University of Natural Resources
and Life Sciences, Vienna

Utilizing the full potential of dielectric spectroscopy as a PAT tool in cell culture fed-batch processes

MASTER THESIS

for the acquisition of the academic degree

Master of Science (MSc)/Diplom-Ingenieur (DI)

Submitted by

Veronika Költringer-Noppinger, BSc

Supervisor

External: DI Christoph Posch

Internal: Gerald Striedner, Ass. Prof. Dipl.-Ing. Dr.nat.techn.

Department of Biotechnology

University of Natural Resources and Life Sciences

Kufstein, June 2021

ABSTRACT

The production of active pharmaceutical ingredients (APIs) passed through major advancements over the last decades, since biopharmaceuticals, also called biologics, gained importance as therapeutic agents in modern age. By the aid of biological expression systems as for instance mammalian cells, insect cells, etc., biopharmaceuticals are manufactured at industrial scale. This calls for a necessity of characterizing and quantifying the manufacturing process and the product itself since these molecules are distinguished by structural and chemical complexity. Both, the U.S Food and Drug administration (FDA) and the European Medicines Agency (EMA) appealed to expand the knowledge in order to improve the process understanding, as well as the control of bioprocesses. The aim thereof is to produce more consistent and thus safer biopharmaceuticals. The modern approach to achieve these regulatory demands is to extend analytical and computational methods and furthermore create suitable process analytical technologies (PAT).

Within this thesis a strategy to develop and compare prediction models for important process parameters during the upstream production process (USP) by the combination of two different technologies is described. Investigations of online monitoring by dielectric capacitance probes and Raman spectroscopy were exploited. For dielectric capacitance measurements, basic models dealing with single frequency mode are well known and even in use. Using the same probes in frequency scan mode, the obtained information promises more capabilities. Thus, research for the utilization of the full potential of dielectric spectroscopy was done. Dielectric capacitance probes feature measurements at several frequencies, and therefore enable the construction of a spectrum that encourages the prediction of several parameters concerning the characterization of the physiology of cells and the cell density in a fermentation process. The aim of this thesis is to analyze whether the use of frequency scan mode is capable of predicting these cell characteristics more precisely and provides a guideline on how to develop suitable models for different PAT technologies facing a multivariate statistical approach.

Partial least squares (PLS) models based on dielectric spectroscopy and respectively simple linear regression models based on coefficients computed from a dielectric spectrum show promising predictions of cell density and viability, not only in the growth phase, but also in the late stage of a bioprocess. By adding Raman spectra to the predictive model, the results especially in the late

stage of the fermentation process could be considerably improved. Also, additional parameters characterizing the physiological state of cells in a culture like average diameter, average circularity, early and late apoptosis were elaborated, however those models lack adequate predictive power. The application of a dielectric capacitance probe was investigated in two different bioreactor scales, namely the 20L development scale and the 13kL large-scale production bioreactor.

ZUSAMMENFASSUNG

Die Produktion von aktiven pharmazeutischen Inhaltstoffen (API) erfuhr in den vergangenen Jahrzehnten einige Weiterentwicklungen, nicht zuletzt, weil Biopharmazeutika, auch „biologics“ genannt, in industriellem Maßstab hergestellt werden. Dadurch entstand die Notwendigkeit, den Herstellungsprozess und das Produkt selbst zu charakterisieren und zu quantifizieren, denn die produzierten Moleküle zeichnen sich durch strukturelle und chemische Komplexität aus. Sowohl die U.S Food and Drug Administration (FDA) und die European Medicines Agency (EMA) riefen dazu auf das Wissen rund um Prozessverständnis auszuweiten und auch erweiterte Prozesskontrolle zu erforschen. Alles in allem führen diese Maßnahmen dazu konstantere und dadurch sichere Biopharmazeutika zu produzieren. Der moderne Zugang, um diesen Anforderungen gerecht zu werden ist es, die analytischen und rechnerischen Methoden zu erweitern und dadurch geeignete Prozessanalytische Technologien zu entwickeln.

In dieser Arbeit wird eine Strategie beschreiben, um Vorhersagemodelle für wichtige Prozessparameter im Upstream Prozess (USP) durch eine Kombination zweier verschiedener Technologien zu entwickeln und zu vergleichen. Es werden zusätzliche Möglichkeiten auf dem Gebiet online Monitoring mit dielektrischen Kapazitätssonden ausgeschöpft. Einfache Modelle zur Messung mit dielektrischer Kapazität, welche sich auf die Signale einer einzelnen Frequenz beziehen, sind gut erforscht und bereits weit verbreitet. Durch die Verwendung der gleichen Sonden im „frequency scan“ Modus erhält man Daten, in welchen zusätzliche Informationen zu dem Prozess enthalten sind. Daher wird daran geforscht, das volle Potential der dielektrischen Spektroskopie auszuschöpfen. Dielektrische Kapazitätssonden ermöglichen Messungen bei verschiedenen Frequenzen und dadurch die Erstellung eines Spektrums, das die Vorhersage einiger Parameter zur Charakterisierung des physiologischen Zustands einer Zellkultur und der Zelldichte in einem Fermentationsprozess verspricht. Das Ziel dieser Masterarbeit ist es zu analysieren, ob die Verwendung des „frequency scan“ Modus die Vorhersage solcher Parameter ermöglicht und stellt einen Leitfaden zur Verfügung mithilfe eines multivariaten statistischen Zugangs geeignete Modelle zu entwickeln.

Partial least squares (PLS) Modelle basierend auf den Messdaten der dielektrischen Spektroskopie, beziehungsweise einfache lineare Regressionsmodelle basieren auf Koeffizienten, die vom dielektrischen Spektrum errechnet werden zeigen vielversprechende Vorhersagen der

Zelldichte und Viabilität, sowohl in der Wachstums- als auch in der späteren Phase eines Bioprozesses. Durch das hinzuziehen von Raman Spektren zu den Vorhersagemodellen kann das Ergebnis, vor allem in der späteren Phase des Bioprozesses, noch einmal gravierend verbessert werden. Zusätzlich wurde versucht weitere Parameter, die eine Aussage um den physiologischen Zustand einer Zellkultur geben, versucht vorherzusagen. Dazu zählt der durchschnittliche Zelldurchmesser, die durchschnittliche Rundheit und späte Apoptose. Diesen Modellen fehlt es allerdings an angemessener Vorhersagekraft. Die Anwendung von dielektrischen Kapazitätssonden wurde an zwei verschiedenen Reaktorgrößen erprobt, und zwar im 20L Entwicklungsmaßstab sowie im 13kL Produktionsmaßstab.

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources. This written work has not yet been submitted anywhere else.

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre eidesstattlich, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Diese schriftliche Arbeit wurde noch an keiner Stelle vorgelegt.

Kufstein, 01.06.2021 Veronika Költringer

Acknowledgements

This work was developed at Sandoz GmbH Austria in USP Development department in Schaftenau, Tirol.

First and foremost, my thanks got to my direct supervisor Christoph Posch, who faced me with patience, practical advice, and scientific input during the completion of this thesis. Besides that, I would like to thank Josef Stettner, who ultimately provided me the opportunity to join his team in midst of the special times of the global COVID19 pandemics and further the whole team for their support despite of social distancing and home office throughout the timespan of my internship.

I would like to take this opportunity to thank the complete team from the working group of Microbial Fermentation at BOKU Vienna, guided by Gerald Striedner, which has defined my time at the BOKU, as I had the opportunity of assisting several projects and learning from experienced biotechnologists.

To conclude, last but not least, I want to thank my family and my closest friends and my fellow students, who became friends as well, for their emotional support throughout my studies. Also, I want to thank the lovely people I met in Kufstein, who companioned me during the creation of this master thesis, for having pleasant times.

Table of Contents

1 Introduction

1.1 Biopharmaceuticals and the development of Biosimilars

Biopharmaceuticals, also known as Biologics, are pharmaceutical products derived from living organisms by using biotechnological methods. Biological systems, such as bacteria, yeast or mammalian cells that are manipulated in their DNA, aimed at producing therapeutic and medical diagnostic products are utilized in industrial processes. Conventional chemical medicines like Generics, tend to have small size and simple structure, and appear in identical copies generated by predictable chemical process, whereas Biopharmaceutics are derived from one unique cell line, making it impossible to ensure identical copies. Biotechnological synthesis enables the production of molecules of large size and complex structure, where living cells feature specific posttranslational modifications which is of crucial importance for the activity as well as the immune tolerance of a biological molecule. [1]

The class of biopharmaceuticals has been available for over 30 years. Over the last years, monoclonal antibodies, hormones, clotting factors enzymes, vaccines, nucleic acid- based products as well as cell-based products represent the most important products in the Biopharma industry. [1],[2] The demand for approved biopharmaceuticals produced from animal cell culture processes increases, not least due to the efficacy of several humanized monoclonal antibodies that are required in large doses. The number of sales reflects the importance of this sector. In 2017, $80.2 billion were generated by the top ten biopharmaceuticals, which represent about 44% of the total revenues in this category. [3]

So-called biosimilars can enter the market as soon as a biopharmaceutical loses its patent protection after 20 years. Biosimilars are replications of approved biopharmaceutical products, which are very similar to their originator products in terms of quality, safety, and efficacy. Still, it is impossible to generate entirely identical products in a new bioprocess. The business of biosimilars gains importance, due to the surge of biologics with expired patents. In 2006, Omnitrope® entered the European marked as a first biosimilar preparation, distributed by Sandoz. [1],[3],[4]

To successfully develop a process for manufacturing a biosimilar, excellent process understanding and strong control strategies are required, since these processes are equally

complex as the originator process, but with pre-defined target ranges for different quality attributes. [4]

For marketing authorization, clinical trial data must be generated, to demonstrate the comparability of the biosimilar to the reference biopharmaceutical. Although there is no need to repeat all trials of the reference biopharmaceutical, the necessity to conduct some biosimilar trials implies considerable expense and time. [1],[3],[5]

It is reported that the biosimilar competition has a significant economic impact with EU-wide price reductions of 8% to 34% on different product classes. Considering 30% saving across the board of biosimilars worldwide, a hypothetical calculation predicts about $22 billion to the global healthcare systems. [2]

1.2 Process analytical technologies (PAT)

The US federal development agency (FDA) first presented the PAT initiative for modern and advanced process control to reach higher health and economic benefits in pharmaceutical manufacturing. The development of new biopharmaceutical processes is tightly connected with the objective to increase process robustness and understanding. [6] While conventional pharmaceutical manufacturing is monitored by laboratory testing of collected samples, there are significant opportunities for quality assurance, product and process development, analysis, and control. PAT has been defined as a system for designing, analyzing and controlling manufacturing through timely measurements (i.e. during processing) of critical quality and performance attributes of raw and in-process materials and processes, with the goal of ensuring final product quality. The goal of PAT is to raise process understanding and control the manufacturing process to obtain a consistent quality.[7]

PAT can appear as process control based on real-time measurement of a critical quality attribute (CQA), of parameters that directly correlate with CQAs of a parameter, which ensures the fit for purpose of several unit operations. [8] Application of PAT remains as a special challenge for cell culture processes as the biological processes are inherently complex, leading to poor understanding of important variables and interactions between variables.

Nevertheless, there are some sensors, which have proven reasonable correlations to relevant process parameters and CQAs. For example, dielectric spectroscopy probes can be used to establish a mathematical correlation to biomass or Raman spectroscopy detects media components and metabolites. Those two applications will be discussed in detail below. [8]

The implementation of PAT in an industrial environment requires the use of a combination of multiple of the PAT tools described below:

- Process analyzers collect all types of process data. This can be at-line (sample is isolated from process stream and analyzed in close proximity), on-line/in-line (sample not diverted from the process stream an analyzed either invasive or non-invasive) or off-line (sample separated and measured independently from time and place).
- Multivariate tools for design, data acquisition and analysis for making us of the relationship between several parameters such as multivariate statistical tools.
- Process control tools
- Continuous improvement and knowledge tools [7]

A process that is designed to enable complete understanding throughout all phases, guided by the PAT framework, will boost the fulfilling of consistent quality requirements and thus be the base of the Quality by Design (QbD) tenet. The risk to quality and regulatory concerns is reduced by applying these tools. [7]

1.2.1 Dielectric spectroscopy

Dielectric spectroscopy probes are also called capacitance probes. This points at the measurement principle behind this type of probes. An alternating electrical field is applied to shortly polarize the sample and the charge stored therein, called capacitance, is measured. This signal gives information about the viable cell density (VCD) in the bioreactor. Dielectric spectroscopy is therefore an appropriate method for online measurement of the biomass in a bioprocess. Unlike other biomass monitoring methods, e.g. optical density, dielectric spectroscopy does measure intact cell membrane and thus dead cells are not accounted. [9]

1.2.1.1 Theoretical background (electric field and frequency scan)

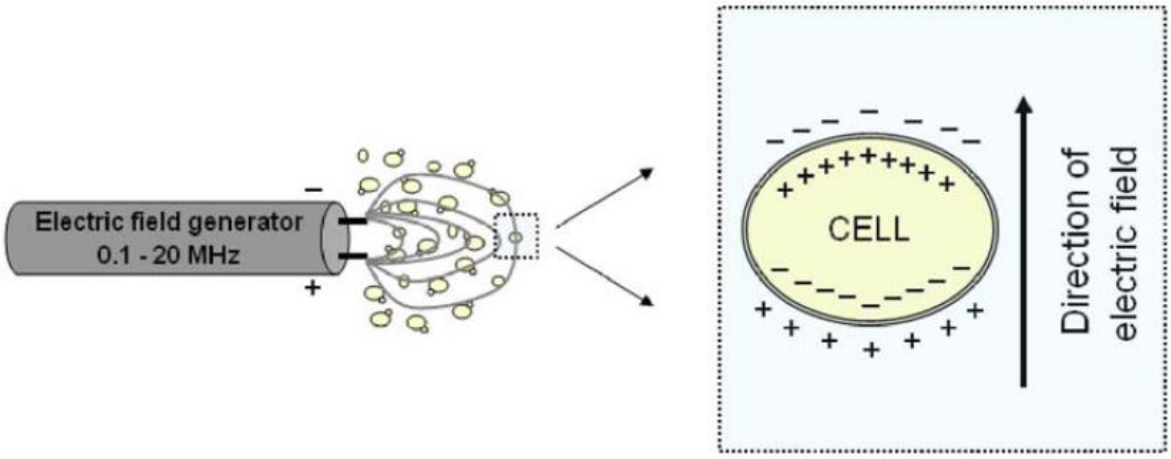

Figure 1: Cell polarization at the plasma membrane in an electric field applied by a dielectric probe. [10]

The basic principle of dielectric spectroscopy is based on the membrane properties of a viable cell. If an electric field is applied to a suspension of cells in an aqueous ionic solution, a polarization profile at the membrane is formed, as all ions migrate to the opposite pole, but are hindered by the intact membrane barrier of a cell. A schematic graphic of a polarized cell is shown in Figure 1. Thus, a cell can be described as a capacitor in this application since it stores a defined amount of charge outside its cell membrane. Consequently, the capacitance of the suspension is a measure of the amount of cell plasma membrane within the suspension and can be correlated to the VCD of the suspension. This field induced polarization is measured in Farads per meter (F/m). [10], [11] Conductivity, the charge that can pass through the medium is measured in Siemens per meter (S/m). The permittivity (ε) and the conductivity (σ) are described by the following formulas.

$$\sigma' = G * (d/A)$$

$$\varepsilon' = C * (d/A\,\varepsilon_0)$$

G ... conductance [S]
C ... capacitance [F]
d/A ... probe constant [cm^{-1}]
ε_0 ... permittivity of vacuum [$8.854 * 10^{-12}$ F/cm^{-1}]

Equation 1: relative conductivity and permittivity

The probe constant is taken into account for the relative conductivity and permittivity, as the signal is highly dependent on the geometry of the two electrodes in the used probe. [12]

Once the electric field switches poles, all ions are pushed to the opposite of the cell, but the magnitude stays the same. The complete formation of this polarization profile takes a finite amount of time. This is where the rate of reversing the electric field is coming into play. A low frequency enables a high number of ions to travel to the opposite pole, whereas only few ions succeed reaching the new pole if the time between reversions is short. In frequency scan applications, the capacitance at different frequencies is recorded, whereby the electric field is changed within radiofrequencies between 50 kHz to 20.000 kHz. At very high frequencies, the timespan for ion movement is so small, that only a few ions can make it. Thus, the contribution of the cells to the signal is very small, and almost only background is measured. The decrease in capacitance that is caused by increasing frequencies is called β-dispersion, whereas the curve at frequencies below this range is called α- dispersion and at higher frequencies, there is γ- and δ-dispersion. [11], [12]

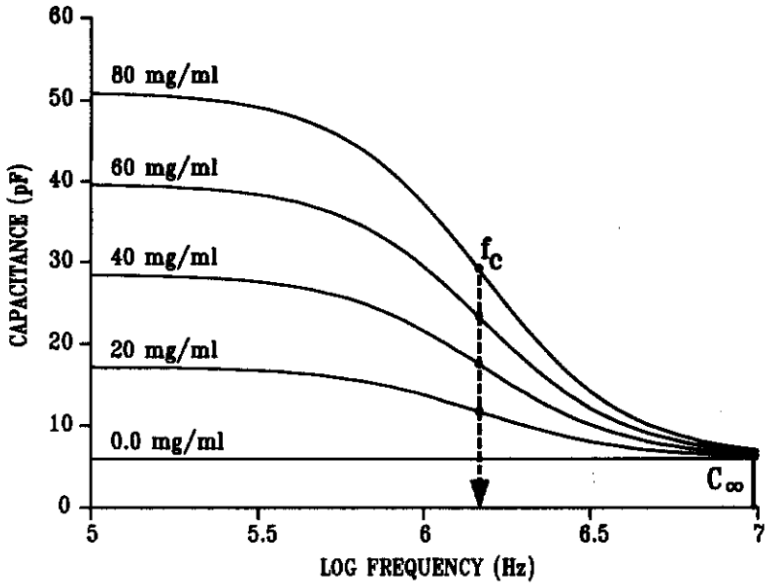

Figure 2: The effect of increasing cell densities on the ß-dispersion. [11]

The shape of a β-dispersion curve, which is relevant for biomass quantification, is drafted in Figure 2. Such a curve can be described as negative sigmoid function specified by a few characteristic parameters, the so-called Cole-Cole parameters. As shown in the Figure 2,

there are two plateaus, one at very low frequencies and the other at very high frequencies. The difference of the capacitance values of these two plateaus is called capacitance increment (ΔC) and rises with higher VCD in suspension. The frequency at the inflection point of the beta-dispersion curve is called critical frequency (f_c). The slope of the curve at the f_c point is called α. [10],[13]

$$C(f) = \frac{\Delta C \left(1 + \left(\frac{f}{f_c}\right)^{(1-\alpha)} \sin\left(\frac{\pi}{2}\alpha\right)\right)}{\left(1 + \left(\frac{f}{f_c}\right)^{(2-2\alpha)} + 2\left(\frac{f}{f_c}\right)^{(1-\alpha)} \sin\left(\frac{\pi}{2}\alpha\right)\right)} + C_\infty$$

Equation 2: Cole-Cole equation based on the Debye equation

The Cole-Cole equation is derived from the Debye equation, by introducing an empirical value alpha. The equation assumes that the polarization of material decays exponentially by removing an applied electric field. [10]

Biomass monitoring by using linear capacitance models correlating a single frequency signal to VCD is already of use in industrial production process. [14] This model provides VCD estimation during the linear growth phase of a fed-batch fermentation, since the physiological state of the cells, as well as cell morphology are nearly stable. Once considerable changes occur in these parameters, the signal no longer follows the trend. [15] This limitation was shown to be overcome by applying multiple frequencies of dielectric pulses, a frequency scan, to gain more information about the capacitance characteristic of the sample and predict the VCD by the use of a multivariate statistical model. [10]

Several researchers investigated in the field of dielectric capacitance application for biomass prediction fed-batch processes before. Some of these studies are stated below.
Dabros et.al. evaluated three techniques of calibrating capacitance spectrometers for biomass prediction, namely using the theoretical Cole-Cole equation, linear regression of dual-frequency capacitance measurements and multivariate modeling of a dielectric frequency scan. The PLS model turned out to be the most robust model in handling signal noise. [10] Konakovsky et.al. investigated the model transfer of multivariate prediction models for biomass prediction between cell lines and process conditions. A novel approach is

able to estimate VCD in real time without further off-line analytics after one biomass measurement at inoculation for offset correction. [16] Opel et. al. examined online prediction of different routine biomass measurements by linear modeling, Cole-Cole modeling and PLS regression. In this paper it is demonstrated, that the definition of viability is critical when analyzing biomass online, and also that dielectric spectroscopy is a complementary measurement of viable biomass, providing useful information about the physiological state of a culture. [17] Cannizzaro et. al. was monitoring CHO perfusion culture experiments by dielectric spectroscopy and was able to correlate as well the viable cell number up to 10^7 cells/mL as the median cell diameter with minor deviations. [13] Another work demonstrating the prediction of VCD by different models was done by Párta et. al., who obtained good predictions throughout the whole fed-batch process (early and decline phases) by PLS and Cole-Cole models, whereas the linear model only convinced in the early state. [15] Braasch et. al. was measuring the viability of a cell culture by several methods including dielectric spectroscopy, flow cytometry and a dielectrophoretic cytometer. They found out, that dielectric probes are sensitive to the early apoptotic changes in cells. [18]

1.2.2 Raman spectroscopy

Raman spectroscopy is a vibrational spectroscopy technique that makes use of the Raman optical activity of molecules of the sample. Considering the possibility of nondestructive, rapid analysis, Raman spectroscopy is an appropriate PAT tool. It is established as such in bioprocesses, as well as in other applications including polymorph identifications, in situ crystallization monitoring or real-time release testing for the last three decades. The “molecular fingerprint” that is recorded by a Raman probe can be used to monitor and model important process performance parameters. Therefore, it enables better understanding and control of bioprocesses. [19],[20] Not only several metabolites like glutamine, lactate, ammonia or glutamate can be detected and quantified with this method, but also information about viable and total cell densities and debris is found in these spectra. [21]

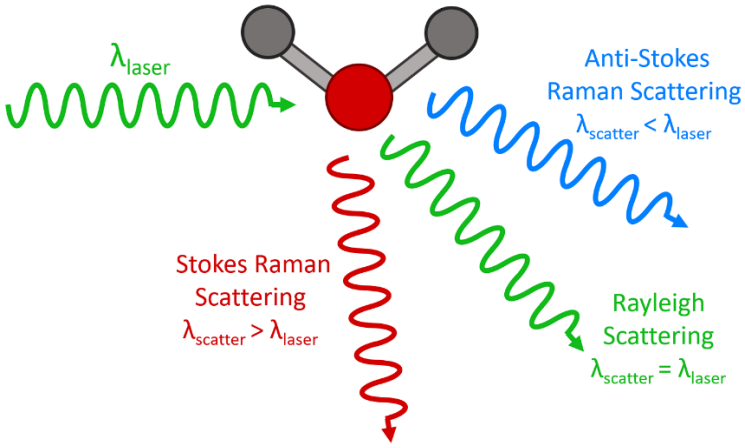

Figure 3: An incoming monochromatic laser light is scattered by the vibration of a molecule. The light is scattered with either the same energy (Rayleigh scatter), higher energy (Anti-Stokes Raman scatter) or lower energy (Stokes Raman scatter). [22]

Raman spectroscopy is a method based on the vibrational transition of molecules. A laser light is scattered, and molecules selectively absorb small amounts of the irradiating light. For pharmaceutical application, the light wavelength is within near-infrared range (λ=785 or 830nm). This light has higher energy than needed, to bring the molecules to a higher vibrational state.

Most of the incident rays are scattered in the same frequency as the energy of the molecules. This light is called elastic Rayleigh radiation. Only a small number of photons is elastically scattered by the molecule, which indicates an energy exchange between the incident light and the sample. This light can either have a lower or higher energy than the incident radiation and describes the "Raman shift". Thus, if a molecule is brought to a higher vibrational state by the incident radiation it consumes energy. The light returning with lower energy is called Stokes scatter. Vice versa, molecules that already are in a higher vibrational state can release energy to the incident light and go back to normal state. This phenomenon is termed anti-Stokes scatter. Anti-Stokes scatter occurs very rarely in biotechnological applications, because molecules tend to appear in normal state at room temperature. Consequently, Raman scatter can be measured by observing the energy changes from one molecular energy level to another. The "Raman shift" is very specific to the composition and the concentration of the molecules occurring in the sample. Thus, the spectra obtained from a Raman probe can serve to detect metabolites, products or even cell densities in a fermentation broth. [23], [24]

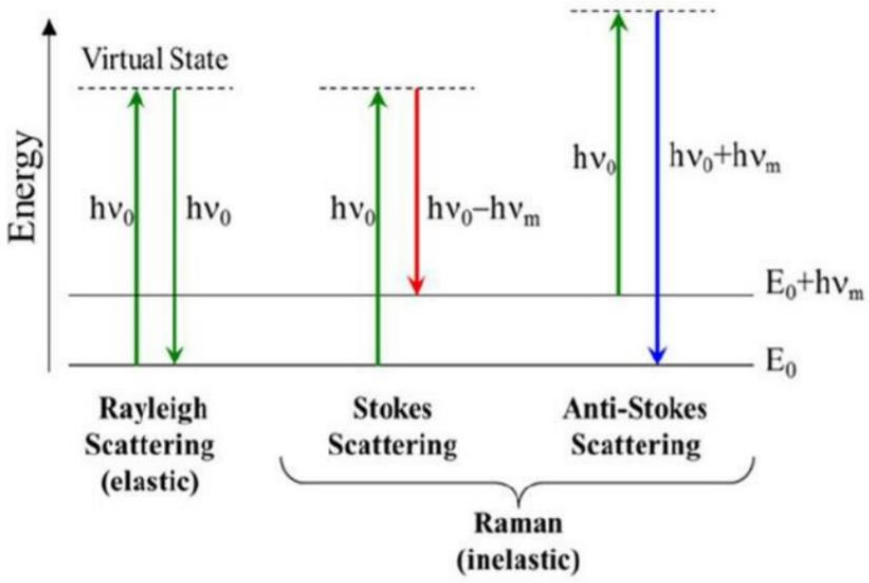

Figure 4: A schematic representation of elastic and inelastic/Raman scattering. Rayleigh scattering with the same energy in both directions whereas stokes scattering comes from a molecule ending up at a higher state and Anti-Stokes scattering results from the loss of an excited state. [25]

1.2.3 Soft sensors

The designation soft sensor is derived from the two words “software” and “sensor” and describes a technique to model a functional relationship between process variables (easy-to-measure variables) and quality variables (difficult to measure variables). [26], [27]

It can be distinguished between model based and data driven soft sensors. If a first principal model accurately describes the process, a model based soft sensor can be used. Indeed, many of these first principal models are computationally complicated, what makes a real-time application difficult. As opposed to this, the large amount of data measured throughout a process can be combined and used for a statistical model that predicts performance or quality parameter, which usually demand offline analysis. This approach is called data driven soft sensors. Furthermore, many process mechanisms are not well understood, thus empirical models, such as multivariate statistical models are used to build regression models. By evaluating the use of different combinations of datasets to predict performance and quality parameters, data-driven soft sensors can be optimized and can be implemented in processes

as PAT tool. [28] There are several applications of soft sensors in use. An example is the use of standard bioreactor on-line data like Gas-flow rate, oxygen and carbon dioxide concentrations, dissolved oxygen tension (DOT) together with feeding and titration rates, which are statistically modelled to predict biomass in microbial applications. [29]

1.3 Monitoring in cell culture

1.3.1 Cell density and viability

Since the monitoring of viable and total cell density in a bioprocess are of high importance and thus basic parameters to be analysed, many different methods are developed, to measure these parameters. Staining with viability dyes or flow cytometry are two widespread options for determining the cell density in a cell culture.

Trypan Blue Exclusion

Trypan blue exclusion is a test based on the principle that live cells possess intact cell membranes that exclude certain dyes, such as trypan blue, eosin or propidium, whereas dead cells do not. When mixing cell suspension with trypan blue, it can be visually examined to determine whether cells take up or exclude dye. [30]

1.3.2 Apoptosis in cell culture

Apoptosis is one form of the programmed cell death which is actively regulated by the cell. Several intra- and extracellular signals are activating an interconnecting cascade of events involving various families of proteins. An early indication of apoptosis in mammalian cells is the presentation of phosphatidyl-serine on the cell surface. Also, cell shrinkage occurs due to ionic cell content regulation and chromatin condensation. Such pro-apoptotic signals trigger the cascade of caspase activity within the cell, leading to the late stage of apoptosis, which is characterized by DNA fragmentation and loss of membrane integrity. [31], [18]

The classical approach to detect these stages is by using specific markers with a flow cytometer, which is described later. A new approach is to connect the ionic flux over the cell membrane with the onset of apoptosis. By the phenomenon of dielectrophoresis, where the uncharged particles move along the gradient of a non-uniform electric field, a monitoring for dielectric changes is possible, which correlates with the physiological and metabolic changes of cells. Further, the heterogeneity of the growth cycle phase of individual cells within a bulk

allows only the monitoring of the average dielectric properties of the cell suspension. Investigations on the online detection to determine early apoptotic events by dielectric spectroscopy are described by Braasch et al. [18] Other studies report of a way to even control apoptosis in a bioprocess by the signal of dielectric spectroscopy, as the media composition is altered if early apoptotic events start to occur more often. [32]

Flow cytometry

By the technology of flow cytometry intrinsic and evoked optical signals from single cells in a moving fluid system can be measured. The final data product is generated by the interplay of three general systems: a fluidic system, an optical system, and a computer. The fluidic system takes up a cell suspension and passes them through a nozzle, so that the cells access the flow cell in single file. By passing the flow cell, the optical system is coming into play. The optical system consists of a laser as light source, focusing lenses, color-selective mirrors, filters and a signal detection and evaluation module with photomultiplier tubes or photodiodes to detect the optical signal followed by analog and/or digital electronics to process and evaluate the signals. For receiving, storing, further processing and displaying the resulting data a computer is required. [33]

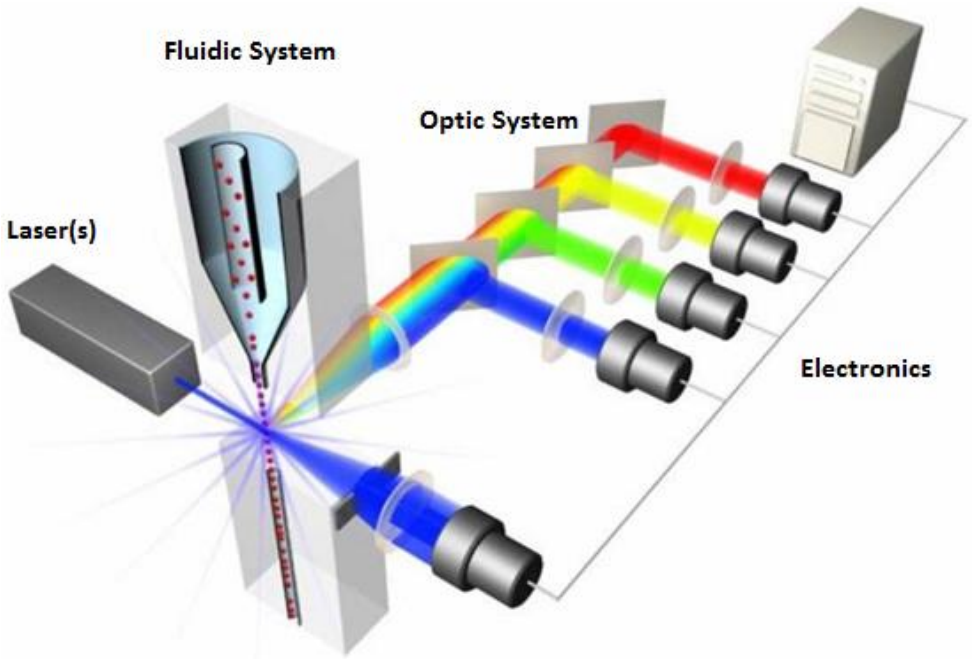

Figure 5: A schematic illustration of the basic components of a flow cytometer. The three main components are the laser system, fluidic System and optic system. [34]

The single cells are passing the laser beam, where the light gets scattered by the cells. Two kinds of scattering can be measured: the forward scatter (FSC), giving information about the particle size and the side scatter (SSC) inferring the granularity of cells. [35]

Beyond that, by addition of fluorescent dyes specific cell properties can be detected. As an example, 7-aminoactionmycin D (7-AAD) is an often-used stain for discriminating living and dead cells. As the membrane integrity of dead cells is lost, 7-AAD can enter the cell and intercalates in the DNA. If the marked cell passes the laser beam the fluorescent molecule is induced and an extinction spectrum can be measured. Annexin V-PE represents another fluorescent marker, that binds phosphatidyl-serin which is presented on the surface of early apoptotic cells. Treating and detecting a sample with those two markers, a diagram like in Figure 6 allows to cluster the cells or also called events in different stages of life cycle. [35] [18]

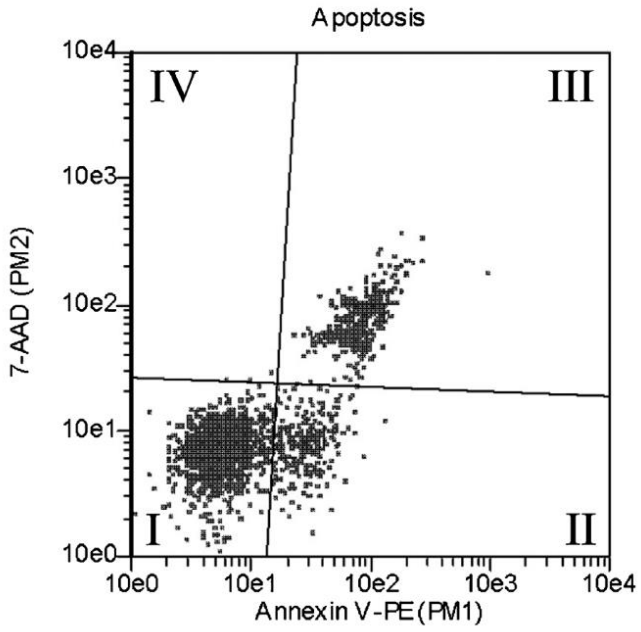

Figure 6: Example of CHO assayed using the Guava Nexin assay. The sample is treated with 7-AAD and Annexin V-PE. This allows identification of the different apoptotic stages identified by the Nexin assay: I viable/non-apoptotic cells [Annexin V-PE (-) and 7-AAD (-)]; II early-apoptotic cells [Annexin V-PE (+) and 7-AAD (-)]; III late stage apoptotic/dead cells [Annexin V-PE (+) and 7-AAD (+)]; IV nuclear debris [Annexin V-PE (-) and 7-AAD (+)]. [18]

1.4 Multivariate data analysis and chemometric modeling

Other than the classic approach, where one single selective signal is measured to provide the desired information, modern process analytics commonly deliver many non- selective signals. Chemometrics is an interdisciplinary science that enables to combine such signals in a multivariate model. It is essential in understanding and diagnosing real-time processes and keeping them under multivariate statistical control. Since cell culture processes are complex in terms of composition and events occurring throughout a bioprocess within the bulk arising from cell growth, death, and metabolism, it is often impossible to match specific process values to single signals or peaks of a spectrum. Multivariate data analysis (MVDA) provides many tools enabling to screen the essential information within a large multidimensional dataset. [36], [37] Prominent examples of such tools are principal component regression (PCR), partial least squares (PLS) or multiple linear regression (MLR). [38]

1.4.1 Data pretreatment

The quality of MVDA tools can be improved if data pretreatments and signal correction are executed in advance. Model building can be needlessly complicated, if redundant signals or especially for spectral data unrelated regions are included in the dataset. Also, some pretreatment methods manage to highlight significant information within the data. There are many approaches for extending the density of information. The common goal is to remove undesired systematic variations like baseline drifts or multiplicate polarization and scatter effects. In this chapter some frequently used pretreatments are described.

Data selection

In dielectric spectroscopy problems due to electrode polarization and baseline distortions can appear. Those effects are and can be probe specific and thus can disturb the quality of a model including data from several probes. Yardley et. al. discusses strategies to clear these errors. A very easy and effective method is to exclude the capacitance data at frequencies below 300 kHz. [13], [39]

Raman devices mostly measure spectra in the range of 100 to 3425cm^{-1} but not all spectral parts are relevant for MVDA and model building as some areas are sensitive to scattering and fluorescence effect or instrument instability. Methods to select important regions are described in literature. [40]

Mean centering

Mean centering is an approach to alleviate multicollinearity in linear regression approaches. In matrix applications it might provide a simpler interpretation of data analysis. The variable's mean is subtracted from all other observations belonging to this variable in the dataset attaining the new variable's mean to be zero. [41]

Scaling

This data pretreatment approach divides each variable by a scaling factor, which is different for every variable in the dataset. Thus, measured values become values relative to the scaling factor, which results in the inflation of small values. [42]

Standard normal variate

Standard normal variate (SNV) is a method to normalize data column wise. As it is a weighted normalization method not all points contribute to the normalization equally. [43]

First and second derivation

First and second order derivation are row wise pretreatments that are used to reduce scatter effects for continuous spectra. First derivatives are used to remove additive baselines and result in a spectrum of slopes of the original spectra in each point.

1.4.2 Partial least squares Regression (PLS-R)

PLS or projection to latent structures by means of partial least squares was developed by Herman Wold in the late 1960s and brought to chemometrics by his son Svante Wold in the 1980s. Nowadays it is one of the most widely used technique in chemometrics, particularly for problems where p > n, like spectral data. For performing PLS, there are two different algorithms in use, namely NIPALS and SIMPLS, whereas NIPALS is used in this application.

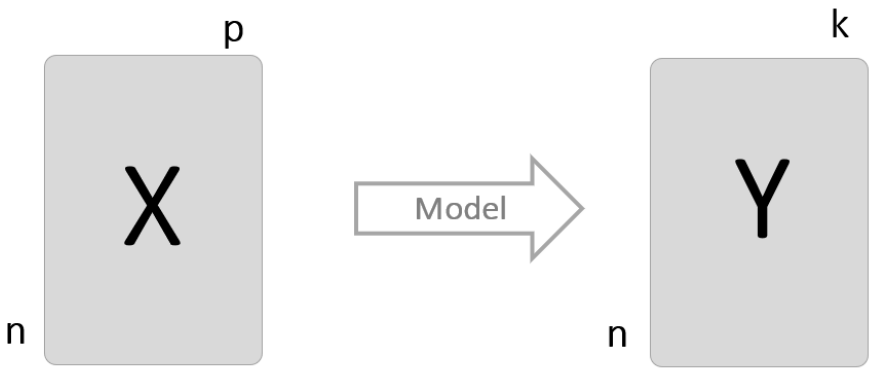

Figure 7: schematic depiction of the dataset needed for a PLS calibration. X is a table containing n observations of p variables of a new (often spectral) method and Y is a table of the observations measured by conventional methods.

Generally spoken, PLS is a supervised technique relating two sets of variables, namely $X_{n\,x\,p}$ and $Y_{n\,x\,k}$, where one can differentiate between PLS1 with a single response (*k* equals 1), and PLS2 with more than one response variables arranged column wise.
The data block X is used to predict a variable Y which might be either expensive or time consuming to measure. Once a model is built, there is not any more need to measure Y, but predict it with X.

Equation 3: necessary steps to develop a PLS calibration model.

$$X_{n\times p} \xrightarrow{PCA} T_{n\times m} \xrightarrow{model} Y$$

Ad interim, a principal component analysis is performed on $X_{n\,x\,p}$, which searches for orthogonal directions in p-dimensional space with a maximum of information. This so-called loading vector $T_{n\,x\,m}$ determines a new coordinate system containing the most information in the first few coordinates. Compared to the original matrix, the new coordinate system is rotated. Other than in a classic unsupervised PCA, PLS also uses the information of the response *Y*. The *m* linear combinations are latent variables extracted in such, as the covariance between the scores and Y is maximized.

After preprocessing, the data X is called V_1 and Y is called U_1.

Equation 4: PLS Step 1 - Preprocessing

$$X \xrightarrow{center} V_1 \qquad and \qquad Y \xrightarrow{center} U_1$$

Now *p* univariate regression models for each response against the predictors are built and the resulting regression coefficients b_{1j} are given by the scalar products.

Equation 5: PLS Step 2 – get T_1

$$U_1 \sim V_1 \qquad for \qquad j = 1, \dots, p$$

$$\hat{U}_{1(j)} = b_{1j} V_{1j}$$

$$b_{1j} = \frac{v_{1j}^T u_1}{v_{1j}^T v_{1j}} = \frac{Cov(v_{1j}, u_1)}{Var(v_{1j})}$$

Each of the *p* regression equations provides an estimate of the response. The first PLS component T_1 is now the weighted average.

Equation 6: PLS Step 3 – Weighting

$$T_1 = \sum_{j=1}^{p} w_{1j} b_{1j} V_{ij} \qquad with \qquad w_{1j} = Var(V_{1j})$$

There might be residual information in Y not explained by T_1 which is expressed in U_2, and residual information in X_j that is not explained by T_1 which is expressed in V_{2j}

Equation 7: PLS Step 4 – Preparation for T_2

$$U_2: \ residual\ of\ U_1 \sim T_1$$

$$V_{2j}: \ residual\ of\ V_{1j} \sim T_1$$

Steps 2 and 3 are repeated using U_2 and V_{2j} instead of U_1 and V_{1j}, to obtain the second PLS component T_2. The same process is continued until all min *{n, p}* components are extracted.

Equation 8: PLS Step 6 – OLS Regression

$$\hat{y} = \alpha_0 + \alpha_1 \times T_1 + \alpha_2 \times T_2 + \cdots + \alpha_m \times T_m$$

By an ordinary least square regression the final model is fit.. X is reproduced exactly, if we extract the maximum number of components, but for creating a model that fits for new data as well, a reduction of components *m* is required. Thus, an error term must be added to the

formular if the number of components *m* is reduced. The number of components m used for predicting y is determined via k-fold cross validation. [44], [45]

1.4.3 Cross validation (CV)

Cross validation (CV) is an important tool in chemometrics to evaluate a predictive model. Basically, the predictive ability of a model is tested by comparing predicted values with actual Y values. There are plenty of different CV methods sharing the same purpose, which is estimating the possible prediction error of an unknown dataset on the one hand, but also for determining how many components are needed to characterize the data.

Leave one out CV, k-fold CV and groupwise CV are three commonly used examples of Cross validation, that all are based on the same method. The available dataset is split into a training and a test set. The training set is used for building a statistical model, whilst the data box X of the test set is used to predict new values by the help of the model. [38]

A k-fold CV randomly divides the dataset in k groups and keeps excluding one group in order to create a training dataset, whereas groupwise CV splits the dataset by any other criteria, which could be batchwise splitting. [44]

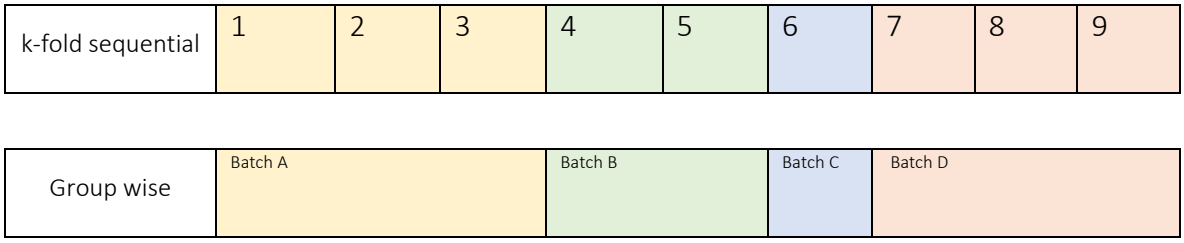

Figure 8: schematic depiction of methods for splitting a dataset into test and training set to perform a cross validation.

Once there are scores predicted for the test set, there are two values to be calculated to enable a comparison. The root mean squared error of prediction (RMSEP), the mean absolute prediction error (MAE) and the mean absolute percentage prediction error (MAPE) are widely used parameters for evaluating the goodness of prediction.

Equation 9: Calculation of root mean squared error of prediction

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

Equation 10: Calculation of mean absolute prediction error

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)$$

Equation 11: Calculation of mean absolute percentage prediction error

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(\frac{(\hat{y}_i - y_i)}{y_i} * 100)$$

For determining the optimum number of components to use in a PLS model, the RMSEP is calculated by CV for a model with 1 to *m* components. The number components whose RMSEP is not significantly worse than the RMSEP with an additional component is chosen as number of latent variables.

2 Aim of this work

The use of dielectric capacitance probes in single frequency mode is an established method for monitoring biomass in development and production processes. To optimize the prediction of viable cell density especially during stationary phase and death phase, the use of dielectric frequency scan data and generation of multivariate models should be investigated. Also, it should be tested if the addition of Raman data to the multivariate models significantly improves the prediction of viable cell density and viability.

Since it is reported in literature, that other interesting cellular parameter can be related to the signal of a dielectric frequency scan probe, investigations in correlating average cell diameter, average circularity and apoptotic events with capacitance data will be further evaluated.

Thus, the full potential of dielectric spectroscopy as a PAT tool in cell culture Fed-batch process should be exploited and evaluated to enable the most efficient use of dielectric spectroscopy.

3 Materials and methods

3.1 Cell culture processes

For testing and comparing several PAT tools, data from 38 mainstage fermentation processes were collected. Those cultivations were CHO cell culture processes, which are used for producing monoclonal antibodies in development scale and production scale. All used vessels are equipped with online sensors for temperature, pH, pO2, as well as control units, which utilize a fully integrated process control system (PCS) that enables to automatically keep the relevant operating parameters at the defined set points. The PCS of all bioreactors also feature running automated cleaning in place (CIP) and steaming in place (SIP) programs, which are performed before and after each mainstage cultivations according to the company internal SOPs.

3.1.1 Development Process

The experiments are performed with 20L or 30L Sartorius stainless steel bioreactor systems (BIOSTAT® D-DCU, Satorius Stedim). Whereas the entire USP process from vial break to harvest takes about 34 days, the main-stage fed-batch process takes 14 days. PH, DO and temperature are controlled at setpoint via closed-loop control. The feeds are added as daily bolus shots. All runs are part of a study focusing on improvement of process robustness to reduce variability in process performance concerning VCD, harvest titer and product quality. The process parameters that were varied purposely during the study were the seeding VCD (0.35, 0.60 and 0.85 [10^6 cells/mL]), the pH after T-shift (pH6.70, pH6.72 and pH6.74), VCD at T-shift (2.0, 4.0 and 6.0 [10^6 cells/mL]), the gassing strategy and the feeding regime. Also, the implementation of capacitance probes will be tested. For this study, the standard main-stage medium and the standard feed solutions were used and prepared according to company internal SOPs. For doing at-line and offline measurements, daily samples were taken from each batch and the relevant process parameters were measured according to the company internal SOPs (more details in chapter 2.2 offline analytics). 26 batches of this study were used for this thesis.

3.1.2 Production Process

The production Process is a highly standardized process with the aim of producing plenty batches with equal product quality of monoclonal antibodies for pharmaceutical use. The main stage of this process is performed in a 13000 L stainless steel tank reactor. After vial break 17 days of a cascade of preculture steps are performed to prepare the mainstage inoculum. The mainstage fermentation is then harvested 10 days after inoculation. PH, DO and temperature are controlled at setpoint via closed-loop control. The feeds are added as daily bolus shots. A mainstage medium and the feed solution for re-supplementation of nutrients, prepared according to the company internal SOPs. For doing offline VCD measurements, daily samples are taken from each batch according to the company internal SOPs. 19 batches of this process were used for this thesis.

3.2 Data acquisition and modeling workflow

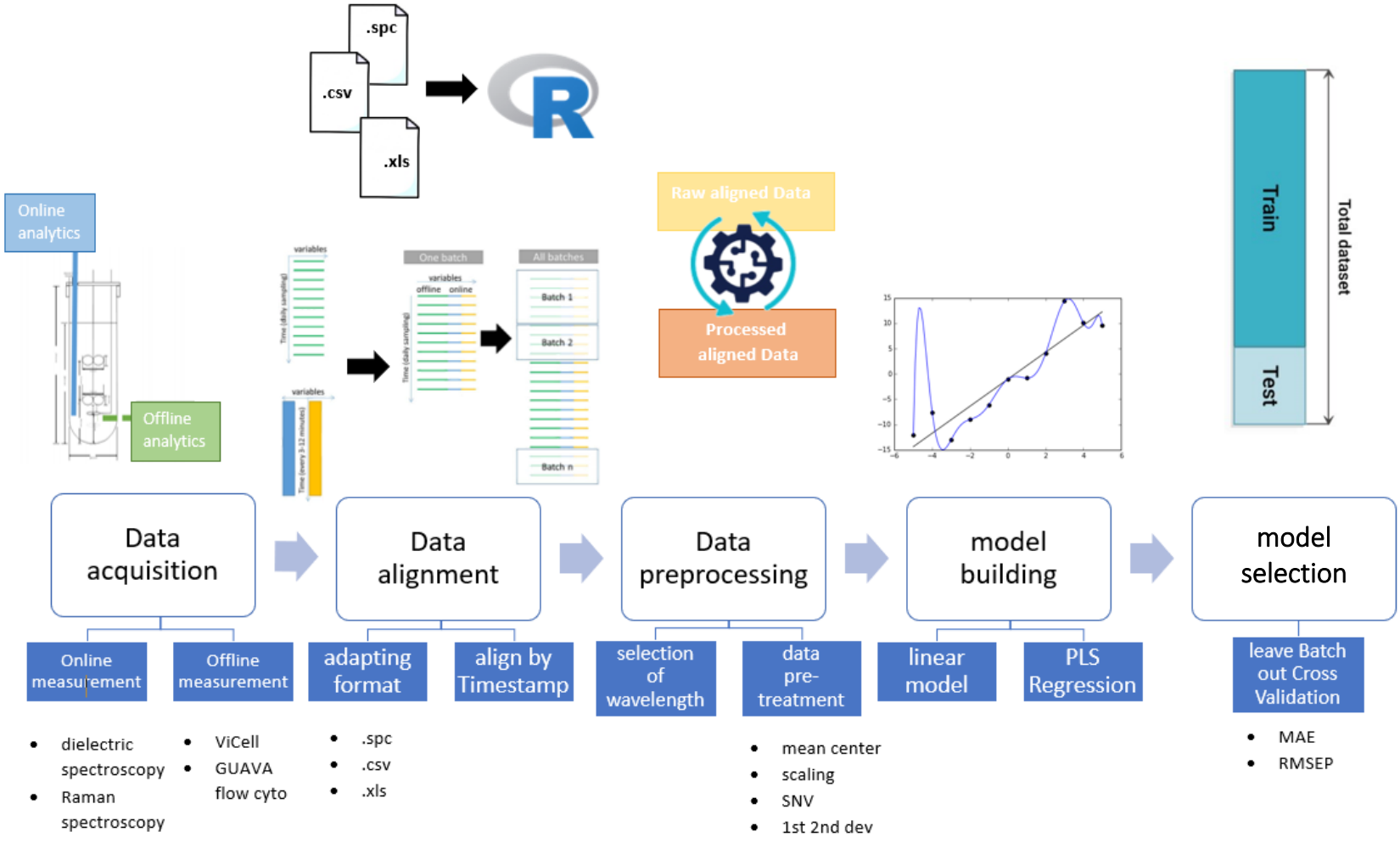

Figure 9: A sketch of the workflow from measuring data of a bioprocess to model selection and evaluation

The general workflow sketched out in Figure 9, from collecting data of a bioprocess to generating a model to predict cell densities and viability includes the steps described below.

Data acquisition

For the generation of online data like dielectric spectroscopy and Raman spectroscopy the probes must be assembled before the sterilization of the bioreactor. Once the fermentation run has stopped all data can be extracted from the data source. For all offline measurements, the daily samples are collected. These samples are analyzed by ViCell^{TM} and the GUAVA Easy Cyte.

Data alignment

Since those data arise from different devices, different data formats must be connected. First all data are imported to the statistical program R and brought to the same format. Then the online data and offline data are aligned to the closest common timepoint. Later, all batches are connected to one big data frame.

Data preprocessing

Especially for spectral data it appears that redundant information exists in the large dataset, which makes modelling unnecessarily complicated. Therefore, columns with redundant and meaningless information are excluded. All other data undergo specific pretreatments like mean centering, scaling, standard normal variation 1st derivation and 2nd derivation to amplify the information content.

Model building

Several simple linear regression (lm) and partial least squares (PLS) models are built varying the data input, the data pretreatments and the response parameter.

Model selection

To compare and validate the quality of those different models, “leave one batch out” cross validation was performed.

3.3 Offline analytics

As mentioned before, each fermentation process is sampled daily to perform associated offline analytics. An overview of the process parameters measured as standard, and the comprehensive methods and devices are listed below. Additionally, flow cytometric measurements are performed for some batches.

Table 1: offline standard process analytics

Parameter	Device / Method
VCD, Viability	Vi-Cell XR
pH, pCO_2	Nova pHOx
Osmolality	Advanced Instruments Fiske 2020
GLUC, GLN, NH_4^+, LAC	Nova Bioprofile 100+

For the modeling performed in this thesis the relevant measurements are the Vi-Cell measurements and the flow cytometric measurements by the GUAVA system.

3.3.1 Vi-Cell™

The reference values for VCD and viability are measured using the Vi-CELL™ Cell Viability Analyzer (Beckmann Coulter), which is an automated system based on trypan blue exclusion. After pipetting 0.6 mL of the sample into a Vi-CELL™ vial, the device automatically mixes the sample with a reagent in a 1:1 ratio and incubates the sample for a defined time. Later the system takes 50 pictures of the sample and analyzes the suspension by different grey scales. Living cells appear light grey, whereas dead cells are dark dots on the picture. Further, cell diameter and average circularity can be measured by the Vi-Cell device. The viability is calculated as it is the quotient of the viable cell density and the total cell density.

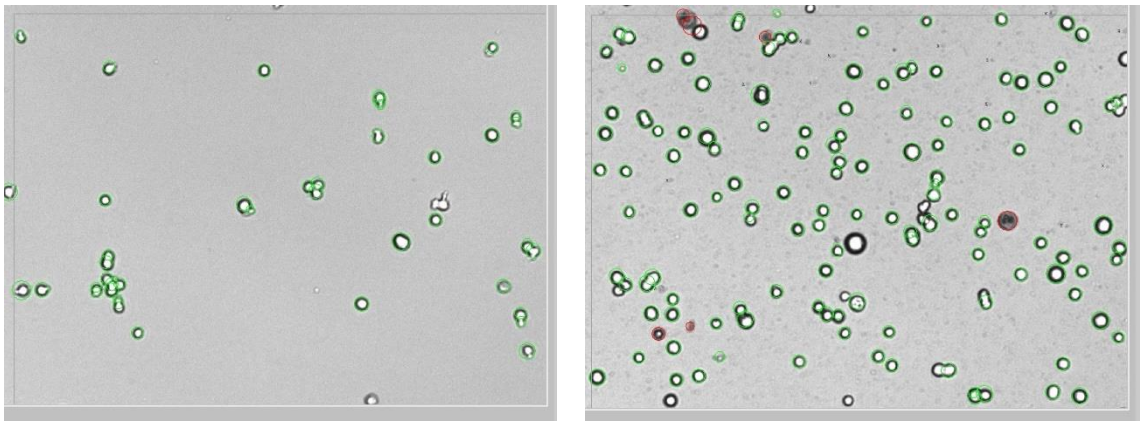

Figure 10: pictures of the cell suspension analyzed by Vi-Cell

In Figure 10 two exemplary pictures of a sample in an early fermentation step (left) and a late state (right) are shown. The green circles present cells recognized as viable cells, whereas the red circles point out the dead cells. The Vi-Cell Analyzer also computes the diameter of those circles.

3.3.2 Guava technology

All flow cytometric analysis were done by the help of the GUAVA Easy cyte™ Plus, which is a tabletop device providing several Kits for predefined assay. The device features an 488nm Argon Laser, forward scatter (FSC), fluorescence detectors in yellow (PM1) at 583 +/- 26 nm, red (PM2) at 680 +/- 30 nm and green (PM3) at 525 +/- 30 nm, samples in 1.5 mL tubes or U-shaped 96-well plates, sample volume of less than 20 µL, counting accuracy of +/- 10 % and precision of < 10% CV. The Guava Easy cyte™ Plus can be controlled by a connected PC running the GUAVA CytoSoft™ software. The data acquisition can be evaluated by the GUAVA Express® Pro software module and the obtained data can be exported as Excel-file for further evaluations. To check the accuracy of the device daily before starting a measurement, a GUAVA Check Kit, consisting of fluorescent beads is provided.

GUAVA ViaCount Assay

To determine the cell count and the viability, the GUAVA ViaCount Assay is performed. The ViaCount reagent contains two different DNA-staining dyes:

- LDS 751 staining all cells and detectable in red fluorescence channel (PM2)
- 7-AAD staining dead cells and detectable in yellow fluorescence channel (PM1)

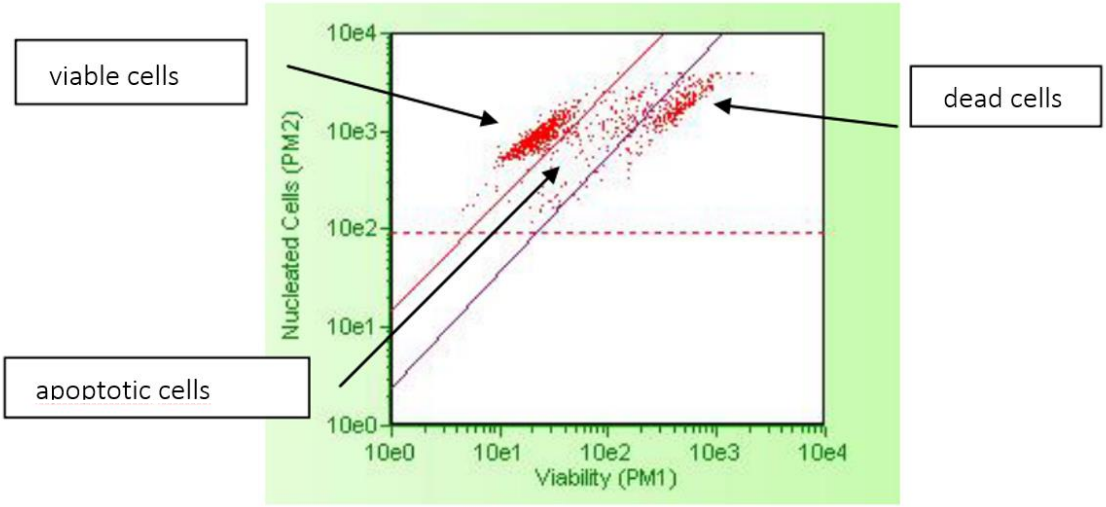

Figure 11: dotplot of a ViaCount measurement

In Figure 11 a dotplot of PM1 against PM3 is depicted, where every event represents a cell in viable, apoptotic, or dead state. Since the ViaCount Assay does not specifically stain apoptotic cells, it cannot be classified as apoptosis Assay.

GUAVA Nexin Assay

The GUAVA Nexin Assay is a specific apoptosis assay to distinguish between vital, early apoptosis, late apoptosis, and dead cells. A vital cell stores phosphatidylserine on the inside of the cell membrane. Once a cell turns into apoptotic state phosphatidylserine is transferred to the cell surface, where the reagent Annexin V can bind. In the Nexin-reagent Annexin V is conjugated to a fluorescent marker.

- Annexin-V-PE staining apoptotic cells and detectable in yellow fluorescence channel (PM1)
- 7-AAD staining dead cells and detectable in red fluorescence channel (PM2)

In Figure 12 the four different populations of cell can be identified and assigned to its actual vital or apoptotic state.

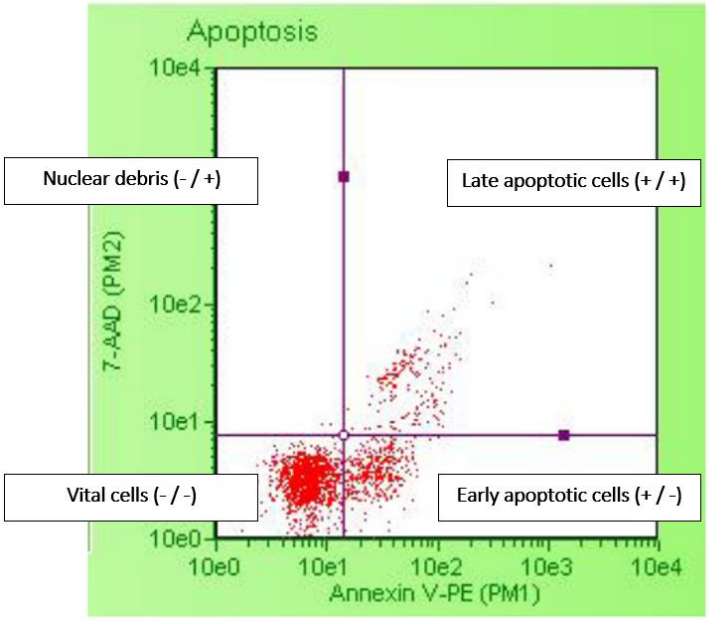

Figure 12: Dotplot of a Nexin Assay measurement

3.4 Aber FUTURA

For dielectric spectroscopy, an ABER Futura biomass monitoring System was chosen. The System consists of four major components.

First, there is the 25mm probe, which is placed into the solution and can be replaced for cleaning purposes. The probe is attached to an amplifier, the so called Futura, which is the main processing engine in the system. Up to four FUTURA probes can be connected to the Aber Hub, which facilitates the communication of the Probe with a PC. The PC runs a software called FUTURA Scada that enables to control the system. This software provides data collection for any number of Futura systems with an event timeline. It also supplies frequency scanning and directly calculates additional parameters like delta C, critical frequency, and Cole-Cole alpha. Data can be exported as .csv file for further processing.

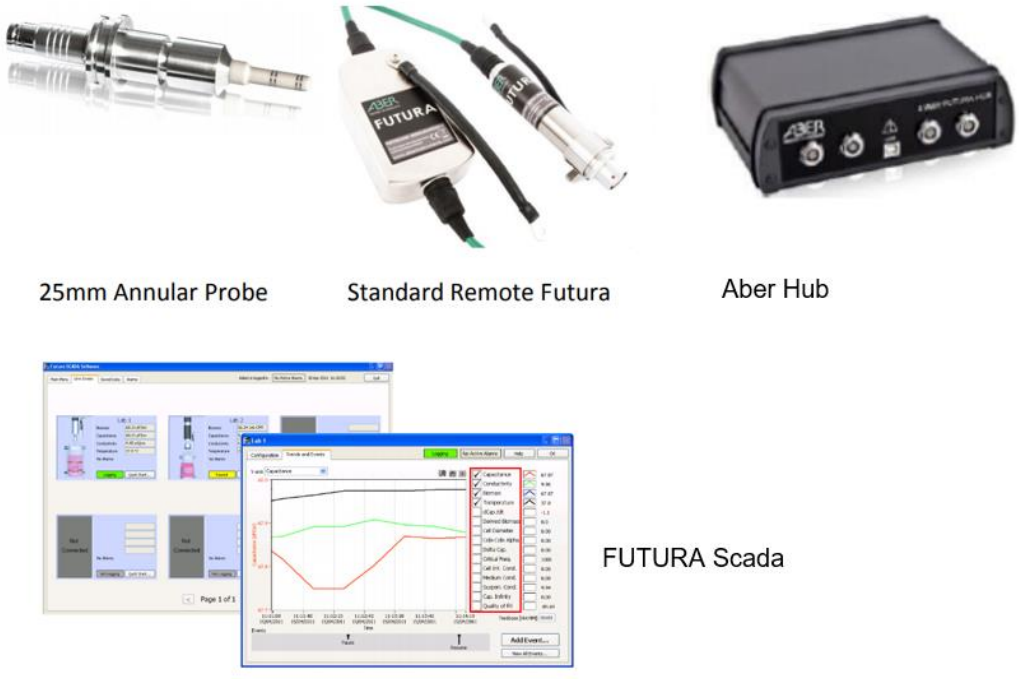

Figure 13: Components of the Aber Futura System

The probe is built into the bioreactor before the sterilization process is started. Once the medium is filled the Zero function is applied to tare the capacitance output of the instrument. The stability of the signal is monitored until inoculation. To start data logging cell culture mode is chosen in the Scada software.

This mode is optimized for cell suspensions of large cells and features a polarization correction that treats electrochemical effects at low frequencies, a noise correction of level 30. In this mode typically 0.1 pF/cm resolution on the instrument represents 10^5 cells/mL. The data are exported to a .csv file by adding no additional filter and including data in the interval of 30 minutes. The exported parameters are Capacitance at 50 kHz, 64 kHz, 82 kHz, 106 kHz, 136 kHz, 174 kHz, 224 kHz, 287 kHz, 368 kHz, 473 kHz, 580 kHz, 779 kHz, 1000 kHz, 1120 kHz, 1648 kHz, 2115 kHz, 2714 kHz, 3484 kHz, 4472 kHz, 5740 kHz, 7368 kHz, 9457 kHz, 12139 kHz, 15650 kHz, 20000 kHz, which are logarithmically distributed frequencies and the Cole-Cole parameters delta Cap, Alpha and critical frequency.

3.5 Raman spectroscopy

For the Raman spectra acquisition, a Kaiser Raman System is used. Therefore, four bio-Pro Raman Probes were attached to a multi-channel RamanRXN2 analyzer that is equipped with a 785 nm excitation laser system (Kaiser Optical Systems Inc., MI).

Before the first start of the bioreactor runs, intensity calibrations were performed using the routine calibration set recommended by the vendor. Before the insertion of the probes into the Bioreactors (20L and 30L, Biostat D DCU, Satorius), a spectra validation was performed using 70% Ethanol. Once the probes are inserted, the fermentation vessels were cleaned and sterilized according to company's internal standard operating procedures (SOPs). The four probes work independently from each other in a sequential way. Spectra acquisition was performed using the RunTime software with 75 scans of 10s each per resulting spectrum and a spectrum range between 100 to 3425 cm^{-1}. After the end of each batch, spectra acquisition was halted, and the spectra were exported as .spc files. For further processing, the spectrum range of 400 to 1800 cm^{-1} is extracted, because the rest of the spectrum covers mostly redundant information.

3.6 Statistics software

All statistical tasks occurring in this thesis are handled via R. R is an environment for statistical computing and graphics, that evolved from the statistical programming language S. As it is an open source, R became a popular tool for statistical computing and thus is highly advanced. RStudio is an integrated development environment that is in use as user interface. By the installation of several additional packages the capabilities of R are enriched. In Table 2 below some important packages used for this thesis are listed.

Table 2: List of some important R packages

Package name	usage
pls	multivariate regression methods
data.table	handling of large data
ggplot, ggplot2	declaratively creating graphics
prospectr	processing spectrometric data
hyperspec	way to work with hyperspectral data like .spc files

3.6.1 data preparation

Alignment

Once a batch is finished all data from the different on- and offline sources are adapted to a corporate data format. Subsequently it is possible to align all data to the closest common timepoint. Thereby a data table arises, that presents the online signals and the reference values, which are the limiting data in terms of observation densities. Also, the batch name is included. Thus, all available batches can be connected to one large data set.

Pretreatment

To increase the information content of especially spectral data, different pretreatments like mean centering, scaling, standard normal variation and first- or second derivation are applied to selected columns of the data frame.

For model creation, the data set is split into a training- and a test dataset to have the possibility to evaluate a model, which will be explained later. All observations of one or more batches are separated to a test-dataset.

3.6.2 linear and multivariate models

linear model

In the simple linear regression model a single numerical variable y, the so-called response is correlated by a single variable x, the so-called predictor. A mean function described by an intercept β_0 and a slope β_1 is created to enable the calculation of a response for new predictors.

multivariate model

For the regression of a wide dataset, partial least squares (PLS) models are chosen. Therefor all relevant parameters found in the data frame are predictors and get correlated to a response variable. The principle behind a PLS model is described in detail in chapter 1.4.2 .

The number of latent variables used for the final models was chosen by calculating the RMSEPs of one to twenty latent variables by cross validation. One component less than the point when the RMSEP stops improving was determined as optimal number of latent variables.

3.6.3 Evaluation

The evaluation of all created models is performed by batch wise cross validation. Therefore, the model, which is built based on the training data set gets applied on the test data set. The obtained response is compared to the original response value. MAE and RMSEPs are calculated to compare the quality of the models.

4 Results

4.1 Pretreatment of the dielectric frequency scan

In order to receive the most possible information from the spectral data of the dielectric frequency scans, the most effective data pretreatment must be figured out. Therefore, a set of spectral data from 26 Batches of the Development Process was submitted to different methods of pretreatment followed by fitting a PLS model that correlates VCD with the pretreated spectra. By performing a batch wise cross-validation and comparing RMSEPs, the best performing pretreatment can be chosen for further model construction. The RMSEPs of all versions of pretreated data are depicted in the bar chart below.

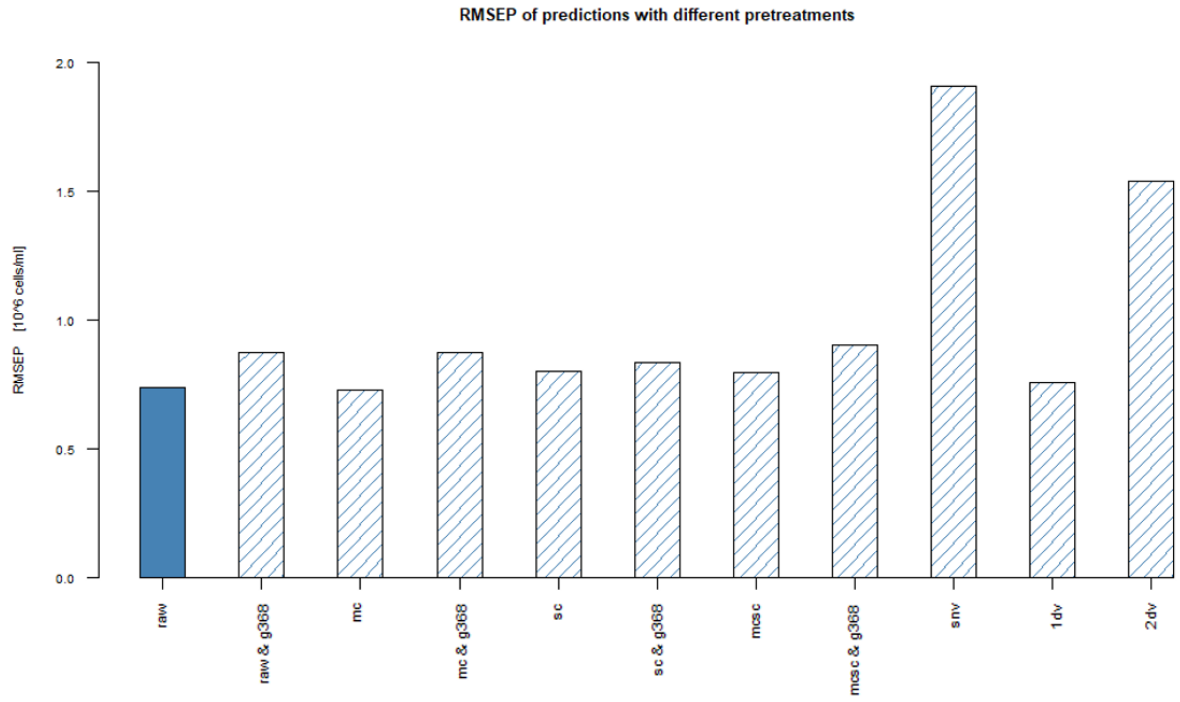

Figure 14: RMSEP of untreated frequency spectra (raw), frequency limited spectra (raw & g368), mean centered (mc), mean centered and frequency limited spectra (mc & g368), scaled spectra (sc), scaled and frequency limited spectra (sc & g368), mean centered and scaled data (mcsc), mcsc and frequency limited spectra (mcsc & g368), standard normal variated data (snv), first differentiation of Spectra (1dv) and second differentiation (2dv). The bar of the untreated model is blue.

Mean centering (mc), scaling (sc), standard normal variate (snv) and first or second differentiation (1dv, 2dv) were applied to the full data of the frequency scan and on a selection of capacitance data at frequencies above 368 kHz. Therefore, the limitation of the

capacitance data is examined, as there are probe specific distinctions at low frequencies due to polarization effects.

Only the RMSEP values of the model with the mean centered data and the first derivative data are smaller than the RMSEP of the untreated data. Considering the dynamics of the first derivative data model, this treatment does not support the prediction of the growth curve. The gap between mean centered and untreated data is not very high. Nevertheless, since this treatment copes the probe specific differences of capacitance at low frequencies, mean centering is taken as suitable pretreatment for dielectric frequency scan data.

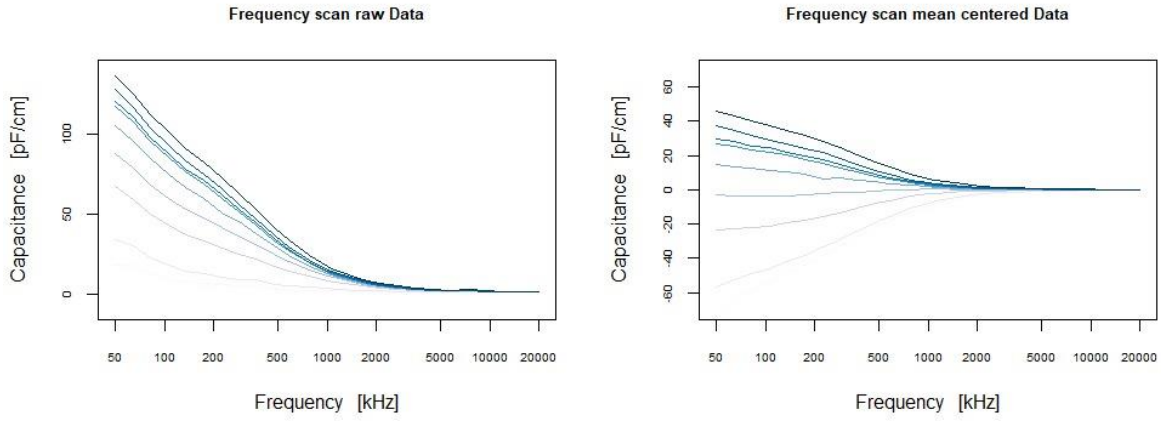

Figure 15: Plot of the raw (left) and mean centered (right) dielectric frequency scan data. Each line represents the Spectrum at a certain time point. The bright lines represent early time points whereas dark lines arise of data from the end of the process. Raw data are plotted on the left side. Mean centered data are plotted on the right. The scale of the X-Axis is logarithmic to emphasize the characteristics of a frequency scan.

Comparing the plots of raw and mean centered capacitance data in Figure 15, it is visible that the scale of capacitance per frequency at different time points don´t change, but the absolute values are shifted towards the negative direction. This effect eliminates the probe specific differences in the low frequencies and thus standardizes the input data of the seven different probes used in the development bioreactors.

4.2 Model fitting for the Development Process

4.2.1 VCD prediction

To determine the best model for predicting VCD, the same 26 development batches are used. A dataset with the mean centered frequency scan data, the three Cole-Cole parameters and the reference data of those batches are aligned. Several linear and multivariate models are constructed and applied to an exemplary chosen test batch.

First, the “state of the art” model, a simple linear regression model, later on mentioned as linear model (lm), correlating the VCD with the capacity of one single frequency (580kHz) is built. Next, each of the cole-cole parameters, namely the cole-cole Alpha (CCAlpha), the critical Frequency (cF) and the delta Capacitance (dCap) is separately correlated to the VCD in a lm. To correlate the VCD to all the frequency scan data, a PLS model was created (fs). Another PLS model uses all, the three Cole-Cole parameters and the frequency scan data in one model (fscc).

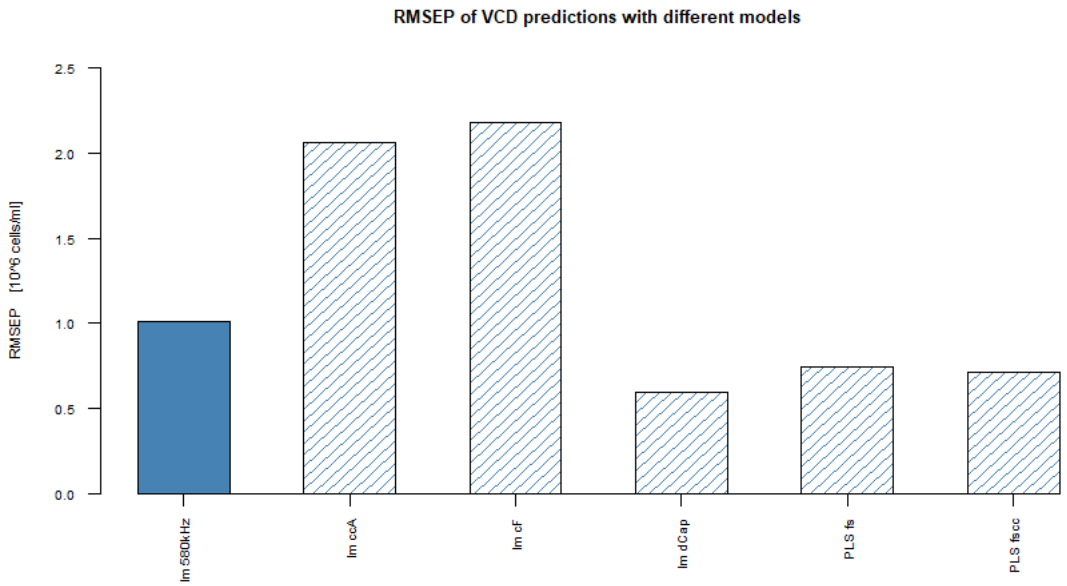

Figure 16: Plot of RMSEPs of the different liner (lm) and partial least squares (PLS) models for VCD prediction. The bar of the “state of the art” model is marked in blue.

Per cross validation, the RMSEP were calculated to enable a first assessment of the different models.

There are two models, the linear model of the Cole-Cole Alpha and the linear model with the

critical Frequency that end up with a significantly high RMSEP value, to be seen in Figure 16. Thus, those two models are no longer considered since there does not seem to be a good correlation with the VCD.

The other four models, namely lm 580kHz, lm dCap, PLS fs and PLS fscc are further evaluated in terms of dynamics.

Table 3: RMSEP values belonging to the chart presenting the different VCD models. The number of latent variables of the PLS models is listed in the bracket.

VCD Model	lm 580	lm ccA	lm cF	lm dCap	PLS fs (2)	PLS fscc (3)
RMSEP [10^6 cells/mL]	1,006	2,060	2,177	0,592	0,742	0,715

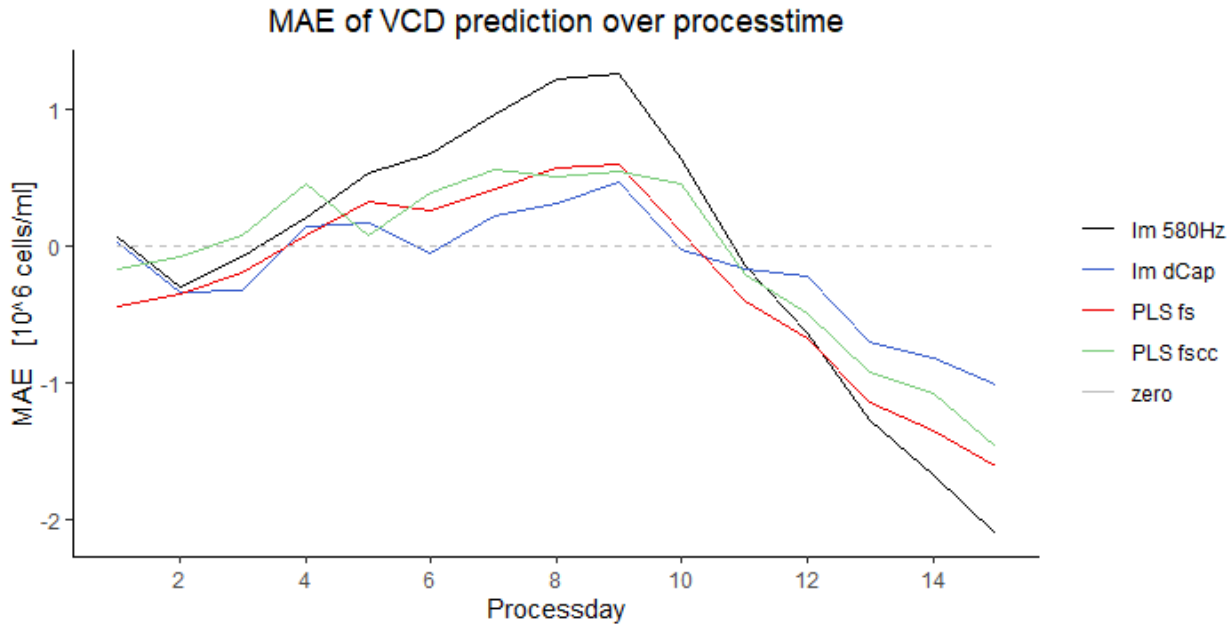

Figure 17: Plot of the mean average error (MAE) of VCD prediction by different models over process time. The grey dashed zero line indicates whether the model is underestimating (positive values) or overestimating (negative value) the VCD.

The MAE of prediction of several batches were calculated for each day and plotted in Figure 17 against the process time to get an insight of the model dynamics throughout the bioprocess. Values above the zero line are underestimations of the actual VCD, whereas values below the grey dashed line are overestimations. Until day five, there are no significant differences between the models, showing, that during the linear growth phase it is easy to predict the VCD. As soon as the curve flattens, all models start to underestimate the actual VCD. In death phase, during the last two to four days, the VCD is overestimated. This reveals the fact that it is hard to fit a model for VCD prediction that performs well during all stages in

bioprocess. Nevertheless, all tested models proof a better performance, than the "state of the art" model lm 580kHz, depicted in black. The best performing model is the lm with dCap, leaving the assumption that two frequencies would suffice for a strong improvement of VCD prediction. The PLS models show a very similar and good performance in the beginning of the process. Later, the addition of the Cole-Cole parameters have a visible influence on the model. The PLS fscc Model shows slightly better results than the PLS fs.

Table 4: MAPE values belonging to the different VCD models. The number of latent variables of the PLS models is listed in the bracket.

VCD Model	lm 580	lm ccA	lm cF	lm dCap	PLS fs (2)	PLS fscc (3)
MAPE [%]	15,43	57,40	59,46	11,92	11,87	12,39

Looking at Table 4, the mean absolute error is expressed as percentage error (MAPE) in relation to the actual measurement value, which allows another assessment of the goodness of fit of the prediction by the models. The values again point out, that lm dCap and PLS fs are equally strong models.

Looking at Figure 18, the plot shows observed and predicted values over process time for three batches confirming that especially during the first phase of the process (growth phase) there is close agreement between predicted and the observed values.

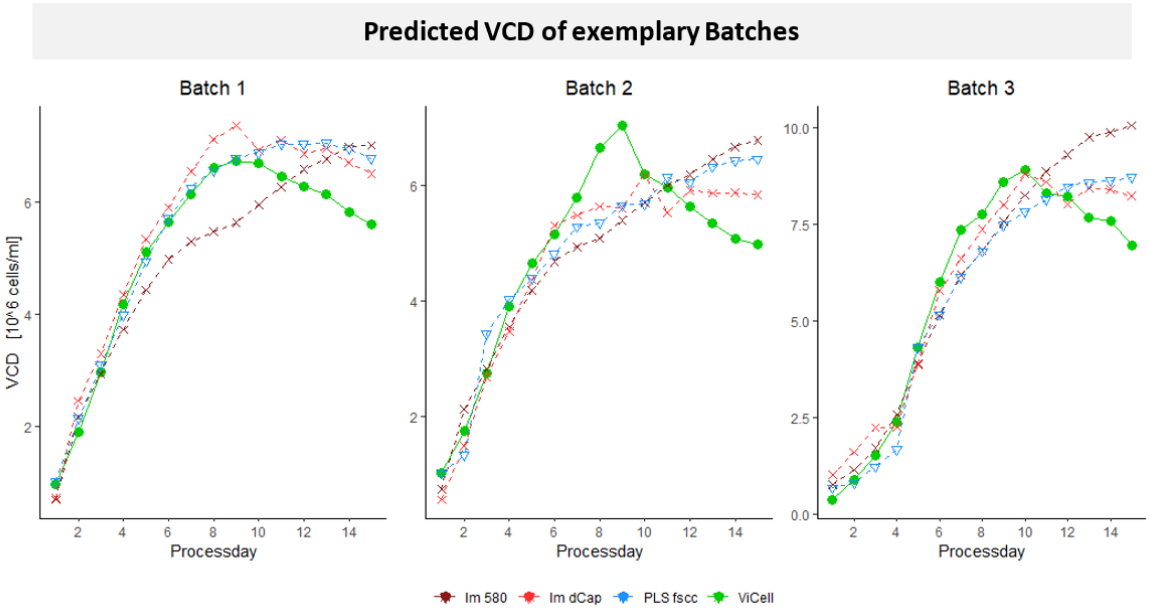

Figure 18: Plot of VCD from three different development batches. The green curve presents the reference values, the brown graph presents the "state of the art" model, the red and blue graph present the new models.

To get an idea of the noise obtained by the online data, the PLS fscc model was applied to the mean centered online data of the frequency scan in Figure 19. As expected, the noise is within an acceptable range, so there is no need to use a filter to even the output curve.

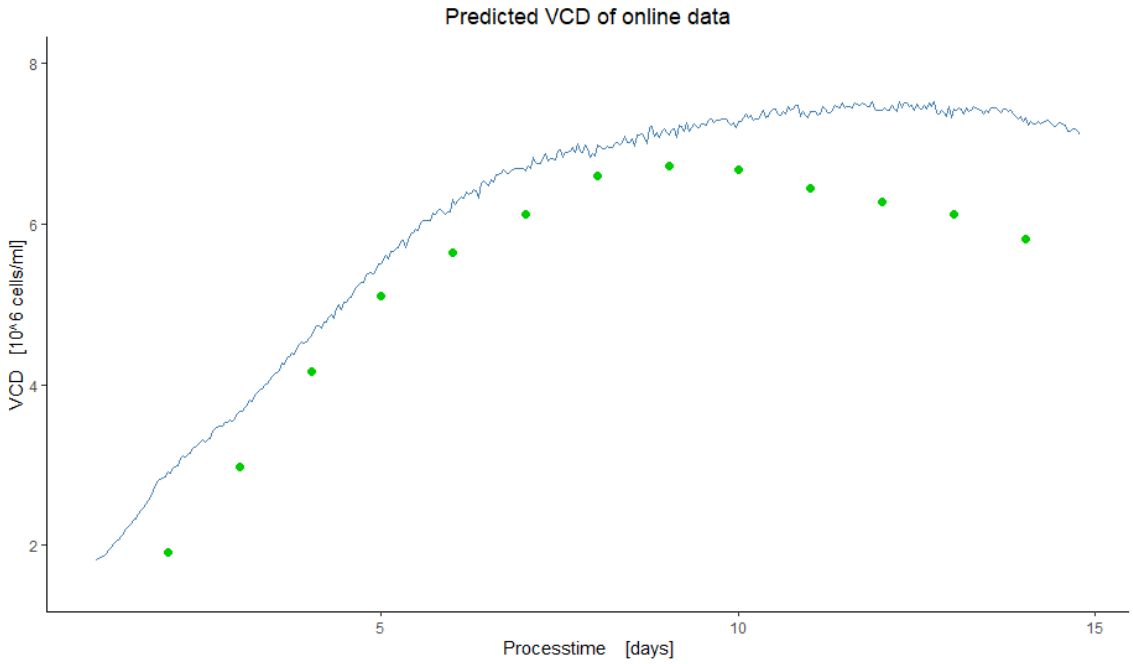

Figure 19: Application of the PLS fscc model on the online data set (blue) of a random development batch. The green points indicate the corresponding Vicell measuring data.

4.2.2 Viability prediction

Again, the 26 development batches were utilized, to create models for viability (VIA) prediction. Linear models (lm) with single frequency (580 kHz), Cole-Cole Alpha (ccAlpha), critical Frequency (cF) and the delta Capacitance (dCap) as well as multivariate PLS models of frequency scan (fs) data and frequency scan data including Cole-Cole parameter (fscc) were correlated to the percental viability, measured with the ViCell technology.

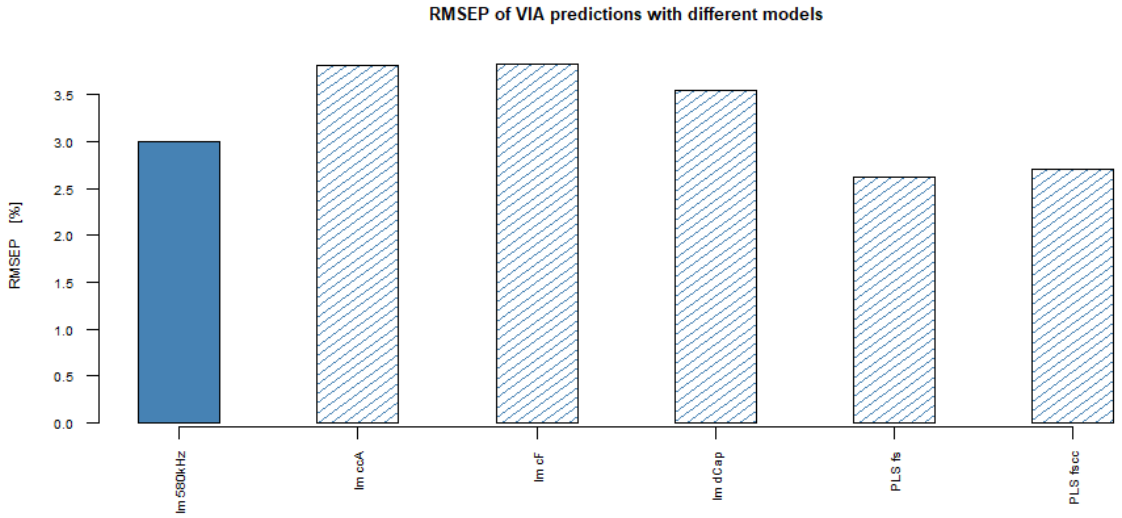

Figure 20: Plot of RMSEPs of the different linear (lm) and partial least squares (PLS) models for viability (VIA) prediction. The bar of the "state of the art" model is marked in blue.

Unlike the VCD prediction, there seems to be significant information in the frequency scan data, that cannot be pooled in a single parameter. Whereas a linear correlation with the dCap worked well for VCD prediction, viability needs more specific information from the capacitance data at different frequencies that cannot be summarized by one of the Cole-Cole parameters. The PLS models perform better than all the linear models. The PLS fs and PLS fscc seem to predict similarly alike. Thus, the MAE of these two models, the lm dCap and the lm 580 kHz were plotted over process time, to view the dynamic behavior of these models.

Table 5: RMSEP values belonging to the chart presenting the different VIA models. The number of latent variables of the PLS models is listed in the bracket.

VIA Model	lm 580	lm ccA	lm cF	lm dCap	PLS fs (6)	PLS fscc (8)
RMSEP [%]	2,995	3,814	3,822	3,549	2,617	2,709

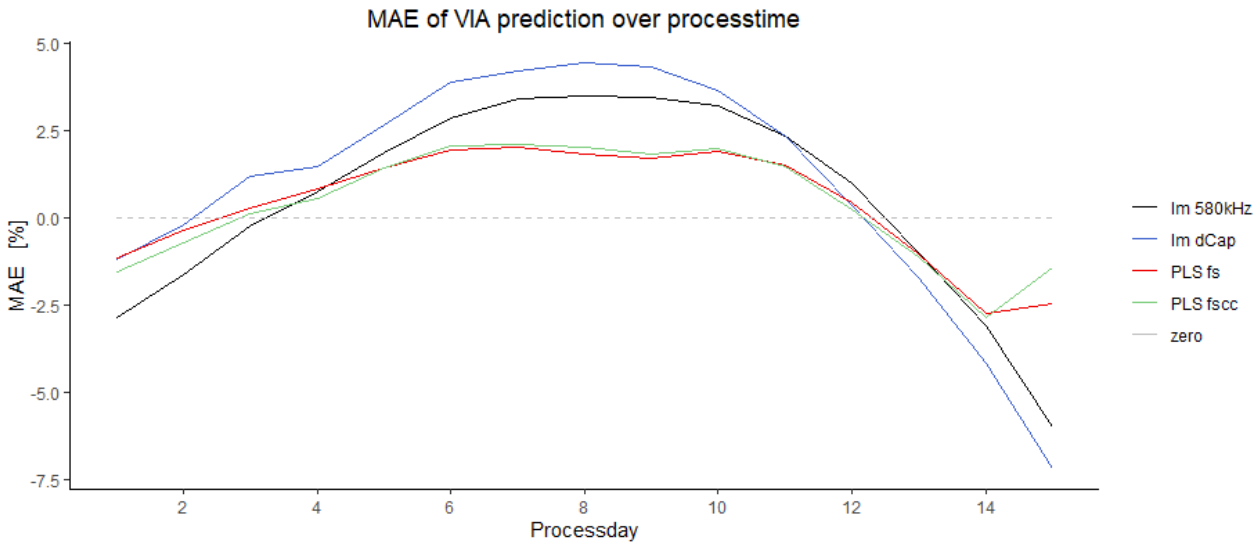

Figure 21: A plot presenting the mean average error (MAE) of VIA prediction by different models over process time. The grey dashed zero line indicates whether the model is underestimating (positive values) or overestimating (negative value) the viability measured by the ViCell technology.

Looking at the MAE plot in Figure 21, the error range of the linear models is very large, reaching from -7.5% to 5.0 %. The MAE curves for PLS models are still big, but closer to the zero line throughout the process. It is seen, that in the end of the Process the PLS models can mimic the trend again.

Table 6: MAPE values belonging to the different VIA models. The number of latent variables of the PLS models is listed in the bracket.

VIA Model	lm 580	lm ccA	lm cF	lm dCap	PLS fs (6)	PLS fscc (8)
MAE [%]	3,03	3,30	3,23	3,27	2,32	2,43

In Table 6 the mean absolute percentage Error is presented as a measure for the goodness of fit.

In the next Figure, the same four models, are applied to four different exemplary batches of the dataset and plotted over process time. For reference, the ViCell viability measurement is printed in green. To be noticed, the prediction quality differs a lot within the four batches. In batch one and four there is a quite strong correlation between the frequency scan data and the viability of the cell culture, whereas prediction for batch two and three is hardly able to follow the viability profile of the reference. It can also be seen that some of the calculated

viability values are greater than 100 %, which cannot happen in real life and naturally is an artefact of the model.

It must be mentioned that all the development batches have remarkably high viability until the end of the process. It is not common, that a process is ended at a viability of more than 85 %. Thus, the range of viability change in the training data set for model creation is little as well. It is supposed, that in this instance the PLS models could perform remarkably better if the training dataset would exhibit more variability in terms of viability.

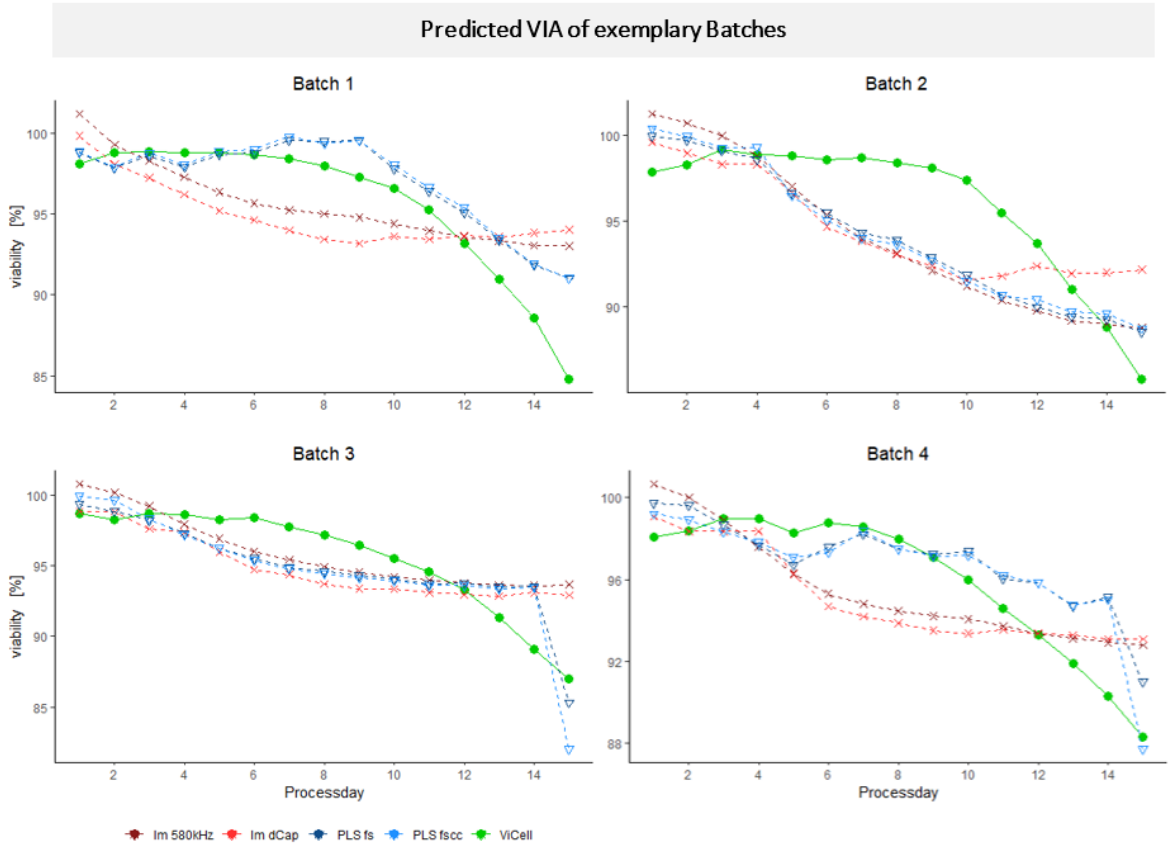

Figure 22: Plot of VIA from four different development batches. The green curve presents the reference values, the brown graph presents the "state of the art" model, the red (linear) and blue (multivariate) graphs present the new models. A strong divergence between viability predictions by the same models in different batches is visible.

4.2.3 TCD prediction

Basically, TCD is the result from dividing VCD by viability. As there already are models for predicting those two parameters, the TCD could be calculated from those values. Since the viability does not deliver satisfyingly accurate results investigations have been done to create models for predicting the TCD separately.

Executing the same approach as described earlier, the mean centered data of the 26 development batches were used to create the three linear models (lm 580kHz, lm ccAlpha, lm cF) and the two multivariate models (PLS fs, PLS fscc) for predicting the TCD.

Table 7: RMSEP values belonging to the chart presenting the different TCD models. The number of latent variables of the PLS models is listed in the bracket.

TCD Model	lm 580	lm ccA	lm cF	lm dCap	PLS fs (4)	PLS fscc (2)
RMSEP [10^6 cells/mL]	0.902	1.993	2.272	0.651	0.901	0.655

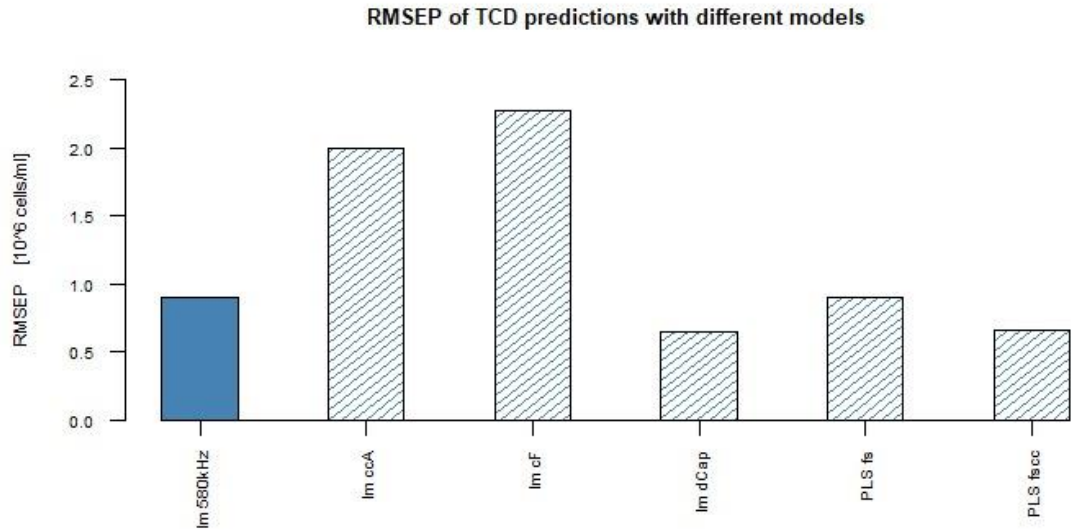

Figure 23: Plot of RMSEPs of the different linear (lm) and partial least squares (PLS) models for TCD prediction. The bar of the "state of the art" model is marked in blue.

Looking at the models, the TCD models behave in analogy to the VCD prediction. The linear correlation of Cole-Cole Alpha and the critical frequency produce extremely high RMSEPs, whereas the lm dCap seems to improve the quality compared to the "state of the art" model correlating a single frequency (lm 580 kHz). Likewise, smaller RMSEP values arise from the lm dCap and the PLS fscc.

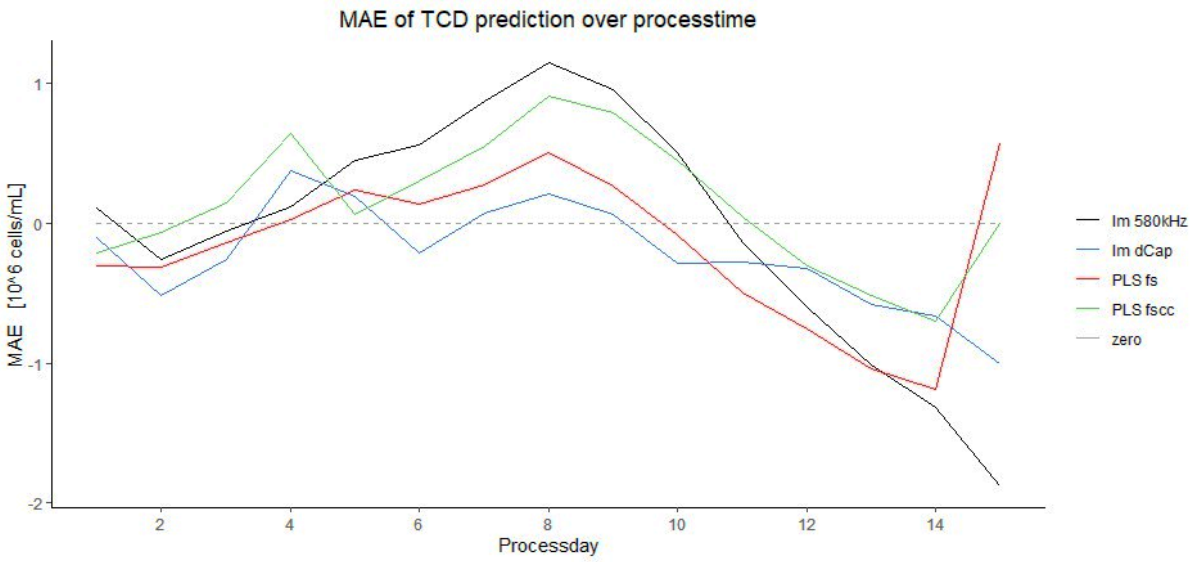

Figure 24: A plot presenting the mean average error (MAE) of TCD prediction by different models over process time. The grey dashed zero line indicates whether the model is underestimating (positive values) or overestimating (negative value) the TCD measured by ViCell.

To distinguish the dynamic manner of those models, the MAE is calculated and plotted over process time. Again, there is similarly good performance of all models in the early stage of the process. The lm dCap sticks out with being close to the zero-line throughout the whole process. Also, the PLS fscc shows good performance in the late stage of the process.

Applying these models on four exemplary batches to be seen in Figure 25, it can be recognized, that the prediction is difficult in the death phase of the cultivation. For batch two, there seems to be an outlier in the end of the process, whereas in batch one and four the PLS fscc model plotted in light blue shows consistently good TCD prediction. Overall, the lm dCap, printed in light red, seems to deliver the most constant and nicely fitting TCD prediction.

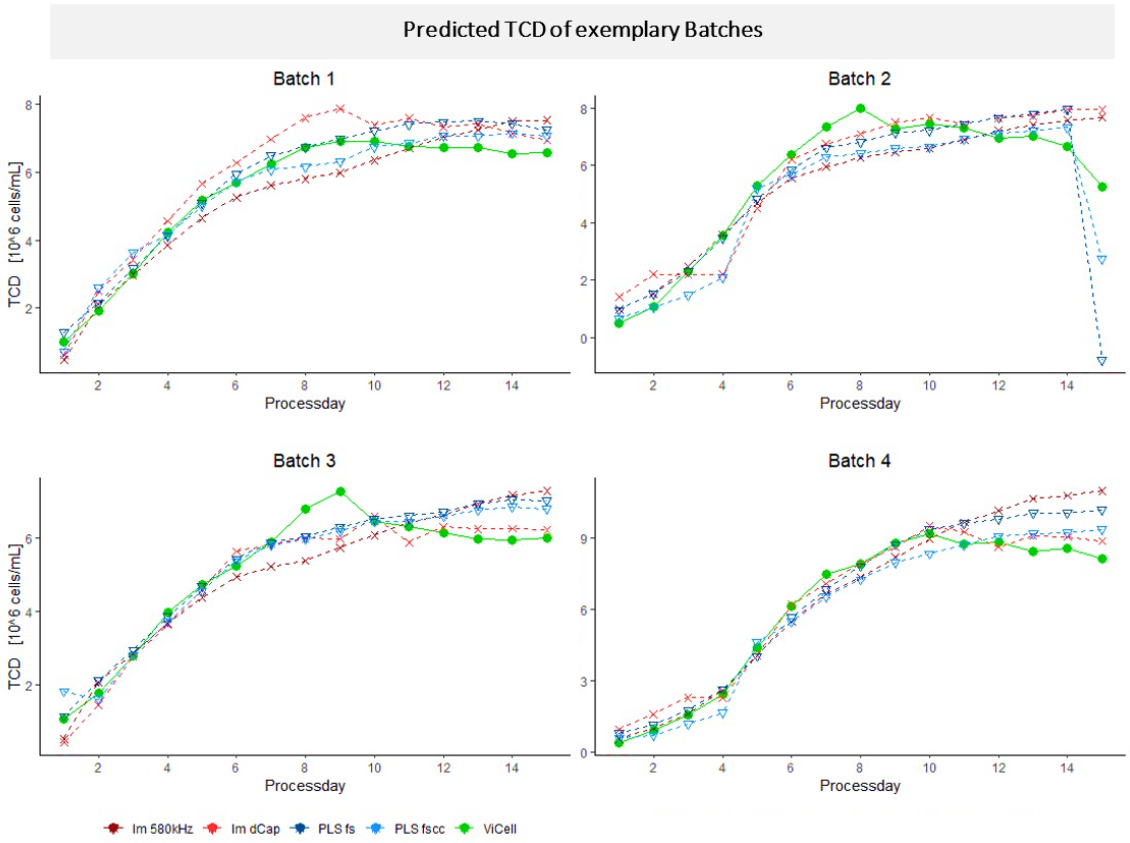

Figure 25: Plot of TCD from four different development batches. The green curve presents the reference values, the brown graph presents the "state of the art" model, the red (linear) and blue (multivariate) graphs present the new models. A good correlation between TCD predictions is visible, but it is recognizable, that the TCD prediction in during the death phase of the cultivation is hard to be captured by the prediction models.

4.2.4 Average diameter prediction

Using the same 26 mean centered frequency scan data sets of the development batches, several linear (lm 580kHz, lm ccA, lm cF, lm dCap) and multivariant models (PLS fs, PLS fscc) were created to predict the average diameter of the cell culture throughout the process. Getting a first idea for the fit of those models, RMSEPs produced by batchwise cross validation are compared.

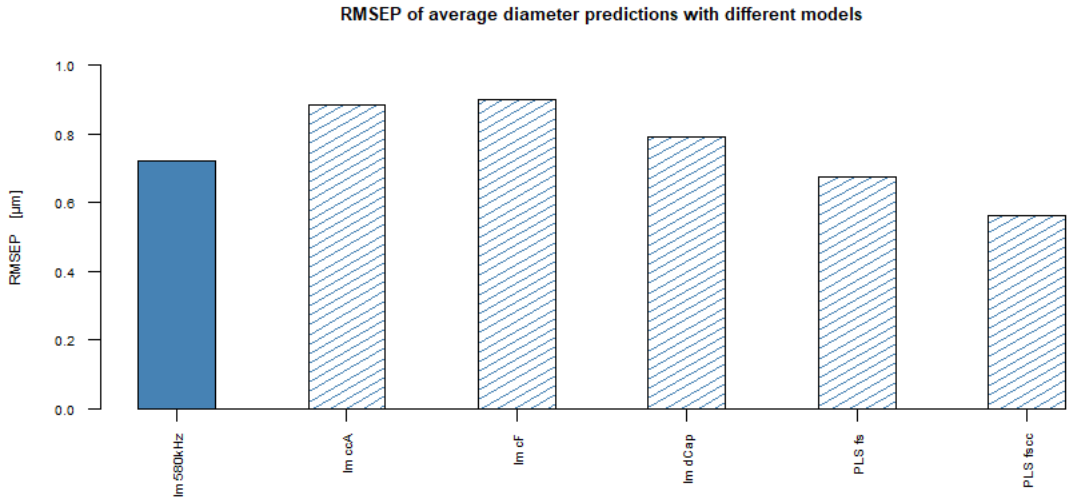

Figure 26: Plot of RMSEPs of the different linear (lm) and partial least squares (PLS) models for average DIA prediction. The bar of the "state of the art" model is marked in blue.

Guided by Cannizzaro et. al we would expect a strong correlation of the Cole-Cole parameter critical Frequency with the average cell diameter, as in this study large differences occur at intermediate frequencies. Thus, a decrease in cell size will cause the characteristic frequency to shift to a higher value because the cells are polarized faster. Conversely, for larger cells the characteristic frequency will be lower. [13] Nevertheless, this assumption was not confirmed by this study, as the RMSEP is the highest of all six tested models. Once again, the combined PLS fscc has the most promising RMSEP value of all models for average diameter prediction.

Table 8: RMSEP values belonging to the chart presenting the different average DIA models. The number of latent variables of the PLS models is listed in the bracket.

DIA Model	lm 580	lm ccA	lm cF	lm dCap	PLS fs (6)	PLS fscc (4)
RMSEP [µm]	0.720	0.885	0.898	0.790	0.676	0.561

To prove the dynamic fit of these models, again the MAE is calculated and plotted over process time in Figure 27. Except of the PLS fscc model, all the models depicted in the Figure below show similarly large prediction errors. The error line of the PLS fscc model (green) is the only one, being slightly closer to the zeroline, indicating a stronger correlation with the average diameter value measured by ViCell. Nevertheless, all the models produce high error values.

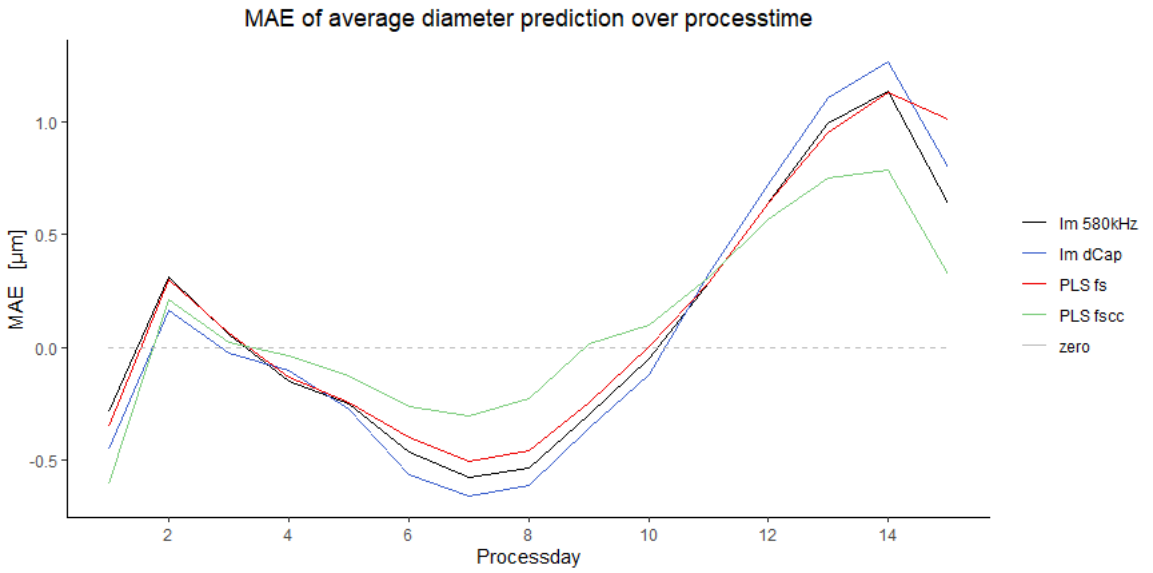

Figure 27: A plot presenting the mean average error (MAE) of average diameter prediction by different models over process time. The grey dashed zero line indicates whether the model is underestimating (positive values) or overestimating (negative value) the average diameter measured by ViCell.

Considering the graphs in Figure 28 which present the average diameter values produced by applying those models on four different, exemplary chosen batches it is obvious, that the linear models, as well as the PLS fs model, cannot at all predict the average diameter in any of the test batches. Only the PLS fscc model, shown in light blue, reflects the curve shape in understated form in batch number two and three. Even in these two test batches, the correlation is too week to predict the measured average diameter of the cell culture adequately well.

All in all, the prediction of average diameter by any of the tested models is not satisfying and thus dielectric capacitance data cannot be used to estimate this parameter.

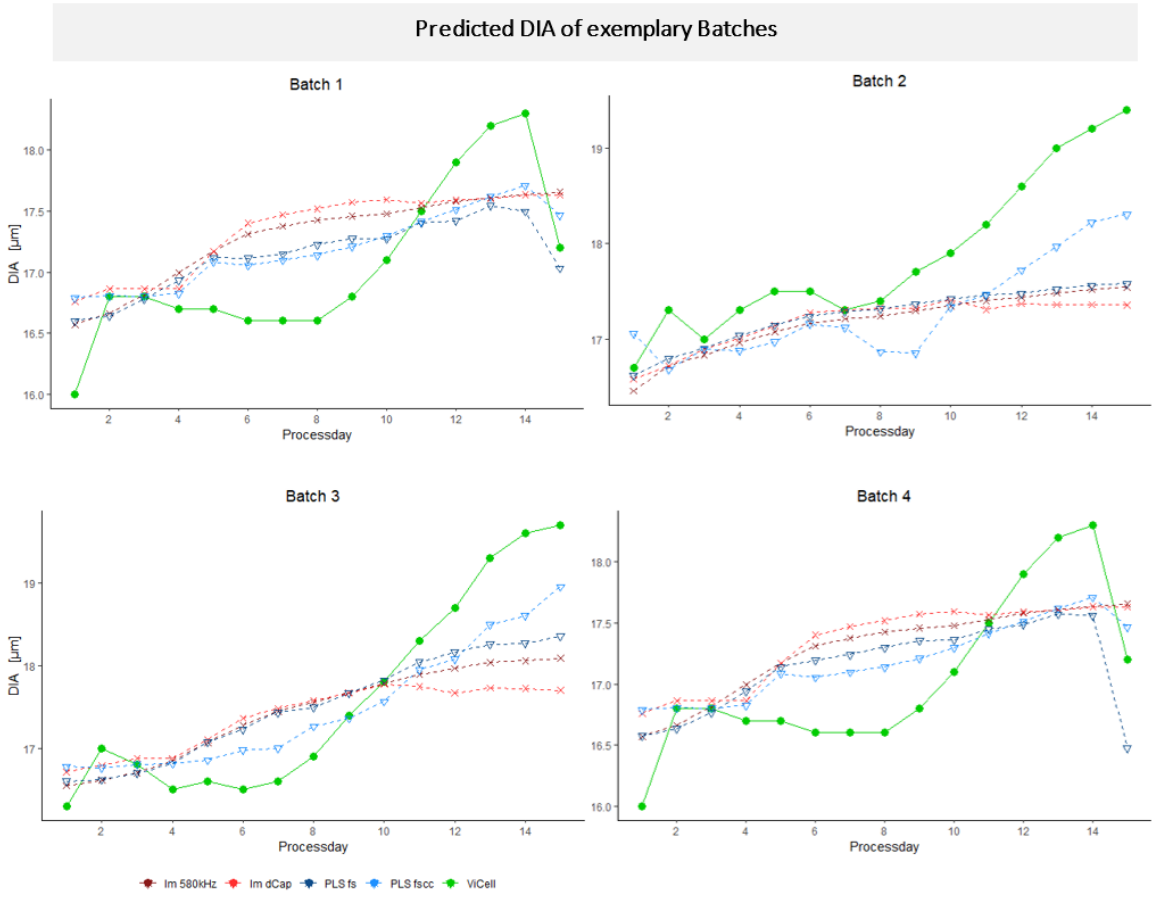

Figure 28: Plot of DIA from four different development batches. The green curve presents the reference values, the brown graph presents the "state of the art" model, the red (linear) and blue (multivariate) graphs present the new models. A week correlation between DIA predictions is visible, but there is no satisfying DIA prediction model within.

4.2.5 Average circularity prediction

The last parameter provided by ViCell, namely the average circularity (CIRC), will be correlated to the 26 development batches in the same way as described earlier. Linear (lm 580, lm ccA, lm cF, lm dCap) and multivariate (PLS fs and PLS fscc) are created and compared by calculation of the RMSEP values. All the CIRC models produce a similarly high RMSEP, mentioning, that the PLS fs delivers the best results in this case.

Table 9: RMSEP values belonging to the chart presenting the different average CIRC models. The number of latent variables of the PLS models is listed in the bracket.

CIRC Model	lm 580	lm ccA	lm cF	lm dCap	PLS fs (2)	PLS fscc (3)
RMSEP	0.022	0.020	0.026	0.021	0.017	0.018

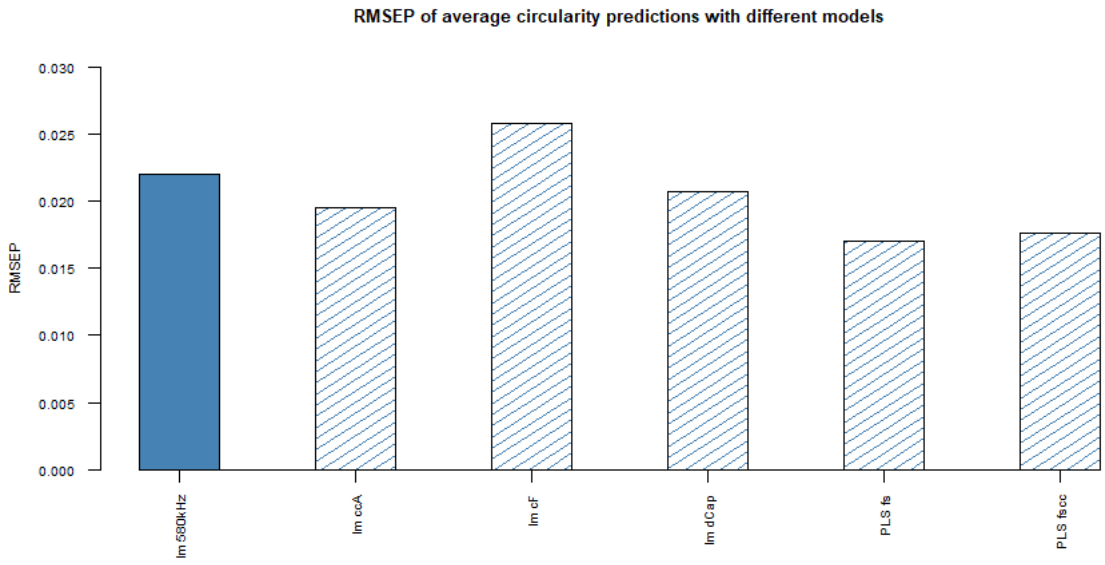

Figure 29: Plot of RMSEPs of the different linear (lm) and partial least squares (PLS) models for average CIRC prediction. The bar of the "state of the art" model is marked in blue.

Calculating the MAE and plotting it over the process time in Figure 30 enables an insight to the dynamic performance of the models. Even though all models have curves strongly divergent from the zero line, the PLS fs (red) and the PLS fscc (green) tend to be closer to the reference during the late stage.

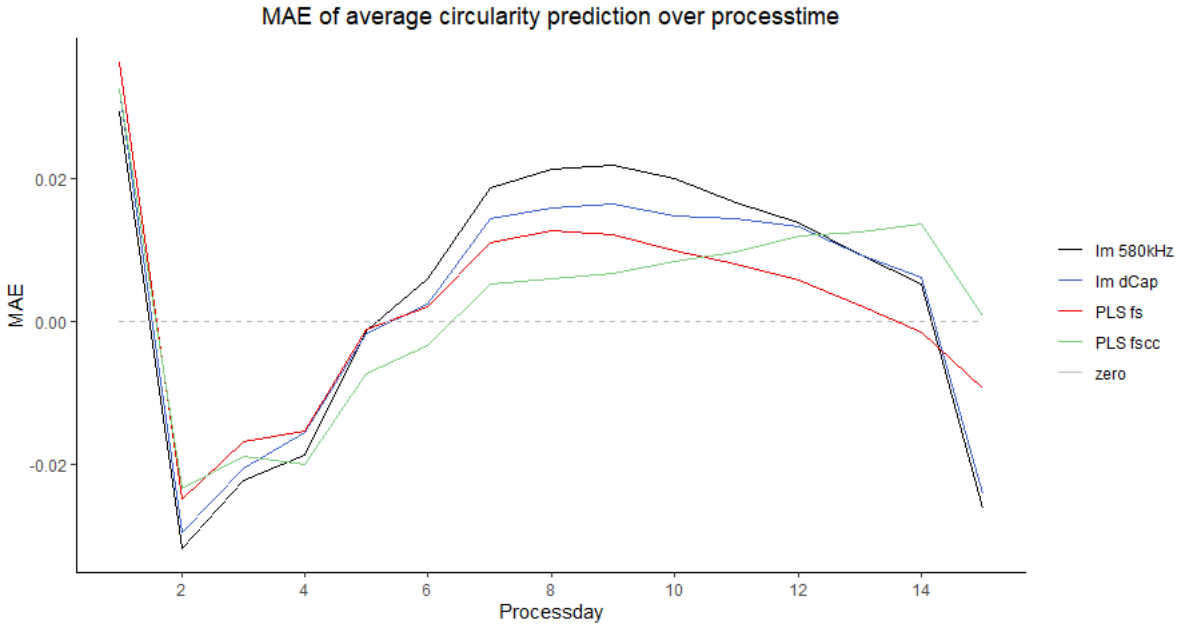

Figure 30: A plot presenting the mean average error (MAE) of average CIRC prediction by different models over process time. The grey dashed zero line indicates whether the model is underestimating (positive values) or overestimating (negative value) the average CIRC measured by ViCell.

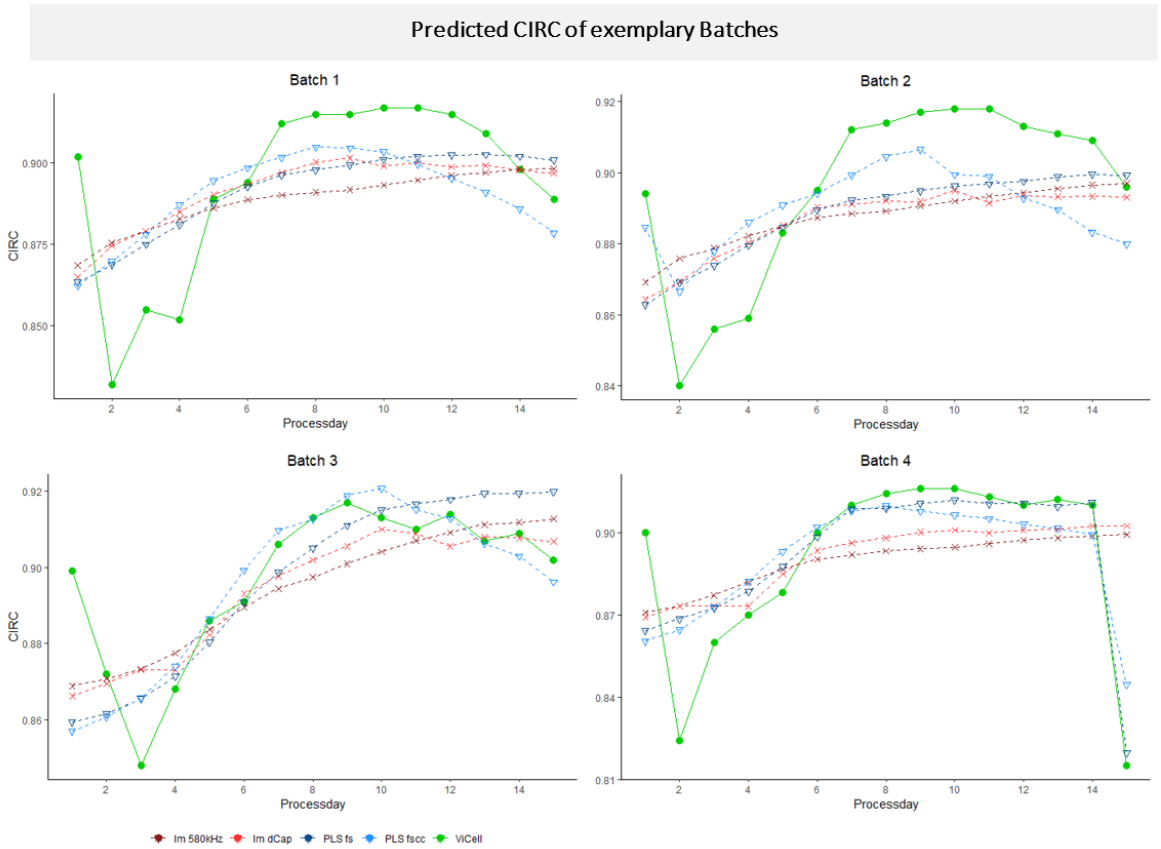

Figure 31: Plot of DIA from four different development batches. The green curve presents the reference values, the brown graph presents the "state of the art" model, the red (linear) and blue (multivariate) graphs present the new models. A week correlation between DIA predictions is visible, but there is no satisfying DIA prediction model within.

Looking at the reference values, printed in green in Figure 31, there seems to be a high circularity straight after inoculation of the preculture to the main stage, followed by a massive drop within the first days. Almost all the prediction models are incapable of predicting this phenomenon. Anyways, in the prediction of batch three and four, the prediction of circularity by the PLS fscc model results in impressive results after this starting phase. For the batches one and two it can be said, that the PLS fscc is the strongest but still not convincing.

Analogue to the prediction of the average diameter, a certain relationship between the dielectric frequency scan data and the average circularity is captured. Since the prediction does not work for all batches, the quality of the PLS fscc model is not satisfying.

4.3 Inclusion of Raman Signal

Even if the dielectric frequency scan enables an enhancement of biomass prediction models, there is still much room for improvement in terms of predicting all phases of the cell culture adequately well. A strategy to approach more accurate prediction dynamics, especially during stationery and death phases of the cell culture, is, to include the signal of the Raman probe to the multivariate models. Since Raman probes have the ability of capturing properties of other components within the fermentation brew than active cell membrane like metabolites and the chemical environment of cells, there is an assumption of improving the prediction during death phase when the data of Raman spectroscopy are used simultaneously to the capacitance data.

4.3.1 VCD prediction

A dataset providing dielectric frequency scan data including Cole-Cole-parameters, raman spectra and referential ViCell data of eleven batches is used to create new prediction models for the VCD in a bioprocess. These data were generated by the development process.

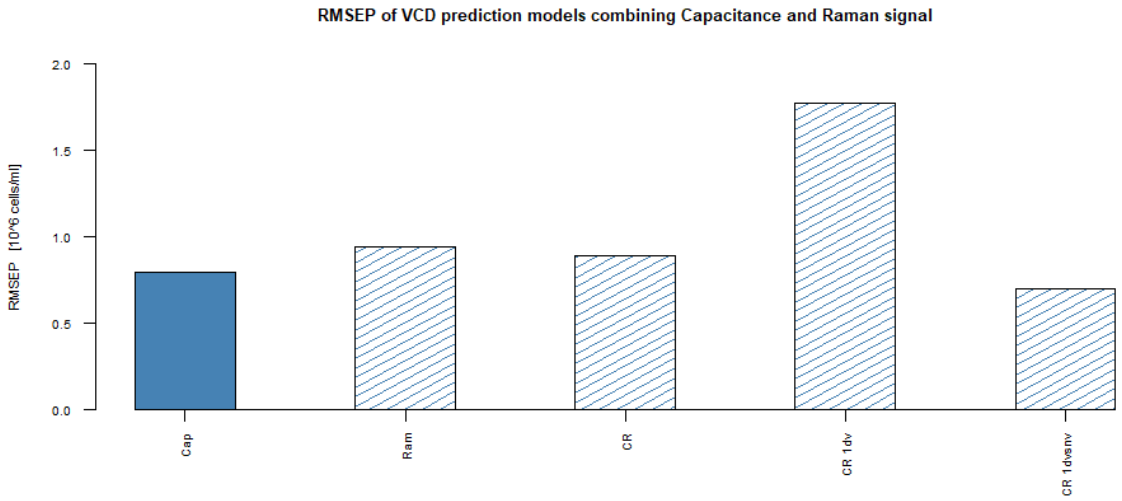

Figure 32: Plot of RMSEPs of the different PLS models for VCD prediction, comparing different input datasets. The bar marked in blue represents the "PLS fscc" model of the previous chapter.

As the PLS fscc, a multivariate model correlating the mean centered capacitance data and the Cole-Cole parameters of the dielectric frequency scan with the VCD measured by ViCell, was the best performing PLS model in the earlier investigation, this dataset is used as base for including further data. In this chapter, a set of five different PLS models is created and compared by cross validation.

In this Comparison, the PLS model with the frequency scan and the Cole-Cole parameters is called (Cap). Another PLS model correlates the raw Raman spectra to the VCD (Ram). Next, the dataset of Cap and Ram is combined in one model (CR). Two more models were tested, in which the raman spectra were submitted to different pretreatments, namely the first derivation (CR 1dv) and the first derivation with standard normal variation (CR 1dvsnv).

Table 10: RMSEP values belonging to the chart presenting the different VCD models by usage of capacitance and raman data. The number of latent variables of the PLS models is listed in the bracket.

VCD Model	Cap (2)	Ram (13)	CR (15)	CR 1dv (6)	CR 1dvsnv (3)
RMSEP [10^6 cells/mL]	0.792	0.937	0.887	1.774	0.699

For a first impression, the RMSEP values are calculated by cross validation. The Cap model and the CR 1dvsnv model seems to perform best. The RMSEP bar of the CR 1dv is about double as high as the others. Thus, this model is not further investigated. Since it is of interest if the models can predict the VCD during steady and late stage of the fermentation, the MAE is calculated and plotted over time again, to be seen in Figure 33. Those two models that had a low RMSEP ended up in crossing the zero line for two times, whereas the other models (CR in blue and Ram in green) stay close to the zero line, implying a rather consistent prediction offset over time. This result is promising, as it indicates that these models can at least reproduce the dynamics of the VCD of the cell culture at any stage. The Cap and the CR 1dvsnv model have big absolute errors in the late stage.

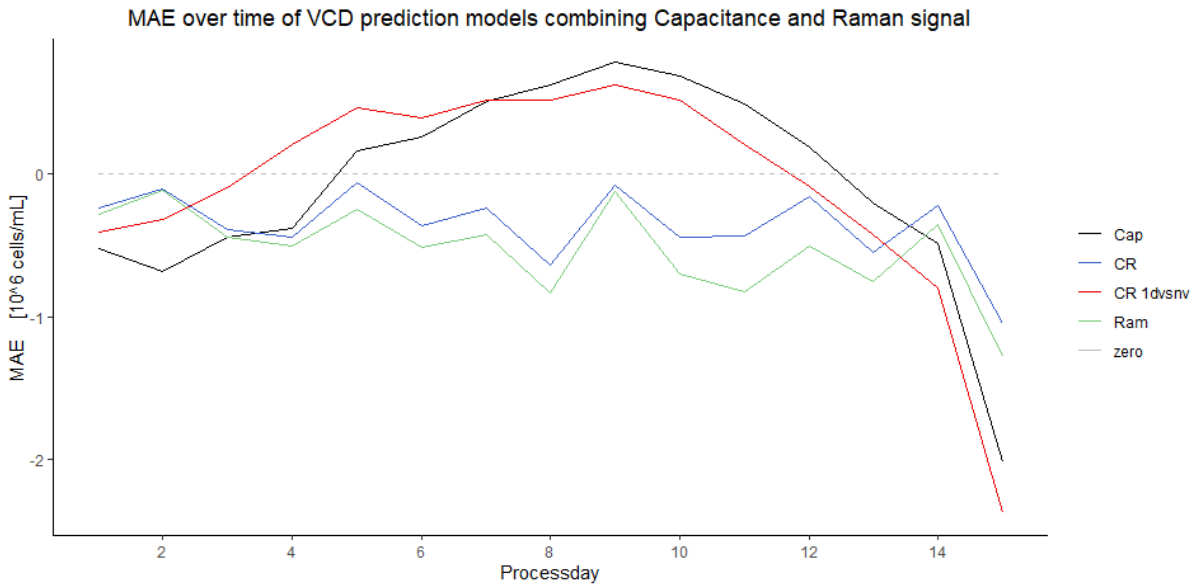

Figure 33: A plot presenting the mean average error (MAE) of VCD prediction by different models over process time. The grey dashed zero line indicates whether the model is underestimating (positive values) or overestimating (negative value) the VCD measured by ViCell.

Table 11: MAPE values belonging to the different VCD models. The number of latent variables of the PLS models is listed in the bracket.

VCD Model	Cap (2)	Ram (13)	CR (15)	CR 1dv (6)	CR 1dvsnv (3)
MAPE [%]	17.80	14.66	10.75	26.33	18.72

The MAPES stated in Table 11 present another value to measure the goodness of fit of the prediction models.

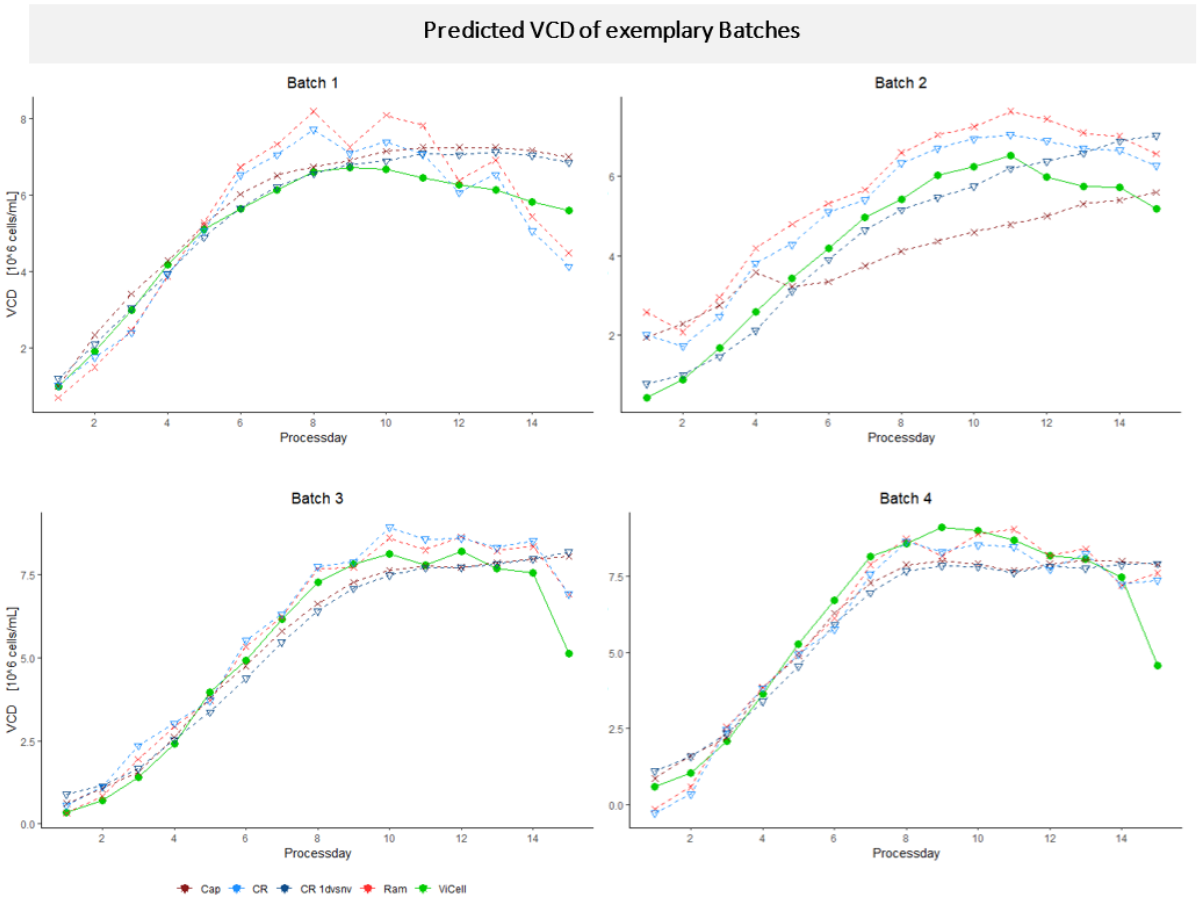

Figure 34: Plot of VCD from four exemplary chosen development batches. The green curve presents the reference values, the brown graph presents the best model created from capacitance data only, the red and blue graphs present the new models. A good correlation with VCD predictions throughout the whole process is visible.

All the presented models perform better than the linear VCD models from the prior chapter. Nevertheless, the best model arising in this comparison is CR. Applying these models to some exemplary batches some processes appear very well predicted (batch 3 and 4), whereas some inconsistent deviations appear in other processes (batch 1 and 2), but still, the trend is well predicted by the CR model. Looking at the predictions of all available batches (data not shown), it can be concluded, that less than a third of the prediction batches show such abnormalities, meaning, that the overall prediction capability still lacks a certain robustness.

4.3.2 Viability prediction

The prediction of another important factor in a bioprocess, the viability, could benefit from adding Raman spectra to the capacitance data as well. In the following chart the same models as in the previous chapter are compared for viability prediction. The mean centered

frequency scan model (Cap) representing the best PLS model of the dielectric dataset, the raw Raman spectra (Ram) and the combined PLS model (CR), the model with pretreated Raman signal (CR 1dv and CV 1dvsnv) are compared by RMSEP values (Figure 35).

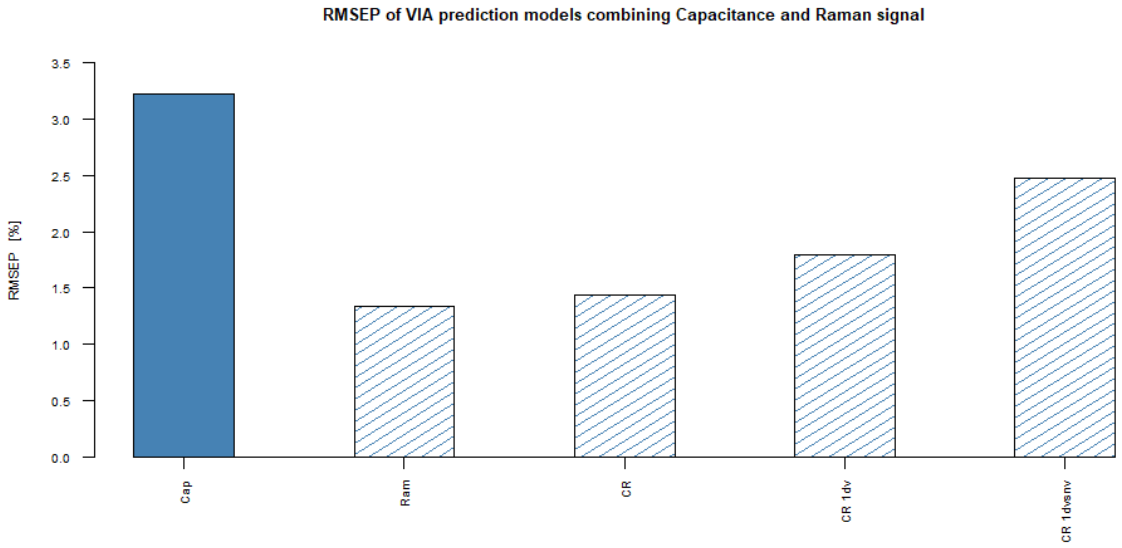

Figure 35: Plot of RMSEPs of the different partial least squares models for VIA prediction, comparing different input datasets. The bar marked in blue represents the "PLS fscc" model of the previous chapter.

Reviewing the bar chart, the addition of the Raman signal to the capacitance data seem to have a massive impact on the prediction quality, as all of the models containing Raman data have a significant lower RMSEP value than the Cap model. In this case, the raw Raman signal seems to predict the viability within the cell culture best.

Table 12: RMSEP values belonging to the chart presenting the different VIA models by usage of capacitance and Raman data. The number of latent variables of the PLS models is listed in the bracket.

VIA Model	Cap (3)	Ram (5)	CR (5)	CR 1dv (5)	CR 1dvsnv (4)
RMSEP [%]	3.223	1.330	1.433	1.797	2.477

To evaluate the dynamic fit of the models, the MAE is calculated and plotted over process time. There are prominently big errors of the Cap (black) models crossing the zero line several times. Two models (Ram and CR) tend to cling to the zero line throughout most of the process. Remarkably, the viability prediction with these models turns out being even more precise in the late stage of the process.

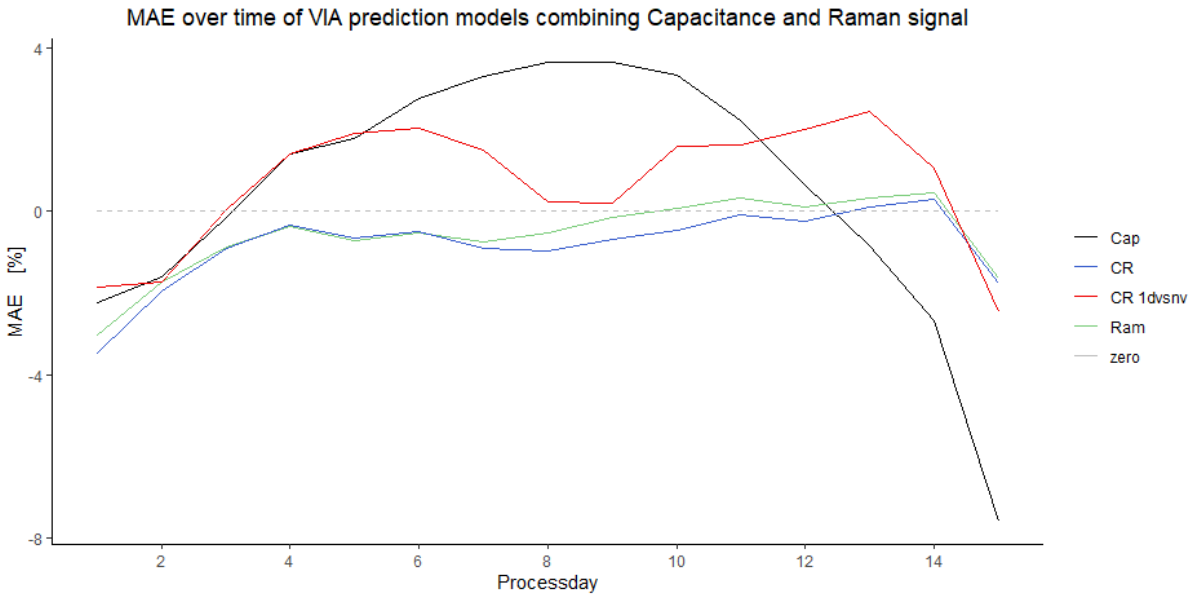

Figure 36: A plot presenting the mean average error (MAE) of VIA prediction by different models over process time. The grey dashed zero line indicates whether the model is underestimating (positive values) or overestimating (negative value) the VIA measured by ViCell.

In Table 13 the MAPE is listed for the different models to get another idea of the goodness of fit.

Table 13: MAPE values belonging to the different VIA models. The number of latent variables of the PLS models is listed in the bracket.

VIA Model	Cap (2)	Ram (13)	CR (15)	CR 1dv (6)	CR 1dvsnv (3)
MAE [%]	2,43	0.96	0.96	1,19	1,85

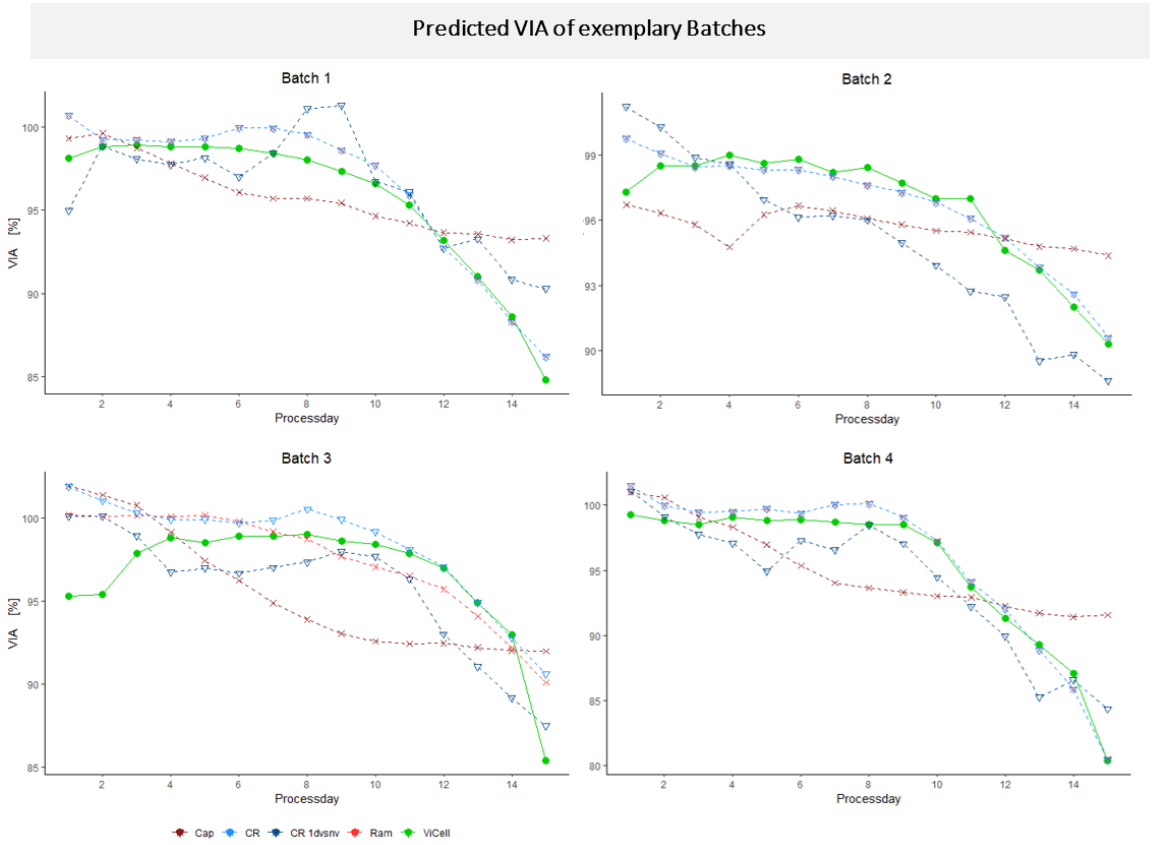

Figure 37: Plot of VIA from four exemplary chosen development batches. The green curve presents the reference values, the brown graph presents the best model created from capacitance data only, the red and blue graphs present the new models. A very nice correlation with VIA predictions throughout the whole process is visible.

Applying these models on four exemplary batches in Figure 37, it is seen that in three out of four batches the Ram and the CR models are overlying each other. This points out, that the addition of capacitance data does not add value to this model. Also, it is obvious, that none of the other models can reproduce the dynamic of the viability as well as those two models. It can also be seen that some of the calculated viability values are greater than 100%, which cannot happen in real life. These faults are due to the prediction errors.

4.4 Prediction of apoptosis

As the programmed cell death has effect on the cell membranes of living cells, a relation between the dielectric capacitance signal and apoptosis is hypothesized. Therefore, the apoptosis is measured via two different flow cytometric assays. On one hand there is the Guava Nexin assay, delivering information about percentual early apoptosis, late apoptosis, debris, and healthy cells. On the other hand, the Guava Viacount assay is performed, giving the percentual viability.

There are measurements of samples from 20 different development batches, mentioning that it was not possible to perform the experiments daily. Thus, the data density is not as high as the standard Vi-Cell values.

A multivariate model was chosen to predict the different parameters describing the viability status of the cultivated cells. A dataset of mean centered dielectric frequency scan data and the appropriate Cole-Cole parameters was correlated to the following parameter: % healthy cells, % early apoptotic cells, % late apoptotic cells, % cell debris and % viability (by Viacount).

To outline the quality of the resulting models, two different sets of plots were created. In the first set (Figure 38), the % early apoptosis, % late apoptosis and % cell debris are plotted, accompanied by the respective predicted value by a PLS model. Three exemplary picked batches are shown. In the second set of plots (Figure 39) the % viability measured by Viacout and the % healthy cells measured by nexin assay, again accompanied by the respective predicted values by a PLS model of the same batches are presented.

The PLS model for early Apoptosis has 3, for late apoptosis has 4, for debris and healthy cells has 5, and for viability has 3 latent variables.

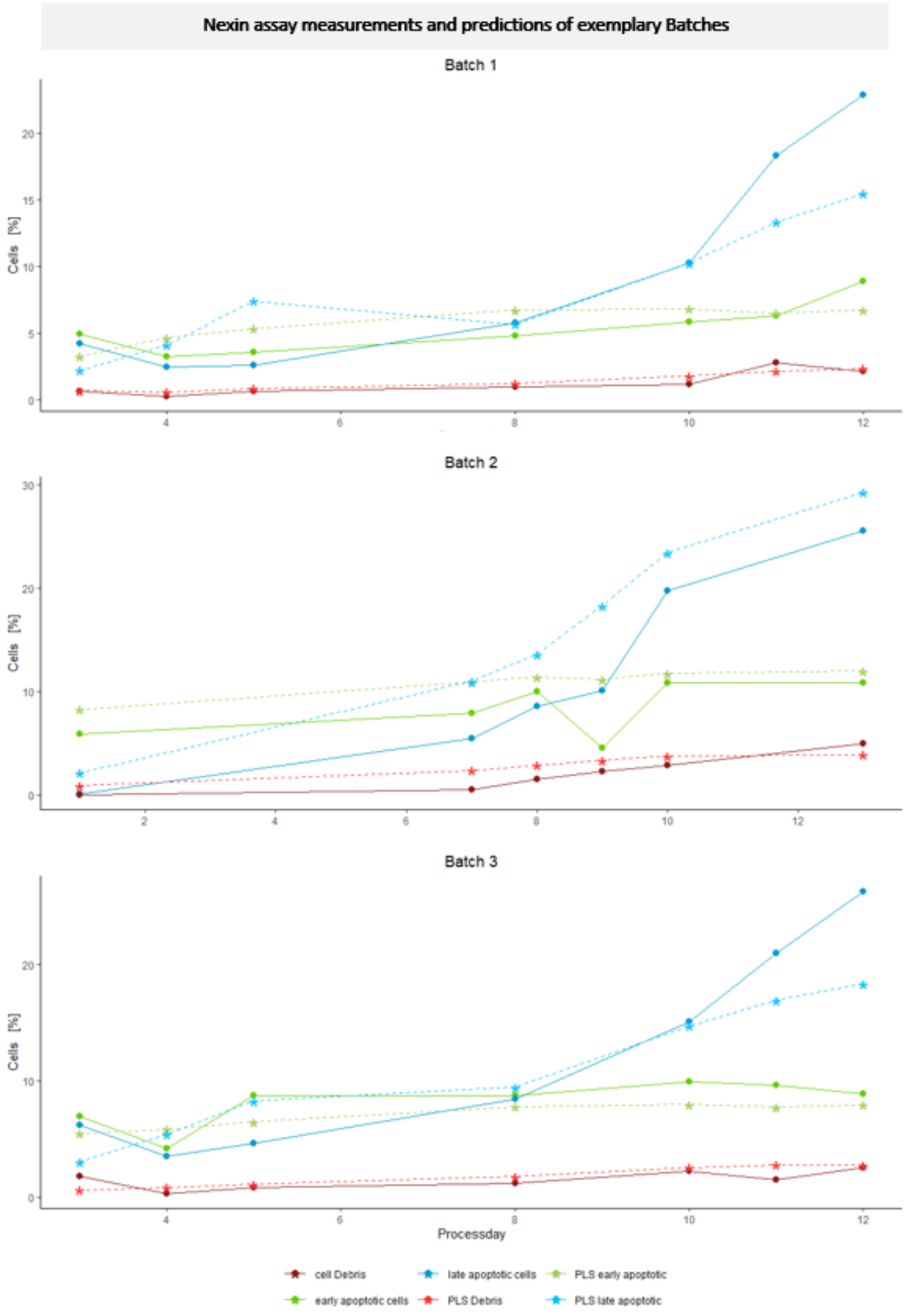

Figure 38: Depiction of three data pairs presenting the measured (continuous line) and predicted (dashed line) percentages of early apoptosis (green), late apoptosis (blue) and cell debris (red) obtained from Guava Nexin assay. Three batches are exemplary chosen for illustrative plotting.

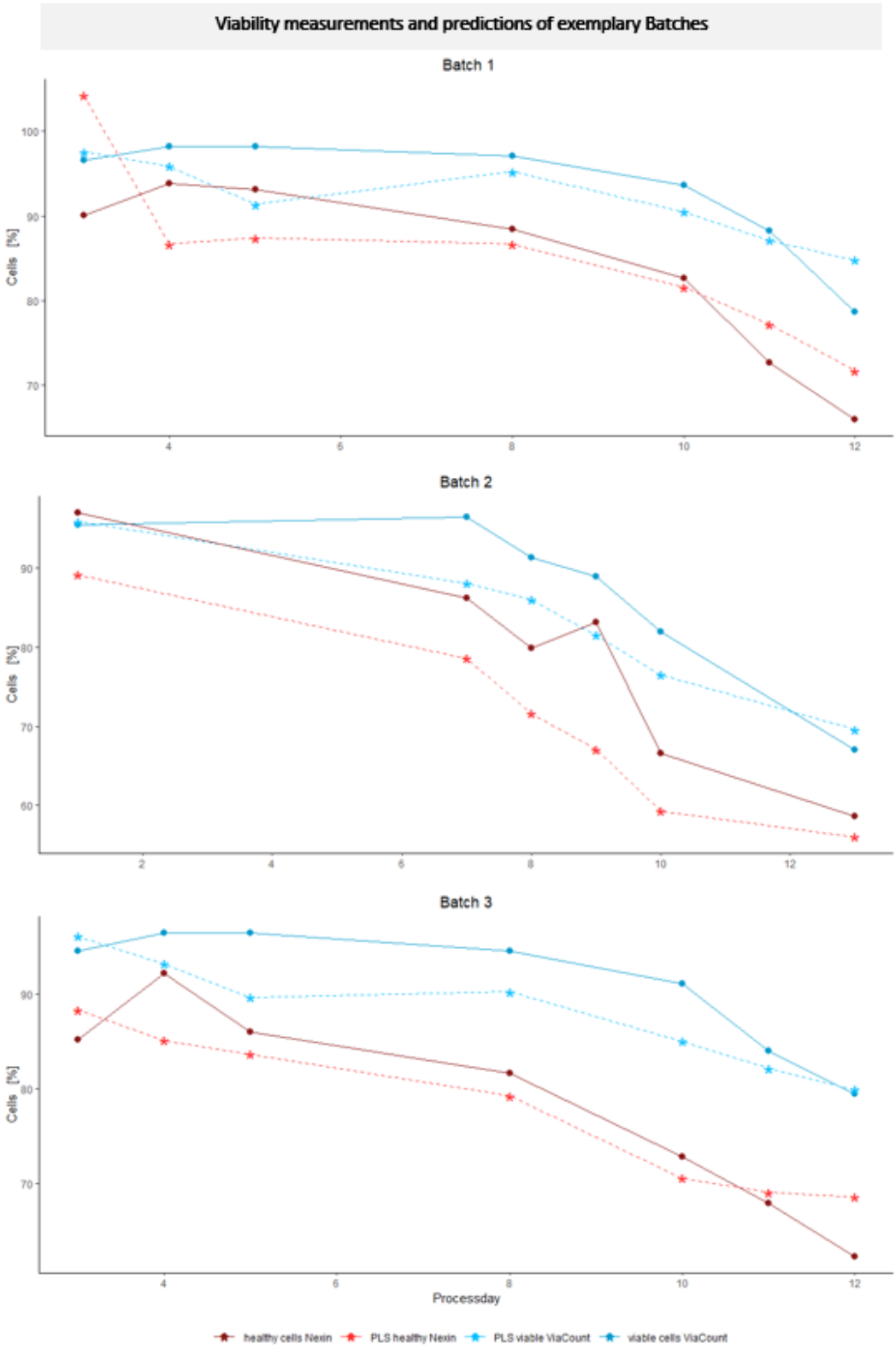

Figure 39: Depiction of three data pairs presenting the measured (continuous line) and predicted (dashed line) percentages of healthy cells from Guava Nexin assay (red) and viability by Guava Viacount assay (blue) Three batches are exemplary chosen for illustrative plotting.

Looking at the first set of plots in Figure 38 it needs to be said, that the quality of this Nexin assay seems to be quite poor. There are almost no cell Debris found throughout the process. Also, it would be expected, that the late apoptosis curve is a timewise shift from early apoptosis as these states follow each other in the biological life cycle. Talking about the second set of plots, the measured values seem to be a bit more trustworthy, since there is, except for the outliers in batch two and three, a shared trend of those two parameters.

Generally speaking, the prediction of the measured parameters by a PLS model delineate the trends, even if the absolute error is relatively high. Considering that the precision of the method to measure the reference values might not be sufficiently good at any timepoints, these models are not very reliable.

4.5 Model testing for large scale production process

Regarding the prediction of biomass parameter in 13 kL production scale, some models for a specific production unit (PU) were created and compared. Generating a global model was not tested here, since the available data from production and development processes arise from different processes for different products, process parameters and varying working procedure for the capacitance probe. There are several batches of one process available, to train and test statistical models. Dielectric frequency scan data including Cole-Cole parameter are available for all these batches. VCD and VIA measured by Vi-Cell serve as reference values.

Utilizing the knowledge from the previous chapters five different models were built to predict firstly the VCD and later the viability of the cell culture in the bioreactor. The dielectric frequency data were mean centered. Mentioning that there is no possibility to zero the probes between two runs in the production bioreactor, those data are lacking a baseline correction in contrast to the development batches. To recreate the model currently used in the PU, a linear model correlating the capacitance at 580 kHz was made (lm 580kHz). As the Cole-Cole parameter dCap convinced with good performance earlier, again a linear correlation was built (lm dCap). Next, the best performing PLS model including all frequency scan and Cole-Cole data was created for the production scale (PLS PU). Lastly, the model from

the development batches was transferred to the production data (PLS 20L) to check whether the model is transferrable between different scales and processes.

4.5.1 VCD prediction

Using the same procedure as in previous chapters, the RMSEP was calculated for all four models via cross validation and depicted in Figure 40.

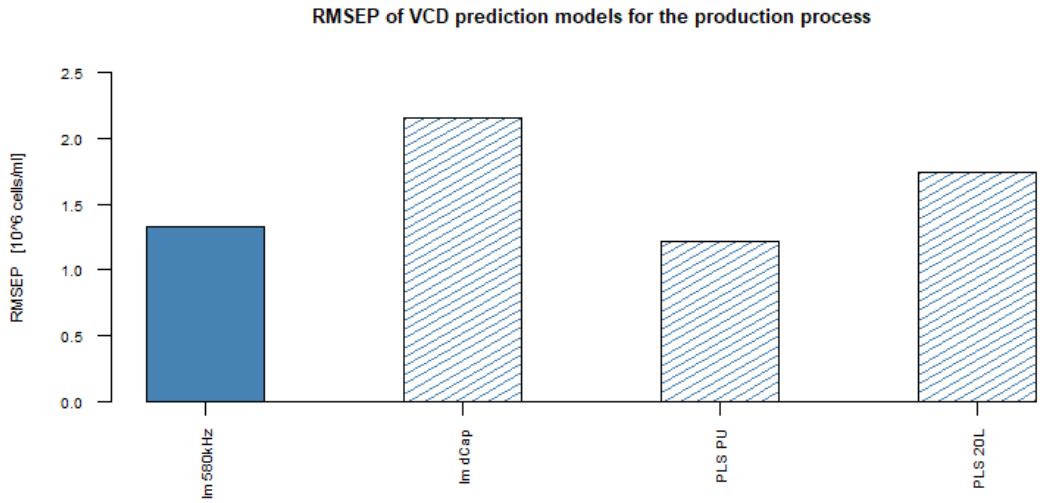

Figure 40: Plot of RMSEPs of the different linear (lm) and partial least squares (PLS) models for VCD prediction in production scale. The bar of the "state of the art" model is marked in blue.

Surprisingly, the lm dCap has the highest error bar of all tested models. Thus, the Cole-Cole parameter seems not to be as significant for non-zeroed data, because the baseline probably irritates the computation for the dCap. At the first look, the PLS seems to predict the VCD more precisely than the linear "state of the art" model, whereas the transferred PLS model has a high RMSEP.

Table 14: RMSEP values belonging to the chart presenting the different VCD models by usage of capacitance and raman data. The number of latent variables of the PLS models is listed in the bracket.

VCD Model	lm 580kHz	lm dCap	PLS PU (8)	PLS 20L (2)
RMSEP [10^6 cells/mL]	1.332	2.159	1.211	1.742

For having a closer insight on the dynamics of those models, the MAE is calculated again. Whilst the first few days the lm 580 kHz plotted in black is very close to the zero-line indicating a smooth VCD prediction. Later in process, the green model (PLS PU) performs best

in having the smallest deviation from the reference values. The lm dCap and the PLS 20L neither promise a good prediction in the early, nor in the late stage of the process.

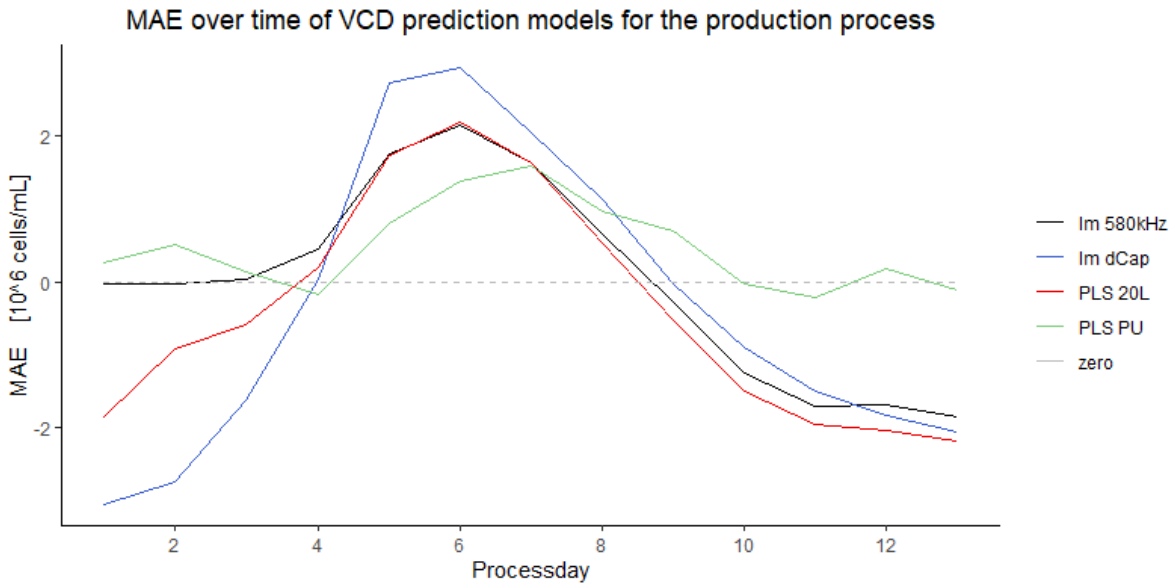

Figure 41: A plot presenting the mean average error (MAE) of VCD prediction by different models over process time. The grey dashed zero line indicates whether the model is underestimating (positive values) or overestimating (negative value) the average VCD measured by ViCell.

Again, the MAPE is calculated for the prediction of VCD by the four different models to measure the goodness of fit.

Table 15: MAPE values belonging to the different VCD models. The number of latent variables of the PLS models is listed in the bracket.

VCD Model	lm 580kHz	lm dCap	PLS PU (8)	PLS 20L (2)
MAPE [%]	29.31	89.65	20.89	44.13

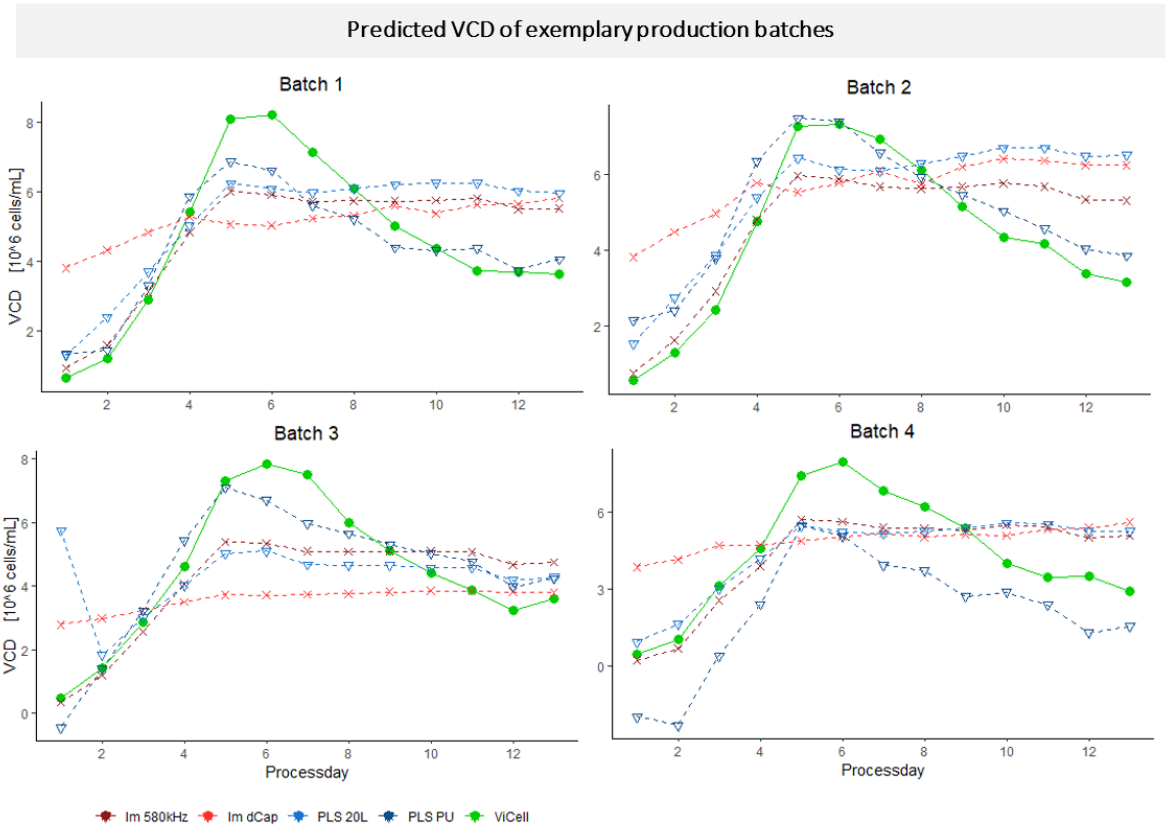

Figure 42: Plot of VCD from four exemplary chosen production batches. The green curve presents the reference. The best correlation with VCD predictions throughout the whole process is observed for the PLS PU model.

Applying the models to four exemplary chosen batches of production process it is obvious, that the PLS PU is the only model following the trend of the measured reference values. The lm dCap has almost no correlation throughout the whole process. The other models fail in predicting the death phase of the process.

Summarizing, the PLS PU delivers the most promising results to predict the VCD by an online technology during the whole process.

4.5.2 Viability prediction

Calculations of RMSEPs by cross validation were done for the four models described before predicting the percentual viability in the production process.

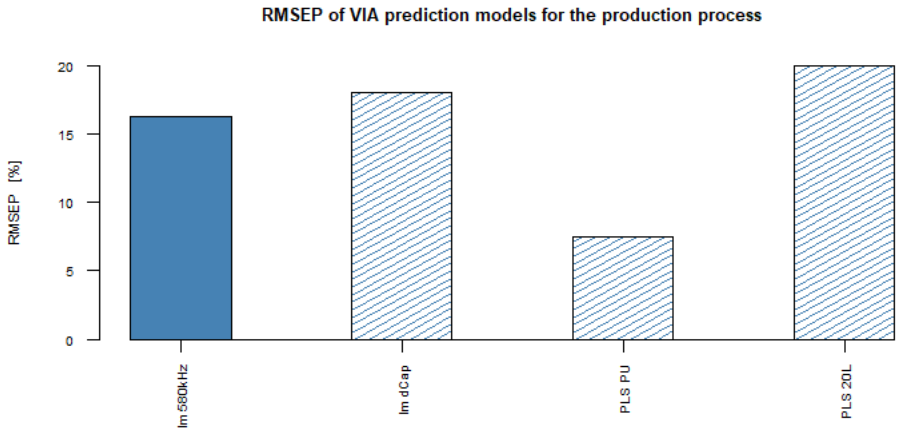

Figure 43: Plot of RMSEPs of the different linear (lm) and partial least squares (PLS) models for VIA prediction in production scale. The bar of the "state of the art" model is marked in blue.

The bar chart in Figure 43 points out, that there is a big reduction of error by the PLS PU, compared to the linear model. Comparing the value 7.51 % to the RMSEP received in chapter 3.2.2. (VIA prediction for development process), it seems still very high. Looking at the measured range of viability in the production process the viability decreases towards ~50 % (development process: ~85 %), allowing larger absolute errors.

Table 16: RMSEP values belonging to the chart presenting the different VIA models by usage of capacitance data. The number of latent variables of the PLS models is listed in the bracket.

VIA Model	lm 580kHz	lm dCap	PLS PU (9)	PLS 20L (2)
RMSEP [%]	16.31	18.02	7.51	20.03

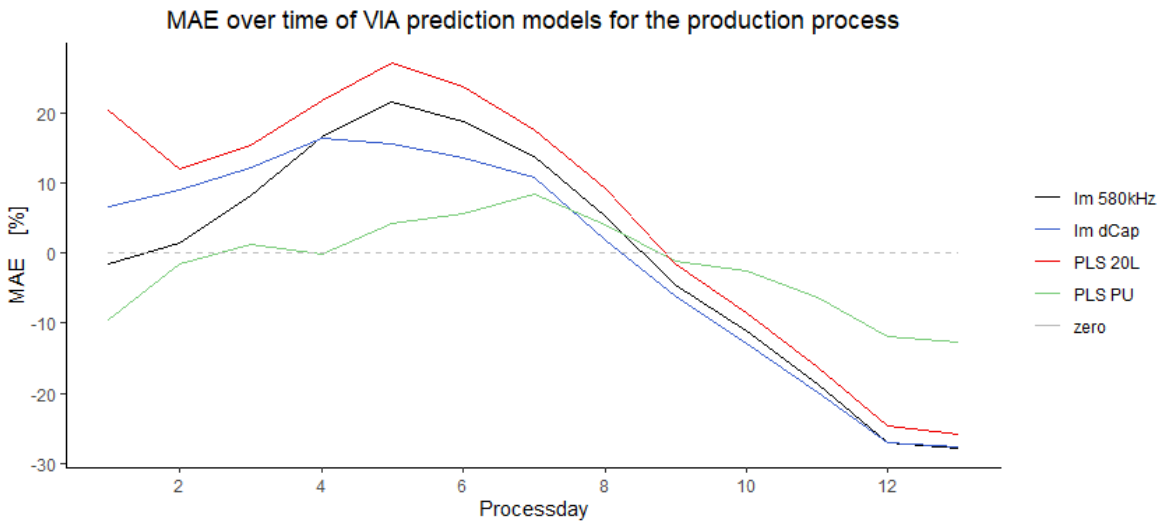

Figure 44: A plot presenting the mean average error (MAE) of VIA prediction by different models over process time. The grey dashed zero line indicates whether the model is underestimating (positive values) or overestimating (negative value) the average VIA measured by ViCell.

The MAPE for the prediction by the 5 prediction models are listed in Table 17.

Table 17: MAPE values belonging to the different VIA models. The number of latent variables of the PLS models is listed in the bracket.

VIA Model	lm 580kHz	lm dCap	PLS PU (9)	PLS 20L (2)
MAE [%]	18.16	18.85	7.90	21,99

Figure 44 showing the MAE of the four VIA models over process time emphasize the relatively good fit of the PLS PU model for the whole process. The green graph is much closer to the reference line than any of the other models. It can also be seen that some of the calculated viability values are greater than 100 %, which is an artefact of the model and cannot happen in real life. These faults are due to the prediction errors.

As the error of the PLS 20L model is the highest in the plot, it can be concluded, that the transfer of a model between different processes is not suitable to predict process parameters like viability or cell density. This was expected, as besides the use of differing process parameters a different working procedure for the capacitance probe has been applied.

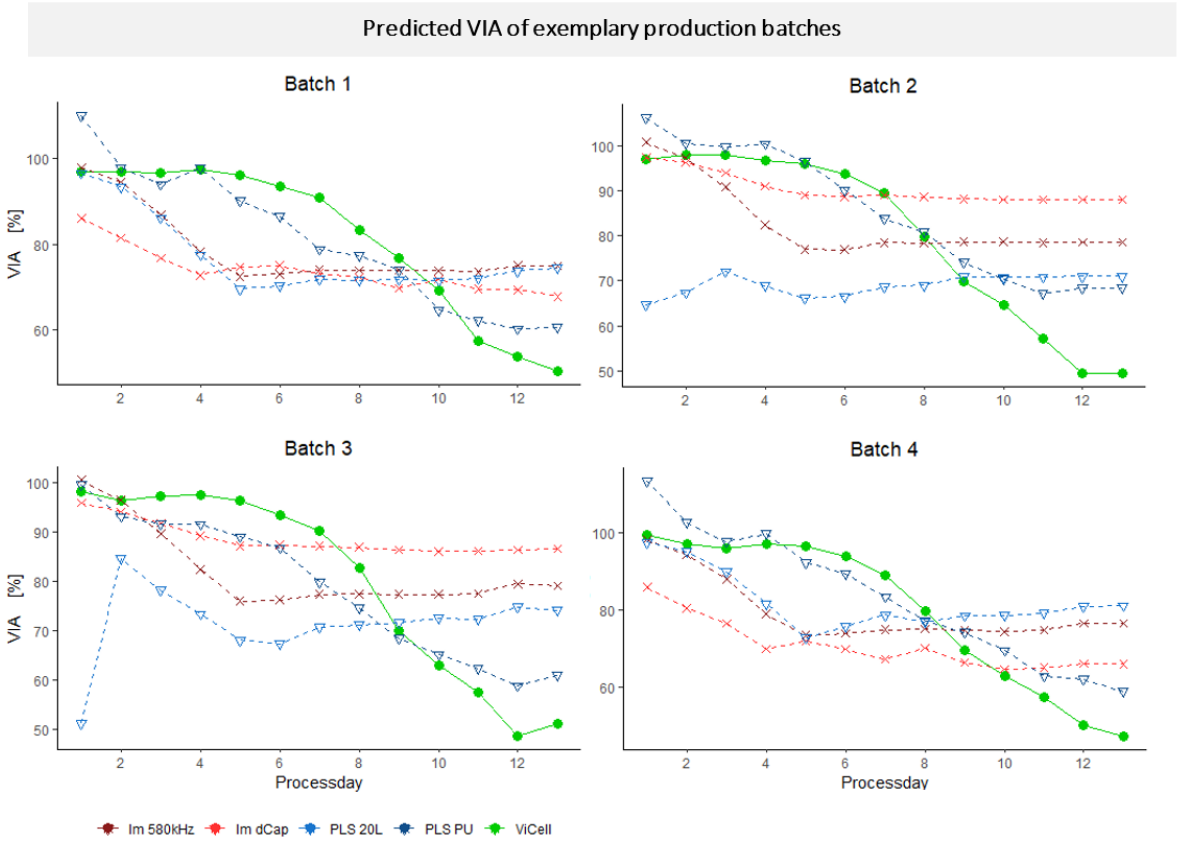

Figure 45: Plot of VIA from four exemplary chosen production batches. The green curve presents the reference. A very nice correlation with VIA predictions throughout the whole process is visible in the PLS PU model.

Applying the models to some exemplary batches in Figure 45, the PLS PU model is the only model capturing the trend of the VIA.

5 Discussion

5.1 VCD prediction

The performance of the VCD prediction models generated from capacitance data is varying a lot. The two linear models based on critical frequency and the Cole-Cole alpha are remarkably worse, compared to the other models, which was expectable, as those parameters are described to comprise information about cell size in the literature [13] and [32], which is not necessarily correlating with the VCD or the total biomass volume. Compared to the linear model based on the single frequency data, which represents the state-of-the-art model, the linear model correlating VCD with a single frequency measurement and the PLS model based on the combined data with the delta capacitance perform better. Considering the overall performance, those two models are similarly good with MAPEs of 11.92% (lm dCap) and 12.39 % (PLS fscc), but viewing the performance during specific phases, the lm dCap model performs better in the early stage whereas the PLS fscc model performs better in the end of the fermentation process. Thus, it is hard to create a model that fits well throughout the whole process. Nevertheless, all these models enhance the prediction quality against the state-of-the-art model. Especially in the late stage of the fermentation, a systematic overestimation of the viable cell density is noticed.

Extending the input data with the spectra of the Raman probe, the overall RMSEP value did not improve, but viewing the dynamic performance of the combined models (CR and CR 1dvsnv), outstandingly good results are seen, as the MAE is only slightly below the zero line throughout the whole process. Thus, it is of note that the overall RMSEP can sometimes trigger insufficient conclusions, as it represents an average error and does not provide the information on how well the growth curve and, hence, the process dynamics are captured by the model. The generated models show a good fit to the trend of the offline measurements, but still an overestimation is visible. Also, the fact that the model including only Raman data and the model including Raman and Capacitance data are showing similar performance indicates, that the Raman signal gives very good information about the VCD in a process. But still, the inclusion of capacitance data improves the resulting prediction quality. The MAPEs of the two best models are 14.66 % (Ram) and 10.75 % (CR).

Considering the models for the production process, the lm dCap turns out as not suitable at all. Since this model fits well for predicting VCD in the development process, it is expected that this might be due to the fact, that the capacitance probe is not zeroed before each process in the production scale. Thus, the baseline might have an impact on the computation of the Cole-Cole parameter delta capacitance, which makes it impossible to predict the VCD from these values. However, the PLS model based on the dielectric frequency scan data and the Cole-Cole parameter is capable of predicting the VCD with a MAPE of 20.89 % (PLS PU), which is better than the state-of-the-art model with a MAPE of 29.31 % (lm 580).

Table 18: Overview of Error values derived in this thesis, the reference method and comparable studies [16].

	Source	Comment	Error [%]
Results	Frequency scan	lm dCap PLS fscc	11.92 % (MAPE) 12.39 % (MAPE)
	Raman	Ram CR	14.66 % (MAPE) 10.75 % (MAPE)
	Large scale	PLS PU	20.89 % (MAPE)
Reference method	Vi-Cell		5 % (Error estimate) [46]
Literature	Cannizzaro, 2003	Batch Phase	9 – 22 % (CVRMSE)
	Opel, 2010	Batch and Fed-batch	7 – 23% (CVRMSE)
	Konakovsky, 2014	Fed-batch	7 – 38 % (CVRMSE)

In Table 18, the mean percentual Error (MAPE) of the best performing models generated in this thesis, the estimated error of the reference method for VCD measurement, which is the day-to-day precision and rather underestimated in this application, and some comparable results from literature are stated. [46] It has to be considered, that the models are based on the ViCell data and thus these approximately 5 % error are an inherent part of the error of the prediction models. Comparing the errors to the literature, it is seen that the models created in other studies have percentual errors in the same or equally larger dimension of the selected new models. Comparing the MAPE values of the capacitance models and the Raman

models, it is seen that the overall performance of the Ram model is not better than the lm dCap or PLS fscc.

5.2 Viability prediction

Comparing the models created for viability prediction, it is seen that the linear model correlating a single frequency to the viability does not follow the trend of the offline viability measurement ending up with a MAPE of 3.03 % (lm 580), whereas the PLS models fs and fscc show better performance in some batches with a MAPE of 2.32 % (PLS fs) and 2.43 % (PLS fscc). Crucial for the quality of the prediction is the significant decrease of the viability, which is adequately predicted in most of the fed-batch processes.

Apparently, the signal to noise ratio is too low, if batches exhibit high viability until the end of the process. Viability values decreasing to less than 90 % are relevant for the establishment of a significant model. That phenomenon is reported by Ma et al. [32], as they observed a correlation between Cole-Cole parameters and viability in batch processes, where the viability drop is much greater than in a fed-batch process.

Adding the Raman signal enhanced the quality of the prediction considerably. The PLS models of Capacitance and Raman data appear to be almost identical to the model of Raman data only. The MAPE of those models is 0.96 % (Ram and CR). This shows that Raman spectroscopy is a more promising tool for online viability prediction in cell culture processes. Different from the VCD prediction models, for viability prediction the data pretreatments reduce the quality of viability prediction.

Confirming the signal-to-noise issue discussed for the viability prediction in the development process by capacitance data, the viability prediction works out well for the production process. In these processes, the viability drops down to 50 % and this enables a good viability prediction with a MAPE of 7.9 % which looks a lot compared to the other models. Considering the range of viability in these processes, it is reasonable that the error is higher.

Table 19: Overview of error values derived in this thesis and the method for reference measurements.

	Source	Comment	Error [%]
Results	Frequency scan	PLS fs PLS fscc	2.32 % (MAPE) 2.43 % (MAPE)
	Raman	Ram CR	0.96 % (MAPE) 0.96 % (MAPE)
	Large scale	PLS PU	7.9 % (MAPE)
Reference method	Vi-Cell		0.8 % (Error estimate)

In Table 19 the MAPEs of the best models derived in this thesis and the error estimate of the reference method, which is an estimate of the combined instrument-to-instrument and operator-to-operator precision and again rather underestimated in this case, which are 0.8 %. [46] This error again influences all the other models, since they are based on the viability values measured by ViCell. As already discussed, the frequency scan models from the development process show inadequate performance, also due to the relatively narrow viability range covered by the data set, whereas the capacitance models in the production process convince with better predictions. dynamics Adding Raman data to the frequency scan improves the viability prediction in batches with high signal-to-noise ratio. Nevertheless, the production process indicates the applicability of the capacitance data for viability prediction.

5.3 TCD prediction

For TCD prediction models effects similar to the VCD prediction are found, as the linear models with the Cole-Cole alpha and the critical frequency again deliver unsatisfying predictions. The predictions provided by the pls fscc model and the lm dcap are close to the reference values. This is not surprising, as the VCD and TCD parameter are closely related in the development process, since the viability is high throughout the whole process.

5.4 Average cell diameter prediction

Successful predictions of cell size are reported by Ma et. al. [32], as they found a correlation between the Cole-Cole parameter critical frequency with the average cell diameter. This

observation didn´t occur in this study, as the lm ckrit did not follow the trend of the reference data at all. The only model that seems to correlate with the reference data is the PLS fscc, but still the deviation is too high, to take it as a successful model. Hence, the prediction of the average cell diameter cannot be done by measurement of dielectric capacitance in a bioprocess.

5.5 Average circularity prediction

Similar observations as for average cell diameter predictions are made with the prediction of the average circularity of a cell culture in a bioprocess. The only model that features a slightly correlation to the reference values is the PLS fscc. But again, the predicted values have large deviations, so it is not useful for predicting this parameter.

5.6 Prediction of apoptosis

Before talking about the models for predicting the different stages of apoptosis it must be said, that within the measured data some irregularities are noticed. It would be expected that the measured early apoptotic fraction turns into late apoptotic cells later. So, the function of late apoptotic cells should be a delayed function of the early apoptotic cells which did not occur in the experiments for this thesis. Altogether only little data is available which makes it hard to create a convincing model. The batches underlying these measurements show, as discussed earlier, high viability and thus the fraction of apoptotic cells is small, making it even harder to generate a meaningful model. Considering these effects, the created PLS models are not very reliable.

6 Conclusion and Outlook

Dielectric spectroscopy provides a practicable opportunity to be used as a PAT tool for the continuous online monitoring during USP process of CHO cells. Its nondestructive properties and the ability to discriminate between different conditions and densities of the cells in a reactor are notable properties, where only one probe is capable of generating online data for various different process parameters and cell characteristics. The use of dielectric spectroscopy prediction models makes the reduction of frequent offline sampling as well as the implementation of quick and smart online control strategies possible. This thesis overall present a strategy of generating and assessing the reliability of these prediction models. The milestones within the workflow comprise the gathering of all available online and offline data to one data frame and the extraction and preparation of the relevant information for building a prediction model. For the assessment of the predictive capability of the models, the batch wise CV approach for calculating the RMSEP and the MAE/MAPE, which is described in this thesis, offers a better comparison between different models unlike the standard k-fold CV approach. This provides more realistic errors as output. Thus, by applying these steps the most reliable and useful prediction models can be selected.

A relevant process parameter for online prediction in a bioprocess is the VCD. The simple linear regression of the Cole-Cole parameter dCap provides promising results in all stages of the process. Also, the pls fscc model delivered good results, but requires more sophisticated modeling approaches. Thus, by implementing a software that calculates the Cole-Cole parameters of the dielectric frequency scan in real-time and which is also connected to the control system, the lm dCap model can be used for online control strategies.

Another crucial process parameter, namely the viability is revealed as solid predictable parameter by a prediction model based on a capacitance probe. However, it is essential to have a training dataset that includes batches with strong variations in the percentual viability. This is affirmed, as the model for the production process shows a better fit than the model for the development process. Nevertheless, a PLS model is required for predicting the viability of a cell culture Since the TCD is the result from dividing VCD by viability, the same is valid for TCD prediction.

Even though in literature successful modelling approaches for predicting average cell diameter are reported, the correlation results obtained in this thesis are weak. Hence, this effect reported in literature is not confirmed. Also, an online prediction of average cell circularity did no prove realizable for process monitoring. Online monitoring of apoptosis by dielectric spectroscopy would need further investigation, as the data set for early and late apoptosis available for this thesis is very small and, similar to the viability prediction in the production process, due to high viabilities the number of apoptotic cells was too small for generating a qualitative model. The resolution of this method makes a capture of apoptotic events impossible in this experiment.

Adding a Raman signal to the dataset for predicting VCD and viability increased the quality of the prediction models. At the same time, processing the spectral data of a Raman probe to generate predicted VCD or viability values are more complex than a prediction by a capacitance model described before. Keeping in mind that the implementation of additional Raman probes is an expensive acquisition, the improved analysis and prediction capabilities needs to be balanced against the additional cost.

Nevertheless, considering the fact that many other process parameters like metabolites can be predicted by Raman spectroscopy, it is a promising technology for process monitoring in upstream processing.

In conclusion, it can be stated, by the comparison of the results with the reference method and literature, that the use of capacitance probes in frequency scan mode can improve the VCD prediction, especially in the middle stage of a fed-batch process, in comparison to the state-of-the-art model. Adding Raman data to the prediction model improves the prediction in the later stage, whereas the prediction in early stage does not benefit from this upgrade. Since the early stage is of high importance in bioprocesses, the generated models by capacitance data are reliably applicable.
For online viability prediction, the capacitance models are sufficient, if the viability range in training data is large, whereas for high viability processes the resolution of the capacitance models is not good enough. In this case, a Raman probe is a good remedy to provide online viability prediction with high resolution.

For the prediction of average circularity, average cell diameter and apoptotic events using dielectric frequency scan data, no adequate model could be generated in this thesis.

To further improve the models developed in this thesis further investigations should be done. One approach could be the development of phase specific models for predicting the VCD and viability during growth phase, stable phase and death phase separately. The models could be improved in terms of robustness, if more variation in the process settings of the training batches would be introduced. Since these models are built on optimized processes, batches with completely divergent growth behaviour are missing.
Other machine learning methods could be tested for more precise models.
Also, models based on data pre-processing that are feasible for all processes and scales can be developed, to enable a generalized model for model transfer between different scales.
As there are some batches that turned out having significantly higher errors than other ones, investigations in finding a pattern behind this effect can be done.

7 List of Abbreviations

1dv	First derivative
2dv	Second derivative
7-AAD	7-aminoactionmycin D
API	Active pharmaceutical ingredients
cc	Cole-Cole parameters
CCAlpha, cca	Cole-Cole Alpha
cF	critical Frequency
CHO	Chinese hamster ovary
CIP	cleaning in place
CIRC	Average circularity of cells
CQA	Critical quality attribute
CV	Cross validation
CVRMSE	coefficient of variation of the root mean square error
dcap	Delta capacitance
DIA	Average cell diameter
EMA	European Medicines Agency
FDA	U.S. Food and Drug administration
fs	Frequency scan
FSC	Forward scatter
fscc	Frequency scan and Cole-Cole parameters
g368	Greater than 368 kHz
lm	Simple linear regression model
MAE	Mean absolute error
MAPE	Mean average percentage error
mc	Mean centering
MLR	multiple linear regression
MVDA	Multivariate data analysis
PAT	Process analytical technology

PCR	principal component regression
PCS	process control system
PLS	Partial least squares
PU	Production Unit
QbD	Quality by Design
RMSEP	Root mean squared error of prediction
sc	Scaling
SIP	steaming in place
snv	Standard normal variate
SNV	Standard normal variate
SOP	standard operating procedures
SSC	Side scatter
TCD	Total cell density
USP	Upstream production process
VCD	Viable cell density

8 List of Figures

9 List of Tables

10 List of Equations

11 References

[1] S. Simoens, “Health economics of market access for biopharmaceuticals and biosimilars,” *J. Med. Econ.*, vol. 12, no. 3, pp. 211–218, 2009, doi: 10.3111/13696990903260094.

[2] G. Walsh, “Biopharmaceutical benchmarks 2018,” *Nat. Biotechnol.*, vol. 36, no. 12, pp. 1136–1145, 2018, doi: 10.1038/nbt.4305.

[3] M. Butler, “Animal cell cultures: Recent achievements and perspectives in the production of biopharmaceuticals,” *Appl. Microbiol. Biotechnol.*, vol. 68, no. 3, pp. 283–291, 2005, doi: 10.1007/s00253-005-1980-8.

[4] B. Sekhon and V. Saluja, “Biosimilars: an overview,” *Biosimilars*, vol. Volume 1, pp. 1–11, 2011, doi: 10.2147/bs.s16120.

[5] Y. Ingrasciotta, P. M. Cutroneo, I. Marcianò, T. Giezen, F. Atzeni, and G. Trifirò, “Safety of Biologics, Including Biosimilars: Perspectives on Current Status and Future Direction.,” *Drug Saf.*, vol. 41, no. 11, pp. 1013–1022, Nov. 2018, doi: 10.1007/s40264-018-0684-9.

[6] M. C. Flickinger, A. S. Rathore, and G. Kapoor, “Process Analytical Technology: Strategies for Biopharmaceuticals,” *Encycl. Ind. Biotechnol.*, no. 1, 2013, doi: 10.1002/9780470054581.eib652.

[7] F. and D. Administration, “Guidance for Industry, PAT-A Framework for Innovative Pharmaceutical Development, Manufacturing and Quality Assurance,” no. September, 2004, [Online]. Available: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm070305.pdf.

[8] E. K. Read, J. T. Park, R. B. Shah, B. S. Riley, K. A. Brorson, and A. S. Rathore, “Process analytical technology (PAT) for biopharmaceutical products: Part I. Concepts and applications,” *Biotechnol. Bioeng.*, vol. 105, no. 2, pp. 276–284, 2010, doi: 10.1002/bit.22528.

[9] B. Moore, R. Sanford, and A. Zhang, “Case study: The characterization and implementation of dielectric spectroscopy (biocapacitance) for process control in a commercial GMP CHO manufacturing process,” *Biotechnol. Prog.*, vol. 35, no. 3, 2019, doi: 10.1002/btpr.2782.

[10] M. Dabros *et al.*, “Cole-Cole, linear and multivariate modeling of capacitance data for on-line monitoring of biomass,” *Bioprocess Biosyst. Eng.*, vol. 32, no. 2, pp. 161–173, 2009, doi: 10.1007/s00449-008-0234-4.

[11] C. L. Davey, H. M. Davey, D. B. Kell, and R. W. Todd, “Introduction to the dielectric estimation of cellular biomass in real time, with special emphasis on measurements at high volume fractions,” *Anal. Chim. Acta*, vol. 279, no. 1, pp. 155–161, 1993, doi:

10.1016/0003-2670(93)85078-X.

[12] G. H. Markx and C. L. Davey, “The dielectric properties of biological cells at radiofrequencies: Applications in biotechnology,” *Enzyme Microb. Technol.*, vol. 25, no. 3–5, pp. 161–171, 1999, doi: 10.1016/S0141-0229(99)00008-3.

[13] C. Cannizzaro, R. Gügerli, I. Marison, and U. Von Stockar, “On-Line Biomass Monitoring of CHO Perfusion Culture With Scanning Dielectric Spectroscopy,” *Biotechnol. Bioeng.*, vol. 84, no. 5, pp. 597–610, 2003, doi: 10.1002/bit.10809.

[14] B. J. Downey, L. J. Graham, J. F. Breit, and N. K. Glutting, “A novel approach for using dielectric spectroscopy to predict viable cell volume (VCV) in early process development,” *Biotechnol. Prog.*, vol. 30, no. 2, pp. 479–487, 2014, doi: 10.1002/btpr.1845.

[15] L. Párta, D. Zalai, S. Borbély, and Á. Putics, “Application of dielectric spectroscopy for monitoring high cell density in monoclonal antibody producing CHO cell cultivations,” *Bioprocess Biosyst. Eng.*, vol. 37, no. 2, pp. 311–323, 2014, doi: 10.1007/s00449-013-0998-z.

[16] V. Konakovsky *et al.*, “Universal capacitance model for real-time biomass in cell culture,” *Sensors (Switzerland)*, vol. 15, no. 9, pp. 22128–22150, 2015, doi: 10.3390/s150922128.

[17] C. F. Opel, J. Li, and A. Amanullah, “Quantitative modeling of viable cell density, cell size, intracellular conductivity, and membrane capacitance in batch and fed-batch CHO processes using dielectric spectroscopy,” *Biotechnol. Prog.*, vol. 26, no. 4, pp. 1187–1199, 2010, doi: 10.1002/btpr.425.

[18] K. Braasch *et al.*, “The changing dielectric properties of CHO cells can be used to determine early apoptotic events in a bioprocess,” *Biotechnol. Bioeng.*, vol. 110, no. 11, pp. 2902–2914, 2013, doi: 10.1002/bit.24976.

[19] K. A. Forbes, “Raman Optical Activity Using Twisted Photons,” *Phys. Rev. Lett.*, vol. 122, no. 10, 2019, doi: 10.1103/PhysRevLett.122.103201.

[20] K. A. Esmonde-White, M. Cuellar, C. Uerpmann, B. Lenain, and I. R. Lewis, “Raman spectroscopy as a process analytical technology for pharmaceutical manufacturing and bioprocessing,” *Anal. Bioanal. Chem.*, vol. 409, no. 3, pp. 637–649, 2017, doi: 10.1007/s00216-016-9824-1.

[21] J. Whelan, S. Craven, and B. Glennon, “In situ Raman spectroscopy for simultaneous monitoring of multiple process parameters in mammalian cell culture bioreactors,” *Biotechnol. Prog.*, vol. 28, no. 5, pp. 1355–1362, 2012, doi: 10.1002/btpr.1590.

[22] E. I. Ltd., “No Title,” 2021. https://www.edinst.com/blog/what-is-raman-spectroscopy/ (accessed Jan. 13, 2021).

[23] T. De Beer, A. Burggraeve, M. Fonteyne, L. Saerens, J. P. Remon, and C. Vervaet, “Near

infrared and Raman spectroscopy for the in-process monitoring of pharmaceutical production processes," *Int. J. Pharm.*, vol. 417, no. 1–2, pp. 32–47, 2011, doi: 10.1016/j.ijpharm.2010.12.012.

[24] R. A. Carlton, "Infrared and Raman Microscopy," in *Pharmaceutical Microscopy*, Springer, 2011, pp. 131–156.

[25] C. Reads, "Developing a Novel Approach for Layer Controlled Graphene Synthesis and Tailoring the Properties for Applications," no. August, 2018, doi: 10.13140/RG.2.2.26125.18407.

[26] P. Kadlec, B. Gabrys, and S. Strandt, "Data-driven Soft Sensors in the process industry," *Comput. Chem. Eng.*, vol. 33, no. 4, pp. 795–814, 2009, doi: 10.1016/j.compchemeng.2008.12.012.

[27] X. Sheng and W. Xiong, "Soft sensor design based on phase partition ensemble of LSSVR models for nonlinear batch processes," *Math. Biosci. Eng.*, vol. 17, no. 2, pp. 1901–1921, 2020, doi: 10.3934/mbe.2020100.

[28] B. Lin, B. Recke, J. K. H. Knudsen, and S. B. Jørgensen, "A systematic approach for soft sensor development," *Comput. Chem. Eng.*, vol. 31, no. 5–6, pp. 419–425, 2007, doi: 10.1016/j.compchemeng.2006.05.030.

[29] H. Sundström and S. O. Enfors, "Software sensors for fermentation processes," *Bioprocess Biosyst. Eng.*, vol. 31, no. 2, pp. 145–152, 2008, doi: 10.1007/s00449-007-0157-5.

[30] W. Strober, "Trypan blue exclusion test of cell viability.," *Curr. Protoc. Immunol.*, vol. Appendix 3, pp. 2–3, 2001, doi: 10.1002/0471142735.ima03bs21.

[31] S. Delhalle, "An Introduction to the Molecular Mechanisms of Apoptosis," 2008.

[32] F. Ma, A. Zhang, D. Chang, O. D. Velev, K. Wiltberger, and R. Kshirsagar, "Real-time monitoring and control of CHO cell apoptosis by in situ multifrequency scanning dielectric spectroscopy," *Process Biochem.*, vol. 80, no. February, pp. 138–145, 2019, doi: 10.1016/j.procbio.2019.02.017.

[33] J. L. Weaver, "Introduction to flow cytometry," *Methods*, vol. 21, no. 3, pp. 199–201, 2000, doi: 10.1006/meth.2000.1000.

[34] T. Rowley, "Flow Cytometry - A Survey and the Basics," *Mater. Methods*, vol. 2, Aug. 2012, doi: 10.13070/mm.en.2.125.

[35] P. McCoy, "Basic principles of flow cytometry," vol. 16, pp. 229–243, 2002.

[36] S. Wold, "Chemometrics, why, what and where to next?," *J. Pharm. Biomed. Anal.*, vol. 9, no. 8, pp. 589–596, 1991, doi: 10.1016/0731-7085(91)80183-A.

[37] S. Roussel, S. Preys, F. Chauchard, and J. Lallemand, "Multivariate Data Analysis (Chemometrics)," 2014, pp. 7–59.

[38] R. G. Brereton, “Introduction to multivariate calibration in analytical chemistry,” *Analyst*, vol. 125, no. 11, pp. 2125–2154, 2000, doi: 10.1039/b003805i.

[39] J. E. Yardley, R. Todd, D. J. Nicholson, J. Barrett, D. B. Kell, and C. L. Davey, “Correction of the influence of baseline artefacts and electrode polarisation on dielectric spectra,” *Bioelectrochemistry Bioenerg.*, vol. 51, no. 1, pp. 53–65, 2000, doi: 10.1016/S0302-4598(99)00069-0.

[40] B. Li, B. H. Ray, K. J. Leister, and A. G. Ryder, “Performance monitoring of a mammalian cell based bioprocess using Raman spectroscopy,” *Anal. Chim. Acta*, vol. 796, pp. 84–91, 2013, doi: 10.1016/j.aca.2013.07.058.

[41] D. Iacobucci, M. J. Schneider, D. L. Popovich, and G. A. Bakamitsos, “Mean centering helps alleviate ‘micro’ but not ‘macro’ multicollinearity,” *Behav. Res. Methods*, vol. 48, no. 4, pp. 1308–1317, 2016, doi: 10.3758/s13428-015-0624-x.

[42] R. A. van den Berg, H. C. J. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf, “Centering, scaling, and transformations: Improving the biological information content of metabolomics data,” *BMC Genomics*, vol. 7, pp. 1–15, 2006, doi: 10.1186/1471-2164-7-142.

[43] R. M. Santos, J. M. Kessler, P. Salou, J. C. Menezes, and A. Peinado, *Monitoring mAb cultivations with in-situ raman spectroscopy: The influence of spectral selectivity on calibration models and industrial use as reliable PAT tool*, vol. 34, no. 3. 2018.

[44] T. Scharl-Hirsch, “Multivariate Statistics - Partial least Squares,” 2020. doi: 10.4135/9781446214565.n11.

[45] A. Hoskuldsson, “PLS Regression Methods,” *Data Handl. Sci. Technol.*, vol. 2, no. C, pp. 165–189, 2003, doi: 10.1016/S0922-3487(08)70226-0.

[46] S. Taferner, “Development of Calibration models for three different PAT Technologies and their potential for online monitoring of cell culture bioprocesses,” *Master thesis*, 2019.