steps. Although targeted LC-MS methods allow for high confidence identification of metabolites, a combined targeted-untargeted method was used for this analysis, and as a result, the metabolite identification level from Figure 4.2 are a combination of annotated and identified compounds.

A few compounds such as D-glucono-1,5-lactone, acetolactate and glutaric acid follow a similar profile to glucose, starting with a high intensity at the beginning of the aerobic phase and gradually decreasing at an exponentially-increasing rate (corresponding to the exponential growth of biomass). At the beginning of the anaerobic production phase, these compounds show a sudden increase in intensity, corresponding to the bolus addition of glucose externally added to the fermentation broth to promote succinate formation.

Thymine, succinoylpyridine, aspartyl-threonine and 3-hydroxysuberic acid are examples of metabolites that start being produced towards the end of the growth phase, when there is glucose limitation.

Another pattern to highlight would be the one followed by metabolites such as pyruvic acid, $\alpha$-ketoisovaleric acid, 4-methyl-2-oxopentanoate, succinic acid and sedoheptulose 7-P, which start increasing immediately at the beginning of the production phase and their production continues until the end of the fermentation.

Finally, (R)-2,3-dihydroxy-3-methylpentanoate, (R)-2,3-dihydroxy-isovalerate and (S)-2-aceto-2-hydroxybutanoate, all belonging to branched-chain amino acid (BCAA) metabolic pathways, increase in intensity exponentially during the aerobic growth phase, showing a good correlation with the measured biomass WCW (Figure 4.3). However, L-proline and L-valine, the two amino acids annotated in the results, do not show this exponential increase during the aerobic phase, even though L-valine is one of the three proteinogenic amino acids from BCAA metabolism (L-isoleucine, L-leucine and L-valine). Instead, L-valine and L-proline were only detected during the anaerobic production phase.



| Metabolites | Identification_level | Intensity_profile |
|---|---|---|
| L-Valine | Identified+fragment | |
| D-Glucose | Identified | |
| N-Acetylleucine | Annotated | |
| (S)-3-Hydroxybutyric acid | Annotated | |
| (S)-2-Aceto-2-hydroxybutanoate | Annotated | |
| 3-Mercaptolactic acid | Annotated | |
| Thymine | Identified | |
| Citramalic acid | Annotated+fragment | |
| Adenosine | Identified+fragment | |
| L-Proline | Identified+fragment | |
| 2,5-Dimethyl-1H-pyrrole | Annotated | |
| N-Methyl-D-aspartic acid | Annotated | |
| Dimethylethanolamine | Annotated | |
| Alanyl-alanine | Annotated | |
| 2-(1-Ethoxyethoxy)propanoic acid | Annotated | |
| succinoylpyridine | Annotated | |
| gamma-Caprolactone | Annotated | |
| N-Acetylornithine | Annotated | |
| Endothion | Annotated | |
| 3-Methoxy-4,5-methylenedioxycinnamaldehyde | Annotated | |
| D-Glucono-1,5-lactone | Annotated | |
| Acetolactate | Annotated | |
| Glutaric acid | Annotated | |
| Dihydrolipoate | Annotated | |
| 2-Oxoadipate | Annotated | |
| 5-Hydroxy-4-pentenoic acid d-lactone | Annotated | |
| Glyceric acid | Annotated | |
| DL-Lactic acid | Annotated+fragment | |
| DL-p-Hydroxyphenyllactic acid | Annotated+fragment | |
| 6-Deoxy-L-galactose | Annotated | |
| 5-Aminopentanal | Annotated | |
| (R)-2,3-Dihydroxy-3-methylpentanoate | Annotated | |
| 1-Hexylglycerol | Annotated | |

| Metabolites | Identification_level | Intensity_profile |
|---|---|---|
| 3-Hexanone | Annotated | |
| 4-Hydroxynonenal | Annotated | |
| (R)-2,3-Dihydroxy-isovalerate | Annotated | |
| Aspartyl-Threonine | Annotated | |
| Inosine | Identified | |
| Mannitol | Annotated+fragment | |
| 5-Aceto valeric acid | Annotated | |
| Deoxyribose 5-P | Annotated+fragment | |
| 3-Phospho-D-erythronate | Annotated | |
| D-glucose 6-P | Identified | |
| Orotidine | Identified | |
| Pyrrolidine | Annotated | |
| 3-Hydroxysuberic acid | Annotated | |
| Succinic anhydride | Annotated | |
| 2,3-Dihydroxybenzoate | Annotated | |
| Phenylpyruvic acid | Identified | |
| 4-Hydroxyphenylpyruvic acid | Annotated | |
| Allopurinol | Annotated+fragment | |
| Xanthine | Identified+fragment | |
| 2-Isopropylmalate | Annotated+fragment | |
| 8-Amino-7-oxononanoate | Annotated | |
| Pyruvic acid | Identified | |
| Pantothenic acid | Annotated+fragment | |
| Aminoadipic acid | Annotated | |
| alpha-D-Ribose 1-methylphosphonate 5-P | Annotated | |
| 3-Hydroxy-2-methyl-[S-(R,R)]-butanoic acid | Annotated | |
| alpha-Ketoisovaleric acid | Annotated | |
| 4-Methyl-2-oxopentanoate | Annotated | |
| D-Phenyllactic acid | Annotated | |
| Dichloroacetate | Annotated | |
| Pantetheine | Annotated | |
| D-Sorbitol | Annotated+fragment | |
| Succinic acid | Identified+fragment | |
| Sedoheptulose 7-P | Annotated | |

**Figure 4.2:** Signals monitored with on-line metabolomics matching the dynamic profile and accurate mass of the results obtained by LC-MS analysis with 5 ppm error. The annotations from the LC-MS analysis were used due to their higher annotation confidence. For easier visualisation, the signals were clustered according to their patterns. The vertical black dashed line indicates the transition from the aerobic growth phase to the anaerobic succinate production phase.

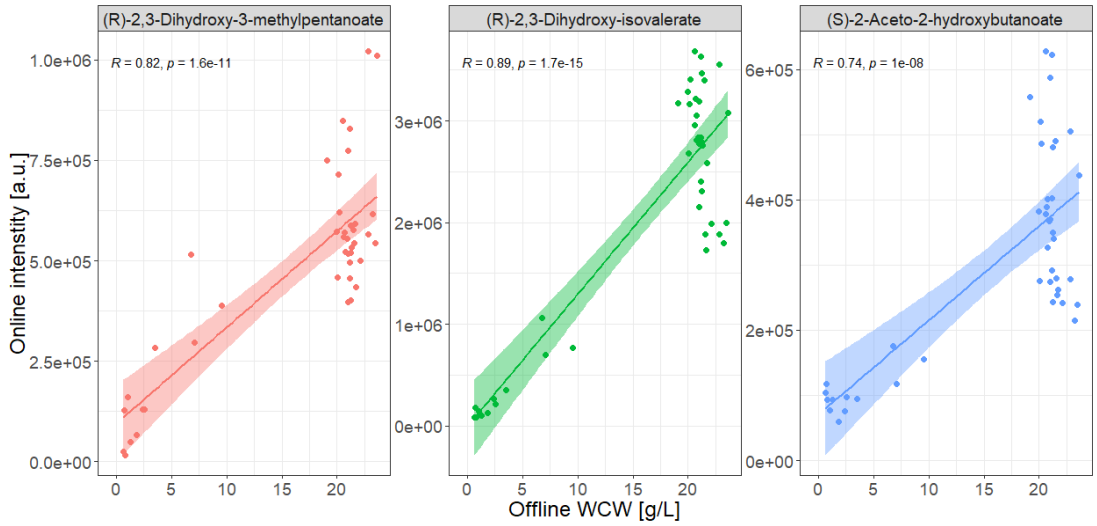



**Figure 4.3:** Pearson correlation between the biomass WCW data measured off-line and the on-line data for the three potential biomass biomarkers annotated as (R)-2,3-dihydroxy-3-methylpentanoate, (R)-2,3-dihydroxy-isovalerate and (S)-2-aceto-2-hydroxybutanoate measured on a triple quadrupole mass spectrometer for a triplicate *E. coli* succinate fermentation experiment. R is the Pearson correlation coefficient and p shows the p-value of the test.

### 4.2.3 Biomarkers of carbon starvation

The second reason to do an LC-MS analysis of the fermentation process was to perform multivariate and pathway analysis to find useful bioprocess biomarkers and potential candidates for bioprocess optimisation.

For bioprocess monitoring reasons, it might be very useful to know whether the cells in the bioreactor have run out of carbon source. For this reason, multivariate statistical analysis was performed to identify biomarkers of carbon starvation. First, the fermentation samples were analysed by HPLC and the glucose concentration was used as a criterium to split the data from the aerobic biomass formation phase into two categories: “unstarved” and “starved” with extracellular glucose concentrations above and below the limit of detection of the method (0.1959 g/L), respectively. Once split into the two categories, the intracellular metabolites detected by LC-MS with a level 1 identification (accurate mass and retention time matched to a reference standard) were analysed with PCA and PLS-DA. It should be mentioned that, as PCA is an unsupervised method, splitting the data into two categories was only necessary for PLS-DA analysis.

### PCA

The PCA scores plot of the samples shows good separation of the two groups (Figure 4.4) and the first two components explain 65.2 % of the variation of the data, with PC1 explaining nearly 50 % (47.1 %) of the total variation (Figure 4.5). The high percentage of variance explained in the first PCs indicates that the metabolites with the highest loadings (Figure 4.6), as absolute values, are those found in most significantly different levels in both sample groups, and therefore these could be good indicators of nutrient availability in the fermentation medium used. Not surprisingly (and possibly a good check), D-glucose and citrate, two substrates added to the fermentation medium, have a high loading score for PC1, having a high contribution to the "unstarved" sample group. Other metabolites with a high loading score for PC1 are cis-aconitate, pyrophosphate and oxalate. On the opposite side of the x-axis, xanthine, pantothenate, L-aspartate, hypoxanthine and UMP are some examples of compounds with a high negative loading score on PC1, thus having a high contribution to the "starved" sample group.

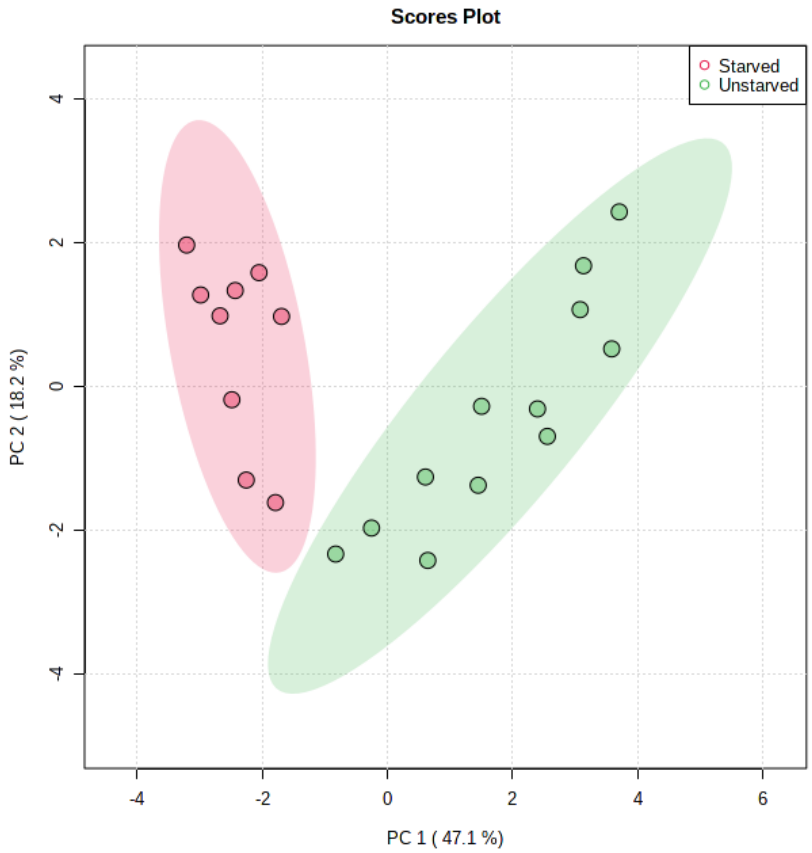



**Figure 4.4:** PCA scores plot PC1 and PC2 for samples during the aerobic biomass formation phase with the respective explained variances shown in brackets. Unstarved and starved groups are annotated in **green** and **red**.

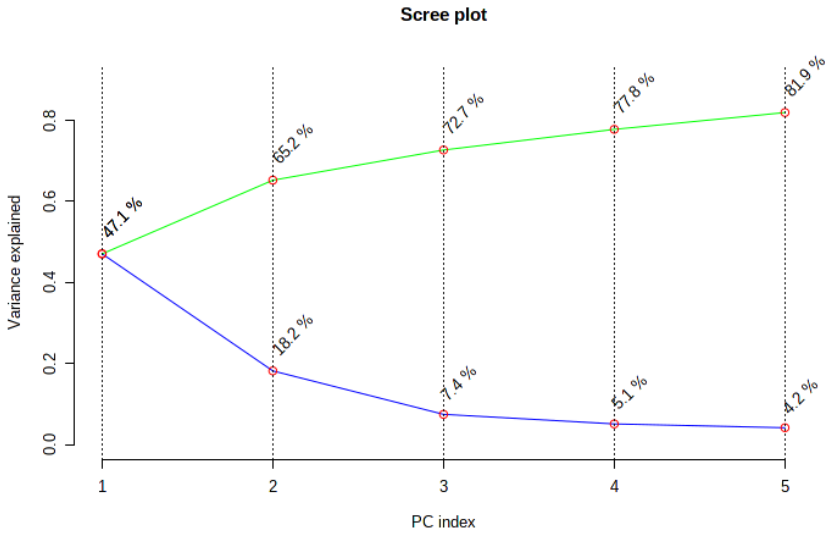



**Figure 4.5:** PCA scree plot for samples during the aerobic biomass formation phase showing the variance explained by each individual PC in **blue**, and the cumulative variance in **green**.

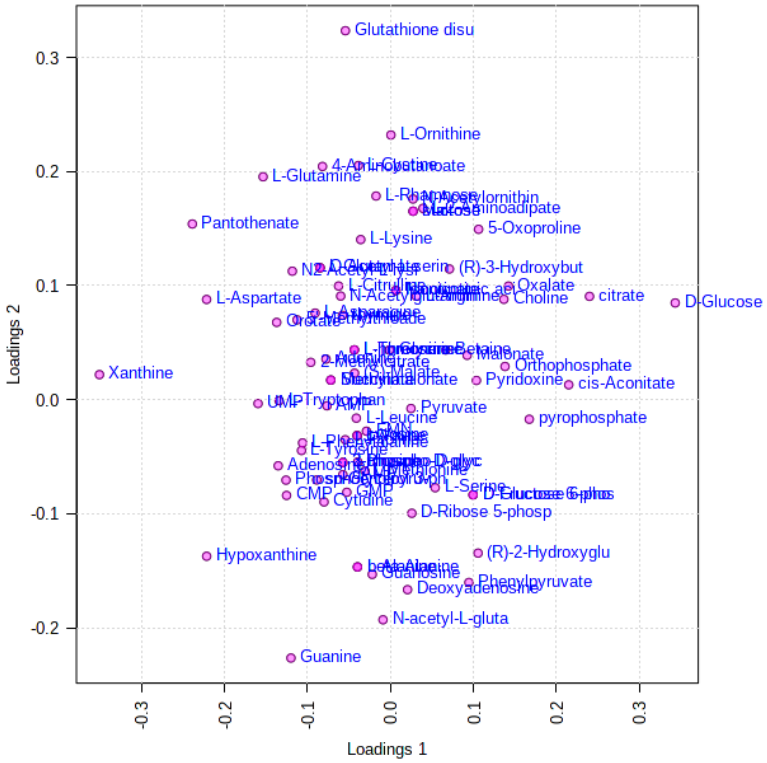



**Figure 4.6:** PCA loadings plot of unstarved and starved samples during the aerobic biomass formation phase. The loadings for PC1 and PC2 are shown on the x- and y-axes, respectively.

### PLS-DA

Very similar results were obtained by PLS-DA. The scores plot shows good separation of the two groups (Figure 4.7), indicating that the most significant latent components could be good indicators of carbon starvation in the fermentation bioprocess used. D-glucose, citrate, cis-aconitate, pyrophosphate and (R)-2-hydroxyglutarate are the five compounds with the highest loading scores on latent component 1 (Figure 4.8 top), and also the compounds with the highest VIP scores among the metabolites more present in the unstarved samples (Figure 4.8 bottom). Xanthine, pantothenate, L-aspartate, hypoxanthine and L-glutamine are the five compounds with the highest negative loading scores on latent component 1, and also the compounds with the highest VIP scores among the metabolites more present in the starved samples. These results are in high agreement with the findings from PCA. All top 15 VIP values were above 1, which indicates high significance in class discrimination.

Model cross-validation determined the optimal number of components to be 5, based on the highest $Q^2$ value, which was close to 0.9, indicating good model predictability in class discrimination (Figure 4.9).

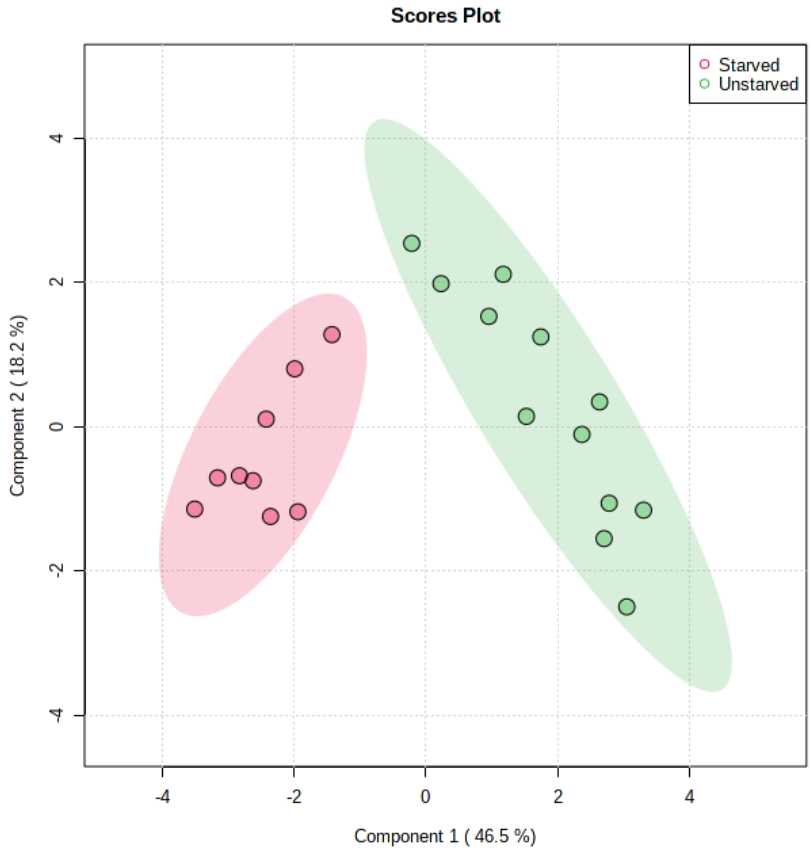



**Figure 4.7:** PLS-DA scores plot of unstarved and starved samples during the aerobic biomass formation phase for latent components 1 and 2 with the respective explained variances shown in brackets.

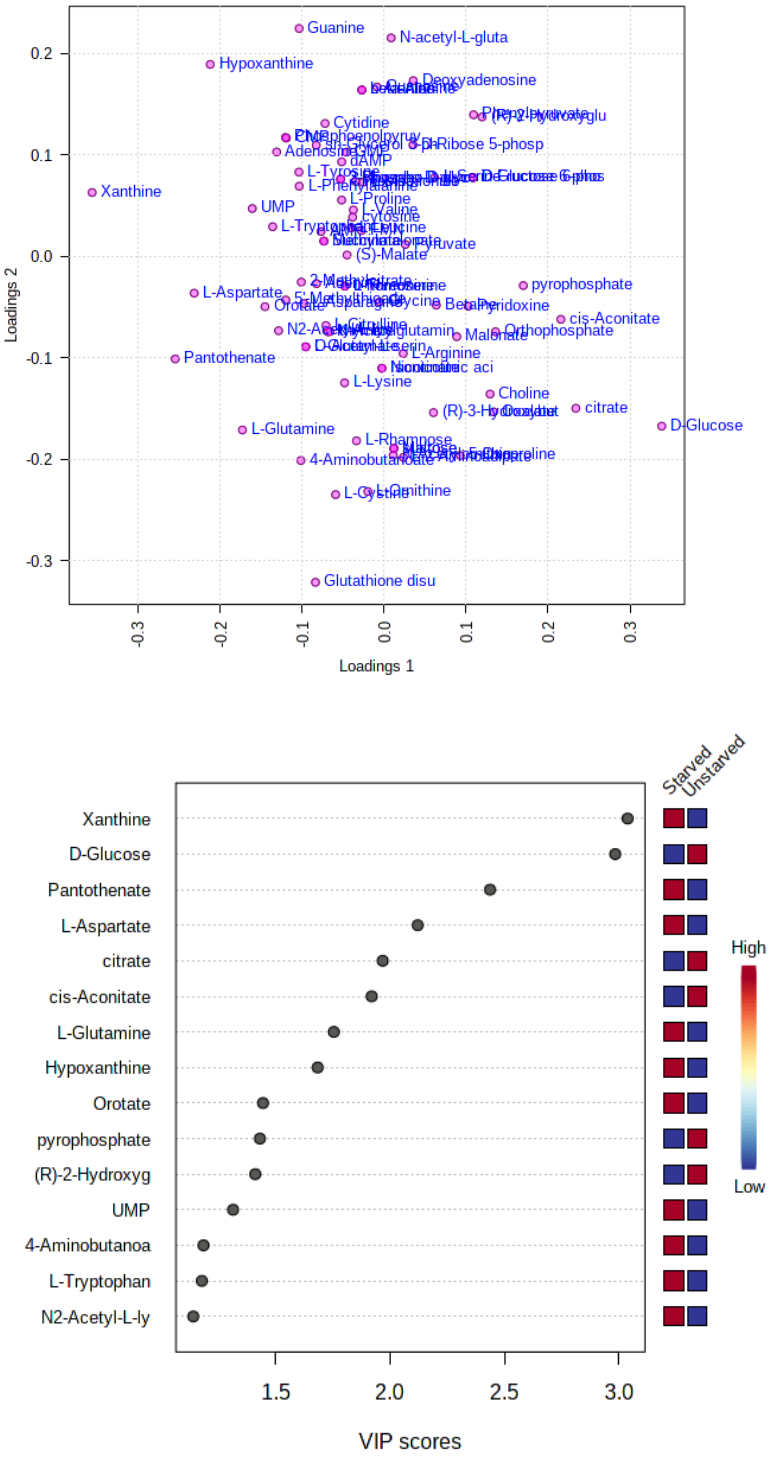



**Figure 4.8:** PLS-DA loadings plot of unstarved and starved samples during the aerobic biomass formation phase (**top**). The loadings for latent component 1 and 2 are shown on the x- and y-axes, respectively. Top 15 important metabolites identified by PLS-DA (**bottom**). The coloured boxes on the right indicate the relative concentrations of each metabolite in each study group.

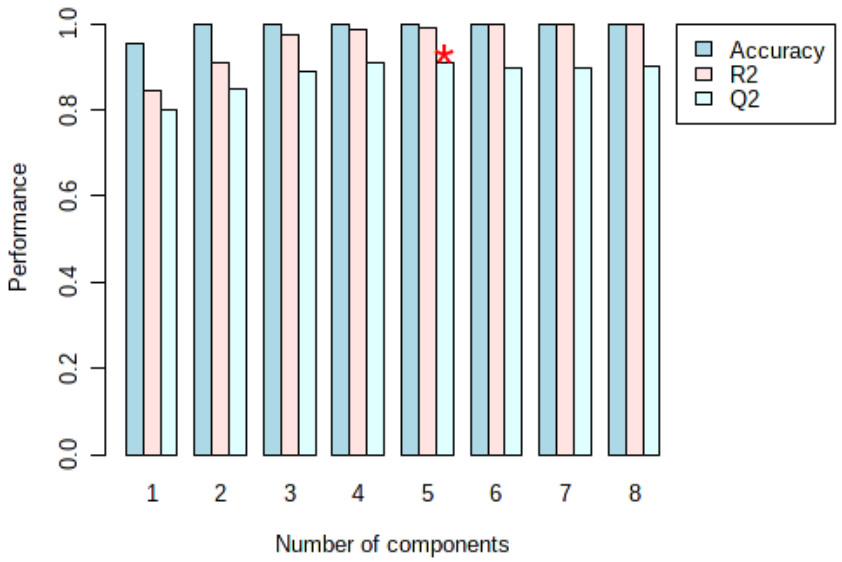



**Figure 4.9:** PLS-DA model evaluation of unstarved and starved samples during the aerobic biomass formation phase based on accuracy, $R^2$ and $Q^2$ for different numbers of PLS-DA components. The red star indicates the best classifier, based on the highest $Q^2$ score.

### 4.2.4 Pathway analysis

As previously mentioned, the second reason to do an LC-MS analysis of the fermentation process was to perform multivariate and pathway analysis to find useful bioprocess biomarkers and potential candidates for bioprocess optimisation. In this section pathway analysis was performed on the LC-MS results to identify potential targets for process optimisation.

Pathway analysis was performed using the PiMP online platform (Gloaguen *et al.*, 2017) using the PALS algorithm and the KEGG library. For every time point of the fermentation, the algorithm evaluates pathways that are significantly different when compared to the first time point. This way, every metabolic pathway assessed is assigned a significance level (p-value) for every fermentation sample (except for the first time point). In order to find the most significant pathways across the time-course samples, an average p-value was calculated for each pathway and the ten ones with the lowest mean p-value were represented as heatmaps in Figure 4.10. These include several pathways from amino acid, carbohydrate, nucleotide and vitamin metabolism.

Lower p-values were obtained for the extracellular fraction samples. This could be because the significance test was calculated at each time point in comparison to the first time point of the fermentation, which in the case of the extracellular fraction would have a rather modest metabolic profile (especially as defined medium was used). Nevertheless, it was not expected to find in the extracellular fraction pathways with essential metabolites such as nucleotides, vitamins and amino acids, as these are not expected to be secreted.

Pathways highly up-regulated during the succinate production phase could be potential targets for metabolic reshuffling. If carbon flux down any of these pathways can be re-channelled towards product formation, this could potentially lead to higher titres or yields. Three pathways were identified to have particularly low p-values during the succinate production phase in the intracellular fraction: the pentose phosphate pathway, ascorbate and aldarate metabolism and pantothenate and CoA biosynthesis (marked with red boxes in Figure 4.10). The criteria to select these three particularly was that they all had at least three time points in the succinate production phase with a p-value $\leq 0.01$ in the intracellular fraction (marked as dark green in Figure 4.10). Out of these three pathways, the pentose phosphate pathway was investigated further as a potential route to try to improve succinate production because of its involvement in the formation of anabolic intermediates required for biomass growth, such as ribose-5P, erythrose-4P and reduced NADPH cofactors (Matano *et al.*, 2014). As there is no biomass growth during the succinate production phase, it might be possible to re-channel some of the carbon flux going to the PPP into the reductive TCA cycle.

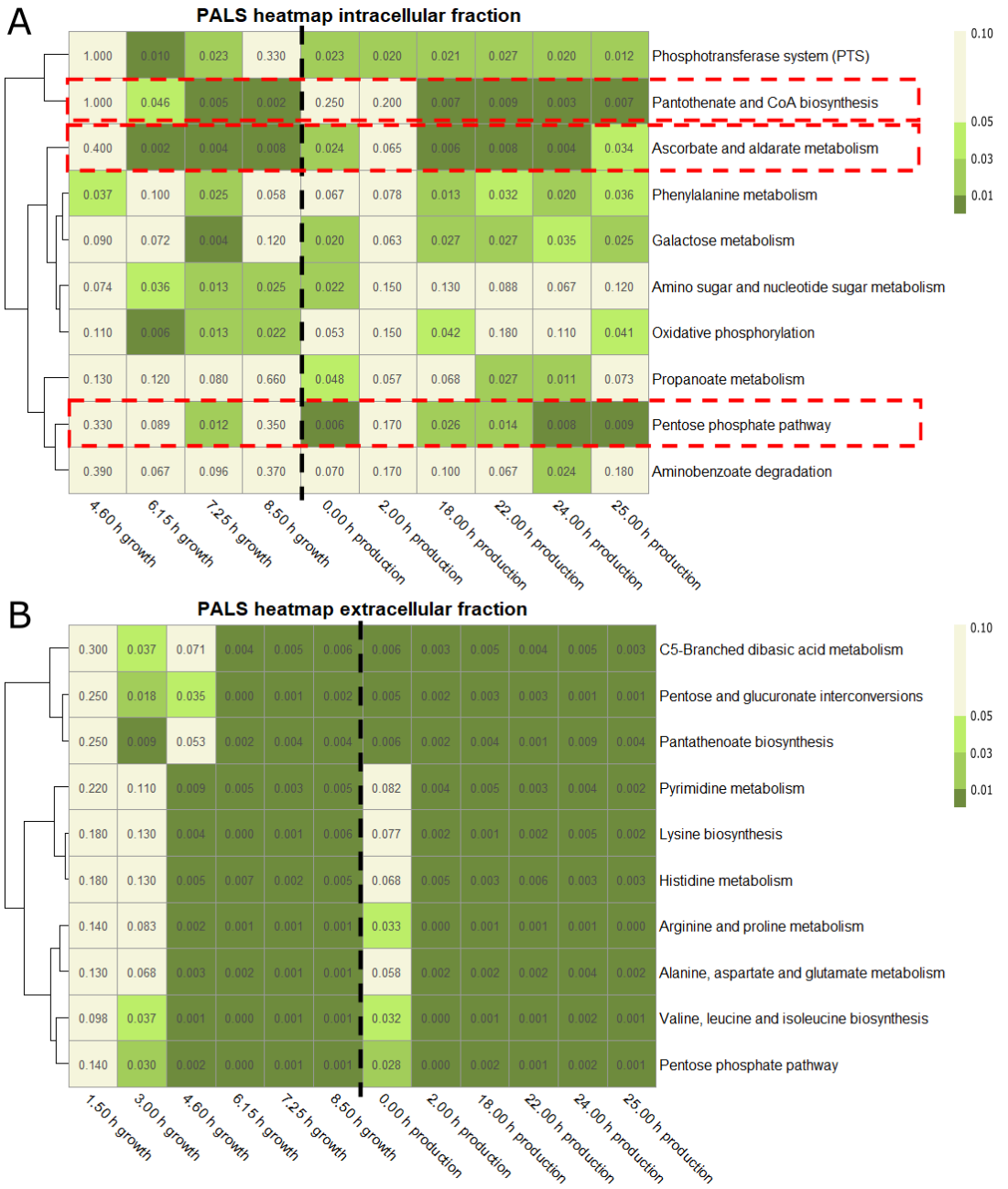



**Figure 4.10:** Top 10 pathways with the lowest average p-value based on pathway analysis using the Pathway Activity Level Scoring (PALS) algorithm for the samples from the intracellular fraction (**A**) and extracellular fraction (**B**). Pathway names are shown on the y-axis (in increasing average p-value going down) and time points of the fermentation process on the x-axis. The values of the cells are the p-values comparing each time point to the first time point of the fermentation. The resulting colour corresponds to the level of significance: green (p-values $\leq$ 0.05) and beige (p-values > 0.05). The vertical black dashed line indicates the transition from the aerobic growth phase to the anaerobic succinate production phase. The horizontal red dashed lines indicate the three pathways highlighted as potential targets for strain engineering.

### 4.2.5 Pentose phosphate pathway

The pentose phosphate pathway (PPP) was identified by pathway analysis as a potential target for bioprocess optimisation. The LC-MS results for the individual metabolites of the PPP were analysed for the intracellular and extracellular fractions (Figure 4.11 and Figure 4.12, respectively).

The intracellular LCMS fraction shows high levels of intermediates from the Embden-Meyerhof-Parnas (EMP) pathway and the PPP during both the aerobic cell growth phase and the anaerobic succinate production phase (Figure 4.11). During the aerobic growth phase, the intracellular levels of most metabolites from these two pathways are relatively high and start decreasing towards the end of the growth phase, when extracellular glucose starts to be depleted. During the anaerobic succinate production phase there is a rapid increase in the intensity of metabolites from both the EMP and PPP pathways after additional extracellular glucose is added, indicating metabolic activity in both directions. Throughout the production phase, the intracellular level of glucose, erythrose 4-P, xylulose 5-P, ribulose 5-P, ribose 5-P and glyceraldehyde 3-P remains high, while the level of glucose 6-P, fructose 6-P and sedoheptulose 7-P drops down after the initial increase in intensity and then stays low.

The extracellular fraction shows low levels of PPP intermediates during the aerobic growth phase, with appreciable intensity levels starting to appear after 4.6 h of cellular growth. This is followed by an increase in intensity during the anaerobic production phase. Regarding the EMP metabolites, extracellular glucose 6-P and fructose 6-P follow a similar decreasing patter to glucose during the growth phase, although – unlike glucose – their levels increase during the production phase. Extracellular glyceraldehyde 3-P was detected at relatively low levels during the growth phase and much higher levels during the production phase.

For bioprocess optimisation purposes, it is worth highlighting the rapid increase in intensity of the PPP metabolites in the intracellular fraction at the beginning of the succinate production phase, as well as the accumulation of PPP metabolites extracellularly towards the end of the succinate production phase.

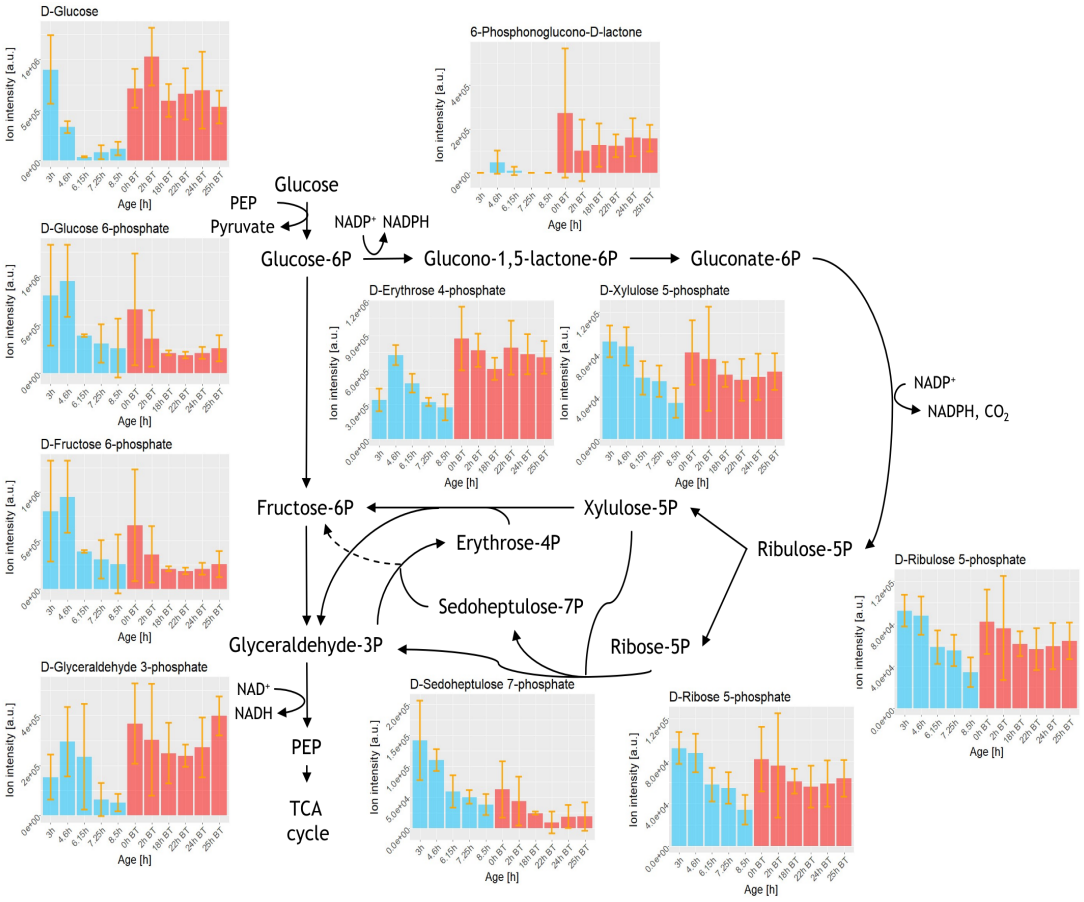



**Figure 4.11:** Dynamic evolution of **intracellular metabolites** of the Embden-Meyerhof-Parnas and pentose phosphate pathway during the bacterial growth phase (**blue**) and succinate production phase (**red**). Orange error bars represent the 95 % confidence intervals. Both phases are initiated with a glucose addition. During the growth phase, there is an increase of all intracellular metabolites coupled to cell growth, followed by a signal drop corresponding to cell starvation. In the production phase, there is a fast increase in the intensity of metabolites from both EMP and PPP pathways after glucose is added. Ribose-5P, ribulose-5P and xylulose-5P are isomers indistinguishable with the method used, hence they have the same pattern. The same applies for glucose-6P and fructose-6P.

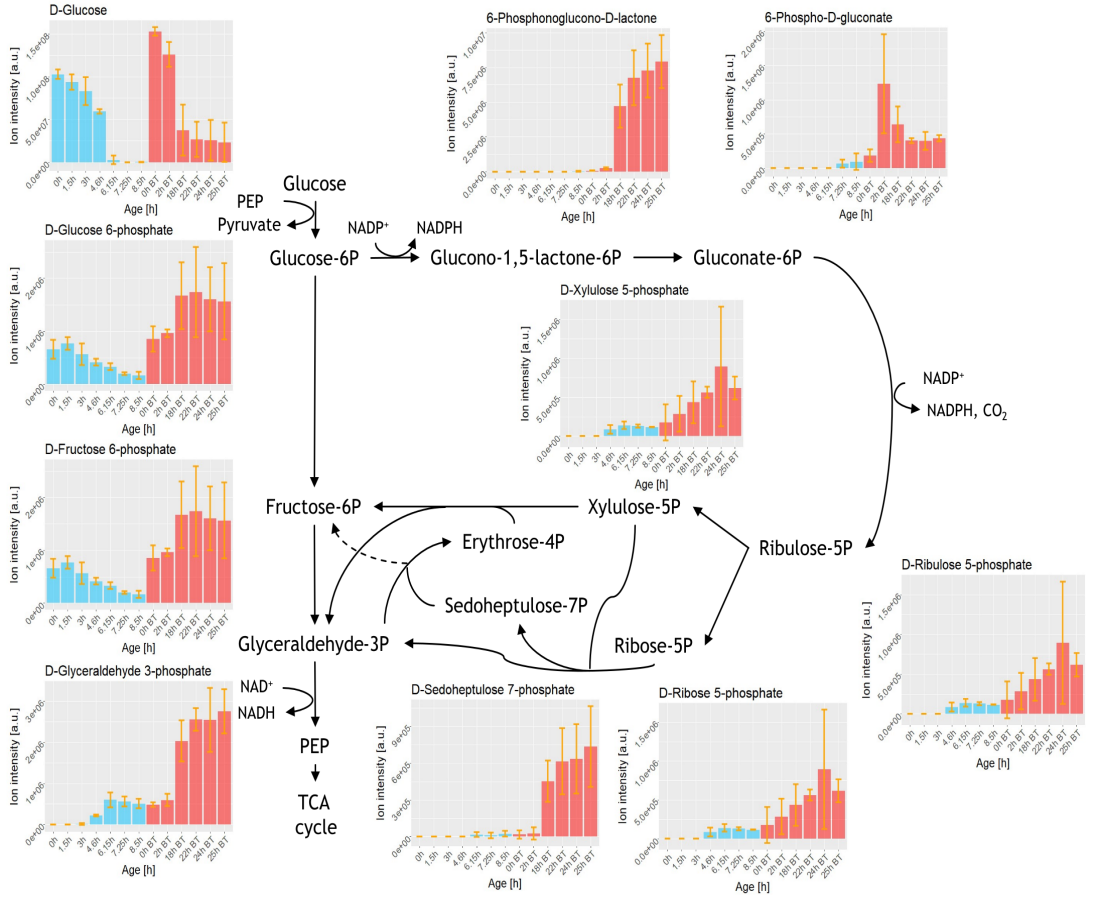



**Figure 4.12:** Dynamic evolution of **extracellular metabolites** of the Embden-Meyerhof-Parnas and pentose phosphate pathways during the bacterial growth phase (**blue**) and succinate production phase (**red**). Orange error bars represent the 95 % confidence intervals. Both phases are initiated with a glucose addition. During the growth phase, there is more extracellular presence of EMP metabolites than PPP ones. In the production phase, there is a fast increase in the secretion of metabolites from both EMP and PPP pathways after glucose is added. Ribose-5P, ribulose-5P and xylulose-5P are isomers indistinguishable with the method used, hence they have the same pattern. The same applies for glucose-6P and fructose-6P.

## 4.3 Discussion

The *E. coli* succinic acid fermentation was analysed by LC-MS to obtain useful information for bioprocess monitoring and optimisation, addressing three main points: complementing metabolite annotation of on-line metabolomics data, finding useful biomarkers for bioprocess monitoring and identifying targets for strain engineering and bioprocess optimisation.

### 4.3.1 Complementing on-line metabolomics data with LC-MS

On-line metabolomics is a powerful tool for bioprocess monitoring that allows the measurement of hundreds of different compounds at the same time. However, the on-line metabolomics system presented in this thesis does not have a chromatographic step. This reduces the complexity of the system and the chances of blocking it in the middle of a fermentation, but it also means that the metabolites are not separated before analysing them in the mass spectrometer. Consequently, metabolite annotation relies on accurate mass and or fragmentation.

Nearly 900 different m/z signals were monitored during the succinate fermentation process, many of them describing interesting dynamic patterns that could be useful for bioprocess monitoring or to identify potential targets for process optimisation (such as by-products or biomass biomarkers). Performing a time-course LC-MS analysis of the fermentation process it was possible to compare the data from on-line metabolomics and LC-MS analysis and annotate with higher confidence certain signals from the on-line monitoring experiments. This was done by comparing the time evolution of signals with the same m/z measured with both systems (within 5 ppm error).

Flow injection mass spectrometry is usually done for high-throughput analysis, biomarker discovery or sample profiling using high-resolution mass spectrometry (Beckmann *et al.*, 2008; Fuhrer *et al.*, 2011; Sarvin *et al.*, 2020; Vaidyanathan *et al.*, 2002). In these cases, metabolite annotation relies on accurate mass and fragmentation, but these are often not contrasted with LC-MS data. To the author's knowledge, performing a time-course LC-MS analysis to improve metabolite annotation of flow injection mass spectrometry data (such as on-line metabolomics) is a novel approach that has not been reported in the literature and it can be helpful to identify biomarkers and strain engineering targets found by flow injection data.

### 4.3.2 Process biomarkers

One of the key strengths of using metabolomics for fermentation process analysis is that the vast detection capacity of mass spectrometry allows the detection of useful biomarkers for bioprocess monitoring. Different biomarkers were identified by on-line metabolomics and LC-MS.

#### Biomass biomarkers

Monitoring the biomass concentration is important for bioprocess analysis because the cells are the catalyst of the bioreactor and therefore, they govern most of the reactions taking place inside the vessel. The concentration of cells and their health (viability and metabolic state) will, then, have a critical effect on the kinetics of substrate consumption, product formation and on the overall fermentation performance. For this reason, the concentration and growth rate of biomass are often used to determine many parameters of bioprocess kinetic models (see Chapter 6 and Choi *et al.*, 2014; Craven *et al.*, 2013; Klimek and Ollis, 1980; Schubert *et al.*, 2007). The signals 133.0506, 145.0505 and 147.0662 m/z were identified as potential biomarkers of biomass concentration using on-line untargeted metabolomics (Chapter 3). After doing LC-MS analyses on the fermentation samples, these ions were annotated as (R)-2,3-dihydroxy-isovalerate, (S)-2-aceto-2-hydroxybutanoate and (R)-2,3-dihydroxy-3-methylpentanoate, all three members of the BCAA biosynthesis metabolism, a metabolic pathway that will be active for protein and biomass synthesis.

Biomass monitoring for bioprocess analysis is currently done with optical density (turbidity), dielectric spectroscopy, microscopy, flow cytometry, fluorescence spectroscopy, calorimetry or vibrational spectroscopy (see Section 1.1.4). Using an intermediate from BCAA metabolism as a proxy of biomass concentration would be similar to using a fluorescent molecule, such as tryptophan, with the advantage that mass spectrometry can measure specific compounds independently, whereas fluorescence monitoring suffers from interference by any other fluorophore in the medium, which is the case of many aromatic compounds such as pyridoxine, FAD, FMN, NAD(P)H or riboflavin (Kiviharju *et al.*, 2008). Furthermore, monitoring biomass with mass spectrometry has the advantage that, at the same time, it also allows the monitoring of hundreds of metabolites with the same technology, rather than using multiple measuring systems (such as a biomass probe and a NIR probe), whose information might be difficult to combine. This might be the case if, for instance, a control system based on the concentration of glucose and biomass is to be designed.

### Product and by-products

Several metabolites were found to follow a similar pattern to succinate, the main fermentation product, such as pyruvic acid, $\alpha$-ketoisovaleric acid, 4-methyl-2-oxopentanoate and sedoheptulose 7-P. This information can help identify by-products that compete with succinate formation and therefore compromise the bioprocess product yield. Having a higher confidence in the annotation of these by-products can help direct metabolic engineering strategies for strain improvement. In this case, targeted identification of the compounds is desired in order to have confirmation of their identity before dedicating laborious efforts to strain engineering.

Succinate can be monitored well with on-line ESI mass spectrometry, especially as it is a compound that ionises well in negative mode and has a mass of 117.0193 m/z, well above 50-70 m/z, which is the common scanning range lower limit for LC-MS and flow injection analysis. Signals below 50 m/z can have a lot of interference caused by water, inorganic salt ions and the LC-MS buffers. However, certain fermentation processes might have a main product that is less suitable for ESI mass spectrometry, for instance molecules smaller than 50 Da or very non-volatile, such as ethanol and glycerol, respectively (Võ and Morris, 2014). In these cases, it might be convenient to find a biomarker metabolite for product formation, like the ones identified for succinate in this case.

### Carbon starvation biomarkers

Special attention was paid to identifying biomarkers of carbon starvation. Cellular systems undergo significant metabolic changes when they transition from a nutrient-rich environment to a nutrient-starved one. Many genes are up- and down-regulated upon substrate depletion leading to a reshuffle of resources. For this reason, from an industrial biotechnology point of view, in order to maximise product yield, titre and productivity, it is desired to avoid substrate starvation, especially for any significant period of time. Therefore, bioprocess monitoring can help to detect early signs of nutrients depletion and trigger control systems to correct this. Although it might be sufficient to monitor the different substrates of the bioprocess, it is worth noting that perhaps all substrates might not be easily detected on a single analytical instrument. For instance, ESI mass spectrometry might not be optimal for detecting poor-ionising compounds, such as glycerol, or small molecules such as $NH_4^+$. For this reason, it might be very useful to be able to determine general biomarkers of nutrient starvation and monitor these as well as the detectable specific bioprocess substrates.

Samples of the growth phase of an *E. coli* succinate fermentation process were analysed by LC-MS and HPLC-UV/Vis-RI. These were then split into “starved” and “unstarved” based on their glucose concentration, which was determined by HPLC-UV/Vis-RI analysis. Multivariate analysis of these samples suggested that D-glucose, citrate, cis-aconitate, pyrophosphate and (R)-2-hydroxyglutarate are good indicators of carbon-rich conditions in the bioprocess studied. As previously mentioned, D-glucose and citrate are two of the substrates present in

the defined fermentation medium, therefore it is not surprising that these were indicative of carbon-rich conditions. Nevertheless, the identification of other potential biomarkers could result useful. Cis-aconitate might be a good biomarker due to its enzymatic proximity to citrate in the TCA cycle. Therefore, the level of the two metabolites might be highly correlated. Pyrophosphate is the by-product of many biosynthetic pathways, including nucleotides, coenzymes, DNA, RNA, proteins and lipids, among others (Russell, 1976). For this reason, during carbon-rich conditions when the cells may have high anabolic activity, there might be an excess of free pyrophosphate. As far as (R)-2-hydroxyglutarate is concerned, this metabolite was not expected to be found in the *E. coli* strain used, according to the K-12 MG1655 KEGG library. The isomer (S)-2-hydroxyglutarate, however, is present in the lysine degradation pathway, so its involvement in anabolic metabolism is unclear. A possible explanation could be that it acts as a precursor for $\alpha$-ketoglutarate for the TCA cycle while this is highly active during carbon-rich conditions, but this could not be confirmed in the literature.

Furthermore, the multivariate analysis found hypoxanthine, xanthine, L-aspartate, L-glutamine and pantothenate as the top five indicators of carbon starvation. Hypoxanthine is a purine nucleobase that represses the production of other purines by feed-back inhibition (M. Liu *et al.*, 2020; Meng and Nygaard, 1990), therefore, it might be a cellular mechanism to stop DNA synthesis when nutrients are not available. Xanthine is another purine nucleobase and it is involved in purine interconversion pathways. The degradation of xanthine has been suggested to generate allantoin, a potential salvage route to generate a nitrogen source for catabolic purposes (Xi *et al.*, 2000). Increased xanthine levels could appear from the conversion of other purines, or as a salvage route to fight carbon starvation.

Increased levels of free amino acids have been reported in *E. coli* during stress conditions, which could be caused by the degradation of misfolded proteins caused by cellular stress, or as a salvage route to generate a pool of amino acids for the synthesis of stress-specific proteins (Jozefczuk *et al.*, 2010).

The precursors of pantothenate are $\alpha$-ketoisovalerate and L-aspartate (Cronan *et al.*, 1982). The increased levels of L-aspartate during carbon starvation could lead to high levels of pantothenate. $\alpha$-ketoisovalerate belongs to the branched chain amino acid synthesis pathway and, although the levels of this metabolite did not rise until the anaerobic production phase, other compounds of this pathway increased throughout the biomass formation phase. This is the case of the metabolites identified as possible biomass biomarkers, *i.e.* (R)-2,3-dihydroxy-3-methylpentanoate, (R)-2,3-dihydroxy-isovalerate and (S)-2-aceto-2-hydroxybutanoate.

In the analysis carried out, biomarkers for carbon starvation were looked at. However, the cellular response over other limiting conditions such as oxygen, phosphate or nitrogen limitation might be very different. For instance, oxygen limitation turns a metabolism based on the full oxidation of substrates into $CO_2$ and reduced cofactors (NADH and $FADH_2$) via the clockwise TCA cycle, into a fermentative metabolism, where ethanol, $H_2$ and organic acids such as

acetate, succinate, formate and lactate are accumulated (Spiro and Guest, 1991). As far as phosphate is concerned, phosphate is present in many metabolites, such as phosphosugars, phospholipids and nucleotides, it is involved in the regulation of many enzymatic reactions (Schastnaya *et al.*, 2021) and is in the core of energy metabolism, for example storing high amounts of energy in the form of P-O phosphoanhydride bonds in molecules such as ATP and ADP. The effect of phosphate limitation, therefore, affects many metabolic reactions, including the hampering of cellular growth and the induction of synthesis of many proteins, such as the ones under the control of the phoA promoter (Y. Wang *et al.*, 2005; Y. Zhang *et al.*, 2010). Pandi *et al.* (2020) even observed increase recombinant protein production under phosphate limited conditions using the widely-used T7 promoter, which, unlike phoA, is *a priori* not phosphate-inducible. Finally, nitrogen is also present in proteins and many metabolites, such as amino acids, nucleotides and cell wall compounds like peptidoglycan, glycoproteins and some phospholipids (Switzer *et al.*, 2020). For this reason, nitrogen limitation has a big effect on the metabolism of *E. coli*, including the activation of at least 75 genes under the control of the nitrogen regulatory protein C (NtrC), with the major consequences of slowing down bacterial growth and the scavenging of nitrogen-containing compounds, mainly glutamate and glutamine (Brown *et al.*, 2014; Switzer *et al.*, 2020; Zimmer *et al.*, 2000).

In conclusion, the biomarkers identified in this section are specific for carbon limitation and are bioprocess-specific. The batch fermentation medium used is designed to be carbon-limited and to have nitrogen and phosphate in excess. Furthermore, the medium dissolved oxygen was controlled to be over 30 % during the biomass formation phase. Therefore, the biomarkers found during the starvation period at the end of the batch phase should be specific for carbon starvation. It is also important to point out that if a different strain or microorganism were used, these would probably generate a different metabolic response to carbon limitation and consequently have different biomarkers for carbon limitation.

### 4.3.3 Targets for strain engineering

The pentose phosphate pathway was significantly active in the fermentation process, both at the beginning of the aerobic biomass formation phase and during the succinate production phase. The main functions of the PPP are to generate anabolic intermediates such as ribose-5P, erythrose-4P and reduced NADPH cofactors (Matano *et al.*, 2014), which are required for processes such as fatty acid synthesis, cell membrane formation (W. Li *et al.*, 2018), ribonucleotide synthesis (Wamelink *et al.*, 2005) and aromatic amino acid biosynthesis (S. Y. Lee *et al.*, 2019). As there is no biomass formation during the succinate production phase, the PPP was considered a potential target for strain engineering.

Examples of previous work describing PPP activity under aerobic and anaerobic conditions are given below.

#### Aerobic conditions

Carbon flux towards the PPP under aerobic conditions has previously been quantified in *E. coli* by J. Zhao and Shimizu (2003) on a wild type K-12 *E. coli* strain with $^{13}$C-labelled acetate and glucose as carbon sources using GC-MS. As mentioned above, the PPP is required under growing conditions for the formation of anabolic intermediates.

#### Anaerobic conditions

Activity in the PPP during succinate production under anaerobic conditions has been reported by S. Lu *et al.* (2009), who calculated carbon fluxes using LC-MS with fermentations carried using $^{13}$C-labelled glucose as a carbon source on an AFP111 *E. coli* strain – which is a mutant of strain NZN111, with deletions of genes *pflB*, *ldhA* (same deletions as the strain used in this study) and a mutation on the *ptsG* gene encoding a membrane glucose permease that results in the ability to grow anaerobically on glucose (Chatterjee *et al.*, 2001). However, the ability to grow under anaerobic conditions could indicate that the PPP is active to generate anabolic intermediates.

Zhu *et al.*, 2014 also reported activity in the PPP during anaerobic succinate production using *E. coli* variant strains of ATCC 8739 by determining the enzymatic activity of the enzyme transketolase, which catalyses the two reactions of the non-oxidative phase of the PPP involving xylulose 5-P (Comín-Anduix *et al.*, 2001; Mariadasse *et al.*, 2016). The authors did not specify if the *E. coli* strains could grow anaerobically on glucose.

The strain used in this thesis is not able to grow anaerobically on glucose, and therefore, the diversion of carbon towards the PPP (and potentially towards connected pathways) during the anaerobic succinate production phase could result in an undesired escape route of carbon in the formation of by-products, such as nucleotide biosynthesis. This led to the hypothesis that rechannelling some of this carbon spent on the PPP towards the anaerobic TCA cycle could be a possible target for improving succinate production. This hypothesis will be addressed in Chapter 7 by inhibiting the PPP using oxythiamine, a known inhibitor of the enzyme transketolase during succinate production.

### 4.3.4 The extracellular fraction

A final point of discussion is the fact that a very large number of metabolites (4801) were annotated in the extracellular fraction, which was unexpected, as many of these are essential compounds for the cell (amino acids, nucleotides, vitamins, *etc.*) and, therefore, are not expected to be excreted. Other authors reported much fewer metabolites when analysing the exometabolome (the extracellular metabolome). For instance, Villas-Bôas *et al.* (2005) reported the detection of hundreds of peaks intra- and extracellularly by GC-MS, leading to up to 29 identifications in the extracellular fraction, mostly including metabolites from amino

acid metabolism and the TCA cycle. To give another example, Kim *et al.* (2019) reported 53 extracellular metabolites identified by GC-MS, including sugars, sugar phosphates, organic acids, amino acids and nucleobases among others. In both cases, cells were grown in minimal media not supplemented with amino acids, indicating that finding these, sugar phosphates and nucleobases extracellularly was caused by either cell lysis or secretion.

Cell lysis and metabolite secretion are considered below as potential reasons to explain the large number of metabolites observed extracellularly.

### Cell lysis

The fact that many essential metabolites were found in the extracellular fraction could be an indication of cell lysis, especially as these appear after 6.15 h of fermentation (at the end of the biomass formation phase) and throughout the production phase. This cell lysis could be caused by high agitation rates or by substrate limitation. After the first 4-5 hours of fermentation the impeller speed increases from 500 to 900 rpm to maintain the dissolved oxygen (DO) levels above 30 % saturation. Then, towards the end of the batch phase, glucose becomes limiting. Finally, after the batch phase, succinate is produced under anaerobic conditions, which cannot sustain cell growth and cause cell death.

As far as agitation is concerned, although industrial strains are commonly selected to be robust in their operational conditions, high values or sudden changes in agitation might cause excessive shear stress and lead to some cell lysis. However, Hewitt *et al.* (1998) determined by flow cytometry that the cell membrane and cytoplasmatic membrane potential of *E. coli* W3110 were intact (a sign of healthy cells) when cultured in a 5 L glass bioreactor in continuous mode at impeller speeds of up to 1200 rpm. Concomitantly, the cell physiology was also found to be unchanged at 400 and 1200 rpm agitation speeds when inspected by transmission electron microscopy and scanning electron microscopy, although a layer of polysaccharide was found outside the cells at 400 rpm and it was washed away at 1200 rpm. This study was done in a similar set-up to the experiments carried out in this thesis, suggesting that cell death due to agitation might not be the case. Even though a different *E. coli* strain was used by Hewitt *et al.* (1998), the strain used in this thesis is employed in industrial applications and is, therefore, not expected to be especially fragile to high agitation speeds. It was also considered whether the fluctuations between 500 and 900 rpm caused by the dissolved oxygen control system could cause an additional cellular stress. However, the effect of fluctuating agitation speeds on cell viability could not be confirmed in any published work.

As far as substrate limitation is concerned, both glucose and oxygen limitation could cause cell lysis due to cellular stress (nitrogen and phosphate limitation are not considered because the batch fermentation medium used is designed to be carbon-limited and to have nitrogen and phosphate in excess). Although many *E. coli* strains can grow anaerobically on glucose, the strain used cannot, as observed experimentally (Figure 3.1). Oxygen limitation probably

causes cell lysis in this strain due to insufficient ATP formation for cell maintenance. The *E. coli* strain is Δ*ldhA*, consequently, making succinate it only generates 1 ATP molecule from 1 molecule of glucose, whereas 3 ATP molecules would normally be generated to make lactate (Figure 1.2). The main difference is in the conversion of PEP. When *E. coli* makes lactate, PEP is converted into pyruvate catalysed by the enzyme pyruvate kinase, which generates ATP. However, to make succinate, PEP is carboxylated into oxaloacetate with no ATP generation. Some authors such as (Singh *et al.*, 2011) have increased ATP production during succinate production in a Δ*ldhA E. coli* by cloning a transgenic ATP-generating PEP carboxylase from *Actinobacillus succinogenes*, resulting in an increased biomass and succinate formation.

### Secretion of metabolites

Another reason that could explain the presence of compounds in the extracellular media could be secretion. Although small uncharged molecules can diffuse freely through the outer membrane of *E. coli*, larger (up to 600 Da) polar molecules and ions are transported via porins into the periplasmic space. Therefore, molecules such as sugars, amino acids and sugar phosphates, need primary and secondary active transporters to cross the cytoplasmatic membrane against the concentration gradient (Alva *et al.*, 2020; Cooper, 2000). In *E. coli*, the preferred method of glucose import is via the phosphoenolpyruvate (PEP)-sugar phosphotransferase (PTS) system, which phosphorylates glucose into glucose 6-phosphate using PEP as phosphate donor as glucose crosses the cytoplasmic membrane (Alva *et al.*, 2020).

To the author's knowledge, there is no transporter system in *E. coli* for the secretion of sugar phosphates into the extracellular space, and therefore, the detection of these compounds outside of the cell would be an indication of cell lysis. It was previously mentioned that Kim *et al.* (2019) also found sugar phosphates in the extracellular fraction of *E. coli* cultures. However, their sample preparation for metabolomics analysis started with a 10-minute centrifugation step at 16,100 x g, which could be an excessive speed that could lyse the cells, as indicated by Pembrey *et al.* (1999), who reported 79 % viability in *E. coli* cells after two rounds of centrifugation at 15,000 x g for 10 minutes and cell resuspension in saline solution. Indeed, this could also be the case for the LC-MS analysis carried out in this thesis. After taking the samples from the bioreactor, the cells were centrifuged twice at 13,000 x g for 10 min. No viability data was found for centrifugation speeds of 13,000 x g, but it is possible that a small percentage of cell lysis is caused in these conditions. An argument to think that this is not the case is that comparing Figures 4.11 and 4.12, relatively high intracellular levels of D-glyceraldehyde 3-P, D-sedoheptulose 7-P, D-ribose 5-P, D-ribulose 5-P and D-xylulose 5-P are seen at 3 h of fermentation, but none of these appear in the extracellular fraction at this point, although this could be because the latter might have been under the limit of detection.

**The extracellular fraction - Summary**

To sum up, many metabolites were found in the extracellular fraction, which could be explained by a few reasons, including cell lysis and secretion. On one hand cell lysis could be caused by substrate limitation (glucose or oxygen), high agitation rates (although this seems unlikely based on other literature findings and the fact that the strain used is an industrial strain) or high centrifugation speeds during sample preparation for LC-MS analysis. On the other hand, secretion of many essential metabolites would require the action of active transporters.

Most likely, a combination of more than one of these elements is taking place. Based on the evidence mentioned above, cell lysis due to substrate limitation and to high centrifugation speeds during sample preparation are probably the main causes of intracellular metabolites being released into the extracellular medium.

## 4.4 Conclusions

This chapter focused on three main strategies to improve bioprocess monitoring and optimisation using LC-MS. Specifically, LC-MS analysis allowed complementing metabolite annotation of untargeted on-line metabolomics data and to perform multivariate and pathway analysis. Biomarkers of carbon starvation were identified with multivariate analysis (PCA and PLS-DA), complementing the findings obtained with untargeted on-line metabolomics. Biomarkers can be extremely useful for bioprocess monitoring. For instance, a biomarker for biomass can be of great importance for understanding if the cells in the bioreactor are behaving as they should in order to meet process targets and product specifications. Similarly, monitoring substrate depletion, product formation or the appearance of by-products is of great importance for bioprocess monitoring, regardless of whether these are monitored by direct measurement or via a biomarker.

Pathway analysis performed on the LC-MS results allowed the identification of three potential pathways that could be targeted for strain engineering and bioprocess optimisation: the pentose phosphate pathway, ascorbate and aldarate metabolism and pantothenate and CoA biosynthesis. Of these three, the pentose phosphate pathway was investigated further and PPP metabolites were observed to rapidly increased at the beginning of the anaerobic succinate production phase. Previous authors have shown activity in the PPP under anaerobic conditions using *E. coli* strains capable of sustaining anaerobic growth. However, as the strain used in this thesis is not able to grow anaerobically on glucose, the expenditure of carbon towards the PPP could result in the formation of by-products and therefore carbon losses. This suggested that rechannelling some of this carbon spent on the PPP towards the anaerobic TCA cycle could be a possible target for improving succinate production, which will be investigated in Chapter 7.

## 4.5 Materials and methods

The materials and methods are as described in Chapter 2. Any differences and particular methods used exclusively in this chapter are described below.

### 4.5.1 Sample extraction for LC-MS analysis

As described in Chapter 2, but metabolism was quenched by putting the microtubes containing the samples on dry ice instead of using $H_2SO_4$. Also, once extracted with 1:3:1 C:M:W, intracellular fractions for samples 3-13 were further diluted by adding 200 $\mu$L 1:3:1 C:M:W to compensate for the low biomass concentration of the first two samples (which had 400 $\mu$L (instead of 200 $\mu$L) of 1:3:1 C:M:W for every 5 mg of WCW pellet to allow for proper mixing during solvent extraction).

### 4.5.2 Correlation analysis between off-line WCW and on-line metabolomics

A Pearson test was used to analyse the correlation between the off-line WCW biomass data and the on-line metabolomics data of the compounds identified as potential biomarkers. This test was performed using the ggpubr package (Version 0.4.0; Kassambara, 2020) in the statistical software environment R (Version 4.0.4).

### 4.5.3 Multivariate statistical analysis

Samples from the aerobic biomass formation phase of an *E. coli* succinate bioprocess were analysed by PCA and PLS-DA as described below. Data pre-processing consisted in $\log_{10}$ transformation of the mass spectrometry peak areas and samples were normalised by the median.

#### PCA

The PCA was performed in the online platform MetaboAnalyst (Version 5.0) (Pang *et al.*, 2021; Xia *et al.*, 2009), which uses the prcomp function from the stats package in the statistical software environment R (Version 4.1.1).

#### PLS-DA

The PLS-DA regression was performed in the online platform MetaboAnalyst (Version 5.0) (Pang *et al.*, 2021; Xia *et al.*, 2009), which uses the plsr function from the pls package (Version 2.1-0; Liland *et al.*, 2021) in the statistical software environment R (Version 4.1.1). The classification and cross-validation are performed using the caret package (Version 3.45; Kuhn *et al.*, 2021). The $Q^2$ parameter was calculated using the LOOCV method.

### 4.5.4 Pathway analysis

Pathway analysis was performed in PiMP (date of use: 08 Jul 2020) using the pathway activity level scoring (PALS) algorithm. Pathways with p-values of 0.05 or below are considered statistically significant across experimental groups compared. The experimental groups compared were the different time points, and each pairwise comparison was done comparing any given time point with the first sample of the fermentation. In order to reduce the dimensionality of the analysis, for each pathway, the p-values of all the samples were averaged together and this mean p-value was used to find the 10 “most significant” pathways (lowest mean p-value). These 10 “most significant” selected pathways were displayed as a heatmap showing the p-value of each pathway for each sample. Heatmaps were created using the gplots package (Version 3.1.1; Warnes *et al.*, 2020) in the statistical software environment R (Version 4.0.4). Pathway analysis for intracellular and extracellular samples was performed separately.

# PART III

# Targeted on-line metabolomics analysis of an *Escherichia coli* succinate fermentation process and steps towards bioprocess improvement

Chapter 5

# On-line targeted metabolomics for monitoring an *Escherichia coli* succinate fermentation process

The reference standards used to create the MRM method were kindly provided by Ingenza Ltd. and Prof. Stephen C. Fry from the School of Biological Sciences at the University of Edinburgh.

## Abstract

Chapter 3 described the development of an on-line metabolomics platform for monitoring fermentation processes. This tool was used to monitor an *Escherichia coli* dual-phase succinate production fermentation process, and metabolites following patterns of potential interest for process monitoring and optimisation were identified. A higher degree of confidence in the annotation of these compounds was achieved by off-line liquid chromatography-mass spectrometry analysis, as demonstrated in Chapter 4. With this information, a targeted analysis method was developed to monitor 41 metabolites of interest with a triple quadrupole mass spectrometer and univariate linear regression models were developed to correlate the on-line mass spectrometry peak area of 13 of these metabolites with their concentration in the bioreactor, allowing their monitoring in real-time. One of these metabolites, (R)-2,3-dihydroxy-isovalerate, was used as a biomass biomarker and its peak area was linearly correlated to the biomass wet cell weight concentration, allowing the on-line monitoring of biomass.

Three fermentation replicates were used to build these models. The data was randomly split into a training set and a test set. The training set was used to build the models using 10-fold cross validation and model performance was assessed on the cross-validation and test data. These on-line metabolomics univariate models performed comparably to the vibrational spectroscopy models reported in the literature, which typically are multivariate partial least squares regressions. Furthermore, the on-line metabolomics models developed in this chapter have a lower risk of overfitting and are much simpler to build and interpret.

## 5.1 Introduction

In Chapter 3, an on-line metabolomics monitoring system was developed and coupled to an Orbitrap mass spectrometer. Due to its high-resolution accurate mass characteristics, the Orbitrap can detect a vast array of metabolites and distinguish very similar masses. This makes the system well-fitted for untargeted analysis, finding ions following interesting patterns throughout the course of the fermentation. This makes it possible to identify key process metabolites, potential targets for strain or process engineering and useful process biomarkers. Confident annotation of these compounds can be achieved by LC-MS, as it was demonstrated in Chapter 4. Once these compounds have been identified, a targeted method can be developed to specifically monitor them and build quantitative models to correlate their mass spectrometry signal to their concentration, enhancing bioprocess monitoring and control.

Targeted metabolomics is the analysis of a predefined specific set of compounds using metabolomics. For bioprocess monitoring, it might be relevant to analyse only a small set of compounds, namely those ones identified as playing an important role in the process, such as substrates, products, by-products, inhibitors and biomarkers. This is particularly the case at production scale, when the most important process compounds have been identified and the process conditions have been well-studied and established (strain, pH, temperature, medium recipe, *etc.*).

In the field of metabolomics, triple quadrupole (QQQ) mass spectrometers are the most popular option for performing targeted analyses, mainly due to their high linear dynamic range and speed in transitioning between the different target metabolites to analyse (Gross, 2017). In this chapter, the on-line metabolomics system was coupled to a QQQ and the compounds of interest found in Chapters 3 and 4 were monitored with a targeted method. Then, univariate linear regression models were built to monitor the concentration of these metabolites. Multiple reaction monitoring is the most common method for the analysis of multiple compounds in a single QQQ experiment.

### 5.1.1 Multiple reaction monitoring

Selected reaction monitoring (SRM) is a mode of operating a QQQ instrument exploiting the filtering and fragmentation capacities of quadrupoles. Particularly, the first quadrupole (Q1) is used to select a specific precursor ion of interest, while filtering out ions of different m/z values. Then, the second quadrupole (Q2) is used as a collision cell, where an inert gas (usually $N_2$ or Ar) fragments the precursor ion with a predefined collision energy to generate a characteristic set of product ions in a process called collision induced dissociation (CID). Finally, the third quadrupole (Q3) selects only a specific product ion, allowing it to reach the ion detector (Gross, 2017; Pisitkun *et al.*, 2007). This mode of operation allows to distinguish

coeluting species with the same m/z that have different fragmentation patterns. When more than one metabolite is sequentially monitored by SRM, the mode of operation is called multiple reaction monitoring (MRM), as depicted in Figure 5.1. In this chapter, MRM will be used to monitor a list of metabolites by on-line and off-line metabolomics.

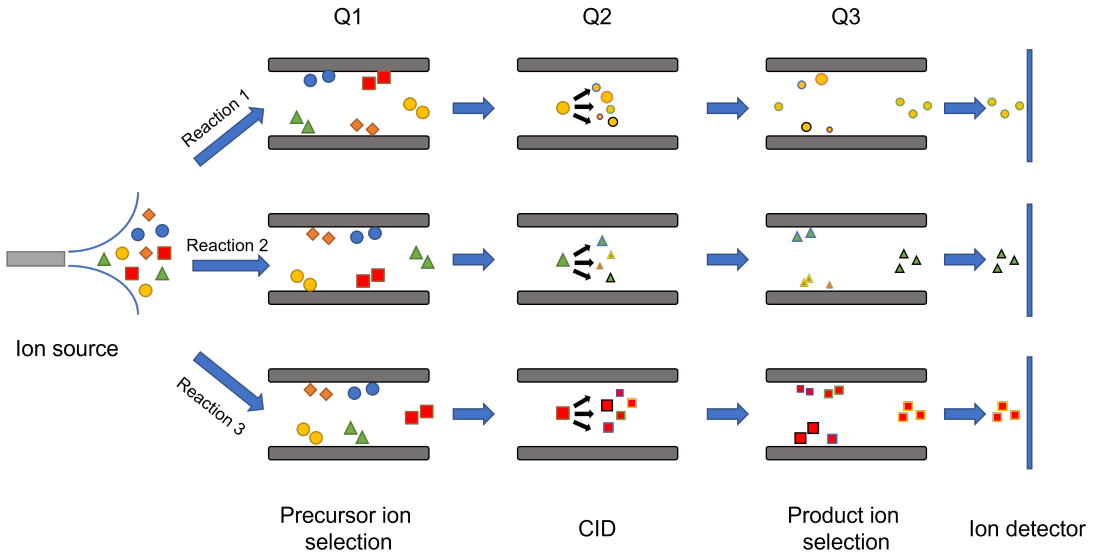



**Figure 5.1:** Multiple reaction monitoring diagram. Different metabolites are represented as different shapes. Different product ions as a result of fragmentation in Q2 are represented with different sizes and outline colours.

### 5.1.2 Aims and objectives

In Chapter 3, the *E. coli* succinate fermentation process was monitored with on-line untargeted metabolomics and different potential metabolites of interest were identified, including compounds of glycolysis, the TCA cycle, the PPP, nucleotide metabolism, metabolites of anaerobic fermentation, amino acids and biomass biomarkers. The main aims of this chapter are:

- To build an MRM method for targeted metabolomics analysis of the potential metabolites of interest found with untargeted metabolomics.
- To monitor three succinate fermentation replicates with the on-line targeted metabolomics MRM method.
- To build regression models to correlate the on-line mass spectrometry signals with metabolite concentration so the on-line monitoring system can be used to directly measure the concentration of the target metabolites. To create these regression models, off-line fermentation samples will also be analysed and used as reference metabolite concentration at different points of the fermentation.

A schematic of the experimental setup to develop the regression models for monitoring the concentration of metabolites on-line is shown in Figure 5.2.

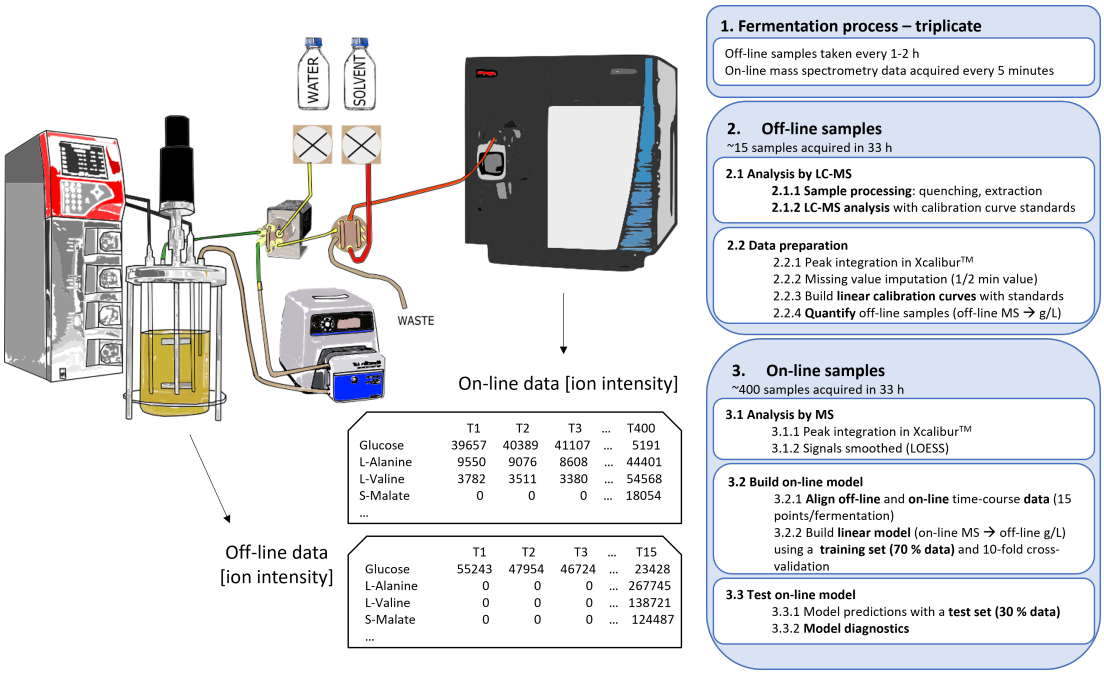



**Figure 5.2:** Process to build the regression models to monitor the concentration of targeted metabolites on-line. On-line and off-line data values shown are only an example (not necessarily realistic numbers).

## 5.2 Results and discussion

### 5.2.1 MRM method

An MRM method was developed in the QQQ mass spectrometer to monitor 41 metabolites of interest during the succinate fermentation. These include compounds of glycolysis, the TCA cycle, the PPP, anaerobic fermentation, amino acids and biomass biomarkers identified in Chapters 3 and 4. Compounds of nucleotide metabolism were not included in the monitoring method due to the lack of access to their reference standards. To build the MRM, reference standards of these compounds of interest are introduced into the mass spectrometer by direct infusion and the instrument is tuned to find different CID energies that generate high-intensity fragment ions for each metabolite. Five fragments were monitored for each molecule. A few examples of proposed fragmentation mechanisms are shown in Figure 5.3. Two product ions were selected for each molecule for the MRM method, based on the intensity of the product ion and that a possible chemical structure could be proposed for it. Considering the structure of the product ions is important to avoid choosing fragments of potential contaminant ions, or choosing non-specific ions, such as $CO_2$, $H_2O$ or $H_2PO_4^-$, which are product ions of many molecules, therefore, these are not specific of the parent ion (as shown in Figure 5.3).

For metabolites with the same unit mass or very similar mass (< 1 m/z difference), product ions with a different m/z were sought, as QQQ are low-resolution instruments and cannot distinguish between peaks within a $\pm$ 0.5 m/z window. However, it was not always possible to find unique combinations of precursor and product ions for all metabolites in the MRM list. For example, ribose-5P, ribulose-5P and xylulose-5P are three isomers that did not generate distinctive product ions that would allow to identify them separately. Therefore, without chromatography these three metabolites will be measured together as a collective signal. In this case it might not be a big problem because the internal fluxes within the PPP were not studied.

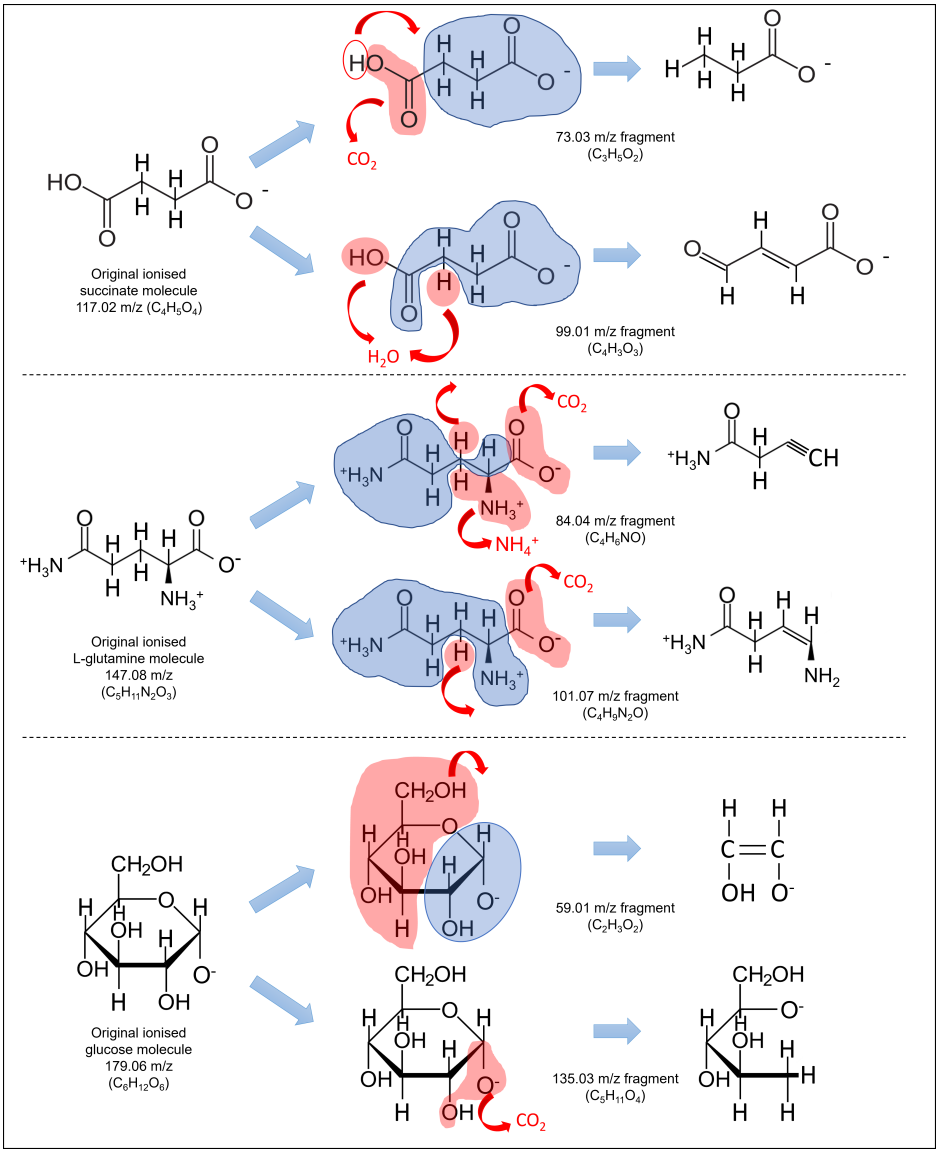



**Figure 5.3:** Examples of proposed fragments observed by CID for the developed MRM method: succinate (**top**), L-glutamine (**middle**) and glucose (**bottom**). The regions highlighted in blue remain in the product fragment ion, while the regions in red do not. Final structures shown are only possible suggestions matching the observed m/z values, *i.e.* they were not empirically elucidated.

### 5.2.2 Targeted on-line metabolomics

#### Triplicate fermentations

Three replicate *E. coli* succinate fermentation experiments were carried out and monitored with on-line metabolomics using a QQQ mass spectrometer with the MRM method discussed above. The samples analysed with on-line metabolomics were extracted via a biomass filtration probe and, therefore, contain extracellular medium (filtration efficiency was measured at >99.94 % by plating filtered and non-filtered samples on selective plates, see Section 5.5). At different time points, off-line samples were taken for LC-MS analysis and biomass measurements. Figure 5.4 shows the DO, pH, temperature and off-gas $CO_2$ and $O_2$ of the three replicates. The first spike in DO and drop in $CO_2$ two hours before the beginning of the succinate production phase are a sign of glucose starvation and coincide with the end of biomass growth and glucose depletion measured off-line (Figure 5.5).

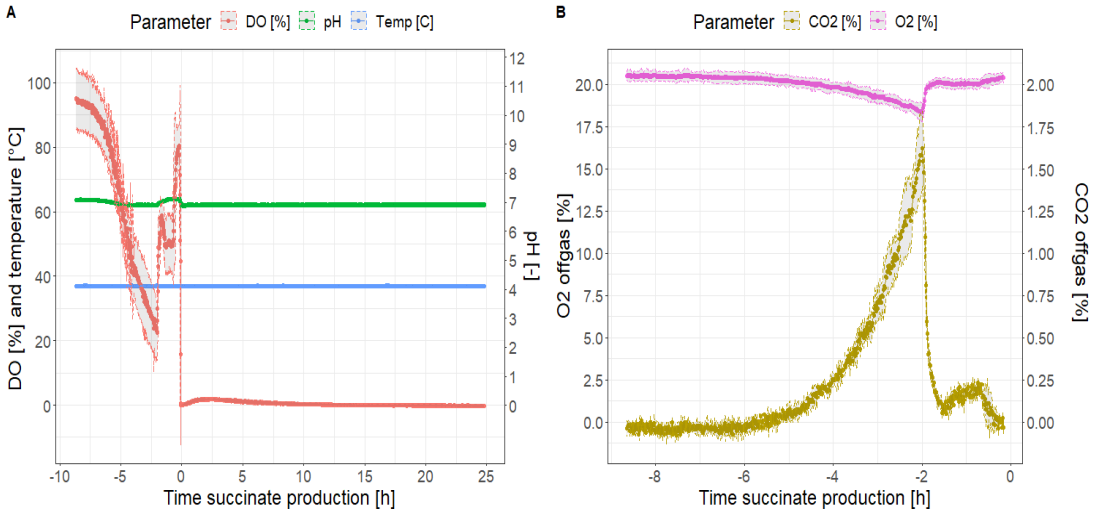



**Figure 5.4:** DO, pH and temperature (**A**) of the three fermentation replicates that were monitored with targeted on-line metabolomics. The off-gas $CO_2$ and $O_2$ (**B**) is only shown during the batch growth phase because during the production phase it was constant at 100 % $CO_2$. The 95 % confidence intervals are shown with a grey area. Time is indicated with respect to the beginning of the succinate production phase for a better comparison of the three replicates.

### Concentration of off-line samples

In order to build calibration models to quantify the on-line metabolomics measurements, first it is necessary to determine the concentration of metabolites in the off-line samples (Figure 5.2).

The extracellular metabolite concentrations of the off-line samples were measured by LC-MS using calibration curves of reference standards. These concentrations are shown in Figure 5.5 together with the biomass levels measured off-line. Only extracellular LC-MS analysis was carried out for comparison with the on-line measurements, which were done using a biomass filtration probe. The concentration in the fermentation samples of 14 compounds was below the range of their corresponding calibration curves, which made it not possible to reliably determine their concentration. The ion intensity of these metabolites in the off-line samples is shown in Figure 5.6. Even though the concentration could not be determined for these 14 compounds, their dynamic profile can be followed throughout the fermentation as a relative quantitation measurement.

Both Figure 5.5 and 5.6 show an increase in the extracellular concentration for many metabolites during the succinate production phase. This is the case for fumarate, L-lactate, pyruvate, S-malate, succinate, most amino acids and sedoheptulose-7P. Based on the concentration, pyruvate, L-alanine and L-lactate were identified as the main by-products of succinate formation. Interestingly, 0.4 g/L of L-lactate are formed despite the strain carrying the genomic deletion of lactate dehydrogenase (*ldhA*l), the main gene for L-lactate production in *E. coli* (see Section 2.2).

A sudden increase in extracellular acetate and L-glutamate is observed starting around five hours before the beginning of the production phase, and both are consumed when glucose is depleted. Under aerobic conditions with glucose excess, acetate is formed in *E. coli* due to overflow metabolism and the Crabtree effect, and part of this excess acetate is excreted out of the cell to prevent osmotic stress due to the accumulation of negative charges in the cytoplasm (Akesson *et al.*, 2001; De Mey *et al.*, 2007; Wolfe, 2005; B. Xu *et al.*, 1999). L-glutamate is one of the main ions involved in regulating osmotic pressure in gram-negative bacteria. Roe *et al.* (1998) demonstrated in *E. coli* that increasing the intracellular concentration of acetate causes the cells to excrete L-glutamate, explaining the L-glutamate peak in Figure 5.6.

Details of the calibration curves developed to quantify the metabolites in Figure 5.5 are given in the next subsection.

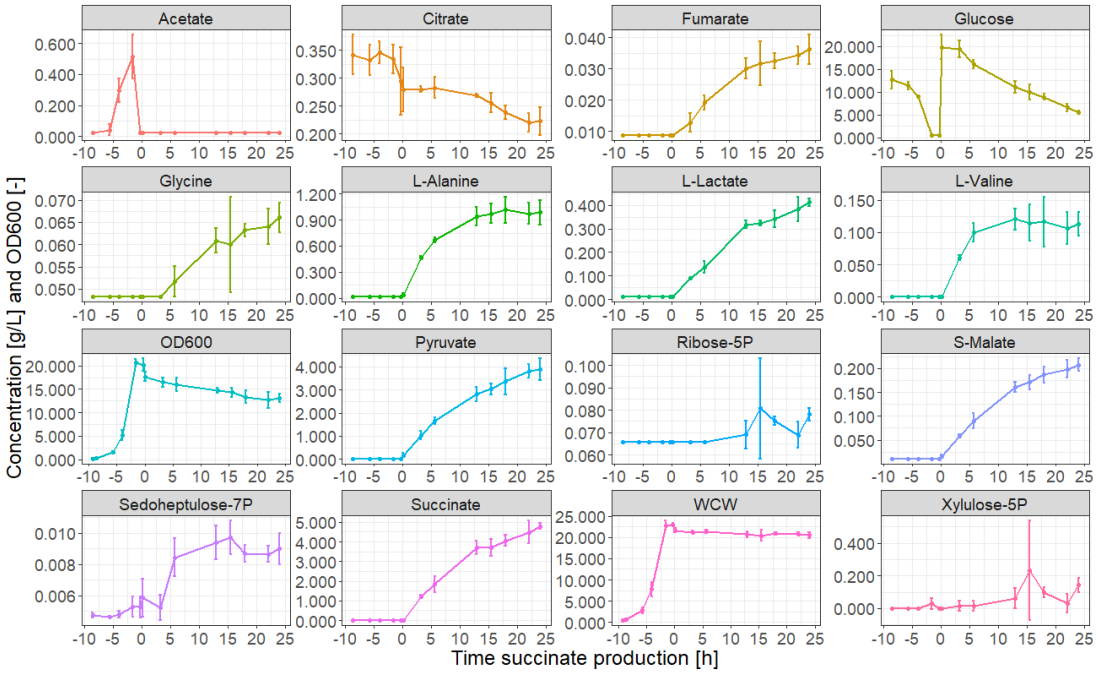



**Figure 5.5:** Off-line extracellular metabolite and biomass concentrations measured by LC-MS at the different stages of the fermentation. Time is indicated with respect to the beginning of the succinate production phase for a better comparison of the three replicates. The 95 % confidence intervals are shown as error bars.

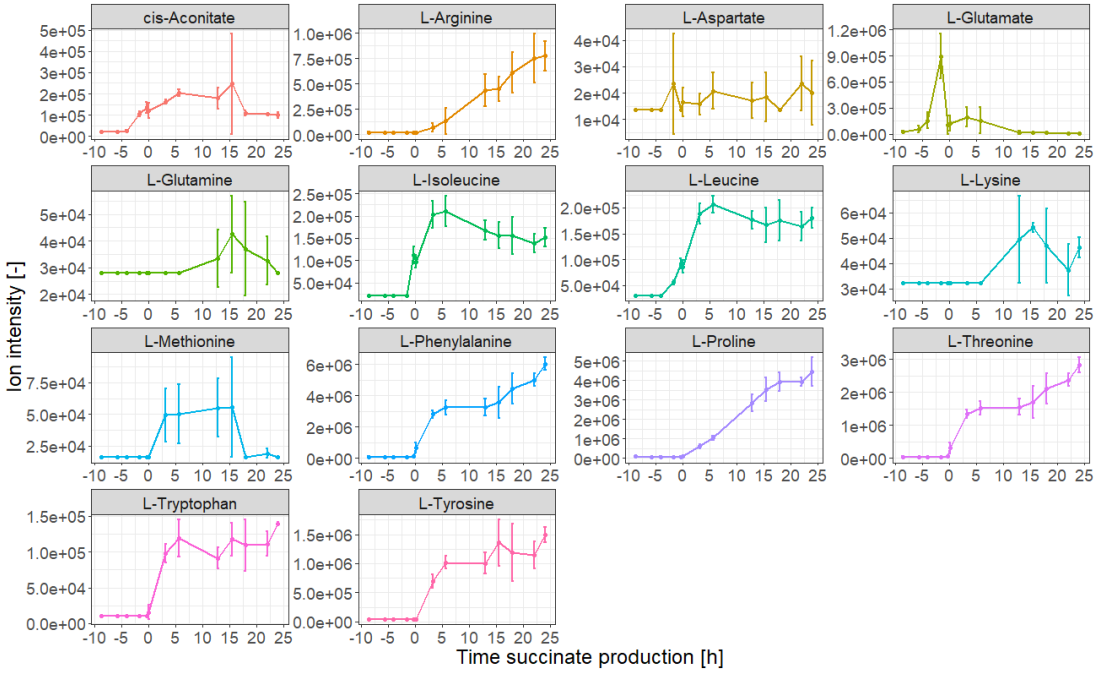



**Figure 5.6:** Off-line ion intensity of the metabolites for which the concentration could not be reliably determined by LC-MS. Time is indicated with respect to the beginning of the succinate production phase for a better comparison of the three replicates. The 95 % confidence intervals are shown as error bars.

*Calibration curves for calculating the concentration of metabolites analysed by LC-MS*

A few examples of the calibration curves built to determine the metabolite concentrations by LC-MS are shown in Figure 5.7 and a summary of the calibration curve parameters for the full list of reference standards is gathered in Table 5.1. To build these calibration curves, a series of dilutions of reference standards of the compounds of interest were analysed by LC-MS and the concentration of each compound was correlated using linear regression to the corresponding peak area measured by LC-MS.

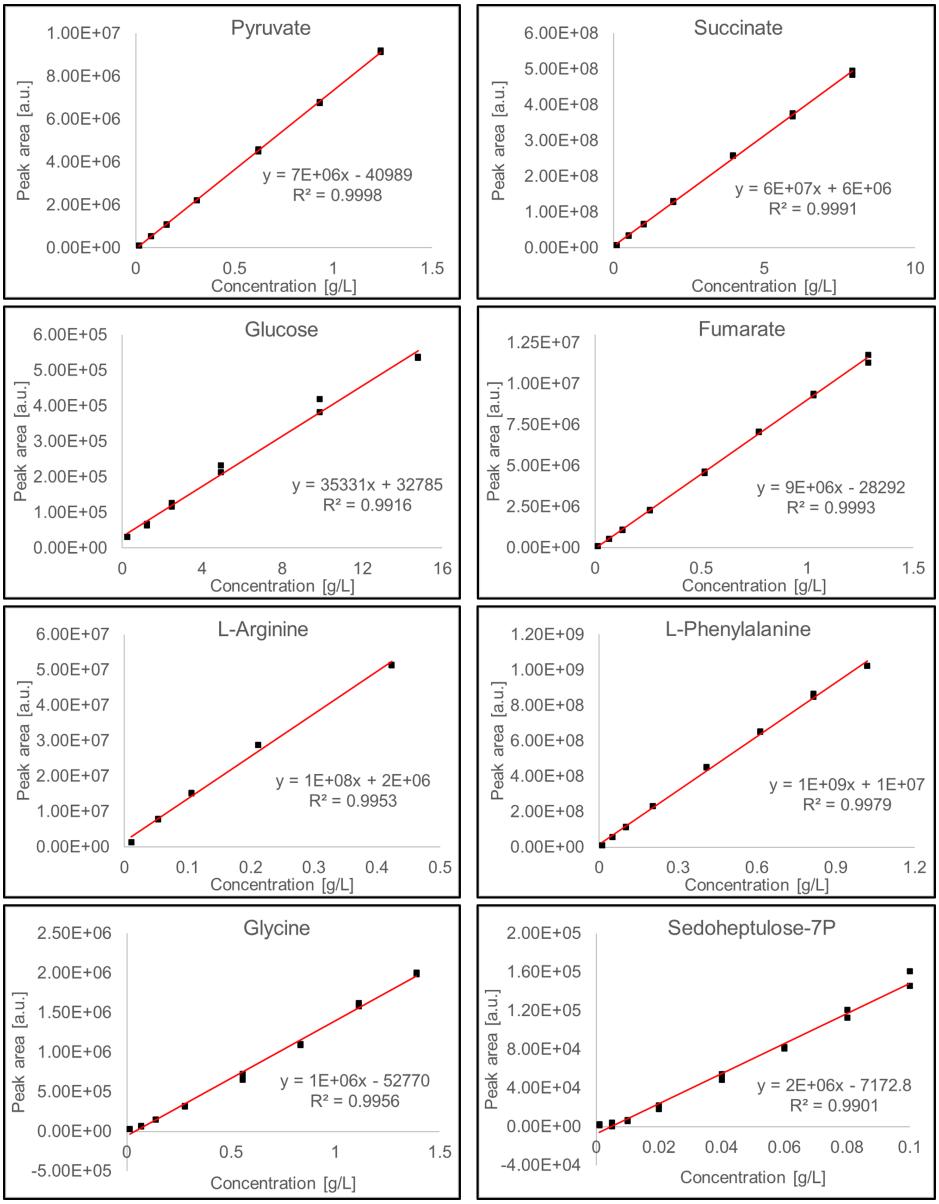



**Figure 5.7:** Calibration curve of reference standards for determining metabolite concentration of the extracellular off-line fermentation samples. Only eight of the 28 measured metabolites are shown.

**Table 5.1:** Parameters and summary statistics of the linear regression calibration curves correlating analyte concentration and the off-line LC-MS peak area for the 28 reference standards analysed. Metabolites in the MRM method that were not found in the samples or that were not included in the list of reference standards for this analysis are not shown on the table.

| Compound | Slope | Intercept | $R^2$ | LOD [g/L] |
|---|---|---|---|---|
| Acetate | 2.05E+07 | 2.82E+06 | 0.9952 | 0.4179 |
| cis-Aconitate | 1.84E+07 | 8.04E+05 | 0.9903 | 0.1046 |
| Citrate | 6.51E+06 | -1.11E+06 | 0.9984 | 0.0414 |
| Fumarate | 9.05E+06 | -2.83E+04 | 0.9993 | 0.0410 |
| Glucose | 3.53E+04 | 3.28E+04 | 0.9916 | 1.7177 |
| Glycine | 1.45E+06 | -5.28E+04 | 0.9956 | 0.1135 |
| L-Alanine | 2.72E+07 | -1.60E+05 | 0.9997 | 0.0293 |
| L-Arginine | 1.20E+08 | 1.62E+06 | 0.9953 | 0.0375 |
| L-Aspartate | 1.35E+07 | 7.91E+05 | 0.9903 | 0.1201 |
| L-Glutamate | 1.00E+07 | 1.25E+06 | 0.9847 | 0.1405 |
| L-Glutamine | 4.93E+08 | 6.89E+06 | 0.9919 | 0.0335 |
| L-Isoleucine | 5.57E+07 | 4.45E+05 | 0.9986 | 0.0463 |
| L-Lactate | 5.06E+06 | -2.97E+04 | 0.9994 | 0.0277 |
| L-Leucine | 4.84E+07 | 6.12E+05 | 0.9985 | 0.0609 |
| L-Lysine | 6.13E+08 | 6.43E+06 | 0.9923 | 0.0252 |
| L-Methionine | 3.22E+08 | 6.74E+06 | 0.9980 | 0.0589 |
| L-Phenylalanine | 1.02E+09 | 1.49E+07 | 0.9979 | 0.0575 |
| L-Proline | 1.02E+09 | 7.02E+07 | 0.9919 | 0.1660 |
| L-Threonine | 4.76E+08 | 6.88E+06 | 0.9976 | 0.0623 |
| L-Tryptophan | 2.39E+08 | 6.82E+06 | 0.9934 | 0.0998 |
| L-Tyrosine | 1.20E+08 | 2.20E+06 | 0.9920 | 0.0689 |
| L-Valine | 2.59E+08 | 5.62E+06 | 0.9983 | 0.0821 |
| Pyruvate | 7.39E+06 | -4.10E+04 | 0.9998 | 0.0189 |
| Ribose-5P | 1.08E+06 | -7.08E+04 | 0.9818 | 0.1998 |
| S-Malate | 5.97E+07 | -4.64E+05 | 0.9998 | 0.0208 |
| Sedoheptulose-7P | 1.55E+06 | -7.17E+03 | 0.9901 | 0.0122 |
| Succinate | 6.17E+07 | 5.63E+06 | 0.9991 | 0.2940 |
| Xylulose-5P | 6.75E+04 | 4.10E+03 | 0.9275 | 0.2569 |

Looking at Table 5.1, the majority of metabolites have a coefficient of determination ($R^2$) higher than 0.99 and the regression curves look linear over the concentration ranges selected, which are desirable characteristics of calibration curves. The LODs were calculated based on the mass spectrometry signal at the intercept of the curve and the standard deviation on the y-direction. However, it is worth pointing that these calculated LODs were, on average, five times higher than the minimum concentration reliably[1] measured for each metabolite, with the exception of citrate. For example, glucose has an LOD of 1.72 g/L but the calibration curve was linear down to 1.23 g/L, L-phenylalanine has an LOD of 0.06 g/L but the calibration curve was linear down to 0.01 g/L, and L-tryptophan has an LOD of 0.10 g/L but the calibration curve was linear down to 0.01 g/L.

The concentration of metabolites in the off-line samples was used to model the metabolite concentrations on-line using on-line metabolomics, as it will be detailed in the next section.

### Concentration of on-line samples

The three fermentation replicates were monitored with on-line metabolomics and the sampling was done via a biomass filtration probe. The concentration of metabolites measured off-line was correlated to the on-line mass spectrometry signal. First, the on-line mass spectrometry data was smoothed using LOESS (see Figure 3.12 for a LOESS example). Then, time point alignment between the off-line and on-line data was used to correlate metabolite[2] concentration with the respective on-line mass spectrometry intensity. A total of 43 data points from each data sets (on-line and off-line) were used, dictated by the number of off-line data points available (*ca.* 15 per fermentation).

Then, to build the linear regression model, the aligned data from the three fermentation replicates was randomly split into a training set with 70 % of the samples (31 data points) and a test set with the remaining 30 % of the samples (12 data points). The training set was used to fit the regression model using 10-fold cross-validation and the test set was used to check the model performance on data that was new to the model. The summary statistics of the linear regression are shown in Table 5.2. The regression models developed were used to estimate the concentration of metabolites based on the on-line measurements. Figure 5.8 shows the comparison between the measured concentrations determined by LC-MS and the estimated concentrations from the on-line metabolomics signal for both the train set (cross-validation) and test set.

---

1. The measurement was in the linear range of the calibration curve, determined by a duplicate injection.
2. Biomass WCW concentration was correlated to (R)-2,3-dihydroxy-isovalerate, which was identified in Chapters 3 and 4 as a potential biomass biomarker by untargeted metabolomics. Although biomass is not a metabolite, it will be referred to as one to facilitate reading.

**Table 5.2:** Parameters and summary statistics of the linear regression calibration curves correlating the off-line concentration and the on-line mass spectrometry signal for the 12 metabolites and the biomass quantified. Metabolites not quantified because the concentration was not in range with the calibration curve are not reported.

| Compound | Slope | Intercept | $R^2$ | RMSECV [g/L] | RMSEP [g/L] | RMSRECV [%] | RMSREP [%] |
|---|---|---|---|---|---|---|---|
| Fumarate | 7.53E+07 | -6.54E+05 | 0.9748 | 0.0035 | 0.0046 | 18.78 | 24.19 |
| Glucose | 1.79E+04 | 4.66E+04 | 0.8181 | 4.6857 | 5.9220 | 48.56 | 49.86 |
| Glycine | 9.71E+04 | -2.46E+03 | 0.4001 | 0.0233 | 0.0206 | 43.23 | 38.11 |
| L-Alanine | 2.76E+06 | 2.45E+04 | 0.8763 | 0.3708 | 0.4063 | 84.60 | 75.94 |
| L-Lactate | 1.64E+06 | 9.66E+04 | 0.9505 | 0.0403 | 0.0341 | 27.09 | 20.75 |
| L-Valine | 3.05E+07 | 1.48E+05 | 0.9032 | 0.0372 | 0.0421 | 70.76 | 64.02 |
| Pyruvate | 1.57E+06 | 2.24E+04 | 0.9928 | 0.2546 | 0.2497 | 17.99 | 14.95 |
| Ribose-5P | -3.08E+05 | 2.78E+04 | 0.5662 | 0.0174 | 0.0181 | 25.04 | 26.82 |
| S-Malate | 4.32E+06 | 6.87E+04 | 0.8151 | 0.0879 | 0.1539 | 106.42 | 167.49 |
| Sedoheptulose-7P | -8.60E+04 | 3.02E+03 | 0.5148 | 0.0215 | 0.0271 | 313.00 | 392.76 |
| Succinate | 6.73E+06 | 4.93E+05 | 0.9930 | 0.4057 | 0.2368 | 23.72 | 12.17 |
| WCW | 1.33E+05 | -1.93E+04 | 0.8536 | 4.0428 | 4.5483 | 26.08 | 25.09 |
| Xylulose-5P | -2.28E+04 | 7.70E+03 | 0.6357 | 0.2349 | 0.2420 | 446.70 | 937.96 |

Fumarate, L-lactate, pyruvate and succinate show a good linear relationship between measured and preidcted values in Figure 5.8 and have root mean squared relative error of cross-validation (RMSRECV) and root mean squared relative error of prediction (RMSREP) values close to 15-25 %. The relative error gives an overall indication of how big the RMSE of a metabolite is in comparison to its average concentration. Furthermore, these four metabolites - together with L-valine - have an $R^2$ higher than 0.9, although some predictions for L-valine were clearly overestimating the measured concentration. Glucose, L-alanine, S-malate and WCW biomass have an $R^2$ between 0.8 and 0.9. The root mean squared relative error (RMSRE) of glucose and WCW are between 25 and 50 %, whereas for L-alanine and S-malate are between 75 and 170 %. For L-alanine and S-malate, a similar overestimation to the one that was found for L-valine was observed. Further investigation revealed that the overestimated values for these three metabolites all belonged to the first fermentation replicate, potentially caused by a difference in instrument calibration between the three fermentation runs. Finally, the models for glycine, ribose-5P, sedoheptulose-7P and xylulose-5P were not satisfactory, with low $R^2$ values between 0.4 and 0.6, even though glycine and ribose-5P had a relatively low RMSRE (25-40 %). Sedoheptulose-7P and xylulose-5P had very high RMSRE values, higher than 300 %.

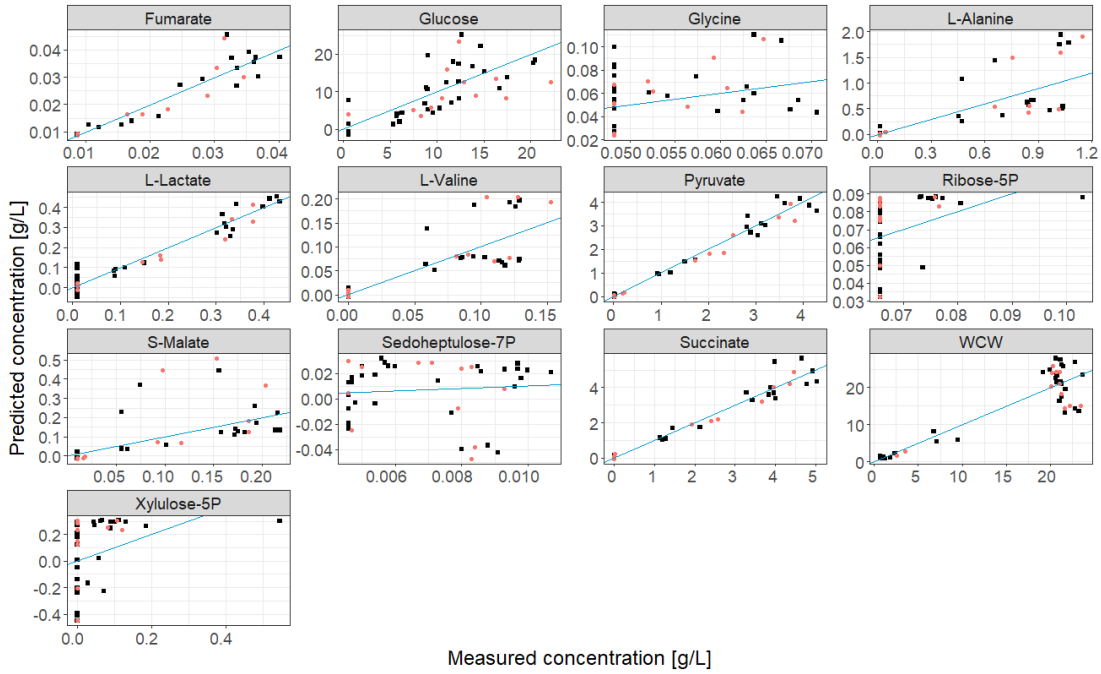



**Figure 5.8:** Comparison of compound concentration measured by LC-MS and predicted by on-line metabolomics for the cross-validation training set (**black squares**) and test set (**red circles**). Perfect matches between measured and predicted concentrations would lie on the blue line, which has a slope of 1.

These models had lower $R^2$ values for all metabolites compared to the calibration curves built to determine the off-line concentration using reference standards (Table 5.1). This effectively means that in the on-line metabolomics models, a lower percentage of the total data variance is explained by the model (that is, a higher percentage of the total variance is explained by noise). A few reasons could explain that. One reason could be that both the reference standards and the off-line samples were run using chromatography (LC-MS), whereas the on-line metabolomics system is operated without chromatography. Although the QQQ selects specific ions of interest to reach the detector one at a time, without chromatography all molecules reach the ESI source at the same time, causing some ion suppression, which can compromise the linearity between metabolite concentration and mass spectrometry signal (noise). Furthermore, contamination of the MRM method is also possible. That is, without chromatography, there is a greater chance of coeluting compounds with the same m/z value interfering with the analyte of interest, for example isomers, adducts and in-source fragments of larger molecules. Another reason could be that the reference standards and the off-line samples were run all together on the same analytical batch, whereas the on-line metabolomics experiments were run separate from the reference standards, on three different days. The batch-to-batch instrumental variation could also introduce some noise to the calibration model.

Compared to similar models from the literature, do Nascimento *et al.* (2017) developed a NIR spectroscopy PLS model to monitor an ethanol fermentation and respectively reported for biomass and glucose $R^2$ of cross-validation of 0.978 and 0.920, RMSECV values of 0.38 and 4.65 g/L and RMSEP values of 0.28 and 7.14 g/L. Although their biomass predictions were more accurate, the glucose ones were similar to the ones presented here. Vann *et al.* (2017) also used a NIR spectroscopy PLS model to monitor an ethanol fermentation and reported an $R^2$ of 0.64 for biomass, which is lower to the one obtained in this chapter.

Rodrigues *et al.*, 2018 developed a MIR spectroscopy PLS model to monitor an ethanol fermentation and respectively reported for biomass and glucose an $R^2$ of 0.953 and 0.998, RMSECV values of 2.78 and 0.79 g/L and RMSEP values of 1.64 and 1.06 g/L, which are significantly better than the predictions from this chapter. M.-Y. Li *et al.* (2018) built a Raman spectroscopy PLS model to monitor a process of monoclonal antibody production using CHO cells and respectively reported for total and non-glycosylated antibody $R^2$ of cross-validation values of 0.88 and 0.89 and RMSRECV values of 9 and 7 % and RMSREP values of 5 and 6 % (although these RMSREs were calculated with respect to the highest concentration of antibody instead of the average concentration, obtaining a lower relative error).

Overall, these examples show that similar predictive models reported in the literature can also have $R^2$ values below 0.99 and similar relative errors to the ones reported here for on-line metabolomics. Although the models in the literature examples might seem slightly better (higher $R^2$ and smaller errors), the on-line metabolomics models developed in this thesis have the great advantage that are simple linear models, rather than multivariate PLS models. These PLS models try to use the information from many variables to model each metabolite. In this process, a bigger percentage of the variance might be explained (higher $R^2$), but there is a high risk of model overfitting by correlating a metabolite signal to non-related physical properties. For example, Surribas, Geissler *et al.* (2006) correlated fluorescence to glycerol production (even though glycerol is not a fluorophore) by PLS because glycerol production was linked to biomass formation (biomass produces a fluorescent signal). Indeed, as it has been previously mentioned, these multivariate chemometric models (like PLS) have little transferability and are only robust in the conditions they are developed (see Section 1.1.4). Simpler linear models like the on-line metabolomics ones developed here are expected to be more flexible and transferable to other configurations (different strain, bioreactor, organism, fermentation medium, temperature, *etc.*).

Schenk *et al.*, 2008 claimed that using a library of pure component spectra it is possible to reduce the time to build IR chemometric models and increase their transferability. However, the authors tested all their fermentations with the same temperature and pH conditions as they built the library, and one of the three different *E. coli* strains they used resulted in unsatisfactory predictions (presumably because the strain produced a different compound during growth, according to the authors). All together, it is questionable whether this method really increases the transferability of vibrational spectroscopy chemometric models.

The model diagnostics from Table 5.2 can be used to quantitatively assess model performance and compare it with different models. Another way to check the performance of the model predictions is to plot the metabolite concentrations predicted from the on-line mass spectrometry data and overlay on top of it the extracellular concentrations measured off-line (Figure 5.9). These on-line predictions generally show similar results to the off-line measurements with the exception of glycine, ribose-5P, sedoheptulose-7P and xylulose-5P. However, on-line metabolomics reveals that these metabolites show an exponential increase during the biomass growth phase, which was not captured by the off-line extracellular LC-MS data. This could be because the concentration of these metabolites was close to their LOD (Table 5.1). Therefore, for these metabolites, the discrepancies of the two methods, the low $R^2$ values and high RMSE might be caused by this detection limit during the off-line analysis, rather than being a limitation of the on-line monitoring system. The poor correlation for sedoheptulose-7P and xylulose-5P also resulted in predictions of negative concentrations, which might also be a consequence of their poor calibration curves caused by the concentration of these metabolites in the off-line samples being too small, especially during the batch phase.

Another important observation to make from Figure 5.9 is that, although the total travelling time from the bioreactor to the mass spectrometer was 31.5 minutes (due to the low flow rate required for the use of the filtration probe, see Section 5.4.3), there is no evidence that the sample was transformed during this time. Firstly, products of anaerobic fermentation such as fumarate, pyruvate and succinate were not observed during the aerobic biomass formation phase - which indicates that the sample did not suffer any significant biological transformation during this travelling time - and secondly, there is a good correlation between the off-line and on-line signals, particularly for compounds such as fumarate, pyruvate and succinate. However, it is important to bear in mind that this long travel time might become an issue for fermentation processes where the conversion of some substrates takes place in the extracellular space, for example in the hydrolysis of starch using extracellular amylases (Favaro *et al.*, 2015). In these cases, the on-line system should be used without the filtration probe, or it should be adapted to reduce the travel time (*e.g.* optimising the tubing system or using a flow-through cell as described by Koch *et al.* (2016)).

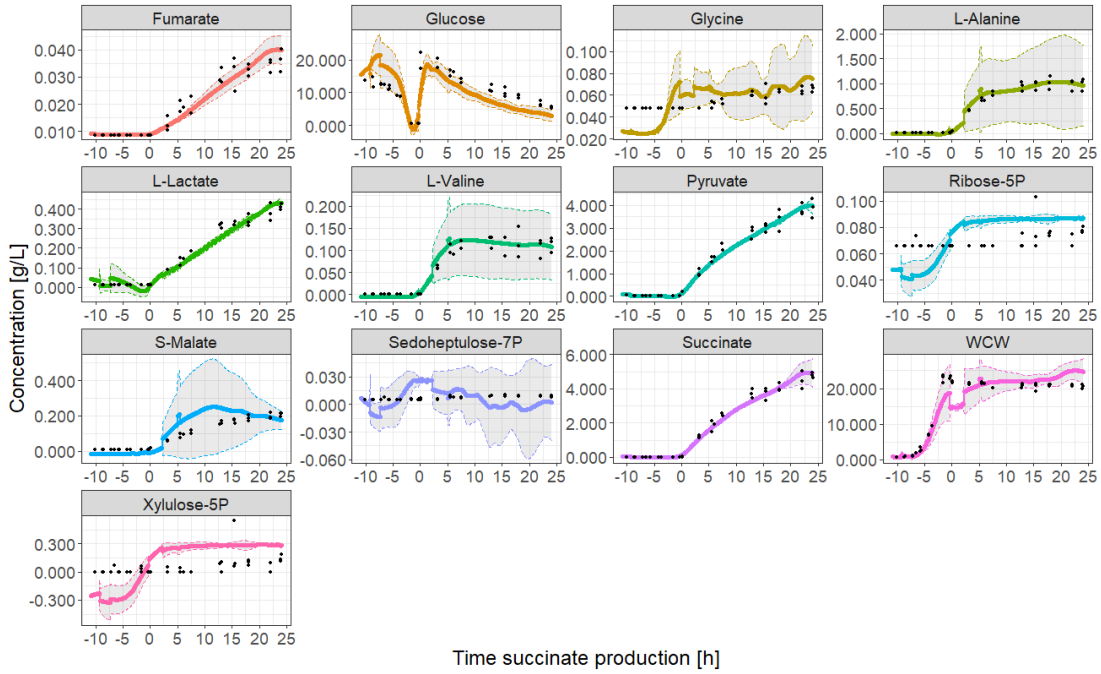



**Figure 5.9:** On-line concentration determined by metabolomics (**coloured lines**) using the calibration models for the three fermentation replicates. The coloured line is the average of the three replicates and the 95 % confidence intervals are shown with a grey area. The measured extracellular off-line data for the three replicates is marked with black circles. Time is indicated with respect to the beginning of the succinate production phase for a better comparison of the three replicates.

Figure 5.9 shows the monitored concentration of targeted metabolites calculated from the on-line mass spectrometry data. However, there were 27 other metabolites in the MRM method that were also monitored on-line even though their concentration could not be quantified for three possible reasons: either because their respective calibration curves did not cover a wide-enough range, because the compounds were not included in the list of reference standards analysed, or because the compounds were not found in the off-line samples. The monitoring of these metabolites is shown in Figure 5.10 and, although the concentration of these metabolites could not be determined, their mass spectrometry signal (ion intensity) and time evolution is a very informative measurement of relative quantitation. For example, Figure 5.10 shows how the mass spectrometry signal for some sugar phosphates involved in the PPP - such as 6-phosphogluconate, erythrose-4P and fructose-6P - rapidly increases soon after the beginning of the succinate production phase. An exponential increase in the abundance of many amino acids during the biomass formation phase is also observed in Figure 5.10, such as L-arginine, L-asparagine, L-aspartate, L-glutamate, L-glutamine, L-isoleucine and L-leucine, whereas other amino acids experience an increase in abundance more towards the succinate production phase, for example L-histidine, L-phenylalanine, L-proline, L-serine, L-threonine, L-tryptophan and L-tyrosine.

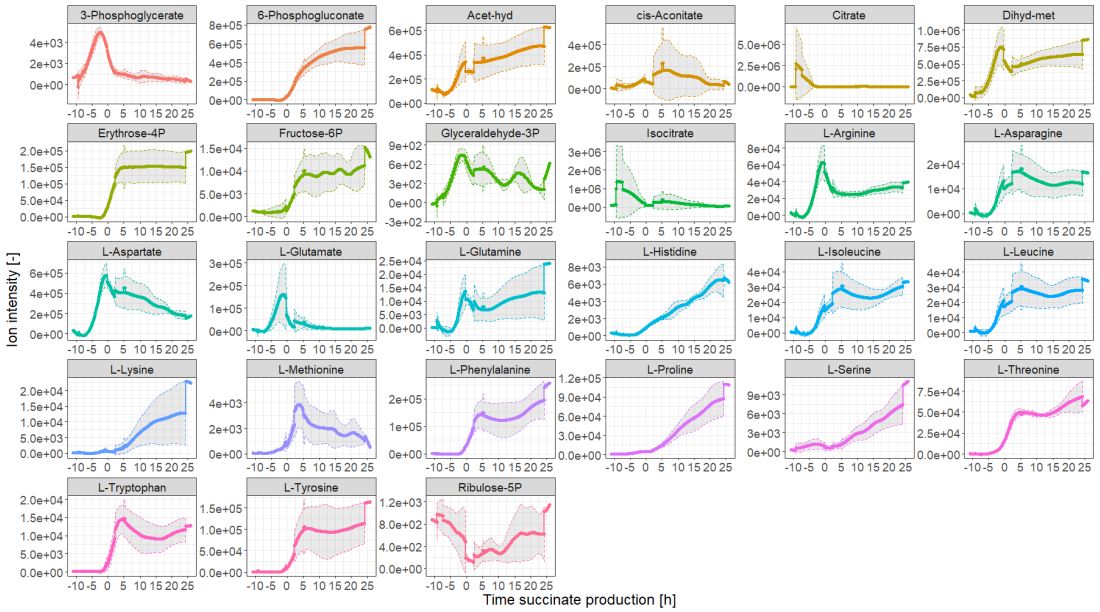



**Figure 5.10:** On-line ion intensity signal measured with metabolomics. The coloured line is the average of the three replicates and the 95 % confidence intervals are shown with a grey area. Time is indicated with respect to the beginning of the succinate production phase for a better comparison of the three replicates. Abbreviations: Acet-hyd: (S)-2-aceto-2-hydroxybutanoate; Dihyd-met: (R)-2,3-dihydroxy-3-methylpentanoate.

In summary, these results show how on-line metabolomics can be used for absolute and relative quantitative monitoring of the fermentation process, even though no chromatography is used. As mentioned earlier, the lack of chromatography could compromise the linearity of the mass spectrometer response to the concentration of metabolites due to ion suppression in the ESI source. However, the results show that good linear responses can be observed.

## 5.3 Conclusions

This chapter focused on building an on-line targeted metabolomics method with a triple quadrupole, using it to monitor metabolites of interest for the succinate fermentation process and the concentration of *E. coli* cells, and to build regression models between the on-line mass spectrometry signal of these metabolites and their concentration in the bioreactor.

An MRM method was developed for targeted analysis of 41 metabolites in the QQQ mass spectrometer. This includes compounds of glycolysis, the TCA cycle, the PPP, metabolites of anaerobic fermentation, amino acids and biomass biomarkers identified in Chapters 3 and 4, all of interest for the succinate fermentation process. At least two fragment ions were identified for most metabolites on the MRM method and a few examples of possible fragmentation mechanisms were proposed.

Three succinate fermentation replicates were monitored with on-line targeted metabolomics using the MRM method. Taking extracellular off-line samples throughout the fermentation it was possible to quantify 14 metabolites from the MRM method. The quantified off-line concentration for these 14 metabolites was used to create univariate linear regression models to correlate the on-line metabolomics signal of 12 of these metabolites to their concentration, as well as (R)-2,3-dihydroxy-isovalerate, a biomass biomarker identified in Chapters 3 and 4, which was correlated to the WCW concentration of biomass. The other metabolites from the MRM that were found in the samples could be qualitatively monitored throughout the fermentation. Twenty-seven compounds from the MRM method were qualitatively monitored on-line, even though their concentration could not be reliably measured. This was caused by either the calibration curve not covering a long enough range of analyte concentration, the compounds not being found in the off-line samples, or because the compounds were not included in the list of reference standards used i the off-line analysis. Therefore, it is important to note that additional calibration curves and adding more metabolites to the MRM method would allow the quantitative monitoring of many more compounds by on-line metabolomics.

To build the univariate linear regression models, the data of the three fermentation replicates was randomly split into a training set and a test set. The training set was used to build the models using 10-fold cross validation, and model performance was assessed on the cross-validation and test data. These on-line metabolomics univariate models performed comparably to the vibrational spectroscopy models reported in the literature, which typically are multivariate PLS regressions. Although the predictions of vibrational spectroscopy generally had higher $R^2$ and lower RMSE values, this is to be expected, as multivariate PLS regression will naturally explain a larger proportion of the experimental variance (by using more variables to fit the model). However, this has a higher risk of overfitting, which makes these models less robust to changes in the bioprocess (organism, temperature, medium, *etc.*). The univariate linear models built for on-line metabolomics, on the other hand, are much simpler and more direct, only correlating the concentration of a metabolite to its single mass spectrometry peak area. Therefore, the models for on-line metabolomics monitoring developed in this chapter are simpler, easier to build and should be more robust to changes in the bioprocess.

## 5.4 Materials and methods

The materials and methods are as described in Chapter 2. Any differences and particular methods used exclusively in this chapter are described below.

### 5.4.1 Building the on-line metabolomics model of metabolite concentration

A schematic of the process followed to build the on-line metabolomics model of metabolite concentration is described in Figure 5.2. In short, three replicate fermentation experiments were performed. Off-line extracellular samples were taken every 1-2 h and on-line measurements were automatically taken every 5 min. Off-line samples were processed and analysed as described in Section 2.7.3 and metabolite concentrations were determined using calibration curves built using reference standards (see Section 5.4.5).

To determine the concentration of metabolites from the on-line metabolomics data a second round of univariate linear models were built, this time using the metabolite concentration of the off-line samples as the independent variable (x) and on-line mass spectrometry peak area as the dependent variable (y). The on-line mass spectrometry data was previously smoothed to reduce the effect of outliers. As the number of data points in the on-line data set is about 27 times larger than the off-line one (approximately 400 to 15), both data sets were aligned by the time of fermentation and only the 15 data points of the on-line metabolomics data set that corresponded to the 15 off-line samples were used to build and test the regression models (43 data points in total with the three replicates).

The 43 data points aligned (off-line *v.s.* on-line) from the three fermentation replicates was randomly split into a training set with 70 % of the samples (31 data points) and a test set with the remaining 30 % of the samples (12 data points). The training set was used to create the correlation using 10-fold cross-validation (Section 5.4.6) and the test set was used to check the model performance on data that was new to the model. The performance of the correlation was assessed with the $R^2$, RMSECV, RMSEP, RMSRECV and RMSREP.

#### Training set

The training set for building the correlation model and for cross-validation consisted of the following randomly selected samples (31 in total):

- Ferm 1: sample 1, 2, 4, 5, 6 (batch phase), 7, 8, 11, 13 and 14 (production phase).
- Ferm 2: sample 1, 2, 3, 4 (batch phase), 7, 8, 10, 11, 13 and 14 (production phase).
- Ferm 3: sample 1, 2, 3, 5, 6, 7 (batch phase), 9, 11, 12, 13 and 15 (production phase).

#### Test set

The test set for testing the correlation model consisted of the following randomly selected samples (12 in total):

- Ferm 1: sample 3 (batch phase), 9, 10 and 12 (production phase).
- Ferm 2: sample 5 (batch phase), 6, 9 and 12 (production phase).
- Ferm 3: sample 4 (batch phase), 8, 10 and 14 (production phase).

### 5.4.2 Fermentation process conditions

All fermentation experiments were carried out in a 5 L Applikon stirred tank fermenter controlled with a FerMac 320 (Electrolab Biotech Ltd.) coupled to a FerMac 368 gas analyser for off-gas analysis. The fermentation consisted of a dual-phase succinate production process as described in Chapter 2. A filtration probe (TRACE Analytics GmbH) equipped with a polypropylene membrane was installed in the fermenter so the on-line metabolomics samples were free of cells.

#### Inoculum

Fermentation inocula were prepared by inoculating 25 $\mu$L of cell bank into 100 mL of growth medium in a 500 mL baffled shake flask and incubated at 37 ℃ and 215 rpm for 18.14 $\pm$ 0.52h.

#### Aerobic batch phase for biomass growth

As described in Chapter 2 but the starting $OD_{600}$ in the fermenter after inoculation was 0.20 +/- 0.04 and the agitation during the fermentation was 600-1200 rpm, as 1200 rpm was the maximum speed possible with the FerMac 320 control tower.

#### Anaerobic succinate production phase

As described in Chapter 2 but agitation was maintained at 600 rpm to avoid fouling of the biomass filtration probe.

### 5.4.3 On-line metabolomics method

As described in Chapter 2 but using a biomass filtration probe (TRACE Analytics GmbH) equipped with a polypropylene membrane submerged in the fermentation broth. Therefore, the samples analysed should consist of cell-free medium. The Masterflex$^{TM}$ L/S ® (Cole-Parmer) high-performance pump was operated at a low flow rate of 1.5 mL/min - as indicated by the filtration probe manufacturer - to prevent cells from entering the filtration probe. Due to the slow flow rate of the pump, the total travelling time from the bioreactor to the mass spectrometer was 31.5 minutes.

### 5.4.4 Missing value imputation

Metabolites with missing data values were imputed using the half of the minimum intensity detected for that metabolite (Wei *et al.*, 2018). These missing values are common in metabolomics and might happen due to different reasons, for example the metabolite being below the limit of detection or by errors occurring during data acquisition or pre-processing.

### 5.4.5 Calculating metabolite concentration from mass spectrometry data

Univariate linear models were built to calculate metabolite concentration from mass spectrometry data. A series of dilutions of reference standards of known concentrations were used to build the models; concentration data was used as the independent variable (x) and mass spectrometry peak area the dependent variable (y). The relationship between both variables was modelled by linear regression using the least squares method in Microsoft® Excel® (Version 2201) using the LINEST function. Each dilution of reference standard was analysed by LC-MS in duplicate. To determine the concentration of an off-line sample, the mass spectrometry peak area is measured by LC-MS, then, the metabolite concentration can be directly calculated using equation 5.1.

$$Concentration = \frac{MS_{signal} - b_{cc}}{m_{cc}} \tag{5.1}$$

where $MS_{signal}$ is the area under the peak measured by mass spectrometry and $b_{cc}$ and $m_{cc}$ are the intercept and slope of the calibration curve built with the reference standards, respectively. Negative concentration values were imputed as zeros. The reference standards and the off-line fermentation samples were analysed on the same LC-MS run.

### 5.4.6 Model cross-validation

Ten-fold cross-validation was performed using the caret package (Kuhn *et al.*, 2021) in the statistical software environment R (Version 4.1.2) using the "trainControl" function with the "repeatedcv" method with three repeats. Therefore, 10-fold cross-validation was performed three times splitting the training data set in three different ways and selecting the best performing cross-validation split (*i.e.* the one with the lowest RMSE).

## 5.5 Filtration efficiency

Filtration efficiency was determined by plating 100 $\mu$L of filtered and unfiltered fermentation samples on selective LB agar plates containing 100 mg/L kanamycin and 34 mg/L chloramphenicol to prevent the growth of contaminants. Before plating, a series of dilutions was prepared for both samples. Filtered samples were plated neat (undiluted) and up to $10^3$x diluted. Unfiltered samples were plated $10^6$-$10^8$x diluted. After two days of incubation at 37 ℃, the colony forming units (cfu) on the plates were used to calculate the filtration efficiency using equation 5.2.

$$Filtration_{efficiency}[\%] = \frac{cfu_{unfiltered} \cdot 10^{dil_{unfiltered}} - cfu_{filtered} \cdot 10^{dil_{filtered}}}{cfu_{unfiltered} \cdot 10^{dil_{unfiltered}}} \cdot 100 \qquad (5.2)$$

where $dil_{filtered}$ and $dil_{unfiltered}$ are the dilution factors used in the plating of the filtered and unfiltered samples, respectively.

Filtration efficiency was measured at the two last points of the fermentation, when the biomass concentration was high and probe fouling more likely.

Chapter 6

# Modelling the *Escherichia coli* succinate fermentation process and data forecasting


## Abstract

In Chapter 5, univariate linear regression models were developed for the quantitative monitoring of metabolites in an *Escherichia coli* (*E. coli*) dual-phase succinate production fermentation process with on-line metabolomics. Three fermentation replicates were used to build these monitoring models. In this chapter, a kinetic model was developed to describe the dynamic evolution in time of the main compounds of this bioprocess during the biomass formation phase: glucose, biomass and acetate. The model parameters were estimated by minimising the root mean squared relative error between the measured off-line fermentation data and the corresponding values predicted by the model using nonlinear regression. Some examples of how the interpretation of the fitted model parameters can improve the understanding of the bioprocess were given, for example with the identification of yields, production rates and affinity constants. Finally, the kinetic model was used to forecast the evolution of glucose, biomass and acetate at three different time points of the biomass formation batch phase from the on-line metabolomics data generated in Chapter 5. This simulates a scenario of real-time forecasting the evolution of these compounds during process monitoring using on-line metabolomics.

Overall, the model was able to describe and predict different biological phenomena such as acetate accumulation during *E. coli* growth on glucose, acetate consumption upon glucose depletion and the switch from biomass exponential growth to stationary phase after consumption of both carbon sources. One of the major factors affecting the quantitative accuracy of the model predictions is the quality of the on-line metabolomics measurements. That is, when the on-line glucose measurements were overestimated, a larger predictive error was found on the biomass and acetate predictions.

## 6.1 Introduction

Modelling is a way of mathematically describing a system of interest with a set of parameters. In fermentation analysis, mathematical models are typically used to describe the evolution of the different important process compounds in time and their stoichiometric relation, which can provide insightful information to better understand, control and improve the process. For example, Salvy and Hatzimanikatis (2021) modelled the diauxic behaviour of *E. coli* growth on a mixture of glucose and lactose, and Anane *et al.* (2017) modelled acetate accumulation due to overflow metabolism in an *E. coli* fed-batch process. Birou *et al.* (1987) and Maskow and Paufler (2015) are two examples of how stoichiometric models can be used with thermodynamics to monitor the concentration of biomass using calorimetry, and Klimek and Ollis (1980) developed a model for substrate consumption, biomass growth and exopolysaccharyde (product) formation for *Pseudomonas sp.*, *Azotobacter vinelandii*, and *Aureobasidium pullulans*. These models are often based on stoichiometry, enzyme kinetics and differential equations.

Stoichiometric equations provide the quantitative relation between the different main process compounds, typically using substrate- or biomass-specific yields (Rieger *et al.*, 1983; Warth *et al.*, 2010; Wechselberger *et al.*, 2010). These reactions describe biomass - and sometimes also product - formation as a process requiring carbon, nitrogen, oxygen and hydrogen (equation 6.1).

$$\mathrm{C_6H_{12}O_6} + Y_{o/s}\mathrm{O_2} + Y_{n/s}\mathrm{NH_4^+} -> Y_{x/s}C_1H_aO_bN_c + Y_{c/s}\mathrm{CO_2} + Y_{w/s}\mathrm{H_2O} + Y_{h/s}\mathrm{H^+} \qquad (6.1)$$

where $Y_{i/s}$ is the yield of production or consumption of compound *i* on substrate and $C_1H_aO_bN_c$ is an elemental formula commonly used to represent dried biomass normalised to one carbon atom. This formula has been reported for several organisms (Roels, 1980; Von Stockar and Liu, 1999), including average microbial values such as $C_1H_{1.8}O_{0.5}N_{0.2}$. Elemental, charge and redox balances can be used to solve these stoichiometric reactions.

Regarding enzyme kinetics, asymptotic functions are often used to describe enzymatic reaction rates as a function of the concentration of their substrates. Particularly, substrate consumption is commonly described using Michaelis-Menten kinetics (Michaelis and Menten, 2013) (equation 6.2) and biomass formation is often described using the Monod model (Monod, 1949) (equation 6.3). Modifications of the Monod model have been proposed to account for phenomena such as product inhibition (Levenspiel, 1980) (equation 6.4) or substrate inhibition (Krishnan *et al.*, 1999). Similarly, an expression for diauxic growth on two substrates can be described based on the expression used by Sonnleitner and Käppeli (1986), who expressed the inhibition of the growth rate on ethanol in *S. cerevisiae* caused by the presence of glucose. This expression can be applied to the substrate consumption rate (equation 6.5) and it contains an inhibitory effect on the consumption of the secondary substrate, particularly

when $S_1 > K_{S1}$.

$$q_S = q_{Smax} \cdot \frac{S}{S + K_S} \tag{6.2}$$

$$\mu = \mu_{max} \cdot \frac{S}{S + K_S} \tag{6.3}$$

$$\mu = \mu_{max} \cdot \frac{S}{S + K_S} \cdot \left(1 - \frac{P}{K_P}\right)^c \tag{6.4}$$

$$q_{S_2} = q_{S_2max} \cdot \frac{S_2}{S_2 + K_{S_2}} \cdot \frac{K_{S_1}}{S_1 + K_{S_1}} \tag{6.5}$$

where $q_S$ is the biomass-specific consumption rate in mol S/(C-mol X · h), (S means substrate and X biomass), $q_{S\,max}$ is the maximum biomass-specific consumption rate in mol S/(C-mol X · h), S is the concentration of substrate in mol S/L, $K_S$ is the half-saturation constant of substrate consumption in mol S/L, $\mu$ is the biomass-specific formation of biomass in $h^{-1}$, $\mu_{max}$ is the maximum biomass-specific formation of biomass in $h^{-1}$, P is the concentration of product in mol P/L, $K_P$ is the limiting concentration of inhibitory product in mol P/L, *c* is a power constant, $S_1$ is the concentration of the preferential substrate under diauxic conditions, $K_{S_1}$ is the half-saturation constant of $S_1$, $S_2$ is the concentration of the secondary substrate under diauxic conditions and $K_{S_2}$ is the half-saturation constant of $S_2$.

These kinetic expressions are used in ordinary differential equations representing the mass balance of the corresponding compounds. These differential equations describe the change in time of the amount of a substance as a function of the inputs and outputs of the bioreactor and the reaction rate. For a batch process like the *E. coli* succinate fermentation studied in this thesis, there are no inputs and outputs in the liquid phase and, therefore, the rate of change of a given compound is determined by its reaction rate. Equation 6.6 shows the ordinary differential equation for substrate in mol S/h, with the rate being expressed as the biomass-specific rate multiplied by the moles of biomass ($N_X$). Alternatively, the rate can be expressed using stoichiometric yields, as shown in equation 6.7. A specific behaviour of product formation was introduced by Luedeking and Piret (1959) while studying lactic acid fermentation with *Lactobacillus delbrueckii*. The authors described product formation as a process with two components: one proportional to the biomass growth rate and another independent from it (equation 6.8).

$$\frac{dN_S}{dt} = q_S \cdot N_X \tag{6.6}$$

$$\frac{dN_X}{dt} = (Y_{X/S} \cdot q_S - k_d) \cdot N_X \tag{6.7}$$

$$\frac{dN_P}{dt} = (\alpha \cdot \mu + \beta) \cdot N_X \tag{6.8}$$

where $N_i$ is amount of *i* in moles, $k_d$ is a term for biomass decay rate in $h^{-1}$, $\alpha$ is the factor proportional to the biomass growth rate in mol P/C-mol X and $\beta$ is the factor independent from the biomass growth rate in mol P/(C-mol X · h).

Models based on mechanistics and first-principles like the ones described above (matter and energy conservation laws, thermodynamics, kinetics, *etc.*) are called parametric models - also known as fundamental or mechanistic models (von Stosch *et al.*, 2014). These use parameters with a physical meaning, for example the enthalpy of formation of a substance or the half-saturation constant of an enzyme. It is important to mention that other types of models can also be used such as non-parametric and semi-parametric models (von Stosch *et al.*, 2014). Non-parametric models - also known as data-driven models - are obtained by purely fitting experimental data, therefore the parameters they use have no physical meaning. For example, Costello and Martin (2018) modelled process dynamics of several metabolites of the mevalonate pathway in *E. coli* using machine learning with metabolomics and proteomics data as the model input, rather than using mechanistic kinetic equations. Finally, semi-parametric models - also known as hybrid models - are a mix of the two. For example, L. H. Harada *et al.* (2002) described a *S. cerevisiae* fermentation process with kinetic and mass balance expressions (parametric) and modelled the biomass growth rate using a neural network using the concentration of substrate, ethanol and temperature as inputs (data-driven).

In this chapter, the focus will be put on modelling the evolution of the key fermentation compounds during the biomass formation phase of the *E. coli* succinate bioprocess using some of the mechanistic descriptions mentioned in this introduction. The kinetic parameters will be fitted minimising the total RMSRE in a data-drive manner, thus resulting in a hybrid model. A bit more details on parameter fitting are given below.

### 6.1.1 Parameter fitting

Fitting parameters for bioprocess modelling is commonly done with least squares minimisation (equation 6.9), which consists in finding those parameters ($\boldsymbol{p}$) that minimise the squared difference between the measured values ($\boldsymbol{y_i}$) and the estimated values by the model ($\boldsymbol{\hat{y}_i(p)}$), which is a function of the fitted $\boldsymbol{p}$ model parameters. For example, Choi *et al.* (2014) developed a kinetic model of a lactic acid bioprocess using *Lactobacillus rhamnosus* grown with date juice, using least squares minimisation to fit the kinetic parameters. However, the minimisation function contains multiple compounds (*e.g.* substrate, biomass and product), which not always have the same concentration magnitude. Consequently, to avoid the minimisation process to favour those compounds with the highest concentration, weighing factors are often applied to each compound of the minimisation function. For example, Sadino-Riquelme *et al.* (2020)

used weighted least squares minimisation to fit the kinetic parameters for a model of alginate production, substrate consumption and biomass formation with *Azotobacter vinelandii*. The weighing factors for each compound of the model was the highest concentration measured for that compound.

$$f_{obj} = min\left[\sum_{j=1}^{k}\left(\sum_{i=1}^{n}(y_{i,j} - \hat{y_{i,j}}(p))^2\right)\right] \tag{6.9}$$

where *i* are the different samples and *j* the different compounds included in the optimisation.

In this chapter, the RMSRE was used in the minimisation objective function as a way of producing weighted least squares (equation 6.10). Indeed, the RMSRE of a compound *j* calculates the squared difference between the measured and estimated values by the model and then it calculates the mean value across all the (*i*) data points (which is a way of correcting for data sets of different lengths). Then, the square root of the resulting value is calculated and, finally, the result is normalised as a percentage of the average measured value of that compound *j*. The overall objective function minimises the sum of the RMSRE of all the *j* compounds of interest (equation 6.22).

$$f_{obj} = min\left[\sum_{j=1}^{k}\left(100 \cdot \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{i,j} - \hat{y_{i,j}}(p))^2}}{\bar{y}_j}\right)\right] \tag{6.10}$$

where *i* are the different samples and *j* the different compounds included in the optimisation.

### 6.1.2 Aims

The aims of this chapter are:

- To develop a kinetic model describing the biomass production phase of the *E. coli* succinate fermentation process. The model should include acetate formation due to overflow metabolism and acetate consumption, as observed experimentally.
- To fit the parameters of the kinetic model using the three fermentation replicates used in Chapter 5.
- To use the kinetic model to forecast the evolution of glucose, biomass and acetate at three different time points of the biomass formation batch phase from the on-line metabolomics data generated in Chapter 5.

## 6.2 Results and discussion

In this chapter, kinetic models were developed to describe the dynamic evolution in time of the main compounds of the *E. coli* succinate fermentation during the biomass formation phase: glucose, biomass and acetate. This part of the fermentation was chosen (as opposed to the succinate production phase) as there are more interesting patterns for modelling and control, such as biomass growth and the accumulation of acetate during overflow metabolism, followed by acetate consumption after glucose starvation. On the other hand, the succinate production phase is a simpler stage of the process where succinate and pyruvate are produced as a linear function of the biomass concentration.

The process modelling consisted of two stages: first, the development of the kinetic model and second, the estimation of the model parameters. Finally, the model was used to forecast the evolution of glucose, biomass and acetate at three different time points of the biomass formation batch phase of the bioprocess using the on-line metabolomics data.

### 6.2.1 The kinetic model

The process reaction during aerobic biomass growth was described by two stoichiometries, the first one based on cell growth on glucose (equation 6.11), which is coupled to acetate formation due to overflow metabolism and the second one based on acetate consumption after glucose starvation (equation 6.12). The average elemental formula for dried biomass normalised to one carbon atom reported by Roels (1980) was used.

$$\mathrm{C_6H_{12}O_6} + Y_{o/s}\mathrm{O_2} + Y_{n/s}\mathrm{NH_4^+} -> Y_{x/s}CH_{1.8}O_{0.5}N_{0.2} + Y_{a/s}\mathrm{C_2H_3O_2^-} + Y_{c/s}\mathrm{CO_2} + Y_{w/s}\mathrm{H_2O} + Y_{h/s}\mathrm{H^+} \tag{6.11}$$

$$\mathrm{C_2H_3O_2^-} + Y_{Ace_{o/s}}\mathrm{O_2} + Y_{Ace_{n/s}}\mathrm{NH_4^+} -> Y_{Ace_{x/s}}CH_{1.8}O_{0.5}N_{0.2} + Y_{Ace_{c/s}}\mathrm{CO_2} + Y_{Ace_{w/s}}\mathrm{H_2O} + Y_{Ace_{h/s}}\mathrm{H^+} \tag{6.12}$$

where $Y_{i/s}$ is the yield of a substance *i* on glucose and $Y_{Ace_{i/s}}$ is the yield of a substance *i* on acetate. Note that these stoichiometries were only described to establish the relationship between the different compounds, but were not used to constrain the yields based on elementary and redox balances.

Once the stoichiometries were defined, the mass balances of the compounds of interest were defined with the ordinary differential equations 6.13- 6.15.

$$\frac{dN_{Glu}}{dt} = q_{Glu} \cdot N_X \tag{6.13}$$

$$\frac{dN_X}{dt} = (\mu - k_d) \cdot N_X \tag{6.14}$$

$$\frac{dN_{Ace}}{dt} = (q_{Ace_{cons}} + q_{Ace_{prod}}) \cdot N_X \tag{6.15}$$

where $q_{Ace_{cons}}$ is the biomass-specific acetate consumption rate and $q_{Ace_{prod}}$ is the biomass-specific acetate production rate, both in mol Ace/(C-mol X · h).

Finally, the different kinetic biomass-specific rates were described in equations 6.16- 6.21.

$$q_{Glu} = q_{Glu_{max}} \cdot \frac{Glu}{Glu + K_{Glu}} \tag{6.16}$$

$$q_{Ace_{cons}} = q_{Ace_{cons_{max}}} \cdot \frac{Ace}{Ace + K_{Ace}} \cdot \frac{K_{Glu}}{Glu + K_{Glu}} \tag{6.17}$$

$$\mu = \mu_{Glu} + \mu_{Ace} \tag{6.18}$$

$$\mu_{Glu} = Y_{X/S} \cdot q_{Glu} \tag{6.19}$$

$$\mu_{Ace} = Y_{Ace_{X/S}} \cdot q_{Ace_{cons}} \tag{6.20}$$

$$q_{Ace_{prod}} = \alpha \cdot \mu_{Glu} + \beta \tag{6.21}$$

Equation 6.16 describes the biomass-specific glucose consumption rate, equation 6.17 models the diauxic behaviour of acetate consumption, inhibited by the presence of glucose, equation 6.18 describes the biomass formation rate, which depends on both glucose and acetate, equations 6.19 and 6.20 respectively describe the biomass formation rate proportional to the glucose and acetate consumption rates, and equation 6.21 is the Luedeking-Piret equation of acetate formation. The component proportional to the biomass growth rate was only linked to the growth rate on glucose because, if acetate is assumed to be produced due to overflow metabolism under glucose excess, it should only be produced while the cells are growing on glucose and not when they are growing on acetate.

There are many different possible expressions that can be used for kinetic modelling. For instance, the biomass could be modelled using a logistic function with a defined maximum value and it could also incorporate a term of glucose consumption for cell maintenance (Klimek and Ollis, 1980). Another option could have been to include growth inhibition due to product (acetate) formation (Choi *et al.*, 2014; L. H. Harada *et al.*, 2002), or glucose and acetate could have been described as mutually inhibiting and include the dissolved oxygen in the kinetic expressions (Anane *et al.*, 2017). Both substrate and product inhibition could have been considered (Krishnan *et al.*, 1999), diauxic growth could have been modelled as a function of time (Liquori *et al.*, 1981) and many more compounds could have been considered in the process stoichiometry (Teixeira *et al.*, 2007).

As it can be appreciated, one fermentation process can be mathematically described at different levels of complexity. The model used in this chapter is relatively simple because the main goal here is to demonstrate how these kinetic models can be used to better understand the process and, ultimately, how fermentation monitoring with on-line metabolomics can be used with these kinetic models to anticipate the evolution of different process metabolites of interest.

### 6.2.2 Model parameter estimation

The three fermentation replicates described in Chapter 5 were used to build and test the kinetic model. The model parameters were estimated by minimising the RMSRE between the fermentation measurements and the model predictions for the three compounds of interest (glucose, biomass and acetate). The RMSRE was used because it provides a percentage of the error for each of the three compounds, thus normalising the differences in magnitude between the three different compounds. The fermentation measurements used for this minimisation were from the off-line analysis of the three fermentation replicates, as the quantitation of the LC-MS data was more accurate than the concentrations estimated with on-line metabolomics (compare Table 5.1 to Table 5.2).

The results from the optimisation for the three fermentation replicates are shown in Figure 6.1 and a summary of the fitted parameters and the RMSRE values are shown in Table 6.1. Some of the kinetic parameters show good agreement between the three replicates, such as $q_{Glu_{max}}$, $Y_{X/S}$, $Y_{Ace_{X/S}}$, $q_{Ace_{max}}$, $K_{Ace}$ and $\alpha$, while the others ($K_{Glu}$,$k_d$ and $\beta$) show larger differences. Nevertheless, a good fit was found for glucose and biomass with a RMSRE around 10 %, whereas the RMSRE for acetate was slightly higher (33.75 %). Overall, the model fits the data well. The largest discrepancy in the acetate prediction comes from the second and third data points, which were measured as 0 g/L. This could be caused by these two points being below the limit of detection of acetate (calculated as 0.42 g/L, see Table 5.1), or it could be that acetate has been produced by the cells but has yet not been secreted outside the cells. In either case, if acetate is produced due to overflow metabolism caused by glucose excess, it should start being produced from the starting conditions - as the model predicts - when the glucose concentration is at its highest.

Some kinetic parameters from Table 6.1 were compared with values from the literature in Table 6.2. These literature values were obtained for similar kinetic models of *E. coli* growth on glucose and acetate. Based on this comparison, it looks like some parameters such as $q_{Glu_{max}}$, $K_{Ace}$ and $Y_{X/S}$ were very similar to the values found in the literature, whereas the fitted $q_{Ace_{max}}$, $K_{Glu}$ and $Y_{Ace_{X/S}}$ were one order of magnitude higher (in absolute value) than the literature values (Table 6.2). This could be avoided by refitting the model with more

constrained boundaries for the parameters, based on the literature values. The model could have also been forced to comply with the elemental, charge and redox balances described by the stoichiometries of equations 6.11 and 6.12 to avoid overestimating the yields used in the model.

Although some parameters were one order of magnitude inaccurate - taking the literature values as correct - the model can be used to extract some useful information from the process. For instance, the model yield of biomass formation on glucose (3.50 $\pm$ 1.15 C-mol X/ mol S) can be used to calculate how much substrate is needed to achieve a specific amount of biomass[1]. Similarly, the fitted biomass-specific consumption rate of glucose and acetate indicate that glucose is consumed faster than acetate (especially if the literature values are taken into account). $K_{Glu}$ was found to be 1.76 $\pm$ 0.83 mM (0.32 $\pm$ 0.15 g/L), or 0.37 $\pm$ 0.32 mM (0.07 $\pm$ 0.06 g/L) based on the literature. According to equation 6.17, a glucose concentration above this $K_{Glu}$ inhibits acetate consumption (50 % inhibition when Glu = $K_{Glu}$), which could be useful information for process control in a continuous or fed-batch process depending whether acetate consumption is desired or not.

**Table 6.1:** Summary of the fitted parameters of the kinetic model and the RMSRE values for glucose, biomass and acetate for the three replicate fermentations.

| Model parameter | Ferment 1 | Ferment 2 | Ferment 3 | Average | CI 95 % |
|---|---|---|---|---|---|
| $q_{Glu_{max}}$ [mol Glu/ (C-molX · h)] | -0.1900 | -0.1694 | -0.1281 | -0.1625 | 0.0357 |
| $q_{Ace_{max}}$ [mol Ace/ (C-molX · h)] | -0.1081 | -0.1138 | -0.1919 | -0.1379 | 0.0530 |
| $k_d$ [$h^{-1}$] | 0.0174 | 0.0928 | 0.1805 | 0.0969 | 0.0924 |
| $K_{Glu}$ [mol Glu/L] | 9.26E-04 | 2.31E-03 | 2.04E-03 | 1.76E-03 | 8.32E-04 |
| $K_{Ace}$ [mol Ace/L] | 1.00E-04 | 1.00E-04 | 1.00E-04 | 1.00E-04 | 0[a] |
| $Y_{X/S}$ [C-molX/mol Glu] | -2.4026 | -3.6837 | -4.4083 | -3.4982 | 1.1493 |
| $Y_{Ace_{X/S}}$ [C-molX/mol Ace] | -4.1372 | -3.7488 | -2.3313 | -3.4058 | 1.0756 |
| $\alpha$ [mol Ace/C-mol X] | 0.0508 | 0.0503 | 0.0500 | 0.0504 | 0.0004 |
| $\beta$ [mol Ace/(C-mol X · h)] | 0.0030 | 0.0087 | 0.0025 | 0.0047 | 0.0039 |
| $RMSRE_{Glu}$ [%] | 8.22 | 10.28 | 9.13 | 9.21 | 1.17 |
| $RMSRE_X$ [%] | 11.11 | 5.96 | 12.39 | 9.82 | 3.85 |
| $RMSRE_{Ace}$ [%] | 36.13 | 32.32 | 32.80 | 33.75 | 2.35 |

[a] The optimisation upper and lower boundaries of $K_{Ace}$ were both accidentally set to 1.00E-04, hence there was no difference between the three replicates.

1. Within certain limits, as substrate inhibition and overflow metabolism might be accentuated at too high substrate concentrations and biological regulatory mechanisms such as quorum sensing might also limit going beyond certain concentration limits of biomass.

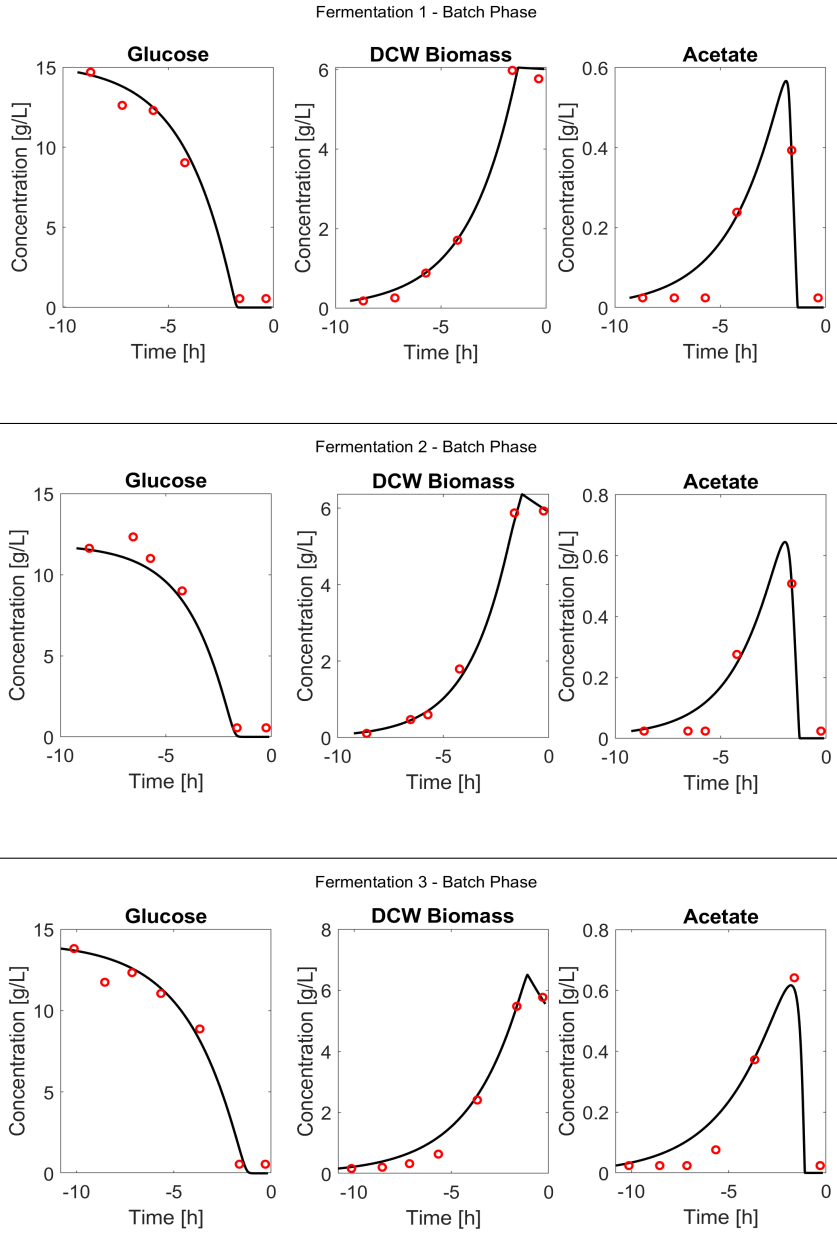



**Figure 6.1:** Results of fitting the parameters of the kinetic model for the three fermentation replicates during the biomass formation phase of the *E. coli* succinate process. The model data (**black line**) is compared to the experimental off-line measurements (**red dots**) of glucose, DCW biomass and acetate.

**Table 6.2:** Comparison of the fitted kinetic model parameters with values from the literature.

| Model parameter | Ref value | Source | Value fitted in this work |
|---|---|---|---|
| $q_{Glu_{max}}$ [mol Glu/ (C-molX·h)] | -0.1845<br>-0.0869 | B. Xu *et al.* (1999)<br>Anane *et al.* (2017) | -0.1625 $\pm$ 0.0357 CI 95% |
| $q_{Ace_{max}}$ [mol Ace/ (C-molX·h)] | -0.0479<br>-0.0208 | Anane *et al.* (2017)<br>B. Xu *et al.* (1999) | -0.1379 $\pm$ 0.0530 CI 95% |
| $K_{Glu}$ [mol Glu/L] | 2.06E-04<br>2.78E-04<br>7.93E-04<br>1.94E-04 | Anane *et al.* (2017)<br>B. Xu *et al.* (1999)<br>Santos *et al.* (2012)<br>Lin *et al.* (2001) | 1.76E-03 $\pm$ 8.32E-04 CI 95% |
| $K_{Ace}$ [mol Ace/L] | 2.27E-04<br>8.47E-04<br>1.69E-03 | Anane *et al.* (2017)<br>B. Xu *et al.* (1999)<br>Lin *et al.* (2001) | 1.00E-04 $\pm$ 0.00 CI 95% |
| $Y_{X/S}$ [C-molX/mol Glu] | -3.07<br>-3.72<br>-3.73<br>-3.60 | Birou *et al.* (1987)<br>Roels (1980)<br>B. Xu *et al.* (1999)<br>Paalme *et al.* (1997) | -3.50 $\pm$ 1.15 CI 95% |
| $Y_{Ace_{X/S}}$ [C-molX/mol Ace] | -0.76<br>-0.86<br>-0.96<br>-0.72 | Roels (1980)<br>Monod (1949)<br>B. Xu *et al.* (1999)<br>Paalme *et al.* (1997) | -3.41 $\pm$ 1.08 CI 95% |

### 6.2.3 Forecasting the concentration of process compounds

Kinetic models are a powerful tool to better understand and control bioprocesses. A good example is the use of these models to forecast the evolution of the model compounds in time. This was tested with the *E. coli* bioprocess. Particularly, the on-line metabolomics data from the three fermentation replicates was used to predict at three time points (6 h, 4 h and 2 h before the beginning of the succinate production phase) the evolution of glucose, biomass and acetate. The on-line metabolomics data was used simulating a process monitoring scenario, where the forecasting predictions could be used for process control.

The on-line metabolomics data was acquired using a triple quadrupole mass spectrometer with an MRM method as detailed in Section 2.8. The compound (R)-2,3-dihydroxy-isovalerate was used as a biomarker of biomass WCW concentration (see Table 5.2), which was converted into DCW as detailed in Section 6.4.1. Acetate could not be successfully monitored with the MRM method due to the presence of an ion of the same m/z value (59 m/z) that was identified as a potential in-source glucose fragment (because it followed the same dynamic profile as glucose). For this reason, the concentration of acetate was modelled using the kinetic

model and the concentration of glucose and biomass measured on-line (see Section 6.4.2). Figure 6.2 shows the results from the predictions for the three fermentation replicates and the RMSE and RMSRE values obtained are summarised in Figure 6.3. These results will be analysed below, one fermentation replicate at a time.

Looking at the first replicate in Figure 6.2, it can be appreciated that although the concentration of glucose was highly overestimated with on-line metabolomics, the prediction at four hours before the end of the growth phase was already capable of predicting glucose depletion within a 45-minute error (depletion predicted at -0.86 h and measured at -1.61 h) and the prediction at two hours before the end of the growth phase could predict glucose depletion within an 8-minute error (depletion predicted at -1.48 h and measured at -1.61 h). Looking at the biomass predictions for the first replicate, the DCW concentration is in good agreement with the measurements until the penultimate data point, however, the model did not manage to predict in time the end of the growth phase, which is probably caused by an overestimation of the acetate production (because the excess acetate would prolong biomass growth). Figure 6.3 shows that in the first replicate, the forecasted acetate concentration was overestimated by 0.80 - 1.40 g/L. The overestimation of acetate is, in turn, probably caused by the overestimated glucose concentration measured by on-line metabolomics. Indeed, the forecasted glucose concentration in the first replicate is overestimated with a RMSE of up to 11 g/L (200 % RMSRE), as shown in Figure 6.3, and it progressively drops down to 0.62 g/L at two hours before the end of the biomass growth phase. Despite the overestimation of acetate, the model properly describes that acetate is consumed upon glucose starvation.

Looking at the second replicate in Figure 6.2, it can be seen that the on-line metabolomics glucose fits much better the off-line analysis, with much lower RMSE values below 3 g/L at six hours before the beginning of succinate production (Figure 6.3). This results in better predictions of acetate and biomass DCW - in agreement with the idea that was introduced in the previous paragraph that the overestimation of glucose led to an overestimation of acetate and biomass in the first fermentation replicate. Indeed, in the second replicate, biomass is predicted to stop growing before the beginning of the succinate production phase, as the off-line data indicates. Interestingly, the prediction for biomass is better at four hours before the end of biomass growth than at two hours before that point (RMSE of 1.04 g/L compared to 1.82 g/L), as shown in Figure 6.3. This is caused by the last prediction forecasting glucose starvation slightly early, which leads to less acetate formation and a premature switch towards growth on acetate, resulting in less biomass formation overall.

Looking at the third replicate in Figure 6.2, the concentration of glucose is overestimated with on-line metabolomics, but not as much as in the first replicate, which causes the acetate and biomass predictions to be closer to the off-line measurements with RMSE around 0.3 - 0.6 g/L and 1.4 - 1.6 g/L, respectively (Figure 6.3). Nevertheless, despite the initial overestimation of the glucose concentration, at two hours before the end of the biomass growth phase, the model is capable of successfully predicting glucose starvation, the consumption of acetate and the end of biomass growth.

In summary, using the kinetic model it was possible to forecast events such as acetate accumulation during cell growth on glucose, glucose starvation, acetate consumption upon glucose depletion and the end of biomass growth after consumption of both carbon sources. The on-line glucose measurement was overestimated for two of the three replicates - particularly for the first one - and this seemed the highest cause for inaccuracy in the predictions. This discrepancy in glucose measurements was not expected, as the three replicates were used to build the monitoring model (Figure 5.8).

A possible explanation to the overestimation of the first replicate could be that the mass spectrometer was not properly conditioned before the beginning of the fermentation. Plekhova *et al.* (2021) recommend using 10 - 20 QC samples for instrument conditioning before starting a mass spectrometry analysis. As samples are injected into the mass spectrometer every 5 minutes, injecting 10 - 20 samples means that it would take 50 - 100 min of analysis to condition the instrument, which is very close to the time that the on-line glucose signal in the first replicate takes to stop increasing at the beginning of the fermentation (from -9.3 to -7.5 h before the succinate production phase, *i.e.* 108 min). This could be easily tested in the future by injecting sufficient media samples before starting the fermentation and, if conditioning is the problem as it is suggested, the quality of the monitoring model and the forecasting capacity of the kinetic model could be further enhanced.

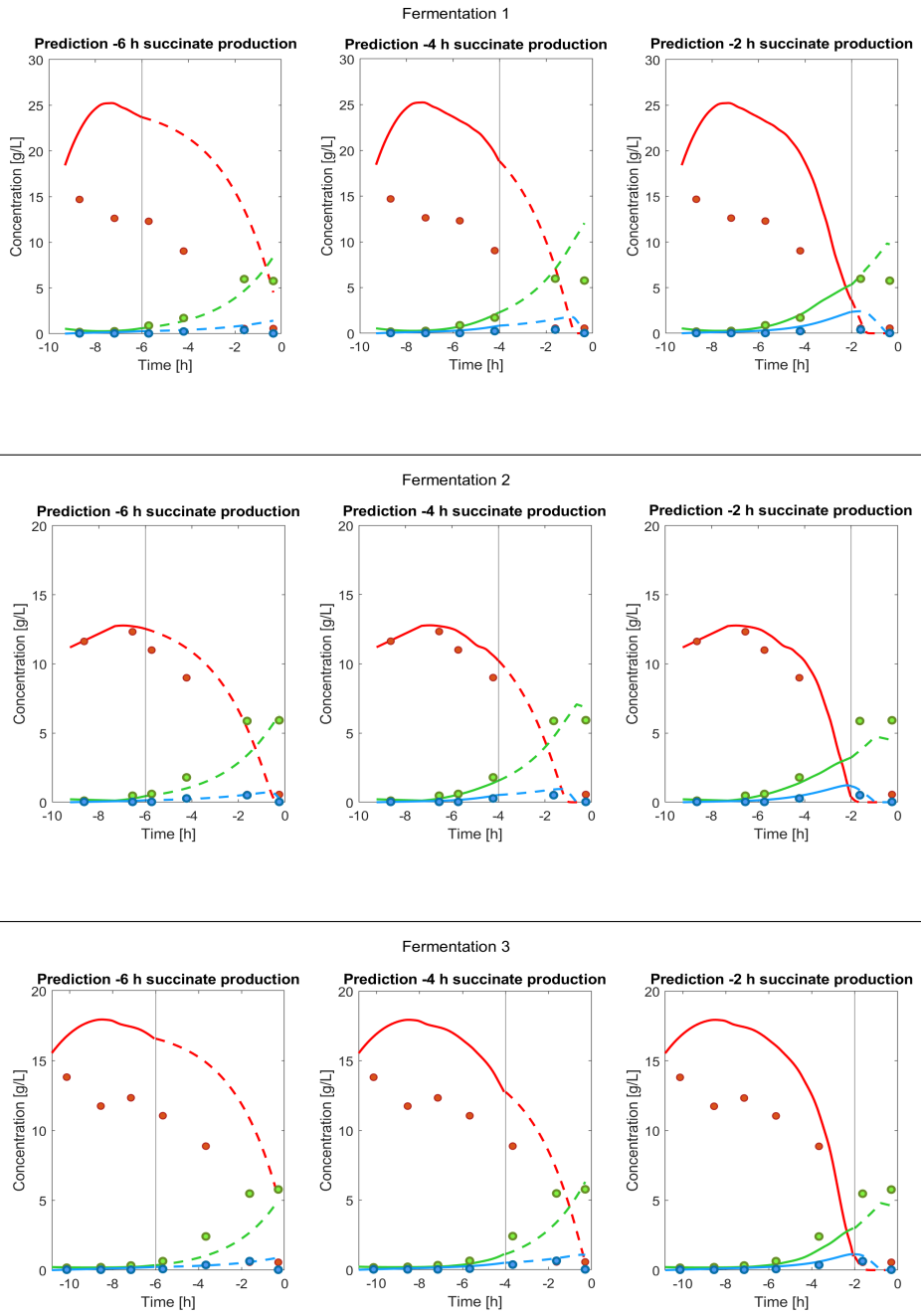



**Figure 6.2:** Forecast of the evolution of glucose, biomass and acetate using the kinetic model and the on-line metabolomics measurements during biomass formation of the *E. coli* succinate process. The figure shows on-line monitoring data until the point of starting the forecast (**solid line**), the off-line measurements (**circles**) and the forecasted values (**dotted line**) for glucose (**red**), biomass DCW (**green**) and acetate (**blue**).

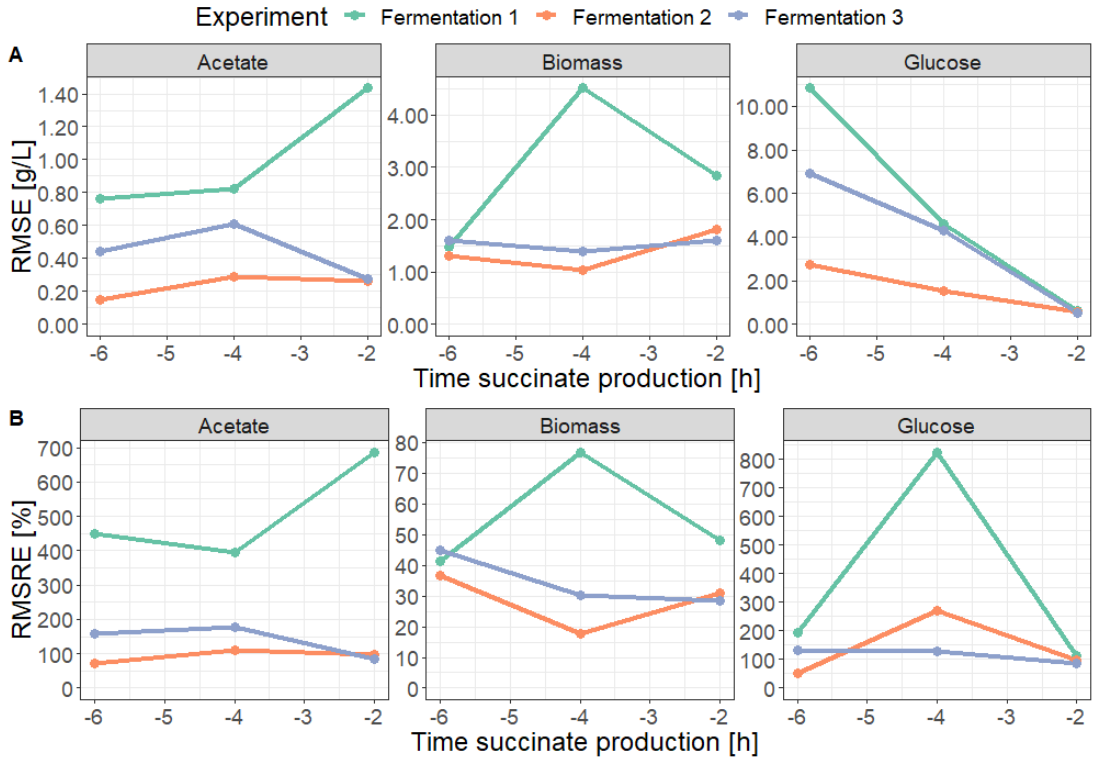



**Figure 6.3:** Forecasting RMSE (**A**) and RMSRE (**B**) values for glucose, biomass and acetate using the kinetic model and on-line metabolomics monitoring for the three replicate fermentations, evaluated at 6 h, 4 h and 2 h before the beginning of the succinate production phase.

Other authors have used kinetic modelling for forecasting the evolution of bioprocesses in the field of predictive control. For example, Santos *et al.* (2012) used bioprocess control to adjust the feed flow rate of an *E. coli* fed-batch culture in order to minimise acetate formation due to overflow metabolism. Cos *et al.* (2006) controlled the concentration of methanol in a *P. pastoris* fed-batch process for the production of *Rhizopus oryzae* lipase. Finally, an interesting approach was taken by Cabaneros Lopez *et al.* (2020), who used a kinetic model to describe an ethanol *S. cerevisiae* process using cellulose in the presence of furfural and acetic acid - two common process inhibitors derived from lignocellulosic materials. The authors used a novel adaptive approach, where the fitted kinetic parameters were updated at different time points of the fermentation based on the glucose concentration calculated by PLS regression using MIR spectroscopy measurements. At these different time points, the updated kinetic models were used to forecast the evolution of glucose, xylose and ethanol for the rest of the fermentation, similar to the work shown in this chapter. The authors obtained RMSE between approximately 0.7 and 5.0 g/L for glucose (the maximum glucose concentration was *ca.* 40 g/L), 0.7 and 3.0 g/L for xylose (the maximum xylose concentration was *ca.* 20 g/L) and 0.7

and 3.0 g/L for ethanol (the maximum ethanol concentration was *ca.* 20 g/L). These error values are below the ones obtained in this chapter, which suggests that the adaptive strategy used by Cabaneros Lopez *et al.* (2020) might be a powerful way of improving the predicting capacity of kinetic models when using fermentation monitoring.

## 6.3 Conclusions

This chapter demonstrates how kinetic models can be used to describe and predict different biological phenomena such as acetate accumulation during *E. coli* growth on glucose, glucose starvation, acetate consumption upon glucose depletion and the switch from biomass exponential growth to stationary phase after consumption of both carbon sources. This can improve the understanding of the bioprocess, for example facilitating the identification of important parameters such as yields, production rates and affinity constants. Furthermore, these models can be employed in combination with fermentation monitoring to anticipate the future evolution of key process compounds and use this information for process control.

On-line metabolomics monitoring was used in combination with the kinetic model developed allowing the forecasting - at different time points - of the evolution of glucose, biomass and acetate during the biomass production phase of the process. This simulates a scenario of real-time forecasting the evolution of these compounds during process monitoring using on-line metabolomics. One of the major factors affecting the quantitative accuracy of the model predictions was the quality of the on-line metabolomics measurements. That is, when the on-line glucose measurements were overestimated, a larger predictive error was found on the biomass and acetate predictions. It was suggested that the overestimation of glucose by on-line metabolomics could have been caused by insufficient conditioning of the mass spectrometer before starting monitoring the fermentation.

## 6.4 Materials and methods

The materials and methods are as described in Chapter 2. Any differences and particular methods used exclusively in this chapter are described below.

### 6.4.1 Biomass dry weight calculation

The biomass DCW was calculated by multiplying the biomass WCW measurements by 0.2528. This was done instead of using the DCW measurements directly because these were very prone to error for values below 1 g/L DCW. The linear conversion was determined experimentally by linear regression of all the DCW and WCW biomass concentration measurements of

the three fermentation replicate experiments used to build the kinetic models (Figure 6.4). The linear regression was done in Microsoft® Excel® (Version 2202) using the LINEST function and negative DCW measurements were imputed as 0 g/L.

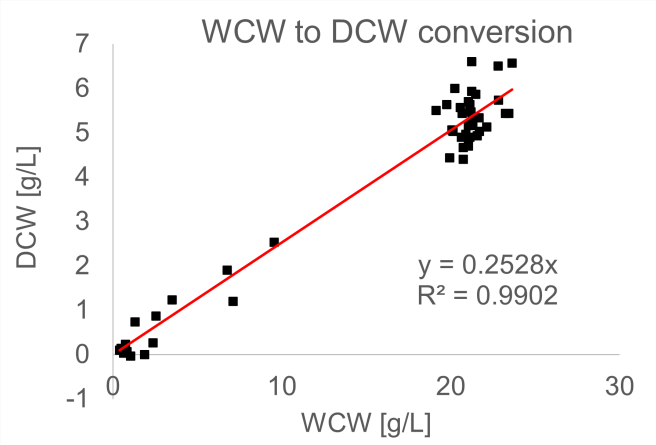



**Figure 6.4:** Linear regression of biomass DCW and WCW.

### 6.4.2 On-line acetate concentration calculation

The concentration of acetate was modelled solving the ordinary differential equation 6.15 using the kinetic expressions of equations 6.17, 6.19 and 6.21 of the kinetic model, the concentration of glucose and biomass measured on-line and a starting concentration of 0 mol/L acetate at the beginning of the fermentation. The on-line DCW was calculated from the measured on-line WCW using the linear conversion from Section 6.4.1.

### 6.4.3 Model parameter estimation

The three fermentation replicates described in Chapter 5 were used to build and test the kinetic model. The model parameters were estimated in Matlab R2019b® by minimising the RMSRE using nonlinear regression with the *fmincon* function. The objective function used for this minimisation is described in equation 6.22.

$$f_{obj} = min\left[\sum_{j=1}^{k} RMSRE_j\right] = min\left[RMSRE_{Glu} + RMSRE_X + RMSRE_{Ace}\right] \tag{6.22}$$

where *j* are the compounds of interest. In this case, glucose, biomass and acetate were used.

### 6.4.4 Ordinary differential equations

The system of ordinary differential equations was written in Matlab R2019b® and solved with the ode45 function.

Chapter 7

# Exploration of the pentose phosphate pathway to improve succinate production in *Escherichia coli*

The HPLC analysis of the PPP inhibition experiments was carried out by Ingenza Ltd.

## Abstract

The metabolomics results from Chapters 3 and 4 revealed an increase of pentose phosphate pathway (PPP) metabolites during the anaerobic succinate production phase of an *Escherichia coli* (*E. coli*) bioprocess (Figures 3.10, 4.11 and 4.12). Although one of the main roles of the PPP is to generate precursors for biomass growth (ribose-5P for nucleotides, NADPH for lipids and erythrose-4P for aromatic amino acids), the *E. coli* strain used does not grow on glucose under anaerobic conditions. This led to the hypothesis that the carbon flux going to the PPP during the succinate production phase could be rechannelled to the anaerobic tricarboxylic acid (TCA) cycle to increase succinate production without causing harm to the cells. This was tested by adding oxythiamine, a known inhibitor of the enzyme transketolase in the PPP, and no increase in succinate titre was observed for oxythiamine concentrations up to 10 mM. In fact, a decrease in succinate titre was observed using 10 mM oxythiamine, suggesting a necessary role in the PPP for the production of succinate. This is probably the consequence of the requirement of reduced NADH cofactors for the anaerobic formation of succinate via the anti-clockwise reductive TCA cycle. This NADH needs to be generated either with the production of phosphoenolpyruvate, pyruvate, or via the PPP, which generates two moles of NADPH for every mole of glucose-6P that enters the pathway. This NADPH can then be converted into NADH by the enzyme transhydrogenase, which is present in *E. coli*.

In order to measure PPP inhibition, samples were taken for LC-MS analysis of the PPP metabolites. However, it was not possible to reliably measure these metabolites, probably due to metabolite loses during the extraction process causing the concentration of the sugar phosphates to be below their limit of detection.

## 7.1 Introduction

In Chapter 4, an *E. coli* dual-phase succinate fermentation process was analysed by LC-MS. The process consists of an aerobic batch phase of biomass growth followed by an anaerobic succinate production phase. Pathway analysis revealed a few metabolic pathways that were particularly up-reagulated during succinate production: the pentose phosphate pathway, ascorbate and aldarate metabolism and pantothenate and CoA biosynthesis.

The pentose phosphate pathway (PPP) is involved in many core metabolic reactions and the formation of anabolic intermediates required for biomass growth, such as ribose-5P, erythrose-4P and reduced NADPH cofactors (Matano *et al.*, 2014). Ribose-5P is a precursor of ribonucleotide synthesis (Wamelink *et al.*, 2005). Erythrose-4P is together with phosphoenolpyruvate the precursor of shikimate, the first intermediate of the shikimate pathway for the formation of aromatic amino acids (S. Y. Lee *et al.*, 2019). Finally, NADPH is an essential cofactor for the formation of fatty acids, cell membrane components and for protecting the cell against reactive oxygen species (ROS) (W. Li *et al.*, 2018; Nordberg and Arnér, 2001). The PPP is typically split into two sections: an oxidative part resulting in the formation of NADPH, and a non-oxidative part (Figure 7.1). The formation of reduced NADPH takes part in the oxidative part, with the oxidation of glucose-6P and gluconate-6P.

As one of the main roles of the PPP is to generate intermediates for biomass growth, it was hypothesised that a high carbon flux down this pathway was not necessary during succinate production (as this strain does not grow anaerobically). Instead, this flux could potentially be re-channelled towards succinate production by partially or totally inhibiting the PPP.

### 7.1.1 Inhibition of the PPP

Several authors have studied the effects of inhibiting the PPP for different applications. A PPP inhibitor commonly used in the literature is oxythiamine, a thiamine (vitamin B1) analogue. Thiamine and oxythiamine are pyrophosphorylated by the enzyme thiamine pyrophosphokinase into thiamine pyrophosphate and oxythiamine pyrophosphate, respectively. Thiamine pyrophosphate is the natural cofactor of the enzyme transketolase, involved in the non-oxidative PPP, and oxythiamine pyrophosphate competes with it causing inhibition of the enzyme and of the PPP (F. Zhang *et al.*, 2016).

Some examples of inhibiting the PPP with oxythiamine include the inhibition of growth and proliferation of the malaria parasite *Plasmodium falciparum* (Chan *et al.*, 2013), arresting growth of Ehrlich's tumour cells (Raïs *et al.*, 1999) and thyroid cancer cells (M. Liu *et al.*, 2020). There are many fewer studies showing the effect of oxythiamine on prokaryotes. In

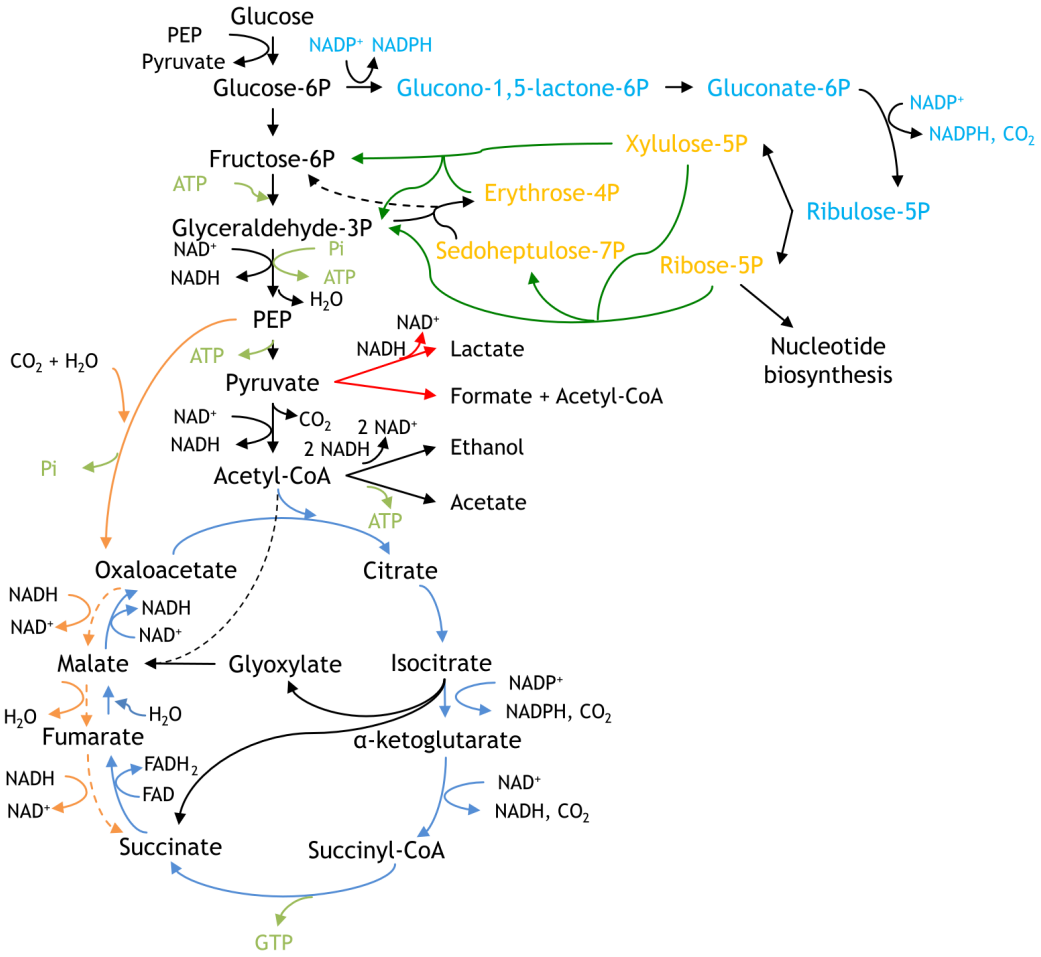



**Figure 7.1:** Main metabolic reactions involved in succinate production in *E. coli*, highlighting the oxidative and non-oxidative PPP in light blue and orange, respectively. **Dark green** lines indicate reactions in the PPP catalysed by the enzyme transketolase. Dark blue lines indicate the oxidative TCA cycle under aerobic conditions. Dark orange lines indicate the reductive TCA cycle under anaerobic conditions. Red lines indicate deleted reactions in the industrial strain used. Arrows crossing other reactions are marked with black dashes. ATP, GTP and free phosphate (Pi) are indicated in light green.

one of these studies, Kawasaki and Esaki (1971) showed that oxythiamine inhibits the uptake of thiamine in growing *E. coli* K-12 and KG33, a K-12 derivate auxotrophic for thiamine. As these examples show, due to its wide range of metabolic functions, PPP inhibition tends to cause the stop of cell growth.

## 7.2 Results and discussion

### 7.2.1 Effect of using oxythiamine on succinate production

Two possible ways of reducing the flux into the PPP could be genetic engineering or using a pathway inhibitor. The former is fundamentally a better long term solution for production scale (particularly for a platform chemical such as succinate), as the use of inhibitors could significantly increase the cost of production. However, finding a strain that grows well aerobically (which requires PPP activity) and has a low PPP flux anaerobically might not be trivial. More importantly, genetic engineering can be a significantly time consuming option. Using an inhibitor, on the other hand, has the advantage that it can be added at the production phase without compromising cellular growth during the batch phase, and different degrees of inhibition can potentially be tested much quicker by simply adjusting the concentration of inhibitor. If an interesting result is found using an inhibitor, efforts can be made to try to replicate the same phenotype with genetic engineering.

For the reasons mentioned above, the effect of PPP inhibition on succinate production was studied by adding oxythiamine, an inhibitor of the enzyme transketolase that catalyses the two reactions of the non-oxidative phase of the PPP involving xylulose-5P as substrate (Figure 7.1) (Comín-Anduix *et al.*, 2001; Mariadasse *et al.*, 2016). This was studied in Schott bottles as a scale-down version of the bioreactor in order to increase the throughput of the experiment. Oxythiamine was added at the beginning of the succinate production phase at different concentrations and the resulting succinate and pyruvate titres and yields were determined by HPLC (Figure 7.2A to D).

The results showed no significant change in pyruvate titres across all concentrations of oxythiamine tested and no change in succinate titre for concentration of up to 5000 $\mu$M oxythiamine. However, the highest concentration of oxythiamine (10,000 $\mu$M) showed a significant decrease in succinate titre (p-value 0.020). This might be explained by a redox imbalance in the production of succinate from glucose via the reductive TCA cycle (Martinez Basterrechea, 2010; Singh *et al.*, 2011; Tan *et al.*, 2016). Specifically, the formation of two moles of succinate from one mole of glucose requires two moles of NADH (stoichiometry shown in equations 7.1 and 7.2) and therefore this conversion requires additional reduced cofactors.

Singh *et al.*, 2009 found a 2:1 NADH:$NAD^+$ ratio during anaerobic succinate production using the similar *E. coli* strain NZN111 and a similar dual-phase fermentation process without PPP inhibition. This indicates that, under normal conditions (no PPP inhibition), NADH is not limiting. This supports the idea that the PPP is critical in generating reduced NADH cofactors, potentially via the formation of NADPH, which can in turn be converted into NADH by the action of the enzyme transhydrogenase, which is present in *E. coli* (Csonka and Fraenkel, 1997; W. Li *et al.*, 2018; Sauer *et al.*, 2004). Alternatively, NADH can be produced from glycolysis with pyruvate or PEP as end products. The formation of three moles of pyruvate from one mole

of glucose generates two moles of NADH (stoichiometry shown in equations 7.3 and 7.4), and the formation of a mole of PEP also generates two moles of NADH (stoichiometry shown in equations 7.5 and 7.6). Consequently, if the PPP is inhibited, the formation of succinate needs to be coupled with pyruvate and/or PEP formation in order to close the redox balance. This would suggest that under PPP inhibiting conditions, the yield of succinate on glucose would decrease in favour of an increase in the yield of pyruvate or PEP. However, no change in the yield of either succinate or pyruvate was observer across all the oxythiamine conditions tested (Figure 7.2C and D).

Interestingly, the succinate and pyruvate yields on glucose using the bioreactor were 0.24 [g succinate/ g glucose] and 0.19 [g pyruvate/ g glucose] (based on the data from Figure 5.5). However, in this scaled-down experiment, the yield of pyruvate on glucose increased around three times, even when no oxythiamine was used (0.69 [g pyruvate/ g glucose]). The yield of succinate on glucose also increased, but less significantly (0.31 [g succinate/ g glucose]). This indicates that moving from the bioreactor setup to the Schott bottle scaled-down bioprocess, the biology was affected favouring a shift towards pyruvate formation. One of the main differences between both scales is that $CO_2$ is continuously sparged in the 5 L bioreactor, while it was added all at once in the Schott bottles. However, because the yield of succinate on glucose did not significantly change between both setups, $CO_2$ might not be the cause of this difference. The basis of this reasoning is that succinate production requires $CO_2$ (see Figure 7.1 and equation 7.1).

Stoichiometry of succinate formation with PPP inhibition:

$$\begin{aligned} &\text{Glucose} + \text{PEP} + 2\text{CO}_2 + \text{ADP} + 2\text{NADH} + 2\text{H}^+ \\ &\longrightarrow 2\text{Succinate} + \text{Pyruvate} + \text{ATP} + 2\text{NAD}^+ + 2\text{H}_2\text{O} + 3\text{H}^+ \end{aligned} \tag{7.1}$$

which in chemical formula is

$$\begin{aligned} &\text{C}_6\text{H}_{12}\text{O}_6 + \text{C}_3\text{H}_2\text{O}_6\text{P}^{3-} + 2\text{CO}_2 + \text{C}_{10}\text{H}_{12}\text{N}_5\text{O}_{10}\text{P}_2{}^{3-} + 2\text{C}_{21}\text{H}_{27}\text{N}_7\text{O}_{14}\text{P}_2{}^{2-} + 2\text{H}^+ \\ &\longrightarrow 2\,\text{C}_4\text{H}_4\text{O}_4{}^{2-} + \text{C}_3\text{H}_3\text{O}_3{}^- + \text{C}_{10}\text{H}_{12}\text{N}_5\text{O}_{13}\text{P}_3{}^{4-} + 2\,\text{C}_{21}\text{H}_{26}\text{N}_7\text{O}_{14}\text{P}_2{}^- + 2\,\text{H}_2\text{O} + 3\,\text{H}^+ \end{aligned} \tag{7.2}$$

Stoichiometry of pyruvate formation with PPP inhibition:

$$\begin{aligned} &\text{Glucose} + \text{PEP} + 2\text{NAD}^+ + 3\text{ADP} + 2\text{Pi} + 3\text{H}^+ \\ &\longrightarrow 3\text{Pyruvate} + 2\text{NADH} + 2\text{H}^+ + 3\text{ATP} + 2\text{H}_2\text{O} \end{aligned} \tag{7.3}$$

which in chemical formula is

$$\begin{aligned} &\text{C}_6\text{H}_{12}\text{O}_6 + \text{C}_3\text{H}_2\text{O}_6\text{P}^{3-} + 2\text{C}_{21}\text{H}_{26}\text{N}_7\text{O}_{14}\text{P}_2{}^- + 3\text{C}_{10}\text{H}_{12}\text{N}_5\text{O}_{10}\text{P}_2{}^{3-} + 2\text{PO}_4{}^{3-} + 3\text{H}^+ \\ &\longrightarrow 3\,\text{C}_3\text{H}_3\text{O}_3{}^- + 2\,\text{C}_{21}\text{H}_{27}\text{N}_7\text{O}_{14}\text{P}_2{}^{2-} + 2\,\text{H}^+ + 3\text{C}_{10}\text{H}_{12}\text{N}_5\text{O}_{13}\text{P}_3{}^{4-} + 2\,\text{H}_2\text{O} \end{aligned} \tag{7.4}$$

Stoichiometry of PEP formation with PPP inhibition:

$$\begin{aligned} &\text{Glucose} + \text{PEP} + 2\text{NAD}^+ + \text{ADP} + 2\text{Pi} + \text{H}^+ \\ \longrightarrow\ &2\text{PEP} + \text{Pyruvate} + 2\text{NADH} + 2\text{H}^+ + \text{ATP} + 2\text{H}_2\text{O} \end{aligned} \tag{7.5}$$

which in chemical formula is

$$\begin{aligned} &\text{C}_6\text{H}_{12}\text{O}_6 + \text{C}_3\text{H}_2\text{O}_6\text{P}^{3-} + 2\text{C}_{21}\text{H}_{26}\text{N}_7\text{O}_{14}\text{P}_2{}^- + \text{C}_{10}\text{H}_{12}\text{N}_5\text{O}_{10}\text{P}_2{}^{3-} + 2\text{PO}_4{}^{3-} + \text{H}^+ \\ \longrightarrow\ &2\text{C}_3\text{H}_2\text{O}_6\text{P}^{3-} + \text{C}_3\text{H}_3\text{O}_3{}^- + 2\,\text{C}_{21}\text{H}_{27}\text{N}_7\text{O}_{14}\text{P}_2{}^{2-} + 2\,\text{H}^+ + \text{C}_{10}\text{H}_{12}\text{N}_5\text{O}_{13}\text{P}_3{}^{4-} + 2\,\text{H}_2\text{O} \end{aligned} \tag{7.6}$$

The hypothesis that the PPP plays an important role for succinate production by providing reduced coffactors is supported by the work from Tan *et al.* (2016), who generated a library of PPP genes under the regulation of a constitutive M1-93 promoter and different ribosome binding sites to generate different *E. coli* variants with different expression levels of the PPP genes. Their results indicate that increasing the expression levels of PPP genes can lead to higher succinate yields and titres, but that a fine tuning of the expression levels gives the best results. Interestingly, the highest succinate titres were obtained with medium expression levels of genes from the oxidative part of the pathway and different combinations of medium and high expression levels of genes from the non-oxidative part.

In conclusion, the drop in succinate titre after inhibition of the PPP suggests that this pathway plays an important role for succinate production, probably via the formation of reduced NADPH. Furthermore, if the PPP is inhibited, succinate production will depend on pyruvate and PEP formation to produce the required NADH. Nevertheless, it is important to note that the scaled-down version of the bioprocess used in this experiment favoured the production of pyruvate in comparison to the 5 L bioreactor setup, even when the PPP was not inhibited. This could mean that the conclusions from this experiment might not necessarily be transferable to the bioreactor.

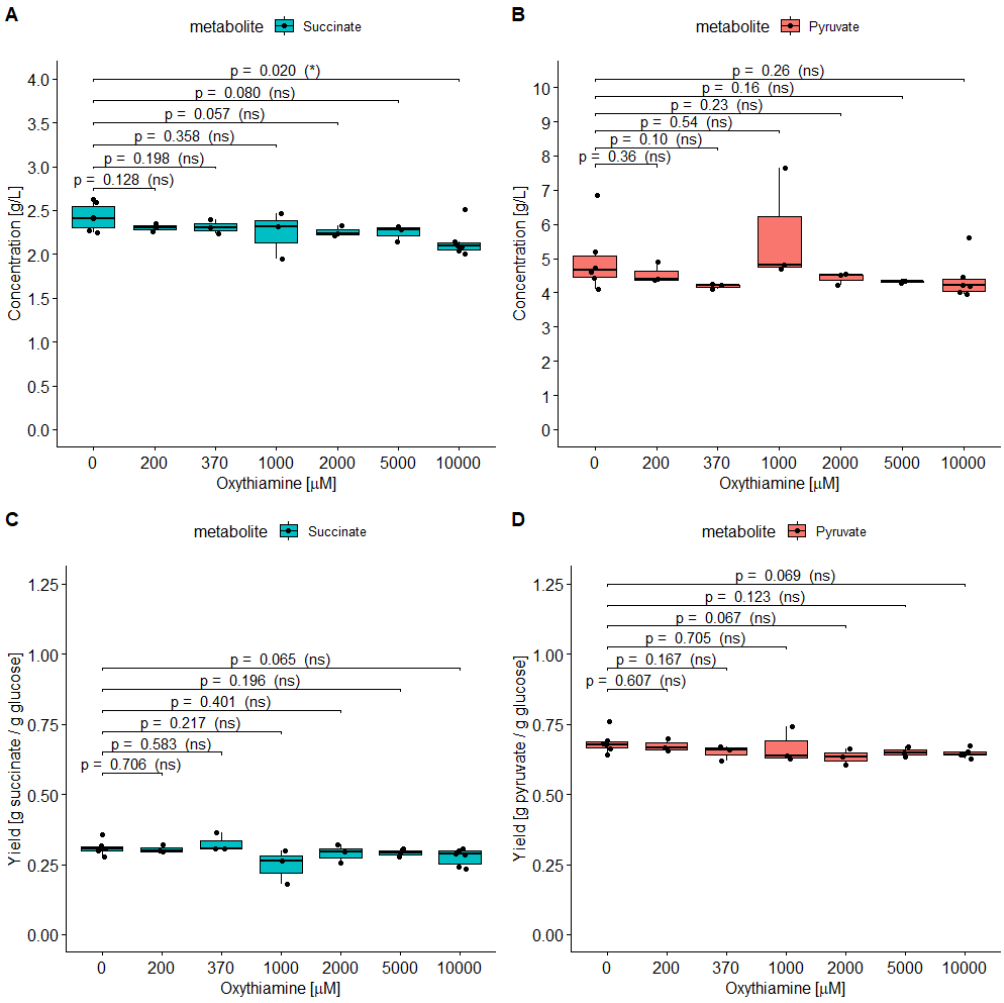



**Figure 7.2:** Succinate concentration (**A**), pyruvate concentration (**B**), succinate yield (**C**) and pyruvate yield (**D**) as a function of the concentration of oxythiamine added at the beginning of the succinate production phase. Data measurements are marked as black dots. Both extreme conditions (0 and 10,000 $\mu$M oxythiamine) were performed as sextuplicates, and the rest of conditions as triplicates. A t-test at 95 % confidence level was done to compare the mean of each condition to the reference of no oxythiamine addition. P-values are shown for each comparison and the significance level is shown in brackets. Ns: not-significant; *: p-value $\leq$ 0.05.

### 7.2.2 Determining inhibition of the PPP

Although oxythiamine has been extensively used in the literature as a transketolase inhibitor, samples were taken at the beginning and end of the succinate production phase for LC-MS analysis to measure the metabolites from the PPP and determine if the pathway had been successfully inhibited. The accumulation of metabolites from the oxidative part, or low levels of metabolites from the non-oxidative part could be indicators of PPP inhibition. Unfortunately, the reliable measurement of PPP metabolites in the samples was not possible, despite trying different approaches, as described below. That is, noisy and non-gaussian chromatographic peaks were observed for these compounds in the biological samples. The metabolites in the reference standard samples showed better peak shapes above the noise level (Figure 7.3).

Different strategies were followed to try to measure the PPP metabolites, using two different extraction methods, two different mass spectrometers and two different chromatographic columns and methods.

The two extraction methods tested were 1:3:1 chloroform:methanol:water (which had been previously used successfully to measure PPP metabolites in Chapter 4) and boiling ethanol (75 % v/v), which is another method commonly used for metabolites extraction for LC-MS analysis (Campos *et al.*, 2017; Gonzalez *et al.*, 1997; Yang *et al.*, 2008). Canelas *et al.* (2009) compared five different extraction methods on 44 intracellular metabolites, including sugar phosphates, and concluded that boiling ethanol and chloroform-methanol were the two methods yielding the highest extraction recoveries and reproducibility. However, the analysis of the PPP inhibition samples with the two extraction methods tested did not result in clear chromatographic peaks for the PPP sugar phosphates.

Two different instruments were used to try to analyse the PPP metabolites in the samples: an ion mobility quadrupole time-of-flight (IM-Q-TOF) and a triple quadrupole (QQQ). On the IM-Q-TOF, a HILIC-Z column was used with a 4-minute chromatographic method as described by Pičmanová *et al.* (2022), whereas a long ZIC p-HILIC column was used with the QQQ with a 65-minute chromatographic method (Teleki *et al.*, 2015). However, neither of the two methods resulted in clear chromatographic peaks in the biological samples.

The unsuccessful attempts to measure the PPP metabolites could potentially be explained due to intracellular concentrations of PPP metabolites being below the LOD. A. Buchholz *et al.* (2001) determined the LOD and intracellular concentration of several PPP metabolites in a glucose-limited chemostat experiment using *E. coli* K-12 using a Finnigan LCQ ion trap mass spectrometer. Namely, they found the LOD of gluconate-6P, ribose-5P and fructose-6P to be 15.2, 8.8 and 20.7 $\mu$M and their respective intracellular concentrations 450, 260 and 250 $\mu$M. Hiller *et al.* (2007) reported similar findings using a similar chemostat setup with *E. coli* K-12 and a Finnigan LCQ Advantage ion trap mass spectrometer. The authors reported the LOD of fructose-6P and glyceraldehyde-3P to be around 8 and 10 $\mu$M and their intracellular

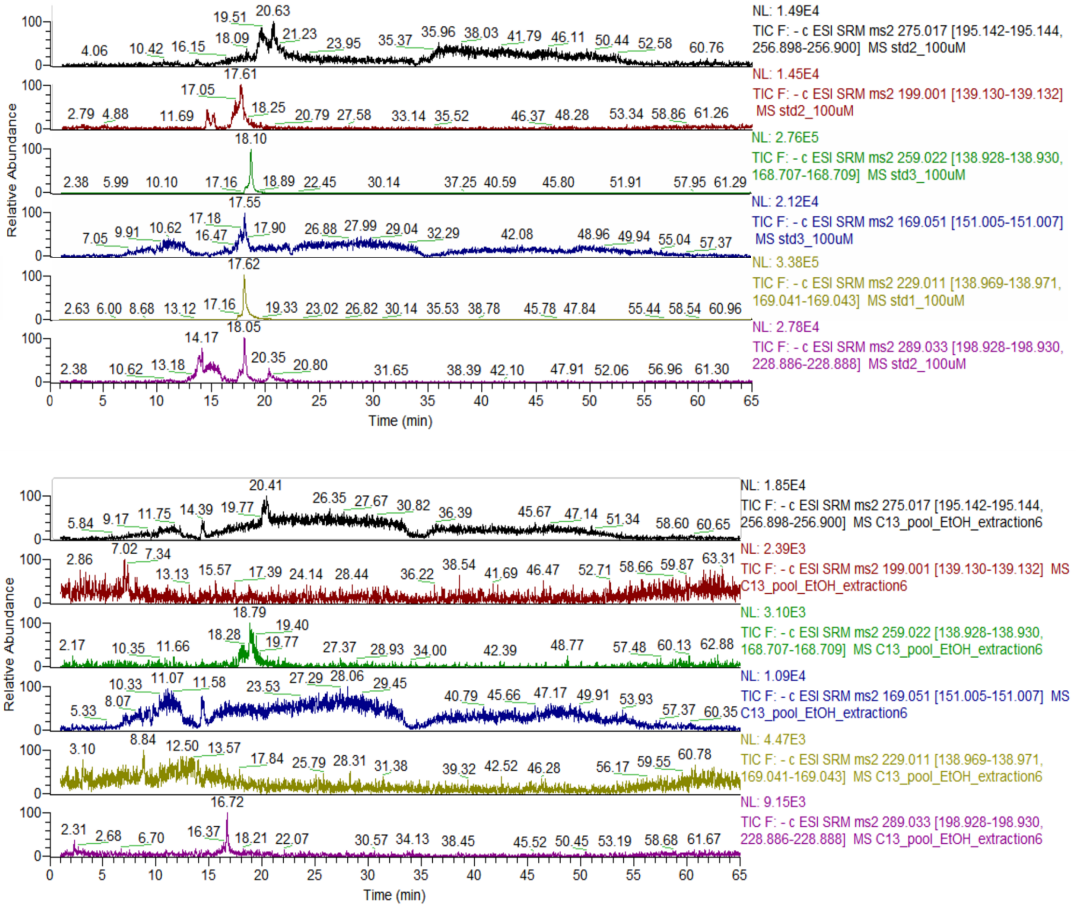



**Figure 7.3:** Chromatogram of some sugar phosphates in the reference standard mixes at concentrations of approximately 10-100 $\mu$M (**top**) and in the pooled sample of the PPP inhibition experiment extracted with boiling ethanol (**bottom**). In descending order, the metabolites are 6-phosphogluconate, erythrose-4P, fructose-6P, glyceraldehyde-3P, ribose-5P and sedoheptulose-7P. The peaks look better for the reference standards and the intensity is higher.

concentration 80 and 200 $\mu$M. In both studies the LODs of the sugar phosphates reported were roughly 10-20 $\mu$M and the intracellular concentrations one order of magnitude higher, roughly 100-400 $\mu$M. As a consequence, any loses in the metabolite levels due to inadequate metabolism quenching or sample dilution during the process of extraction prior to LC-MS analysis could compromise the detection of these compounds.

The LOD of an analyte is determined using the standard deviation of the intercept and slope of its calibration curve (equation 2.1) and, therefore, it will depend on the sensitivity of the mass analyser and, importantly, on the extraction method used, as reported by Hiller *et al.* (2007). This is the case because different extraction methods might result in different metabolite recoveries, consequently changing the calibration curve. The LOD of ribose-5P, sedoheptulose-7P and xylulose-5P in the QQQ were, respectively, 0.1998, 0.0122 and 0.2569

g/L (868, 42 and 1116 $\mu$M) using 1:3:1 chloroform:methanol:water extraction (see Chapter 5, Table 5.1). These LODs for ribose-5P and xylulose-5P are above the intracellular concentration reported by the literature works mentioned above (100-400 $\mu$M), providing a reason why these compounds could not be measured successfully.

Metabolites of the PPP were successfully measured by LC-MS in Chapter 4. In that case, the analysis was done in a Q Exactive$^{TM}$ Orbitrap mass spectrometer and the bioprocess was performed in a bioreactor, but the 1:3:1 C:M:W metabolite extraction method was the same in both analyses. One possible explanation of why the PPP compounds were detected in that experiment could be that the concentration of biomass reached in the fermenter was higher (up to 25 g/L WCW, see Figure 3.1) than the concentration reached in the PPP inhibition experiments, which was 5-6 g/L WCW (data not shown).

A few PPP metabolites, such as glucose-6P, sedoheptulose-7P and ribose-5P were also successfully measured in the Exactive$^{TM}$ Orbitrap and in the QQQ instruments when these were used for on-line metabolomics (see Chapter 3, Figure 3.10 and Chapter 5, Figure 5.9). This indicates that the detection issues described in this chapter probably arise from metabolite loses during the extraction method.

## 7.3 Conclusions

Production and extracellular accumulation of PPP metabolites was observed in *E. coli* during anaerobic succinate production, leading to the hypothesis that the inhibition of the PPP at this stage of the bioprocess could cause the rechannelling of carbon into the Embden-Meyerhof-Parnas pathway and improve succinate production. However, the addition of oxythiamine - a transketolase inhibitor often used in the literature to inhibit the PPP - at the beginning of the succinate production phase caused a significant decrease in succinate titre for oxythiamine concentrations of 10 mM. This is probably the consequence of the requirement of two moles of reduced NADH for the anaerobic formation of two moles of succinate via the anti-clockwise reductive TCA cycle. This NADH needs to be generated either with the production of PEP, pyruvate, or via the PPP, which generates two moles of NADPH for every mole of glucose-6P that enters the pathway. This NADPH can then be converted into NADH by the enzyme transhydrogenase, which is present in *E. coli*.

Samples were taken before and after the addition of oxythiamine, as well as at the end of the succinate production phase to measure the PPP metabolites by LC-MS and try to observe PPP inhibition. However, it was not possible to reliably measure the PPP metabolites in the biological samples. This is probably due to the concentration of the metabolites being below the LOD of the method used, potentially due to metabolite losses during the LC-MS extraction method.

## 7.4 Materials and methods

The materials and methods are as described in Chapter 2. Any differences and particular methods used exclusively in this chapter are described below.

### 7.4.1 Experimental conditions PPP inhibition

#### Growth conditions

Biomass grow-up for PPP inhibition studies was performed in six 2 L baffled shake flasks containing 330 mL of the growth media (see Section 2.3), inoculated with 100 $\mu$L of *E. coli* cell bank. These were incubated for 18.5 h at 37 ℃ and 200 rpm to an $OD_{600}$ of 4.40, and then combined into one single pre-culture before starting the succinate production phase.

#### Succinate production conditions

The combined culture was divided in 60 mL fractions into different 100 mL Schott bottles containing glucose, sodium bicarbonate and oxythiamine at a final concentration of 20 g/L glucose, 5 g/L sodium bicarbonate, and a different concentration of oxythiamine depending on the bottle. The Schott bottles were sparged for 10 seconds with $CO_2$ in the bottle headspace before closing the lid and incubating them at 37 ° C and 150 rpm for 24 h. The $CO_2$ was filtered through a Midisart® 2000 (Sartorius Stedim) 0.2 $\mu$m pore size filter to avoid microbial contamination. The different concentrations of oxythiamine used were 0, 200, 370, 1000, 2000, 5000 and 10,000 $\mu$M, and the stock of oxythiamine used was prepared at 200 mM. These concentrations were chosen as 100 $\mu$M oxythiamine had been reported to cause 82.6 % inhibition of thiamine uptake in *E. coli* (Kawasaki and Esaki, 1971). Samples were taken for LC-MS analysis right before and after starting the succinate production phase, as well as at the end of the succinate production phase.

### 7.4.2 Boiling ethanol sample extraction method for LC-MS analysis

A boiling ethanol method was used for intracellular metabolites extraction as a comparison with the 1:3:1 C:M:W method described in Section 2.7.1. For the boiling ethanol method, cell pellets were resuspended in 1 mL of cold 60 % methanol/water (v/v) buffered with 10 mM ammonium acetate (pH 7.1). At this point the samples were incubated on dry ice for three minutes and then spun down for 5 min at 5000 g and -10 ℃. The supernatant was discarded and 1 mL of 75 % ethanol/water (v/v) at 90 ℃ buffered with 10 mM ammonium acetate (pH 7.8) was added to the pellet. At this point the samples were incubated in a water bath at 90 ℃ for three minutes. Then, the samples were cooled down in dry ice for three minutes and spun down for 5 min at 5000 g and -10 ℃. Finally, 800 $\mu$L of supernatant were transferred into a new microtube and stored at -80 ℃ until LC-MS analysis.

### 7.4.3 LC-MS analysis

#### Ion mobility Q-TOF

LC-MS analysis using the IM-Q-TOF was performed as indicated by Pičmanová *et al.* (2022). Briefly, metabolites were separated in an InfinityLab Poroshell 120 zwitterionic hydrophilic interaction liquid chromatography (HILIC-Z) column (Agilent Technologies) (2.1 mm × 50 mm, 2.7 $\mu$m particle size) using an Agilent 1290 Infinity II UHPLC system. The mobile phases used in negative mode were 10 mM ammonium acetate (pH 9.0) for solvent A, and 10 % 10 mM ammonium acetate with 90 % acetonitrile for solvent B (pH 9.0). A linear gradient was applied from 93 to 80 % solvent B for 1.80 min, followed by a liner gradient from 80 to 70 % solvent B for 0.20 min and a further 0.30 min at 70 % solvent B. Then, the column was re-equilibrated with a linear gradient from 70 to 93 % solvent B for 0.05 min and the mobile phase was held at 93 % solvent B for 1.15 min. The total flow rate was 800 $\mu$L/min, column temperature was maintained at 30 ℃, sample injection volume was 1 $\mu$L and samples were maintained at 4 ℃ for the duration of the analysis.

Metabolite detection was done in an Agilent 6560 IM-Q-TOF with a Dual Agilent Jet Stream Electron Ionisation source. More details of the mass spectrometry parameters can be found on the original paper (Pičmanová *et al.*, 2022).

#### Triple quadrupole

Metabolite separation was performed adapted from Teleki *et al.* (2015). Briefly, metabolites were separated in a polymeric zwitterionic hydrophilic interaction liquid chromatography (ZIC®-pHILIC) column (Merck SeQuant®) (150 mm x 4.6 mm, 5 $\mu$m particle size) using an Thermo Scientific$^{TM}$ UltiMate$^{TM}$ 3000 UHPLC system. The mobile phases used were 10 % 10 mM ammonium acetate with 90 % acetonitrile for solvent A, and 90 % 10 mM ammonium acetate with 10 % acetonitrile for solvent B. Both adjusted to pH 9.2 and percentages are given in volume/volume ratios. The mobile phase gradient applied was 0 % solvent B for 1 min, followed by a linear gradient running from 0 to 75 % solvent B for 30 min. Then a linear gradient from 75 to 100 % solvent B for 4 min. Afterwards, the column was washed with 100 % solvent B for 5 min and then re-equilibrated using a linear gradient running from 100 to 0 % solvent B for 10 min and held at 0 % solvent B for 15 min. The total flow rate was 500 $\mu$L/min, column temperature was maintained at 40 ℃, sample injection volume was 10 $\mu$L, samples were maintained at 4 ℃ for the duration of the analysis and a HESI probe was used on the ion source.

Metabolite detection was done in a Thermo Scientific$^{TM}$ TSQ Quantiva triple quadrupole mass spectrometer using the MRM method in polarity switching mode (see Section 2.8) with a spray voltage of ±3.5 kV. Capillary temperature was set to 150 ℃, ion transfer tube (sheath gas) temperature set to 325 ℃, sheath gas flow rate 40 a.u., auxiliary gas 5 a.u. and sweep gas 0 a.u.

# PART IV

# General discussion and conclusions

# Chapter 8

# General discussion and conclusions

Industrial biotechnology has the potential to play a key role in the transition from an economic model based on non-renewable fossil fuels to a model based on the use of renewable sources to produce energy and goods. One example is succinic acid, a dicarboxylic acid that can be synthesised chemically using fossil fuels or by microbial fermentation using renewable feedstocks. However, the successful production of succinic acid by fermentation depends on it being competitive with the equivalent chemical processes. Although fermentation processes can be more environmentally friendly than chemical processes, they usually achieve lower product titres, which require more expensive purification costs. This might not be a decisive factor for high-value products that are challenging or impractical to produce chemically - such as monoclonal antibodies, fusion proteins, enzymes or large peptide hormones - but it often is for commodity chemicals, which are produced in large quantities at a low price per volume. For this reason, it is important to develop better technologies that can help optimise and increase the success rate of fermentation processes, as well as ensuring that the end product quality meets the required market standards.

## 8.1 Development of an on-line metabolomics platform

In this thesis, a system for on-line metabolomics monitoring of undiluted fermentation samples was developed and implemented with an *E. coli* fermentation process of succinate production. The monitoring of biomass, metabolites and proteins is gaining increased attention in the biotech industry in order to optimise and increase the success rate of bioprocesses and allow the transition from batch-mode manufacturing methods to more efficient pipelines such as continuous mode.

Using a high-resolution accurate mass Exactive Orbitrap mass spectrometer it was possible to monitor 886 different **untargeted** signals of the succinate bioprocess every 5 min with on-line metabolomics. The vast detection capacity of mass spectrometry and the ability to use it without the need to develop extensive calibration models make this technology **particularly useful for process development**. More specifically, alternative technologies used for bioprocess monitoring, such as vibrational spectroscopy - especially NIR, MIR and

Raman spectroscopy - require the development of chemometric models like PLS regression to deconvolute the measured spectra into different compounds. More importantly, these PLS regression models are very specific to the conditions used to develop them and, thus, have little transferability if something in the process is changed, such as the microbial strain or cell line, the bioreactor, or the process conditions (temperature, pH, medium composition, *etc.*). This is a big limitation for bioprocess optimisation, when different parameters might need to be iteratively changed in order to improve the process. Furthermore, real-time metabolite monitoring techniques commonly used today (on-line HPLC, vibrational spectroscopy and enzymatic analysers) are used in a targeted fashion for the monitoring of a pre-defined set of compounds, which limits the exploratory potential of untargeted metabolomics.

With on-line untargeted metabolomics, the time evolution of 886 ions was monitored during the succinate bioprocess, describing patterns that can help identify biomarkers of interest (such as for biomass or carbon starvation) and relevant pathways for the bioprocess, which can be targeted for genetic or process engineering (such as the pentose phosphate pathway).

It must be said that using the chromatography-free on-line metabolomics system with a mass spectrometer without fragmentation (such as the Exactive Orbitrap) has certain limitations. Despite its high-resolution accurate mass properties, the confident annotation of any of the signals observed is very limited. In this work, it was necessary to validate the annotation of the different compounds by off-line LC-MS using chromatography and reference standards. However, for a given organism, most of these annotations will still be valid if the process conditions are changed (temperature, pH, medium composition, *etc.*). If a different microbial strain or cell line is used, these annotations might still be valid, as long as it is the same organism species. On the other hand, if the organism is changed, larger metabolic differences should be expected and, therefore, the annotations determined by LC-MS might need to be validated again.

The potential of on-line metabolomics as a targeted monitoring method was also studied, as it is detailed below.

## 8.2 A targeted method for on-line metabolomics

Based on compounds of interest identified by untargeted on-line metabolomics - and further confirmation of these identifications by LC-MS - a method of **targeted** on-line metabolomics was developed to specifically monitor these compounds in a more quantitative manner using a triple quadrupole mass spectrometer in MRM mode. This way, 41 metabolites were monitored with the targeted on-line metabolomics method during three *E. coli* succinate fermentation replicates and univariate linear regression models were built to correlate the mass spectrometry signal to the concentration of 12 metabolites and the biomass, demonstrating how

metabolomics can be used to monitor the concentration of multiple bioprocess metabolites of interest on-line. At this point, it is important to mention that additional calibration curves and a more extended MRM method could allow the quantitative monitoring of a larger number of compounds.

The performance of these monitoring models was comparable to monitoring models described in the literature for vibrational spectroscopy, although the predictions for some metabolites were not fully satisfactory. The predictions of vibrational spectroscopy generally had higher $R^2$ and lower RMSE values, but this is to be expected because these models are typically built with multivariate PLS regression, which will naturally explain a larger proportion of the experimental variance (by using more variables to fit the model). However, this has a higher risk of overfitting, which makes these models less robust to changes in the bioprocess. The univariate linear regression models built for on-line metabolomics, on the other hand, are much simpler and more direct, only correlating the concentration of a metabolite to its single mass spectrometry peak area. Therefore, the models developed in this thesis for on-line metabolomics monitoring are simpler, easier to build and should be more robust to changes in the bioprocess.

Targeted on-line metabolomics was performed using a biomass filtration probe, which should prolong the duration that the monitoring system can be used without blocking. It is worth mentioning that intracellular metabolites, such as sugar phosphates and amino acids, were observed even though a filtration probe was used. This could likely be caused by either an imperfect filtration capacity (which was measured at >99.94 % efficiency) or by the release of intracellular metabolites into the extracellular medium in the bioreactor due to cell lysis. The likelihood of the latter is higher during the anaerobic succinate production phase, as the cells used cannot grow on glucose anaerobically and experience cell death.

Manufacturers claim that these biomass filtration probes can be used for weeks without fouling, particularly for eukaryotic cells, as these are bigger than prokaryotic ones. However, due to the slow flow rates required for the proper operation of the probe, its implementation significantly extends the time delay of measuring the sample (31.5 min) even though the time gap between samples is still of 5 min. A few things need to be mentioned. First, it should be possible to shorten the time delay by making the sampling lines shorter and changing the diameter of the tubing. And secondly, targeted analysis can also be implemented without the filtration probe. In fact, biomass accumulation in the lines started to appear towards the end of the fermentation process, becoming a potential threat of blocking the system.

Overall, it was demonstrated that on-line metabolomics can be used for the **quantitative monitoring** of metabolites in the liquid phase of fermentation processes. This is specially important at manufacturing stages of the bioprocess, as this technology can help to quantitatively assess the state of many key compounds of the bioprocess and this can be implemented in process control, as it is mentioned below.

## 8.3 Kinetic modelling

A kinetic model was implemented to describe the biomass formation phase of the *E. coli* fermentation process, including biomass growth on glucose, acetate formation due to overflow metabolism, acetate consumption upon glucose depletion and the switch from biomass exponential growth to stationary phase after consumption of both carbon sources. In combination with the on-line targeted metabolomics data, it was possible to forecast the future evolution of key process compounds at different stages of the process. The successful forecasting of these compounds can be used with predictive control during product manufacturing in order to ensure proper bioprocess performance.

The predictions made in this thesis are a proof of concept of the potential of on-line metabolomics for bioprocess monitoring during product manufacturing. However, the error of these predictions was, in some instances, significantly large. In part, this is because the three compounds forecasted - acetate, biomass and glucose - were modelled as a function of glucose, which was particularly overestimated in one of the three fermentation replicates. This might have been caused by insufficient conditioning of the mass spectrometer before starting monitoring the fermentation, which could easily be tested in the future by incorporating more conditioning samples before the beginning of the fermentation.

Alternatively, other forecasting models could be used. For example, the monitoring of additional compounds could be incorporated into the model inputs (not just glucose), in order to better constrain the model predictions, for instance using the process stoichiometry. Another option could be to develop a machine learning predictive model to use many mass spectrometry signals as inputs and generate the forecast of key process metabolites as outputs.

## 8.4 Future perspectives

New technologies will be required in the coming decades in order to reduce the use of fossil fuels and minimise the effects of climate change. As biotechnology has the potential to play a key role in this transition, the development of technologies that have the potential to improve the efficiency, productivity and robustness of bioprocesses are expected in the near future. The use of technologies for monitoring key process compounds in fermentation processes has gained a lot of popularity in the last two-three decades, particularly vibrational spectroscopy. It is expected that these technologies will have a more extended use in the biotech industry in the near future. This is also the case for metabolomics.

This thesis demonstrates that monitoring the metabolites in the liquid phase of a fermentation processes in real-time can be used to improve the process in many different ways. The main applications of on-line monitoring explored in this thesis include gaining a better understanding of the process and finding engineering targets during process development, and being a tool for quantitative monitoring of key process compounds during product manufacturing.

As far as the on-line metabolomics sampling system detailed in this thesis is concerned, there are a couple of recommendations that could be implemented in the future to improve the system. For example, it might be possible to shorten the washing step to 2.50 min (three volume changes of the longest section of tubing with both water and solvent mixture) without leading to signal loss or blockage of the system. Depending on the bioprocess, this might not have a large effect - particularly for organisms with a slow metabolism - but it might be of great importance to increase the resolution of faster-changing processes. To give a specific example, in the targeted metabolomics fermentations, it would have been useful to analyse samples more frequently at the beginning of the fermentation, when there was more uncertainty in the glucose metabolomics measurements, potentially due to incomplete instrument conditioning. Another possible suggestion to improve the sampling system would be to add a mixing chamber to combine the fermentation sample with organic solvent to test if more intracellular metabolites are observed due to better cell lysis. In this thesis it was assumed that the cells from the fermentation sample lyse due to either contact with the organic solvent that is used to introduce the sample into the mass spectrometer, the high voltages in the ESI chamber, the vacuum in the mass spectrometer, or a combination of the three. The observation of intracellular metabolites on-line was used as evidence that cell lysis takes place, despite the fact that their observation could also be the result of cell lysis happening in the bioreactor. However, it is not straightforward to test if the current setup causes cell lysis because mass spectrometry is a sample destructive technique, making it impossible to recover the sample and test the state of the cells (for example by plating or flow cytometry). Finally, it must be noted that adding a mixing chamber might require adding more solvent to the sample, and therefore might result in diluting the intensity of the metabolites detected and decreasing the sensitivity of the system.

Concerning untargeted on-line metabolomics, this technology was applied in a pre-established fermentation process. The complete potential of this technology would become more apparent in the future when tested with new processes, particularly if different strains or cell lines are compared. In this scenario, untargeted on-line metabolomics would allow to very quickly determine which variants are more attractive candidates for the process studied based on the observation of different levels of products, by-products and the increased or decreased activity

of certain metabolic pathways. Furthermore, this technology could also be tested in the future to develop a new bioprocess from scratch, as it can be used to identify key process events - substrate limitation, product formation, *etc*. - and calculate production and consumption rates, as has been shown throughout the thesis.

As far as targeted on-line metabolomics is concerned, this thesis has shown how it can be used to monitor the concentration of specific metabolites and the biomass. As it has been mentioned above, a more extensive calibration should be performed in the future in order to monitor the concentration of many more metabolites, which is one of the main advantages of this technology compared to alternative ones, such as spectroscopic techniques. Furthermore, it would be important to test if the calibrations developed using the three fermentation replicates described in Chapter 5 are transferable to further succinate fermentation runs, and even to other bioprocesses using other organisms or media. In the future, it is also very important that the mass spectrometer is properly conditioned - particularly for targeted experiments - to ensure a stable signal at the beginning of the experiment. Another future suggestion would be to investigate if a better estimate of the biomass concentration could be found. In this thesis, the signal of (R)-2,3-dihydroxy-isovalerate was used as the unique parameter to calculate the biomass concentration with linear modelling. However, other models could be tested, including considering other biomarkers, or developing multivariate models using multiple metabolites. Finally, demonstrating that this technology can be used in combination with process control would reinforce its position as a candidate technology to be used in a production setup. For this reason, a strong recommendation would be to implement a bioprocess control system that can act responding to the on-line metabolomics signal measured in real-time.

As a final remark, it is worth mentioning that the complete strategy of starting with an untargeted metabolomics approach with a high-resolution mass spectrometer to find monitoring targets, validate them with LC-MS and transition to a quantitative targeted monitoring approach is laborious and requires of expensive and specialised equipment and personnel. For example, being able to fully dedicate a high-resolution and a QQQ mass spectrometers to bioprocess monitoring is not accessible to all companies or research institutes. Moreover, staff members trained in mass spectrometry and fermentation would be required to operate and understand this system (although a similar situation applies for alternative monitoring technologies). Although this scenario might be more feasible in a laboratory setup than in a manufacturing environment, repeated demonstration that this technology works at laboratory scale will be an instrumental driving force for its adoption for process manufacturing. For this reason, the application of on-line metabolomics monitoring to other bioprocesses and other production hosts would be of upmost importance in extending the use of this technology.

# References

# References


Abu-Absi, N. R., Kenty, B. M., Cuellar, M. E., Borys, M. C., Sakhamuri, S., Strachan, D. J., Hausladen, M. C. & Li, Z. J. (2011). Real time monitoring of multiple parameters in mammalian cell culture bioreactors using an in-line Raman spectroscopy probe. *Biotechnology and Bioengineering*, *108*(5), 1215–1221. https://doi.org/10.1002/bit.23023

Abu-Absi, N. R., Zamamiri, A., Kacmar, J., Balogh, S. J. & Srienc, F. (2003). Automated flow cytometry for acquisition of time-dependent population data. *Cytometry Part A*, *51A*(2), 87–96. https://doi.org/10.1002/CYTO.A.10016

Akesson, M., Hagander, P. & Axelsson, J. P. (2001). Avoiding acetate accumulation in *Escherichia coli* cultures using feedback control of glucose feeding. *Biotechnology and Bioengineering*, *73*(3), 223–230. https://doi.org/10.1002/bit.1054

Altuntaş, E., Krieg, A., Baumgaertel, A., Crecelius, A. C. & Schubert, U. S. (2013). ESI, APCI, and MALDI tandem mass spectrometry of poly(methyl acrylate)s: A comparison study for the structural characterization of polymers synthesized via CRP techniques and the software application to analyze MS/MS data. *Journal of Polymer Science, Part A: Polymer Chemistry*. https://doi.org/10.1002/pola.26529

Alva, A., Sabido-Ramos, A., Escalante, A. & Bolívar, F. (2020). New insights into transport capability of sugars and its impact on growth from novel mutants of *Escherichia coli*. *Applied Microbiology and Biotechnology*, *104*, 1463–1479. https://doi.org/10.1007/s00253-019-10335-x

Anane, E., López C, D. C., Neubauer, P. & Cruz Bournazou, M. N. (2017). Modelling overflow metabolism in *Escherichia coli* by acetate cycling. *Biochemical Engineering Journal*, *125*, 23–30. https://doi.org/10.1016/J.BEJ.2017.05.013

Armstrong, E. F. (1933). Enzymes: A discovery and its consequences. *Nature*, *131*(3311), 535–537. https://doi.org/10.1038/131535A0

Arnold, S. A., Crowley, J., Woods, N., Harvey, L. M. & McNeil, B. (2003). In-situ near infrared spectroscopy to monitor key analytes in mammalian cell cultivation. *Biotechnology and Bioengineering*, *84*(1), 13–19. https://doi.org/10.1002/bit.10738

Avery, C. E. (1881). Manufacture of lactates. https://patents.google.com/patent/US243827

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L. & Mori, H. (2006). Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular Systems Biology*, *2*(1), 1–11. https://doi.org/10.1038/MSB4100050

Baeshen, N. A., Baeshen, M. N., Sheikh, A., Bora, R. S., Ahmed, M. M. M., Ramadan, H. A., Saini, K. S. & Redwan, E. M. (2014). Cell factories for insulin production. *Microbial Cell Factories*, *13*(141), 1–9. https://doi.org/10.1186/S12934-014-0141-0

Barnett, J. A. (2003). Beginnings of microbiology and biochemistry: The contribution of yeast research. *Microbiology*, *149*(3), 557–567. https://doi.org/10.1099/MIC.0.26089-0

Bayer, B., von Stosch, M., Melcher, M., Duerkop, M. & Striedner, G. (2020). Soft sensor based on 2D-fluorescence and process data enabling real-time estimation of biomass in *Escherichia coli* cultivations. *Engineering in Life Sciences*, *20*(1-2), 26–35. https://doi.org/10.1002/ELSC.201900076

Beckmann, M., Parker, D., Enot, D. P., Duval, E. & Draper, J. (2008). High-throughput, nontargeted metabolite fingerprinting using nominal mass flow injection electrospray mass spectrometry. *Nature Protocols*, *3*(3), 486–504. https://doi.org/10.1038/nprot.2007.500

Behrendt, U., Koch, S., Gooch, D. D., Steegmans, U. & Comer, M. J. (1994). Mass spectrometry: A tool for on-line monitoring of animal cell cultures. *Cytotechnology*, *14*(3), 157–165. https://doi.org/10.1007/BF00749612

Bélanger, J. M., Paré, J. R. J. & Sigouin, M. (1997). High performance liquid chromatography (HPLC): Principles and applications. In J. M. Bélanger & J. R. J. Paré (Eds.), *Techniques and instrumentation in analytical chemistry* (pp. 37–59). Elsevier. https://doi.org/10.1016/S0167-9244(97)80011-X

Belini, V. L., Suhr, H. & Wiedemann, P. (2020). Online monitoring of the morphology of an industrial sugarcane biofuel yeast strain via in situ microscopy. *Journal of Microbiological Methods*, *175*, 105983. https://doi.org/10.1016/J.MIMET.2020.105973

Berbegal, C., Khomenko, I., Russo, P., Spano, G., Fragasso, M., Biasioli, F. & Capozzi, V. (2020). PTR-ToF-MS for the online monitoring of alcoholic fermentation in wine: Assessment of VOCs variability associated with different combinations of *Saccharomyces*/Non-*Saccharomyces* as a case-study. *Fermentation*, *6*(2), 55. https://doi.org/10.3390/FERMENTATION6020055

Birou, B., Marison, I. W. & Stockar, U. V. (1987). Calorimetric investigation of aerobic fermentations. *Biotechnology and Bioengineering*, *30*(5), 650–660. https://doi.org/10.1002/BIT.260300509

Bluma, A., Höpfner, T., Lindner, P., Rehbock, C., Beutel, S., Riechers, D., Hitzmann, B. & Scheper, T. (2010). In-situ imaging sensors for bioprocess monitoring: state of the art. *Analytical and Bioanalytical Chemistry*, *398*(6), 2429–2439. https://doi.org/10.1007/S00216-010-4181-Y

Boesl, U. (2017). Time-of-flight mass spectrometry: Introduction to the basics. *Mass Spectrometry Reviews*, *36*(1), 86–109. https://doi.org/10.1002/mas.21520

Bp. (2021). *Statistical Review of World Energy 2021* (tech. rep.). https://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy.html

Brecker, L., Rg Weber, H., Griengl, H. & Ribbons, D. W. (1999). In situ proton-NMR analyses of *Escherichia coli* HB101 fermentations in 1H2O and in D2O. *Microbiology*, *145*, 3389–3397. https://doi.org/10.1099/00221287-145-12-3389

Bremus, C., Herrmann, U., Bringer-Meyer, S. & Sahm, H. (2006). The use of microorganisms in l-ascorbic acid production. *Journal of Biotechnology*, *124*(1), 196–205. https://doi.org/10.1016/J.JBIOTEC.2006.01.010

Broger, T., Odermatt, R. P., Huber, P. & Sonnleitner, B. (2011). Real-time on-line flow cytometry for bioprocess monitoring. *Journal of Biotechnology*, *154*(4), 240–247. https://doi.org/10.1016/J.JBIOTEC.2011.05.003

Brown, D. R., Barton, G., Pan, Z., Buck, M. & Wigneshweraraj, S. (2014). Nitrogen stress response and stringent response are coupled in *Escherichia coli*. *Nature Communications*, *5*(1), 1–8. https://doi.org/10.1038/ncomms5115

Bruins, A. P., Covey, T. R. & Henion, J. D. (1987). Ion spray interface for combined liquid chromatography/atmospheric pressure ionization mass spectrometry. *Analytical Chemistry*. https://doi.org/10.1021/ac00149a003

Buchholz, A., Takors, R. & Wandrey, C. (2001). Quantification of Intracellular Metabolites in *Escherichia coli* K12 Using Liquid Chromatographic-Electrospray Ionization Tandem Mass Spectrometric Techniques. *Analytical Biochemistry*, *295*(2), 129–137. https://doi.org/10.1006/ABIO.2001.5183

Buchholz, K. & Collins, J. (2013). The roots - A short history of industrial microbiology and biotechnology. *Applied Microbiology and Biotechnology*, *97*(9), 3747–3762. https://doi.org/10.1007/S00253-013-4768-2

Buchner, E. (1897). Alkoholische Gährung ohne Hefezellen. *Berichte der deutschen chemischen Gesellschaft*, *30*(1), 117–124. https://doi.org/10.1002/CBER.18970300121

Buszewski, B. & Noga, S. (2012). Hydrophilic interaction liquid chromatography (HILIC)—a powerful separation technique. *Analytical and Bioanalytical Chemistry*, *402*(1), 231–247. https://doi.org/10.1007/S00216-011-5308-5

Cabaneros Lopez, P., Feldman, H., Mauricio-Iglesias, M., Junicke, H., Huusom, J. K. & Gernaey, K. V. (2019). Benchmarking real-time monitoring strategies for ethanol production from lignocellulosic biomass. *Biomass and Bioenergy*, *127*(105296), 1–14. https://doi.org/10.1016/j.biombioe.2019.105296

Cabaneros Lopez, P., Udugama, I. A., Thomsen, S. T., Roslander, C., Junicke, H., Mauricio-Iglesias, M. & Gernaey, K. V. (2020). Towards a digital twin: a hybrid data-driven and mechanistic digital shadow to forecast the evolution of lignocellulosic fermentation. *Biofuels, Bioproducts and Biorefining*, *14*(5), 1046–1060. https://doi.org/10.1002/BBB.2108

Cajka, T. & Fiehn, O. (2015). Toward Merging Untargeted and Targeted Methods in Mass Spectrometry-Based Metabolomics and Lipidomics. *Analytical Chemistry*, *88*(1), 524–545. https://doi.org/10.1021/ACS.ANALCHEM.5B04491

Cameron, A. E. & Eggers, D. F. (1948). An Ion "Velocitron". *Review of Scientific Instruments*, *19*, 607. https://doi.org/10.1063/1.1741336

Campos, C. G., Veras, H. C. T., de Aquino Ribeiro, J. A., Costa, P. P. K. G., Araújo, K. P., Rodrigues, C. M., de Almeida, J. R. M. & Abdelnur, P. V. (2017). New Protocol Based on UHPLC-MS/MS for Quantitation of Metabolites in Xylose-Fermenting Yeasts. *Journal of the American Society for Mass Spectrometry*, *28*(12), 2646–2657. https://doi.org/10.1007/s13361-017-1786-9

Canelas, A. B., Ten Pierick, A., Ras, C., Seifar, R. M., Van Dam, J. C., Van Gulik, W. M. & Heijnen, J. J. (2009). Quantitative Evaluation of Intracellular Metabolite Extraction Techniques for Yeast Metabolomics. *Analytical Chemistry*, *81*(17), 7379–7389. https://doi.org/10.1021/AC900999T

Cech, N. B. & Enke, C. G. (2001). Practical implications of some recent studies in electrospray ionization fundamentals. *Mass Spectrometry Reviews*, *20*(6), 362–387. https://doi.org/10.1002/MAS.10008

Chambers, M. C., MacLean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T. A., Brusniak, M. Y., Paulse, C., Creasy, D., . . . Mallick, P. (2012). A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology*, *30*(10), 918–920. https://doi.org/10.1038/nbt.2377

Chan, X. W. A., Wrenger, C., Stahl, K., Bergmann, B., Winterberg, M., Müller, I. B. & Saliba, K. J. (2013). Chemical and genetic validation of thiamine utilization as an antimalarial drug target. *Nature Communications*, *4*(1), 2060. https://doi.org/10.1038/ncomms3060

Chatfield, C. & Collins, A. J. (2018). *Introduction to Multivariate Analysis*. Routledge. https://doi.org/10.1201/9780203749999

Chatterjee, R., Millard, C. S., Champion, K., Clark, D. P. & Donnelly, M. I. (2001). Mutation of the ptsG gene results in increased production of succinate in fermentation of glucose by *Escherichia coli*. *Applied and Environmental Microbiology*, *67*(1), 148–154. https://doi.org/10.1128/AEM.67.1.148-154.2001

Chi, A., Bai, D. L., Geer, L. Y., Shabanowitz, J. & Hunt, D. F. (2007). Analysis of intact proteins on a chromatographic time scale by electron transfer dissociation tandem mass spectrometry. *International journal of mass spectrometry*, *259*(1-3), 203. https://doi.org/10.1016/J.IJMS.2006.09.030

Choi, M., Al-Zahrani, S. M. & Lee, S. Y. (2014). Kinetic model-based feed-forward controlled fed-batch fermentation of *Lactobacillus rhamnosus* for the production of lactic acid from Arabic date juice. *Bioprocess and Biosystems Engineering*, *37*(6), 1007–1015. https://doi.org/10.1007/S00449-013-1071-7

Chong, I.-G. & Jun, C.-H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, *78*(1-2), 103–112. https://doi.org/10.1016/J.CHEMOLAB.2004.12.011

Clementschitsch, F. & Bayer, K. (2006). Improvement of bioprocess monitoring: Development of novel concepts. *Microbial Cell Factories*, *5*(1), 1–11. https://doi.org/10.1186/1475-2859-5-19

Cocchi, M., Biancolillo, A. & Marini, F. (2018). Chemometric Methods for Classification and Feature Selection. *Comprehensive analytical chemistry* (pp. 265–299). Elsevier. https://doi.org/10.1016/BS.COAC.2018.08.006

Coene, K. L. M., Kluijtmans, L. A. J., van der Heeft, E., Engelke, U. F. H., de Boer, S., Hoegen, B., Kwast, H. J. T., van de Vorst, M., Huigen, M. C. D. G., Keularts, I. M. L. W., Schreuder, M. F., van Karnebeek, C. D. M., Wortmann, S. B., de Vries, M. C., Janssen, M. C. H., Gilissen, C., Engel, J. & Wevers, R. A. (2018). Next-generation metabolic screening: targeted and untargeted metabolomics for the diagnosis of inborn errors of metabolism in individual patients. *Journal of Inherited Metabolic Disease*, *41*(3), 337–353. https://doi.org/10.1007/s10545-017-0131-6

Cohen, S. N., Chang, A. C., Boyer, H. W. & Helling, R. B. (1973). Construction of Biologically Functional Bacterial Plasmids In Vitro. *Proceedings of the National Academy of Sciences*, *70*(11), 3240–3244. https://doi.org/10.1073/PNAS.70.11.3240

Collen, D. & Lijnen, H. R. (2004). Tissue-type plasminogen activator: a historical perspective and personal account. *Journal of thrombosis and haemostasis*, *2*(4), 541–546. https://doi.org/10.1111/J.1538-7933.2004.00645.X

Colomban, P. & Gouadec, G. (2008). Raman Scattering Theory and Elements of Raman Instrumentation. In M. S. Amer (Ed.), *Raman spectroscopy for soft matter applications* (pp. 11–29). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470475997.CH2

Comín-Anduix, B., Boren, J., Martinez, S., Moro, C., Centelles, J. J., Trebukhina, R., Petushok, N., Lee, W. N. P., Boros, L. G. & Cascante, M. (2001). The effect of thiamine supplementation on tumour roliferation: A metabolic control analysis study. *European Journal of Biochemistry*, *268*(15), 4177–4182. https://doi.org/10.1046/j.1432-1327.2001.02329.x

Cooper, G. M. (2000). *Cell Membranes*. Sinauer Associates. https://www.ncbi.nlm.nih.gov/books/NBK9928/

Cos, O., Ramon, R., Montesinos, J. L. & Valero, F. (2006). A simple model-based control for *Pichia pastoris* allows a more efficient heterologous protein production bioprocess. *Biotechnology and Bioengineering*, *95*(1), 145–154. https://doi.org/10.1002/BIT.21005

Costello, Z. & Martin, H. G. (2018). A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. *npj Systems Biology and Applications*, *4*(1), 1–14. https://doi.org/10.1038/s41540-018-0054-3

Craven, S., Shirsat, N., Whelan, J. & Glennon, B. (2013). Process model comparison and transferability across bioreactor scales and modes of operation for a mammalian cell bioprocess. *Biotechnology Progress*, *29*(1), 186–196. https://doi.org/10.1002/BTPR.1664

Creek, D. J., Jankevics, A., Breitling, R., Watson, D. G., Barrett, M. P. & Burgess, K. E. (2011). Toward global metabolomics analysis with hydrophilic interaction liquid chromatography-mass spectrometry: Improved metabolite identification by retention time prediction. *Analytical Chemistry*, *83*(22), 8703–8710. https://doi.org/10.1021/ac2021823

Cronan, J. E., Littel, K. J. & Jackowski, S. (1982). Genetic and biochemical analyses of pantothenate biosynthesis in *Escherichia coli* and *Salmonella typhimurium*. *Journal of Bacteriology*, *149*(3), 916. https://doi.org/10.1128/jb.149.3.916-922.1982

Csonka, L. N. & Fraenkel, D. G. (1997). Pathways of NADPH Formation in *Escherichia coli*. *The Journal of Biological Chemistry*, *252*(10), 3382–3391. https://doi.org/10.1016/s0021-9258(17)40402-9

Custer, T. G., Wagner, W. P., Kato, S., Bierbaum, V. M. & Fall, R. (2003). Potential of On-Line CIMS for Bioprocess Monitoring. *Biotechnology Progress*, *19*(4), 1355–1364. https://doi.org/10.1021/bp025730k

da Silva, T. L., Roseiro, J. C. & Reis, A. (2012). Applications and perspectives of multiparameter flow cytometry to microbial biofuels production processes. *Trends in Biotechnology*, *30*(4), 225–232. https://doi.org/10.1016/J.TIBTECH.2011.11.005

Davidson, M. R. & Harvie, D. J. (2007). Electroviscous effects in low Reynolds number liquid flow through a slit-like microfluidic contraction. *Chemical Engineering Science*, *62*(16), 4229–4240. https://doi.org/10.1016/j.ces.2007.05.006

De Mey, M., De Maeseneire, S., Soetaert, W. & Vandamme, E. (2007). Minimizing acetate formation in *E. coli* fermentations. *Journal of Industrial Microbiology and Biotechnology*, *34*(11), 689–700. https://doi.org/10.1007/s10295-007-0244-2

Demain, A. L., Vandamme, E. J., Collins, J. & Buchholz, K. (2016). History of Industrial Biotechnology. In C. Wittmann & J. C. Liao (Eds.), *Industrial biotechnology : Microorganisms.* (pp. 3–84). John Wiley & Sons, Incorporated. https://doi.org/10.1002/9783527807796.ch1

Deshmukh, K., Sankaran, S., Ahamed, B., Sadasivuni, K. K., Pasha, K. S., Ponnamma, D., Rama Sreekanth, P. S. & Chidambaram, K. (2017). Dielectric Spectroscopy. *Spectroscopic methods for nanomaterials characterization* (pp. 237–299). Elsevier. https://doi.org/10.1016/B978-0-323-46140-5.00010-8

do Nascimento, R. J. A., de Macedo, G. R., dos Santos, E. S. & de Oliveira, J. A. (2017). Real time and in situ near-infrared spectroscopy (NIRS) for quantitative monitoring of biomass, glucose, ethanol and glycerine concentrations in an alcoholic fermentation. *Brazilian Journal of ChemicalEngineering*, *34*(02), 459–468. https://doi.org/10.1590/0104-6632.20170342s20150347

Doran, P. M. (2013). Reactor Engineering. *Bioprocess engineering principles* (Second, pp. 761–852). Elsevier. https://doi.org/10.1016/C2009-0-22348-8

E4tech, RE-CORD & WUR. (2015). *From the Sugar Platform to biofuels and biochemicals* (tech. rep.). European Commission. https://ec.europa.eu/energy/sites/ener/files/documents/EC%20Sugar%20Platform%20final%20report.pdf

Farrell, A., McLoughlin, N., Milne, J. J., Marison, I. W. & Bones, J. (2014). Application of multi-omics techniques for bioprocess design and optimization in Chinese hamster ovary cells. *Journal of Proteome Research*, *13*(7), 3144–3159. https://doi.org/10.1021/pr500219b

Favaro, L., Viktor, M. J., Rose, S. H., Viljoen-Bloom, M., van Zyl, W. H., Basaglia, M., Cagnin, L. & Casella, S. (2015). Consolidated bioprocessing of starchy substrates into ethanol by industrial *Saccharomyces cerevisiae* strains secreting fungal amylases. *Biotechnology and Bioengineering*, *112*(9), 1751–1760. https://doi.org/10.1002/BIT.25591

FDA. (2004). *Guidance for Industry PAT - A Framework for Innovative Pharmaceutical Development, manufacturing, and Quality Assurance* (tech. rep.). https://www.fda.gov/regulatory-information/search-fda-guidance-documents/pat-framework-innovative-pharmaceutical-development-manufacturing-and-quality-assurance

Fiehn, O. (2002). Metabolomics – the link between genotypes and phenotypes. *Plant Molecular Biology*, *48*(1), 155–171. https://doi.org/10.1023/A:1013713905833

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., ... Venter, J. C. (1995). Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd. *Science*, *269*(5223), 496–512. https://doi.org/10.1126/SCIENCE.7542800

Fleming, A. (1929). On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of *B. influenzæ*. *British journal of experimental pathology*, *10*(3), 236. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2048009/

Fuhrer, T., Heer, D., Begemann, B. & Zamboni, N. (2011). High-throughput, accurate mass metabolome profiling of cellular extracts by flow injection-time-of-flight mass spectrometry. *Analytical Chemistry*, *83*(18), 7074–7080. https://doi.org/10.1021/ac201267k

Fuhrer, T. & Zamboni, N. (2015). High-throughput discovery metabolomics. *Current Opinion in Biotechnology*, *31*, 73–78. https://doi.org/10.1016/j.copbio.2014.08.006

Fukui, K., Nanatani, K., Hara, Y., Yamakami, S., Yahagi, D., Chinen, A., Tokura, M. & Abe, K. (2017). *Escherichia coli* yjjpb genes encode a succinate transporter important for succinate production. *Bioscience, Biotechnology and Biochemistry*, *81*(9), 1837–1844. https://doi.org/10.1080/09168451.2017.1345612

Garthwaite, P. H. (1994). An Interpretation of Partial Least Squares. *Journal of the American Statistical Association*, *89*(425), 122–127. https://doi.org/10.1080/01621459.1994.10476452

Ge, Y., Lawhorn, B. G., ElNaggar, M., Strauss, E., Park, J. H., Begley, T. P. & McLafferty, F. W. (2002). Top Down Characterization of Larger Proteins (45 kDa) by Electron Capture Dissociation Mass Spectrometry. *Journal of the American Chemical Society*, *124*(4), 672–678. https://doi.org/10.1021/JA011335Z

Gloaguen, Y., Morton, F., Daly, R., Gurden, R., Rogers, S., Wandy, J., Wilson, D., Barrett, M. & Burgess, K. (2017). PiMP my metabolome: An integrated, web-based tool for LC-MS metabolomics data. *Bioinformatics*, *33*(24), 4007–4009. https://doi.org/10.1093/bioinformatics/btx499

Goh, H.-Y., Sulu, M., Alosert, H., Lewis, G. L., Josland, G. D. & Merriman, D. E. (2020). Applications of off-gas mass spectrometry in fed-batch mammalian cell culture. *Bioprocess and Biosystems Engineering*, *43*(3), 483–493. https://doi.org/10.1007/s00449-019-02242-2

Gokarn, R. R., Eiteman, M. A. & Altman, E. (2000). Metabolic analysis of *Escherichia coli* in the presence and absence of the carboxylating enzymes phosphoenolpyruvate carboxylase and pyruvate carboxylase. *Applied and Environmental Microbiology*, *66*(5), 1844–1850. https://doi.org/10.1128/AEM.66.5.1844-1850.2000

Golabgir, A. & Herwig, C. (2016). Combining Mechanistic Modeling and Raman Spectroscopy for Real-Time Monitoring of Fed-Batch Penicillin Production. *Chemie-Ingenieur-Technik*, *88*(6), 764–776. https://doi.org/10.1002/cite.201500101

Gonzalez, B., Franc, J. & Renaud, M. (1997). A Rapid and Reliable Method for Metabolite Extraction in Yeast using Boiling Buffered Ethanol. *Yeast*, *13*, 1347–1356. https://doi.org/10.1002/(sici)1097-0061(199711)13:14%7B%5C%%7D3C1347::aid-yea176%7B%5C%%7D3E3.0.co;2-o

Gorrochategui, E., Jaumot, J., Lacorte, S. & Tauler, R. (2016). Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow. *Trends in Analytical Chemistry*, *82*, 425–442. https://doi.org/10.1016/J.TRAC.2016.07.004

Goudar, C. T., Piret, J. M. & Konstantinov, K. B. (2011). Estimating cell specific oxygen uptake and carbon dioxide production rates for mammalian cells in perfusion culture. *Biotechnology Progress*, *27*(5), 1347–1357. https://doi.org/10.1002/btpr.646

Grobbelaar, J. U. (2009). Turbidity. In G. E. Likens (Ed.), *Encyclopedia of inland waters* (pp. 699–704). Academic Press. https://doi.org/10.1016/B978-012370626-3.00075-2

Gross, J. H. (2017). *Mass Spectrometry* (Third Edit). Springer.

Haack, M. B., Eliasson, A. & Olsson, L. (2004). On-line cell mass monitoring of *Saccharomyces cerevisiae* cultivations by multi-wavelength fluorescence. *Journal of Biotechnology*, *114*(1-2), 199–208. https://doi.org/10.1016/J.JBIOTEC.2004.05.009

Hamilton, S. E., Mattrey, F., Bu, X., Murray, D., McCullough, B. & Welch, C. J. (2014). Use of a miniature mass spectrometer to support pharmaceutical process chemistry. *Organic Process Research and Development*, *18*(1), 103–108. https://doi.org/10.1021/OP400253X

Harada, L. H., Da Costa, A. C. & Maciel Filho, R. (2002). Hybrid neural modeling of bioprocesses using functional link networks. *Applied Biochemistry and Biotechnology*, *98*(1), 1009–1023. https://doi.org/10.1385/ABAB:98-100:1-9:1009

Harada, Y., Sakata, K., Sato, S. & Takayama, S. (2014). Fermentation Pilot Plant. *Fermentation and biochemical engineering handbook: Principles, process design, and equipment: Third edition* (Third Edit, pp. 3–15). Elsevier Inc. https://doi.org/10.1016/B978-1-4557-2553-3.00001-5

Harris, C. M., Todd, R. W., Bungard, S. J., Lovitt, R. W., Morris, J. G. & Kell, D. B. (1987). Dielectric permittivity of microbial suspensions at radio frequencies: a novel method for the real-time estimation of microbial biomass. *Enzyme and Microbial Technology*, *9*(3), 181–186. https://doi.org/10.1016/0141-0229(87)90075-5

Hashimoto, S.-I. (2017). Discovery and History of Amino Acid Fermentation. In A. Yokota & M. Ikeda (Eds.), *Amino acid fermentation* (pp. 15–34). Springer, Tokyo. https://doi.org/10.1007/10_2016_24

Hewitt, C. J., Boon, L. A., Mcfarlane, C. M. & Nienow, A. W. (1998). The Use of Flow Cytometry to Study the Impact of Fluid Mechanical Stress on *Escherichia coli* W3110 During Continuous Cultivation in an Agitated Bioreactor. *Biotechnology and Bioengineering*, *59*(5), 612–620. https://doi.org/10.1002/(SICI)1097-0290(19980905)59:5

Hiller, J., Franco-Lara, E. & Weuster-Botz, D. (2007). Metabolic profiling of *Escherichia coli* cultivations: Evaluation of extraction and metabolite analysis procedures. *Biotechnology Letters*, *29*(8), 1169–1178. https://doi.org/10.1007/S10529-007-9384-8

Hisiger, S. & Jolicoeur, M. (2005). A multiwavelength fluorescence probe: Is one probe capable for on-line monitoring of recombinant protein production and biomass activity? *Journal of Biotechnology*, *117*(4), 325–336. https://doi.org/10.1016/J.JBIOTEC.2005.03.004

Hoffmann, E. & Stroobant, V. (2007). *Mass Spectrometry Principles and Applications* (Third Edit). John Wiley & Sons Ltd.

Houbraken, J., Frisvad, J. C. & Samson, R. A. (2011). Fleming's penicillin producing strain is not *Penicillium chrysogenum* but *P. rubens*. *IMA Fungus*, *2*(1), 87–95. https://doi.org/10.5598/IMAFUNGUS.2011.02.01.12

Hu, Q., Noll, R. J., Li, H., Makarov, A., Hardman, M. & Cooks, R. G. (2005). The Orbitrap: A new mass spectrometer. *Journal of Mass Spectrometry*, *40*(4), 430–443. https://doi.org/10.1002/jms.856

Huang, N., Siegel, M. M., Kruppa, G. H. & Laukien, F. H. (1999). Automation of a Fourier Transform Ion Cyclotron Resonance Mass Spectrometer for Acquisition, Analysis, and E-mailing of High-Resolution Exact-Mass Electrospray Ionization Mass Spectral Data. *Journal of the American Society for Mass Spectrometry*, (10), 1166–1173. https://doi-org.ezproxy.is.ed.ac.uk/10.1016/S1044-0305(99)00089-6

Idle, J. R. & Gonzalez, F. J. (2007). Metabolomics. *Cell Metabolism*, *6*(5), 348–351. https://doi.org/10.1016/j.cmet.2007.10.005

IPCC. (2014). *Chapter Climate Change 2014 Synthesis Report Summary for Policymakers* (tech. rep.). The Intergovernmental Panel on Climate Change. https://www.ipcc.ch/report/ar5/syr/

Irving, H., Freiser, H. & West, T. (1978). Nomenclature, symbols, units and their usage in spetrochemical analysis–II. Data interpretation. In H. Irving, H. Freiser & T. West (Eds.), *Compendium of analytical nomenclature* (pp. 114–117). Pergamon. https://doi.org/10.1016/B978-0-08-022008-6.50020-8

Iversen, J. A., Berg, R. W. & Ahring, B. K. (2014). Quantitative monitoring of yeast fermentation using Raman spectroscopy. *Analytical and Bioanalytical Chemistry*, *406*(20), 4911–4919. https://doi.org/10.1007/s00216-014-7897-2

Jestel, N. L. (2010). Raman Spectroscopy. In K. A. Bakeev (Ed.), *Process analytical technology: Spectroscopic tools and implementation strategies for the chemical and pharmaceutical industries* (Second, pp. 195–243). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470689592.CH7

Jiang, M., Ma, J., Wu, M., Liu, R., Liang, L., Xin, F., Zhang, W., Jia, H. & Dong, W. (2017). Progress of succinic acid production from renewable resources: Metabolic and fermentative strategies. *Bioresource Technology*, *245*, 1710–1717. https://doi.org/10.1016/j.biortech.2017.05.209

Jozefczuk, S., Klie, S., Catchpole, G., Szymanski, J., Cuadros-Inostroza, A., Steinhauser, D., Selbig, J. & Willmitzer, L. (2010). Metabolomic and transcriptomic stress response of *Escherichia coli*. *Molecular Systems Biology*, *6*(1), 364–379. https://doi.org/10.1038/MSB.2010.18

Kacmar, J. & Srienc, F. (2005). Dynamics of single cell property distributions in Chinese hamster ovary cell cultures monitored and controlled with automated flow cytometry. *Journal of Biotechnology*, *120*(4), 410–420. https://doi.org/10.1016/J.JBIOTEC.2005.06.031

Kamiloglu, S., Sari, G., Ozdal, T. & Capanoglu, E. (2020). Guidelines for cell viability assays. *Food Frontiers*, *1*(3), 332–349. https://doi.org/10.1002/FFT2.44

Kassambara, A. (2020). 'ggplot2' Based Publication Ready Plots. https://cran.r-project.org/package=ggpubr

Kawasaki, T. & Esaki, K. (1971). Thiamine uptake in *Escherichia coli*. III. Regulation of thiamine uptake in *Escherichia coli*. *Archives of Biochemistry and Biophysics*, *142*(1), 163–169. https://doi.org/10.1016/0003-9861(71)90271-2

Kim, J., Cheong, Y. E., Jung, I. & Kim, K. H. (2019). Metabolomic and Transcriptomic Analyses of *Escherichia coli* for Efficient Fermentation of L-Fucose. *Marine Drugs*, *17*(82), 1–14. https://doi.org/10.3390/MD17020082

Kiviharju, K., Salonen, K., Moilanen, U. & Eerikäinen, T. (2008). Biomass measurement online: the performance of in situ measurements and software sensors. *Journal of Industrial Microbiology and Biotechnology*, *35*(7), 657–665. https://doi.org/10.1007/S10295-008-0346-5

Klimek, J. & Ollis, D. F. (1980). Extracellular microbial polysaccharides: Kinetics of *Pseudomonas* sp., *Azotobacter vinelandii*, and *Aureobasidium pullulans* batch fermentations. *Biotechnology and Bioengineering*, *22*(11), 2321–2342. https://doi.org/10.1002/BIT.260221109

Koch, C., Posch, A. E., Herwig, C. & Lendl, B. (2016). Comparison of Fiber Optic and Conduit Attenuated Total Reflection (ATR) Fourier Transform Infrared (FT-IR) Setup for In-Line Fermentation Monitoring. *Applied Spectroscopy*, *70*(12), 1965–1973. https://doi.org/10.1177/0003702816662618

Kohler, R. (1971). The background to Eduard Buchner's discovery of cell-free fermentation. *Journal of the History of Biology*, *4*(1), 35–61. https://doi.org/10.1007/BF00356976

Köhler, G. & Milstein, C. (1975). Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature*, *256*(5517), 495–497. https://doi.org/10.1038/256495a0

Koliander, W., Arnezeder, C. & Hampel, W. A. (1990). A Simple and Versatile System for Fermentation Control by On-Line HPLC Analysis of Medium Components. *Acta Biotechnol*, *10*(5), 387–394. https://doi.org/10.1002/abio.370100502

Konermann, L., Ahadi, E., Rodriguez, A. D. & Vahidi, S. (2013). Unraveling the Mechanism of Electrospray Ionization. *Analytical Chemistry*, *85*(1), 2–9. https://doi.org/10.1021/AC302789C

Kreyenschulte, D., Paciok, E., Regestein, L., Blümich, B. & Büchs, J. (2015). Online monitoring of fermentation processes via non-invasive low-field NMR. *Biotechnology and Bioengineering*, *112*(9), 1810–1821. https://doi.org/10.1002/bit.25599

Krishnan, M. S., Ho, N. W. Y. & Tsao, G. T. (1999). Fermentation kinetics of ethanol production from glucose and xylose by recombinant *Saccharomyces* 1400(pLNH33). *Applied Biochemistry and Biotechnology*, *77-79*(1), 373–388. https://doi.org/10.1385/ABAB:78:1-3:373

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B. & core team, R. (2021). Classification and Regression Training. https://cran.r-project.org/package=caret

Lander, E. S. (2011). Initial impact of the sequencing of the human genome. *Nature*, *470*(7333), 187–197. https://doi.org/10.1038/NATURE09792

Lange, O., Damoc, E., Wieghaus, A. & Makarov, A. (2014). Enhanced Fourier transform for Orbitrap mass spectrometry. *International Journal of Mass Spectrometry*, *369*, 16–22. https://doi.org/10.1016/J.IJMS.2014.05.019

Larkin, P. (2011a). Basic Principles. *Infrared and raman spectroscopy* (pp. 7–25). Elsevier. https://doi.org/10.1016/B978-0-12-386984-5.10002-3

Larkin, P. (2011b). Introduction: Infrared and Raman Spectroscopy. *Infrared and raman spectroscopy* (pp. 1–5). Elsevier. https://doi.org/10.1016/B978-0-12-386984-5.10001-1

Lee, H. L. T., Boccazzi, P., Gorret, N., Ram, R. J. & Sinskey, A. J. (2004). In situ bioprocess monitoring of *Escherichia coli* bioreactions using Raman spectroscopy. *Vibrational Spectroscopy*, *35*(1-2), 131–137. https://doi.org/10.1016/j.vibspec.2003.12.015

Lee, L. C., Liong, C.-Y. & Jemain, A. A. (2018). Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. *Analyst*, *143*(15), 3526–3539. https://doi.org/10.1039/C8AN00599K

Lee, S. Y., Kim, H. U., Chae, T. U., Cho, J. S., Kim, J. W., Shin, J. H., Kim, D. I., Ko, Y. S., Jang, W. D. & Jang, Y. S. (2019). A comprehensive metabolic map for production of bio-based chemicals. *Nature Catalysis*, *2*(1), 18–33. https://doi.org/10.1038/s41929-018-0212-4

Legner, R., Wirtz, A., Koza, T., Tetzlaff, T., Nickisch-Hartfiel, A. & Jaeger, M. (2019). Application of green analytical chemistry to a green chemistry process: Magnetic resonance and Raman spectroscopic process monitoring of continuous ethanolic fermentation. *Biotechnology and Bioengineering*, *116*(11), 2874–2883. https://doi.org/10.1002/BIT.27112

Levenspiel, O. (1980). The monod equation: A revisit and a generalization to product inhibition situations. *Biotechnology and Bioengineering*, *22*(8), 1671–1687. https://doi.org/10.1002/BIT.260220810

Li, M.-Y., Ebel, B., Paris, C., Chauchard, F., Guedon, E. & Marc, A. (2018). Real-time monitoring of antibody glycosylation site occupancy by *in situ* Raman spectroscopy during bioreactor CHO cell cultures. *Biotechnology Progress*, *34*(2), 486–493. https://doi.org/10.1002/btpr.2604

Li, W., Wu, H., Li, M. & San, K.-Y. (2018). Effect of NADPH availability on free fatty acid production in *Escherichia coli*. *Biotechnology and Bioengineering*, *115*(2), 444–452. https://doi.org/10.1002/bit.26464

Liland, K. H., Mevik, B.-H., Wehrens, R. & Hiemstra, P. (2021). Partial Least Squares and Principal Component Regression. https://cran.r-project.org/package=pls

Lin, H. Y., Mathiszik, B., Xu, B., Enfors, S. O. & Neubauer, P. (2001). Determination of the maximum specific uptake capacities for glucose and oxygen in glucose-limited fed-batch cultivations of *Escherichia coli*. *Biotechnology and Bioengineering*, *73*(5), 347–357. https://doi.org/10.1002/BIT.1068

Lindinger, W. & Hansel, A. (1997). Analysis of trace gases at ppb levels by proton transfer reaction mass spectrometry. *Plasma Sources Science and Technology*, *6*, 111–117.

Link, H., Fuhrer, T., Gerosa, L., Zamboni, N. & Sauer, U. (2015). Real-time metabolome profiling of the metabolic switch between starvation and growth. *Nature Methods*, *12*(11), 1091–1097. https://doi.org/10.1038/nmeth.3584

Liquori, A. M., Monroy, A., Parisi, E. & Tripiciano, A. (1981). A Theoretical Equation for Diauxic Growth and Its Application to the Kinetics of the Early Development of the Sea Urchin Embryo. *Differentiation*, *20*(1-3), 174–175. https://doi.org/10.1111/J.1432-0436.1981.TB01173.X

Liu, M., Fu, Y., Gao, W., Xian, M. & Zhao, G. (2020). Highly Efficient Biosynthesis of Hypoxanthine in *Escherichia coli* and Transcriptome-Based Analysis of the Purine Metabolism. *ACS Synthetic Biology*, *9*(3), 525–535. https://doi.org/10.1021/ACSSYNBIO.9B00396

Liu, Y.-C., Wang, F.-S. & Lee, W.-C. (2001). On-line monitoring and controlling system for fermentation processes. *Biochemical Engineering Journal*, *7*(1), 17–25. https://doi-org.ezproxy.is.ed.ac.uk/10.1016/S1369-703X(00)00100-5

Liu, Y., Wu, H., Li, Q., Tang, X., Li, Z. & Ye, Q. (2011). Process development of succinic acid production by *Escherichia coli* NZN111 using acetate as an aerobic carbon source. *Enzyme and Microbial Technology*, *49*(5), 459–464. https://doi.org/10.1016/j.enzmictec.2011.07.006

Lombardino, J. G. (2000). A brief history of Pfizer central research. *Bulletin for the History of Chemistry*, *25*(1), 10–15.

Lu, R. M., Hwang, Y. C., Liu, I. J., Lee, C. C., Tsai, H. Z., Li, H. J. & Wu, H. C. (2020). Development of therapeutic antibodies for the treatment of diseases. *Journal of Biomedical Science*, *27*(1), 1–30. https://doi.org/10.1186/S12929-019-0592-Z

Lu, S., Eiteman, M. A. & Altman, E. (2009). Effect of CO2 on succinate production in dual-phase *Escherichia coli* fermentations. *Journal of Biotechnology*, *143*(3), 213–223. https://doi.org/10.1016/j.jbiotec.2009.07.012

Luchner, M., Gutmann, R., Bayer, K., Dunkl, J., Hansel, A., Herbig, J., Singer, W., Strobl, F., Winkler, K. & Striedner, G. (2012). Implementation of proton transfer reaction-mass spectrometry (PTR-MS) for advanced bioprocess monitoring. *Biotechnology and Bioengineering*. https://doi.org/10.1002/bit.24579

Luedeking, R. & Piret, E. L. (1959). A kinetic study of the lactic acid fermentation. Batch process at controlled pH. *Journal of Biochemical and Microbiological Technology and Engineering*, *1*(4), 393–412. https://doi.org/10.1002/JBMTE.390010406

Majors, P. D., McLean, J. S. & Scholten, J. C. M. (2008). NMR bioreactor development for live in-situ microbial functional analysis. *Journal of Magnetic Resonance*, *192*, 159–166. https://doi.org/10.1016/j.jmr.2008.02.014

Makarov, A. (2000). Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis. *Analytical Chemistry*, *72*(6), 1156–1162. https://doi.org/10.1021/ac991131p

Mamyrin, B. A., Karataev, V. I., Shmikk, D. V. & Zagulin, V. A. (1973). The mass-reflectron, a new nonmagnetic time-of-flight mass spectrometer with high resolution. *Journal of Experimental and Theoretical Physics*, *37*(1), 45–48.

Mandenius, C.-f. & Titchener-Hooker, N. J. (2013). *Measurement, Monitoring, Modelling and Control of Bioprocesses* (Vol. 132). Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-642-36838-7

Mariadasse, R., Biswal, J., Jayaprakash, P., Rao, G. R., Choubey, S. K., Rajendran, S. & Jeyakanthan, J. (2016). Mechanical insights of oxythiamine compound as potent inhibitor for human transketolase-like protein 1 (TKTL1 protein). *Journal of Receptors and Signal Transduction*, *36*(3), 233–242. https://doi.org/10.3109/10799893.2015.1080272

Marison, I., Hennessy, S., Foley, R., Schuler, M., Sivaprakasam, S., Freeland, B., Marison, I., Hennessy, Á. S., Foley, Á. R., Schuler, Á. M., Sivaprakasam, Á. S., Freeland, Á. B., Biochem, A. & Biotechnol, E. (2012). The Choice of Suitable Online Analytical Techniques and Data Processing for Monitoring of Bioprocesses. *Advances in biochemical engineering/biotechnology* (pp. 249–280). Springer, Berlin, Heidelberg. https://doi.org/10.1007/10_2012_175

MarketsandMarkets™. (2018). *Succinic Acid Market Worth $183 Million by 2023* (tech. rep.). New York. https://www.marketsandmarkets.com/Market-Reports/succinic-acid-market-402.html

Markx, G. H. & Davey, C. L. (1999). The dielectric properties of biological cells at radiofrequencies: applications in biotechnology. *Enzyme and Microbial Technology*, *25*(3-5), 161–171. https://doi.org/10.1016/S0141-0229(99)00008-3

Marose, S., Lindemann, C. & Scheper, T. (1998). Two-Dimensional Fluorescence Spectroscopy: A New Tool for On-Line Bioprocess Monitoring. *Biotechnology Progress*, *14*(1), 63–74. https://doi.org/10.1021/BP970124O

Martinez Basterrechea, I. (2010). *Metabolic Engineering of the Flow of Reducing Equivalents for the Production of Biochemicals in Escherichia coli* (Doctoral dissertation). Rice University.

Maskow, T. & Paufler, S. (2015). What does calorimetry and thermodynamics of living cells tell us? *Methods*, *76*, 3–10. https://doi.org/10.1016/J.YMETH.2014.10.035

Matano, C., Meiswinkel, T. M. & Wendisch, V. F. (2014). Amino Acid Production from Rice Straw Hydrolyzates. *Wheat and rice in disease prevention and health* (pp. 493–505). Elsevier Inc. https://doi.org/10.1016/B978-0-12-401716-0.00038-6

May, O. (2019). Industrial Enzyme Applications – Overview and Historic Perspective. In A. Vogel & O. May (Eds.), *Industrial enzyme applications* (pp. 1–24). John Wiley & Sons, Ltd. https://doi.org/10.1002/9783527813780.CH1_1

McCoy, M. (2019). Succinic acid, once a biobased chemical star, is barely being made. *C&EN Chemical & Engineering news*, *97*(12), 15–15. https://doi.org/10.1021/cen-09712-buscon5

McGovern, A. C., Broadhurst, D., Taylor, J., Kaderbhai, N., Winson, M. K., Small, D. A., Rowland, J. J., Kell, D. B. & Goodacre, R. (2002). Monitoring of complex industrial bioprocesses for metabolite concentrations using modern spectroscopies and machine learning: Application to gibberellic acid production. *Biotechnology and Bioengineering*, *78*(5), 527–538. https://doi.org/10.1002/bit.10226

McGovern, P. E. (2019). Wine of the World's First Cities. *Ancient wine* (pp. 148–166). Princeton University Press. https://doi.org/10.1515/9780691198965-010

McGovern, P. E., Zhang, J., Tang, J., Zhang, Z., Hall, G. R., Moreau, R. A., Nuñez, A., Butrym, E. D., Richards, M. P., Wang, C. S., Cheng, G., Zhao, Z. & Wang, C. (2004). Fermented beverages of pre- and proto-historic China. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(51), 17593–17598. https://doi.org/10.1073/PNAS.0407921102

McKinnon, K. M. (2018). Flow Cytometry: An Overview. *Current protocols in immunology*, *120*, 5.1.1–5.1.11. https://doi.org/10.1002/CPIM.40

McLuskey, K., Wandy, J., Vincent, I., van der Hooft, J. J. J., Rogers, S., Burgess, K. & Daly, R. (2021). Ranking Metabolite Sets by Their Activity Levels. *Metabolites*, *11*(2), 1–15. https://doi.org/10.3390/METABO11020103

Meng, L. M. & Nygaard, P. (1990). Identification of hypoxanthine and guanine as the co-repressors for the purine regulon genes of *Escherichia coli*. *Molecular Microbiology*, *4*(12), 2187–2192. https://doi.org/10.1111/J.1365-2958.1990.TB00580.X

Michaelis, L. & Menten, M. M. L. (2013). The kinetics of invertin action: Translated by T.R.C. Boyde Submitted 4 February 1913. *FEBS Letters*, *587*(17), 2712–2720. https://doi.org/10.1016/J.FEBSLET.2013.07.015

Miller, J. C. & Miller, J. (2005). *Statistics and Chemometrics for Analytical Chemistry*. Pearson Education UK.

Monod, J. (1949). The growth of bacterial cultures. *Annual Review of Microbiology*, *3*(1), 371–394. https://doi.org/10.1146/ANNUREV.MI.03.100149.002103

Montgomery, D. C. (2012). *Design and analysis of experiments* (John Wiley & Sons, Ed.; 8th). https://doi.org/10.1002/9783527809080.cataz11063

Müller, M., Meusel, W., Husemann, U., Greller, G. & Kraume, M. (2018). Application of heat compensation calorimetry to an *E. coli* fed-batch process. *Journal of Biotechnology*, *266*, 133–143. https://doi.org/10.1016/J.JBIOTEC.2017.12.002

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. & Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor symposia on quantitative biology*, *51 Pt 1*(1), 263–273. https://doi.org/10.1101/SQB.1986.051.01.032

Mulvaney, R. (1993). Mass Spectrometry. *Nitrogen isotope techniques* (pp. 11–57). Academic Press. https://doi.org/10.1016/B978-0-08-092407-6.50007-9

Nengsih, T. A., Bertrand, F., Maumy-Bertrand, M. & Meyer, N. (2019). Determining the number of components in PLS regression on incomplete data set. *Statistical Applications in Genetics and Molecular Biology*, *18*(6), 1–28. https://doi.org/10.1515/SAGMB-2018-0059

Nghiem, N. P., Kleff, S. & Schwegmann, S. (2017). Succinic acid: Technology development and commercialization. *Fermentation*, *3*(26). https://doi.org/10.3390/fermentation3020026

Nordberg, J. & Arnér, E. S. (2001). Reactive oxygen species, antioxidants, and the mammalian thioredoxin system. *Free Radical Biology and Medicine*, *31*(11), 1287–1312. https://doi.org/10.1016/S0891-5849(01)00724-9

Nuoffer, C., Zanolari, B. & Erni, B. (1988). Glucose permease of *Escherichia coli*. The effect of cysteine to serine mutations on the function, stability, and regulation of transport and phosphorylation. *Journal of Biological Chemistry*, *263*(14), 6647–6655. https://doi.org/10.1016/S0021-9258(18)68691-0

Oeggerlit, A. & Heinzle, E. (1994). On-Line Exhaust Gas Analysis of Volatiles in Fermentation Using Mass Spectrometry. *Biotechnology Progress*, *10*, 284–290. https://doi.org/10.1016/S1474-6670(17)50375-4

Oliver, S. G., Winson, M. K., Kell, D. B. & Baganz, F. (1998). Systematic functional analysis of the yeast genome. *Trends in Biotechnology*, *16*(9), 373–378. https://doi.org/10.1016/S0167-7799(98)01214-1

Paalme, T., Elken, R., Kahru, A., Vanatalu, K. & Vilu, R. (1997). The growth rate control in *Escherichia coli* at near to maximum growth rates: the A-stat approach. *Antonie van Leeuwenhoek*, *71*(3), 217–230. https://doi.org/10.1023/A:1000198404007

Pandi, K., Chauhan, A. S., Khan, W. H. & Rathore, A. S. (2020). Phosphate starvation controls lactose metabolism to produce recombinant protein in *Escherichia coli*. *Applied Microbiology and Biotechnology*, *104*(22), 9707–9718. https://doi.org/10.1007/S00253-020-10935-Y

Pang, Z., Chong, J., Zhou, G., De Lima Morais, D. A., Chang, L., Barrette, M., Gauthier, C., Jacques, P. É., Li, S. & Xia, J. (2021). MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. *Nucleic Acids Research*, *49*(W1), W388–W396. https://doi.org/10.1093/NAR/GKAB382

Pasteur, L. (1995). Mémoire sur la fermentation appelée lactique (Extrait par l'auteur). *Molecular Medicine*, *1*(6), 601. https://doi-org.ezproxy.is.ed.ac.uk/10.1007/BF03401599

Paul, W. & Raether, M. (1955). Das elektrische Massenfilter. *Zeitschrift für Physik*, *140*(3), 262–273. https://doi.org/10.1007/BF01328923

Pembrey, R. S., Marshall, K. C. & Schneider, R. P. (1999). Cell surface analysis techniques: What do cell preparation protocols do to cell surface properties? *Applied and Environmental Microbiology*, *65*(7), 2877–2894. https://doi.org/10.1128/AEM.65.7.2877-2894.1999

Peralta-Yahya, P. P., Zhang, F., del Cardayre, S. B. & Keasling, J. D. (2012). Microbial engineering for the production of advanced biofuels. *Nature*, *488*(7411), 320–328. https://doi.org/10.1038/nature11478

Pérez-Enciso, M. & Tenenhaus, M. (2003). Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. *Human Genetics*, *112*(5), 581–592. https://doi.org/10.1007/S00439-003-0921-9

Pičmanová, M., Moses, T., Cortada-Garcia, J., Barrett, G., Florance, H., Pandor, S. & Burgess, K. (2022). Rapid HILIC-Z ion mobility mass spectrometry (RHIMMS) method for untargeted metabolomics of complex biological samples. *Metabolomics*, *18*(16), 1–12. https://doi.org/10.1007/S11306-022-01871-1

Pisitkun, T., Hoffert, J. D., Yu, M.-J. & Knepper, M. A. (2007). Tandem Mass Spectrometry in Physiology. *Physiology*, *22*(6), 390–400. https://doi.org/10.1152/physiol.00025.2007

Plekhova, V., Van Meulebroek, L., De Graeve, M., Perdones-Montero, A., De Spiegeleer, M., De Paepe, E., Van de Walle, E., Takats, Z., Cameron, S. J. & Vanhaecke, L. (2021). Rapid ex vivo molecular fingerprinting of biofluids using laser-assisted rapid evaporative ionization mass spectrometry. *Nature Protocols*, *16*(9), 4327–4354. https://doi.org/10.1038/s41596-021-00580-8

Plum, A. & Rehorek, A. (2005). Strategies for continuous on-line high performance liquid chromatography coupled with diode array detection and electrospray tandem mass spectrometry for process monitoring of sulphonated azo dyes and their intermediates in anaerobic-aerobic bioreactors. *Journal of Chromatography A*, *1084*(1-2), 119–133. https://doi.org/10.1016/j.chroma.2005.03.001

Prince, J. T. & Marcotte, E. M. (2006). Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. *Analytical Chemistry*, *78*(17), 6140–6152. https://doi.org/10.1021/ac0605344

Pu, Y. Y., O'Donnell, C., Tobin, J. T. & O'Shea, N. (2020). Review of near-infrared spectroscopy as a process analytical technology for real-time product monitoring in dairy processing. *International Dairy Journal*, *103*, 1–11. https://doi.org/10.1016/J.IDAIRYJ.2019.104623

Raïs, B., Comin, B., Puigjaner, J., Brandes, J. L., Creppy, E., Saboureau, D., Ennamany, R., Paul Lee, W. N., Boros, L. G. & Cascante, M. (1999). Oxythiamine and dehydroepiandrosterone induce a G1 phase cycle arrest in Ehrlich's tumor cells through inhibition of the pentose cycle. *FEBS Letters*, *456*(1), 113–118. https://doi.org/10.1016/S0014-5793(99)00924-2

Ranke, J. (2021). Calibration Functions for Analytical Chemistry. https://cran.r-project.org/web/packages/chemCal/index.html

Redenbaugh, K., Hiatt, W., Martineau, B. & Emlay, D. (1994). Regulatory assessment of the FLAVR SAVR tomato. *Trends in Food Science & Technology*, *5*(4), 105–110. https://doi.org/10.1016/0924-2244(94)90197-X

Rehorek, A., Urbig, K., Meurer, R., Schafer, C., Plum, A. & Braun, G. (2002). Monitoring of azo dye degradation processes in a bioreactor by on-line high-performance liquid chromatography. *Journal of Chromatography A*, *949*(1-2), 263–268. https://doi-org.ezproxy.is.ed.ac.uk/10.1016/S0021-9673(01)01427-3

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C., Merico, D. & Bader, G. D. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protocols 2019 14:2*, *14*(2), 482–517. https://doi.org/10.1038/s41596-018-0103-9

Rende, U., Niittylä, T. & Moritz, T. (2019). Two-step derivatization for determination of sugar phosphates in plants by combined reversed phase chromatography/tandem mass spectrometry. *Plant Methods*, *15*(1), 1–10. https://doi.org/10.1186/s13007-019-0514-9

Revuelta, J. L., Ledesma-Amaro, R., Lozano-Martinez, P., Díaz-Fernández, D., Buey, R. M. & Jiménez, A. (2017). Bioproduction of riboflavin: a bright yellow history. *Journal of Industrial Microbiology and Biotechnology*, *44*(4-5), 659–665. https://doi.org/10.1007/S10295-016-1842-7

Rieger, M., Kappeli, O. & Fiechter, A. (1983). The role of limited respiration in the incomplete oxidation of glucose by *Saccharomyces cerevisiae*. *Journal of General Microbiology*, *129*(3), 653–661. https://doi.org/10.1099/00221287-129-3-653

Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology*, *26*(3), 303–304. https://doi.org/10.1038/nbt0308-303

Rittmann, B. E. & McCarty, P. L. (2020). Moving Toward Sustainability. *Environmental biotechnology: Principles and application* (pp. 9–49). McGraw-Hill Education. https://www-accessengineeringlibrary-com.ezproxy.is.ed.ac.uk/content/book/9781260441604/chapter/chapter1

Rodrigues, K. C., Sonego, J. L., Bernardo, A., Ribeiro, M. P., Cruz, A. J. & Badino, A. C. (2018). Real-Time Monitoring of Bioethanol Fermentation with Industrial Musts Using Mid-Infrared Spectroscopy. *Industrial and Engineering Chemistry Research*, *57*(32), 10823–10831. https://doi.org/10.1021/acs.iecr.8b01181

Roe, A. J., McLaggan, D., Davidson, I., O'Byrne, C. & Booth, I. R. (1998). Perturbation of anion balance during inhibition of growth of *Escherichia coli* by weak acids. *Journal of Bacteriology*, *180*(4), 767–772. https://doi.org/10.1128/JB.180.4.767-772.1998

Roels, J. A. (1980). Application of macroscopic principles to microbial metabolism. *Biotechnology and Bioengineering*, *22*(12), 2457–2514. https://doi.org/10.1002/BIT.260221202

Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmond, A. & Jent, N. (2007). A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *Journal of Pharmaceutical and Biomedical Analysis*, *44*(3), 683–700. https://doi.org/10.1016/J.JPBA.2007.03.023

Russell, R. (1976). Metabolism of inorganic pyrophosphate (PPi). *Arthritis and Rheumatism*, *19*(3), 465–478. https://doi.org/10.1002/1529-0131(197605/06)19:3+%7B%5C%%7D3C465::aid-art1780190722%7B%5C%%7D3E3.0.co;2-%7B%5C#%7D

Sadino-Riquelme, M. C., Rivas, J., Jeison, D., Hayes, R. E. & Donoso-Bravo, A. (2020). Making sense of parameter estimation and model simulation in bioprocesses. *Biotechnology and Bioengineering*, *117*(5), 1357–1366. https://doi.org/10.1002/BIT.27294

Salvy, P. & Hatzimanikatis, V. (2021). Emergence of diauxie as an optimal growth strategy under resource allocation constraints in cellular metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(8), 1–11. https://doi.org/10.1073/PNAS.2013836118

Santos, L. O., Dewasme, L., Coutinho, D. & Wouwer, A. V. (2012). Nonlinear model predictive control of fed-batch cultures of micro-organisms exhibiting overflow metabolism: Assessment and robustness. *Computers & Chemical Engineering*, *39*, 143–151. https://doi.org/10.1016/J.COMPCHEMENG.2011.12.010

Sarvin, B., Lagziel, S., Sarvin, N., Mukha, D., Kumar, P., Aizenshtein, E. & Shlomi, T. (2020). Fast and sensitive flow-injection mass spectrometry metabolomics by analyzing sample-specific ion distributions. *Nature Communications*, *11*(1), 1–11. https://doi.org/10.1038/s41467-020-17026-6

Sauer, U., Canonaco, F., Heri, S., Perrenoud, A. & Fischer, E. (2004). The soluble and membrane-bound transhydrogenases UdhA and PntAB have divergent functions in NADPH metabolism of *Escherichia coli*. *The Journal of biological chemistry*, *279*(8), 6613–6619. https://doi.org/10.1074/JBC.M311657200

Saxena, R. K., Saran, S., Isar, J. & Kaushik, R. (2017). Production and Applications of Succinic Acid. *Current developments in biotechnology and bioengineering: Production, isolation and purification of industrial products* (pp. 601–630). Elsevier Inc. https://doi.org/10.1016/B978-0-444-63662-1.00027-0

Schastnaya, E., Raguz Nakic, Z., Gruber, C. H., Doubleday, P. F., Krishnan, A., Johns, N. I., Park, J., Wang, H. H. & Sauer, U. (2021). Extensive regulation of enzyme activity by phosphorylation in *Escherichia coli*. *Nature Communications*, *12*(1), 1–11. https://doi.org/10.1038/s41467-021-25988-4

Schenk, J., Viscasillas, C., Marison, I. W. & von Stockar, U. (2008). On-line monitoring of nine different batch cultures of *E. coli* by mid-infrared spectroscopy, using a single spectra library for calibration. *Journal of Biotechnology*, *134*(1-2), 93–102. https://doi.org/10.1016/J.JBIOTEC.2007.12.014

Schrimpe-Rutledge, A. C., Codreanu, S. G., Sherrod, S. D. & McLean, J. A. (2016). Untargeted Metabolomics Strategies—Challenges and Emerging Directions. *Journal of the American Society for Mass Spectrometry*, *27*(12), 1897–1905. https://doi.org/10.1007/S13361-016-1469-Y

Schubert, T., Breuer, U., Harms, H. & Maskow, T. (2007). Calorimetric bioprocess monitoring by small modifications to a standard bench-scale bioreactor. *Journal of Biotechnology*, *130*(1), 24–31. https://doi.org/10.1016/j.jbiotec.2007.02.013

Schürrle, K. (2018). History, Current State, and Emerging Applications of Industrial Biotechnology. In Fröhling Magnus & Hiete Michael (Eds.), *Sustainability and life cycle assessment in industrial biotechnology. advances in biochemical engineering/biotechnology* (pp. 13–51). Springer, Cham. https://doi.org/10.1007/10_2018_81

Schüürmann, G., Ebert, R. U., Chen, J., Wang, B. & Kühne, R. (2008). External Validation and Prediction Employing the Predictive Squared Correlation Coefficient — Test Set Activity Mean vs Training Set Activity Mean. *Journal of Chemical Information and Modeling*, *48*(11), 2140–2145. https://doi.org/10.1021/CI800253U

Schwann, T., Smith, H. & Schleiden, M. J. (1847). *Microscopical researches into the accordance in the structure and growth of animals and plants.* The Sydenham Society. https://doi.org/10.5962/BHL.TITLE.11431

Scigelova, M., Hornshaw, M., Giannakopulos, A. & Makarov, A. (2011). Fourier Transform Mass Spectrometry. *Molecular & Cellular Proteomics*, *10*(7), 1–19. https://doi.org/10.1074/MCP.M111.009431

Shalabaeva, V., Lovato, L., La Rocca, R., Messina, G. C., Dipalo, M., Miele, E., Perrone, M., Gentile, F. & De Angelis, F. (2017). Time resolved and label free monitoring of extracellular metabolites by surface enhanced Raman spectroscopy. *PLoS ONE*, *12*(4), 1–16. https://doi.org/10.1371/journal.pone.0175581

Siano, S. A. & Mutharasan, R. (1989). NADH and flavin fluorescence responses of starved yeast cultures to substrate additions. *Biotechnology and Bioengineering*, *34*(5), 660–670. https://doi.org/10.1002/BIT.260340510

Silva, A. C., Nuno Moreira, J., Sousa Lobo, J. M. & Almeida, H. (2020). *Current Applications of Pharmaceutical Biotechnology* (A. Catarina Silva, J. Nuno Moreira, J. M. Sousa Lobo & H. Almeida, Eds.; 1st ed., Vol. 171). Springer International Publishing. https://doi.org/10.1007/978-3-030-40464-2

Singh, A., Cher Soh, K., Hatzimanikatis, V. & Gill, R. T. (2011). Manipulating redox and ATP balancing for improved production of succinate in *E. coli*. *Metabolic Engineering*, *13*(1), 76–81. https://doi.org/10.1016/j.ymben.2010.10.006

Singh, A., Lynch, M. D. & Gill, R. T. (2009). Genes restoring redox balance in fermentation-deficient *E. coli* NZN111. *Metabolic Engineering*, *11*(6), 347–354. https://doi.org/10.1016/j.ymben.2009.07.002

Six, S., Andrews, S. C., Unden, G. & Guest, J. R. (1994). *Escherichia coli* possesses two homologous anaerobic c4-dicarboxylate membrane transporters (dcua and dcub) distinct from the aerobic dicarboxylate transport system (dct). *Journal of Bacteriology*, *176*(21), 6470–6478. https://doi.org/10.1128/jb.176.21.6470-6478.1994

Solomon, B. D., Barnes, J. R. & Halvorsen, K. E. (2007). Grain and cellulosic ethanol: History, economics, and energy policy. *Biomass and Bioenergy*, *31*(6), 416–425. https://doi.org/10.1016/J.BIOMBIOE.2007.01.023

Sonnleitner, B. & Käppeli, O. (1986). Growth of *Saccharomyces cerevisiae* is controlled by its limited respiratory capacity: Formulation and verification of a hypothesis. *Biotechnology and Bioengineering*, *28*(6), 927–937. https://doi.org/10.1002/BIT.260280620

Sonnleitner, B. (2012). Automated Measurement and Monitoring of Bioprocesses: Key Elements of the M3C Strategy. *Advances in biochemical engineering/biotechnology* (pp. 1–33). Springer, Berlin, Heidelberg. https://doi.org/10.1007/10_2012_173

Spekreijse, J., Lammens, T., Parisi, C., Ronzon, T. & Martijn, V. (2019). *Insights into the European market for bio-based chemicals* (tech. rep.). European Commission. https://doi.org/10.2760/673071

Spéranido, M. & Paul, E. (1997). Determination of Carbon Dioxide Evolution Rate Using On-Line Gas Analysis During Dynamic Biodegradation Experiments. *Biotechnology and Bioengineering*, *53*(3), 243–252. https://doi.org/10.1002/(SICI)1097-0290(19970205)53:3

Spiro, S. & Guest, J. R. (1991). Adaptive responses to oxygen limitation in *Escherichia coli*. *Trends in Biochemical Sciences*, *16*(100), 310–314. https://doi.org/10.1016/0968-0004(91)90125-F

Stanbury, P. F., Whitaker, A. & Hall, S. J. (2017). Instrumentation and control. *Principles of fermentation technology* (pp. 487–536). Elsevier. https://doi.org/10.1016/B978-0-08-099953-1.00008-9

Stretton, A. O. (2002). The first sequence. Fred Sanger and insulin. *Genetics*, *162*(2), 527–532. https://doi.org/10.1093/genetics/162.2.527

Stuart, B. H. (2005). *Infrared Spectroscopy: Fundamentals and Applications* (D. J. A. Ando, Ed.). wiley. https://doi.org/10.1002/0470011149

Suitor, J. T., Varzandeh, S. & Wallace, S. (2020). One-Pot Synthesis of Adipic Acid from Guaiacol in *Escherichia coli*. *ACS Synthetic Biology*, *9*(9), 2472–2476. https://doi.org/10.1021/ACSSYNBIO.0C00254

Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W.-M., Fiehn, O., Goodacre, R., Griffin, J. L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A. N., Lindon, J. C., Marriott, P., Nicholls, A. W., ... Viant, M. R. (2007). Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics : Official journal of the Metabolomic Society*, *3*(3), 211–221. https://doi.org/10.1007/S11306-007-0082-2

Surribas, A., Geissler, D., Gierse, A., Scheper, T., Hitzmann, B., Montesinos, J. L. & Valero, F. (2006). State variables monitoring by in situ multi-wavelength fluorescence spectroscopy in heterologous protein production by *Pichia pastoris*. *Journal of Biotechnology*, *124*(2), 412–419. https://doi.org/10.1016/J.JBIOTEC.2006.01.002

Surribas, A., Montesinos, J. L. & Valero, F. F. (2006). Biomass estimation using fluorescence measurements in *Pichia pastoris* bioprocess. *Journal of Chemical Technology & Biotechnology*, *81*(1), 23–28. https://doi.org/10.1002/JCTB.1352

Svendsen, C., Skov, T. & van den Berg, F. W. J. (2015). Monitoring fermentation processes using in-process measurements of different orders. *Journal of Chemical Technology & Biotechnology*, *90*(2), 244–254. https://doi.org/10.1002/jctb.4483

Switzer, A., Burchell, L., McQuail, J. & Wigneshweraraj, S. (2020). The Adaptive Response to Long-Term Nitrogen Starvation in *Escherichia coli* Requires the Breakdown of Allantoin. *Journal of Bacteriology*, *202*(17), 1–11. https://doi.org/10.1128/JB.00172-20

Synoground, B. F., McGraw, C. E., Elliott, K. S., Leuze, C., Roth, J. R., Harcum, S. W. & Sandoval, N. R. (2021). Transient ammonia stress on Chinese hamster ovary (CHO) cells yield alterations to alanine metabolism and IgG glycosylation profiles. *Biotechnology Journal*, *16*(7), 2100098. https://doi.org/10.1002/BIOT.202100098

Szymańska, E., Saccenti, E., Smilde, A. K. & Westerhuis, J. A. (2012). Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics*, *8*(1), 3–16. https://doi.org/10.1007/S11306-011-0330-3

Takahashi, S., Miyachi, M., Tamaki, H. & Suzuki, H. (2021). The *Escherichia coli* CitT transporter can be used as a succinate exporter for succinate production. *Bioscience, biotechnology, and biochemistry*, *85*(4), 981–988. https://doi.org/10.1093/bbb/zbaa109

Tamis, J., Mulders, M., Dijkman, H., Rozendal, R., van Loosdrecht, M. C. M. & Kleerebezem, R. (2018). Pilot-Scale Polyhydroxyalkanoate Production from Paper Mill Wastewater: Process Characteristics and Identification of Bottlenecks for Full-Scale Implementation. *Journal of Environmental Engineering*, *144*(10). https://doi.org/10.1061/(ASCE)EE.1943-7870.0001444

Tan, Z., Chen, J. & Zhang, X. (2016). Systematic engineering of pentose phosphate pathway improves *Escherichia coli* succinate production. *Biotechnology for Biofuels*, *9*(1), 262–274. https://doi.org/10.1186/s13068-016-0675-y

Teixeira, A. P., Alves, C., Alves, P. M., Carrondo, M. J. & Oliveira, R. (2007). Hybrid elementary flux analysis/nonparametric modeling: Application for bioprocess control. *BMC Bioinformatics*, *8*(1), 1–15. https://doi.org/10.1186/1471-2105-8-30

Tejero Rioseras, A., Garcia Gomez, D., Ebert, B. E., Blank, L. M., Ibáñez, A. J. & Sinues, P. M. (2017). Comprehensive Real-Time Analysis of the Yeast Volatilome. *Scientific Reports*, *7*, 1–9. https://doi.org/10.1038/s41598-017-14554-y

Teleki, A., Sánchez-Kopper, A. & Takors, R. (2015). Alkaline conditions in hydrophilic interaction liquid chromatography for intracellular metabolite quantification using tandem mass spectrometry. *Analytical Biochemistry*, *475*, 4–13. https://doi.org/10.1016/j.ab.2015.01.002

Thakker, C., Martínez, I., San, K. Y. & Bennett, G. N. (2012). Succinate production in *Escherichia coli*. *Biotechnology Journal*, *7*(2), 213–224. https://doi.org/10.1002/biot.201100061

Thermo Fisher Scientific. (2017). Orbitrap LC-MS Comparison Chart. Retrieved March 31, 2018, from https://www.thermofisher.com/ng/en/home/industrial/mass-spectrometry/liquid-chromatography-mass-spectrometry-lc-ms/lc-ms-systems/orbitrap-lc-ms/orbitrap-lc-ms-comparison-chart.html

Tohmola, N., Ahtinen, J., Pitkänen, J. P., Parviainen, V., Joenväärä, S., Hautamäki, M., Lindroos, P., Mäkinen, J. & Renkonen, R. (2011). On-line high performance liquid chromatography measurements of extracellular metabolites in an aerobic batch yeast (*Saccharomyces cerevisiae*) culture. *Biotechnology and Bioprocess Engineering*, *16*(2), 264–272. https://doi.org/10.1007/S12257-010-0147-3

Triba, M. N., Le Moyec, L., Amathieu, R., Goossens, C., Bouchemal, N., Nahon, P., Rutledge, D. N. & Savarin, P. (2015). PLS/OPLS models in metabolomics: the impact of permutation of dataset rows on the K-fold cross-validation quality parameters. *Molecular BioSystems*, *11*(1), 13–19. https://doi.org/10.1039/C4MB00414K

Vaidyanathan, S., Kell, D. B. & Goodacre, R. (2002). Flow-injection electrospray ionization mass spectrometry of crude cell extracts for high-throughput bacterial identification. *Journal of the American Society for Mass Spectrometry*, *13*(2), 118–128. https://doi.org/10.1016/S1044-0305(01)00339-7

Van Poucke, C., Dumoulin, F. & Van Peteghem, C. (2005). Detection of banned antibacterial growth promoters in animal feed by liquid chromatography–tandem mass spectrometry: optimisation of the extraction solvent by experimental design. *Analytica Chimica Acta*, *529*(1-2), 211–220. https://doi.org/10.1016/J.ACA.2004.08.076

Vann, L., Layfield, J. B. & Sheppard, J. D. (2017). The application of near-infrared spectroscopy in beer fermentation for online monitoring of critical process parameters and their integration into a novel feedforward control strategy. *Journal of the Institute of Brewing*, *123*(3), 347–360. https://doi.org/10.1002/JIB.440

Vemuri, G. N., Eiteman, M. A. & Altman, E. (2002). Effects of growth mode and pyruvate carboxylase on succinic acid production by metabolically engineered strains of *Escherichia coli*. *Applied and Environmental Microbiology*, *68*(4), 1715–1727. https://doi.org/10.1128/AEM.68.4.1715-1727.2002

Villas-Bôas, S. G., Moxley, J. F., Åkesson, M., Stephanopoulos, G. & Nielsen, J. (2005). High-throughput metabolic state analysis: the missing link in integrated functional genomics of yeasts. *Biochemical Journal*, *388*(2), 669–677. https://doi.org/10.1042/BJ20041162

Võ, U.-U. T. & Morris, M. P. (2014). Nonvolatile, semivolatile, or volatile: Redefining volatile for volatile organic compounds. *Journal of the Air & Waste Management Association*, *64*(6), 661–669. https://doi.org/10.1080/10962247.2013.873746

Vogel, H. C. & Todaro, C. L. (2014). *Fermentation and biochemical engineering handbook : principles, process design, and equipment*. Elsevier Inc. https://www.sciencedirect.com/science/book/9781455725533

Von Stockar, U. & Liu, J. S. (1999). Does microbial life always feed on negative entropy? Thermodynamic analysis of microbial growth. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, *1412*(3), 191–211. https://doi.org/10.1016/S0005-2728(99)00065-1

von Stosch, M., Davy, S., Francois, K., Galvanauskas, V., Hamelink, J. M., Luebbert, A., Mayer, M., Oliveira, R., O'Kennedy, R., Rice, P. & Glassey, J. (2014). Hybrid modeling for quality by design and PAT-benefits and challenges of applications in biopharmaceutical industry. *Biotechnology Journal*, *9*(6), 719–726. https://doi.org/10.1002/BIOT.201300385

Wamelink, M. M. C., Struys, E. A., Huck, J. H. J., Roos, B., Van Der Knaap, M. S., Jakobs, C. & Verhoeven, N. M. (2005). Quantification of sugar phosphate intermediates of the pentose phosphate pathway by LC-MS/MS: application to two new inherited defects of metabolism. *Journal of Chromatography B*, *823*, 18–25. https://doi.org/10.1016/j.jchromb.2005.01.001

Wang, Y., Du, P., Gan, R., Li, Z. & Ye, Q. (2005). Fed-batch cultivation of *Escherichia coli* YK537 (pAET-8) for production of phoA promoter-controlled human epidermal growth factor. *Biotechnology and Bioprocess Engineering*, *10*(2), 149–154. https://doi.org/10.1007/BF02932585

Wang, Z. J., Wang, H. Y., Li, Y. L., Chu, J., Huang, M. Z., Zhuang, Y. P. & Zhang, S. L. (2010). Improved vitamin B12 production by step-wise reduction of oxygen uptake rate under dissolved oxygen limiting level during fermentation process. *Bioresource Technology*, *101*(8), 2845–2852. https://doi.org/10.1016/j.biortech.2009.10.048

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., Venables, B. & Galili, T. (2020). Package 'gplots' Title Various R Programming Tools for Plotting Data. https://github.com/talgalili/gplots

Warth, B., Rajkai, G. & Mandenius, C. F. (2010). Evaluation of software sensors for on-line estimation of culture conditions in an *Escherichia coli* cultivation expressing a recombinant protein. *Journal of Biotechnology*, *147*(1), 37–45. https://doi.org/10.1016/j.jbiotec.2010.02.023

Watson, J. D. & Crick, F. H. C. (1953). THE STRUCTURE OF DNA. *Cold Spring Harbor Symposia on Quantitative Biology*, *18*, 123–131. https://doi.org/10.1101/SQB.1953.018.01.020

Weastra. (2012). *Determination of market potential for selected platform chemicals* (tech. rep.). Weastra. www.bioconsept.eu

Wechselberger, P., Seifert, A. & Herwig, C. (2010). PAT method to gather bioprocess parameters in real-time using simple input variables and first principle relationships. *Chemical Engineering Science*, *65*(21), 5734–5746. https://doi.org/10.1016/J.CES.2010.05.002

Wei, R., Wang, J., Su, M., Jia, E., Chen, S., Chen, T. & Ni, Y. (2018). Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Scientific Reports*, *8*(1), 663–672. https://doi.org/10.1038/s41598-017-19120-0

Werpy, T. & Petersen, G. (2004). *Top Value Added Chemicals from Biomass Volume I-Results of Screening for Potential Candidates from Sugars and Synthesis Gas* (tech. rep.). United States. https://doi.org/10.2172/15008859

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. https://ggplot2.tidyverse.org

Wishart, D. S. (2016). Emerging applications of metabolomics in drug discovery and precision medicine. *Nature Reviews Drug Discovery*, *15*, 473–484. https://doi.org/10.2165/00126839-200809050-00002

Wold, S., Sjöström, M. & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

Wolfe, A. J. (2005). The Acetate Switch. *Microbiology and Molecular Biology Reviews*, *69*(1), 50. https://doi.org/10.1128/MMBR.69.1.12-50.2005

Wolff, M. M. & Stephens, W. E. (1953). A Pulsed Mass Spectrometer with Time Dispersion. *Review of Scientific Instruments*, *24*, 617. https://doi.org/10.1063/1.1770801

Wu, H., Li, Z. M., Zhou, L. & Ye, Q. (2007). Improved succinic acid production in the anaerobic culture of an *Escherichia coli* pflB ldhA double mutant as a result of enhanced anaplerotic activities in the preceding aerobic culture. *Applied and Environmental Microbiology*, *73*(24), 7837–7843. https://doi.org/10.1128/AEM.01546-07

Wu, L., Lange, H. C., Van Gulik, W. M. & Heijnen, J. J. (2003). Determination of in vivo oxygen uptake and carbon dioxide evolution rates from off-gas measurements under highly dynamic conditions. *Biotechnology and Bioengineering*, *81*(4), 448–458. https://doi.org/10.1002/bit.10480

Xi, H., Schneider, B. L. & Reitzer, L. (2000). Purine Catabolism in *Escherichia coli* and Function of Xanthine Dehydrogenase in Purine Salvage. *Journal of Bacteriology*, *182*(19), 5332. https://doi.org/10.1128/JB.182.19.5332-5341.2000

Xia, J., Psychogios, N., Young, N. & Wishart, D. S. (2009). MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Research*, *37*(suppl_2), W652–W660. https://doi.org/10.1093/NAR/GKP356

Xu, B., Jahic, M. & Enfors, S. O. (1999). Modeling of overflow metabolism in batch and fed-batch cultures of *Escherichia coli*. *Biotechnology Progress*, *15*(1), 81–90. https://doi.org/10.1021/bp9801087

Xu, Y. .-., Lu, W. & Rabinowitz, J. D. (2015). Avoiding misannotation of in-source fragmentation products as cellular metabolites in liquid chromatography-mass spectrometry-based metabolomics. *Analytical Chemistry*, *87*(4), 2273–2281. https://doi.org/10.1021/ac504118y

Yamashita, M. & Fenn, J. B. (2002). Negative ion production with the electrospray ion source. *Journal of Physical Chemistry*, *88*(20), 4671–4675. https://doi.org/10.1021/J150664A046

Yang, W. C., Sedlak, M., Regnier, F. E., Mosier, N., Ho, N. & Adamec, J. (2008). Simultaneous quantification of metabolites involved in central carbon and energy metabolism using reversed-phase liquid chromatography-mass spectrometry and in vitro 13C labeling. *Analytical Chemistry*, *80*(24), 9508–9516. https://doi.org/10.1021/ac801693c

Zhang, F., Masania, J., Anwar, A., Xue, M., Zehnder, D., Kanji, H., Rabbani, N. & Thornalley, P. J. (2016). The uremic toxin oxythiamine causes functional thiamine deficiency in end-stage renal disease by inhibiting transketolase activity. *Kidney International*, *90*, 396–403. https://doi.org/10.1016/j.kint.2016.03.010

Zhang, Y., Li, Z. & Ye, Q. (2010). Enhanced production of human epidermal growth factor under control of the phoA promoter by acetate-tolerant *Escherichia coli* DB15 in a chemically defined medium. *Biotechnology and Bioprocess Engineering*, *15*(4), 626–634. https://doi.org/10.1007/S12257-009-3049-5

Zhao, J. & Shimizu, K. (2003). Metabolic flux analysis of *Escherichia coli* K12 grown on 13 C-labeled acetate and glucose using GC-MS and powerful flux calculation method. *Journal of Biotechnology*, *101*, 101–117. https://doi.org/10.1016/S0168-1656(02)00316-4

Zhao, R., Natarajan, A. & Srienc, F. (1999). A Flow Injection Flow Cytometry System for On-Line Monitoring of Bioreactors. *Biotechnology and Bioengineering*, *62*(5), 609–617. https://doi.org/10.1002/(SICI)1097-0290(19990305)62:5

Zhu, X., Tan, Z., Xu, H., Chen, J., Tang, J. & Zhang, X. (2014). Metabolic evolution of two reducing equivalent-conserving pathways for high-yield succinate production in *Escherichia coli*. *Metabolic Engineering*, *24*, 87–96. https://doi.org/10.1016/j.ymben.2014.05.003

Zhuang, J., Gutknecht, R., Flükiger, K., Hasler, L., Erni, B. & Engel, A. (1999). Purification and Electron Microscopic Characterization of the Membrane Subunit (IICBGlc) of the *Escherichia coli* Glucose Transporter. *Archives of Biochemistry and Biophysics*, *372*(1), 89–96. https://doi.org/10.1006/ABBI.1999.1458

Zientz, E., Six, S. & Unden, G. (1996). Identification of a third secondary carrier (DcuC) for anaerobic C4-dicarboxylate transport in *Escherichia coli*: Roles of the three Dcu carriers in uptake and exchange. *Journal of Bacteriology*, *178*(24), 7241–7247. https://doi.org/10.1128/jb.178.24.7241-7247.1996

Zimmer, D. P., Soupene, E., Lee, H. L., Wendisch, V. F., Khodursky, A. B., Peter, B. J., Bender, R. A. & Kustu, S. (2000). Nitrogen regulatory protein C-controlled genes of *Escherichia coli*: Scavenging as a defense against nitrogen limitation. *Proceedings of the National Academy of Sciences*, *97*(26), 14674–14679. https://doi.org/10.1073/PNAS.97.26.14674

Zu, T. N., Liu, S., Gerlach, E. S., Germane, K. L., Servinsky, M. D., Mackie, D. M. & Sund, C. J. (2017). Real-time metabolite monitoring of glucose-fed *Clostridium acetobutylicum* fermentations using Raman assisted metabolomics. *Journal of Raman Spectroscopy*, *48*(12), 1852–1862. https://doi.org/10.1002/jrs.5264