Advanced modeling of industrial-scale fermentation process for antibiotic production

Magnusson, Atli Freyr

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

DTU Chemical and Biochemical Engineering
Department of Chemical and Biochemical Engineering

Advanced modeling of industrial-scale fermentation process for antibiotic production

Atli Freyr Magnússon - PhD Thesis

Preface

This thesis is submitted as a partial fulfillment of the requirements for the degree of Doctor of Philosophy at the Technical University of Denmark. The work has been carried out at the Process and Systems Engineering Centre (PROSYS) at the Department of Chemical and Biochemical Engineering from September 2019 to November 2022, and supervised by Prof. Gürkan Sin, and co-supervised by Dr. Stuart M Stocks and Dr. Jari Pajander.

This project is classified as an Industrial PhD program and is a collaboration between LEO Pharma A/S and the PROSYS research center at DTU Chemical Engineering. The project was co-financed by Innovationsfond as part of Denmark's Industrial Researcher Programme for the purpose of creating growth and employment in Denmark. (Egrant: Innovationsfond 9065-00134B, Denmark).

Acknowledgements

First of all, I would like to express my most real gratitude to my DTU supervisor Gürkan Sin for giving me the opportunity and for all the guidance and input, as well as all his support and confidence in the project. A special deepest thanks go to Stuart Stocks, whom I probably spent the most time with over the last three years. He's been the guiding hand over the entire period, and I would like to express my deepest gratitude for his straightforward and sincere approach. Furthermore, I'd like to acknowledge my final co-supervisor, Jari Pajander, for all his scientific input regarding chemometrics. There were not many direct meetings, but I enjoyed them.

From LEO Pharma, I would also like to thank the people who supported and helped me during my experiments. I appreciate the industrial insight they provided, which was very important for this project and encouraged me to continue researching this topic. In particular to those within the MSAT Fermentation team, Kasper Israelsen, Claus Christiansen, and Jarno Robin.

I also have to acknowledge the team working on the production site for training me to collect data directly from production, Rasmus Mortensen for helping us with experimental setups, and everyone in Biological Process and Analytics department for allowing us to use the laboratory space and equipment.

The last acknowledgment from the LEO side goes to Peter Molnár, an MSc student who I always saw more as a co-worker and friend than my masters student. The data collection in the early morning would have been extremely boring without his presence to lighten the mood.

I have also greatly enjoyed being part of the work environment at the PROSYS center, and I would like to express my warmest gratitude to everyone I met there. I would not have made it through this PhD journey without you. Special thanks go to Adem Aouichaoui for all the great times. I would also like to acknowledge Nikolous, Simon, Peter, Mads, Jesper (Both of them), Fiammetta, Johan, Amalie, Vicente, and Sebastian for making PROSYS such a unique but enjoyable workplace. I wish I could have been there more. Thanks to our administrative coordinators, Gitte Læssøe and Anja Jensen, for all their support and assistance when needed.

Moving away from direct colleagues, my heartfelt gratitude and thanks to all my friends

back home and elsewhere worldwide. Our online voice chats were the only thing keeping my spirits high when stuck in quarantine alone in a foreign country. Last but not least, I would like to remember the unconditional support and encouragement from my large, loving family.

I have always had the critical flaw of not remembering names. If I failed to acknowledge anyone, I sincerely apologize. Rest assured that your contributions are not forgotten, and I appreciate everyone who supported me during this thesis. This one is dedicated to all of you.

Abstract

Biotechnology plays an integral role in the modern economy and is responsible for producing various bulk and specialty chemicals critical to the function of society. The bioprocess involves using microorganisms as a miniature factory to convert substrates into valuable molecules. Antibiotics are antimicrobial substances used to fight bacterial infections and are commonly produced via fermentation. As global antimicrobial resistance becomes a larger and larger threat to humanity, there has been an increased interest in the potential of old-generation antibiotics to address the current need for new antibiotics. Faced with increasing demand, we need to modernize the current production methods. The chemical and biochemical industry is transitioning through the fourth industrial revolution, or Industry 4.0, which is the joining of technologies that blur the lines between the physical, digital, and biological worlds.

Digital twins based on process models are a crucial technology for industries in the changing competitive landscape following the change to Industry 4.0. Digital Twins are a digital representation of a real-world physical product, system, or process. This technology's foundation and primary enabler is a mathematical model that accurately captures the relevant physical and biological phenomena. However, modeling fermentation systems is challenging due to biological complexities. Furthermore, the pharmaceutical industry has always had a strong emphasis on quality. Legal regulations specify the required purity of Active Pharmaceutical Ingredients and place limits on potentially harmful or efficacy-reductive impurities. These impurities may be byproducts of the bioprocess itself and may have very similar chemical and physical properties, making them impossible to separate in the purification process. So far, bioprocess modeling has focused on the ability to predict the productivity of fermentation with little to no focus on product quality.

The objective of this project was to accelerate model development implementations by developing state-of-the-art mathematical models that can analyze and simulate the Fusidic Acid fermentation process. Due to its status as an antibiotic, there is a growing need to improve the yields of Fusidic Acid at the industrial site hosted by the LEO Pharma A/S. However, any attempts to push productivity can lead to increased byproduct formation, which needs to be tightly controlled, especially in a pharmaceutical setting. The developed models are designed with the end goal in mind to predict the harvest of Fusidic Acid

and the formation of a particular byproduct that is extremely difficult to separate. These models can then be applied to gain a deeper process understanding, test new process conditions, and be used as soft sensors. Data is needed to build the models, and one of the essential variables used in virtually all fermentation models is the concentration of cells in the medium. More specifically, the concentration of *living* or viable biomass.

This thesis describes how two models were built to predict concentrations of the main product and the byproduct accurately. One of the models is a purely statistical model that uses batch data collected in real-time from a currently active production and creates an Input-Output correlation. The novelty of the method is that it is the only tool that can directly model batches of various duration without a complicated preprocessing step known as batch trajectory synchronization. It was also the only chemometric model that could predict both batch productivity and quality, whereas traditional methods could only predict productivity.

The second model is a hybrid model, which combines the available scientific knowledge with machine learning in a synergistic way. Biological systems are highly complex, and it can take years of research to gather the available knowledge required to capture all the relevant process phenomena accurately. On the other hand, data-driven models do not require extensive knowledge but infer patterns and knowledge from data. Before developing the first principles model, an experimental procedure was designed to collect the relevant data. A new linear calibration methodology was discovered that allows the conversion of dielectric spectroscopy data to viable biomass concentration, which gives quick and reliable estimates of viable biomass. This information was then used to calibrate a mechanistic model that could accurately and reliably describe biological growth and main product concentrations. With the mechanistic approach, the predictive qualities improved significantly, with an average prediction error of 6.6%. Furthermore, the investigation into model uncertainties revealed that the model structure had limited uncertainty propagation when simulating the process, indicating that the model is a good representative of the physical system. Neural Networks were then directly integrated into the mechanistic model to find the hidden patterns that relate the current batch culture conditions to changes in byproduct concentrations. The final hybrid model is a kinetic description of the process bound by conservation laws that can describe the growth and consumption

profiles of biomass, substrates, main product, and byproducts of the Fusidic Acid fermentation. Byproduct accumulation in the fermenter could be accurately predicted with an average prediction error of 22.8% throughout a full fermentation period. The primary indicator of batch quality is the concentration of byproducts, and this model can determine whether a batch meets quality criteria. Furthermore, simulations can adjust process conditions and discover substrate control methods that completely eliminate the presence of byproducts during harvest or find alternative scenarios that increase the batch outputs of the main product while still maintaining the quality criteria.

Resumé

Bioteknologi spiller en integreret rolle i den moderne økonomi og er ansvarlig for at producere forskellige bulk- og specialkemikalier, der er kritiske for samfundets funktion. Bioprocessen går ud på at bruge mikroorganismer som en miniaturefabrik til at omdanne substrater til værdifulde molekyler. Antibiotika er antimikrobielle stoffer, der bruges til at bekæmpe bakterielle infektioner og produceres almindeligvis via fermentering. Efterhånden som global antimikrobiel resistens bliver en større og større trussel mod menneskeheden, har der været en øget interesse for potentialet i den gamle generation af antibiotika til at imødekomme det nuværende behov for nye antibiotika. Stillet over for stigende efterspørgsel er vi nødt til at modernisere de nuværende produktionsmetoder. Den kemiske og biokemiske industri er på vej gennem den fjerde industrielle revolution, eller Industry 4.0, som er sammenføjningen af teknologier, der udvisker grænserne mellem den fysiske, digitale og biologiske verden.

Digitale tvillinger baseret på procesmodeller er en afgørende teknologi for industrier i det skiftende konkurrencelandskab efter skiftet til Industri 4.0. Digitale tvillinger er en digital repræsentation af et fysisk produkt, system eller proces i den virkelige verden. Denne teknologis fundament og primære muliggører er en matematisk model, der nøjagtigt fanger de relevante fysiske og biologiske fænomener. Imidlertid er modellering af fermenteringssystemer udfordrende på grund af biologiske kompleksiteter. Derudover har medicinalindustrien altid haft stor vægt på kvalitet. Lovlige bestemmelser specificerer den påkrævede renhed af aktive farmaceutiske ingredienser og sætter grænser for potentielt skadelige eller virkningsreducerende urenheder. Disse urenheder kan være biprodukter af selve bioprocessen og kan have meget lignende kemiske og fysiske egenskaber, hvilket gør dem umulige at adskille i rensningsprocessen. Hidtil har bioprocesmodellering fokuseret på evnen til at forudsige produktiviteten af fermentering med lidt eller intet fokus på produktkvalitet.

Formålet med dette projekt var at lægge et af nøglegrundlaget for digitale tvillingeimplementeringer ved at udvikle avancerede matematiske modeller, der kan analysere og simulere Fusidinsyre-fermenteringsprocessen. Disse modeller er designet med det endelige mål for øje at forudsige høsten af fusidinsyre og dannelsen af et bestemt biprodukt, som er ekstremt vanskeligt at adskille. Disse modeller kan derefter anvendes til at opnå

en dybere procesforståelse, teste nye procesforhold og bruges som bløde sensorer. Data er nødvendige for at bygge modellerne, og en af de væsentlige variabler, der bruges i stort set alle fermenteringsmodeller, er koncentrationen af celler i mediet. Mere specifikt koncentrationen af *levende* eller levedygtig biomasse.

Denne afhandling beskriver, hvordan to modeller blev bygget til præcist at forudsige koncentrationer af hovedproduktet og biproduktet. En af modellerne er en rent statistisk model, der bruger batchdata indsamlet i realtid fra en aktuelt aktiv produktion og skaber en Input-Output korrelation. Det nye ved metoden er, at det er det eneste værktøj, der direkte kan modellere batches af forskellig varighed uden et kompliceret forbehandlingstrin kendt som batch-banesynkronisering. Det var også den eneste kemometriske model, der kunne forudsige både batchproduktivitet og kvalitet, hvorimod traditionelle metoder kun kunne forudsige produktivitet.

Den anden model er en hybridmodel, som kombinerer den tilgængelige videnskabelige viden med maskinlæring på en synergistisk måde. Biologiske systemer er meget komplekse, og det kan tage mange års forskning at indsamle den tilgængelige viden, der kræves for at fange alle de relevante procesfænomener præcist. På den anden side kræver datadrevne modeller ikke omfattende viden, men udleder mønstre og viden fra data. Før modellen udvikledes, blev der designet en eksperimentel procedure til at indsamle de relevante data. En ny lineær kalibreringsmetodologi blev opdaget, der tillader konvertering af dielektriske spektroskopidata til levedygtig biomassekoncentration. Denne information blev derefter brugt til at kalibrere en mekanistisk model, der nøjagtigt og pålideligt kunne beskrive biologisk vækst og hovedproduktkoncentrationer. Neurale netværk blev derefter direkte integreret i den mekanistiske model for at finde de skjulte mønstre, der relaterer de nuværende batchkulturbetingelser til ændringer i biproduktkoncentrationer. Den endelige hybridmodel er en kinetisk beskrivelse af processen bundet af bevaringslove, der kan beskrive vækst- og forbrugsprofilerne for biomasse, substrater, hovedprodukt og biprodukter fra Fusidinsyre-fermenteringen.

List of dissemination activities

Submitted Manuscripts

A.F. Magnússon, J. Pajander, G. Sin, S. Stocks. (2022). Determining the linear correlation between dielectric spectroscopy and viable biomass concentration in filamentous fungal fermentations, *Biotechnology Letters*. *Manuscript in review as of October 2022*.

Manuscripts in preparation

A.F. Magnússon, G. Sin, S.M. Stocks, J. Pajander (2023). Multi-modal regression models of batch data without batch trajectory synchronization. *In preparation for Journal of Chemometrics*.

A.F. Magnússon, S.M. Stocks, J. Pajander, G. Sin (2023). Development and reliability assessment of unstructured mechanistic model for filamentous fed-batch processes. *In preparation*.

A.F. Magnússon, S.M. Stocks, J. Pajander, G. Sin (2023). Hybrid modeling extension to simulate byproduct formation in an antibiotic fed-batch process. *In preparation*.

List of conference proceedings

A.F. Magnússon, R. Al, G. Sin, S.M Stocks, (2020). Development and Application of Simulation-based Methods for Engineering Optimization Under Uncertainty, Vol. 48, pp. 451-456.

A.F. Magnússon, J. Pajander, G. Sin, S.M Stocks, (2022). Multimodal modelling of uneven batch data. *14th International Symposium on Process Systems Engineering*, Vol. 14, pp. 2143-2148.

List of conference participations

A.F. Magnússon, R. Al, G. Sin. Development and Application of Simulation-based Methods for Engineering Optimization Under Uncertainty. Oral contribution at Computer Aided Process Engineering (CAPE) Forum at ULiège, Liege, Belgium (2019).

A.F. Magnússon, R. Al, G. Sin. Development and Application of Simulation-based Methods for Engineering Optimization Under Uncertainty. Oral contribution at the

30th European Symposium on Computer-Aided Process Engineering (ESCAPE 30), Virtual meeting (2020).

A.F. Magnússon, S.M. Stocks, J. Pajander, G. Sin. Hybrid modelling of industrial scale fermentation process. Oral and poster contribution at KT Consortium Annual Meeting, Virtual meeting (2021).

A.F. Magnússon, G. Sin, S.M. Stocks, J. Pajander. Multi-Modal modelling methods: Applications in batch bioprocesses. Oral contribution at American Institute of Chemical Engineers (AIChE), Boston, United States of America (2021)

A.F. Magnússon, S.M. Stocks, J. Pajander, G. Sin. Hybrid modelling of industrial scale fermentation process. Poster contribution at Nordic Process Control Workshop, Luleå, Sweden (2022).

A.F. Magnússon, G. Sin, S.M. Stocks, J. Pajander. Three-Dimensional modeling of bioprocess batch data. Poster presentation at KT Consortium Annual Meeting, Helsingør, Denmark (2022)

A.F. Magnússon, S.M. Stocks, J. Pajander, G. Sin. Hybrid modelling of industrial scale fermentation process. Oral and poster contribution at the 32nd European Symposium on Computer-Aided Process Engineering (ESCAPE 32), Toulouse, France (2022)

A.F. Magnússon, G. Sin, S.M. Stocks, J. Pajander. Multimodal modelling of uneven batch data. Oral contribution (Presented virtually due to Covid-19) at Process Systems Engineering (PSE) symposium, Kyoto, Japan (2022)

Contents

List of Figures

List of Tables

Nomenclature

Roman

ΔC	Permittivity increment in the radio frequency region
$\dot{m}_{H_2O}$	Water vapor flow rate
$\hat{f}_{RS}$	Estimated biochemical rate change of related substance
u	Vector containing batch inputs
x	Vector containing batch states
y	Vector containing batch outputs
a	$k_l a$ equation fitted parameter
b	$k_l a$ equation fitted parameter
c	$k_l a$ equation fitted parameter
C_i	Concentration of component i
C_i^*	Solubility of component i in broth
$C_{i,f}$	Feed Rate of component i
C_{RS}	Concentration of component related substance
E	Evaporation rate
F_{CER}	Mass flow of CO_2
$F_{evaporation}$	Evaporation flow rate of water
F_{feed}	Feed Rate
F_{OUR}	Mass flow of O_2
k_d	Specific cell death rate
$k_l a$	Gas to Liquid mass transfer coefficient of component i

K_S	Saturation constant for biological growth
K_{OX}	Oxygen limitation constant
K_{SP}	Substrate limitation constant
K_{SS}	Maintenance saturation constant
K_{SX}	Contois saturation constant
M	Broth mass
m_S	Substrate maintenance term
M_{frac}	Mass fraction of non-heat-treated broth in diluted samples
$p^*_{H_2O}$	Satured vapor pressure for water
P_o	Pressure at bottom of the tank
$P_{agitator}$	Energy dissipation from agitator
P_{air}	Energy dissipation from aeration
P_{power}	Energy dissipation
$P_{Precipitated}$	Concentration of precipitated non-biological solids
Q	Process air flow rate
q_D	Cell death rate
q_i	Biochemical rate of component i
q_{CER}	Carbon evolution rate
q_{OUR}	Oxygen uptake rate
R	Ideal gas constant
S	Main carbon source concentration
T	Temperature
t	Time
t_{lag}	Lag time

v_g	Superficial gas velocity
X	Biomass concentration
X_d	Cell debris concentration
X_{TDW}	Total Dry Weight concentration
X_{Viable}	Viable biomass concentration
Y_{SP}	Product substrate yield coefficient
Y_{SX}	Biomass substrate yield coefficient
Z	Broth liquid height

Greek

α	Dry weight viability fraction
α_{opt}	Optimal dry weight viability fraction
β_1	Slope of dielectric spectroscopy linear calibration
β_2	Dielectric spectroscopy linear calibration offset
ϵ	Resdiual
μ_i	Specific biochemical rate of component i
μ_{CER}	Specific carbon evolution rate
μ_{max}	Maximum specific growth rate
μ_{OUR}	Specific oxygen uptake rate
$\mu_{P,max}$	Maximum specific product synthesis rate
μ_P	Specific product synthesis rate
μ_S	Specific substrate consumption rate
$\mu_{X,max}$	Maximum specific biomass growth rate
μ_X	Specific growth rate
ϕ	Relative Humidity

ρ	Broth density
θ	Parameter vector

Abbreviations

ALS	Alternating Least Squares
ANN	Artifical Neural Networks
API	Active Pharmaceutical Ingredient
CDW	Cell Dry Weight
CER	Carbon Evolution Rate
CFD	Computational Fluid Dynamics
CPP	Critical Process Parameter
CQA	Critical Quality Attribute
DNN	Deep Neural Networks
GMP	Good Modeling Practices
$GTucker2$	Generalized Tucker2
$HOPLS$	Higher Order Partial Least Squares
LHS	Latin Hypercube Sampling
$MCOVR$	Multimodal Covariate Regression
$MPCA$	Multiblock Principal Component Analysis
$MPLS$	Multiblock Partial Least Squares
$NPLS$	Multilinear Partial Least Squares
ODE	Ordinary Differential Equation
OUR	Oxygen Uptake Rate
PAT	Process Analytical Technology
PCA	Principal Component Analysis

PLS	Partial Least Squares
PSE	Process Systems Engineering
QbD	Quality by Design
RE	Relative Error
$RMSECV$	Root mean squared error for Cross Validation
$RMSEP$	Root mean squared error for Prediction
$RMSSE$	Root mean square
$RMSSE$	Root mean sum of squared error
$SCREAM$	Shifted Covariates Regression Analysis of Multiway data

1 Introduction

From the beginning of 2019 to the end of 2022, the world was heavily affected by the COVID-19 pandemic. At the time of writing, the disease is estimated to have led to the loss of life of millions worldwide. Furthermore, the crisis caused severe economic recessions due to necessary countermeasures from which the world is still recovering.
The pandemic clearly shows the disastrous effects of a global health crisis which we are not well equipped to deal with. The World Health Organization (WHO) has identified antibiotic resistance as one of today's biggest threats to global health, food security, and development[1]. Antibiotic resistance occurs when bacteria develop new resistance mechanisms to commonly prescribed antibiotics, threatening the ability to treat common infectious diseases. Without urgent action, the world could head for a post-antibiotic era, in which common infections can kill again.

There has been renewed interest in the potential of old-generation antibiotics[2] recently due to the global problem of advancing antimicrobial resistance. Even though they are not a permanent solution, an increased variety of antibiotics can buy valuable time required while implementing more long-term solutions such as social behavior changes.
A spore-forming filamentous fungus produces Fusidic acid. It has been in use at LEOPharma A/S for commercial production of Fusidic Acid since 1962[3], and there has always been a strong focus on the quality of the medicine, especially concerning its related substances, of which there are many[4]. To date, the production process has only been internally modeled via empirical means (i.e., in a lab setting) because it is tough to understand the kinetics of the production of fusidic acid and its related substances. This difficulty largely stems from a lack of research into the microorganism itself. Indeed there are little to no examples in literature that explore microbiology or the biochemistry relating to the fermentation of the Fusidic Acid production strain.

With the advance of new digital tools, the Process Systems Engineering (PSE) commu-

nity is shifting its focus to more sustainable solutions for engineering problems. Modern process systems must comply with more stringent regulations to meet today's sustainable development goals (SDGs)[5]. Production of antibiotics is generally not directly associated with the ever-encroaching global problem of climate change. However, sustainability is also focused on people, and within the SDGs is also a focus on "Good Health and Well-Being" to ensure healthy lives and promote well-being for all ages. We need ready access to affordable, high-quality medicine to achieve this goal. Furthermore, while antibiotics are commonly produced in a bioprocess, the process consumes a lot of clean water, becoming a more precious resource every year. The entire production chain, from Active Pharmaceutical Ingredients (APIs) to pharmacies, also consumes a lot of resources, for example, in product purification, which commonly uses solvents sourced from the petrochemical industry. Improving the efficiency of the production process and reducing the failure rate will contribute to these SDG goals by making more affordable medicine while consuming fewer precious resources.

The manufacturing industry is trending toward the fourth industrial revolution, also known as Industry 4.0. This is the propagation of digitalization and the Internet of Things[6]. The backbone of digitalization efforts is a high-fidelity model that can represent the physical system accurately. Unfortunately, the bioprocess industry seems to be adopting it slower than many other industries such as synthetic chemical counterparts. A major obstacle that hinders the full transition of the current bioprocesses to Industry 4.0 is the inherent complexity of biological systems, making high-fidelity model developments a slow and expensive task.

To serve as many patients as possible affordable medicine is needed. This can only be achieved by driving manufacturing costs down while increasing the production of high-quality pharmaceuticals. This requires innovations in managing and controlling production processes to minimize resource consumption. This is best achieved with suitably predictive mathematical models. Therefore, this project aims to develop new and innovative models, not only to predict the production of pharmaceuticals but also, for the first time to predict the production of unwanted related substances that may hamper the product quality and might be dangerous to the customers. The proposed models here will be based on a modern data-driven approach using knowledge-based scientific and engineering prin-

ciples. A verified functional mathematical model can be further incorporated as a digital twin that will eliminate guesswork enabling intelligent decision-making, thereby reducing months or years of experimental verification of changes in all scales from the laboratory to full production scale where failures are catastrophically expensive.

Most bioprocess modeling literature focuses on building models to predict batch productivity, with little to no attention to accumulating impurities in the batch process. However, this is a general problem in all bioprocesses but is especially relevant for pharmaceuticals. Pharmaceuticals are under much stricter quality control. Taking another look at Penicillin, the Europa pharmacopeia specifically states limits of impurity levels. Specifically for Penicillin V, there are six identified impurities that must not exceed a given value[7]. Yet, little to no published literature focused on modeling the growth or synthesis of these substances. Making the process more efficient in terms of the main product is valuable for the economics of the product. Still, if the final product doesn't meet quality standards, it can't even be sold, making any productivity optimizations irrelevant. Furthermore, suppose impurities are a potential issue in the production process. In that case, the models used for monitoring and controlling the batch should consider accumulation to reduce the failure rate and increase batch-to-batch consistency.

1.1 Research goal and scope

The main hypothesis in this PhD thesis is that novel data-driven methods and the integration of machine learning into traditional mechanistic model structures can accelerate model development of complex biological systems such as fermentation. The core objective of this project is to apply state-of-the-art modeling methodologies to an established industrial production using a microorganism with relatively minimal research. The proposed models have to meet the criteria of explaining the formation of the main product and the evolution of a specific impurity. Therefore, the scope is set on creating, validating, and analyzing these models in a specific industrial scale setting. We've divided the project into four more specific research objectives throughout the research. They are as follows:

- **Research Objective 1:**

  Development of an experimental protocol to directly measure viable biomass of filamentous fungi.

- **Research Objective 2:**

  Use Multivariate techniques to create process models that predict the main products and related substances in an industrial environment.

- **Research Objective 3:**

  Development and analysis of a mechanistic model of a novel filamentous fungi strain.

- **Research Objective 4:**

  Integrate data-driven approaches with mechanistic models in a synergistic way to fill in relevant knowledge gaps.

1.2 Thesis structure

Chapters 1,2 serve as an introduction and conclusion to the thesis and the current state-of-the-art relevant to the research topics. Subsequent chapters 3-6 will each be dedicated to a research objective, structured similarly to journal articles. This is because each chapter will be used as a baseline to create a manuscript for publication in research journals. The overview of the different chapters is as follows:

Chapter 1 Introduction

The first chapter briefly introduces this thesis's research goals and motivations and provides an outline of how to read this thesis.

Chapter 2 Background

This chapter gives a brief description of the fed-batch, which is used for the industrial production of antibiotics. Also, the chapter provides an overview of state-of-the-art bioprocess modeling within the three significant categories of data-driven, mechanistic and hybrid modeling. Finally, the chapter provides a brief overview of biomass measurements and the importance of viability.

Chapter 3 Determining viable biomass concentration via cell capacitance

This chapter develops an experimental protocol for measuring biomass with dielectric spectroscopy equipment and proposes a linear methodology for directly measuring viable biomass concentration in filamentous fungi. The chapter focuses on **Research Objective 1**

Chapter 4 Multimodal modeling of industrial-scale fermentation

This chapter explores the use of a novel multi-modal regression technique called Shifted Covariates Regression (SCREAM) and its use in predicting the harvest of main product and final batch quality w.r.t. byproducts in industrial fed-batch data. The method is compared to traditional multiblock and multi-modal techniques used in the bioprocess industry. The chapter focuses on **Research Objective 2**

Chapter 5 Mechanistic Modelling of Industrial scale batches for antibiotics

This chapter sets up an equation structure for a mechanistic fed-batch model of the filamentous organism using viable biomass as a key process variable. The model is identified and further analyzed within the Good modeling framework for PAT applications. The chapter focuses on **Research Objective 3**

Chapter 6 Hybrid Modelling for fermentation batch quality

This chapter focuses on further expanding the model capabilities outlined in chapter 5 by integrating Neural Networks directly into the model structure in a serial hybrid model structure. The Neural Networks thus allow the hybrid model to predict everything outlined in chapter 5 and Byproduct concentration, CO_2 evolution, O_2 uptake, and pH, whose underlying biochemistry is not understood. An example application is testing model outputs with different feed rate strategies. The chapter focuses on **Research Objective 4**

Chapter 7 Conclusions and Future perspectives

This chapter summarizes the main achievements of the projects and concludes by discussing some remaining challenges and future perspectives.

1.3 Note on confidentiality

Most of the data used in this work are confidential by LEO Pharma A/S and can not be published unedited. To minimize the loss in understanding and explanation of phenomena, it has been decided to use a scaling factor for all data related to the product yields present in the main text, tables, and figures. Nominal model parameters that explain the primary process will not be revealed. Specific data acquisition methods for data provided by the company and not directly obtained by the author will not be discussed. Also, the actual numbers for various cultivation conditions and scales are not shown. Certain graphs have axis labels removed to preserve trends and model quality while not showing real numbers.

Bibliography

[1] Ousmane Oumou Diallo, Sophie Alexandra Baron, Cédric Abat, Philippe Colson, Hervé Chaudet, and Jean Marc Rolain. Antibiotic resistance surveillance systems: A review. *Journal of Global Antimicrobial Resistance*, 23:430–438, 2020.

[2] Matthew E Falagas, Alexandros P Grammatikos, and Argyris Michalopoulos. Potential of old-generation antibiotics to address current need for new antibiotics. *Expert Review of Anti-infective Therapy*, 6(5):593–600, 2008. PMID: 18847400.

[3] Wo Godtfredsen, S Jahnsen, L Tybring, K Roholt, and H Lorck. Fusidic acid - new antibiotic. *Nature*, 193(4819):987–, 1962.

[4] WO Godtfredsen, N Rastrup-Andersen, S Vangedal, and WD Ollis. Metabolites of fusidium coccineum. *Tetrahedron*, 35(20):2419–2431, 1979.

[5] Joseph N. Pelton. Un sustainable development goals for 2030. *Handbook of Small Satellites*, pages 1537–1566, 2020.

[6] Carina L. Gargalo, Simoneta Caño de Las Heras, Mark Nicholas Jones, Isuru Udugama, Seyed Soheil Mansouri, Ulrich Krühne, and Krist V. Gernaey. Towards the development of digital twins for the bio-manufacturing industry. *Advances in Biochemical Engineering/biotechnology*, 176:1–35, 2021.

[7] Council of Europe. *European Pharmacopoeia (Ph. Eur.) 10th Edition*. Strasbourg, 2019.

2 Background

2.1 Industrial fermentation of antimicrobics

The production of secondary metabolites has been the subject of many studies because of its industrial importance. For example, antibiotics are commonly the secondary metabolite of filamentous microorganisms. Most of the global supply of antibiotics can be traced to the fermentation of these microorganisms[1]. The consequence of being a secondary metabolite is that the target product is not associated with cell growth. For this reason, it is common practice to grow the cell culture before the main production starts. Industrial bioreactors are extremely large, and a seed train commonly accomplishes the required cell density. This is the scaling of the cell culture density from a small volume of cells in a cell bank vial to a larger volume via repeated batch fermentations in different bioreactor sizes[2]. The main production reactor is usually operated in a fed-batch mode to promote the synthesis of the antibiotic. A fed-batch operation is characterized by predetermined or controlled addition of nutrient medium in an otherwise batch operation. It was developed to increase biomass yield in the production of Baker's yeast[1]. For antibiotic production, Fed-batch modes have notable advantages over a standard batch operation. They allow for tighter control of various cellular processes. Synthesis of secondary metabolites is usually promoted when cell growth is discouraged or limited. Proper control of nutrient additions can thereby lead to conditions where cell growth is discouraged and thus promotion of antibiotic synthesis.

In aerobic fermentations where oxygen is required, the fed-batch allows considerably more feedstock to be converted to a product than in a simple batch. In aerobic fermentations, oxygen supply is often a limiting factor due to its poor solubility in water. In a batch process, all the feedstock will be included in the initial blend, leading to increased demand for oxygen at rates surpassing what can be supplied. This will lead to batches with short durations and low product concentrations[3]. Conversely, a fed batch can supply nutrients at a rate that keeps oxygen demand low, allowing for longer batch durations with higher product concentrations. High product concentration, often referred to as a titer, is

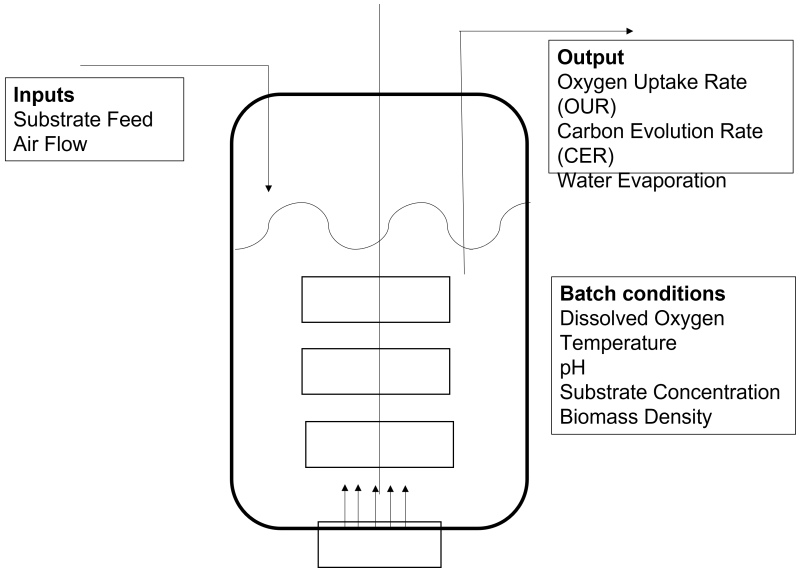

Figure 2.1: Sketch of a fermentor running in a basic fed-batch setup

one of the primary descriptors describing a bioprocess's economics. This is because high product concentrations mean less downstream processing, leading to savings on recovery and water recycling and disposal. Long-duration batches are also preferred because of reduced downtime from restarting short batches, which will increase productivity.

2.2 Fermentation models

Models use the critical process parameters (CPPs) as inputs to predict critical quality attributes (CQAs) of, e.g., the desired product. Once the model structure is established, the model parameters must be determined. Afterward, it's crucial that the model is properly validated by evaluating models on independent test data. Furthermore, it is vital to prove the model's reliability by relying on the eventual application. Several methods and tools exist to determine the credibility of models, including identifiability, uncertainty, and sensitivity analysis[4]. The promising potential of these modeling tools combined with pilot-scale validation has been successfully demonstrated for fungi/yeast fermentation for industrial enzymes[5] and lab-scale fermentations for starter cultures[6]. However, this potential has been untapped for fusidic acid production or the quality of the fermentate. Batch quality can contain a broad scope, so for this work, any mention of fermentation or batch quality will refer to the concentration of related substances. There are multiple approaches to identifying a model structure before any training or calibration is done. The proper method is selected based on model purpose and access to relevant data. Biopro-

cess models can be divided into three major categories: data-driven, mechanistic, and hybrid. Model explainability refers to the concept of being able to understand the model so that the results are trustworthy and any model biases can be explained. The mechanistic model translates existing knowledge about a process into governing equations and provides valuable insights into the underlying behavior of the system. Data-driven methods are more difficult to interpret as model developers are not aware of how the input variables are combined and processed to give a certain output beyond a simple linear regression. However, they can discover hidden patterns without prior knowledge. The explainability of hybrid models is dependent on what extent data-driven methods are incorporated into a mechanistic model structure. They contain all the available process knowledge but can also reveal the underlying structures or patterns in the data regarding poorly understood aspects of the physical system.

2.2.1 Data Driven models

Data-driven models are purely empirical approaches. When there is limited process understanding but a wealth of process data is available, the empirical models are an attractive approach to modeling the process. Data-driven methods include modern machine learning algorithms such as Neural Networks and more traditional multivariate modeling methods.

Multivariate modeling techniques refer to methods such as Principal Component Analysis (PCA) and Partial Least Squares (PLS). Both methods can be applied to identify trends within a large multivariate dataset or be used for process modeling PCA is not an Input-Output model but rather a dimensionality reduction technique. PLS is a tool that can be used for linear calibration and modeling of a multivariate input signal to realize the process variable.

PCA and PLS, on their own, take no definitive account of the ordered nature of the dataset, i.e., that the data is collected in a sequential matter. This is a flaw when using these methods directly on batch data due to the dynamic nature of fermentation. It is expected that process variables change over time. Data collected from batch or fed-batch processes have a three-dimensional structure assuming that process variables are being measured continuously. Multi-way models are thus utilized in practice for data-driven models for batch data. Multi-way principal component analysis (MPCA) and multi-way partial least

squares (MPLS) proposed by Nomikos and MacGregor[7] have been successfully used in modeling batch data[8]. These algorithms do not result in three-dimensional models because the methodology involves unfolding the three-dimensional data structure, thereby converting the data into a standard two-dimensional matrix. Ordinary PCA and PLS algorithms are then applied to the unfolded process data. Despite widespread application, MPCA and MPLS have been criticized due to two primary drawbacks. Unfolded data treats each measurement of the same variable as an independent variable. Thus, unfolded datasets may have a sizeable variable number but a small sample size, leading to unreliable estimates of model parameters[9]. Furthermore, unfolding a dataset destroys the three-way structure meaning that MPCA and MPLS can not offer an explicit description of any potential three-way interactions. These disadvantages may reduce monitoring performance, prediction ability, and interoperability. Another disadvantage of MPCA and MPLS is that these are linear methods, while fermentation is essentially a non-linear process. In case of significant non-linear characteristics in the dataset, model developers must utilize a non-linear modeling method such as Neural Networks[10] or modify the linear formulations of PCA and PLS to use a nonlinear kernel projection into a high-dimensional feature space[11][12].

Recently there has been an increase in research in analyzing batch data using tensor analysis methods[13]. These algorithms keep the batch data structure's three-way representation by explicitly modeling each dimension as tensors. This usually leads to models that have better intuitive interpretability. Tensor models also commonly have much fewer parameters because the data is compressed in three directions which generally leads to more stable models. There are multiple tensor models available, and the selection should be based on the dataset's nature and the model's purpose. Parallel Factor Analysis (PARAFAC) and Tucker decomposition are dimensionality reduction models that are well known and are sometimes described as multi-way generalized PCA[14]. These models act as a substitute for MPCA and are commonly used for batch process monitoring[15]. For regression purposes and multilinear calibration, it is common to use one of these; N-way Partial Least Squares (NPLS), Higher-Order Partial Least Squares (HOPLS), and Multiway Covariates Regression (MCOVR)[16][17].

A case of tensor methods that may be of particular interest is decomposition methods

based on the PARAFAC2 algorithm[18]. PARAFAC2 has been used successfully in fault detection for semiconductor production[19]. Similarly, a generalized Tucker2 (GTucker2) model has been proposed for monitoring a penicillin-fed-batch process[19]. The advantage of these models is that they allow for uneven-length data sets, i.e., batches can have varying durations without needing a complex time-alignment preprocessing step. Multilinear methods that allow a relaxation in one mode are, therefore, almost perfectly suited for modeling industrial fed-batch processes because they handle the two most difficult aspects of the nature of the dataset; They are more likely to avoid the curse of dimensionality induced by data unfolding and naturally solve the uneven length problem without batch trajectory synchronization. For regression purposes, there is a recent development with Shifted Covariates Regression (SCREAM) model[20], which combines PARAFAC2 and MCOVR algorithms into a single method. This is the only uneven tensor algorithm for regression found in the literature and has so far not been applied to fermentation data.
Input-Output models are much less expensive to develop. However, data-driven models only match the conditions in the experiments that were made but do not immediately inform the scientist very much about the underlying principles of the process. Since no scientific knowledge is required to build the models, they are unaware of any first principles. Data-driven models may lead to predictions that conflict with fundamental constraints like conservation principles, particularly when outside the domain of training. Purely data-driven models are usually limited in the application of monitoring and control of an already established process and are not suitable for process design, intensification, and optimization.

2.2.2 Mechanistic models

Mechanistic models are based on fundamental physical principles such as conversation laws like mass, energy, and momentum balances. Mechanistic models, typically consisting of differential equations approximating the kinetics of a process solved by a numerical approach, are preferred as they contain the relevant scientific knowledge needed to describe a system's behavior adequately. The most common model seen in the literature is some variation of the Monod model. This empirical relation shares the mathematical form

of the even more famous Michaelis-Menten enzyme kinetics[21].

$$\frac{dX}{dt} = \mu_{max} \frac{S}{K_S + S} X \tag{2.1}$$

In general mechanistic models describing biomass growth can be classified into unstructured and structured models[22]. Structured models include a fairly detailed description of the important reactions inside the cells. They represent a good understanding of the bioprocess. However, these models are difficult to obtain due to requiring knowledge about reaction kinetics, thermodynamics, transport, and physical properties. Unstructured models are simpler and based on pooling all cellular components into a single representative biomass concentration. The biochemical reaction kinetics are then described as some function of component and biomass concentrations in the media and other cultivation conditions such as pH and temperature. These functions are usually empirical, and the Monod model is an example of an unstructured mechanistic model. Unstructured models are much easier to make while still commonly applicable to real fermentation systems.
The Monod expression is not universally applicable. There are dozens of variations of Monod-type kinetics[23][24] to describe bacterial growth. A way to limit the scope of available model structures and the associated complexities is to review models based on similar microorganisms. To that end, we narrow the focus to Penicillin fermentation models. Penicillin fermentation shares many similar characteristics to the process being studied in this work. The microorganism *Penicillium chrysogenum* is a filamentous fungus with a secondary antibiotic metabolite, and industrial production is usually carried out in a fed-batch process with similar cultivation conditions. One of the prominent early examples is the unstructured model of Bajpai and Reuss[25], which was shown to give good agreement with experimental results from Pirt and Righoletto[26]. This model has been used as a basis for developing modular simulation packages for the penicillin fermentation process [27][28] and is an excellent inspiration for designing an equation structure for modeling different filamentous microorganisms. Further research into mechanistic modeling of *Penicillium chrysogenum* fungus has led to the development of an industrial-scale fed-batch fermentation simulator by Goldrick et al.[29]. It's a highly detailed work showing the applicability of mechanistic models. It has been extensively used to test various batch process monitoring and control methods[30]. The biotechnological industry increas-

ingly applies mechanistic models because it has realized their significance[5][31]; they are advantageous for predicting dynamic system behavior in different scenarios, while data-driven or statistical models sometimes fail[32].

2.2.3 Hybrid model

Biological systems are notoriously complex, so sufficient knowledge is rare and difficult to obtain. For example, the previously mentioned industrial fed-batch simulator utilizes a morphologically structured model from Paul and Thomas[33]. Incorporating such details requires years of extensive research into the morphological dynamics of the fungus. In contrast, modern statistical/data-driven models rely only on empirical or historical data, the correlations between inputs and outputs detected by choice of various Machine Learning algorithms with PLS or Artificial Neural networks (ANN)". As such data-driven models might match the conditions, the experiments were made but did not immediately inform the scientist very much on the underlying principles of the process and thus can be unreliable when exploring new designs.

Hybrid models are not a new invention. Research into the development of modeling approaches that combine mechanistic and data-driven elements started appearing in the 1990s; an example in the bioprocess industry dates back to 1994 with the work of Schubert et al.[34] However, hybrid models are receiving increased interest in recent years[35]. This interest is spurred on by technological developments where we now have easy access to massive amounts of data, advanced analytical tools, and computation resources. This new world, along with the emergent Industry 4.0, is opening up opportunities to provide unique solutions to old problems. Hybrid models are a broad category with many design decisions on identifying appropriate models. White and black-box models can be organized in serial and parallel structures. The type of available information and purpose is used to define the hybrid structure. The parallel structure finds good use in situations where a mechanistic model can predict the relevant phenomena but has limited predictive power or accuracy due to potential unknown effects[36][37]. The final prediction is a fusion of outputs from mechanistic and data-driven models. They are directly combined via addition, subtraction, etc.. or a combination via a weighing function[38]. An established full mechanistic explanation must already be present to make a parallel hybrid model.

For this work, there is more interest in the serial structure. The serial design where the

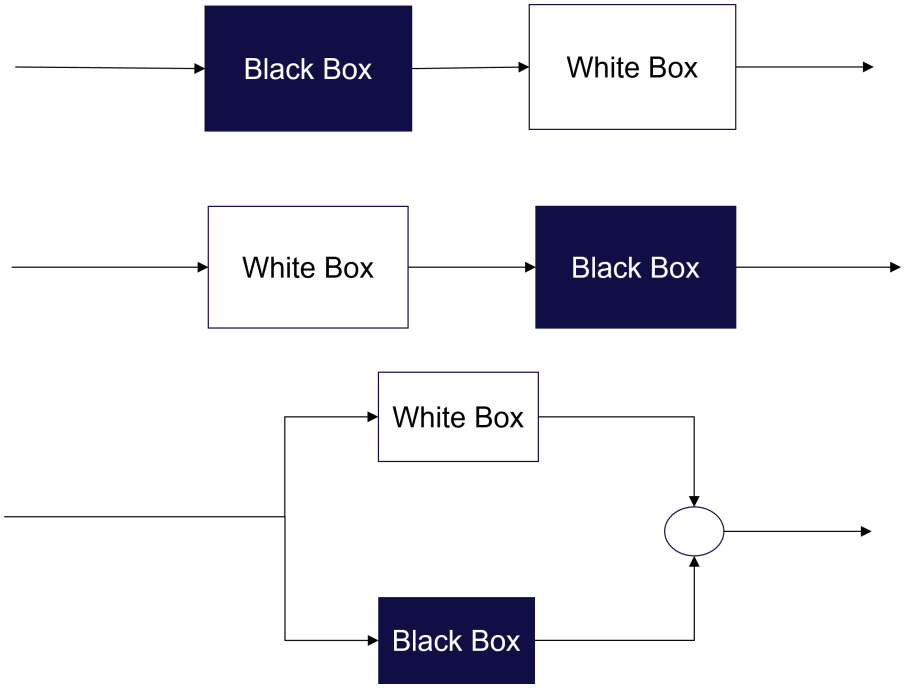

Figure 2.2: Illustration of available hybrid modeling structures commonly used in research and industry today.

black-box model outputs are fed into the mechanistic equations is more common and has seen widespread success for modeling chemical and biochemical processes[39][40][41]. The mechanistic model usually represents this structure's fundamental conservation laws and transport phenomena. The data-driven model is then used for the part of the biological phenomenon that either lack knowledge or is too complex to formulate a proper equation structure. Integration of data-driven models allows the prediction of unknown biochemical reaction kinetics. It is argued that this approach extrapolates better than a pure data-driven model and is more reliable and interpretable. Furthermore, tying ANNs to mechanistic models reduces the complexity of ANN leading to higher accuracy models[42].

While Neural Networks are often used as the data-driven component, it is not required. Other machine learning algorithms, such as Support Vector Machines, have been successfully integrated into a hybrid model structure describing a fermentation process[43]. ANNs have infinite configurations and can theoretically fit linear and non-linear input and output relationships. When looking at what machine learning model to use, it is clear that there is no correct choice or perfect hybrid model. Rather, a model with good enough accuracy to fit the purpose should be considered, which should be decided based on various validation techniques available.

In modern times, hybrid models are constructed using both approaches, where our knowledge is incorporated, and a data-driven approach takes care of what is not understood[44]. The general goal is to gather all available knowledge while improving prediction accuracies by taking advantage of big data. However, all approaches rely on data, and collection and access to data is something that has only happened in recent years, i.e., technology is at a point where we can now consider such modeling approaches.

2.3 Biomass measurements

Biomass is a critical parameter in a fermentation process. It's a key variable to optimize to reach maximum efficiency for many bioprocess products and, in some cases, is the main product[45]. Every single fermentation mechanistic model uses biomass as a fundamental variable and is present in practically all kinetic functions. Thus, any model development relies on accurate biomass measurements. However, it is often difficult to measure despite being a key variable. There are diverse methods for the quantification of biomass which are useful in different cases, depending on the application.

Mechanistic model development typically uses *off-line* quantification measurements, which is when the result is manually obtained because it's not impeded by time delay. The most widely applied biomass estimation method is the cell dry weight method. This is a method where cells in a sample are separated from the broth and weighed after thorough drying. It's a simple but time-consuming method. A significant criticism of this method is that it cannot distinguish viable cells from dead cells[46]. Furthermore, the process is erroneous if the broth contains other insoluble materials.

2.3.1 Viable Biomass

Being able to distinguish between viable and dead biomass is not often considered when developing mechanistic models. However, it should be evident that dead cells do not grow or produce anything and should not contribute to model predictions. Therefore, models that can accurately predict viable biomass should be more reliable and applicable. Microorganisms have a limited lifespan, and in long fermentation processes such as a fed-batch process, some cells are expected to die and leave behind cell debris which impacts the dry weight. Relying on cell dry weight may lead to experimental observations and conclusions that cells grow indefinitely. In reality, if the nutrient amount is kept

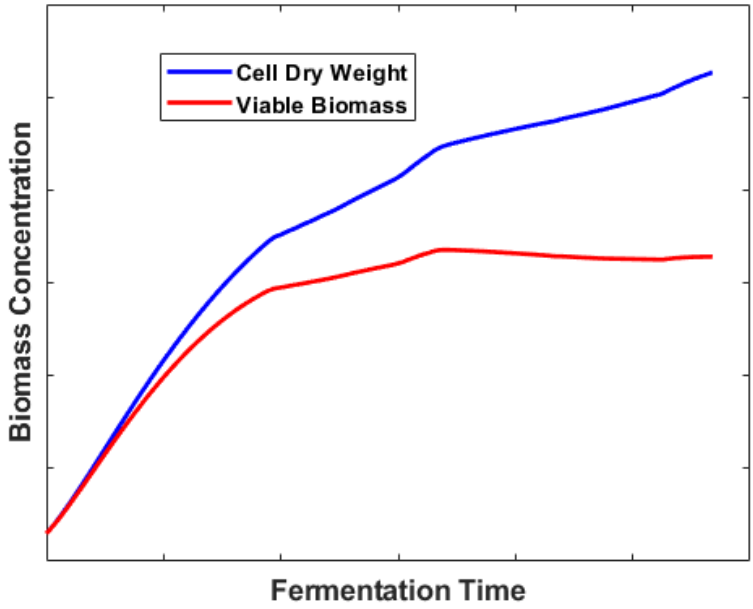

Figure 2.3: Illustration of a potential problem with relying on CDW as the ”gold standard” for biomass

consistent through controlled feeding, the cell culture in the batch will reach a stationary phase where dead cells are replaced by new cells at approximately similar rates[47]. Suppose a mechanistic model predicts indefinite growth because it's calibrated to fit cell dry weight. In that case, it's natural that productivity is overestimated by simply extending a fermentation runtime because the model thinks there are more productive cells when the fermentation duration is extended. We are thus interested in more sophisticated methods that can accurately distinguish between viable and dead biomass.

There have been several definitions of the term ”Viable Biomass.” The debate is complicated due to the presence of biologically active but not culturable cells because the conditions do not favor them[48]. We will limit our interest to all metabolically active cells for developing the bioprocess model in this work.

Measuring viable biomass is not an easy task. Madrid et al.[49] provide an excellent overview of the technologies used to estimate biomass. However, many proposed methods fail due to the physiology of a particular microorganism or the methods being too expensive or subjective to be practically utilized in an industrial setting. One of the main difficulties in collecting data on viable biomass in filamentous organisms' fed-batch culture is the complicated morphological We. The photograph in figure 2.4 illustrates the com-

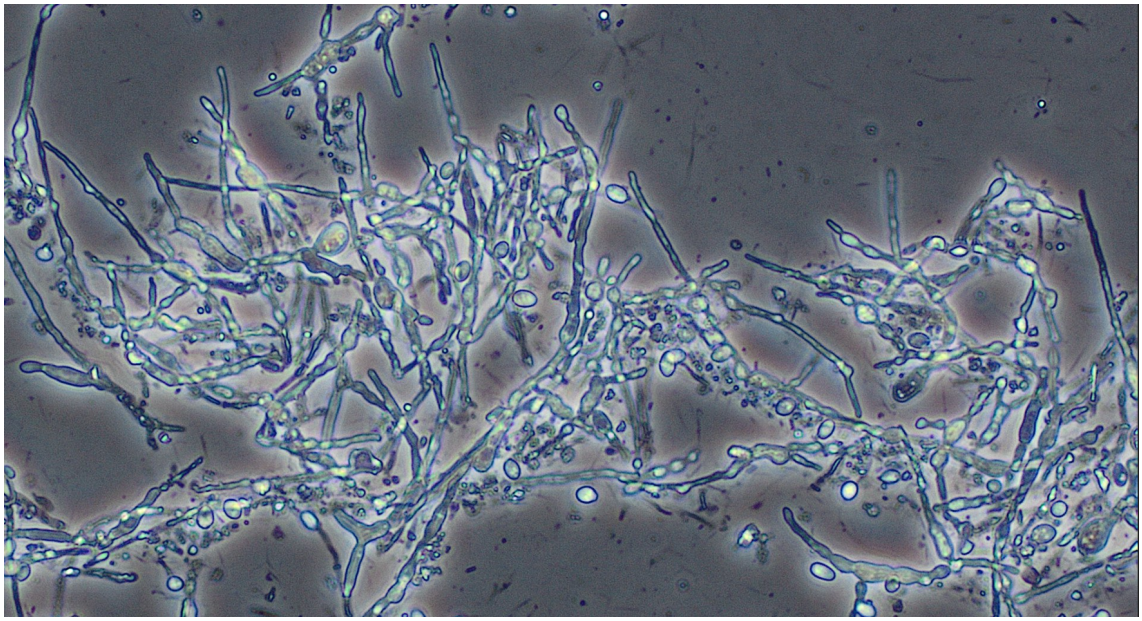

Figure 2.4: Example microscopic image detailing the complex morphological structure of a filamentous fungus during a fed-batch fermentation process

plex morphology. Quantifying viable biomass of filamentous fungi after germination has been only possible with staining methods such as the *Bac*Light[50] followed by rigorous image analysis[33]. Recent developments have been in adapting modern flow cytometry technology for viability estimates in *Penicillium chrysogenum*[51]. However, these very recent developments. Both these methods are incredibly time-consuming for developing appropriate protocols and require specialized equipment and materials.

Dielectric spectroscopy is interesting because of its ease of use and general applicability[52]. Modern probes take advantage of a biological phenomenon called β-disperion to provide an estimate of microbial growth within a fermentation medium. β-disperion is the ability of a biological cell membrane to filter out low-frequency currents and allow the high frequency to pass through[53]. This allows the probe to collect biomass measurements in *real-time* setting by measuring the dielectric properties of the media at high and low current frequencies. β-disperion is only observed when the cell membrane is intact, meaning this measurement technique automatically filters out most dead biomass and other solids in the medium. It's been proven successful in monitoring filamentous organisms[54]. The main problem with dielectric spectroscopy is with the unit itself. It measures dielectric properties usually in *pF/cm* but not biomass concentration, meaning that the collected data can not be directly used in mechanistic models because there is no way to include

dielectric value in a mass conservation balance. Instead, a calibration curve is needed to convert the dielectric measurement to biomass concentration. This, however, has been surprisingly difficult to achieve consistently[55]. It has been established in mammalian cell culture that dielectric spectroscopy measurement correlates with biomass concentrations and not the number of active cells[56]. However, this correlation has never been successfully identified for an entire fed-batch fermentation of filamentous organisms. This is due to a lack of available viability data due to the problematic viability measurements leading researchers to try to correlate dielectric measurements with Cell Dry Weight which starts showing discrepancies as the fermentation reaches later phases.

Bibliography

[1] John Villadsen, Jens Nielsen, and Gunnar Lidén. *Bioreaction engineering principles*. Springer, 2011.

[2] Björn Frahm. Seed train optimization for cell culture. *Methods in Molecular Biology*, 1104:355–367, 2014.

[3] Stuart. M. Stocks. Industrial enzyme production for the food and beverage industries: Process scale up and scale down. *Microbial Production of Food Ingredients, Enzymes and Nutraceuticals*, pages 144–172, 2013.

[4] Gürkan Sin, Krist Gernaey, and Anna Eliasson Lantz. Good modeling practice for pat applications: Propagation of input uncertainty and sensitivity analysis. *Biotechnology Progress*, 25(4):1043–1053, 2009.

[5] Lisa Mears, Stuart M. Stocks, Mads O. Albaek, Gürkan Sin, and Krist V. Gernaey. Mechanistic Fermentation Models for Process Design, Monitoring, and Control. *Trends in Biotechnology*, 35(10):914–924, 2017.

[6] Robert Spann, Christophe Roca, David Kold, Anna Eliasson Lantz, Krist V. Gernaey, and Gürkan Sin. A probabilistic model-based soft sensor to monitor lactic acid bacteria fermentations. *Biochemical Engineering Journal*, 135:49–60, 2018.

[7] P Nomikos and JF MacGregor. Multi-way partial least squares in monitoring batch processes. *Chemometrics and Intelligent Laboratory Systems*, 30(1):97–108, 1995.

[8] T Kourti, P Nomikos, and JF Macgregor. Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway pls. *Journal of Process Control*, 5(4):277–284, 1995.

[9] Qibin Zhao, Cesar F. Caiafa, Danilo P. Mandic, Zenas C. Chao, Yasuo Nagasaka, Naotaka Fujii, Liqing Zhang, and Andrzej Cichocki. Higher order partial least squares (hopls): A generalized multilinear regression method. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1660–1673, 2013.

[10] A.J. Austin, D. Piergentili, J. Glassey, A.C. Ward, and B.V. Kara. Stress monitoring in recombinant fermentations for artificial neural network control. *Ifac Proceedings*

Volumes, 31(8):247–252, 1998.

[11] Dong Sheng Cao, Yi Zeng Liang, Qing Song Xu, Qian Nan Hu, Liang Xiao Zhang, and Guang Hui Fu. Exploring nonlinear relationships in chemical data using kernel-based methods. *Chemometrics and Intelligent Laboratory Systems*, 107(1):106–115, 2011.

[12] Jie Yu. Multiway gaussian mixture model based adaptive kernel partial least squares regression method for soft sensor estimation and reliable quality prediction of nonlinear multiphase batch processes. *Industrial and Engineering Chemistry Research*, 51(40):13227–13237, 2012.

[13] Lijia Luo, Shiyi Bao, Jianfeng Mao, and Di Tang. Quality prediction and quality-relevant monitoring with multilinear pls for batch processes. *Chemometrics and Intelligent Laboratory Systems*, 150:9–22, 2016.

[14] Rasmus Bro. Parafac. tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 38(2):149–171, 1997.

[15] Johan A. Westerhuis, Theodora Kourti, and John F. MacGregor. Comparing alternative approaches for multivariate statistical analysis of batch process data. *Journal of Chemometrics*, 13(3□4):397–413, 1999.

[16] Rasmus Bro. Multiway calibration. multilinear pls. *Journal of Chemometrics*, 10(1):47–61, 1996.

[17] Age K. Smilde and Henk A.L. Kiers. Multiway covariates regression models. *Journal of Chemometrics*, 13(1):31–48, 1999.

[18] Henk A. L. Kiers, Jos M. F. ten Berge, and Rasmus Bro. Parafac2—part i. a direct fitting algorithm for the parafac2 model. *Journal of Chemometrics*, 13(3□4):275–294, 1999.

[19] Barry M. Wise, Neal B. Gallagher, and Elaine B. Martin. Application of parafac2 to fault detection and diagnosis in semiconductor etch. *Journal of Chemometrics*, 15(4):285–298, 2001.

[20] Federico Marini and Rasmus Bro. Scream: A novel method for multi-way regression problems with shifts and shape changes in one mode. *Chemometrics and Intelligent Laboratory Systems*, 129:64–75, 2013.

[21] Jacques Monod. The growth of bacterial cultures. *Selected Papers in Molecular Biology by Jacques Monod*, pages 139–162, 1978.

[22] Victor M. Saucedo and M. Nazmul Karim. Analysis and comparison of input-output models in a recombinant fed- batch fermentation. *Journal of Fermentation and Bioengineering*, 83(1):70–78, 1997.

[23] Mpho Muloiwa, Stephen Nyende-Byakika, and Megersa Dinka. Comparison of unstructured kinetic bacterial growth models. *South African Journal of Chemical Engineering*, 33:141–150, 2020.

[24] Shiying Guo, Bing Li, Wei Yu, David I. Wilson, and Brent R. Young. Which model? comparing fermentation kinetic expressions for cream cheese production. *Canadian Journal of Chemical Engineering*, 99(11):2405–2427, 2021.

[25] RK Bajpai and M Reuss. A mechanistic model for penicillin production. *Journal of Chemical Technology and Biotechnology*, 30(6):332–344, 1980.

[26] S J Pirt and R C Righelato. Effect of growth rate on the synthesis of penicillin by penicillium chrysogenum in batch and chemostat cultures. *Applied Microbiology*, 15(6):1284–1290, 1967.

[27] Gülnur Birol, Cenk Ündey, and Ali Çinar. A modular simulation package for fed-batch fermentation: Penicillin production. *Computers and Chemical Engineering*, 26(11):1553–1565, nov 2002.

[28] J C Menezes, S S Alves, J M Lemos, and S F de Azevedo. Mathematical modelling of industrial pilot-plant penicillin-g fed-batch fermentations. *Journal of Chemical Technology and Biotechnology*, 61(2):123–138, 1994.

[29] Stephen Goldrick, Andrei Ştefan, David Lovett, Gary Montague, and Barry Lennox. The development of an industrial-scale fed-batch fermentation simulation. *Journal of Biotechnology*, 193:70–82, 2015.

[30] Piyush Agarwal, Mohammad Aghaee, Melih Tamer, and Hector Budman. A novel unsupervised approach for batch process monitoring using deep learning. *Computers and Chemical Engineering*, 159:107694, 2022.

[31] Krist Gernaey, Anna Eliasson Lantz, Pär Tufvesson, John Woodley, and Gürkan Sin. Application of mechanistic models to fermentation and biocatalysis for next-generation processes. *Trends in Biotechnology*, 28(7):346–354, 2010.

[32] Julian Kager, Christoph Herwig, and Ines Viktoria Stelzer. State estimation for a penicillin fed-batch process combining particle filtering methods with online and time delayed offline measurements. *Chemical Engineering Science*, 177:234–244, 2018.

[33] G. C. Paul and C. R. Thomas. A structured model for hyphal differentiation and penicillin production using penicillium chrysogenum. *Biotechnology and Bioengineering*, 51(5):558–572, 1996.

[34] Jörg Schubert, Rimvydas Simutis, Michael Dors, Ivo Havlik, and Andreas Lübbert. Bioprocess optimization and control: Application of hybrid modelling. *Journal of Biotechnology*, 35(35):51–68, 1994.

[35] Joel Sansana, Mark N. Joswiak, Ivan Castillo, Zhenyu Wang, Ricardo Rendall, Leo H. Chiang, and Marco S. Reis. Recent trends on hybrid modeling for industry 4.0. *Computers and Chemical Engineering*, 151:107365, 2021.

[36] Belmiro P.M. Duarte and Pedro M. Saraiva. Hybrid models combining mechanistic models with adaptive regression splines and local stepwise regression. *Industrial and Engineering Chemistry Research*, 42(1):99–107, 2003.

[37] Mohammed Saad Faizan Bangi, Katy Kao, and Joseph Sang Il Kwon. Physics-informed neural networks for hybrid modeling of lab-scale batch fermentation for β-carotene production using saccharomyces cerevisiae. *Chemical Engineering Research and Design*, 179:415–423, 2022.

[38] J. Peres, R. Oliveira, and S. Feyo De Azevedo. Knowledge based modular networks for process modelling and control. *Computers and Chemical Engineering*, 25(4-6):783–791, 2001.

[39] Mohammed Saad Faizan Bangi and Joseph Sang Il Kwon. Deep hybrid modeling of chemical process: Application to hydraulic fracturing. *Computers and Chemical Engineering*, 134:106696, 2020.

[40] Rasmus Fjordbak Nielsen, Nima Nazemzadeh, Laura Wind Sillesen, Martin Peter Andersson, Krist V. Gernaey, and Seyed Soheil Mansouri. Hybrid machine learning assisted modelling framework for particle processes. *Computers and Chemical Engineering*, 140:106916, 2020.

[41] Moritz von Stosch, Jan Martijn Hamelink, and Rui Oliveira. Hybrid modeling as a qbd/pat tool in process development: an industrial e. coli case study. *Bioprocess and Biosystems Engineering*, 39(5):773–784, 2016.

[42] Dimitris C. Psichogios and Lyle H. Ungar. A hybrid neural network-first principles approach to process modeling. *Aiche Journal*, 38(10):1499–1511, 1992.

[43] Xianfang Wang, Jindong Chen, Chunbo Liu, and Feng Pan. Hybrid modeling of penicillin fermentation process based on least square support vector machine. *Chemical Engineering Research and Design*, 88(4):415–420, 2010.

[44] Christoph Herwig. *Hybrid Modelling and Multi- Parametric Control of Bioprocesses*. MDPI - Multidisciplinary Digital Publishing Institute, 2018.

[45] Leander A. H. Petersen. Single cell protein production in u-loop bioreactors: Fundamentals, modeling amp; control, 2019.

[46] DB Kell, GH Markx, CL Davey, and RW Todd. Real-time monitoring of cellular biomass - methods and applications. *Trac-trends in Analytical Chemistry*, 9(6):190–194, 1990.

[47] Pauline M. Doran. *Bioprocess engineering principles*. Academic Press, 1995.

[48] Douglas B. Kell, Arseny S. Kaprelyants, Dieter H. Weichart, Colin R. Harwood, and Michael R. Barer. Viability and activity in readily culturable bacteria: A review and discussion of the practical issues. *Antonie Van Leeuwenhoek, International Journal of General and Molecular Microbiology*, 73(2):169–187, 1998.

[49] R. E. Madrid and C. J. Felice. Microbial biomass estimation. *Critical Reviews in Biotechnology*, 25(3):97–112, 2005.

[50] S. M. Stocks. Mechanism and use of the commercially available viability stain, baclight. *Cytometry Part a*, 61(2):189–195, 2004.

[51] Lukas Veiter and Christoph Herwig. The filamentous fungus penicillium chrysogenum analysed via flow cytometry—a fast and statistically sound insight into morphology and viability. *Applied Microbiology and Biotechnology*, 103(16):6725–6735, 2019.

[52] John E. Yardley, Douglas B. Kell, John Barrett, and Christopher L. Davey. On-line, real-time measurements of cellular biomass using dielectric spectroscopy. *Biotechnology and Genetic Engineering Reviews*, 17(1):3–36, 2000.

[53] Herman P. Schwan. Electrical properties of tissue and cell suspensions. volume 5 of *Advances in Biological and Medical Physics*, pages 147–209. Elsevier, 1957.

[54] S. Krairak, K. Yamamura, M. Nakajima, H. Shimizu, and S. Shioya. On-line monitoring of fungal cell concentration by dielectric spectroscopy. *Journal of Biotechnology*, 69(2-3):115–123, 1999.

[55] Thomas Maskow, Anita Röllich, Ingo Fetzer, Jun Yao, and Hauke Harms. Observation of non-linear biomass-capacitance correlations: Reasons and implications for bioprocess control. *Biosensors and Bioelectronics*, 24(1):123–128, 2008.

[56] Juhi Fernandes, Jayme Currie, Kevin Ramer, and An Zhang. Development of capacitance tools: At-line method for assessing biomass of mammalian cell culture and fixed cell calibration standard. *Biotechnology Journal*, 14(4):1800283, 2019.

3 Determining viable biomass concentration via cell capacitance: Linear calibration methodology

*Disclosure: The following chapter is a journal article submitted to Biotechnology letters. The format has been adjusted to fit the thesis.

Abstract

Objectives: Dielectric spectroscopy is commonly used for online monitoring of biomass growth. It is however not often utilized for quantitative viable biomass determination due to poor correlation with dry weight. A calibration methodology is developed that can directly measure viable biomass concentration using dielectric values.

Results: The methodology is applied on an industrial scale fermentation of Acremonium fusidioides. By mixing fresh and heat-killed samples, a linear model including sample viability could be fitted with the dielectric β-dispersion (ΔC) values and total solids concentration. With a total of 26 samples across 21 different cultivations, three different measurement methods were tested: A modern annular probe in an offline setting with two different sample volumes of 2 and 100 mL, and a legacy viable cell analyzer. The linear model provided an R2 value of 0.99 between ΔC and viable biomass across the sample set using either instrument. The difference in ΔC when analyzing 100 mL and 2 mL samples with an annular probe can be adjusted by a scalar factor of 1.33 within the microbial system used in this study, preserving the linear relation with R2 of 0.97.

Conclusions: It is possible to estimate viable biomass utilizing dielectric spectroscopy without excessive viability studies. The same novel method can be applied to calibrate different instruments to measure viable biomass concentration. Small sample volumes are appropriate as long as the sample volumes are kept consistent.

3.1 Introduction

The concentration of active biomass is a key process parameter for any bioprocess. Despite its importance, it is extremely difficult and time-consuming to measure and for some organisms, it is practically impossible. Even for simple organisms, measuring active or viable biomass is subjective and may not be practical[1]. The most widely accepted standard technique for quick estimations of biomass is the use of Optical Density calibrated to Cell Dry Weights (CDW)[2]. However, these techniques measure total biomass, including dead cells which may lead to inaccuracies when determining strain parameters like observed yield. Furthermore, CDW can not differentiate between biomass and non-biological solids present in the media.

Dielectric spectroscopy is the only known method that allows for online monitoring of active biomass[3]. In the radio-frequency range, the media permittivity is dominated by the capacitive behavior of cell membranes of intact cells. The frequency-dependence of the media permittivity in this region is also known as the β-dispersion. Biocapacitance is widely used in the fermentation and cell culture industries for monitoring purposes. The trends can be utilized to detect abnormalities in cellular growth[4] or even directly adapted for developing control strategies[5]. However, there is value in being able to monitor biomass concentrations instead of dielectric properties. Biomass concentration is widely utilized for more accurate process modeling, control, and establishing growth kinetics, yield, and stoichiometry which are critical in process optimization and intensification[6]. Unfortunately, there is no universal relation between dielectric spectroscopy and active biomass concentration that is transferable between different biological systems. Due to difficulties in independently measuring active biomass, most online measurements of biomass rely on establishing a correlation between permittivity increments and CDW. This works well for systems with low biomass concentration due to the measurement technique being extremely robust to changes in media composition and conditions[7]. However, in practice this endeavor frequently fails at later stages of fermentation[8], the likely cause is due to the accumulation of cell debris, non-viable biomass, and other non-soluble solids during the fermentation period[9].

The term viable biomass is not clearly defined, and different viability measurement techniques give different results[10]. Dielectric spectroscopy only measures changes in dielectric properties which are usually directly influenced by the amount of cytoplasm surrounded by an intact plasma membrane and thus indirectly estimates membrane integrity[4]. It has been established for mammalian cells that dielectric spectroscopy can measure viable cell densities regardless of different cell growth phases[11]. Dielectric spectroscopy isn‘t limited to these types of organisms so a similar correlation should exist for most bioprocesses. The viability of a filamentous organism is extremely difficult to measure, the only reliable protocols are expensive and time-consuming to develop[12]. A quick and easy estimation of active biomass concentration for these types of organisms would be invaluable for those interested in going beyond analyzing data trends.

In this work, we will report the use of dielectric spectroscopy applied in the study of industrial filamentous fungal fermentation. We will show the relationship between dielectric spectroscopy and active biomass via viability control by mixing fresh and heat-killed samples, and then establish how dielectric spectroscopy can be used for easy and direct measurement of viable cell dry weights.

3.2 Materials and Methods

3.2.1 Microorganism and media conditions

Samples were obtained from the main bioreactors used for the commercial manufacturing of Fusidic Acid at the Ballerup site of LEO Pharma A/S. Fresh samples were taken from industrial fermentations from various times of cultivation. The conditions and media are similar to the process description of Fusidic Acid fermentation reported by Daehne et al.[13], but updated details regarding present-day operating conditions, component concentrations, and organisms are considered sensitive information and are not disclosed

3.2.2 Cell Dry Weight

Cell Dry Weight (CDW) measurements are performed by passing a known mass of fermentation broth through a 70 mm glass fiber filter paper with an applied vacuum. The filtrate is washed two times by filling the funnel with deionized water. The washed filtrate

was then placed in a drying oven at a temperature of 100 °C for at least 48 hours. Final CDW values are expressed in unit dry weight per unit fermentation broth weight.

3.2.3 Dielectric Spectroscopy

The permittivity of each sample was measured using FUTURA Pico 12 mm annular probe using the microbial setting and dual-frequency mode measuring at 580 kHz and 15650 kHz. Modern instruments usually report permittivity increment (denoted as ΔC from now) as the measurement difference between these two points. Due to sample volume constraints in smaller scales of operation and our need to often utilize offline measurements, the systematic deviation between small and large sample volumes needed to be established. The permittivity of fresh samples was measured with 2 mL cell suspension in a 1 cm diameter tube and a 100 mL suspension in a 10 cm diameter sterilized plastic bottle. The manufacturer recommends the use of a minimum of 8 cm diameter sample container, preventing artificial shifts in capacitance[14]. A 2 mL sample is an approximate practical limit.

For legacy reasons, certain samples were also measured in an older at-line Viable Cell Analyzer (ABER Instruments 822). This device measures permittivity at a single frequency of 1 kHz. The instrument has a built-in sample chamber with a stirrer and temperature control and requires approximately 2 mL for accurate measurements. To measure ΔC a small portion of the fresh sample is filtered, and the permittivity of the cell-free filtrate is recorded and utilized as the media background. A total of 26 samples were generated for this study from 21 independent cultivations. 15 samples were analyzed using the annular probe with 2 mL of broth in a tube and a larger flask containing approximately 100 mL of broth. The rest of the samples were analyzed using the viable cell analyzer.

3.2.4 Viability control: Mixed fresh heat-killed sample

The procedure follows previously published methods used by Véronique et al.[15] as a viability assay. A portion of each sample was moved to a separate container and heat-killed by placing the container in a water bath for 30 minutes at 70 °C. Heat-killed and fresh samples were mixed at approximately 0%, 25%, 50%, 75%, and 100% w/w, with the actual portions being accurately weighed. The permittivity of the samples was then measured.

3.2.5 Experimental Analysis and calibration

Calibration and methodology were applied and developed in MATLAB R2021a. The calibration intends to find a mathematical relationship that directly converts dielectric spectroscopy measurements to viable biomass concentration without the need to measure the cell viability independently. The theory supports a direct linear relationship between ΔC and viable biomass (X_{Viable}). However, we will also introduce an offset (β_2) to account for any residual permittivity.

$$\Delta C = \beta_1 X_{Viable} + \beta_2 \tag{3.1}$$

β_2 is easy to estimate from killed samples. If no residual permittivity persists then $\beta_2 = 0$. Estimating β_1 is a much more challenging task as X_{Viable} is not known. However, X_{Viable} can be written as a relationship to CDW (X_{TDW}) as

$$X_{Viable} = \alpha M_{frac} X_{TDW} \tag{3.2}$$

Where α denotes the viability fraction of a fresh sample which can take a value from 0 to 1 and is unique for each fresh sample. Here we have also introduced the mass fraction of a fresh sample M_{frac}, which allows us to include all the data from the mixed fresh and killed measurements, a completely killed sample has $M_{frac} = 0$. Note that the α value is shared across different sample mixings as long as the same fresh sample is used when mixing. Once a relationship between viable biomass concentration and dielectric spectroscopy is confirmed an optimization routine is used to fit sample viability fractions so that the estimated viable biomass concentration follows the established relationship. If α is known for each sample, then X_{Viable} can be calculated for all samples. Subsequently, β_1 can be estimated via linear regression. The optimal α will result in the best fit to a linear relationship and thus can be estimated via optimization by utilizing the R^2 value as the optimization objective, thus, the optimization is formulated as follows.

$$\begin{aligned} \boldsymbol{\alpha}_{opt} &= \underset{\boldsymbol{\alpha}}{\text{argmax}}\, R^2 \\ 0 &< \boldsymbol{\alpha} < 1 \end{aligned} \tag{3.3}$$

This optimization problem was solved with MATLAB's *fmincon* function which utilizes an Interior Point algorithm. The calibration methodology was later easily ported over to Microsoft Excel using the native GRG Nonlinear which gave the same results for more general use. Note that the data presented in this work have been arbitrarily scaled to preserve industry-sensitive information and thus exact units will not be included on the axes. Instead, axis ticks are preserved to illustrate the overall trend.

3.3 Results and Discussions

3.3.1 Linear relation

Initial measurements of a sample from a bioreactor on the viable cell analyzer show no significant change in the permittivity signal after leaving a sample in an open container for 90 minutes at ambient conditions, indicating that the microbial system is stable throughout the analysis. Subjecting broth to heat treatment by placing the sample container in a water bath at 70 °C led to a rapid loss of permittivity signal. The signal stabilized at low permittivity measurements after 15 minutes, see Supplementary Figure A.1. Some residual broth permittivity remains even after further heat treatment. Still, the microbes are considered dead as permittivity does not decrease further with longer heat treatment. For the rest of this work, each sample portion that is killed will be placed in the 70 °C water bath for a minimum of 30 minutes to ensure that the sample is fully dead.

The classical issue of traditional calibration methods based only on measured Cell Dry Weight (CDW) is seen in Figure 3.1. At low biomass concentrations in the early phase, there is a strong linear relationship between CDW and Permittivity Increment (ΔC). At higher biomass concentrations, corresponding to middle and late phases, this correlation weakens significantly. This can mostly be explained by the total CDW measurement being a poor estimate of viable CDW rather than dielectric spectroscopy failing at high biomass concentrations.

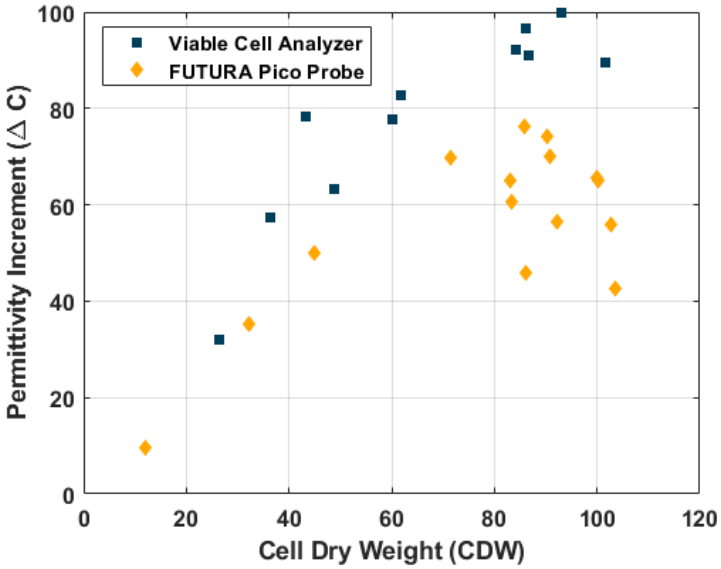

Figure 3.1: Comparison of measured CDW and ΔC. Dielectric spectroscopy of certain samples in ABER Viable Cell Analyzer and others. No single sample is measured on both instruments

One sample was measured at 10% intervals of the fresh mass fraction to verify linearity; the results are shown in Fig 2a). Furthermore, Fig. 2b) shows three of the measured samples subject to viability control. The samples are selected so that each phase of the fermentation is represented. These are expected results when ΔC is linearly correlated with viable biomass. In the early phase, there is usually a direct relationship between CDW and ΔC. Viability control suggests a direct relationship between viable biomass and ΔC during that phase, indicating that CDW is a good estimator for viable biomass in the growth phase. At higher biomass concentrations, which typically occur in the middle and late phases, the direct relationship between CDW and ΔC begins to fail. However, for this microbial system, viability control shows evidence of a linear relationship between viable biomass concentration and ΔC, even at higher biomass concentrations. The results confirm a linear relationship between permittivity and viable biomass concentration for the entire fermentation duration.

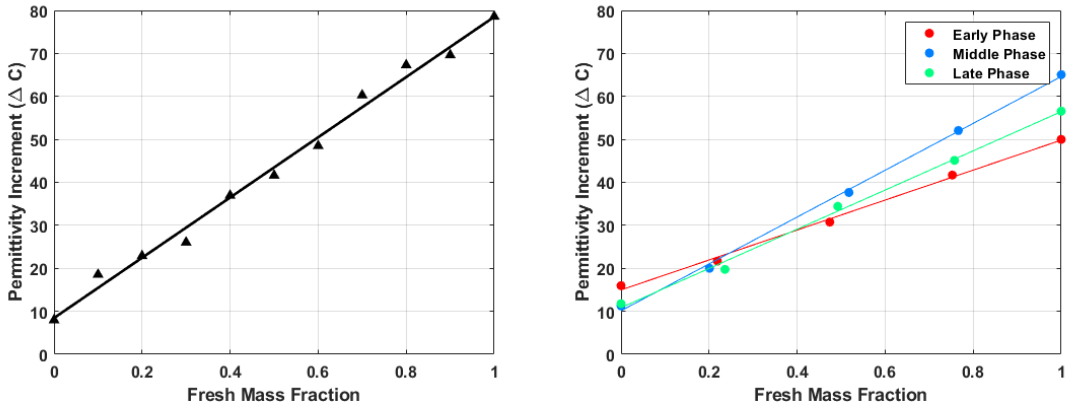

Figure 3.2: Differences in permittivity increment (ΔC) at different dilution levels. Fig. 2a Shows a single sample in the stationary phase measured on the Viable Cell Analyzer at an increment of approx. 0.1 fresh sample mass fraction. Fig. 2b shows the permittivity increment of three selected samples to represent different fermentation phases at different dilution levels, measured on the ABER FUTURA Probe.

3.3.2 Dielectric spectroscopy calculation

Following the calibration methodology, it was possible to select a value for α for each of the analyzed samples that gave an extremely strong linear correlation (Figure 3.3). It is important to consider that the vector α containing viability fractions can be scaled with a single scalar without a drop in R^2 value, thus we have decided to normalize the estimated values so that the highest α value after the optimization routine takes a value of 1. This is equivalent to assuming that at least one of the measured samples consists of only viable biomass as the solid phase. This may not always be a realistic assumption. However, as a correction to CDW, only non-active biomass is subtracted from the measurement and is thus a better approximation to active biomass than the original CDW measurement while at the same time being a much faster measurement technique, requires lower samples volumes and with modern probes, can be used for online data collection. Furthermore, if independent viable cell analysis could be done, it only needs to be done on one sample to calibrate the probe to measure active biomass accurately, as a single viability measurement can be utilized to scale all α values for other samples properly correctly.

Figure 3.3 shows the estimated viable biomass (X_{Viable}) after the estimation of the viability fraction of all samples via optimization and subsequent corrections for both measurement devices. Linear calibration is displayed in Table 1. The probe generally measures lower

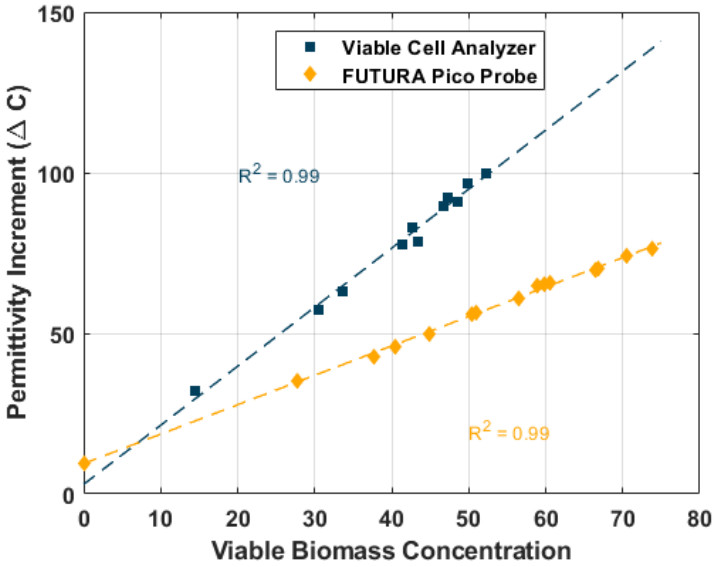

Figure 3.3: Determination of hidden relation between viable biomass and ΔC after applying the viable fraction corrections on CDW measurements. Two independent calibrations are obtained depending on the measurement device used for ΔC

ΔC than the viable cell analyzer for similar viable biomass concentrations and strain. The difference can be explained because the probe and instrument utilized different measurement techniques at different frequencies and also a measured capacitance shift due to small sample containers. The important thing to note is that the signal is still able to establish the linear trend between ΔC and viable biomass; each method can be calibrated easily. These calibrations can be utilized directly to measure viable biomass concentration using dielectric spectroscopy in the current fermentation.

Table 3.1: Parameters for the linear correlation between ΔC and viable biomass across the three different measurement types.

Measurement type	Slope (β_1	Intercept (β_2)
Viable Cell Analyzer	1.67 ± 0.05	9.49 ± 1.39
Annular probe (2 mL)	0.89 ± 0.02	11.91 ± 0.76
Annular probe (100 mL)	1.18 ± 0.12	15.84 ± 0.83

3.3.3 Sample volume required with Annular probe

All fresh samples measured with the annular probe were also measured in a larger sample container to examine the influence of the wall effect on such small sample volumes. As expected, there was a difference in the ΔC reading depending on the container size due to close proximities between the electrode and container wall in smaller tubes. However,

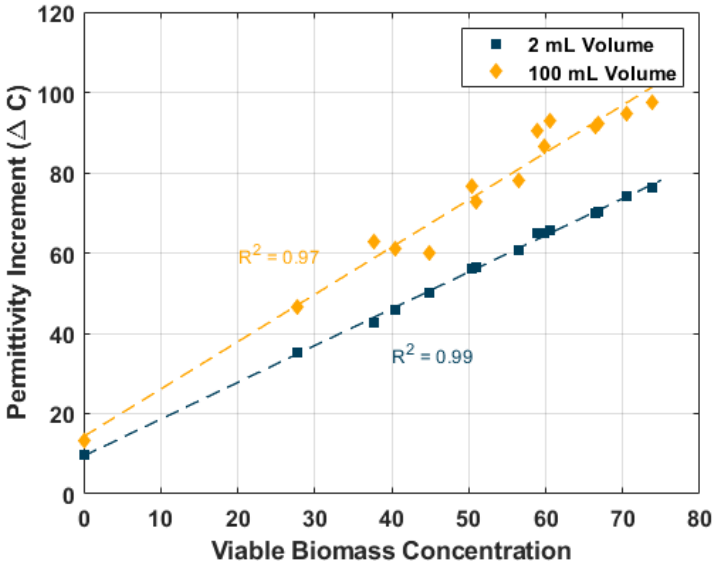

Figure 3.4: The difference in ΔC when measuring the same fresh samples with different sample volumes when using the annular probe. Viable biomass is calculated using the linear correlation calibrated with the 2 mL samples measured with the annular probe.

the ΔC difference between large and small samples is very consistent throughout. Thus, the wall effect is largely the same as long as the container volume is kept consistent throughout a sampling campaign. A small sample of 2 mL can be used for direct offline measurements, but the scale difference must be noted if one intends to utilize the same calibration between different sample volumes or when transitioning to online monitoring.

Figure 3.4 shows the comparison between when the same fresh samples are analyzed using the probe in a small tube container of 2 mL and a larger container of 100 mL, which is large enough to achieve a better resolution of the permittivity signal without artificial shifts. The viable biomass depicted in Fig. 4 is calculated using the linear constants from Table 1 with the 2 mL samples and directly compared to the ΔC measured on 100 mL samples. A new linear relation was constructed, which is also illustrated in Figure 3.4 utilizing the 100 mL samples, and the constants are shown in Table 3.1. It was established by Fernandes et al.[14] that moving from *offline* to *online* could be done with a simple scalar conversion factor. This scaling factor is likely dependent on the microbe and measurement techniques and thus has to be determined individually. While other studies concluded that *offline* analysis could be done with a modern probe as long as the electrode was not close to a surface to measure the full signal, the results here establish that

it is possible to go even further down in the sample volume while retaining the information of interest. The conversion factor for this microbial system when estimating ΔC of a 100 mL sample utilizing 2 mL measurement was estimated to be 1.33 ± 0.07. It is hypothesized that a similar scaling factor method could be utilized for 2 mL offline samples to calibrate the probe for online monitoring.

3.4 Conclusions

This work explored the use of dielectric spectroscopy measurements in an industrial fermentation system utilizing a filamentous fungus, using off-line measurements of various sample volumes, with modern and legacy equipment. Viability control, where a fresh sample containing an unknown fraction of living biomass is mixed with a killed portion of the same sample, indicates a strong linear relationship between viable biomass and Permittivity Increment (ΔC), in all cases, at every stage of the fermentation. Classically, the correlation between ΔC and Cell Dry Weight (CDW) did not exist past a certain fermentation phase. For the first time, a simple calibration methodology was developed using ΔC measurement from mixed fresh and heat-killed samples. This method provides better estimations of viable biomass concentration without going through the long and difficult procedure of calibration with independent viability measurements. The constants for the linear relationship were specific to the instrument, sample volume, and container diameter. The constants may be more generally applicable to other fermentation processes or organism types. Exploring this would require additional work, which could be done in other laboratories. We will apply the present calibration to modeling work presently underway in our laboratories and production facilities.

Bibliography

[1] Ruifei Wang, Bettina Lorantfy, Salvatore Fusco, Lisbeth Olsson, and Carl Johan Franzén. Analysis of methods for quantifying yeast cell concentration in complex lignocellulosic fermentation processes. *Scientific Reports*, 11(1):11293, 2021.

[2] R. E. Madrid and C. J. Felice. Microbial biomass estimation. *Critical Reviews in Biotechnology*, 25(3):97–112, 2005.

[3] John E. Yardley, Douglas B. Kell, John Barrett, and Christopher L. Davey. On-line, real-time measurements of cellular biomass using dielectric spectroscopy. *Biotechnology and Genetic Engineering Reviews*, 17(1):3–36, 2000.

[4] Pareshkumar Patel and Gerard H. Markx. Dielectric measurement of cell death. *Enzyme and Microbial Technology*, 43(7):463–470, 2008.

[5] Daniela Ehgartner, Thomas Hartmann, Sarah Heinzl, Manuela Frank, Lukas Veiter, Julian Kager, Christoph Herwig, and Jens Fricke. Controlling the specific growth rate via biomass trend regulation in filamentous fungi bioprocesses. *Chemical Engineering Science*, 172:32–41, 2017.

[6] Rita Lencastre Fernandes, Vijaya Krishna Bodla, Magnus Carlquist, Anna Lena Heins, Anna Eliasson Lantz, Gürkan Sin, and Krist V. Gernaey. Applying mechanistic models in bioprocess development. *Advances in Biochemical Engineering/Biotechnology*, 132(December 2012):137–166, 2013.

[7] S. Krairak, K. Yamamura, M. Nakajima, H. Shimizu, and S. Shioya. On-line monitoring of fungal cell concentration by dielectric spectroscopy. *Journal of Biotechnology*, 69(2-3):115–123, 1999.

[8] Nanna Petersen Rønnest, Stuart M. Stocks, Anna Eliasson Lantz, and Krist Gernaey. Introducing process analytical technology (pat) in filamentous cultivation process development: comparison of advanced online sensors for biomass measurement. *Journal of Industrial Microbiology and Biotechnology*, 38(10):1679–1690, 2011.

[9] Thomas Maskow, Anita Röllich, Ingo Fetzer, Jun Yao, and Hauke Harms. Observation of non-linear biomass-capacitance correlations: Reasons and implications for

bioprocess control. *Biosensors and Bioelectronics*, 24(1):123–128, 2008.

[10] Douglas B. Kell, Arseny S. Kaprelyants, Dieter H. Weichart, Colin R. Harwood, and Michael R. Barer. Viability and activity in readily culturable bacteria: A review and discussion of the practical issues. *Antonie Van Leeuwenhoek, International Journal of General and Molecular Microbiology*, 73(2):169–187, 1998.

[11] Hae Woo Lee, John Carvell, Kurt Brorson, and Seongkyu Yoon. Dielectric spectroscopy-based estimation of vcd in cho cell culture. *Journal of Chemical Technology and Biotechnology*, 90(2):273–282, 2015.

[12] S. M. Stocks. Mechanism and use of the commercially available viability stain, baclight. *Cytometry Part a*, 61(2):189–195, 2004.

[13] WV Dahene, S Jahnsen, I Kirk, R Larsen, and H Lorck. *Fusidic acid: Properties, biosynthesis, and fermentation*. Biotechnology of industrial antibiotics, 1984.

[14] Juhi Fernandes, Jayme Currie, Kevin Ramer, and An Zhang. Development of capacitance tools: At-line method for assessing biomass of mammalian cell culture and fixed cell calibration standard. *Biotechnology Journal*, 14(4):1800283, 2019.

[15] Véronique Lecault, Nilesh Patel, and Jules Thibault. Morphological characterization and viability assessment of trichoderma reesei by image analysis. *Biotechnology Progress*, 23(3):734–740, 2007.

4 Multi-modal modeling of industrial-scale fermentation

Abstract

The Shifted Covariates Regression Analysis for Multi-way data (SCREAM) modeling tools is applied for quality prediction in batch processes. This model has two prominent advantages. The first one is that it relies on tensor decomposition and thus avoids the potential ”curse of dimensionality” and information loss when the data structure is unfolded. The second advantage is that it can model uneven-length problems without requiring batch trajectory synchronization. The method is tested on simulated and real industrial-scale fed-batch datasets. The model’s performance is compared to traditional multi-way regression models, Unfold Partial Least Squares (Unfold-PLS) and multilinear PLS (NPLS). SCREAM showed comparable performance to established methods when predicting the harvest of the main product with an average prediction error of 12.73%. However, when predicting byproduct concentration in a dataset from an industrial sponsor, SCREAM performs better than other available regression approaches with an average prediction error of 42.74% compared to the error of 93.24% with Unfold-PLS.

4.1 Introduction

Batch and fed-batch processes are widely used in manufacturing specialty chemicals, including food, biochemicals, and pharmaceuticals. Models of batch processes are greatly valued as they can efficiently support batch scheduling and optimization of process performance. While mechanistic models are considered ideal for this purpose[1], model development can be limited due to a lack of understanding of the complex system dynamics and available measurements of key process variables.

Bioprocesses are an excellent example of batch processes with limited system dynamic knowledge while also commonly accompanied by a large amount of data. Data-driven modeling can help interpret these large datasets and describe the process without prior knowledge[2]. This is especially relevant in pharmaceutical settings, which depending on the novelty of the working cell culture, may contain no systematic knowledge. Despite needing no prior knowledge, data-driven models can help identify process trends and thus facilitate mechanistic model building in the future.

On the industrial side, a predictive model that can estimate the quality of the batch at harvest can yield significant benefits. The recovery process‘s efficiency may depend on difficult-to-measure batch quality variables. Depending on the sampling and measurement methods, this information may only be available after commencing recovery to meet production demands due to time constraints. Batches may not be up to specification; thus, additional resources are wasted. Furthermore, suppose the recovery process relies on solvents, the solvent used may be optimized, and the process can be made more efficient if the yield of the main product can be estimated ahead of time.

The process systems engineering community is seeing an increased interest in using machine learning and artificial intelligence algorithms to solve various problems in the biochemical industry[3]. However, it isn‘t easy to use these data-driven techniques directly due to the unique nature of the batch and fed-batch datasets. In standard regression, data are arranged in a two-way structure; a table or a matrix, but batch data is a three-way data consisting of batch $\times$ time $\times$ process variable. Furthermore, unfolding the data to conform to a two-dimensional structure is not trivial due to the varying run times which leads to Fed-batch process data being an uneven three-way data array.

Multivariate tools have been utilized in batch modeling, primarily for monitoring purposes[4].

Most of these tools are based on more traditional bilinear models such as Principal Component Analysis (PCA) or Partial Least Squares (PLS). There has also been success in using specialized three-way models to handle the batch process data structure. A significant motivation for using three-way models is preserving the three-way structure within the model for easier interpretation and to avoid the curse of dimensionality[5]. Traditionally as part of a preprocessing step, time alignment is required to make the dataset „even, “ which is a burdensome task to do correctly[6]. There are novel methods that allow for more flexible modeling of multi-way data. The most widely used is the PARAFAC2 model[7], but recently a specialized version of the Tucker model called GTucker2 has shown promise in monitoring a penicillin process[8].

There can be no conclusions drawn as to which three-way model is the best for batch process data because they might serve different purposes[9][10]. When applied to batch data, the most common applications of multivariate models can fit into the following categories; Process optimization, Process monitoring, or Product quality prediction. This study focuses on quality prediction, where a regression model is built to predict a quality variable **Y** from recorded batch measurements **X**

For Regression purposes, Marini and Bro [11] developed a regression method for GC-MS data to quality variables; this is a difficult task due to baseline drifts which leads to shifts in the dataset. The Shifted Covariates Regression Analysis for Multi-way data (SCREAM) method is a modification to Multivariate Covariates Regression (McovR)[12] to utilize the PARAFAC2 engine when decomposing the data structure over the traditional constrained Tucker models. This allows direct regression of data without correcting the baseline drifts. An additional feature is that since PARAFAC2 is used for decomposing, SCREAM can handle three-way arrays where each slice is of varying lengths. This makes it a promising tool for the regression of batch data as it takes both the three-way structure and the varying runtime of independent batches without requiring time adjustment as a preprocessing step. This study will explore the application of SCREAM and multilinear PLS (NPLS)[13] as regression tools in the fermentation industry while also comparing it to the more traditional method of unfold-PLS.

4.2 Materials and Methods

4.2.1 Model Backgrounds

Unfold-PLS

Unfold Partial Least Squares or Unfold-PLS is a standard tool in literature for multivariate analysis of batch data[14] and is commonly referred to as Multiblock PLS. Despite having a unique name, this is a standard PLS model with an unfolding preprocessing step. Unfolding is the act of turning a 3-D data array into a 2-D matrix by taking slabs from an $(IxJxK)$ data array to create multiple 2-D matrices and then aligning them to give a single 2-D matrix. Since there are different ways to unfold 3-D displays, and it's usually based on the type of data, the unfolding used throughout this work will be discussed during model development.

Given an unfolded array **X** of size $(KxIJ)$, the PLS model seeks to decompose **X** into a series of scores and loadings that maximizes variance explained while having a high correlation with response **Y**.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\top} + \mathbf{E} \tag{4.1}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\top} + \mathbf{F} \tag{4.2}$$

Where **T** consists of R extracted orthonormal score vectors from **X**, and **U** are the score vectors from **Y** having maximum covariance with **T**. Matrices **P** and **Q** represent loadings while **E** and **F** are respectively the residuals for **X** and **UY**. PLS is a subsequent method meaning the components are estimated one at a time by solving the following optimization

$$\min_{\mathbf{w}} \quad [cov(\mathbf{t}, \mathbf{y}|\mathbf{t} = \mathbf{X}\mathbf{w}; ||\mathbf{w}|| = 1] \tag{4.3}$$

The solution to this problem is the following result

$$\mathbf{w} = \frac{\mathbf{X}^{\top}\mathbf{y}}{||\mathbf{X}^{\top}\mathbf{y}||} \tag{4.4}$$

The score vector of this first component is found using $\mathbf{t} = \mathbf{Xw}$. Subsequent components are found by deflating the **X** and **y** blocks. The quality variable **y** is predicted, assuming a linear relationship between the score matrices.

$$\mathbf{U} = \beta\mathbf{T} \tag{4.5}$$

Where β is a regression vector. Further review of Multiblock algorithms can be found in the original paper by Kourti et al.[15].

NPLS

NPLS is the multiway generalization of traditional PLS. Whereas two-way PCA and PLS try to summarize the data matrix by decomposing it into a sum of dyads, the three-way NPLS utilizes triads to summarize the data. Unlike unfold-PLS, the NPLS model is truly multilinear. We will limit the discussion to Tri-PLS1, a three-way PLS regression onto a single univariate variable. For a three-way data array, $\underline{\mathbf{X}}$ and univariate dependent data **y**, A single component of $\underline{\mathbf{X}}$ is a decomposition into a score vector and two weight vectors. Essentially each mode or dimension gets a unique vector. A single-component Tri-PLS model of X is given as

$$\mathbf{X} = \mathbf{t}(\mathbf{w}^J \otimes \mathbf{w}^I) + \mathbf{E} \tag{4.6}$$

$$\mathbf{y} = \mathbf{tb} + \mathbf{f} \tag{4.7}$$

Where $\otimes$ denotes the Kronecker product of the two vectors.
NPLS models are determined by finding the weight vectors $\mathbf{w}^J$ and $\mathbf{w}^I$ such that the covariance between t and y is maximized. N-PLS models are the solution to the following optimization problem

$$\min_{\mathbf{w}^J, \mathbf{w}^I} \quad [cov(\mathbf{t}, \mathbf{y} | \mathbf{t} = \mathbf{X}(\mathbf{w}^J \otimes \mathbf{w}^I); ||\mathbf{w}^J|| = ||\mathbf{w}^I|| = 1] \tag{4.8}$$

Finding the solution to problem 4.8 requires defining a matrix, $\mathbf{Z}(JxK)$ with elements

$$z_{ij} = \sum_{k=1}^{K} y_k x_{ijk} \tag{4.9}$$

The solution to the NPLS optimization problem is found as the left and right singular vectors of **Z**. The score vector **t** is then found via regression of **X** onto **w**.

$$\mathbf{w} = (\mathbf{w}^J \otimes \mathbf{w}^I) \tag{4.10}$$

$$\mathbf{t} = \frac{\mathbf{Xw}}{(\mathbf{w}^\top \mathbf{w})} \tag{4.11}$$

The NPLS algorithm is a sequential method, meaning that the vectors for each subsequent component can be determined by deflating the arrays **X** and **y** and then repeating the steps required to find a single-component model.

NPLS works directly with multi-way data like ones from a batch process and does not require any unfolding. The three-way structure is kept and can be utilized in model interpretation. NPLS uses far fewer parameters and is more strict than its Unfold-PLS counterpart, which could lead to better predictive power. A drawback of the Unfold PLS and NPLS models is that all slabs or batches must contain the same number of data points, i.e., the data array must be even. We refer to the paper by Bro[13] for further details regarding NPLS. This work builds N-PLS models utilizing the scripts available with the N-way toolbox, freely available at http://www.models.life.ku.dk/nwaytoolbox.

SCREAM

The SCREAM model utilizes a PARAFAC2 fitting algorithm based on an Alternating Least Squares (ALS) approach. PARAFAC2 models are expressed as

$$\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}_k^\top + \mathbf{E}_k \qquad k = 1, ..., K \tag{4.12}$$

Here $\mathbf{X}_k$ is a single slab of the entire three-way structure $\underline{\mathbf{X}}$, or the data from a single batch. For a PARAFAC2 model with F components, the matrix **A** is a matrix (IxF) of loadings in the I direction. For batch data, this is usually the variable loadings. $\mathbf{D}_k$ is a diagonal matrix (IxF) containing the $k'th$ row of the matrix $\mathbf{C}(KxF)$ which has the loadings in the K or batch direction. **C** is similar to a score matrix in ordinary 2-way PCA. Finally, $\mathbf{B}_k$ is the loadings in the J direction or the time point direction. Generally, $\mathbf{B}_k$ holds the loadings where the shifts happen. Finally, $\mathbf{E}_k$ contains the residuals. PARAFAC2 models are made

unique by the constraint that the cross-product of each $\mathbf{B}_k$ is the same, i.e., $\mathbf{B}_k^\top \mathbf{B}_k = \mathbf{H}$ for all $k = 1, \ldots K$. The standard PARAFAC does not have a unique loading matrix in the J direction for each slab $\mathbf{X}_k$ but instead uses a single B for the entire three-way structure. Different $\mathbf{B}_k$ loadings allow PARAFAC2 to directly model three-way arrays of batch data of various lengths and make it more flexible when handling shifts in batch data.

Fitting a PARAFAC2 model is the least squares minimization of the following loss function.

$$\sum_{k=1}^{K} ||\mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}_k^\top||^2 \tag{4.13}$$

Note that **C** is a 2D matrix, and consequently a direct multi-linear regression onto **Y** is possible using **C** as the predictors. This would be the multimodal equivalent of Principal Component Regression (PCR). However, there is no guarantee that the **C** score matrix is predictive of **Y** as it attempts to summarize the entire **X**, array. Thus, changes in **X**, which may have no significance on the output **Y**, will still affect the **C** matrix.
For prediction purposes, it is sought to seek a score matrix **C** that is relevant for predicting **Y**. For a single dependent variable **y**, the prediction capabilities are expressed with the following loss function.

$$||\mathbf{y} - \mathbf{C}\mathbf{r}||^2 \tag{4.14}$$

where **r** is a vector of regression coefficients. Making a predictive model relevant for **X** and **y** requires minimizing both loss functions. This is the same setup as in the two-way Principal Covariate Regression (PCovR), where a weighing parameter α between 0 and 1 is introduced. This parameter controls to what degree the fitting should summarize **X** or predict **y**. The SCREAM model is then fitted by minimizing a combination of loss functions 4.13 and 4.14.

$$\alpha \sum_{k=1}^{K} ||\mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}_k^\top||^2 + (1-\alpha)||\mathbf{y} - \mathbf{C}\mathbf{r}||^2 \tag{4.15}$$

The PARAFAC2 direct fitting algorithm[7] is utilized to solve this minimization problem while maintaining the uniqueness constraint. These modeling techniques have two hyperparameters that must be determined, the number of factors F and the value of the weighing parameter α. Improper selection of these parameters leads to models that do not

predict or overfit **y**. The original MATLAB scripts used in this study for building SCREAM models are available at http://www.models.life.ku.dk/scream.

Initial implementations of the model on batch data performed poorly, where even resampling the training data would lead to different predictions. The suspected reason for this discrepancy is due to a local minima issue. PARAFAC2 direct fitting algorithm utilizes A to minimize the fitting function. Prediction of new samples using PARAFAC2 is not trivial. Predicting new samples also uses ALS but keeps the variable matrix **A** and cross-product matrix **H**, fixed. Thus, in the presence of multiple local minima, different initial guesses can give different results. The original implementation utilized a random initial guess for the score matrix **C**. We propose modifying the original SCREAM implementation by changing how **C** is initialized when predicting new samples. Instead of random guesses of **C**, we propose selecting the score values $\mathbf{c}_k$ for each new batch to match the same $\mathbf{c}_k$ of the batch it most closely resembles from the training set. To determine which batch in the calibration set most corresponds new batch in a prediction, we extract the first Principal Component from each batch in the calibration set and the prediction set and compare the geometric distance. The $\mathbf{c}_k$ of the batch in the calibration with the lowest distance from a new batch when comparing the Principal components is used to initialize the PARAFAC2 direct fit algorithm. This modification vastly stabilized the prediction element of the SCREAM model.

4.2.2 Preprocessing

While the SCREAM model does not require time alignment, it is still necessary to perform centering and scaling operations when working with batch process data. Furthermore, PARAFAC2 models are considered more sensitive to noise captured in the dataset compared to other multivariate methods because time profiles are estimated for each batch separately[16]. To address this, a Savitzky-Golay derivative filter is used to smooth each process variable to reduce the overall noise; this is a common technique when modeling batch data[17].

Multi-modal data centering and scaling are more complex and can be done in multiple ways[18]. Improper centering and scaling can reduce model qualities and even lead to degeneracies in multi-way models[19]. Research considering multivariate methods on batch data indicates that mean-centering the data followed by single slab scaling on a

mean-centered array leads to more predictive models[20] and thus will be used throughout this work.

Mean centering is done by computing and subtracting the mean for each time point within the batch period.

$$\overline{x}_{ij} = \frac{\sum_{i=1}^{I} x_{ijk}}{K} \tag{4.16}$$

$$x^*_{ijk} = x_{ijk} - \overline{x}_{ij} \tag{4.17}$$

Single slab scaling is a method in which all time points for a single process variable are scaled for equal cumulative variance. To account for the uneven data array, the data is temporarily made even by extending each batch duration to the longest duration batch (J_{max}) and filling the matrix slabs with missing elements or NaN values. The scaling is then done with the following;

$$RMS_i = \sqrt{\frac{\sum_{k=1}^{K}\sum_{j=1}^{J_{max}} x_{ijk}^2}{IJ_{max} - N_{NaN}}} \tag{4.18}$$

$$x^*_{ijk} = \frac{x_{ijk}}{RMS_i} \tag{4.19}$$

Where N_{NaN} is the number of missing elements when the matrices are artificially extended to make the data structure even.

An additional preprocessing step is performed only before fitting a SCREAM model, where **X** and **Y** blocks are scaled to have a sum of squares of 1. This makes it easier to select optimal alpha values during hyperparameter optimization.

4.2.3 Model implementation

A model implementation workflow was developed, followed throughout the study, and is shown in Figure 4.1. The starting point is a raw dataset containing time series of multiple monitored process variables throughout multiple batches. Whether batches require filtering depends on the level of noise captured in the dataset, and the final model is used. Since PARAFAC2 algorithms are more sensitive to noise and missing data, we especially recommend some sort of data filtering or smoothing if SCREAM is intended to be used.

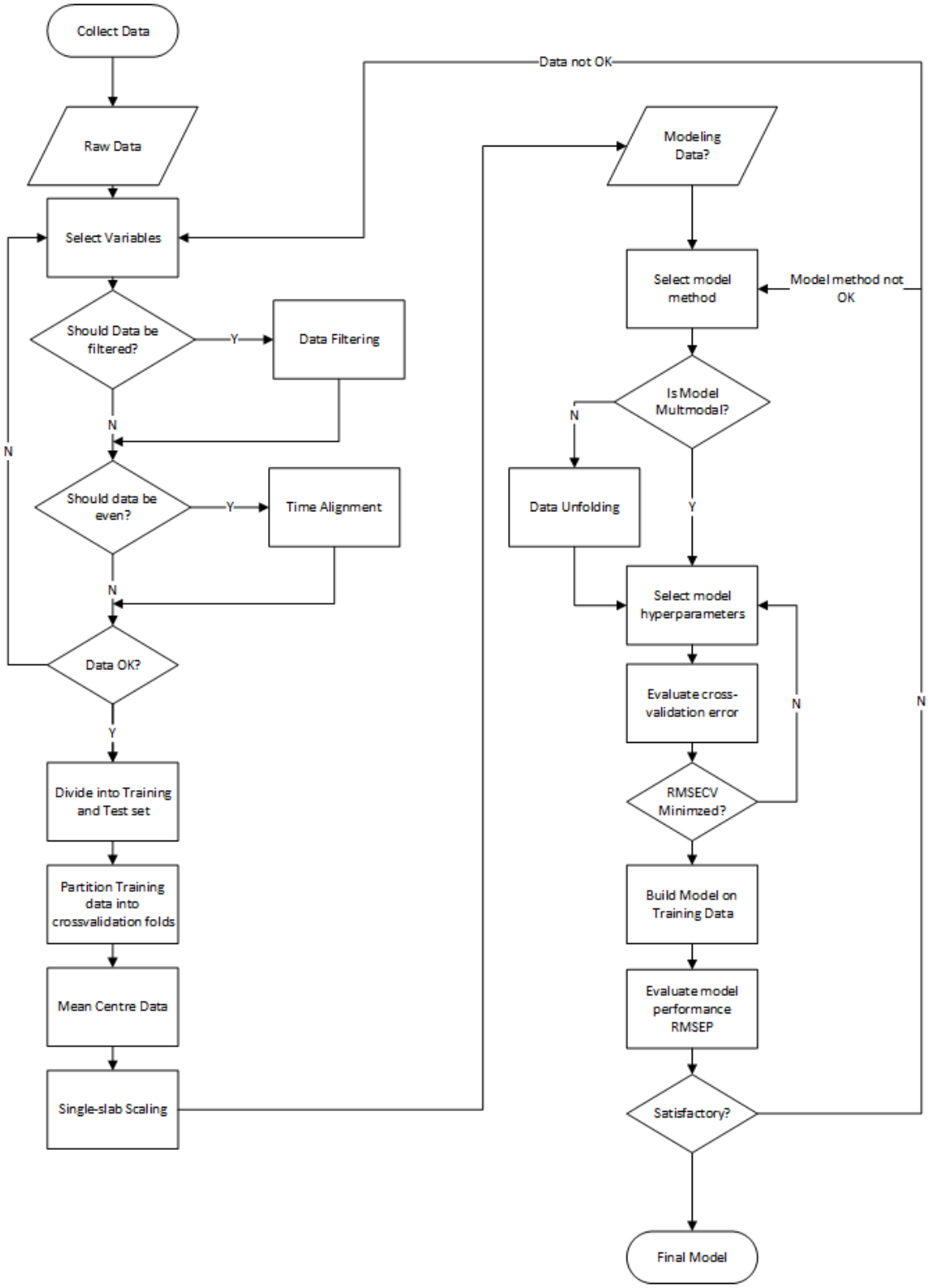

Figure 4.1: Proposed methodology for multimodal or multivariate regression modeling of batch process data

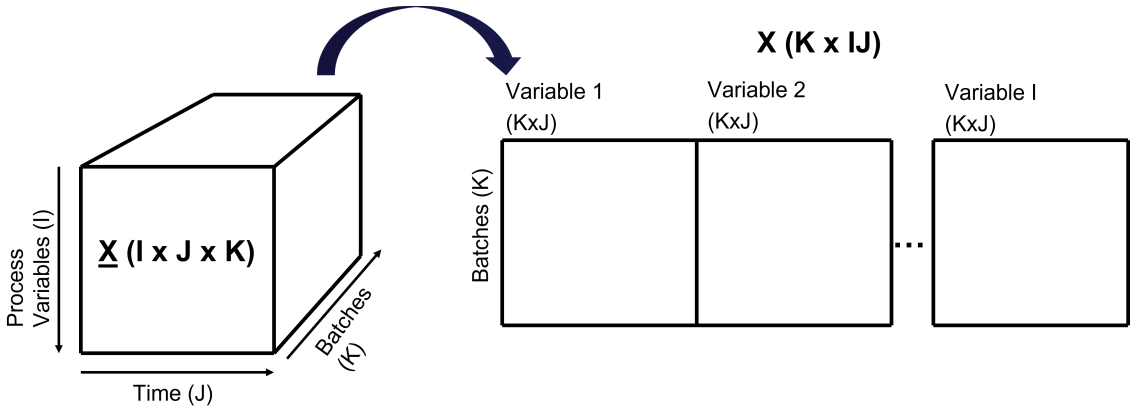

Figure 4.2: Illustration of unfolding three-way batch data into a 2-D matrix

Time alignment is required when NPLS or unfold-PLS models are used but not when selecting SCREAM models. For this work, only the cut-to-shortest method is used to align the data. This is a very simplistic batch trajectory synchronization method where all data that passes a time point exceeding that of the shortest duration batch is simply omitted from the dataset used for modeling.

Cross-validation is the recommended hyperparameter tuning technique when using experimental data[21]. The Venetian blinds segmentation method divides the batches into each segment. Batches are segmented before mean centering and multiway scaling is applied. This ensures that the left-out-batches during cross-validation do not influence the model calibration at all[22]. In the case of Unfold PLS, a separate unfolding step is required. There are multiple ways of unfolding three-way matrices. Still, for this work, we will only consider unfolding to preserve the variable dimension, i.e., all measurements of a single process variable are grouped in the final matrix. Given a 3-D batch data array containing I process variables, J time indices, and K batches, the unfolding returns a 2-D matrix of size ($KxIJ$). This is illustrated in Figure 4.2. When tuning the hyperparameters for the different model types, we recommend a cross-validation and grid search optimization routine. MPLS and NPLS models only have a single integer hyperparameter representing a number of factors or components. Thus a grid is not necessary, but rather the number of factors is increased until the root mean square error in cross-validation (RMSECV) increases which is an indicator of overfitting.

$$RMSECV = \sqrt{\frac{\sum_{k=1}^{N} \left(y_k - \hat{y}_{k,CV}\right)^2}{N}} \tag{4.20}$$

Where $\hat{y}_{k,CV}$ is the predicted value of the response for the i'th sample in cross-validation. SCREAM models utilize the same hyperparameter but also introduce α that can take any value between 0 and 1. When optimizing hyperparameters for SCREAM models, a grid search from 0.5 and 1 with a step size of 0.02 is used to find the α value, which provides the lowest RMSECV for each factor. When hyperparameters are found that minimize RMSECV, a new model is built using all the training data with the selected hyperparameters. The model quality is reported by examining the model root mean square error on the test data prediction error (RMSEP) and associated bias.

$$bias = \frac{\sum_{k=1}^{n}(y_k - \hat{y}_k)}{K} \tag{4.21}$$

4.3 Results and Discussions

4.3.1 Penicillin Data

A generated dataset based on the industrial simulation of Penicillin by Goldrick et al.[23]. This simulator is freely available at www.industrialpenicillinsimulation.com. The simulation has been widely used to analyze the applicability and benchmark new model methodologies of bioprocesses at industrial scales[24]. The dynamic simulations are heavily detailed and have the option of intentionally adding process faults to test model robustness. The list of monitored process variables is shown in Table 4.1.

Table 4.1: List of monitored process variables used for regression using the simulated dataset

Number	Process Variable
1	Sugar feed rate (L/h)
2	Acid flow rate (L/h)
3	Base flow rate (L/h)
4	Cooling water flow rate (L/h)
5	Airhead pressure (bar)
6	Substrate concentration (g/L)
7	Dissolved oxygen concentration (mg/L)
8	Broth Volume (L)
9	Vessel Weight (kg)
10	pH
11	Temperature (K)
12	Generated heat (kJ)
13	Carbon dioxide percent in off-gas $(\%)$
14	PAA flow rate (L/h)
15	Oxygen Uptake rate (g/min)
16	Oxygen percent in off-gas $(\%)$
17	Carbon evolution rate (g/h)

Two types of data are generated, and both are uneven in length. One type was generated by running the simulator repeatedly under normal operating conditions with slight variations. The simulator automatically induces variations when simulating multiple batches without user interventions by random permutations of the internal biochemical kinetic parameters. This dataset consists of 100 reference batches and is used for training and evaluating the model. The second dataset was generated by running the simulator by considering different faults occurring during the process. The description of each fault number is listed in Table 4.2. Each fault number also has slight variations, and ten batches

are simulated for each fault type for a total of 60 faulty batches. This ensures that both minor and significant deviations in batch harvest are observed. The data containing faulty batches are not used for any model calibration and are purely used to test model applicability and robustness.

Table 4.2: List of monitored process variables used for regression using the simulated dataset

Fault no.	Fault	Magnitude	Time (hr)
1	Disturbance in aeration flow rate (L/hr)	20,20	[20,24],[100,110]
2	Disturbance in vessel back pressure (bar)	2,2	[100,104],[200,230]
3	Disturbance in base flow rate (L/hr)	2,20,20	[20,30],[76,92],[200,214]
4	Disturbance in base flow rate (L/hr)	5,10	[80,84],[140,160]
5	Disturbance in coolant flow rate (L/hr)	2	[70,90]
6	All of the above faults		

Dataset being uneven means that the batches had varying runtimes, with the shortest batch length being fermented for 82 hours; thus, when building NPLS and Unfold-PLS, a time-cut of all data after 82 hours of batch runtime is removed to make the data even in length. However, these omitted readings will be included when building a SCREAM model. The models are calibrated to predict the total harvested Penicillin amount at the end of the batch.

During model calibration, the batches with induced faults are not included but rather reserved for testing model robustness and the ability to predict outcomes when the process is not running according to specifications. Before building the model, the data is partitioned by setting 30 of the 100 good batches aside to test model quality during standard operation. The remaining 70 are then divided into equal-sized nine segments for cross-validation in order to select the appropriate model hyperparameters. Unfold-PLS and NPLS models with two components had the lowest prediction errors during cross-validation. The results of hyperparameter tuning of the SCREAM model are shown in Figure 4.3. Utilizing two factors and setting $\alpha = 0.52$ led to the lowest cross-validation error. A total of 160 batches were simulated and subsequently analyzed with the proposed methodology. The batches had varying runtimes, with the shortest batch length being at 82 hours, thus when building NPLS and Unfold-PLS a time-cut of all data after 82 hours of batch runtime is performed to make the data even in length.

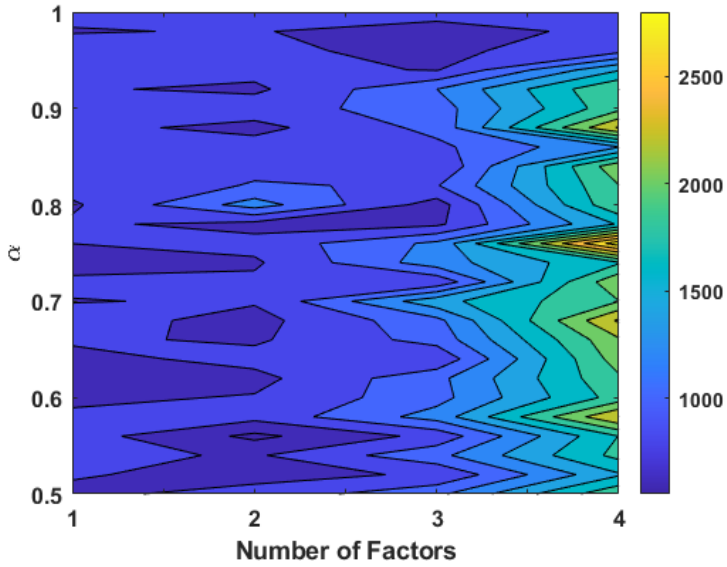

Figure 4.3: Results of grid search hyperparameter tuning of SCREAM model for the simulated penicillin dataset

Table 4.3: Comparison of the results obtained by different regression methods on the simulated industrial Penicillin dataset

Method	RMSEP Normal Batches	Bias	RMSEP Faulty Batches
Unfold-PLS	611	163.4	848
NPLS	647	122.8	1057
SCREAM	588	189.5	20602

During model calibration, the batches with induced faults are not included but rather reserved for testing model robustness. 100 of the simulated batches have no induced faults and 30 of them are set aside to test model quality during standard operation. The remaining 70 are divided into 9 equal-sized segments for cross-validation in order to select the appropriate model hyperparameters. Venetian blinds technique is used to divide the batches into segments. When optimizing hyperparameters for SCREAM models, a grid search from 0.9 and 1 with a step size of 0.01 is used to find the α value which provides the lowest RMSECV for each factor.

The resulting model performances are shown in Table 4.3. SCREAM model is the most accurate when predicting the penicillin harvest when only considering standard operation with a root mean square error of 588 kg when looking at the test data. Figure 4.4 shows the final model regression using SCREAM.

It is interesting to look at the performance metrics when the models are to predict the

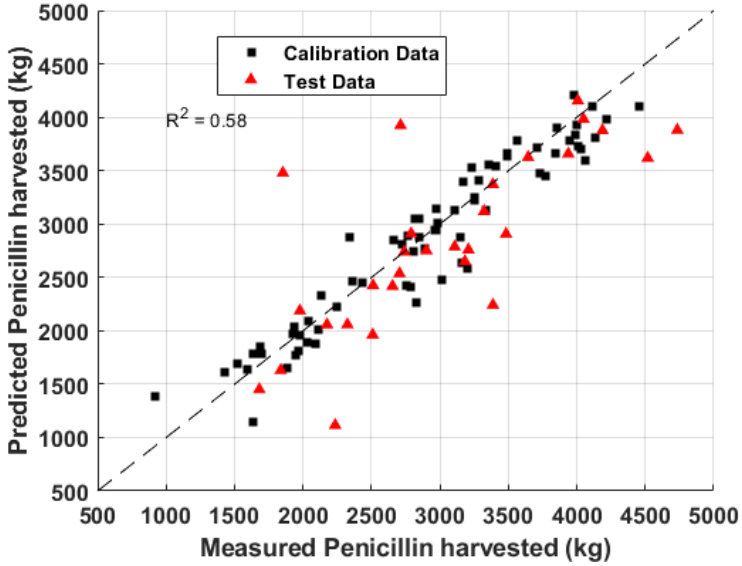

Figure 4.4: Parity-plot showing SCREAM model predictions compared to the measured Penicillin harvest during normal operation

outcome of faulty batches. While SCREAM predicts better during normal operation, the model is entirely unviable when batches are operated in a way that does not conform to normal operation. This indicates that while SCREAM can obtain more accurate predictions within the scenarios, it's calibrated in. Still, the increased flexibility and complexity result in trade-offs regarding outliers or batches with faulty operations. Similar effects can be seen in model bias and two very noticeable outliers in Figure 4.4.

Unfold-PLS performs slightly better than NPLS for this case study. This is not unexpected as Unfold-PLS models will always describe an equal or higher amount of covariance and have a higher number of model parameters than NPLS with an equal number of components. It was a bit more surprising to see that Unfold-PLS models also outperform both SCREAM and NPLS when predicting the outcome of faulty batches. Looking at Figure 4.5, NPLS undershoots significantly when the harvest is under 1000 kg leading to a higher RMSEP value. Otherwise, they have very similar overall prediction performance and similar systematic errors depending on the fault numbers used. While the SCREAM method shows stronger predictive qualities when focusing purely on batches at standard operation, this process may have better overall choices. This is not only due to its incapability to establish any predictive qualities when batches are not running according to specifications but also to the sheer difficulty of building and maintaining such a model. There is a

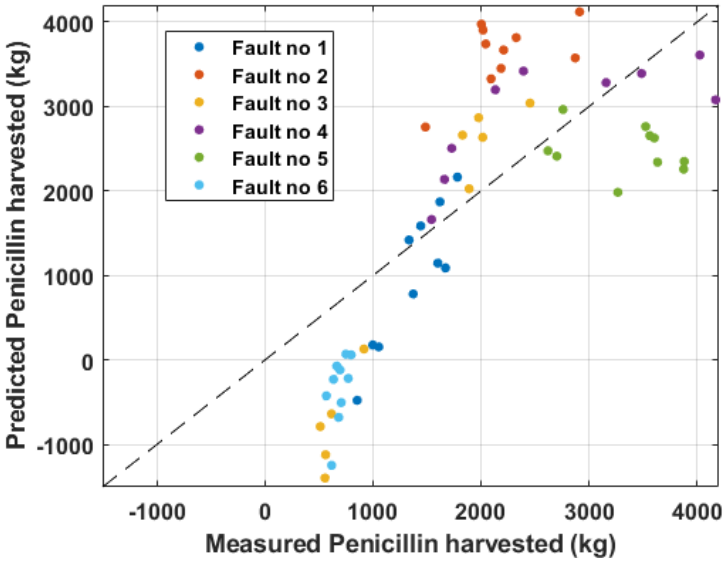

Figure 4.5: Parity-plot showing NPLS model predictions compared to the measured Penicillin harvest during faulty operation

relatively large degree of bias in the SCREAM model, indicating a bit of overfitting to get decent cross-validation. Poor results on faulty batches suggest that this model configuration is susceptible to small changes in the data.

Furthermore, building adequate SCREAM models requires much more resources and time than the PLS counterparts. This is mainly due to the drawbacks of the PARAFAC2 engine compared to the sequential component approach of PLS. PARAFAC2 models generally take longer to build due to the ALS algorithm required to find the optimal loading matrices. It is also recommended to run multiple different starting iterations of the PARAFAC2-ALS due to a risk of local minima, which adds to the runtime. Furthermore, because SCREAM implementations add additional hyperparameters that need to be established for accurate model regression, we are already seeing a grid of 100s of potential model configurations compared to NPLS's meager seventeen possible configurations, i.e., the maximum possible number of components, where each SCREAM configuration takes significantly longer to evaluate correctly. Proper expertise can help lower the search space, but PLS-based models will always be significantly faster to develop.

Suppose the extra prediction accuracy gained during regular operation is not crucial for any other processes or analysis. In that case, we do not recommend selecting the SCREAM modeling method over PLS for the simulated case study. PLS-based models are much

faster and easier to build and can retain some prediction power during faulty operation. In this case study, selecting between NPLS and Unfold-PLS would depend mainly on the model's purpose. If only regression is required, then Unfold-PLS is clearly the correct choice as it outperforms NPLS even on faulty batches. However, if model interpretation is desired, then NPLS has some advantages because of the explicit modeling of variable and time mode.

4.3.2 Industrial Case study

LEO Pharma provides another dataset used in this work for a different case study. The data is obtained from a fed-batch production utilizing a filamentous fungus and consists of 43 batches in total. This dataset consists of two dependent variables to be predicted. The goal of the analysis is to predict the final concentration of the main product and the concentration of related substances, which directly correlate to batch quality. To get a better insight into the applicability of each method for either purpose, separate models are built using the same modeling method to predict one quality variable at a time.
Of the 43 batches, 13 are held over for testing model performance, leaving 30 batches for model calibration. The training dataset is further segmented into six segments for k-fold cross-validation to determine model hyperparameters. Model performance indicators are also only reported on the test set. For this case study, model performance is reported in this article using the root mean sum of square error as a percentage of the mean (RMSSE) for confidentiality reasons.

$$RMSSE(\%) = 100 * \frac{RMSEP}{\hat{y}} \tag{4.22}$$

Where $\hat{y}$ is the mean value of the quality variable for all measurements in the test set.

Predicting main product concentration was possible with all three modeling types within acceptable levels of accuracy. Both PLS-based methods can sufficiently predict main product yields and do it even slightly better than the best-obtained SCREAM model. The parity plot in figure 4.6 doesn't reveal any significant outliers that would severely affect the summary statistics. Still, there seems to be a significant bias in the test data toward with model underestimating the main product concentration. Furthermore, the R^2 value of 0.47 is low but slightly improved over bilinear methods, as shown in table 4.4. The

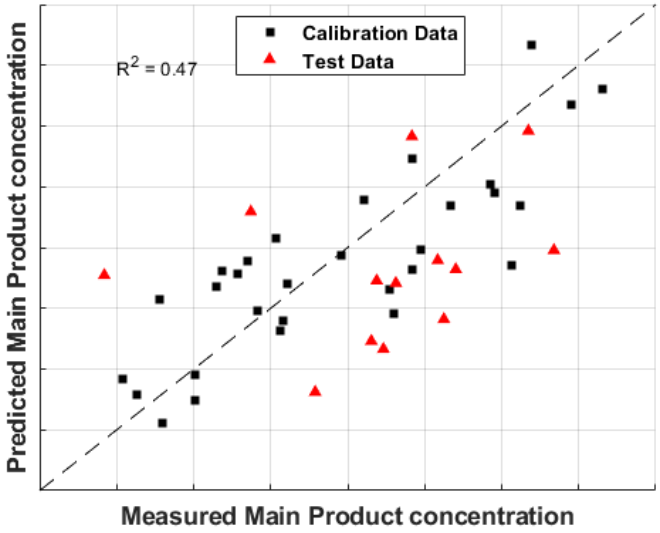

Figure 4.6: Parity-plot showing SCREAM model predictions compared to the measured Main Product concentration at harvest with data from Industrial sponsor

poor correlation may indicate that the dataset exhibits significant non-linear effects when predicting the main product across all modeling methods. This is expected as the main product is the result of fermentation which is a non-linear process. Since all modeling methods discussed here are linear, there is a limitation on the predictive capabilities when utilizing the entire dataset. Better results can be obtained by incorporating a non-linear kernel projection or utilizing first principles models.

Table 4.4: Comparison of the results obtained by different regression methods on the dataset obtained from LEO Pharma

Method	RMSSE (%) Main Product	R^2	RMSSE Related Substances	R^2 (%)
Unfold-PLS	11.21	0.46	93.24	0.35
NPLS	12.55	0.33	98.52	0.39
SCREAM	12.73	0.47	42.74	0.83

However, in the case of Related Substances concentration, there is a significant quality improvement when selecting the SCREAM model over PLS-based methods. The RMSEE drops from 98.5% down to 42.7% purely by using the novel model type. The comparison between SCREAM-predicted and measured related substance concentration is shown in Figure 4.7. Since related substance concentration is directly related to batch quality, it is more important to predict the high concentration. It indicates that the model can be

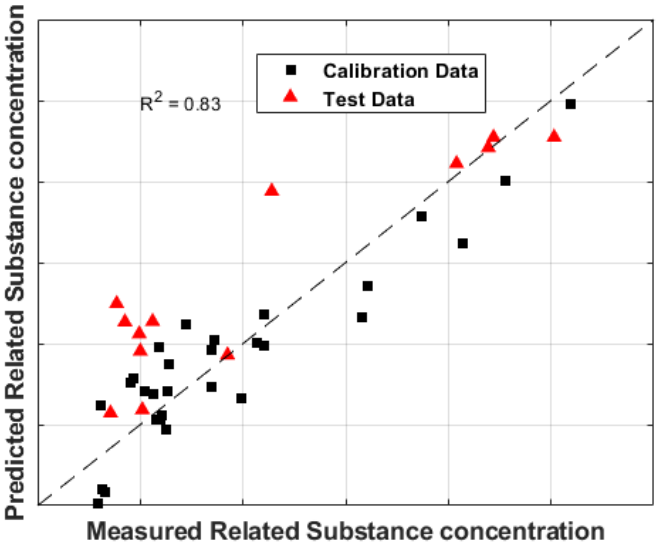

Figure 4.7: Parity-plot showing SCREAM model predictions compared to the measured related substance concentration at harvest with data from Industrial sponsor

directly used to determine if the batch is up to standard before initiating the downstream process. The additional flexibility of the SCREAM method allows the modeling of related substances within acceptable accuracies in the process, which was previously impossible.

For the industrial dataset, SCREAM shows significant improvement over NPLS when looking at the related substance concentration variable. The expected reason is that data lost during the cut-to-shortest time alignment method is crucial for accurate predictions of related substances. PLS methods could achieve similar prediction levels with better time alignment methods if this is the case. However, SCREAM is preferred as more sophisticated time alignment to achieve similar prediction accuracy may destroy the interpretability of the models.

However, the data lost at the end or shifts in the data do not seem to affect the prediction of the main product as severely, and PLS methods can maintain good predictions for that variable. The simulated case study shows that SCREAM is risky to use in general and may have no predictive power when extrapolating outside the scenarios it is calibrated in, such as in cases where an unexpected fault during operation occurs. It is impossible to confirm that this would also be the case for this dataset as it does not contain any batches that have known faulty operations. For this case study, NPLS is recommended

as the modeling tool when analyzing or predicting the productivity or amount of main product obtained from the batch. However, the flexibility of SCREAM is required when investigating the batch quality or accumulation of related substances.

4.4 Conclusions

This work proposes a methodology to create multi-modal regression models for the analysis of bioprocess batch data. Two different multi-way models are examined, and the predictive power is compared to the more traditional multiblock methods. Modifications to the novel SCREAM model are proposed to make model predictions more consistent when handling bioreactor data. When studying the penicillin simulation process, it was determined that the SCREAM method had the lowest error when only normal operating conditions were considered. However, SCREAM has some drawbacks in terms of the difficulty of modeling and its risk of being overly sensitive to minor changes. SCREAM could only predict batches with no significant deviations from the calibration set and lost all predictive qualities when the simulation was run with induced faults.

This work focused on quality variable prediction, and from that point of view, Unfold-PLS is the superior model when predicting the main product in both case studies due to relatively low prediction error on the calibration set and fairly robust when considering faulty batches. Of course, modeling objectives are not mutually exclusive, and it is possible to build models focusing on quality prediction and use them as a basis for process optimization or monitoring. With this in mind, NPLS could be the preferred model type since the prediction errors are not considerably worse. Multi-modal models contain fewer parameters than their unfolded counterparts and preserve the multilinear structure, making it easier to interpret models and generate valuable insights.

This study indicates that the powerful and flexible SCREAM models are not recommended as the first choice of models. However, if severe shifts or varying runtimes are suspected to affect model predictive power negatively, SCREAM is a promising regression method. The most significant advantage of SCREAM is the ability to solve the uneven-length problem naturally. The case study relating to industrial antibiotic fermentation shows that our proposed modifications to the SCREAM modeling methodology allowed the model to accurately predict the batch quality of related substances without any batch trajectory syn-

chronization steps, which is something Unfold-PLS and NPLS model types failed to do. This confirmed the SCREAM method as a practical engineering tool in the data-driven analysis of batch process data.

Bibliography

[1] Krist V. Gernaey, Anna Eliasson Lantz, Pär Tufvesson, John M. Woodley, and Gürkan Sin. Application of mechanistic models to fermentation and biocatalysis for next-generation processes. *Trends in Biotechnology*, 28(7):346–354, 2010.

[2] Frank Westad, Lars Gidskehaug, Brad Swarbrick, and Geir Rune Flåten. Assumption free modeling and monitoring of batch processes. *Chemometrics and Intelligent Laboratory Systems*, 149:66–72, 2015.

[3] Gareth John Macdonald. Stars in alignment for artificial intelligence in bioprocessing. *Genetic Engineering and Biotechnology News*, 41(2):40–44, 2021.

[4] J. M. González-Martínez, J. Camacho, and A. Ferrer. Mvbatch: A matlab toolbox for batch process modeling and monitoring. *Chemometrics and Intelligent Laboratory Systems*, 183:122–133, 2018.

[5] D. J. Louwerse and A. K. Smilde. Multivariate statistical process control of batch processes based on three-way models. *Chemical Engineering Science*, 55(7):1225–1235, 2000.

[6] Yang Zhang, Bo Lu, and Thomas F. Edgar. Batch trajectory synchronization with robust derivative dynamic time warping. *Industrial Engineering Chemistry Research*, 52(35):12319–12328, 2013.

[7] Henk A. L. Kiers, Jos M. F. ten Berge, and Rasmus Bro. Parafac2—part i. a direct fitting algorithm for the parafac2 model. *Journal of Chemometrics*, 13(3□4):275–294, 1999.

[8] Lijia Luo, Shiyi Bao, Jianfeng Mao, and Di Tang. Quality prediction and quality-relevant monitoring with multilinear pls for batch processes. *Chemometrics and Intelligent Laboratory Systems*, 150:9–22, 2016.

[9] William J. Egan, S. Michael Angel, and Stephen L. Morgan. Comments on three-way analyses used for batch process data. *Journal of Chemometrics*, 15(1):19–27, 2001.

[10] Leo H. Chiang, Riccardo Leardi, Randy J. Pell, and Mary Beth Seasholtz. Industrial experiences with multivariate statistical analysis of batch process data. *Chemomet-*

rics and Intelligent Laboratory Systems, 81(2):109–119, 2006.

[11] Federico Marini and Rasmus Bro. Scream: A novel method for multi-way regression problems with shifts and shape changes in one mode. *Chemometrics and Intelligent Laboratory Systems*, 129:64–75, 2013.

[12] Age K. Smilde and Henk A.L. Kiers. Multiway covariates regression models. *Journal of Chemometrics*, 13(1):31–48, 1999.

[13] Rasmus Bro. Multiway calibration. multilinear pls. *Journal of Chemometrics*, 10(1):47–61, 1996.

[14] P Nomikos and JF MacGregor. Multi-way partial least squares in monitoring batch processes. *Chemometrics and Intelligent Laboratory Systems*, 30(1):97–108, 1995.

[15] T Kourti, P Nomikos, and JF Macgregor. Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway pls. *Journal of Process Control*, 5(4):277–284, 1995.

[16] José Manuel Amigo, Thomas Skov, Rasmus Bro, Jordi Coello, and Santiago Maspoch. Solving gc-ms problems with parafac2. *Trac - Trends in Analytical Chemistry*, 27(8):714–725, 2008.

[17] Pau Cabañeros Lopez, Isuru A. Udugama, Sune Tjalfe Thomsen, Christian Roslander, Helena Junicke, Miguel Mauricio□Iglesias, and Krist V. Gernaey. Towards a digital twin: a hybrid data□driven and mechanistic digital shadow to forecast the evolution of lignocellulosic fermentation. *Biofuels, Bioproducts and Biorefining*, 14(5):1046–1060, 2020.

[18] Stephen P. Gurden, Johan A. Westerhuis, Rasmus Bro, and Age K. Smilde. A comparison of multiway regression and scaling methods. *Chemometrics and Intelligent Laboratory Systems*, 59(1-2):121–136, 2001.

[19] Rasmus Bro. Parafac. tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 38(2):149–171, 1997.

[20] Lisa Mears, Rasmus Nørregaard, Gürkan Sin, Krist V. Gernaey, Stuart M. Stocks, Mads O. Albæk, and Kris Villez. Functional unfold principal component regres-

sion methodology for analysis of industrial batch process data. *A I Ch E Journal*, 62(6):1986–1994, 2016.

[21] Marlies Vervloet, Katrijn Van Deun, Wim Van den Noortgate, and Eva Ceulemans. On the selection of the weighting parameter value in principal covariates regression. *Chemometrics and Intelligent Laboratory Systems*, 123:36–43, 2013.

[22] D. J. Louwerse, Age K. Smilde, and Henk A.L. Kiers. Cross-validation of multiway component models. *Journal of Chemometrics*, 13(5):491–510, 1999.

[23] Stephen Goldrick, Andrei Ştefan, David Lovett, Gary Montague, and Barry Lennox. The development of an industrial-scale fed-batch fermentation simulation. *Journal of Biotechnology*, 193:70–82, 2015.

[24] Piyush Agarwal, Mohammad Aghaee, Melih Tamer, and Hector Budman. A novel unsupervised approach for batch process monitoring using deep learning. *Computers and Chemical Engineering*, 159:107694, 2022.

5 Mechanistic Modelling of Industrial scale batches for antibiotic production

Abstract

An unstructured mechanistic model is proposed to describe the industrial-scale production of Fusidic Acid in fed-batch cultivations. The model accounts for differences in dead and viable biomass and the effect of the primary carbon source and oxygen on cell growth and production. The model parameter is calibrated, and performance is tested using experimental data obtained from an operating industrial production in Denmark. Parametric uncertainty and correlation are considered using a statistical bootstrap analysis technique. Model reliability w.r.t. parametric uncertainty is analyzed utilizing a Monte Carlo simulation approach to calculate the effects of uncertainty propagation. The model predicted the main product concentration with a relative mean error of 7%. Model uncertainty in main product harvest is extremely low, indicating reliable model outputs. These successful implementations open up opportunities for soft-sensor implementations for key state variables but also set a good foundation for further model extensions, such as hybrid modeling.

5.1 Introduction

Industrial microbiology has been one of the most common methods for the mass production of antibiotics since the wide-scale production of penicillin. In a very similar sense, Fusidic Acid is a secondary metabolite that is commercially produced using a filamentous microorganism[1]. However, while there is extensive literature on the modeling of penicillin production with varying degrees of complexity[2][3][4][5], there is no published research on the mechanistic modeling of the Fucidin process. The bioprocess industry has seen rapid advancements toward digitalization. This has led to increasing interest in industrial applications of digital twins. These technologies require good-quality process models to act as the foundation. Mechanistic models are the gold standard as they incorporate current knowledge into a first principles mathematical description of the system of interest[6]. While more difficult to develop, these model types are notably beneficial for application in the fermentation industry since they can be better extrapolated to new scenarios than machine learning algorithms and artificial intelligence.

A mechanistic model describes the dynamic behavior of a system with a series of mathematical formulas, typically ordinary differential equations (ODEs). These models are based on prior knowledge of the system phenomena like mass, energy, and moment balances.

The mechanistic model of Bajpai and Reuss [2] is a good starting point when designing a model structure for a poorly researched organism. It's unstructured, has low complexity, few parameters, and can be calibrated based on experimentally measured growth profiles. Furthermore, it was made to model the secondary metabolites of a filamentous organism and showed excellent agreement with experimental data[7]. Extensions and changes will be proposed based on observed phenomena on industrial and lab scales to improve predictive qualities and applications.

When developing and applying models for bioprocesses, it is considered good modeling practice to analyze the model's reliability. To that end, the models developed here will be done through the framework of Good modeling practices proposed by Sin et al. [8], which advocate the use of statistical methods to analyze inherent model uncertainties resulting from experimental and the overall identifiability of the model. This allows judging the model's fitness to the purpose under uncertainty as a proactive solution when dealing

with uncertainties for process development[9].

This study aims to design and evaluate a mechanistic model to quantitatively describe the behavior of an industrial scale Fucidic Acid process. To this end, a mechanistic model structure is inspired by unstructured penicillin models with extensions or modifications to account for observable phenomena seen both in the production setting and during lab experiments performed at LEO Pharma A/S. Industrial-scale data is collected to calibrate and validate model parameters and structure. A statistical parameter estimation technique was performed to get more reliable parameter mean values and quantify parametric uncertainties in the model based on available data. Monte Carlo simulations of the dynamic model were performed using estimated uncertainties from the parameter estimation techniques to propagate model uncertainties to the predicted output. Finally, output distributions of the main product are used to assess the model's applicability in an industrial setting.

5.2 Materials and Methods

Before delving into experimental data collection and model structure, it's important to note that all the data and subsequent modeling work is done on a mass basis rather than a volume basis. It is common in the literature to report component concentrations in g/L, but this work will utilize g/kg_{Broth}. There are three primary reasons for this

1. Measuring fermentation broth mass is significantly easier than broth volume.
2. Broth densities will vary throughout fermentation. This is sometimes overlooked despite being a popular method of estimating alcoholic content in beer production[10].
3. Water evaporation is significantly easier to calculate on a mass basis.

5.2.1 Fermentation Media

Samples were obtained from the main bioreactors used for the commercial manufacturing of Fusidic Acid at the Ballerup site of LEO Pharma A/S. In addition, fresh samples were taken from industrial fermentations from various cultivation times. The conditions and media are similar to the process description of Fusidic Acid fermentation reported by Daehne[11]. Still, exact details regarding present-day operating conditions, component concentrations, and microorganisms are considered sensitive information and are

not disclosed.

5.2.2 Experimental data collection

Cell Dry Weight

Samples are obtained from the production line and are transferred to the laboratory with the filtration equipment and the oven quickly after taking the sample. For filtration,70 mm diameter glass fiber filters are pre-dried for two days at 100°C. Before filtration, one of these filters was put in a holder made of aluminum foil and weighed together. The filter was then transferred to the 70 mm diameter Buchner funnel attached to the 1 L Buchner flask. Around 10-20 mL biomass fraction was slowly poured onto it while the vacuum was on. Next, the bottle with the remaining sample was weighed again, allowing biomass calculation on a mass basis. Next, the filtered biomass was washed with around 40 mL of distilled water at least twice, shaking up the biomass on the filter as much as possible to dissolve any solid particulates like sugar crystals. After washing, the wet biomass was placed in the oven with an aluminum foil container for two days at 100°C. After two days, the aluminum foil, filter, and biomass were placed in a desiccator and weighed after they were cooled down to room temperature.

Main product and byproducts

The main product and all relevant related substances are measured in an *offline* setting. We use a High-Performance Liquid Chromatography (HPLC) method that uses a Waters Acquity CSH Phenyl-Hexyl column. The column temperature is set to 60, and a mobile phase with 70% $MeOH$ with 30% 0.1 w/w% H_3PO_4 is used at 0.6 mL/min. The injection volume is 3 μ L, and the runtime is 6 min. Ethanol was used as the mobile phase solvent.

Viable Biomass

Viable biomass is measured via capacitance by placing approximately 2 mL of fermentation broth sample in a tube flask and using a Futura Pico probe. The capacitance signal is converted to viable biomass concentration using a relation developed explicitly for this strain. For more details on this method and obtaining a calibration curve, see chapter 3. The dataset used in this study did not contain direct measurement of viable biomass as the samples were collected before analysis equipment to measure it was obtained. However, when the equipment was available, and the experimental protocol was developed, a separate sampling campaign was commenced to compare viable biomass and CDW in

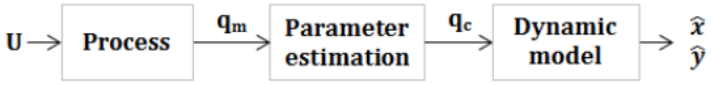

Figure 5.1: Model structure, works as soft sensor

the industrial process over the entire fermentation period. With this data, we could observe biomass's average viability throughout fermentation. To have additional observed variables for better model calibration, we augment the original dataset to introduce an estimated viable biomass concentration.

$$X_{Viable} \approx X_{TDW} - f_{Viability}(t) \tag{5.1}$$

Where $f_{viability}(t)$ is an experimental correlation that describes the cell viability as a function of time, specifically during standard operating conditions. The correlation is not reported here due to confidentiality reasons.

5.3 Model Structure

5.3.1 Component concentration balance

The following extension of the traditional fed-batch equation is utilized for the overall concentration balance. For component i, the instant change in concentration C_i is calculated.

$$\frac{dC_i}{dt} = q_i + \frac{E}{M}C_i + \frac{F_{feed}}{M}(C_{i,f} - C_i) + k_l a_i(C_i^* - C_i) \tag{5.2}$$

Where the first term q_i denotes the biochemical kinetic rate of component i, the second term accounts for increased concentration due to mass loss from water evaporation and offgas balance, where E describes the broth mass change due to evaporation and offgas balance $E = F_{evaporation} + F_{CER} - F_{OUR}$. It is assumed that all relevant components are non-volatile and thus are not present in the offgas. The third term describes the dilution due to feed where F_{feed} is the feed rate, M is the broth weight, and $C_{i,f}$ is the concentration of component i in the feed. The final term describes the mass transfer of component i from the gas phase to the liquid phase, where $k_l a_i$ is the mass transfer coefficient of component i and C_i^* is the equilibrium solubility of component i. For simplifications, it is assumed that kla is 0 for all components except oxygen.

As the model is unstructured, it is assumed that all the biochemical kinetics can be explained with a specific growth rate term. Which can be either positive, indicating growth/production, or negative, indicating consumption.

$$\boldsymbol{q} = \boldsymbol{\mu} X \tag{5.3}$$

A particular case is included with viable biomass, which can degrade based on exponential decay by introducing a death rate term k_d. The death rate is necessary as observations with bio-capacitance indicate that viable biomass decreases during the later stages of fermentation. However, there is never an observed reduction in Total Dry Weight (X_{TDW}), indicating an accumulation of cell debris.

$$q_X = (\mu_X - k_d)X \tag{5.4}$$

Which leads to the accumulation of cell debris X_D as

$$q_D = k_d X \tag{5.5}$$

To further account for the difference between measured total dry weight and measured viable biomass, it is important to note that the main product is Fusidic Acid which is poorly insoluble in aqueous solutions. Thus the precipitated main product $P_{precipitated}$ also contributes to X_{TDW}. Overall this becomes.

$$X_{TDW} = X + X_D + P_{precipitated} \tag{5.6}$$

5.3.2 Biomass Growth

The Contois model kinetics is utilized for biomass growth due to the excellent agreement with experimental data for multiple microbial systems[12]. The specific growth rate is expected to depend on a primary carbon source and dissolved oxygen. The lag phase when the fermenter is initially inoculated will also be considered, as proposed by Sin et

al.[13]. The overall growth is described by

$$\mu_X = \mu_{X,max} \frac{S}{K_{SX}X + S} \frac{O}{K_{OX}X + O} \left(1 - exp\left(\frac{-t}{t_{lag}}\right)\right) \tag{5.7}$$

Temperature effects are not considered due to limited variations in the experimental data. In addition, the pH result is observed to be inconsequential in the current conditions via lab studies and will thus not be included in the equation structure. This data relating to broth pH and growth is confidential and thus not shown or referenced.

5.3.3 Main product

As the fermentation utilizes a filamentous microorganism, it is assumed that the main product is a secondary metabolite and is not associated with growth. Therefore, the Contois kinetics are reused and are used to describe specific product growth. Dissolved oxygen is omitted for product growth as lab experiments have determined that the fungus can synthesize the main product at low oxygen concentrations. The data relating to dissolved oxygen concentrations is confidential and thus not shown or referenced.

$$\mu_P = \mu_{P,max} \frac{S}{K_{SP}X + S} \tag{5.8}$$

The main product is not fully soluble in aqueous solutions and will precipitate out of the liquid phase. Since the main product is a weak acid, solubility is based on pH, and the following relation is utilized.

$$P_{precipitated} = P - 2.453 \cdot 10^{-6} exp(1.808 * pH) \tag{5.9}$$

5.3.4 Substrate consumption

The main carbon source consumption is biomass growth, main product synthesis, and cell maintenance. The specific substrate consumption rate in the process is described as

$$\mu_S = \frac{\mu_X}{Y_{SX}} + \frac{\mu_P}{Y_{SP}} + m_S \frac{S}{K_{SS}X + S} \tag{5.10}$$

It is common to utilize a constant m_S as the consumption required for maintenance and other biological activities[14]. However, this has been criticized as improbable because

it predicts active substrate consumption despite being completely depleted[15]. Thus substrate consumed for maintenance is modeled in a Contois-like manner. As substrate depletes, the microorganism will utilize it less for maintenance purposes, and the specific consumption rate μ_S is zero when no substrate is present.

5.3.5 Mass balance and evaporation

The overall change in broth rate is estimated by considering each element that enters or leaves the fermenter. Since we are working on an industrial scale, any sample-taking is considered negligible and not considered. In an aerobic fed-batch process, there is a constant air supply, and the additional substrate is fed at specific time intervals based on the feeding strategy. Therefore, water evaporation is considerable at industrial scales if no condensers are used and should be included[16]. Furthermore, since the model is based on broth mass over volumes, we must also consider the effect of the gas evolutions. Aerobic fermentations consume O_2 and release CO_2, and there is a weight difference in the molecules which can lead to mass loss in the broth.

$$\frac{dM}{dt} = F_{feed} - F_{evaporation} + F_{OUR} - F_{CER} \tag{5.11}$$

Evaporation can be calculated as the difference between water vapor entering and leaving the fermenter.

$$F_{evaporation} = \dot{m}_{H_2O,out} - \dot{m}_{H_2O,in} \tag{5.12}$$

Water vapor in the process air can be estimated with the ideal gas law.

$$\dot{m}_{H_2O} = \frac{\phi p^*_{H_2O} Q}{RT} \tag{5.13}$$

The relative humidity is not measured in offgas; it is assumed to be saturated ($\phi = 1$) in the outlet. There have been several proposals for estimating saturation pressure, but a simple and effective method is to use a steam table or the Antoine equation.

$$log_{10} p^*_{H_2O} = A - \frac{B}{C + T} \tag{5.14}$$

The parameters used are from Bridgeman and Aldrich[17], $A = 5.08$, $B = 1838.67$, and

$C = -31.74$. Unfortunately, we could not determine an acceptable mechanistic model to describe Oxygen Uptake Rates (OUR) and Carbon Dioxide Evolution Rates (CER). However, these measurements are available with *realtime* offgas analysis via mass spectrometry.

5.3.6 Oxygen Mass Transfer

This model will only estimate the mass transfer of oxygen gas particles into the broth. We will use the common following empirical relation to calculate $k_L a$ of oxygen[18].

$$k_L a = c \left(\frac{P_{power}}{M} \right)^a v_g^b \tag{5.15}$$

Where a,b, and C are empirical constants, M is the total broth weight, P_{power} is the energy dissipation in the fermentation broth, and v_g is the superficial gas velocity.
The energy dissipation is calculated as the summation of agitation power and aeration

$$P_{power} = P_{agitator} + P_{air} \tag{5.16}$$

Where $P_{agitator}$ is the power imposed by the agitator. The energy dissipated due to aeration is calculated using the following relationship by Roels and Heijnen[19]

$$P_{air} = \frac{1}{22.4} \frac{v_g RT}{Z} ln(1 + \frac{\rho g Z}{P_o}) \tag{5.17}$$

Where R is the universal gas constant, T is the vessel temperature, g is the gravitational constant, ρ is the broth density, P_0 is the pressure at the bottom of the tank, and Z is the liquid height.

5.4 Modeling methodology

5.4.1 Parameter Estimation

The working strain contains minimal information in the literature, and no attempts have been made before to develop kinetic models. The biochemical kinetics portion of the model contains 11 parameters, and there is no prior information to draw from regarding potential parameter values. This lack of a previous study means an expert review is impossible, and all parameters must be estimated simultaneously.

We employ the bootstrap method[20] for parameter estimation to calibrate the model. This method is preferable as it can evaluate parameter distribution without assumptions about the underlying error distribution. Furthermore, the covariance and correlation between parameters during estimation can also be calculated easily from the results of bootstrap analysis. The method assumes that the experimental data is a function of the underlying model with added noise. If all the experimental observations are stored in a vector, **y** a frequentist approach assumes the following

$$\mathbf{y} = f(\theta) + \epsilon \tag{5.18}$$

Where θ is a vector containing all the true model parameters and ϵ is a vector containing measurement noise, assuming a stochastic process.
Bootstrap does not assume that a true parameter value exists but rather a parameter estimator θ as a random variable. Establishing the distribution of θ using bootstrap is done with the following steps

1. Perform a reference parameter estimation using non-linear least squares
   In this step, we use the *lsqnonlin* in MATLAB to find a parameter subset that minimizes the sum of square errors from the model predictions and experimental data, assuming no errors in the experimental measurement.

2. Generate synthetic data by bootstrap sampling and repeat parameter estimation
   A residual vector is generated using the model predictions calculated using the parameter set in Step 1. Then, synthetic data is generated by random sampling with replacement from the residual vector and adding it to the model prediction. For each synthetic data, the parameter estimation is repeated with the *lsqnonlin*, and the different parameter sets are recorded.

3. Review and analyze results
   This step revolves around the statistical analysis of all parameter sets obtained in Step 2. This involves finding the parameter mean value standard deviation, calculating the covariance and correlation matrix and plotting the distribution of the parameters, and concluding which parameters are identifiable within the framework.

We are primarily interested in the biochemical parameters described above for this work. Therefore, a complete list is presented in table 5.1.

Table 5.1: Summary of biochemical model parameters

Parameter	Description	Unit
$\mu_{X,max}$	Maximum specific biomass growth	$\frac{g}{kg\,h}$
K_{SX}	Contois saturation constant	$\frac{g}{kg\,h}$
t_{lag}	Lag Time	h
$\mu_{P,max}$	Maximum specific product synthesis	$\frac{g}{kg\,h}$
K_{SP}	Substrate limitation constant	$\frac{g}{kg}$
m_S	Substrate maintenance term	$\frac{g}{kg}$
K_{SS}	Maintenance saturation constant	$\frac{g}{kg}$
Y_{SX}	Biomass substrate yield coefficient	$\frac{g}{g}$
Y_{SP}	Product substrate yield coefficient	$\frac{g}{g}$
k_d	Biomass specific death rate	$\frac{g}{kg\,h}$
K_{OX}	Oxygen limitation constant	$\frac{g}{kg}$

5.4.2 Monte Carlo based Uncertainty analysis

Within the context of uncertainty analysis, which is concerned with estimating the error propagation from a set of inputs to a group of model outputs. The Monte Carlo is one of the most reliable and utilized methods for uncertainty analysis with complex chemical engineering models[21]. The Monte Carlo method is a numerical analysis method and includes three main steps:

1. Define input parameters and distribution
2. Sample from defined distribution using a random number generator
3. Perform the simulations with the generated samples
4. Statistical analysis and interpretation of the results

The definition and identification of the model input uncertainty range depend on the case study. The defined uncertainty range heavily influences the output of the Monte Carlo analysis; thus, each case study must be systematically evaluated to provide accurate uncertainty ranges. For step 1, uncertainties in the biological model parameters are considered and are obtained as a direct result of the bootstrap parameter estimation technique. They will be represented by a covariance matrix which includes a standard deviation and correlation matrix.

For step 2, the Latin Hypercube Sampling (LHS) technique[22] is used for initial sample generation to ensure that the sample space is decently covered. The Iman Conover rank correlation method[23] is subsequentially applied to induce the correlation matrix calculated from the sample parameter estimation. The final sampling step is through the inverted probability distribution to real values (*icdf* function in MATLAB) so that the generated samples follow the same distribution as obtained from the parameter estimation. For this work, we will generate 500 parameter sets for further analysis.
Step 3 repeatedly runs the dynamic simulation utilizing the parameter sets generated from the samplings in step 2. This results in 500 model predictions for each time step.
Step 4 revolves around analyzing the 500 model predictions, such as generating the mean and standard deviations of the model output at each time step and assessing the overall uncertainty propagation.

5.4.3 Implementation

The model implementation, the simulations, and the abovementioned statistical methods are performed in Matlab R2021B (The Mathworks, Natick, Massachusetts). The model equations (ODE and algebraic) were solved using a stiff-solver (ode15s with integration accuracy set to 1.0E-07), while parameter estimation was performed using the *lsqnonlin* algorithm available in Matlab. Model inputs and outputs are summarized in Tables 5.2 and 5.3.

Table 5.2: Summary of biochemical model inputs

Variable	Description	Unit
F	Substrate Feed Rate	$\frac{kg}{h}$
C_f	Concentration of substrate in feed	$\frac{g}{kg}$
$P_{agitator}$	Energy dissipation from agitator	W
Q	Aeration rate	$\frac{L}{hr}$
T_{Air}	Process Air Temperature	K
P	Air pressure	bar_g
ϕ	Relative Humidity	$fraction$

Table 5.3: Summary of biochemical model outputs

Variable	Description	Unit
X	Viable Biomass concentration	$\frac{g}{kg}$
S	Substrate concentration	$\frac{g}{kg}$
FA	Main Product concentration	$\frac{g}{kg}$
X_{TDW}	Total Dry Weight	$\frac{g}{kg}$
O	Dissolved Oxygen concentration	$\frac{g}{kg}$
pH	Broth pH levels	
W	Broth Weight	kg
CER	Carbon Dioxide Evolution Rate	$\frac{mol}{hr}$
OUR	Oxygen uptake rate	$\frac{mol}{hr}$

5.5 Results

5.5.1 Model fit and parameter estimation

Experimental data collected from sampling industrial production are used for parameter estimation. Before any additional analysis, we want to ensure that the mechanistic description provided so far can adequately model the fermentation process. The initial fit can be seen in Figure 5.2. This fit is estimated by minimizing the error via a single use of the *lsqnonlin* before any bootstrap analysis. The prior estimation showed a generally satisfactory fit with experimental data and can accurately describe the concentration profiles of viable biomass and main product. It also captured the concentration of the primary carbon source reasonably.

The overall prediction accuracy is displayed in Table 5.4. Model quality was assessed with the root mean sum of squared errors (RMSSE) but is reported here as a scaled percentage deviation from the experimental measured mean value.

$$RMSSE(\%) = 100 * \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{N}\left(y_{meas,i} - \hat{y}_i\right)^2}}{\bar{y}} \tag{5.19}$$

Initial parameter estimation shows that the Model has the most trouble predicting main carbon source concentrations. If all measurements are included, there is an error of approximately 20%. It's hypothesized that the complex media blend gives rise to multiple available carbon sources. At the same time, the Model focuses only on a singular primary carbon source. This may lead to different consumption rates depending on the variation

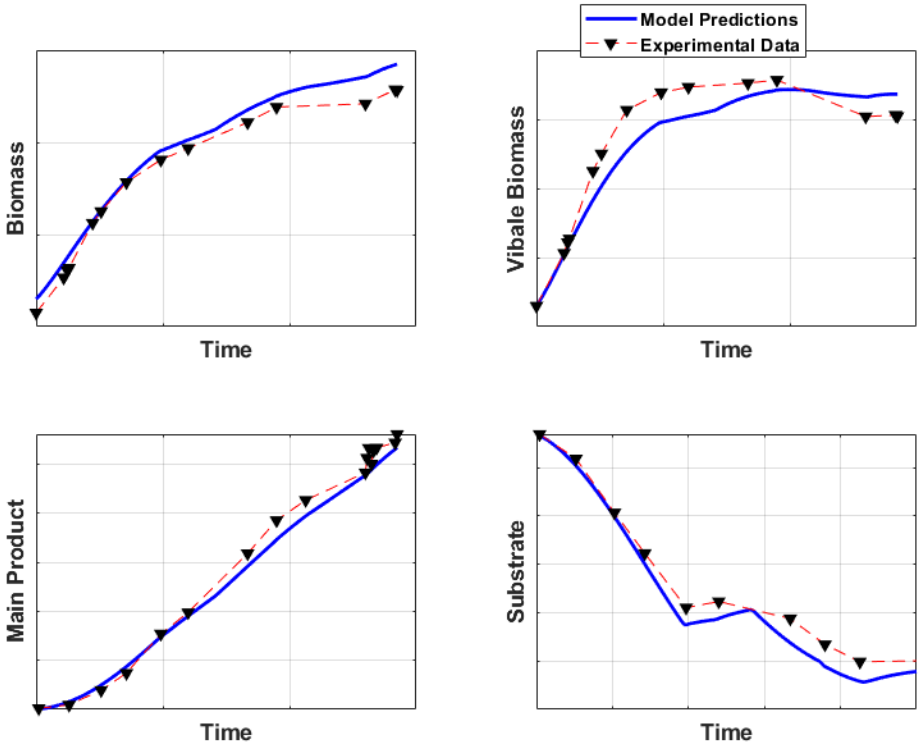

Figure 5.2: Model fits industrial fermentation data when the Model is simulated on the batch left out for validation

Table 5.4: Model evaluation quality for each process variable in the experimental dataset

Process Variable	RMSSE (%)	R^2
CDW	12.75	0.98
Viable Biomass	12.99	0.91
Main Product	6.60	0.99
Substrate	19.52	0.98

in the initial media blending. Both biomass measurements show a similar error of approximately 13%. The R^2 for both biomass measurements exceeded 0.9, indicating a strong correlation between model predictions and measured biomass. Similarly, with the substrate predictions, it may be possible to improve fits by considering the effects of potential alternative carbon sources, but this is a highly complicated data collection and modeling process.

However, for engineering purposes, we are primarily interested in the main product concentration, which is the main economic driver of the entire process. Therefore, the comparison of measured and simulated growth profiles of all sampled batches is shown in

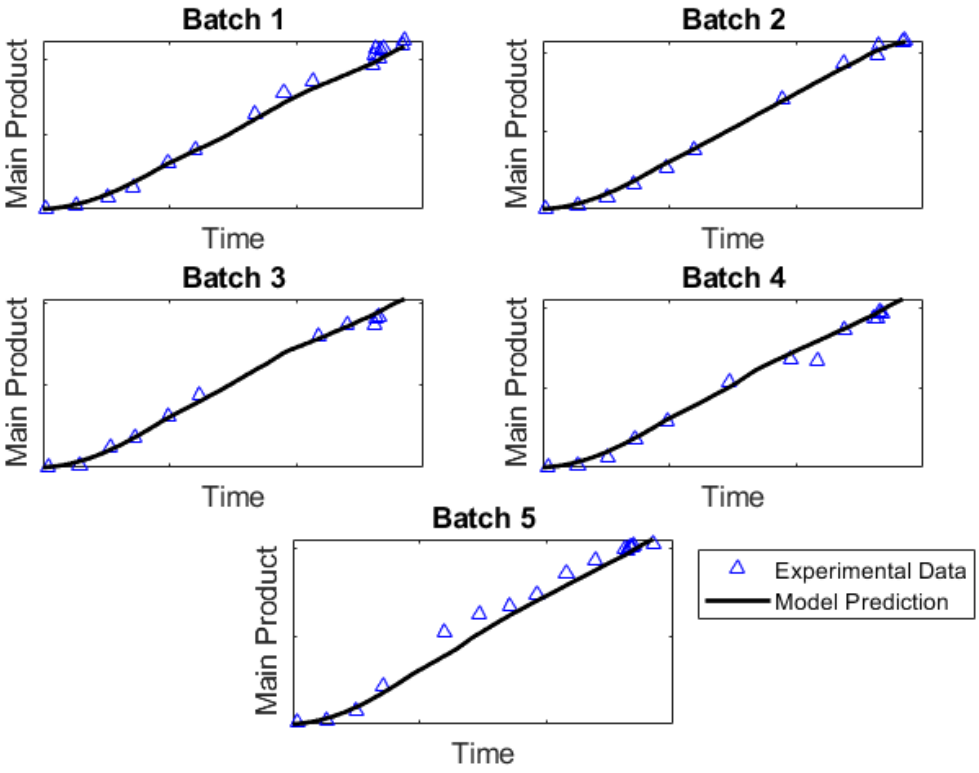

Figure 5.3: Model fits industrial fermentation data when the Model is simulated on the batch left out for validation

Figure 5.3. The model simulation profile follows the measured trend over the entire batch duration. The overall fit to the main product considering all experimental values is depicted in figure 5.4 with an overall prediction error of 6.6%. This is a significant improvement over the data-driven modeling methods utilized in chapter 4, which had the lowest prediction error of 11.21% and vastly exceeded the expected prediction accuracy required for further downstream processing.

This is a fit using 11 parameters. Parameter estimation using the bootstrap methodology is performed to assess model reliability. The parametric mean values and the standard deviation are estimated as well as the covariance and correlation matrices between them. It is considered that the estimated parameters depend on the nominal parameter values obtained through a single use of *lsqnonlin*, model structure, and cultivation conditions. Bootstrap parameter estimation results are shown in Table 5.5, and the Distribution is visualized in figure 5.5. The mean parameter value is not shown to preserve the confidentiality of the working cell bank. Parameter uncertainty is represented as the relative

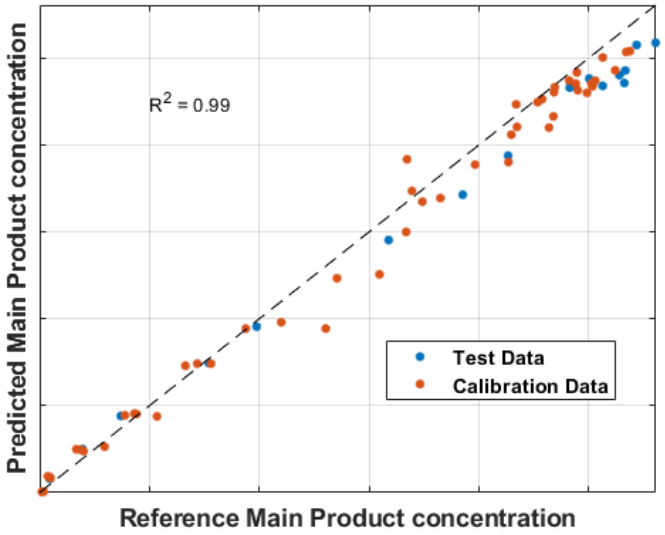

Figure 5.4: Parity plot showing Model fits with experimental data for both calibration and validation batches when predicting main product concentration

error (RE) between the standard deviation of the parameter estimate concerning the estimated mean value.

$$RE_i = \frac{\sigma_{\theta_i}}{\bar{\theta}_i} \tag{5.20}$$

None of the parameters had a relative error of over 15%, and only three were over 10%. However, not all parameters are uniquely identifiable. A correlation coefficient smaller than 0.5 is often considered a statistical threshold for a parameter to be uniquely identifiable. A significant linear dependency of parameters $\mu_{X,max}$ and K_{SX} describe biomass growth. It is common for these parameters to show linear dependency when a biochemical model is fit using raw fermentation data. The issue is usually traced to the model equation structure. Increasing $\mu_{X,max}$ will result in an increased biomass growth rate while raising K_{SX} will reduce the growth rate. Some parameters met the criterion, but since the parameter set is considered whole, the linear dependency between parameters should be considered. Thus in the following Monte Carlo simulations to estimate the effect of propagating parameter uncertainties, we will consider the correlation matrix in table 5.5.

5.5.2 Uncertainty Analysis

Propagation of uncertainty was estimated by simulating the dynamic fermentation models using 500 LHS samples in a Monte Carlo procedure. The experimental data from

Table 5.5: Estimated mechanistic model parameter relative error and correlation matrix from bootstrap parameter estimation method analysis.

Parameter	Relative Error (RE)	Correlation Matrix										
		$\mu_{X,max}$	K_{SX}	t_{lag}	$\mu_{X,max}$	K_{SP}	m_S	K_{SS}	Y_{SX}	Y_{SP}	k_d	K_{OX}
$\mu_{X,max}$	0.037	1.00	**0.80**	0.06	-0.09	-0.46	0.17	-0.40	0.29	-0.30	-0.54	0.05
K_{SX}	0.037		1.00	-0.12	0.07	-0.04	-0.20	-0.07	0.08	-0.40	-0.09	0.10
t_{lag}	0.075			1.00	0.12	-0.22	0.65	-0.15	0.56	-0.15	-0.39	-0.20
$\mu_{P,max}$	0.070				1.00	0.22	-0.02	0.14	-0.02	0.05	0.17	0.14
K_{SP}	0.126					1.00	-0.37	0.75	-0.34	-0.03	0.66	0.05
m_S	0.025						1.00	-0.07	0.79	-0.37	-0.55	-0.01
K_{SS}	0.077							1.00	-0.26	-0.26	0.63	0.19
Y_{SX}	0.091								1.00	-0.57	-0.53	-0.01
Y_{SP}	0.108									1.00	0.16	-0.15
k_d	0.080										1.00	0.07
K_{OX}	0.149											1.00

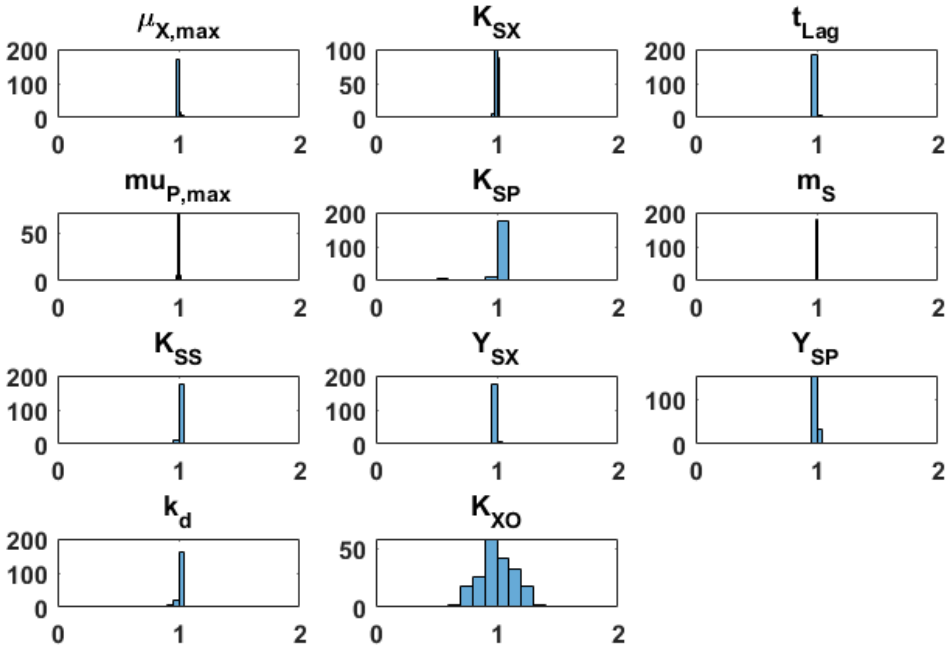

Figure 5.5: Distribution of model parameters plotted as relative to their mean value.

the first batch in the dataset (i.e., the validation batch) is used as a reference. The raw simulation results from the Monte Carlo procedure are plotted in Figure 5.6 along with the experimental measurements. Looking at the viable biomass profile, we can see that most simulations show the typical fermentation curves, consisting of exponential, stationary, and death phases. This is indicated by the mean and confidence bands. Each simulation uses a different set of parameter values. Some simulations show odd behavior of increased biomass growth after the death phase. Overall the uncertainty band for viable biomass measurement is the largest; this can be traced back to the overall large parameter distribution of parameters $\mu_{X,max}$ and K_{SX}. The uncertainty band for the main product is significantly lower than all other process variables, indicating a very reliable model response when predicting the main product.

Despite the wide uncertainty bands for viable biomass, the uncertainty bounds for CDW and Main product concentrations are much narrower. This is likely the result of a narrow uncertainty in the $\mu_{P,max}$ parameter, which is the main driver for product growth.

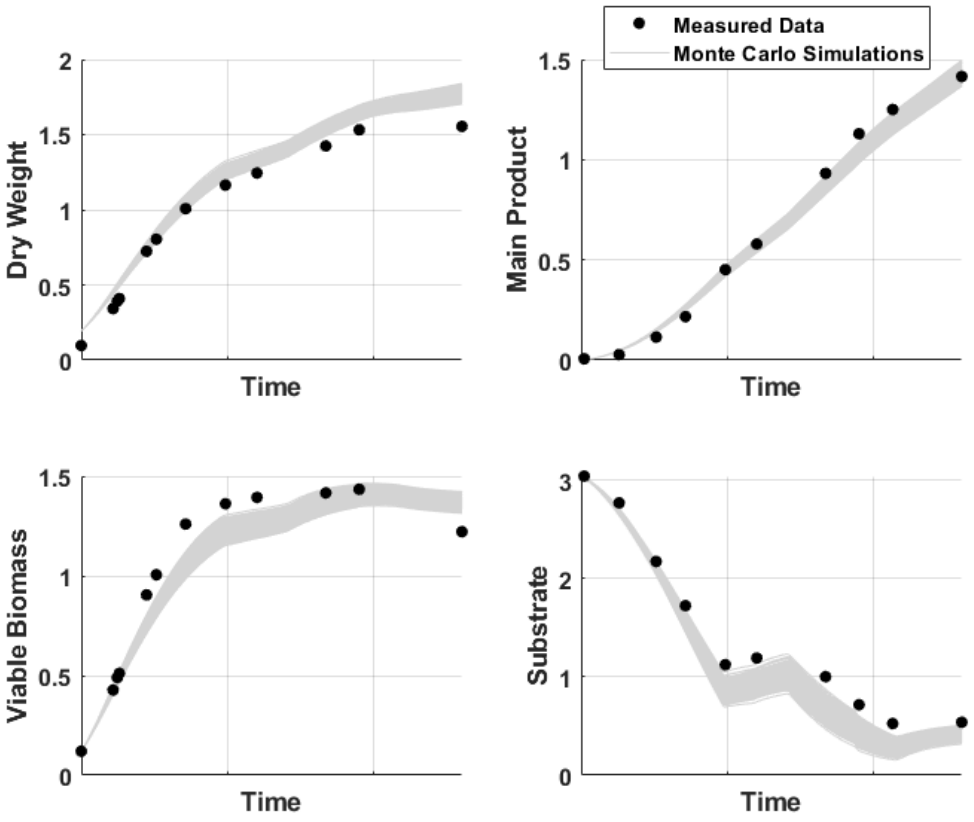

Figure 5.6: 500 Monte Carlo simulations using parameter distribution and correlation determined via bootstrap analysis.

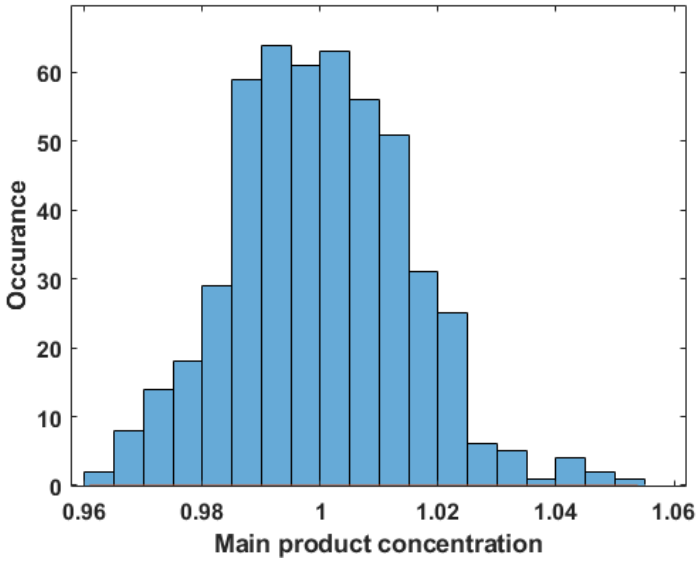

Figure 5.7: Distribution of main product concentration from the Monte Carlo analysis at end-of-batch. Plotted as relative to the mean value of all Monte Carlo simulation outputs

5.6 Discussion

Despite the model not being verified due to not all parameters being uniquely identifiable, it has been validated for the current industrial cultivation conditions by leaving one batch out for testing purposes. Thus it can still be of use for a wide array of applications. The primary performance metric of the process is the final concentration of the main product when the batch is harvested. The previous section showcased low uncertainty propagation due to parametric uncertainties, especially for the main product. Consider uncertainty in model predictions during harvest depicted in figure 5.7. We can see that most simulations harvest prediction clusters within 2% of the experimentally measured mean value.

Previous work on this process yielded a multivariate model that can estimate harvest yields with an error margin of 11.21%, detailed in chapter 4. The mechanistic model can predict with even less error, and the uncertainty in model parameters shows that the model predictions are very reliable. Even considering a worst-case scenario in the uncertainty distribution, the mechanistic model is still expected to be more accurate than a multivariate model. Since the confidence interval is so narrow for both parameter and model output uncertainties, it is likely unnecessary to hunt for identifiable parameter subsets as the model is reliable enough for engineering applications. Furthermore, all essential batch states can be modeled using readily available sensor readings. Thus the model could be

used with a state estimator for process monitoring and control if we were only concerned with the main product yields.

However, the work still has some missing elements if it should be used as the basis of a digital twin. One of the main drawbacks of using the current iteration for batch planning or as the basis for digital twins is that the model cannot run as an independent simulation. This is because the model relies on OUR and CER data to estimate evaporation accurately. Without this information, any optimization that relies on model outputs will probably underestimate final product concentrations while overestimating the current broth weight. This will provide a risk of suggesting feed strategies that will underfill the tank. Furthermore, and this is related to the scope of this research, the model does not predict batch quality concerning related substances. All these missing variables stem from a lack of scientific knowledge required to develop mechanistic descriptions that can accurately describe their rate of change.

In the current state, the mechanistic model uses batch measurements to calculate estimates of key process variables that describe the batch states that are not observable. Furthermore, since parametric uncertainties and correlations matrix have been identified, this model could be used as a probabilistic model-based soft sensor for monitoring[24]. It would be interesting to test the applicability of the current iteration as a soft sensor by investigating the sensitivity to faulty data. The groundwork laid in this chapter will be crucial for further model developments, especially if we consider hybrid models to account for the unknown variables.

5.7 Conclusions

A mechanistic model structure was proposed to describe an antibiotic production process of a novel filamentous fungi strain. The model can be used to simulate growth profiles on an industrial scale. It can simulate biomass growth, main product synthesis, and substrate uptake. Due to the lack of sophisticated hardware sensor technology, these key process variables are not commonly measured *on-line*. The predictions are supported by readily available *on-line* measurements such as offgas mass spectrometry to evaluate carbon evolution and oxygen uptake. The model was built utilizing experimental data via rigorous sampling of industrial scale batches and parameter estimation techniques within

the Good modeling practices framework. Uncertainties in parameters were obtained via statistical analysis, and the propagation of uncertainties is shown via Monte Carlo analysis to have a minimal effect when predicting the main product concentrations at a large scale.

However, we cannot assign proper scientific reasoning to all the model parameters because not all parameters are uniquely identifiable. Thus, care should be used if the model needs to describe the biological characteristics of the working cell bank. Limitations of the model that prevent direct application in the current process are the lack of information regarding oxygen uptake rate and carbon evolution rate, which lead to model errors if these values are ignored. Furthermore, the model cannot simulate the accumulation of related substances, which is crucial for this process. Future work will use this model as a foundation for the integration of machine learning techniques to expand the model capabilities to obtain readily available predictions of the missing key state variables.

Bibliography

[1] Wo Godtfredsen, S Jahnsen, L Tybring, K Roholt, and H Lorck. Fusidic acid - new antibiotic. *Nature*, 193(4819):987–, 1962.

[2] R. K. Bajpai and M. Reuss. Mechanistic Model for Penicillin Production. *Journal of chemical technology and biotechnology*, 30(6):332–344, 1980.

[3] Gülnur Birol, Cenk Ündey, Satish J. Parulekar, and Ali Çinar. A morphologically structured model for penicillin production. *Biotechnology and Bioengineering*, 77(5):538–552, 2002.

[4] G. C. Paul and C. R. Thomas. A structured model for hyphal differentiation and penicillin production using penicillium chrysogenum. *Biotechnology and Bioengineering*, 51(5):558–572, 1996.

[5] Stephen Goldrick, Andrei Ştefan, David Lovett, Gary Montague, and Barry Lennox. The development of an industrial-scale fed-batch fermentation simulation. *Journal of Biotechnology*, 193:70–82, 2015.

[6] Krist Gernaey, Anna Eliasson Lantz, Pär Tufvesson, John Woodley, and Gürkan Sin. Application of mechanistic models to fermentation and biocatalysis for next-generation processes. *Trends in Biotechnology*, 28(7):346–354, 2010.

[7] S J Pirt and R C Righelato. Effect of growth rate on the synthesis of penicillin by penicillium chrysogenum in batch and chemostat cultures. *Applied Microbiology*, 15(6):1284–1290, 1967.

[8] Gürkan Sin, Krist Gernaey, and Anna Eliasson Lantz. Good modeling practice for pat applications: Propagation of input uncertainty and sensitivity analysis. *Biotechnology Progress*, 25(4):1043–1053, 2009.

[9] Gürkan Sin, Anne S. Meyer, and Krist Gernaey. Assessing reliability of cellulose hydrolysis models to support biofuel process design – identifiability and uncertainty analysis. *Computers and Chemical Engineering*, 34(9):1385–1392, 2010.

[10] R. Sawyer and E. J. Dixon. The automatic determination of original gravity of beer. part ii. the determination of alcohol and gravity lost. *Analyst*, 93(1111):680–687, 1968.

[11] WV Dahene, S Jahnsen, I Kirk, R Larsen, and H Lorck. *Fusidic acid: Properties, biosynthesis, and fermentation*. Biotechnology of industrial antibiotics, 1984.

[12] Rubayyi T. Alqahtani, Mark I. Nelson, and Annette L. Worthy. A biological treatment of industrial wastewaters: Contois kinetics. *Anziam Journal*, 56(4):397–415, 2015.

[13] Gürkan Sin, Peter Ödman, Nanna Petersen, Anna Eliasson Lantz, and Krist Gernaey. Matrix notation for efficient development of first-principles models within pat applications: Integrated modeling of antibiotic production with streptomyces coelicolor. *Biotechnology and Bioengineering (print)*, 101(1):153–171, 2008.

[14] S.J. Pirt. *Principles of microbe and cell cultivation*. Blackwell„ 1975.

[15] DB Kell and B Sonnleitner. Gmp - good modelling practice: An essential component of good manufacturing practice. *Trends in Biotechnology*, 13(11):481–492, 1995.

[16] Pauline M. Doran. *Bioprocess engineering principles*. Academic Press, 1995.

[17] O. C. Bridgeman and E. W. Aldrich. Vapor pressure tables for water. *Journal of Heat Transfer*, 86(2):279–286, 1964.

[18] John Villadsen, Jens Nielsen, and Gunnar Lidén. *Bioreaction engineering principles*. Springer, 2011.

[19] JA Roels and JJ Heijnen. Power dissipation and heat-production in bubble-columns - approach based on non-equilibrium thermodynamics. *Biotechnology and Bioengineering*, 22(11):2399–2404, 1980.

[20] B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1):1–26, 1979.

[21] Jerome Frutiger, Jesper Graa Andreasen, Wei Liu, Hartmut Spliethoff, Fredrik Haglind, Jens Abildskov, and Gürkan Sin. Working fluid selection for organic rankine cycles - impact of uncertainty of fluid properties. *Energy*, 109:987–997, 2016.

[22] Wei Liem Loh. On latin hypercube sampling. *Annals of Statistics*, 24(5):2058–2080, 1996.

[23] Ronald L. Iman and W. J. Conover. A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics - Simulation and Computation*, 11(3):311–334, 1982.

[24] Robert Spann, Christophe Roca, David Kold, Anna Eliasson Lantz, Krist V. Gernaey, and Gürkan Sin. A probabilistic model-based soft sensor to monitor lactic acid bacteria fermentations. *Biochemical Engineering Journal*, 135:49–60, 2018.

6 Hybrid Modelling for fermentation batch quality

Abstract

The capabilities of a mechanistic model built to analyze the industrial-scale fed-batch fermentation for the production of Fusidic Acid is extended by incorporating data-driven models. Key missing state variables that prevented the implementation of the model as a state-of-the-art simulator were identified. Two Artificial Neural Networks (ANN) were integrated into the model block flow with the role of underlying learning functions that describe the missing biochemical kinetics. The result is a hybrid model that can accurately predict the concentration profiles of a problematic byproduct with a relative error of 22%. Furthermore, it can also measure carbon dioxide evolution rate (CER), Oxygen uptake rate (OUR), and changes in pH. This model can be exploited to explore new production strategies while taking into account not only the productivity of a batch but also the quality of the batch. An example application is showcased where, with more conservative nutrient-feeding strategies, it is possible to eliminate the byproduct at the fed-batch stage at the cost of a slightly reduced main product harvest. On the other hand, we also show the potential effects w.r.t. batch quality of more aggressive feeding strategies to increase fed-batch throughput.

6.1 Introduction

The biotechnological industry heavily relies on mathematical models as a core component of the business model and manufacturing economics. The gold standard is a mechanistic model. These models are preferred as they contain relevant scientific knowledge to describe the system's behavior. Mathematical models are essential elements of the biopharmaceutical industry for process optimization and intensification. However, biological systems are notoriously complex, and developing such a model can be a massive time and resource investment. Otherwise, they will have difficulty accurately describing the process. To simplify things, first principles models most often utilize an empirical approach to describe the growth of the cell culture. The list of appropriate equations to choose from is large[1], and that only accounts for cell growth; further modeling is needed to account for the production of products, substrate uptakes, gas evolution, and more.

The production of pharmaceuticals has always had a strong focus on quality. A common method for producing Active Pharmaceutical Ingredients (API) is via biological processes or fermentation of high-producing bacteria or fungal strains. These processes have been subject to mechanistic modeling for use in batch optimization, monitoring, and control[2]. However, the metabolic pathways responsible for the main product can lead to the accumulation of related substances and other impurities in the batch that hampers the final product quality and may be impossible to remove in the downstream process[3]. Mechanistic models rarely consider batch quality, and even the most state-of-the-art models are focused on the main product only[4]. There is a strict definition of allowed quantities of defined impurities when a product is used as a pharmaceutical[5]. There is little value in utilizing a mathematical model for process optimization if there is a risk that the model suggested optimal conditions further promote the accumulation of impurities. A high-yield process is worthless if the product can't be sold because it fails quality checks.

Several key drawbacks were highlighted previously in the mechanistic model presented in chapter 5 that prevent its full utilization as a digital twin. Especially if we consider the purpose and scope of such a model. None of the equation structures predict oxygen consumption rates or carbon dioxide evolution rates. Consequently, the model relies on batch measurements obtained from *online* sensors as a replacement. However, this information must be estimated to simulate a batch in a new scenario fully. Otherwise, the mass

balance component will lead to errors due to overestimating the broth weight. Unfortunately, due to the complex broth mixture, the stoichiometry of the biochemical reaction is unknown and is also expected to change throughout the fermentation. Thus a rate model of carbon evolution based on stoichiometric balance is not accessible. Furthermore, the empirical model suggested by Birol et al.[6] did not work for this particular process. Modeling CO_2 production in the same way as the main product and O_2 as another substrate also did not give model structures that fit the profile.

Furthermore, another major component that is especially relevant for the pharmaceutical industry that forms this study's scope is the concentration of a related substance. This compound is integral to the batch's final quality, and strict requirements must be met; otherwise, the batch is considered a failure. No mechanistic description can accurately explain the growth and decay kinetics.

Due to the complexity of biological systems, a hybrid model approach can be a potential alternative. The concept of hybrid modeling in this context combines a first principles mechanistic model and machine learning models into a single model. They've seen an emergence in research in process modeling for Industry 4.0[7] due to increased computational power and the amount of available data. The idea is that a mechanistic model is built that incorporates all the obtained scientific knowledge and integrated machine learning algorithms to predict underlying functions of phenomena that are poorly understood or too complex[8]. Machine learning and Artificial Intelligence algorithms such as Artificial Neural Networks (ANN) have seen an increase in popularity in various research fields, and the use of Hybrid modeling has seen success in chemical engineering, such as in particle processes[9][10][11].

This study focuses on applying a hybrid modeling framework by integrating data-driven or black-box models to support the mechanistic model previously developed for the fermentation of filamentous fungi that produce Fusidic Acid.

6.2 Materials and Methods

6.2.1 Dataset description

The industrial scale dataset used to calibrate the kinetic model for this study is re-used. The dataset already contains information regarding related substances. Refer to chapter

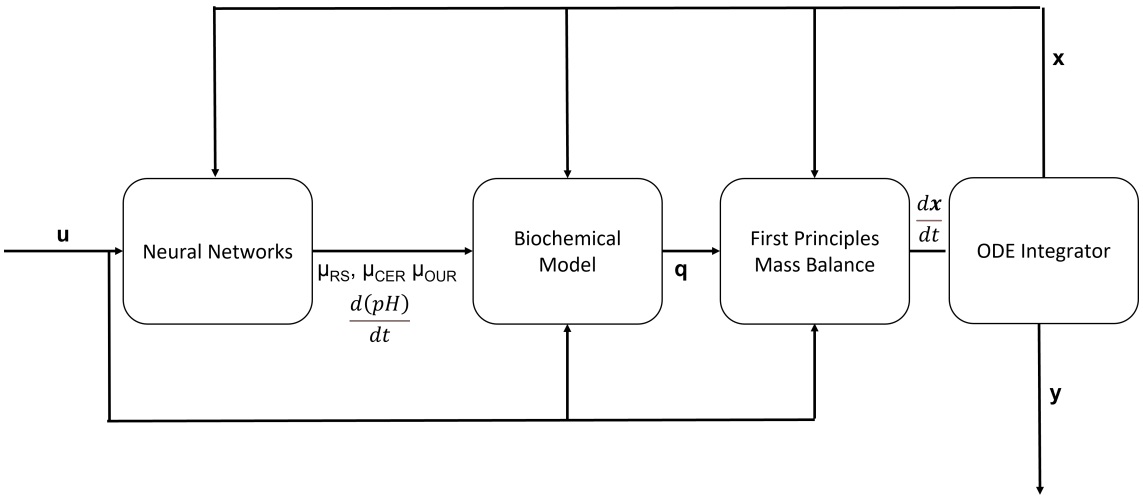

Figure 6.1: Proposed Hybrid model structure after integrating Neural Networks before remaining biochemical kinetic expressions and conservation equations.

5 of this thesis for more detailed information regarding experimental data.

6.2.2 Model Structure

For this particular study, a serial structure configuration is presented in Figure 6.1. This is because we already have a decent mechanistic structure representing mass balance around the fermenter. The kinetic model also explains a variety of crucial process variables, such as viable biomass growth, main product synthesis, and substrate consumption. The data-driven model is then used to account for the part of the phenomenon in which there is no available model.

In this case, the hybrid model can be considered an extension of an existing mechanistic model and is learning the functional relationship between the current batch state and the instant rate of change. A simplified version of how the dynamic model estimates the current state, which forms the basis of the functional relationships the hybrid should learn.

$$\mathbf{x}_{i+1} = \mathbf{x}_i + f(\mathbf{x_i}, \mathbf{u_i}, \theta) \tag{6.1}$$

Where the i denotes a specific time point, we bring up this formulation to emphasize that data-driven models are used to estimate kinetic rates that generally are not directly measurable.

6.2.3 Incorporating Data-Driven model elements

The mechanistic model already describes the majority of the kinetics. However, four process variables currently lack adequate mechanistic descriptions for this process. These are outlined in table 6.1.

Table 6.1: List of observed variables to hybridize

State Variable	Importance	Availability
Concentration of related substances	Main indicator of Batch Quality	*Offline*
Carbon Dioxide Evolution Rate	Evaporation calculations	*Online*
Oxygen Uptake Rate	Evaporation calculations	*Online*
pH	Precipitation of main product	*Online*

Using data-driven models like Neural networks can present a challenge because there are no limits on the output. In many ways, we can improve the reliability of neural networks by incorporating as much scientific knowledge as possible when deciding the overall equation structure. We can improve hybrid models' reliability by carefully considering the "what should the data-driven model predict?"

A good start is looking at three of the four state variables related to biological activity.

$$\mathbf{q} = \mu X_{Viable} \tag{6.2}$$

Where μ is the specific growth or consumption rate of a given component. We can ensure that the rate of change of these components is tied to biological activity by having the data-driven models predict μ as a function of the current batch state variables **x** and input variables **u** rather than the overall rate of change. This inclusion of scientific knowledge and assumptions provides an important example of why hybrid models are preferred over purely data-driven models. In this case, the growth and consumption of specific components are directly tied to biological activity since all activity goes to 0 when X_{Viable} is 0. This will not necessarily be the case if the data-driven models predict the biochemical rate of change **q**.

Simply predicting μ is a good start but isn't perfect since μ is allowed to be negative or positive. This is the desired effect when predicting related substances, but there could be scenarios where a data-driven model predicts the uptake of carbon dioxide and release

of oxygen. For these two cases, it can be fixed with the following

$$\mu_{CER} = max(0, f_{CER}(\mathbf{x}, \mathbf{u})) \tag{6.3}$$

$$\mu_{OUR} = min(0, f_{OUR}(\mathbf{x}, \mathbf{u})) \tag{6.4}$$

Where f is the function describing neural network outputs, note that we are not putting the same restriction on the specific related substance rate μ_{RS} because we can observe both formation and consumption in the experimental data.

The mechanism for pH is unknown. However, Ebrahimpour has successfully modeled pH dynamically in a bioprocess for cream cheese production using a black-box long short-term memory model[12]. For this work, we will rely on a more traditional shallow neural network to predict changes in pH using batch state. It is believed that pH changes correlate to certain biological activities when batch states are correct. Due to the broth's complexity, we could not determine any special rules regarding the change in pH values. It was observed to both increase and decreased over the process, and thus the pH is modeled simply as

$$\frac{d(pH)}{dt} = f_{pH}(\mathbf{x}, \mathbf{u}) \tag{6.5}$$

6.2.4 Neural Networks

While several data-driven different modeling techniques can be incorporated into a mechanistic description, the most commonly used is the Artificial Neural Network (ANN). Because they are universal approximators, they can model any linear and non-linear behavior and require no structural knowledge of the modeled system[13].

ANNs are loosely based on how the human brain processes information. They consist of a series of data processing units called neurons connected by information flows. ANN performs non-linear operations to the information flowing through them so that a problem becomes linearly separable in the final feature space. Each neuron computes the weighted sum of all signals received from its connection from a previous layer plus a bias term. This weighted sum is then used as an input in an activation function to generate

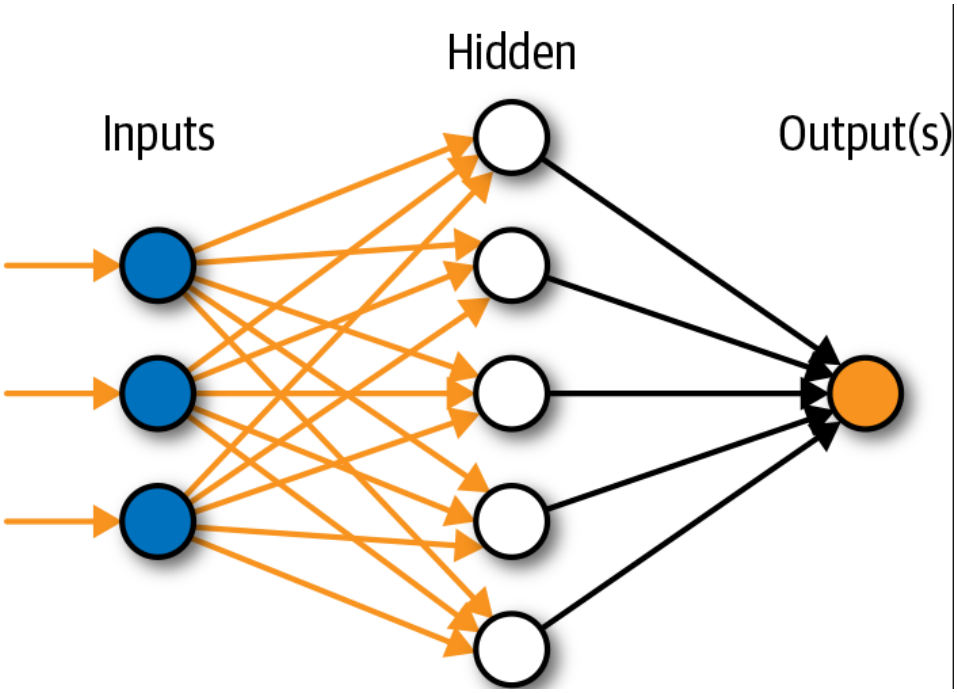

Figure 6.2: Schematic representation of a Feed-Forward ANN with one hidden layer

layer output. The transformation between layers can be summed with

$$\mathbf{a} = f(\mathbf{z}) = f(\mathbf{W}\mathbf{x} + \mathbf{b}) \tag{6.6}$$

Where **a** is referred to as an activation, $f(.)$ is an activation, **W** is a matrix containing neuron weights, and **b** has the neuron biases of the current layer and **x** is the input to the network. The activation **a** can be fed-forward into more hidden layers, ANNs with more than one hidden layer are often called Deep Neural Networks (DNN). There are multiple-choice activation functions, but most ANNs focus on a linear, hyperbolic tangent, sigmoid, or rectified linear units. A single-layer ANN can be sufficient for regression problems, and it's been proven that given enough neurons, they can fit any input-output relationship given. But, of course, more neurons mean more parameters making more complex models and increasing the chance of overfitting.

6.3 Model development

Since the mechanistic model has already been adequately identified, the overall hybrid model identification follows the direct approach[14]. This workflow starts with identifying a mechanistic model without considering a data-driven model and, subsequently, identifying

the data-driven model. Since the development and calibration of the kinetic model are already finished and detailed in the previous chapter, the focus here is the implementation and training of the neural networks.

6.3.1 ANN Development and Training

The ANN model structure used in this study will be restricted to a shallow neural network, i.e., only one hidden layer. These network structures are often sufficient for regression problems. Furthermore, shallow ANNs rely on fewer parameters than their DNN counterparts which makes them computationally easier to train as they require fewer model evaluations to calculate gradients. Shallow neural networks with fewer nodes are also likely to be more robust, as they don't have the same opportunity to overemphasize the effects of measurement noise and errors.

For this work, we will use the hyperbolic tangent function for the activation when through the hidden layer. The nodes in the input and output layers were chosen to have linear transfer functions. Training ANN for regression is done by adjusting network bias and weights to minimize a loss function. This is usually achieved with various gradient-based algorithms. This work will optimize neural network weights and biases using the Levenberg-Marquardt backpropagation algorithm[15]. For deciding on the best network structure and identification of network parameters, the dataset is divided into three partitions, a training, validation, and test partition. For deciding the test set, it was decided to leave the entire first batch in the dataset out, designated as the test data. This is the same test data omitted during the mechanistic model calibration work featured in chapter 5. The remaining data was partitioned randomly into training and validation sets. The purpose of the validation set is to provide an early stopping criterion to prevent overfitting. Different numbers of nodes in the hidden layer were considered and are considered a hyperparameter to optimize. Different network structures were compared based on their performance on the loss function value when considering the test set only. The training environment is summarized in Table 6.2.

We can use a single neural network with multiple outputs as the data-driven model block. However, for convenience, we will develop two separate ANN models labeled as the byproduct model and *online* model. The byproduct model is made for predicting related

Table 6.2: Summary of the ANN training environment

Description	Setting
Training Algorithm	Levenberg-Marquardt backpropagation
Activation function in hidden layer	Hyperbolic tangent
Activation function in output layer	Linear
Loss Function	Mean Squared Error
Maximum number of epochs	1000
Maximum validation failure	6
Training Ratio	0.7
Validation Ratio	0.3
Data partition method	Random split

substance accumulation only, and the *online* model is used to predict state variables captured via *online* sensors, which are oxygen uptake, carbon evolution, and pH changes. We develop two different models because phenomena measured using *online* via frequent measurement, the kinetic terms can be estimated directly from the process data. However, in cases where kinetic terms are not readily available, a particular pre-training phase is adopted for otherwise unnecessary model training.

6.3.2 Byproduct model

One of the significant difficulties in training models is predicting the rate of related substance accumulation. The specific growth rate μ_{RS} can not be calculated directly using the available process data. The primary reason is that the experimental data is obtained via offline analysis, a slow measurement method that has to be done manually, giving a very sparse growth profile. This makes the actual rates impossible to measure; thus, the model has to be trained to fit the final measured concentration.

$$\mathcal{L}(\theta) = \frac{\sum_{i=1}^{N}(C_{RS,pred,i} - C_{RS,measured,i})^2}{N} \tag{6.7}$$

Every time the loss function is evaluated during model training, we require evaluating the entire dynamic model. This involves solving differential equations for every loss function evaluation, which can lead to severe performance issues. The performance impact depends on the numerical solver and the model's stiffness, which depends on parameters and error tolerance specifications. In our implementation, we see that simulating a single batch over an entire fermentation period can take up to 3000 function calls. With four batches in the training set, this requires 12000 evaluations of neural networks for just a

single estimation of the loss function. This makes evaluating the gradient, even for simple models, a much more computationally expensive task.

For this, we advocate for a pre-training step in which the neural network is trained on a set of estimates which we define as $\hat{f}_{RS}$. A similar solution was proposed by Shah et al. when training DNNs for use in a hybrid setting[16]. The idea is that a model that fits through an interpolated fit between data points is already significantly better than a model using completely randomized parameters. Furthermore, any neural network configuration that can not properly predict this curve fit is also unlikely to be a suitable candidate for modeling the actual specific growth rate. The consequences are that fitting to the estimate $\hat{f}_{RS}$ as the pre-training step gives us a vastly better initialization of the neural network, drastically reducing the number of epochs required to train an optimal model. Furthermore, we can also utilize the quality of fit to $\hat{f}_{RS}$ to estimate the number of neurons required in the hidden layer for a suitable model, i.e., this pre-training phase can also be used for hyperparameter tuning.

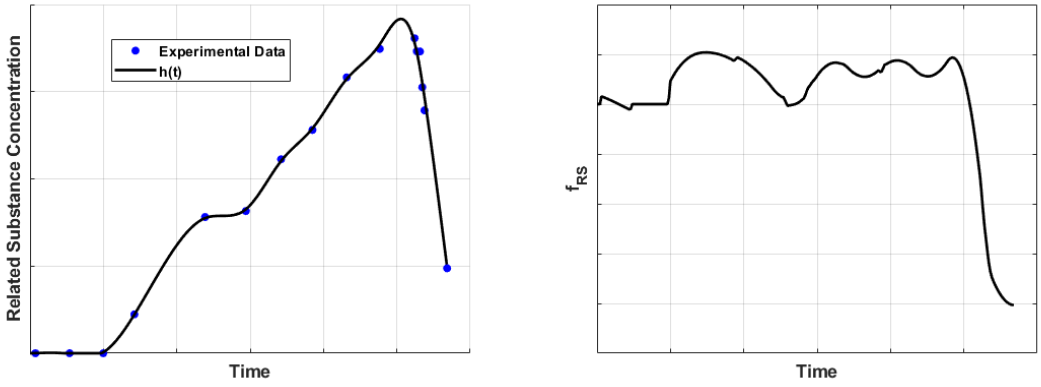

Figure 6.3: Illustration of spline curve fitting used to generate rate estimations used in ANN pre-training phase and the resulting estimate of specific growth rate profile

We define a function $h(t)$ as a curve fit between all experimental points. This study will always use a spline function as the basis for the curve fitting. Primarily because it is an easy choice and it is possible to adjust the fitting with a smoothing parameter to remove any visual artifacts. This was an accepted methodology to estimate underlying kinetics back when the hybrid was first proposed for a bioprocess [17]. This function is then used

as a temporary stand-in for the experimental data by $C_{RS} \approx h(t)$. Since $h(t)$ is a well-defined curve fit, it's easily differentiable. Thus we can define an estimate of the specific growth rate by incorporating the value in the mass balance as

$$\hat{f}_{RS} = \frac{1}{X_{Viable}} \left(\frac{d(h(t))}{dt} - \frac{E}{M} h(t) + \frac{F_{feed}}{M} h(t) \right) \tag{6.8}$$

Note that all these values are available through the previously developed kinetic model. The pre-training step is thus training a neural network where the loss function minimizes the error in $\hat{f}_{RS}$. Once a suitable model that fits $\hat{f}_{RS}$ is identified, we update the network parameter and weight biases by switching the loss function to minimize the error in experimental data using equation 6.7.

6.3.3 *Online* model

The remaining unknown phenomena μ_{CER}, μ_{OUR}, and f_{pH} are easier to develop hybrid models for. This is because all these values can be estimated reasonably well with available process data. While specific rates for oxygen uptake and carbon evolution are not directly measured, the overall rates q_{CER} and q_{OUR} are measured and available in the dataset because of the use of a mass spectrometer. Since a kinetic model has already been calibrated for the process, it is possible to simulate a batch in the dataset as a soft sensor to get estimates of X_{Viable} at every time point. The specific rates can then easily be calculated using equation 6.2.

We also wanted the model to predict change in pH according to equation 6.5. While $\frac{d(pH)}{dt}$ is not directly measured, pH is available *online*, and it is estimated frequently. Note that fermentation processes are slow, with growth parameters often expressed per hour, but pH is recorded every few seconds. Because of the frequent pH measurements, we can do numerical differentiation of the experimental pH curve, which provides a reasonable estimate for f_{pH}. Since all these process variables can be estimated well enough with available process data, updating the network weight and biases is unnecessary after the training phase. Instead, the models are trained to fit the calculated kinetics, and model quality is evaluated directly afterward.

6.3.4 Implementation

The existing mechanistic model code developed in MATLAB R2021b from chapter 5 is modified to include direct predictions from neural networks in the kinetic expressions. To reduce simulation runtime, a custom code for Feed Forward Neural Networks and the Levenberg-Marquardt training algorithm was written from scratch rather than relying on the Mathworks Neural Network Toolbox. ANN weight and biases are randomly initialized. This custom initialization utilizes only homogeneous double floating point number arrays when performing Neural Network calculations, which reduces the average simulation runtime from approximately 30 seconds to approximately 0.5 seconds on an AMD Ryzen 5 1600 CPU. The Neural Network code will be made available at GSI Research Group Github page at github.com/gsi-lab.

6.4 Results

The hybrid modeling method is used to train a model for industrial-scale production. The same batch inputs **u** that are measured are used for simulation purposes. The mechanistic model with measured process data is used to estimate relevant state variables **x** such as Viable Biomass, Substrate concentration, etc., used for training the neural networks. With the combined inputs **u** and states **x**, we estimate the specific rate of change μ_{RS} via curve fitting and numerical differentiation. With the estimated hidden function f_{RS}, the neural network is initialized by training an ANN by minimizing neural network prediction error. A single hidden layer neural network with six nodes in the hidden layer was determined to be optimal in terms of prediction error on the test data.

The best network was further trained by changing the loss function to estimate the error in actual byproduct measurements to more accurately represent the hidden rate functions. The predicted concentration of the related substance compared to the experimental data of all measured batches is illustrated in Figure 6.4. Visually the model growth profile is in excellent agreement with the collected data. Utilizing the same inputs **u** and **x**, we can estimate the rates relevant for the uptake of oxygen and evolution of carbon dioxide using measured data as well as the rate of change of pH. Since these measurements are available in *realtime* with a frequency much shorter than the process timescale, we can assume that the estimated rates are good representations of the hidden rates the neural

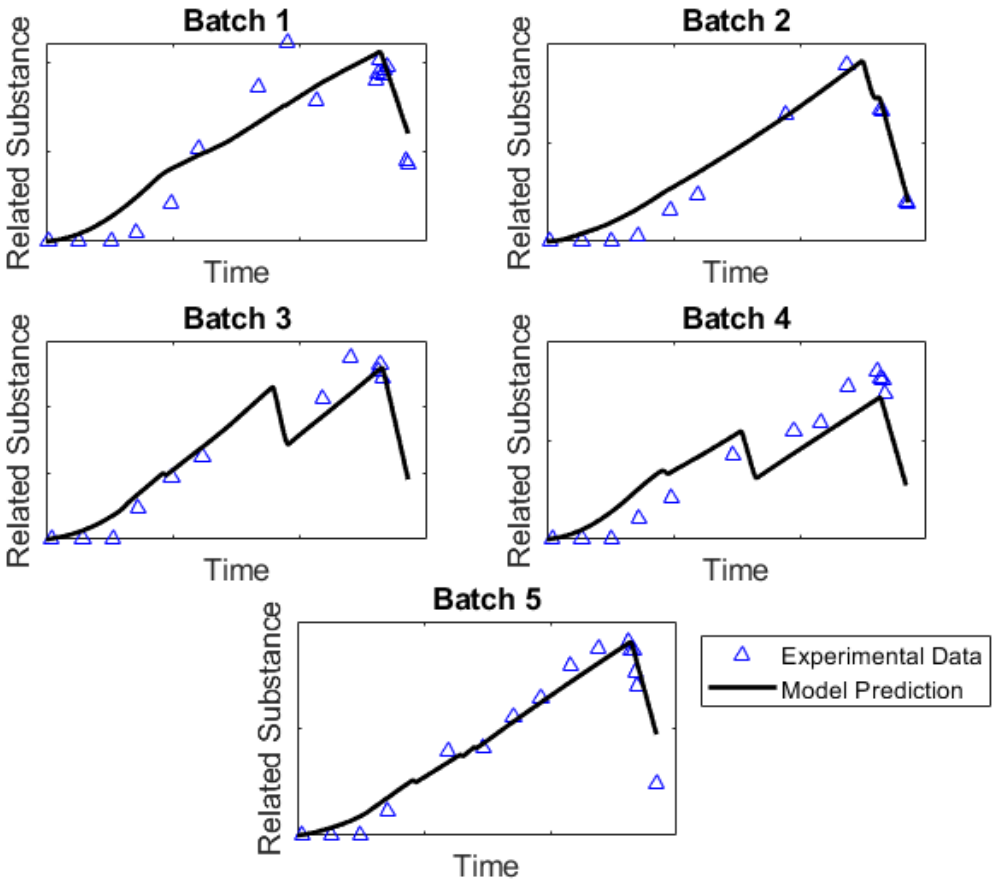

Figure 6.4: Hybrid model performance when predicting related substances across all available industrial data

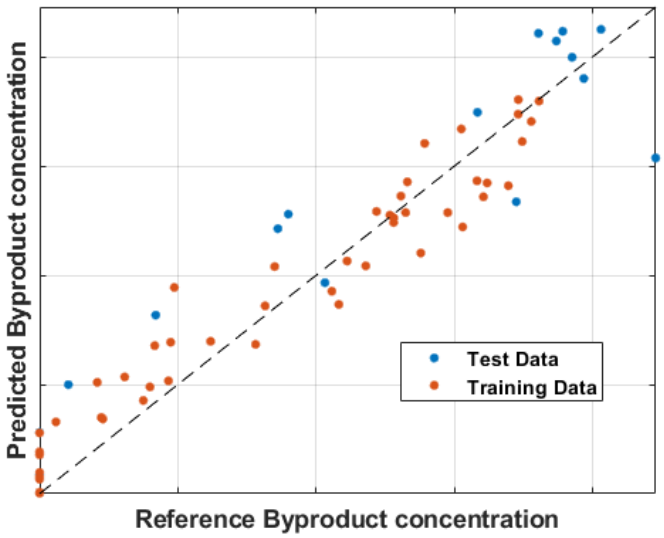

Figure 6.5: Parity-plot showing end-of-batch predictions of related substances when utilizing trained hybrid model

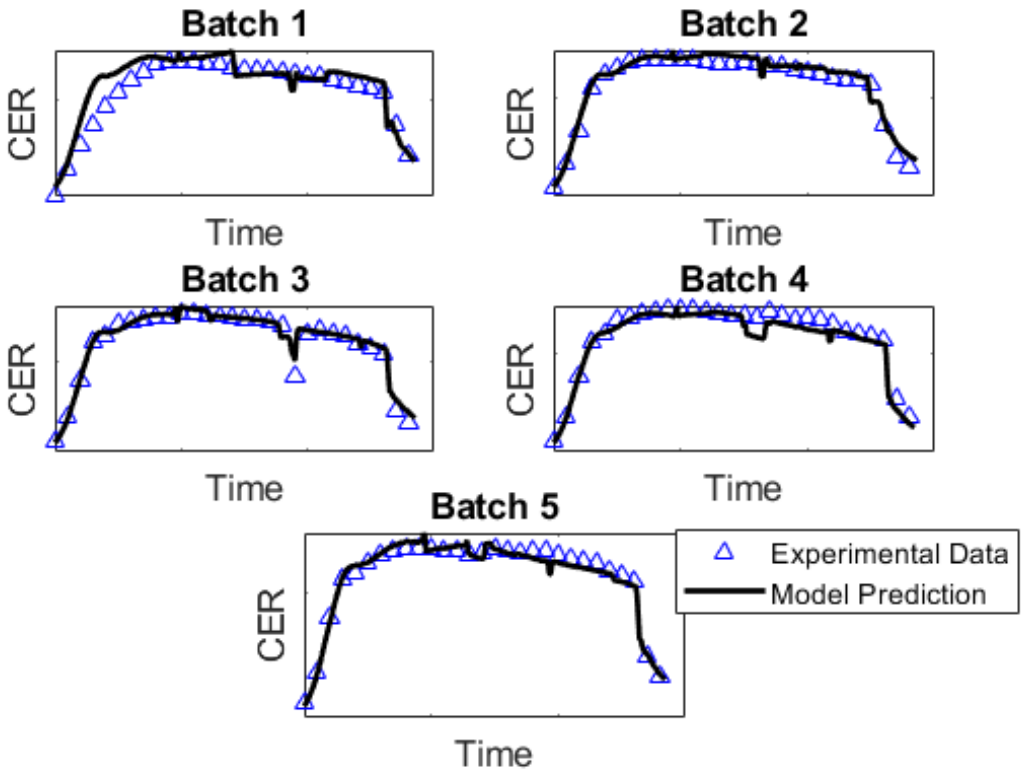

Figure 6.6: comparison of prediction from hybrid model against experimental data for Carbon Evolution Rate (**CER**)

network needs to fit. Therefore, it was determined that a single-layered neural network with 15 nodes in the hidden layer was the most optimal ANN structure for predictions. Figures 6.6 to 6.8 shows the comparison between data measured by online sensors and the hybrid model.

The overall prediction accuracy is displayed in Table 6.3. Model quality was assessed with the root mean sum of squared errors (RMSSE) but is reported here as a scaled percentage deviation from the experimental measured mean value.

$$RMSSE(\%) = 100 * \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{N}\left(y_{meas,i} - \hat{y}_i\right)^2}}{\bar{y}} \tag{6.9}$$

There is a good agreement between all measured process variables and the model predictions. The RMSSE of the related substance is 22.8% which can be alarming, but it's a significant improvement over the SCREAM method from chapter 4. Furthermore, this prediction accuracy is good enough to determine whether a batch is a failure quality-wise

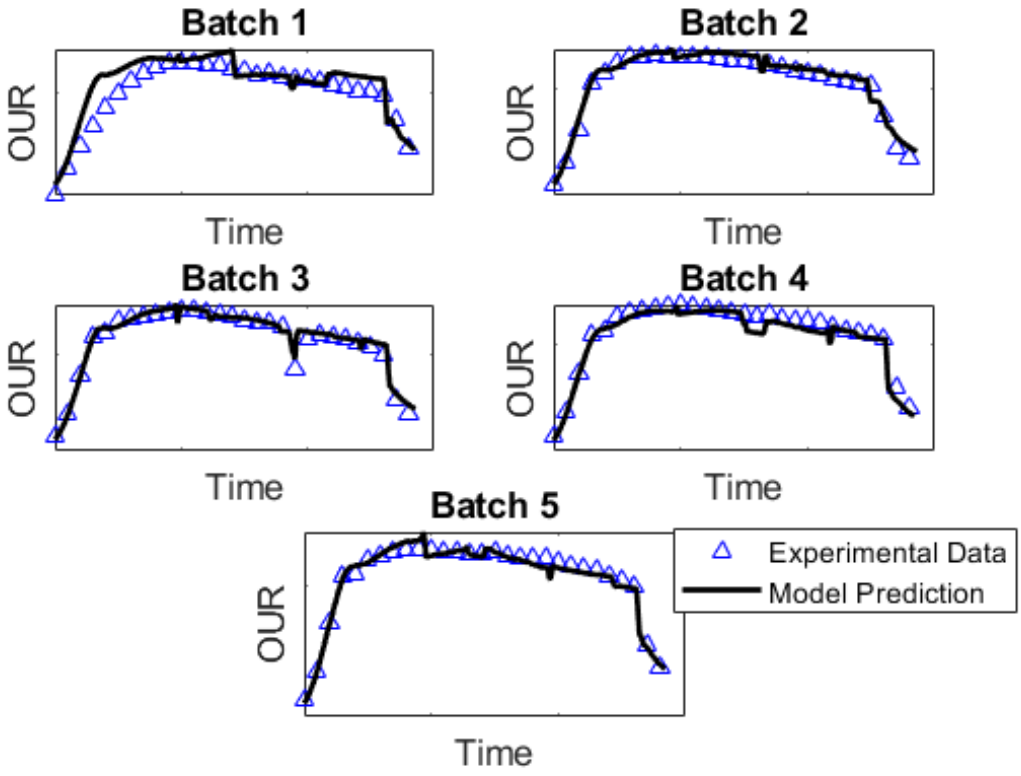

Figure 6.7: comparison of prediction from hybrid model against experimental data for Oxygen Uptake Rate (**OUR**)

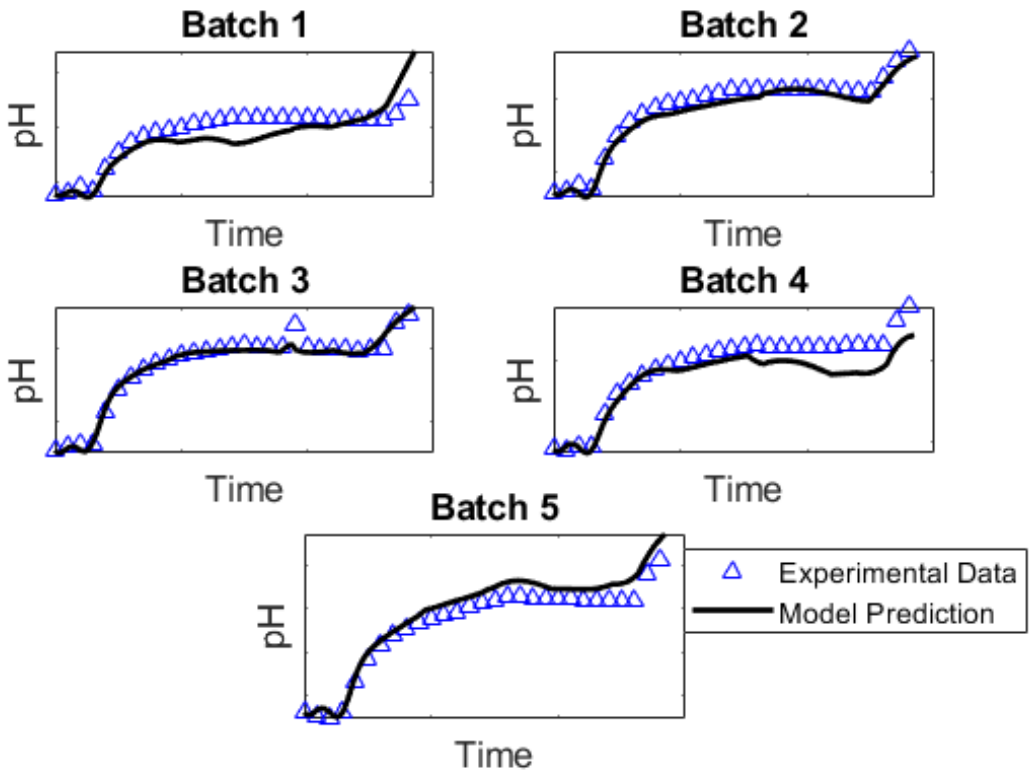

Figure 6.8: comparison of prediction from hybrid model against experimental data for **pH**

Table 6.3: Hybrid model evaluation quality for each process variable in the experimental dataset

Process Variable	RMSSE (%)	R^2
Related Substance	22.79	0.92
CER	7.55	0.94
OUR	7.47	0.94
pH	2.18	0.88

due to accumulation. It can also determine when the process should be stopped to prevent batch failure. The low prediction errors of CER and OUR mean that broth weight can be more reliably simulated without requiring off-gas measurements as model inputs. The worst case deviation when considering the prediction errors on CER and OUR is approximately 1% of the total broth weight.

Looking at the predicted and measured batch profiles, there is a measured reduction in carbon evolution and oxygen uptake at the end of the process, which the model captures. This phenomenon is also experimentally observed in the related substance profiles. It appears at the end of the batch, usually when there is no more carbon source; the biomass consumes the related substance without replenishing it. Despite the limited off-line samples taken during each batch, the model can capture this phenomenon quite well. The pH profiles are interesting, with a steady increase to a steady-state value followed by significant growth as the process reaches the end. This continued rise in pH is unique as most fermentations which is an odd phenomenon, especially considering that the main product is a weak acid. This behavior has so far been unexplained. The spike in pH at the end of the process is also interesting, and the biological explanation for this phenomenon is unknown. The hybrid model, however, can capture and model this slow rise to a steady state and even the spike at the end. The results showcase the power of machine learning algorithms to learn underlying functions of complex biological dynamics with the combination of experimental data and previously established kinetic models.

6.5 Discussions

The previous section highlights the capabilities of the hybrid model to estimate process variables with unknown mechanisms, which can approximate kinetic rates based on batch inputs and currently modeled state variables.

Looking at the predicted related substance profile in figure 6.4, there is an oddity present in Batch 3 and 4. The predicted byproduct is shown to decay rapidly sometime in the middle of the fermentation. Unfortunately, due to the sparse data collection of related substance concentration, this decay was not experimentally observed. At first glance, this may seem like a flaw in the model. However, we can see a similar phenomenon coinciding if we also consider the offgas evolution depicted in figures 6.6 and 6.7. For Batch 3, this is also observed experimentally in the offgas data. This indicates a potential process disturbance captured somewhere in the state variables **x** and batch inputs **u**, and the data-driven models take these disturbances into account in an expected way.

The hybrid model can achieve excellent results on all four missing critical process variables despite limiting the models to a single hidden layer. The recent trend in the literature is to jump to Deep Learning approaches immediately. But this study shows that shallow neural networks can result in highly reliable models even for complex biological phenomena.

The incorporation of the data-driven models in a hybrid setting allows the simulation of an entirely new batch without any reliance on process measurements. This vastly expands the scope of model applications. Before the inclusions, the model was limited to a soft sensor, but now it can be used in many ways, such as process optimization. Furthermore, including the critical process parameter of related substance concentration means that the model is now much better suited as a basis for a digital twin in the pharmaceutical industry as both batch productivity and quality are considered.

6.5.1 Quality vs Productivity

We can already explore one simple application example in terms of process optimization. The major motivation for this research is exploring how to improve batch efficiency and productivity while ensuring proper product quality via strict control of related substances. The experimental data and hybrid model indicate that related substance growth and decay are associated with the batch state, especially the available carbon source. The available substrate can be controlled by adjusting the feed in a fed-batch process. To that end, we will explore the model-predicted implications of adjusting how the batch is fed. Two dosing scenarios are suggested and compared with a reference batch from the experimental dataset. A PI controller is included in the model code to adjust the feed rate based on

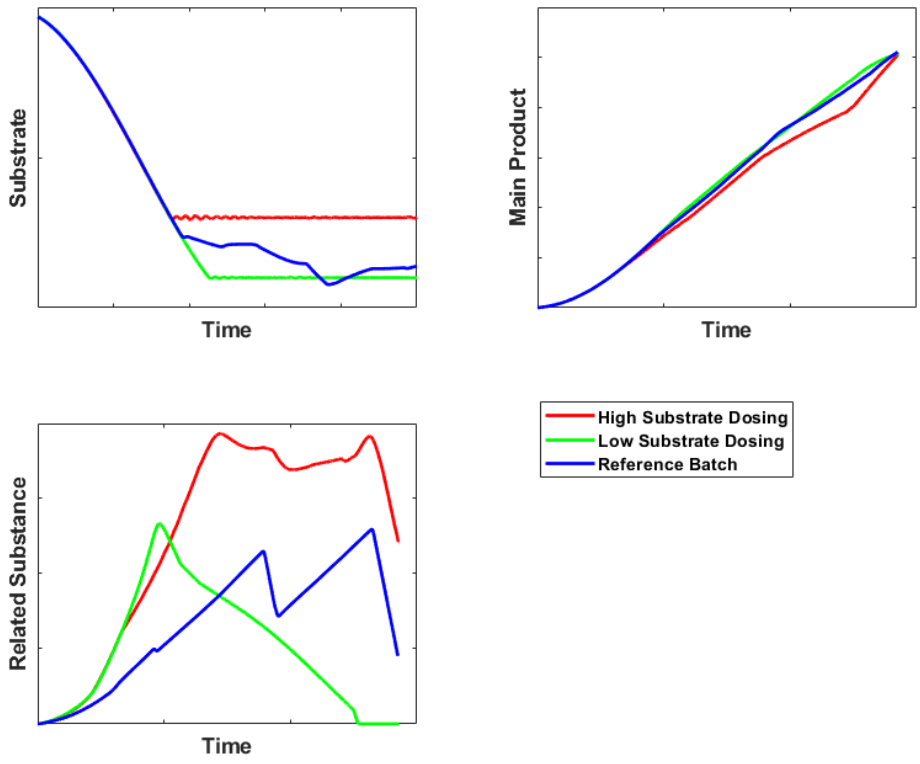

Figure 6.9: Comparison of different concentration profiles while pursuing various nutrient feed strategies

predicted substrate levels. Everything else regarding the initial batch state and process air conditions is similar.

Table 6.4: Comparison of effects of implementing two possible dosing strategies

Feed Strategy	Main Product harvest	Byproduct concentration
Reference Batch	+0 %	+0 %
Low Substrate	-17%	-100%
High Substrate	+55%	+168%

The substrate, main product, and related substance concentration profiles are depicted in figure 6.9. The overall performance of the batches compared to the reference is shown in Table 6.4. The growth profile displays concentration, but the table shows the total harvested product. All dosing strategies show similar levels of the main product concentration. However, the high dosing strategy has a much larger working volume due to more feed and thus results in a larger overall harvest. A high-dosing strategy does have a lower concentration of the main product around the middle phase of the fed batch. This is expected due to increased nutrient feed leading to a larger dilution effect. The hybrid models predict that overfeeding to increase harvest per batch can lead to a dramatic increase in related substance concentration and, thus, a much-increased risk of batch failing qualification. On the other hand, with a more conservative feeding strategy and a slight reduction in the main product harvested, it is possible to eliminate the presence of this related substance at harvest.

Due note that this is a simplistic example of a hybrid model application. So far, we have not considered the tank vessel's total working volume, which in most industrial applications should be filled. Only two scenarios were also considered, and it is unnecessary to eliminate the related substance. There would likely be a fixed constraint in the allowed final concentration, which is also dependent on the recovery process and thus decided by considering the entire process pipeline factoring in various risk management strategies. With a fixed limit of related substance and considering all other constraints, such as tank volume, it would be possible to utilize a variety of modern optimization methods such as MOSKOPT[18] to determine optimal batch operations.

6.6 Conclusions

A hybrid model was developed by incorporating Neural Networks into an existing mechanistic model to simulate an industrial fermentation of filamentous organisms producing Fusidic Acid. Combining machine learning models into a mechanistic framework allows for versatile modeling of the fed-batch process. The model capabilities are thus extended to predict the evolution of unwanted related substances and other key state variables needed to do the overall mass balance work. The byproduct, oxygen uptake, and carbon kinetics release are all estimated by feeding the current cultivation conditions as information into a shallow neural network. The pH levels are also described using a Neural Network to predict the change in pH based on the current batch state.

This modeling approach is demonstrated in an industrial-scale case study, where the model is trained and tested using experimental data obtained from an already operating industrial production. It was determined from error statistics that the models perform well predicting the missing batch states, even with limited training data and a complete lack of kinetic insights.

Model application is demonstrated via a simple example where the nutrient feed rates are adjusted, demonstrating that with different feed-rate set points, it is possible to control the level of the final byproduct concentration. With conservative feeding strategies, the model predicts that it can completely eliminate the related substance if deemed necessary. However, it is also possible to push for higher levels of the main product with more aggressive feeding strategies. Optimization algorithms can use this hybrid model to search for scenarios where the main product harvest is maximized while keeping the quality in check.

It is of interest to further expand the hybrid model by subjecting it to the Modeling framework of Sin et al. [19] to assess the reliability of the hybrid models in terms of uncertainty and sensitivity analysis. There is growing research in determining the uncertainty of Neural Networks[20], but a more efficient training strategy is required to be feasible. Uncertainty and sensitivity analysis of the Neural Network parameters will also give further insights into the robustness of the simulation during different process operating conditions. Furthermore, while the model achieved good results on the validation data obtained during standard operation, it would be relevant to run a test batch with the feed strategy

proposed by the hybrid model simulations to validate the model applicability for process optimization further.

Bibliography

[1] Mpho Muloiwa, Stephen Nyende-Byakika, and Megersa Dinka. Comparison of unstructured kinetic bacterial growth models. *South African Journal of Chemical Engineering*, 33:141–150, 2020.

[2] Krist Gernaey, Anna Eliasson Lantz, Pär Tufvesson, John Woodley, and Gürkan Sin. Application of mechanistic models to fermentation and biocatalysis for next-generation processes. *Trends in Biotechnology*, 28(7):346–354, 2010.

[3] WO Godtfredsen, N Rastrup-Andersen, S Vangedal, and WD Ollis. Metabolites of fusidium coccineum. *Tetrahedron*, 35(20):2419–2431, 1979.

[4] Stephen Goldrick, Andrei Ştefan, David Lovett, Gary Montague, and Barry Lennox. The development of an industrial-scale fed-batch fermentation simulation. *Journal of Biotechnology*, 193:70–82, 2015.

[5] Council of Europe. *European Pharmacopoeia (Ph. Eur.) 10th Edition*. Strasbourg, 2019.

[6] Gülnur Birol, Cenk Ündey, Satish J. Parulekar, and Ali Çinar. A morphologically structured model for penicillin production. *Biotechnology and Bioengineering*, 77(5):538–552, 2002.

[7] Joel Sansana, Mark N. Joswiak, Ivan Castillo, Zhenyu Wang, Ricardo Rendall, Leo H. Chiang, and Marco S. Reis. Recent trends on hybrid modeling for industry 4.0. *Computers and Chemical Engineering*, 151:107365, 2021.

[8] Christoph Herwig. *Hybrid Modelling and Multi- Parametric Control of Bioprocesses*. MDPI - Multidisciplinary Digital Publishing Institute, 2018.

[9] Mohammed Saad Faizan Bangi and Joseph Sang Il Kwon. Deep hybrid modeling of chemical process: Application to hydraulic fracturing. *Computers and Chemical Engineering*, 134:106696, 2020.

[10] Rasmus Fjordbak Nielsen, Nima Nazemzadeh, Laura Wind Sillesen, Martin Peter Andersson, Krist V. Gernaey, and Seyed Soheil Mansouri. Hybrid machine learn-

ing assisted modelling framework for particle processes. *Computers and Chemical Engineering*, 140:106916, 2020.

[11] Moritz von Stosch, Jan Martijn Hamelink, and Rui Oliveira. Hybrid modeling as a qbd/pat tool in process development: an industrial e. coli case study. *Bioprocess and Biosystems Engineering*, 39(5):773–784, 2016.

[12] Misagh Ebrahimpour, Wei Yu, and Brent Young. Artificial neural network modelling for cream cheese fermentation ph prediction at lab and industrial scales. *Food and Bioproducts Processing*, 126:81–89, 2021.

[13] K Hornik, M Stinchcombe, and H White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[14] Abbas Azarpour, Tohid N.G. Borhani, Sharifah R. Wan Alwi, Zainuddin A. Manan, and Mohamed I. Abdul Mutalib. A generic hybrid model development for process analysis of industrial fixed-bed catalytic reactors. *Chemical Engineering Research and Design*, 117:149–167, 2017.

[15] S. He, N. Sepehri, and R. Unbehauen. Modifying weights layer-by-layer with levenberg-marquardt backpropagation algorithm. *Intelligent Automation and Soft Computing*, 7(4):233–247, 2001.

[16] Parth Shah, M. Ziyan Sheriff, Mohammed Saad Faizan Bangi, Costas Kravaris, Joseph Sang Il Kwon, Chiranjivi Botre, and Junichi Hirota. Deep neural network-based hybrid modeling and experimental validation for an industry-scale fermentation process: Identification of time-varying dependencies among parameters. *Chemical Engineering Journal*, 441:135643, 2022.

[17] Jörg Schubert, Rimvydas Simutis, Michael Dors, Ivo Havlik, and Andreas Lübbert. Bioprocess optimization and control: Application of hybrid modelling. *Journal of Biotechnology*, 35(35):51–68, 1994.

[18] Resul Al and Gürkan Sin. Moskopt: A simulation-based data-driven digital twin optimizer with embedded uncertainty quantification. *Computer-aided Chemical Engineering*, 50:649–654, 2021.

[19] Gürkan Sin, Krist Gernaey, and Anna Eliasson Lantz. Good modeling practice for pat applications: Propagation of input uncertainty and sensitivity analysis. *Biotechnology Progress*, 25(4):1043–1053, 2009.

[20] Adem R. N. Aouichaoui, Seyed Soheil Mansouri, Jens Abildskov, and Gürkan Sin. Uncertainty estimation in deep learning□based property models: Graph neural networks applied to the critical properties. *Aiche Journal*, 68(6):e17696, 2022.

7 Conclusions and Future perspectives

7.1 Achievements

Bioprocess modeling plays a key role in designing and maintaining fermentation processes while also giving key understandings of the relationship between current batch states and microbial kinetics. In terms of developing new state-of-the-art models, this thesis has shown that there are still ongoing developments in classic mechanistic and data-driven modeling. But the reemerging focus on hybrid modeling shows excellent potential in solving some old and general problems in the bioprocess industry.

For the first time, a modeling methodology where the focus is shifted from only predicting the productivity of fed-batch processes but also other batch quality indicators, such as the accumulation of harmful byproducts that render the main product useless. We've applied a novel data-driven method that had the flexibility of naturally solving the uneven-batch length problem as a regression tool. This was the only multivariate method that could sufficiently predict the byproduct accumulation without resorting to batch trajectory synchronization methods.

The conservation balance was also developed from the point of view of using viable biomass as the backbone for all kinetic activity. This was only possible due to our efforts in developing experimental protocols and linear calibrations that allow for easy and fast measurements of viable biomass directly through dielectric spectroscopy. The success of the linear calibration has prompted LEO Pharma to adopt the technique in the standard workflow when collecting lab scale data during experiments to gain further process understanding.

The conservation balance and biochemical models inspired by equation structures used in industrial-scale penicillin models with slight modifications proved a success in developing a mechanistic model that describes several key features relevant to producing Fusidic Acid. Testing the model on an individual batch proved a good fit, and the statistical analysis of model parameters and uncertainties showed extremely reliable results giving a strong mechanistic foundation to test a hybrid model framework.

The hybrid model created using a serial structured approach by integrating ANNs into the mechanistic model equation structure provided a method of predicting the kinetics of a related substance as a function of the current culture conditions. During the development, significant issues arose with the training of the models being exceptionally computationally expensive. This is because the model was written in a MATLAB environment where the machine learning algorithms were obtained from the Neural Network toolbox. This made model evaluation a slow process and gradient descent algorithms for training an unfeasible process. We've managed to massively reduce the model runtime by rewriting a custom Neural Network engine with the focus of only using memory-efficient data types, i.e., allowing double precision arrays and no structure or class types as the toolbox version uses. This level of optimization is generally not needed in standard Machine Learning applications. Still, in cases where a single model evaluation may need 10000 Neural Network evaluations, this led to significant performance increases. Splitting the training problem for byproducts in a pre-training phase also reduced the computational cost in model development. The hybrid models performed exceptionally well on all the data we were able to provide and successfully modeled all the missing state variables deemed necessary to simulate the process with the given purpose. We demonstrate the potential of the models by changing some aspects of the process and show the effects on batch efficiency and quality.

7.2 Remaining Challenges

Despite the achievements attained with the work done during the study, several areas are open for further improvement identified throughout the study.

7.2.1 Data Quality

Hybrid and mechanistic models prove potent for describing complex biological phenomena, and we've successfully taken advantage of machine learning algorithms to predict the concentration of related substances. However, in the current framework, they are not as readily available for implementation as purely data-driven ones. This is not necessarily due to a lack of data but the correct data, a problem that is relevant in the bioprocess industry. Biomass is, for example, an essential process variable in all mechanistic fermentation models, whether it's unstructured or structured fashion, and it's natural that any black box

model integration will still use biomass as a critical variable, meaning that this data needs to be readily available for training. However, even with the increased presence of Process Analytical Technology (PAT) packages and Quality by Design (QbD) principles, it's not guaranteed that biomass data is available. So despite having a lot of data to work with, there may not be enough of the correct type of data meaning that manual experimental work is still required, and the number of batches available for model development will be limited. We addressed the viable biomass aspect of this project and how it could be obtained in real-time in production. However, other fermentation metabolites that are key for the overall mechanistic mass balance still need our attention, such as the main product and byproducts.

7.2.2 Lab and Industrial scale differences

This work was done using data collected from an already-established industrial process. However, most experiments used to support process developments are done at a lab scale. This is the scale new scenarios and process optimization suggestions initially tested due to the expenses associated with industrial-scale test batches. If new phenomena are discovered vital for process optimization, it is imperative that a model developed at the lab scale can be transferred to full scale without too much error. Looking at the model proposed in this thesis, some modifications are required to the proposed equation structure. Due to a significantly smaller scale, most fed-batch lab experiments are done with a condenser installed to prevent evaporation of all the media before the experiment is concluded. It is common to completely ignore the effects of evaporation on the lab scale when developing mass conservation balances. However, it's been established that evaporation can be a significant factor even with a condenser setup[1]. The experimental correlation relating to condenser temperature can be used to account for evaporation at the lab scale. Furthermore, oxygen transfer at the lab scale will need to be adjusted. Oxygen transfer relies on experimental correlation for mass transfer coefficient $k_l a$ whose parameters are expected to change due to different tank geometry and impeller type.

In the current facility, the transfer between lab and full scale is, at this moment, an untested issue. Even if the model with proposed modifications can be calibrated for lab scale, there is no guarantee that a direct transfer is possible until proven otherwise. Should further model development reveal discrepancies in model predictions between scales, further re-

search is required depending on the severity of the model error. Note that this may not be a model-specific issue but rather differences in the environment caused by the varying hydrodynamics of scale-up or scale-down. This is a well-known problem that is still being actively researched today. There have been great strides in Computational Fluid Dynamics (CFD) to understand and avoid process heterogeneity at a large scale[2]. Identifying and eliminating process heterogeneities at both industrial and lab-scale could make kinetic models more transferable between scales and support further process developments by combining lab experiments with state-of-the-art models.

7.2.3 Hybrid training

The power of the hybrid model methodology is showcased as it established a predictive tool for kinetics which there was no prior knowledge, and managed to integrate with mechanistic models. But of course, it doesn't explain everything. While researching and developing the hybrid model, we can find myriad ways to extend the capabilities further. As an example for this research, we focused on one troublesome byproduct. However, according to pharmacopeia, ten more identified impurities must be lower than a certain fraction; otherwise, the final product can not be sold[3]. It is possible to use the same hybrid model methodology to account for every single one with the proper data collection. It is also possible to consider using a parallel hybrid approach to enhance further the prediction quality of state variables that we could model adequately. For example, we could use data-driven models to improve the main carbon source consumption estimate, which of the main state variables had the highest error.

Model training has been a consistent bottleneck throughout this work. Before further extending capabilities as new phenomena are observed or more details are required as the process becomes more sophisticated, it's advocated at least factor in the computational costs of training the models. This work largely tried to mitigate the issue by reducing the cost of model evaluations by focusing on code optimizations. However, it's expected that the computation difficulty will again be an issue with further complex models, no matter how much code optimization is implemented. A more permanent solution might be to reevaluate how the gradient in the loss function is calculated. To that end, combining automatic differentiation with backpropagation is interesting and has been successfully implemented in training hybrid models[4]. Compared to finite-difference, this approach

scales significantly better with increasing model parameters. To benefit from his technique, a software environment is required to support automatic differentiation, which isis only become available recently.

7.2.4 Implementation onto current process

The work presented in this thesis was done to create the basis for a digital twin for the Fusidic Acid fermentation process. While we've shown the capabilities of the models to create a simulation of the system, it can't be said that all the work of building a Digital Twin of the process is complete. A logical next step would be applying these process models to generate value in the current production. However, this is a multifaceted problem and would require a different approach than a pure biochemical engineering one. The models could be written into a software package allowing plant personnel to run and explore various scenarios or even analyze batches currently in production. This does require some relevant knowledge in software engineering and User Interface design so that the relevant people can utilize the models.

Since the model no longer relies on measurement, it can be taken beyond soft sensors and utilized for process optimization. Furthermore, the statistical analysis on the mechanistic part gives uncertainties in model outputs. We can take advantage of the uncertainty information directly with newly developed simulation-based framework tools[5] to further analyze the models, hedge the uncertainties, and find optimal process settings after considering everything we can change within the GMP framework. Part of the problem formulation that motivated this research is that the model's focus on productivity and quality can be utilized as objective and constraint functions in an optimization procedure since the hybrid model has been validated to quantify these state variables.

Part of Industry 4.0 uses Digital Twin of soft sensors, which rely on high-quality models due to the difficulties of developing hardware sensors that measure vital process variables. After incorporating the hybrid, we now have a reliable way of generating predictions of all relevant complicated to estimate state variables. Furthermore, complementing the models with a modern state-of-the-art filter for the measurements or incorporating probabilistic elements will give more accurate measurements over a current batch process[6]. This provides more data for crucial process variables that we can take advantage of to improve the Data-driven tensor models presented in this work or even the hybrid models

themselves.

7.3 Future perspective

Although simulation models prove potent for describing a fermentation process, there are still complex phenomena vital in practical applications that can't be explained with current mechanistic modeling knowledge. While the hybrid may be the emergent solution to these old problems, there is still room for improvement before the industry can take full advantage of them. As part of future work, it would be interesting to compare the different frameworks that have been recently proposed and combine them in a software package so that future researchers can readily take advantage of the other modeling tools.

We touched a bit upon the lack of relevant data in the bioprocess industry to fully take advantage of big data in a hybrid framework. Further improvement of online methods, such as taking full advantage of spectroscopic technology or implementing current models as a soft-sensor, might provide a real-time estimation of the formation of various fermentation metabolites. This will allow us to take full advantage of the hybrid model framing methodology in the presence of Big Data by incorporating all recorded batches rather than a select few via a sampling campaign.

I hope that applying the hybrid modeling framework is a success in the PSE community and the bioprocess industry and motivates research and development to further the current state-of-the-art. The combination of mechanistic modeling and machine learning is a way to speedily solve practical problems today instead of simply waiting for the solutions of tomorrow, which could come way down the line. As it stands, hybrid modeling is still a bit of a niche for the general researcher. There are no official toolboxes in MATLAB or standard packages for Python within SciPy or Tensorflow, with a sizeable community focusing on hybrid modeling despite the recent resurgence. I hope this will change and we see more widespread adoption of the hybrid and multimodal models as common tools in the engineering toolbox.

Bibliography

[1] Magnus Ask and Stuart M. Stocks. Aerobic bioreactors: condensers, evaporation rates, scale-up and scale-down. *Biotechnology Letters*, 44(7):813–822, 2022.

[2] Christian Bach. Modelling of gradients in large scale bioreactors, 2018.

[3] Council of Europe. *European Pharmacopoeia (Ph. Eur.) 10th Edition*. Strasbourg, 2019.

[4] Rasmus Fjordbak Nielsen. Hybrid modelling strategies for development of digital twins, 2021.

[5] Resul Al. Simulation-based framework for design and optimization of wastewater treatment plants, 2020.

[6] Robert Spann, Christophe Roca, David Kold, Anna Eliasson Lantz, Krist V. Gernaey, and Gürkan Sin. A probabilistic model-based soft sensor to monitor lactic acid bacteria fermentations. *Biochemical Engineering Journal*, 135:49–60, 2018.

A Supplementary Materials for Chapter 3

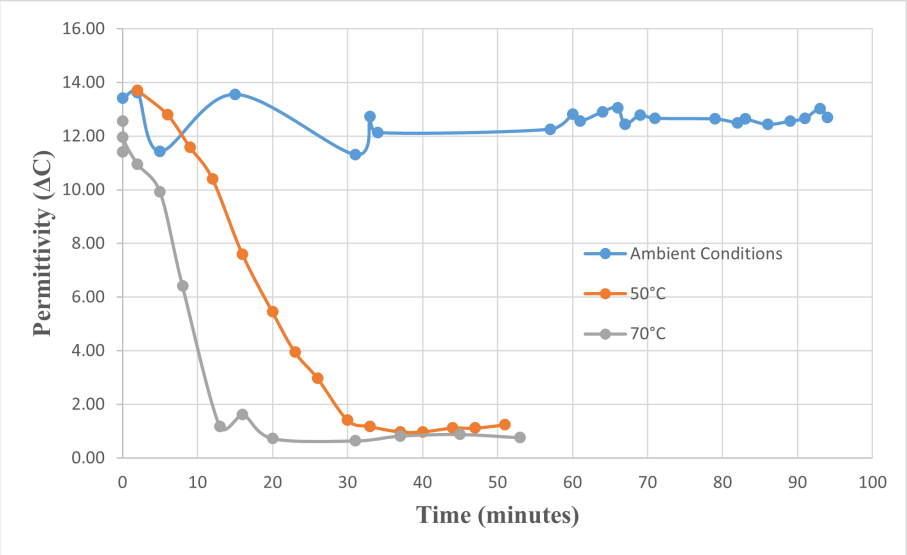

Figure A.1: Dielectric spectroscopy reading on Legacy ABER Cell Analyzer for a sample in ambient conditions and two subsequent samples placed in a water bath at 50°C and 70°C, respectively. It takes approximately 15 minutes to kill a sample in a water bath 70°C and about 30 minutes to kill a sample in a 50°C water bath while a sample in ambient conditions remains stable for over 90 minutes

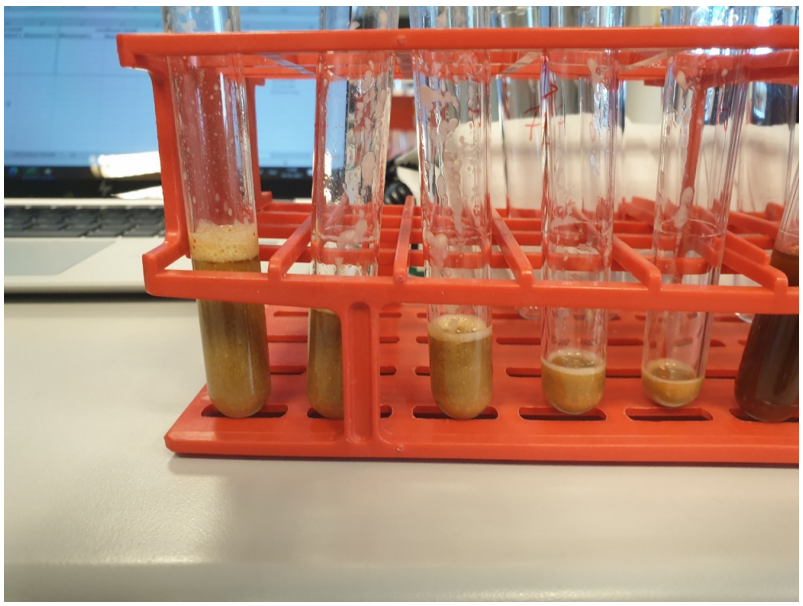

Figure A.2: Various sample volumes considered for this work when measuring with the ABER FUTURA Pico probe. From left to right, the samples are approximately 4 mL, 3 mL, 2 mL, 1 mL and 0.5 mL

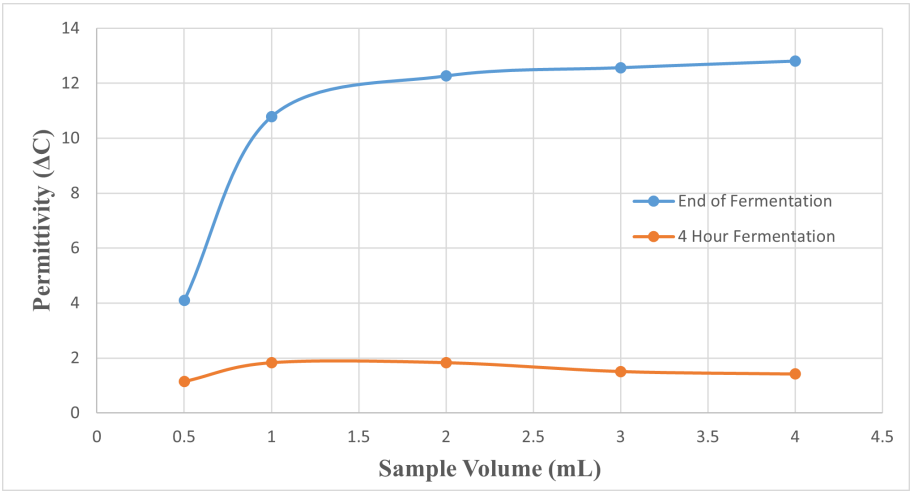

Figure A.3: Dielectric spectroscopy readings with the ABER FUTURA Pico probe with different sample volumes. Two samples are considered, the end of the fermentation sample is taken right before harvest resulting in higher biomass concentrations, and another sample is taken after 4-hour fermentation resulting in lower biomass concentrations. The readings are consistent for 2 mL volumes and above for both high and low biomass concentrations.

Figure A.4: Dielectric spectroscopy reading of approximately 100 mL bulk sample with the ABER FUTURA Pico probe

B Supplementary Materials for Chapter 4

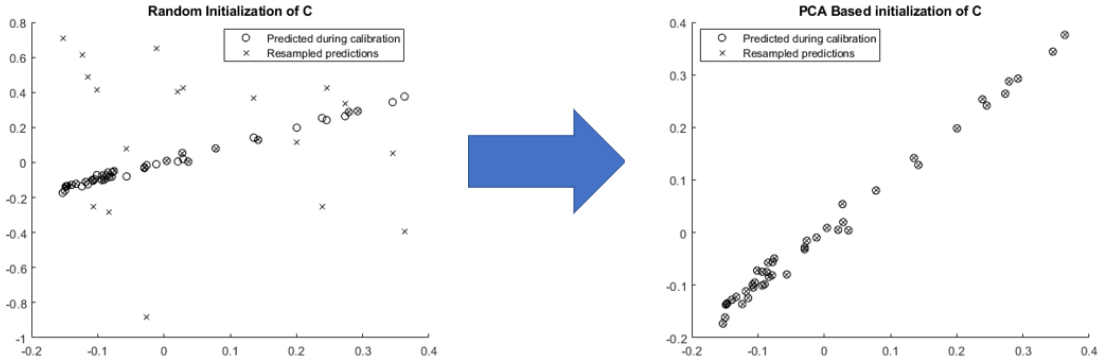

Figure B.1: Comparison of default random initialization for SCREAM prediction on calibrated data and the proposed PCA based initialization

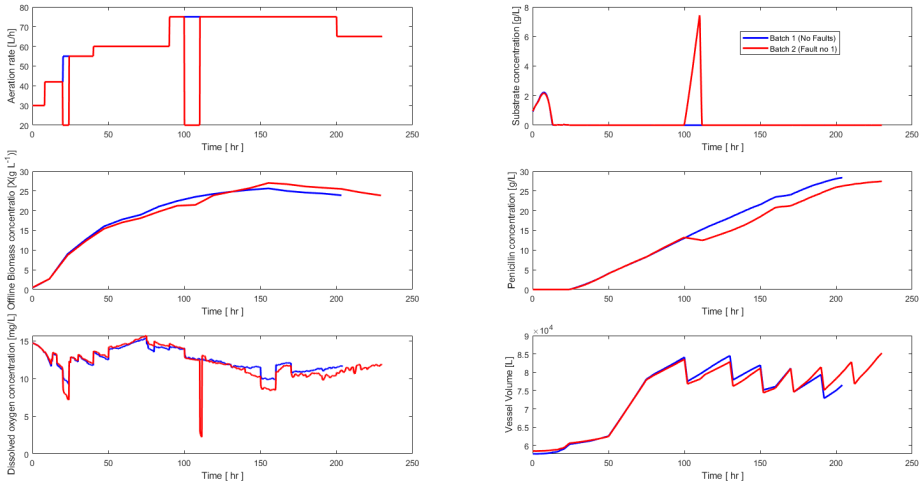

Figure B.2: Example output from the IndPenSim v2.02 software used to generate simulated industrial fed-batch data. Showcased here are batch profiles generated during normal operation and profiles during a fault in aeration rates.

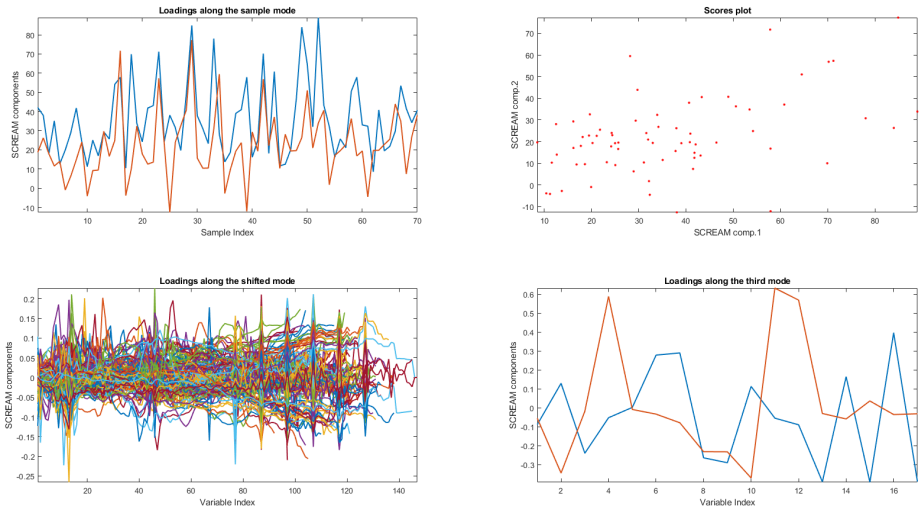

Figure B.3: Model loadings and score matrices generated by the SCREAM source code when modeling the simulated industrial data from the IndPenSim software

C Supplementary Materials for Chapter 5

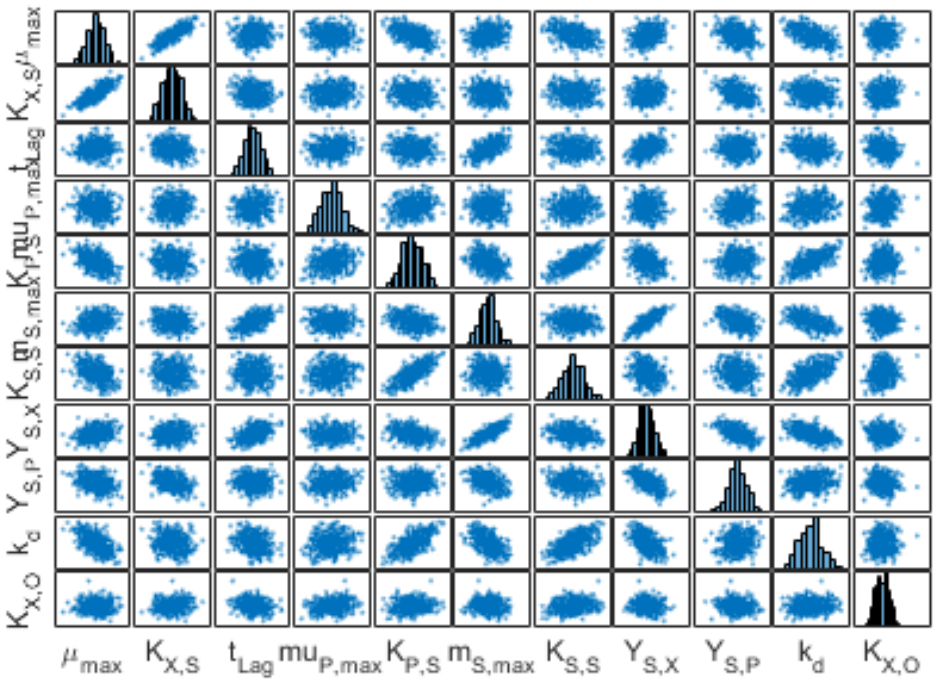

Figure C.1: 500 generated parameter samples used in the Monte Carlo Uncertainty analysis after applying Imon-Conover rank correlation method and inverse probability function on LHS generated samples.

D Supplementary Materials for Chapter 6

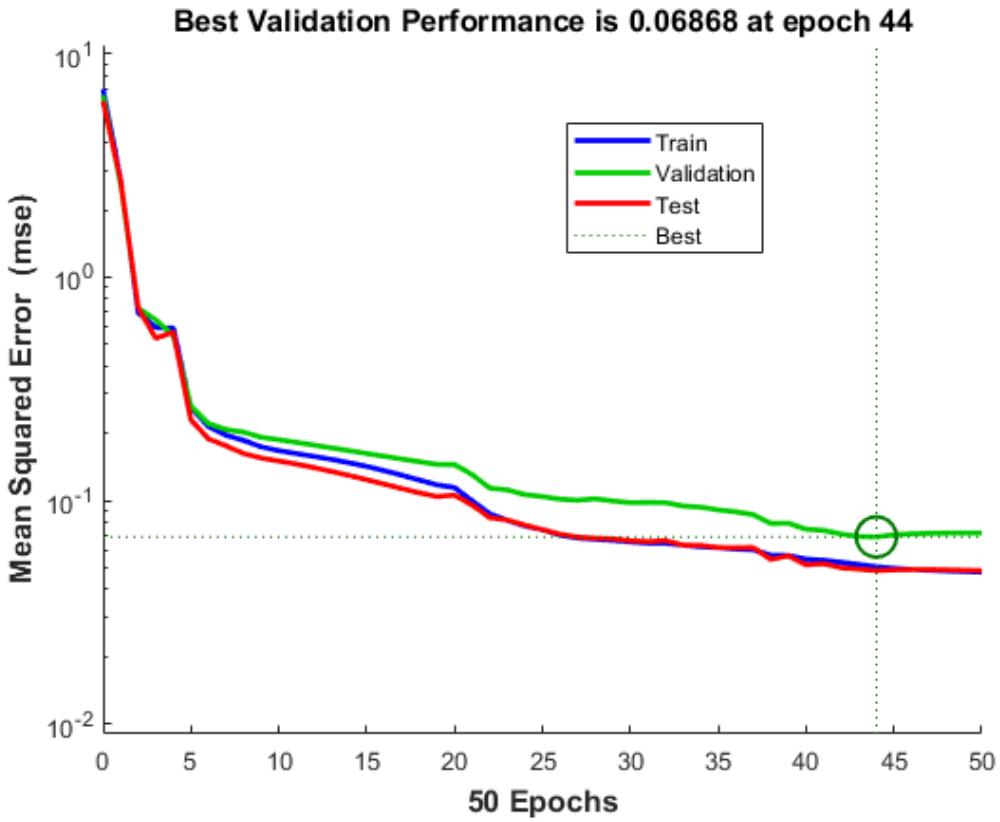

Figure D.1: Training performance of the byproduct model over number of gradient descent epochs

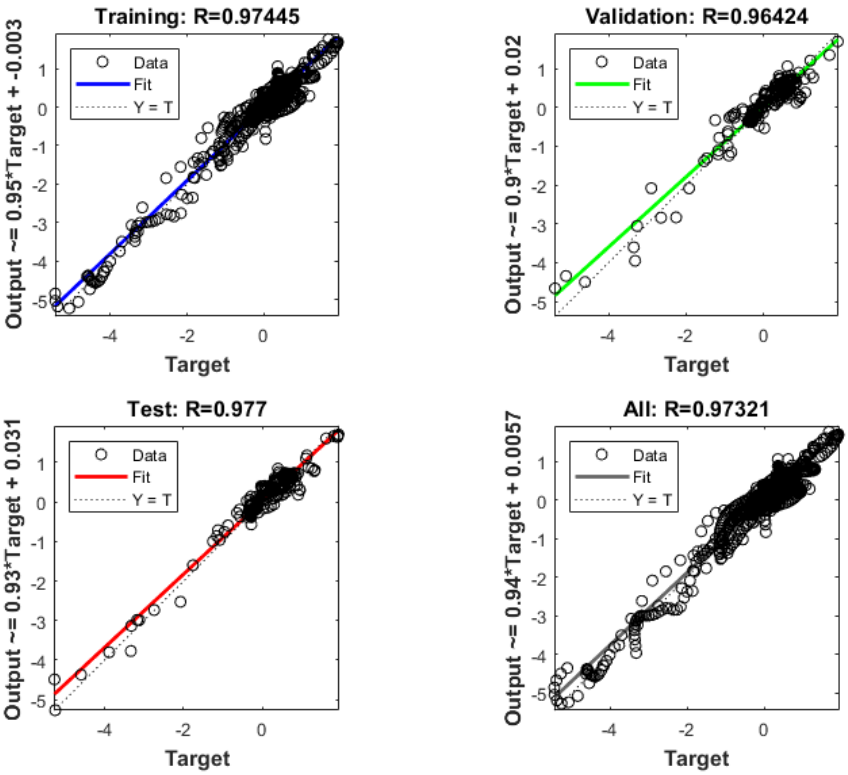

Figure D.2: Regression performance of the byproduct model for the pre-training initialization phase. See main chapter for final regression results

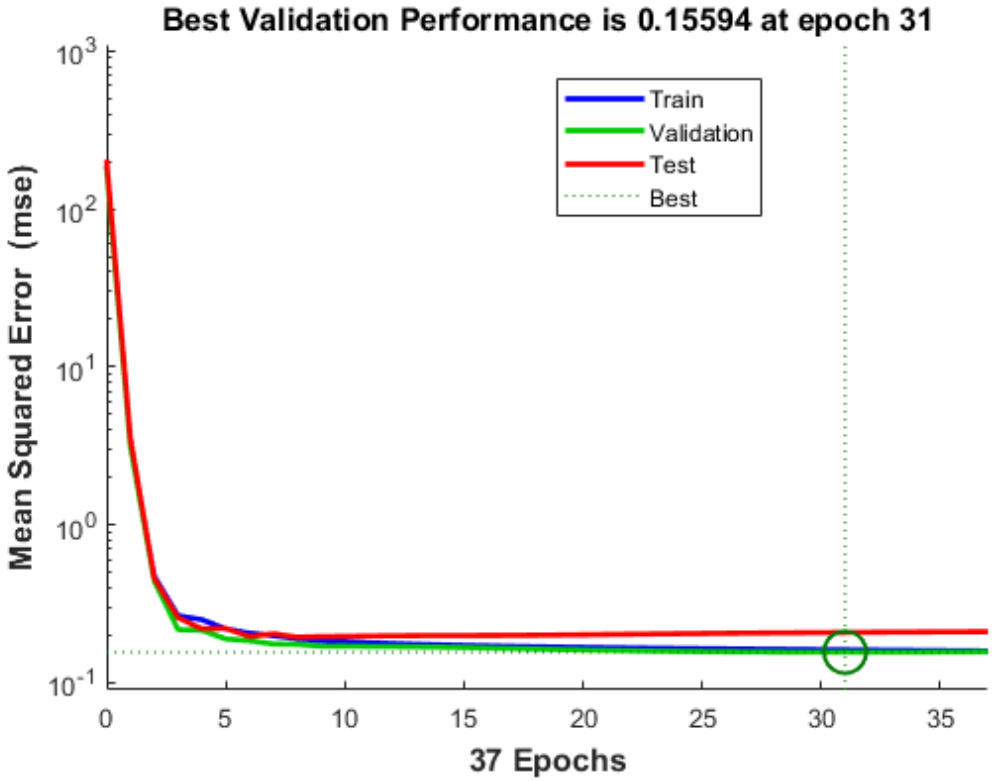

Figure D.3: Training performance of the *Online* model over number of gradient descent epochs

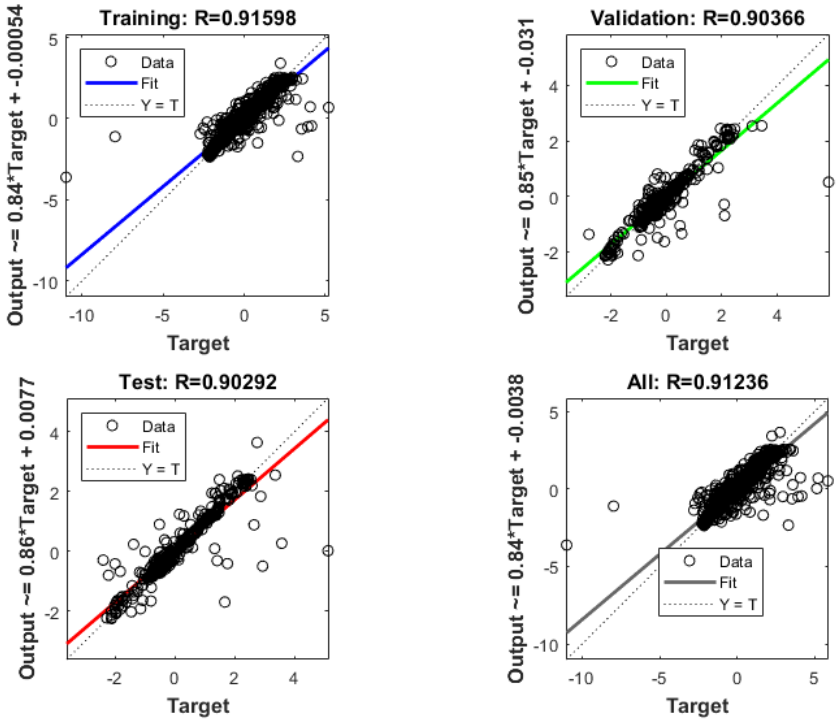

Figure D.4: Regression performance of the *Online* model