Diplomarbeit

Model development and comparison of mechanistic models for *E. coli* to allow for model predictive control

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Master of Science (MSc) unter der Leitung von

Prof. Dr. techn. Dipl.-Ing. Christoph Herwig

betreut durch

Dr. techn. Dipl.-Ing. Sophia Ulonska

Institut für Verfahrenstechnik, Umwelttechnik und Technische Biowissenschaften

Technischen Universität Wien
Fakultät für Technische Chemie

von

Daniel Waldschitz, BSc

Matrikelnummer: 01126111

Wien, am

eigenhändige Unterschrift

Abstract

The complexity of biological systems makes physiological bioprocess control a challenging task. The establishment of physiological process control systems using model based control is key in the development of advanced bioprocess control systems. Advanced bioprocess control strategies, which aim at avoiding unwanted by-products rather then removing them, are recommended by regulatory authorities such as the US Food and Drugs Administration (FDA). In order to do so, models which are able to accurately predict the behaviour of the cells during their cultivation are needed. However, developing such models is a difficult task, since the production of recombinant products causes a lot of cellular stress for the used production hosts, which negatively effects the performance of the cells, resulting in a decline in specific performance. The accurate prediction of the growth and productivity of the cells is necessary in order to set up and facilitate an optimal control strategy for each individual bioprocesses.

For this study two different mechanistic modelling approaches to model the performance of *Escherichia coli*, which allow for model predictive control (MPC), are compared to each other. Thereby, the quality of fit as well as the sensitivity and identifiability of the models is analysed. The first approach, modelled the performance decay of the cells during the cultivation via the cumulative metabolised substrate, and relied only on the gas rates as monitored states of the cultivation. For the second approach, additionally the cell size was monitored as characteristic for the physiological state of the cells used during the process. Hereby, the specific product formation rate of the cells was used to model the performance decay of the cells, instead of using the consumed amount of sugar. The model for the second modelling approach, the cell size model, was newly developed for this study. Both models were analysed and compared towards each other in terms of model fit (using NMRSE values as well as their standard deviation as quantitative measure and time-resolved as well as observed-vs-predictied plots as qualitative measure) and model structure (using local sensitivity analysis and structural identifiability analysis). With the newly developed model a more accurate description of the system with lower NMRSE and standard deviation of NMRSE values (< 11 % NMRSE instead of < 20 % and < 9 % StDev of NMRSE instead of < 15 %) as well as the increased structural identifiability of its parameters (all parameters at once instead of 2 sets of parameters) could be achieved.

It could be shown that both used modelling approaches are giving a valid description of the system. Furthermore, by using the additional monitored state of the cell size, the prediction accuracy and reproducibility could be enhanced. Hereby, more accurate predictions by mechanistic models of industrial relevant bioprocesses, due to an increased insight in cell physiological reactions, can be used to decrease the formation of by-products, to more reliably reach the needed product quality as well as to increase the overall space-time-yields of production plants. However, to apply the newly developed cell size model for industrial bioprocesses, further testing of the model with predictive control needs to be done, in order to compare its control capabilities to commonly applied standard control strategies.

List of Abbreviations

DCW	Dry cell weight
EBC	Elemental balance control
MPC	Model predictive control
GFP	Green fluorescent protein
DoE	Design of experiment
FCM	Flow cytometry
FACS	Fluorescence activated cell sorting
CPP	Critical process parameter
CQA	Critical quality attribute
QbD	Quality by design
HPLC	High pressure liquid chromatography
MW	Molecular weight
IB	Inclusion bodies
M^{3}C	Measuring, monitoring, modelling and control
NRMSE	Normalised root mean square error
StDev	Standard deviation
GMP	Good modelling practice

List of Variables

V	Volume $[L]$
F	Feed $[L/h]$
c_F	Substrate concentration in the feed $[g/L]$
dil	Dilution rate $[L/L/h]$
X	Biomass $[g/L]$
S	Substrate $[g/L]$
Glu	Glucose $[g/L]$
Lac	Lactose $[g/L]$
Gal	Galactose $[g/L]$
P	Product (GFP) $[counts/L]$
IF	Impact factor $[-]$
$S_{met,cum}$	Cumulative metabolized substrate $[g(S)]$
CS	Relative cell size [%]
$Y_{X/S}$	Biomass per substrate yield $[g(X)/g(S)]$
$Y_{O_2/X}$	O_2 per biomass yield $[mol(O_2)/g(X)]$
$Y_{P/X}$	Product per biomass yield $[counts/g(X)]$
K_S	Half-velocity constant of substrate uptake $[g(X)/L]$
q_S	Specific substrate uptake rate $[g(S)/g(X)/h]$
q_X	Specific biomass growth rate $[g(X)/g(X)/h]$
q_P	Specific product formation rate $[counts/g(X)/h]$
OUR	Oxygen uptake rate $[mol(O_2)/h]$
CER	Carbon evolution rate $[mol(C)/h]$

Contents

1 Introduction

1.1 General Field of Study

Bioprocesses are a commonly used tool to produce complex biological compounds such as proteins, hormones, antibodies or secondary metabolites like antibiotics, on an industrial scale for pharmaceutical usage. For their foreseen usage as pharmaceuticals, those compounds need to meet the highest quality standards. Thereby, biotechnological processes are characterised by a huge variety of intracellular interactions, many of which are very complex and still not fully understud [1]. In order to achieve the necessary product quality, the US Food and Drugs Administration (FDA) encourages producers of pharmaceuticals to increase the knowledge of their processes through monitoring and control instead of treating them as a black-box. Thereby, the FDA's Quality by Design (QbD) and Process Analytical Technology (PAT) initiatives [2, 3, 4, 5] aim at avoiding unwanted or harmful by-products during the bioprocesses it self, by monitoring critical quality attributes (CQA) and controlling critical process parameters (CPP), rather than removing them later in additional clean-up unit-operations. The goal the FDA hereby sets, is to achieve the necessary quality by designing the manufacturing process in a way that ensures quality during the development phase of the process.

1.1.1 Industrial Production of Bio-Pharmaceuticals

Commonly a discontinuous industrial cultivation process consists of one or more batch process steps, in order to grow enough biomass out of frozen cell stocks for the production of the desired product. This is followed by un-induced and induced fed-batch process steps to further increase the biomass concentration and for the production of the desired product under limiting conditions [6]. That allows a separate biomass accumulation and product formation phase within a cultivation [7]. Thereby, the nutrient supply during the induction phase of the fed-batch is a powerful tuning factor of the production [8]. The vast majority of products are not native to the used host cells. Gene technology is used in order to facilitate the production in easier cultivatable organisms. A variety of such established standard organisms include *Escherichia coli*, *Saccharomyces cerevisiae*, *Pichia pastoris* as well as *Chinese hamster ovary (CHO)*. The production of such recombinant proteins causes a significant amount of cellular stress, since the cells are forced to use resources, they normally would use for their growth and metabolism, to produce the desired product in large quantities [9]. The incurred amount of cellular stress thereby leads to a decline in cell performance during the induction phase of a fed-batch cultivation [10] compared to similar non-producing cultivations. Furthermore, the formation of non-producing sub-populations has been reported for a variety of hosts [11, 12, 13], since the cells are trying to avoid stressful conditions.

1.2 *Escherichia coli* as Production Host

For this study a well known *E. coli* BL21(DE3) strain, expressing green fluorescent protein (GFP) was used. *E. coli* is a commonly used bacterial host for recombinant production due to its rapid growth up to high cell densities and its high production of recombinant products. The genome of *E. coli* is well known and a lot of strains and cloning vectors for gene engineering are commercially available. One frequently used expression system in *E. coli* is the pET based T7-Lac promoter system. It features a strong induction of the recombinant target genes in the presence of lactose or Isopropyl-β-D-thiogalactopyranosid (IPTG). Due to the high production of the recombinant proteins an accumulation of those proteins into inclusion bodies (IB) often occurs due the capacity folding machinery of *E. coli* being overwhelmed, instead of the formation of the soluble correctly folded native form of the protein. Thereby, the formation of such IB's within *E. coli* causes further stress for the cells [14, 15]. The solubilisation and refolding of such IB's is often associated with a great loss in product [16]. However, the formation of IB's often leads to a greater overall product yield [17]. Furthermore, IB's often have a high purity of the desired product which greatly reduces the effort needed to separate the target protein out of all soluble proteins of the cell [17]. The strain used for this study is well known; it relies on the T7 expression system and has already been used for a variety of previous studies, including:

i) the assessment of differences of induction between lactose and IPTG [18]

ii) the characterisation and quantification of the dependency of the specific lactose uptake rate $q_{S_{Lac}}$ on the specific glucose uptake rate $q_{S_{Glu}}$ for a binary sugar feed [19]

iii) the impact of glycerol as alternative carbon source on the induction [20]

iv) and the assessment of possibilities to tune the production of inclusion bodies (IB) [17].

1.2.1 Effects of Metabolic Stress on *E. coli*

As already mentioned above, the production of recombinant proteins causes a significant amount of cellular stress, since the cells are forced to use resources, they normally would use for their growth and metabolism, to produce the desired product in large quantities [9]. In *E. coli* however, cellular stress leads to physiological changes; the growth of biomass is rather characterised by an increase in cell length instead of regular cell division [21, 22]. Hereby, an intermediate phase between normal cell growth and cell death is reported [21]; standard cell growth up to a certain threshold, followed by cell division and growth back to the threshold (normal cell growth), gets altered to a viable but non-culturable (VNBC) state where cells inflate their size above normal values without dividing (intermediate phase), before cell death and folowing membrane disintegration occurs (lysis phase) [21]. In that intermediate phase, basic cell viability like growth and the synthesis of proteins and metabolites is still present whereas more complex biological functions like cell division is already hampered.

1.2.2 Possible Adaptations to cope with Metabolic Stress

Monitoring such physiological changes can provide additional insight and explanation for the performance of the strain. The cell size can be easily monitored during the bioprocess using flow cytometry (FCM) [23, 24]. Fluorescence activated cell sorting (FACS) has been used to screen for strains which are less stressed by the production of their respective target proteins [25, 26]. In order to do so, fusion proteins of the desired product fused to a fluorescent marker protein are generated. Subsequently, sorting for maximal fluorescence of single cells after a cultivation, is used to screen for cells with higher than average production. However, a complete elimination of the stress caused by producing conditions is not possible. Therefore, even for optimised strains an accurate description of the effects of metabolic stress is necessary. Hereby, process modelling can be used to find and predict the optimal conditions for the production of the desired product with the given host cells.

1.3 Goals of the Models used for this Study

For this study two different approaches to mechanistically model the decline in specific cellular performance (via q_S or q_P) for *E. coli* should be compared. Additionally to the fist modelling approach, an adapted version of the model described in [32], a second model should be developed. Both models should thereby ultimately be usable for model predictive process control with a minimal monitoring effort required.

Hereby, the models should only rely on:
i) strain specific constants, which can be measured (like the sugar uptake rates)
ii) the initial values of the states at the start of the bioprocess (like the initial biomass in the reactor)
iii) the input values of the feed-rates for sugars, base and oxygen which they control
iv) as few as possible measured outputs (like the gas rates or the cell size) as feedback of the current state of the bioprocess, in order to keep the monitoring effort low

The newly developed model thereby should:
i) take the physiological state of the cells into account
ii) model the cellular metabolic stress via the specific product formation rate q_P
iii) expand the monitoring strategy only by measuring the cell size with flow cytometry, in order to not inflate the overall monitoring effort required
iv) deliver a comparable or better description of the process
v) deliver a comparable or better sensitivity and identifiability of the involved model parameters

Thereby, a new model featuring the novel method of modelling the physiologic changes via the specific production rate q_P, combined with the monitoring of the physiologic changes (the relative cell size CS in *E. coli*), should be compared to existing approaches.

1.4 Model Development

In recent years the progress in development of reliable and robust processes is facilitated by advances in measuring, monitoring, modelling and control of bioprocesses (M^3C) [27]. Hereby, the functional relationship between CPP and CQA needs to be investigated [3]. Mathematical models can be used to simulate model such relationships in order to generate further knowledge about the interlinks between specific CPP's to CQA's [28]. Furthermore, such mathematical models can be used as knowledge storing systems [29]. However, currently there is a lack of accepted standardised work-flows to set up the modelling work needed [28]. Luckily, good modelling practice (GMP) guidelines lay out three similar basic steps which are always needed for successful model development [30]:

i) set-up of a modelling project
ii) set-up of a model
iii) analysis of the model

1.4.1 Modelling Project

The basis for every modelling project is a clear and precise definition of the gaols. Thereby, efficient process models should be as simple as possible while being as accurate as necessary [31]. Models can be used to generate new as well as solidify existing knowledge, additionally, their application for facilitate process control is of high interest to researchers [31, 32, 33, 34, 35, 36]. For an advanced modelling strategy, which is needed to facilitate process control, a high amount of knowledge about the specific process is needed. However, it than allows for a efficient reach of the specified CQA's as well as an overall increased space-time yield of the bioprocess [37]. In order to do so, strain specific characteristics, such as the effects of the production, on the production (like the decline in specific performance due to the stressful production of recombinant proteins) need to be better understood, and the developed models, need to account for the decline in biomass growth and product formation during the induction phase. This is necessary to avoid miss-estimations of the biomass which would cause for a loss in control over the feeding strategy of the cultivation [38, 39], when the model is employed for process control.

1.4.2 Model Set-up

Generally models can be classified as static or dynamic models [31]. Dynamic models include differential equations which allow for a time or location dependant prediction of model states. They provide a way to use the information generated by the monitoring strategy with the knowledge about the specific bioprocess in a predictive way. Static models, however, cannot provide time dependant predictions of the system and rather represent correlations between sates. Thereby, different model types exist; frequently they are classified as data-driven, hybrid or mechanistic models [31]. Data-driven models require a sizeable training data set to set up, and can only operate with that dataset. Mechanistic models consist of mathematical equations with physiological meaningful parameters [40]. Thereby, mechanistic models need a lower amount of training-data to set up and validate, but require packing the correlations and connections between the the states of the system into model equations. Hereby, the set-up of models as well as the model analysis are iterative steps [41]. Recently there is some effort to standardise and automatise the work-flow of setting up new models, however, these approaches are still at an rather early stage [28, 42, 43, 44, 45].

1.4.3 Model Analysis

In order to perform the model analysis, first a parameter fit of the model parameter is performed. Hereby, an optimization algorithm is employed to adapt the values of the model parameters to yield an optimal result based on an specified optimisation criteria. The normalised mean root square error (NMRSE) is often used as such an optimisation criteria. Thereby, the optimization algorithm alters the values of the model parameters to minimise the NMRSE values of the model states [31]. Afterwards sensitivity and identifiability analysis are performed to analyse the model. Hereby, the sensitivity of the model parameters is a prerequisite for their identifiability [46]. Not all parameters may be sensitive at all phases during the course of an cultivation. However, it is important that a parameter is sensitive, in at least one phase of the cultivation, in order to be able to accurately fit the value of the parameter to the described system. Insensitive parameters can occur when:

i) more than one parameter influences the system in a way, that a change of the other parameters can compensate the change in the analysed parameter which makes them not directly identifiable in general (for example qs_{max} and Ks)

ii) or the exact value of the parameter is not important for the predictive quality of the model with the given model structure

Determining whether it is possible to identify the model parameters with the given model structure based on its equations, can be used to distinguish between the two types. Insensitive parameters then either:

i) need different experimental conditions to be sensitive and as a result identifiable
ii) need to be taken from literature when the information about their value is already known
iii) or can entirely be replaced by a structural change of the model, if they are superficial for the model

In the end, only a model with only sensitive/identifiable parameters or known/published parameters allows for a valid description of the system [47, 46]. Hence, if this cannot be achieved either:

i) more experiments with different conditions need to be done
ii) or the structure of the model needs to be changed, until it can be achieved, since otherwise no valid model can be derived

To determine the local sensitivity of the model parameters, their values are deviated from the optimal value (calculated with the optimisation algorithm), and the response in the optimisation criteria quantified (for example the change in NMRSE of the model states). Furthermore, a ranking of the parameter importance can be performed to analyse what the critical parameters for the model are [46]. Structural identifiability looks at the collinearity of the model parameters. Too collinear parameters thereby can not be identified at the same time, due to their correlation to each other. Hereby a simple threshold is employed to classify the identifiability of the model parameters [47].

1.5 Modelling of Metabolic Stress

As mentioned above, in commonly used industrial production hosts such as *E. coli*, during cultivations, changes occur on a physiological level. Thereby, it is reported that the physiological capabilities decrease during the cultivation, leading to a decrease in physiological parameters [10]. In order to accurately describe the system, process models need to take such changes into account. One described way to model physiologic changes during the cultivation, uses the cumulative metabolised substrate ($S_{met,cum}$) [48, 32]. Hereby, all metabolic stress causing factors, which lead to the physiological decrease in performance, are summed up and simplified by the amount of energy available to the cells (via the amount of substrate fed q_S), which is used to perform all cellular functions. This results in a highly simplified description of the system, but has already shown its effectiveness in some cases [48, 32]. Additionally, also some work on modelling the change of physiologic parameters such as the cell size, using kinetic models based on the amount of substrate available to the cell, has been done [49].

All those approaches rely on the amount of energy available to the cells, to abstract and simplify the metabolic stress causing effects, instead of featuring a physiologically more meaningful description of the system. Since the performance decay of the cells occurs in cultivations producing the desired target recombinant proteins compared to cultivations were no recombinant proteins are produced, the specific product formation rate q_P seems to be a valuable variable to model physiologic changes [9]. However, no mechanistic models using the the specific product formation rate q_P to model the change of physiologic parameters, exist to the authors knowledge.

1.6 Model Set-up

Cumulative Substrate Model The cumulative substrate model is an version of the model described in [32], featuring minor adoptions and simplifications mainly in the description of the sugar metabolism of the cells. For this modelling approach the cumulative metabolised substrate ($S_{met,cum}$) during the induction phase of the cultivation is used as a measure of the incurred cellular stress. Thereby, the cumulative metabolised substrate ($S_{met,cum}$) is used as a negative feedback onto the biomass per substrate yield ($Y_{X/S}$) [48] to account for the decline in specific performance [32].

This results in a simple method with the advantage that no additional separate measurements have to be performed. However, it relies on the specific substrate consumption rate (q_S) to model the performance decline. Hereby, the causes of the cellular stress which lead to the decay of performance are abstracted by the amount of sugar consumed during stressful conditions. This simplification may not be very physiologically meaningful, since the sugar consumption itself is not the stress causing factor, but it allows for a valid but basic description of the performance decay [48].

Cell Size Model The cell size model was newly developed for this study. To generate a more advanced modelling method, the cell size (CS) is monitored with flow cytometry and used to characterise the physiological state of the *E. coli* cells [50, 51, 52]. Subsequently, that physiological state is used for modelling purposes [49]. Hereby, the specific product formation rate (q_P) is used to model the cell size increase, which serves as a negative feedback onto the biomass per substrate as well as the product per biomass yield ($Y_{X/S}$ and $Y_{P/X}$) to model the performance decline.

1.7 Workfolw of this Study

In the following chapters, the set-up of the two models including the model equations and parameters is discussed in detail [Chapter 3.1]. As optimization criteria for the model parameter fit the normalised mean root square error (NMRSE) of the model states was used. Afterwards the fit of the models to the system is analysed, using the NMRSE values as well as their standard deviation as quantitative measure, and time-resolved as well as observed-vs-predictied plots as qualitative measure [Chapter 3.2].

This is followed by the analysis of the model quality [Chapter 3.3]. As basis for the comparison the model quality, the local sensitivity of the used parameters as well as the structural identifiability of the models was used. Finally the performance of the two models [Chapter 4.1], as well as their structure and real-time applicability [Chapter 4.1.2 - 4.1.3] plus their ability to describe the system [Chapter 4.2] are discussed and compared.

2 Materials and Methods

2.1 Experimental Plan

In order to show the validity of the model for a broad set of conditions, a two feed system with glucose (main C-source) and lactose (C-source and inducer) where both feeds were independently controlled, was chosen. A two level two factor full factorial design of experiments (DoE) with a triplicate centre point [Factor 1: $q_{S_{Glu}}$ (Levels: 0.15; 0.35 g/g/h); Factor 2: $q_{S_{Lac}}$ (Levels: 0.05; 0.1 g/g/h); 7 experiments total] was used as experimental plan. The levels were chosen to lie under the specific maximal lactose uptake rate [Figure 2.1].

2.2 Strain

Fed-batch cultivations performed for this study used an *Escherichia coli* BL21(DE3) strain, expressing GFP as a fluorescent model protein under a T7-promoter. A $pET21a^{+}$ vector with the gene for the GFP thereby provided the T7 induction system. Lactose was used to induce the production of GFP under the T7-promotor.

2.3 Cultivations

The experiments were conducted in a DASGIP parallel bioreactor system (Eppendorf AG, Germany, working volume 2 L) and monitored with potentiometric pH sensors (Hamilton, Switzerland), and optical DO probes (Hamilton, Switzerland). The exhausted gas composition was analysed by a ZrO_2 sensor for O_2 and an infra-red sensor for CO_2 analysis (DASGIP module GA4, Eppendorf AG, Germany). The feeds and base were supplied via peristaltic pumps (DASGIP module MP8, Eppendorf AG, Germany). All experiments were performed with the full synthetic media described in [53] at 35 $°C$ and 1400 rmp stirrer speed (DASGIP module TC4SC4, Eppendorf AG, Germany). The pH was kept constant at 7.2 with 12.5 % NH_4OH and the dissolved oxygen (DO) was kept over 30 % (DASGIP module PH4PO4, Eppendorf AG, Germany) with a gas supply of 2 vvm via a L-sparger (DASGIP module MX4/4, Eppendorf AG, Germany). A pre-culture in shake flasks (5 L, with 500 ml media [53]) was cul-

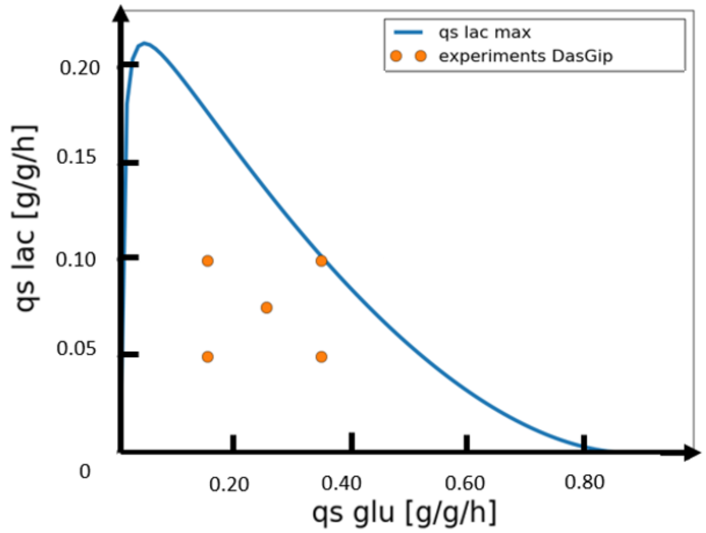


Figure 2.1: The set points for the performed experiments under the maximal specific lactose uptake rate $q_{S_{Lac}}$ dependent on the specific glucose uptake rate $q_{S_{Glu}}$ as proposed by [19].

tivated from frozen glycerol-stocks (-80 $°C$) overnight (16 h at 32 $°C$ and 200 rmp). 10 % of the initial reactor liquid volume (100 ml of 1 L) was inoculated using the grown pre-culture. To acquire a target biomass concentration of 25 g/L for the start of the experiment, cells were grown on glucose as exclusive C-source, first in a batch phase (22 g/L glucose) upon depletion of the C-source and afterwards in a fed-batch (450 g/L glucose in Feed, $q_{S_{Glu}}$ of 0.25 g/g/h). After reaching the target biomass the experiments were started by feeding glucose (450 g/L) and lactose (180 g/L). The online-data was logged by the process management software of the reactors (DASware control software, Eppendorf AG, Germany).

2.4 Set Point Control

The specific sugar uptake rates for the experiments were controlled by elemental balance control (EBC) [32] using the soft-sensor described in [54]. Both feeds as well as the base were individually controlled and supplied from separate storage bottles. The pH and temperature were regulated via a PID controller using the process management software of the reactors (DASware control software, Eppendorf AG, Germany). DO was kept over 30 % by increasing the oxygen content in the inflow air when necessary starting from synthetic air without additional oxygen added (DASware control software, Eppendorf AG, Germany).

2.5 Samples

Dry cell weight (DCW) was determined gravimetrically by centrifugation (4500 g, 10 min, 4 °C) of 5 mL cultivation broth and subsequent drying of the cell pellet for 72 h at 105 °C. Cell-free samples of the cultivation broth were analysed for concentrations of substrates and metabolites by HPLC (UlitMate3000, Thermo Scientific, USA) with a Supelco gel C-610 H ion exchange column (Sigma-Aldrich, USA) and a refractive index detector (Agilent Technologies, USA). A mobile phase of 0.1 % H_2PO_4 with a flow rate of 0.5 mL/min at 30 °C was used. Cell size and GFP content were measured photometrically with flow cytometry (Cube 8, Sysmex, Switzerland). The forward-scatter can thereby be correlated to the cell size, however, it can also be influenced by changes in cell morphology like inclusion body formation [14, 15] which makes it a less reliable tool for our purposes. Thereby, the sideward-scatter can be correlated to the granularity of the cells. Therefore, the cell size was additionally quantified by staining all cells with a fluorescent dye that binds to the cell membrane, which is thereby delivering a signal proportional to the cell size which is independent of the cell morphology. For staining of the cells 2 μl of a 2 mM stock of RH414 (AnaSpec, USA) in dimethyl sulfoxide stored at -20 °C were added to 1 mL of a 1:10000 dilution of the cell brought in 0.9 % NaCl solution [19]. For the measurement of the GFP content, an excitation wave-length of 488 nm and an emission wave-length of 509 nm were used for the photometric measurement.

2.6 Data-Analysis

Evaluation of the data produced by flow cytometry (FCM) was done with CyFlow (Sysmex, Switzerland). Principal component analysis performed in Matlab 2015b (Mathworks, USA). Evaluation of the DoE data was done in MODDE 11 (Umetrics, Sweden). Statistical evaluation of the data was done in SIMCA 13 (Umetrics, Sweden). The experimental data was reposited and exported into Python using Incyght (Exputec, Austira).

2.7 Model-Analysis

The mechanistic models were programmed in Python. Model parameters were estimated using an adapted version of the fmin downhill simplex optimisation algorithm from the SciPy package as described in [55, 56] additionally considering for parameter boundaries. For the analysis of the quality of fit of the models, the normalised root mean square error, and its standard deviation between the different experiments was calculated [Equation (2.1)].

$$NRMSE = \frac{\sqrt{\frac{\sum(\hat{y}-y)^2}{\mathrm{len}(y)}}}{\max(y) - \min(y)} \tag{2.1}$$

Local sensitivity analysis and structural identifiability analysis was performed according to [47] and [46] using Matlab 2015b (Mathworks, USA). Thereby, the local sensitivity of the parameters was determined by the distortion of single parameters (every 10 min for 1 % of the parameter value) and measuring the change in the model outputs relative to no distortion. For ranking the parameter importance (δ), an unscaled version of the method described in [46] using Matlab 2015b (Mathworks, USA) was used. It is important to mention, that the specific scaling method used can lead to very different results; for example dividing by the state value for scaling [46] can lead to strong overestimations of states with very low numeric values ($<$1). The benefits of scaling the parameter importance (δ), by dividing through the state values, are that the results can be interpreted without having the results effected by the numeric values of the states (for states with values not too close to zero). For this study the numeric values for Glu, Lac, CER and OUR are all very low. Therefore, an unscaled parameter importance ranking is used for this study. Only the states X, P, Gal, IF, CER and OUR were used for the parameter importance ranking, since they are the most important states for the model. The numeric values of Glu and Lac were below the limit of quantification (LoQ) of the used HPLC method, therefore they were excluded for the parameter importance ranking. Structural identifiability analysis was performed according to [47] and [46] using Matlab 2015b (Mathworks, USA). For the structural identifiability analysis, a collinearity index (γ) threshold of 10 was applied as suggested by [47]. Below that threshold parameters were classified as structural identifiable, since two too collinear parameters are not identifiable at the same time, due to their correlation to each other.

3 Results

3.1 Mechanistic Models

The mechanistic models used for this study were based upon a basic mechanistic model described in [32]. Both models share a common backbone describing the reactor balances and the principle sugar uptake as well as the specific carbon and oxygen metabolism capabilities of the used strain. The models only differ in the description of the biomass growth and product formation according to the different modelling approaches used for the performance decline of the cells which should be compared. For the first modelling approach only the carbon evolution rate (CER) and the oxygen uptake rate (OUR) are monitored. For the second approach additionally the cell size (CS) is monitored to asses the physiological state of the cells, providing additional information to the system, while not inflating the monitoring effort.

The differential equations for the concentrations of biomass (X), glucose (Glu), lactose (Lac), galactose (Gal), product (P) and volume (V) were as follows:

$$\frac{dX}{dt} = q_X \cdot X - dil \cdot X \tag{3.1}$$

$$\frac{dP}{dt} = q_P \cdot X - dil \cdot P \tag{3.2}$$

$$\frac{dGlu}{dt} = -q_{S_{Glu}} \cdot X + \frac{F_{Glu} \cdot c_{F,Glu}}{V} - dil \cdot Glu \tag{3.3}$$

$$\frac{dLac}{dt} = -q_{S_{Lac}} \cdot X + \frac{F_{Lac} \cdot c_{F,Lac}}{V} - dil \cdot Lac \tag{3.4}$$

$$\frac{dGal}{dt} = q_{S_{Lac}} \cdot \frac{MW_{Gal}}{MW_{Lac}} \cdot X - dil \cdot Gal \tag{3.5}$$

$$\frac{dV}{dt} = F_{in} - F_{out} \tag{3.6}$$

The uptake rates of the two substrates glucose (Glu) and lactose (Lac) were modelled Monod-like. The disaccharide lactose consists of a glucose and a galactose monosaccharide of which only glucose can be metabolised. Therefore, the total metaboliseable substrate $q_{S_{met}}$ needed to be calculated, in order to account for the additional glucose that is available for the cells when they cleave the lactose disaccharide. The gas rates (CER and OUR) were calculated stoichiometrically. Due to the limited range of variation of $q_{S_{Glu}}$ and $q_{S_{Lac}}$ used within the DoE, a more detailed description of the dependency of $q_{S_{Lac}}$ by $q_{S_{Glu}}$, as proposed in [19], could be omitted.

$$
\begin{aligned}
q_{S_{Glu}} &= q_{S_{Glu,\max}} \cdot \frac{Glu}{Glu + K_{s_{Glu}}} && (3.7)\\
q_{S_{Lac}} &= q_{S_{Lac,\max}} \cdot \frac{Lac}{Lac + K_{s_{Lac}}} && (3.8)\\
q_{S_{met}} &= q_{S_{Glu}} + q_{S_{Lac}} \cdot \frac{MW_{Gal}}{MW_{Lac}} && (3.9)\\
CER &= q_{s_{met}} \cdot X \cdot V \cdot \frac{Glu}{MW_{Glu}} - q_X \cdot X \cdot V \cdot \frac{X}{MW_X} \cdot \frac{Y_{X/S,max}}{Y_{X/S}} && (3.10)\\
OUR &= -CER - q_X \cdot X \cdot V \cdot Y_{O_2/X} && (3.11)
\end{aligned}
$$

The equations for the specific growth and product formation rates (q_X and q_P), as well as the used feedback on the yields ($Y_{X/S}$ and $Y_{P/X}$), differed between the two compared approaches:

3.1.1 Cumulative Substrate Model

For the first mechanistic modelling approach of the decline in specific performance, the cumulative metabolised substrate $S_{met,cum}$ is used [32, 48] as an impact factor for the negative feedback. Thereby, the specific metaboliseable sugar uptake rate $q_{S_{met}}$ is cumulated up during the induction phase of the bioprocess, which leads to a constantly increasing impact factor for the negative feedback. Hence, the amount of incurred stress due to the production of the target protein is estimated with the amount of sugar that is taken up during producing conditions.

$$
\begin{aligned}
IF &= S_{met,cum} && (3.12)\\
\frac{dS_{met,cum}}{dt} &= q_{s_{met}} && (3.13)
\end{aligned}
$$

The biomass per substrate yield $Y_{X/S}$ is the target of the negative feedback, resulting in an increased amount of g sugar needed per g biomass produced.

$$
\begin{aligned}
Y_{X/S} &= -Y_{X/S,crit} \cdot \frac{IF}{\sqrt{d_{Y_{X/S}} + IF^2}} + Y_{X/S,max} && (3.14)\\
q_X &= q_{S_{met}} \cdot Y_{X/S} && (3.15)
\end{aligned}
$$

The product formation is modelled as simple growth associated production, since the production occurs whilst the cells are growing. A constant yield for the amount of product per substrate $Y_{P/X}$ had to be assumed, since there is no monitoring strategy for this approach which could asses changes over time of that yield. For doing so, a product related state would need to be monitored.

$$Y_{P/X} = constant \tag{3.16}$$
$$q_P = q_X \cdot Y_{P/X} \tag{3.17}$$

3.1.2 Cell Size Model

For the second mechanistic modelling approach of the decline in specific performance, the cell size (CS) is monitored with flow cytometry (FCM), in order to characterize the physiological state of the cells. Hereby, the specific productivity of the cells q_P is used to model the cell size increase in order to generate a predictive model, since the production of the target protein is the catalyst for the decline in performance compared to non-producing cultivations [21]. The relative cell size increase during the cultivation was used as impact factor, for the negative feedback of the stressful production conditions on the performance of the cells.

$$IF = CS \tag{3.18}$$
$$\frac{dCS}{dt} = \frac{q_P \cdot CS}{Y_{CS/P}} \tag{3.19}$$

Since the cell size increase is correlated to the incurred stress by the production [21, 22], the biomass per substrate yield $Y_{X/S}$ and the product per biomass yield $Y_{P/X}$ can be target of the negative feedback. For this more advanced approach the biomass (X) and the product (P) are both part of the feedback loop and not only the biomass (X) as in the first approach.

$$Y_{X/S} = \frac{Y_{X/S,max}}{IF} \tag{3.20}$$
$$q_X = q_{S_{met}} \cdot Y_{X/S} \tag{3.21}$$
$$Y_{P/X} = \frac{Y_{P/X,max}}{q_{S_{met}} \cdot IF} \tag{3.22}$$
$$q_P = q_X \cdot Y_{P/X} \tag{3.23}$$

3.1.3 Model Parameters

The fitted parameters as well as the parameters taken from literature are listed in Table 3.1. Since the affects of the different calculation method of the specific performance decay should be assessed for this study, the same values were used for the

Table 3.1: Model parameters

	Parameter	Unit	Value	Reference values
Shared Model Backbone				
	MW_X	g/mol	26.54	(based on stoichiometry)
	MW_{Glu}	g/mol	180.16	(based on stoichiometry)
	MW_{Gal}	g/mol	342.29	(based on stoichiometry)
	C_{Glu}	C-mol/mol	6	(based on stoichiometry)
	C_X	C-mol/mol	1	(based on stoichiometry)
	$q_{s_{Glu,\max}}$	g/g/h	0.88	0.88-1.06 [19]
	$q_{s_{Lac,\max}}$	g/g/h	0.36	[32]
	$K_{s_{Glu}}$	g/l	0.099	0.00005-0.099 [57]
	$K_{s_{Lac}}$	g/l	0.058	[32]
	$Y_{X/S,max}$	g/g	0.5	< 0.7 [48]
	$Y_{O_2/X}$	mol/g	0.01	(based on stoichiometry)
Cumulative Substrate Model				
	$d_{Y_{X/S}}$	g^2/g^2	3.93	[32]
	$Y_{X/S,max}$	g/g	0.5	[32]
	$Y_{X/S,crit}$	g/g	0.3	[32]
	$Y_{P/X}$	counts/g	874	[32]
Cell Size Model				
	$Y_{CS/P}$	counts/counts	8481.92	-
	$Y_{X/S,max}$	(g·counts)/g	0.3474	-
	$Y_{P/X,max}$	(g·counts^2)/(h·g^2)	378.44	-

shared parameters of the basic model. Therefore, differences in the model fit and prediction quality of the two models can be discussed only in perspective of the calculation method, and independent of differences in the values of the shared parameters. It should be mentioned, that some of the model specific parameters also occur in both models but have an inherently different unit due to the different calculation of the actual yields, as can be seen in Table 3.1. Furthermore, it is worth mentioning that in the cumulative substrate model some of the parameter values are assumed as physiologic constants, which were not fitted (the $Y_{X/S,max}$ and the $Y_{X/S,crit}$ to be specific).

3.2 Model fit

Several methods are applied to analyse and compare the two different modelling approaches. First of all, the quality of fit of the simulated data to the experimental data was analysed. Hereby, a simple plotting of the time-resolved values is an easy accessible way of qualitatively assessing the performance of the model. Additionally, the normalized root mean square error (NRMSE) can be calculated to quantitatively asses the model performance and the standard deviation of those NRMSE values to asses reproducibility. In the following, the time-resolved plots for a centre point of the DoE are shown and discussed since,

i) the centre point of the DoE represents the medium conditions the model encounters within the DoE

ii) and the centre point is measured in a triplicate ensuring that the consistency of the results can be evaluated.

The time-resolved plots are only practical for each experiment individually. However, for a broad set of varying conditions for the experiments, like in a DoE for example, the quality of fit will probably be influenced by the exact experiment which is examined. Therefore, in a second step the prediction quality of the models should be analysed for all DoE experiments, analysing the capability of the models to cope with the different sets of conditions. For that, observed-vs.-predicted plots are employed as qualitative assessment of the model fit and the NMRSE values and their standard deviation as quantitative measure.

3.2.1 Basic Centre Point Fit

Cultivations are prone to high biological variations of cells. For this reason, testing the reproducibility of the system is very important. The triplicate centre point of the DoE offers a way, for the assessment of the reproducibility of the experiments in general, as well as the used measurements and methods in particular. Furthermore, the models involved need to be able to describe the system in an accurate way for all the replicates equally. To check whether the used models can provide that, the NRMSE values of the triplicate centre point specifically as well as the standard deviation of those NRMSE values are analysed [Table 3.2]. For both models the standard deviation of the NRMSE values is below 5 %, except for OUR where it is slightly above 5 %, which is extremely

Table 3.2: NMRSE & standard deviation values of the triplicate centre point of the DoE for both models in %

		X	P	Gal	CER	OUR	IF
Cumulative Substrate Model	NRMSE (in %)	10.83	9.03	4.39	1.15	13.27	-
	StDev of NRMSE	1.88	0.68	1.26	0.44	5.04	-
Cell Size Model	NRMSE (in %)	5.88	9.60	4.39	1.15	13.71	15.60
	StDev of NRMSE	2.45	2.11	1.26	0.44	5.44	3.80

good given the high biological variations of cells. However, for OUR the slightly higher value can be explained by looking at the equation used to calculate it [Equation 3.11]. The calculated values for CER and X, which both come up with an error, are used for the calculation of OUR. Hence, the error of the derived value will generally be slightly larger.

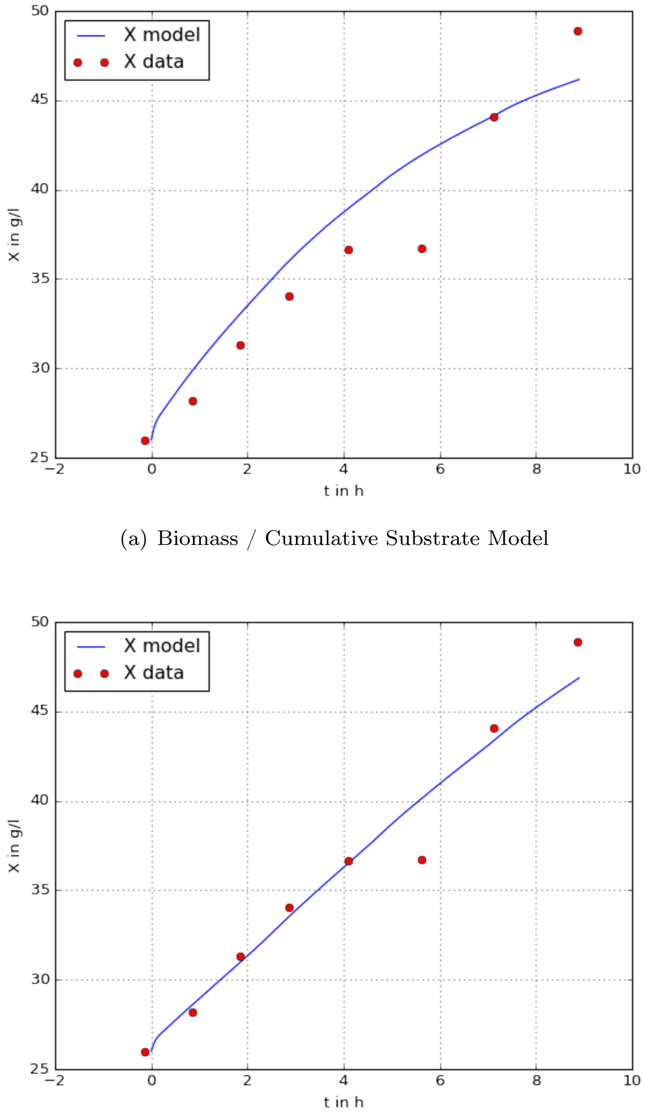
The NMRSE values themselves are also quite promising. The prediction of Gal and CER is excellent. NMRSE values of around 10 %, like for X and P are adequate for biological systems. The values for OUR and CS are slightly higher, however, this is due to the multiple other states which are used for their calculation. Comparing the two different models to each other, the biomass X is much better represented in the cell size model, displaying an improved biomass X description of that modelling approach. For the product P, NRMSE values are comparable, however, the standard deviation of the NMRSE values is slightly higher for the cell size model, probably due to the slightly more complex calculation method used. The values for Gal, CER and OUR are all very similar for both models since their calculation used the same model backbone. The predicted impact factor IF values can only be compared to measured values for the cell size model, illustrating the benefits of the expanded monitoring strategy. In summary, the cell size model shows a comparable or better description of the system as the cumulative substrate model, based on the quantitative assessment with the NMRSE values.

The following part discusses whether the qualitative description of the system (based on the time-resolved plots) reassembles the results of the quantitative assessment (with the NMRSE values), in depth for all the states of the model:

Biomass Both models represent the biomass growth of the experiments adequately. However, differences in the behaviour of the description between the models are observable in the time-resolved plots [Figure 3.1]. The biggest difference in the behaviour of the simulated experiments to the measurements is visible for the biomass X. The decline in biomass growth for the cumulative substrate model is much stronger compared to the cell size model, leading to a stronger curve of the simulated biomass. The cumulative substrate model thereby seems to suffer from the way how the negative feedback of the impact factor IF is implemented. As impact factor IF the cumulative metabolised substrate is used, which leads to a steadily increasing negative feedback over the course of the cultivation. The fit is better at the beginning and the end of the cultivation due to the stronger curvature of simulated biomass. In contrast, the more linear simulated biomass of the cell size model fits the measured data visibly better, which is also supported by the lower NRMSE value of this model [Table 3.2]. Additionally, it should be mentioned that the measured biomass value at approximately 6 h is known to be an outlier.

The calculated values of the specific biomass growth rates q_X are very similar for both models [Figure 3.2]. However, the decrease in the specific rates during the cultivation is generally stronger for the cumulative substrate model. This is also the case for the biomass per substrate yield $Y_{X/S}$, where additionally also stronger bending of the simulated $Y_{X/S}$ is observable. The minimum and maximum values of the $Y_{X/S}$ are determined by the parameters $Y_{X/S,max}$ and $Y_{X/S,crit}$, whereas only an upper limit for the $Y_{X/S}$ is set in the cell size model. The resulting $Y_{X/S}$ values are strikingly lower for the cell size model at the beginning of the cultivation. For the cell size model the difference in values of the $Y_{X/S}$ is in general much smaller, thereby the the values of the yield parameters os the cell size model were fitted whereas the cumulative substrate model used values for $Y_{X/S,max}$ and $Y_{X/S,crit}$ are assumed to be physiologic constants in this model.

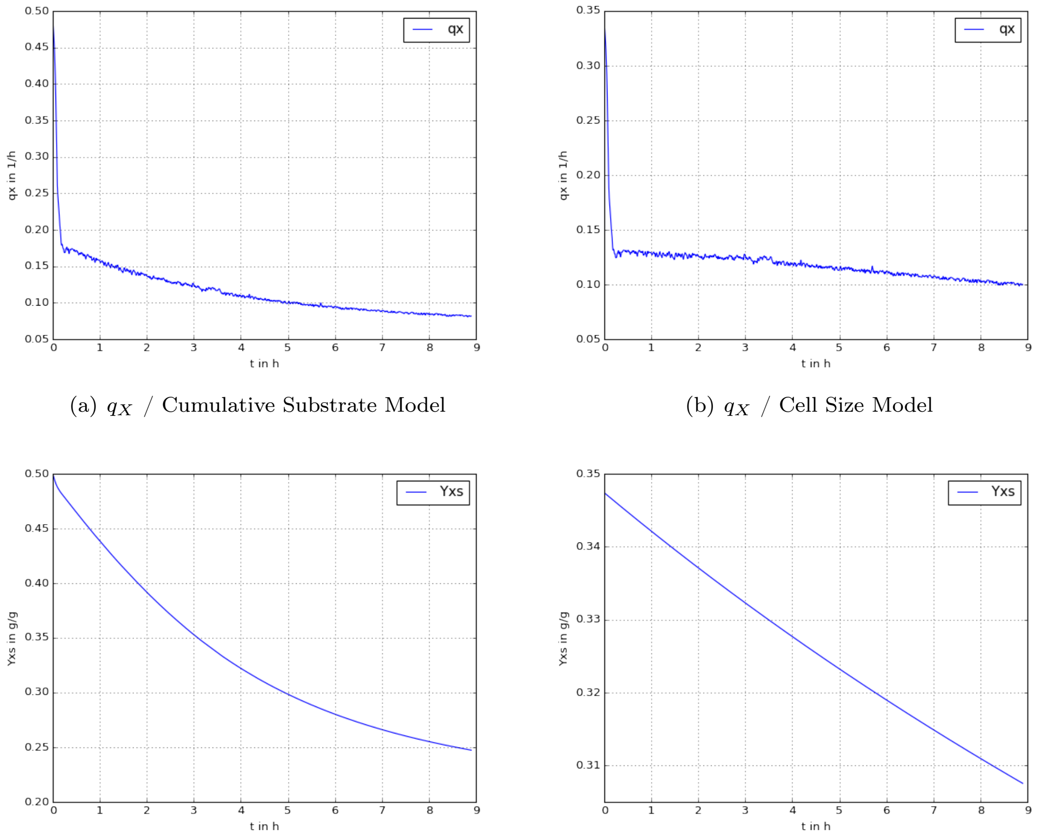
Product The behaviours of the simulated product P values are similar to the ones of the biomass. Although the NRMSE value for the cell size model is lower, the simulated product accumulation of cumulative substrate model reassembles the behaviour of the measured data points closer [Figure 3.3]. The measured data points show a stronger decrease at the end of the experiment for the product than for the biomass. This leads to an optically better resemblance of the product accumulation behaviour with the stronger curved simulated product accumulation of the cumulative substrate model. It could be assumed that the high NRMSE value is caused by the underestimation of P at the end of the cultivation. This is probably due to the need to accurately describe product accumulation for a variety of conditions within the DoE [Table 3.2]. The lower NRMSE value for the cell size model is probably achieved by a more accurate



(a) Biomass / Cumulative Substrate Model

(b) Biomass / Cell Size Model

Figure 3.1: Model fit comparison of measured to simulated biomass X, for a centre point of the DoE with both models respectively. The simulated values are represented by a solid line and the measurements by dots. The measured biomass value at approximately 6 h is known to be an outlier.



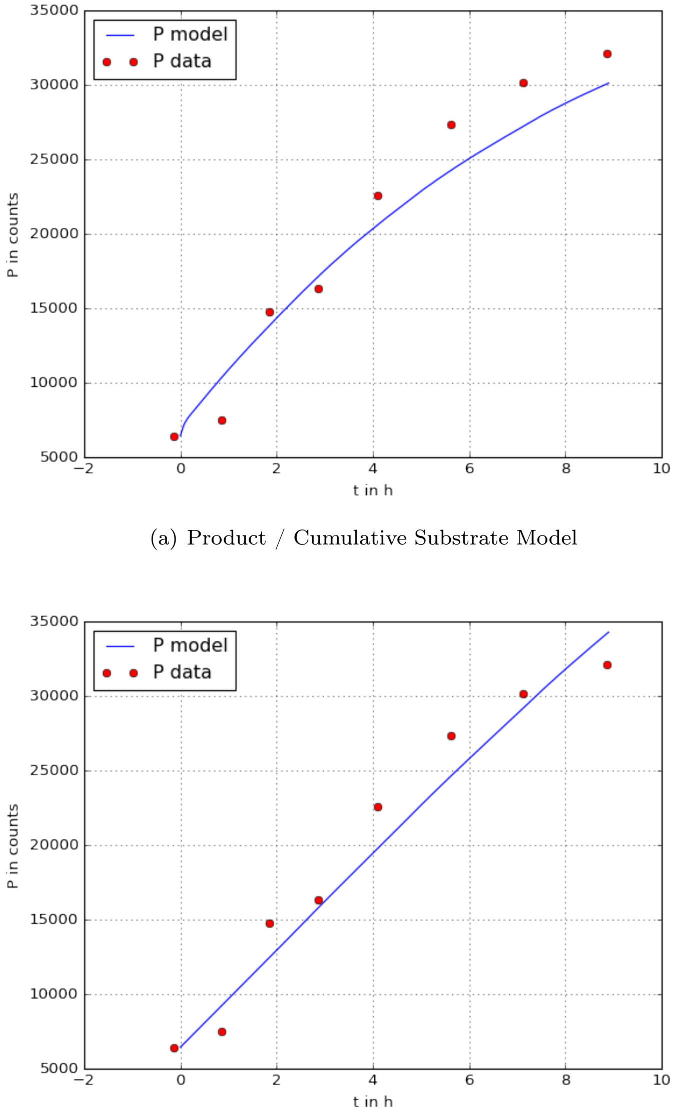
(a) q_X / Cumulative Substrate Model

(b) q_X / Cell Size Model

(c) $Y_{X/S}$ / Cumulative Substrate Model

(d) $Y_{X/S}$ / Cell Size Model

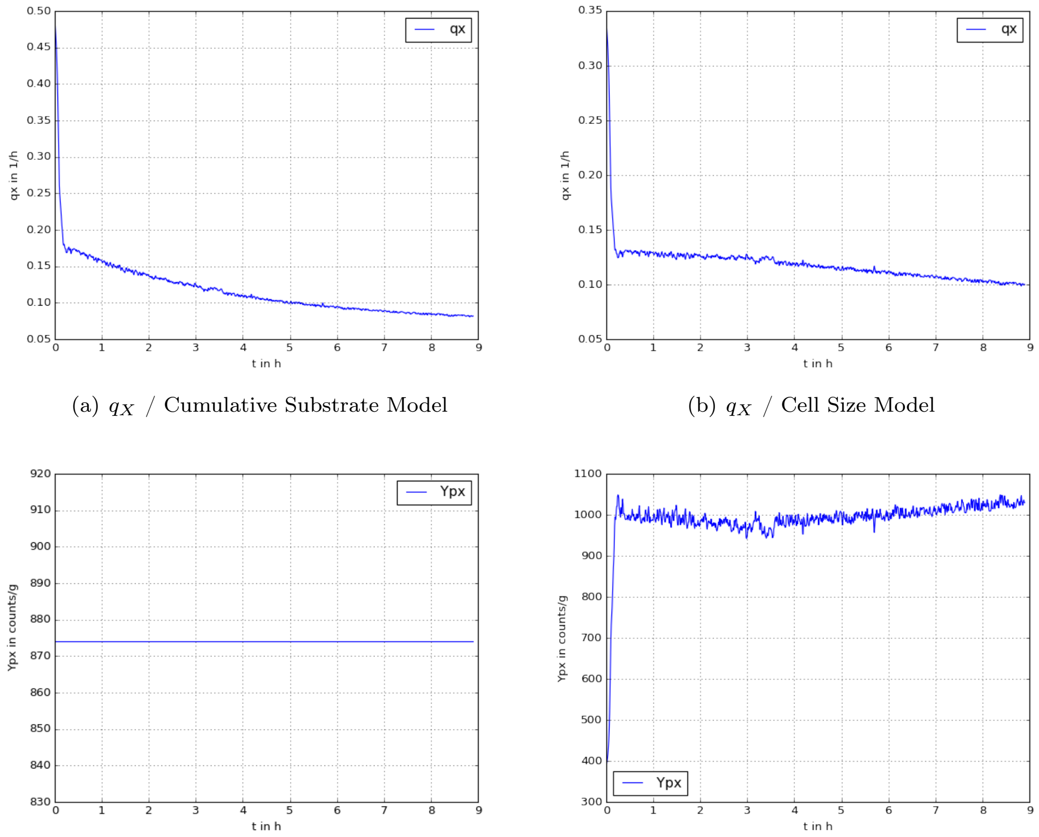
Figure 3.2: Comparison of the calculated values of q_X and $Y_{X/S}$ for both models, for a centre point of the DoE. lots a) and b) show the calculated q_X values for both models respectively. Subplots c) and d) show the calculated $Y_{X/S}$ values for both models respectively.



(a) Product / Cumulative Substrate Model

(b) Product / Cell Size Model

Figure 3.3: Model fit comparison of measured to simulated product P, for a centre point of the DoE with both models respectively. The simulated values are represented by a solid line and the measurements by dots.



(a) q_X / Cumulative Substrate Model

(b) q_X / Cell Size Model

(c) $Y_{P/X}$ / Cumulative Substrate Model

(d) $Y_{P/X}$ / Cell Size Model

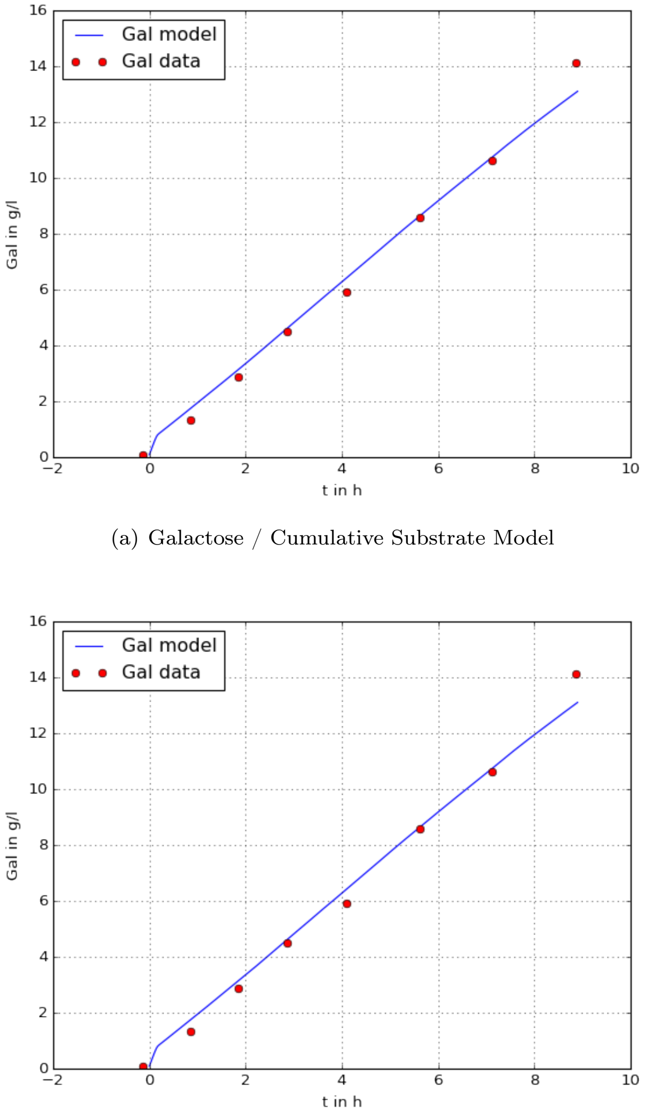
Figure 3.4: Comparison of the calculated values of q_P and $Y_{P/X}$ for both models, for a centre point of the DoE. Subplots a) and b) show the calculated q_P values for both models respectively. Subplots c) and d) show the calculated $Y_{P/X}$ values for both models respectively.

estimation of the amount of product rather than by accurately describe the behaviour of the production.

The calculated values of the specific product formation rates q_P are similar for both models [Figure 3.4]. Again the decrease in the specific rates during the cultivation is generally stronger for the cumulative substrate model. The values of q_P in the cumulative substrate model are higher at the beginning of the cultivation and lower at the end of the cultivation then the values of the cell size model. The product per biomass yield $Y_{P/X}$ for the cumulative substrate model had to be assumed as a constant, due to the lack of a suitable monitoring strategy for the product. However, the $Y_{P/X}$ of the cell size model is also remarkably constant during the cultivation. Since the NMRSE values for both models for X and for P are good, a constant $Y_{P/X}$ seems to be a valid assumption for this specific process. The value of the $Y_{P/X}$ is higher for the cell size model though.

Sugars and Gas Rates For the uptake of the substrates, glucose Glu and lactose Lac, the measured sugar values were below the limit of quantification (LoQ) for the used HPLC method, therefore no comparison to measured values can be done. The simulated values of glucose Glu and lactose Lac are also consistently close to zero for the whole simulation. However, at the beginning of the cultivation there is always a short period of time needed until the simulated values stabilise. Fortunately the production of galactose Gal is directly connected to the consumption of lactose via stoichiometry, which makes an assessment of the accuracy of the lactose uptake possible by looking at the galactose accumulation [Figure 3.5]. Furthermore, the production of CO_2 measured as the carbon evolution rate CER as well as the oxygen demand required to produce CO_2 by metabolising the sugars, measured as the oxygen uptake rate OUR, can be used to indirectly verify the accuracy of the total sugar consumption. Since the lactose consumption can independently verified with the galactose production, the CER can be used to verify the glucose uptake in case of an accurate Gal description. Both models represent the galactose Gal accumulation excellently [Table 3.2]. In addition, the CER is also represented excellent. However, for the OUR the measured values are sightly overestimated (meaning more oxygen is consumed in the simulation than for measured values, leading to even lower values in the plot since the expected uptake is greater) [Figure 3.6].

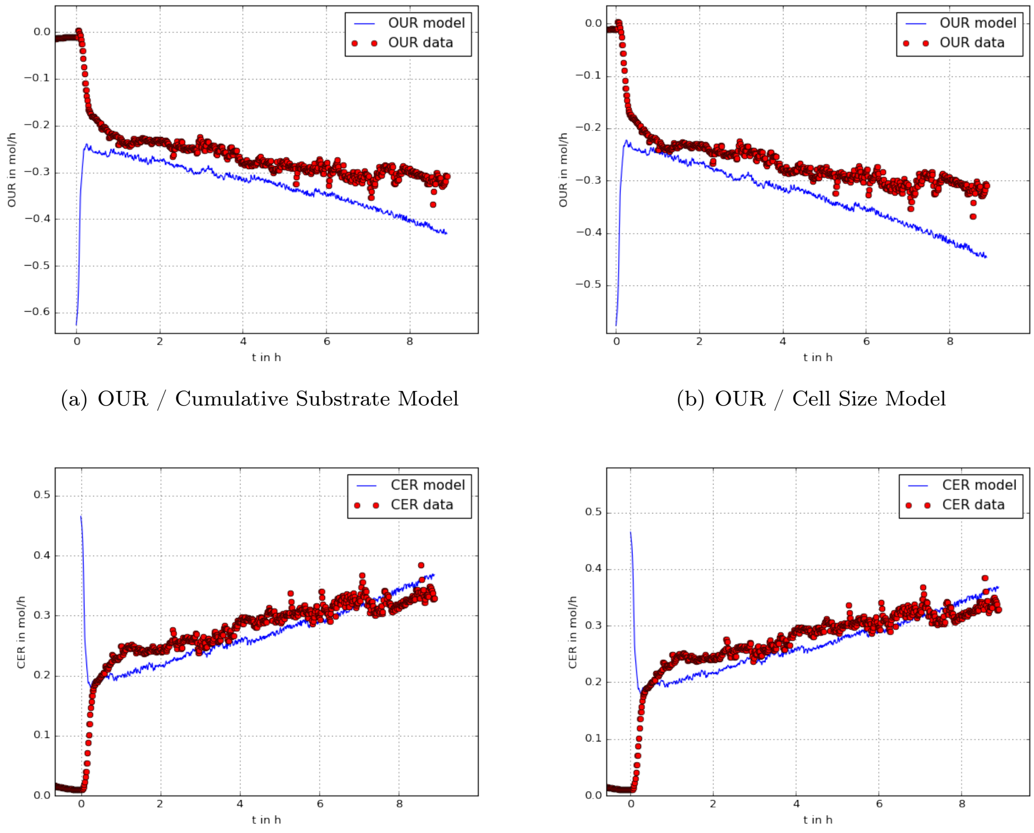
Impact Factor Regarding the impact factor IF, only for the cell size model a comparison to measurements is possible. The NMRSE value of the relative cell size CS is acceptable, though a slightly less linear description of the relative cell size increase would reassemble the behaviour of the measured data points slightly more optimal [Table 3.2; Figure 3.7].



(a) Galactose / Cumulative Substrate Model

(b) Galactose / Cell Size Model

Figure 3.5: Model fit comparison of the *Gal* accumulation, with both models for a centre point of the DoE. The simulated values are represented by a solid line and the measurements by dots.



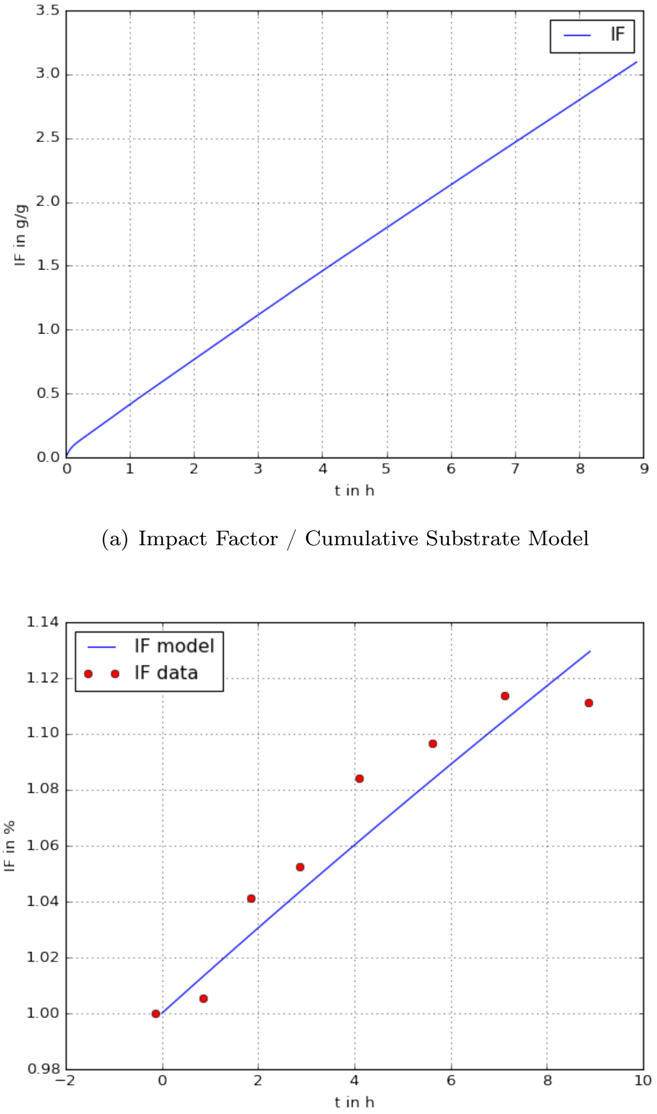
(a) OUR / Cumulative Substrate Model

(b) OUR / Cell Size Model

(c) CER / Cumulative Substrate Model

(d) CER / Cell Size Model

Figure 3.6: Model fit comparison for the OUR and CER estimation, with both models for a centre point of the DoE. The simulated values are represented by a solid line and the measurements by dots. Subplots a) and b) show the calculated OUR values for both models respectively. Subplots c) and d) show the calculated CER values for both models respectively.



(a) Impact Factor / Cumulative Substrate Model

(b) Impact Factor / Cell Size Model

Figure 3.7: Model fit comparison of measured to simulated impact factor IF, for a centre point of the DoE. The cumulative substrate or the cell size respectively are thereby used as impact factor for the models to describe the incurred cellular stress. Note that no comparison to measured values can be done due to the lack of a monitoring strategy for the incurred cellular stress for the cumulative substrate model. The simulated values are represented by a solid line and the measurements by dots.

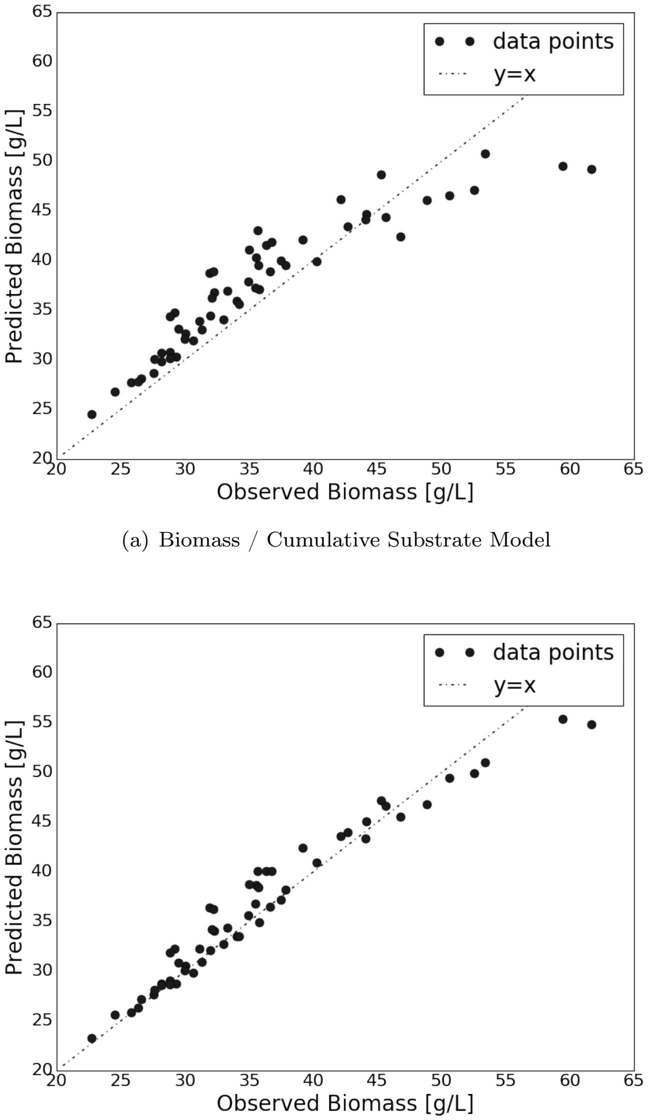
Table 3.3: NMRSE & standard deviation values of the whole DoE in %

		X	P	Gal	CER	OUR	IF
Cumulative Substrate Model							
	NRMSE (in %)	19.72	14.79	7.65	1.36	14.63	-
	StDev of NRMSE	14.12	5.53	5.73	0.97	7.25	-
Cell Size Model							
	NRMSE (in %)	10.78	10.31	7.64	1.36	15.08	15.04
	StDev of NRMSE	8.22	2.88	5.73	0.97	7.97	4.70

3.2.2 Full DoE Model Fit

To asses the capabilities of the models to describe the system accurately for a variety of conditions, the NMRSE values for the whole DoE can be used. In order to do so, all experiments were taken into account for calculating the NMRSE values and their standard deviation, instead of only the experiments of the triplicate centre point [Table 3.3]. For the cumulative substrate model the description of biomass X and product P is adequate. However, an increase in the standard deviation of the NMRSE values compared to the values for the triplicate centre point can be observed due to the increased variety of conditions. The description of Gal, CER and OUR is comparable to the values of the centre point, showing no loss of prediction quality for the expanded set of conditions. For the cell size model significantly lower NMRSE values for the biomass X and product P are obtained. Furthermore, the standard deviation of those NMRSE values is also significantly lower, showing the increased capabilities of the cell size model to cope with the varying set of conditions. Additionally, the impact factor IF (the relative cell size CS) also shows comparable values to the centre point. Hence, its prediction quality is unaffected by the broader set of conditions. In summary, the cell size model shows an increased ability to deliver an accurate description of a broader set of conditions than the cumulative substrate model does.

Biomass Plotting the observed state values against the predict ones (x = y line) can provide further insight into the prediction quality of the models [Figure 3.8]. For the biomass X, at medium cell densities the biomass gets over-estimated for both models, however, this effect is stronger for the cumulative substrate model. Contrary, there is an under-estimation for high cell densities which is especially strong for the



(a) Biomass / Cumulative Substrate Model

(b) Biomass / Cell Size Model

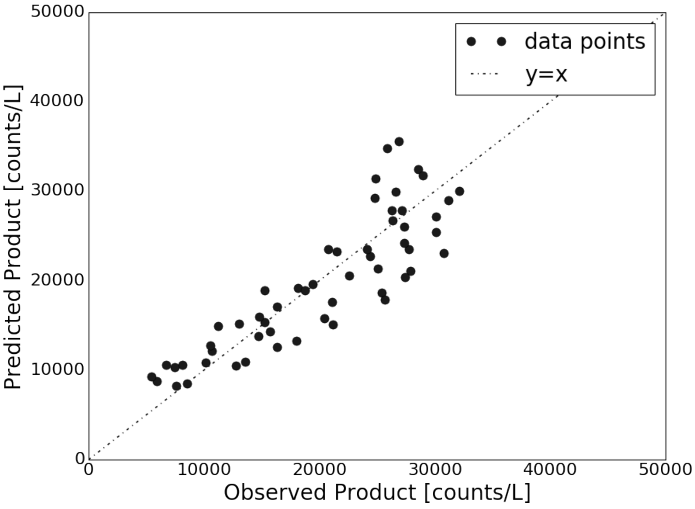
Figure 3.8: The observed values plotted against the predicted values of the biomass X, represented dots in the plots, for all performed samples of all experiments for both models respectively. The predicted values are on the y-axis and the observed values on the x-axis. The x = y line resembles a perfect prediction of the observed values, represented as dotted line in the plots.

cumulative substrate model. This is explaining the significantly better NMRSE values for the biomass X of the cell size model. However, it can be seen that there is no single especially high set of miss-estimations for the biomass X at all cell densities. This shows that both models can facilitate a valid description of the system for the entire set of conditions.

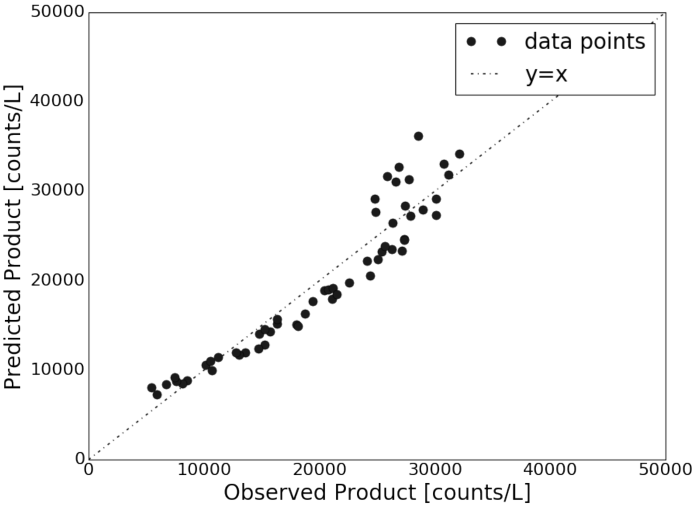
Product For the product P, the data-points are significantly more scattered for the cumulative substrate model than for the cell size model [Figure 3.9]. This shows that the predicted values do not fit the observed values well, for all product concentrations. Except for very high product P concentrations, for the cell size model, the observed values closely reassemble the measured ones. However, at a certain level of product concentration the fit significantly worsens but still being comparable to the cumulative substrate model.

Sugars and Gas Rates Since the measured values of glucose Glu and lactose Lac were below the limit of quantification of the used HPLC method, no observed-vs.-predicted plots can be made. For the galactose Gal no significant differences between the models can be observed, since it is mainly calculated via the shared backbone [Figure 3.10]. However, it is not skewed towards over- or under-estimation, showing that no systematic miss-description of the system is present. The gas rates OUR and CER are on-line measurements; a data point is available every 30 s during the whole cultivation. Thereby, both are calculated via the shared model backbone leading to similar results for the two models [Figure 3.11]. At the very beginning of the experiments, the OUR and CER values needed a few minutes to stabilise (as can be seen in the observed-vs.-predicted plots as a peak in the signal at very low CER values and very high OUR values respectively), after that the measurement gave reliable results. The oxygen uptake rate OUR is systematically underestimated for all experiments, in both models. Hereby, the prediction qualities decreases over the course of the cultivation (lower OUR values are reached in later phases of the cultivation with higher biomass concentrations). That might be caused by the multiple other estimated states used for the prediction of the OUR and the changes in oxygen supply made during the cultivation by the process control software. Thereby, the oxygen supply is adapted in order to keep the dissolved oxygen (DO) in the cultivation media over 30 % to assure no oxygen limitation of the cells. That constant adaptation may cause the systematic under-estimation, since the models do not take it into account. As for the carbon evolution rate CER a similar but mirrored (since one is an uptake rate whereas the other is an emission rate) behaviour is observable. This is unsurprising, due to the strong interconnection between the oxygen demand and carbon dioxide formation.

Impact Factor For the impact factor IF only for the cell size model a observed-vs.-predicted plot can be generated, due to its expanded monitoring strategy. The cumulative metabolised sugar can only be calculated, but not monitored with the monitoring strategy used for the cumulative substrate model. For the relative cell size CS it can be observed that with increased cell size the fit slightly worsens [Figure 3.12].

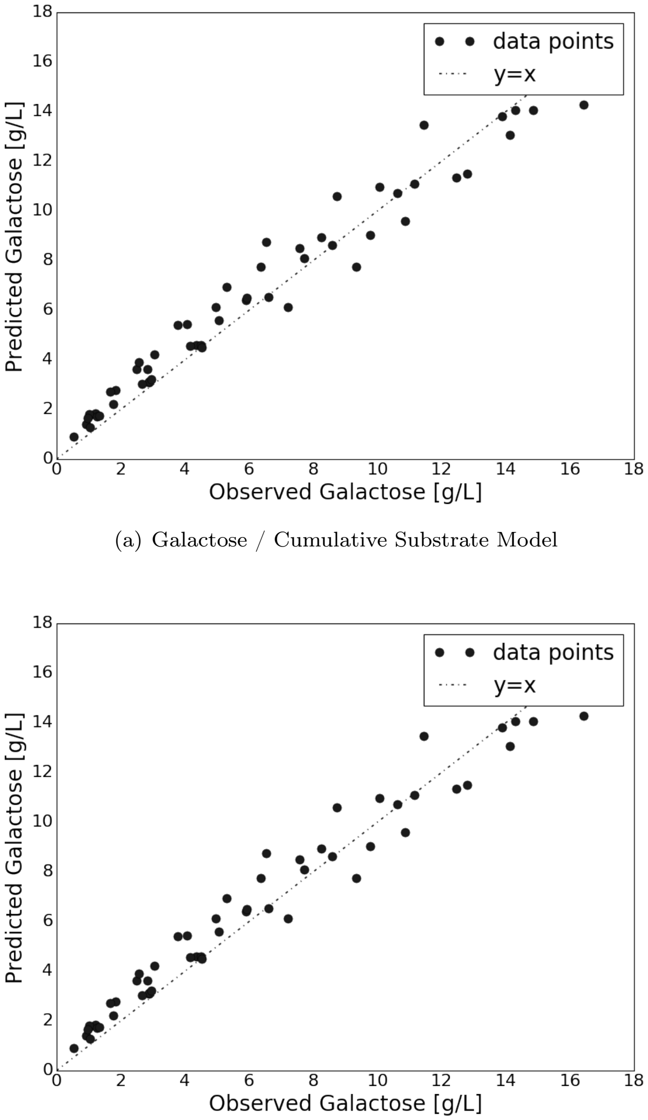


(a) Product / Cumulative Substrate Model



(b) Product / Cell Size Model

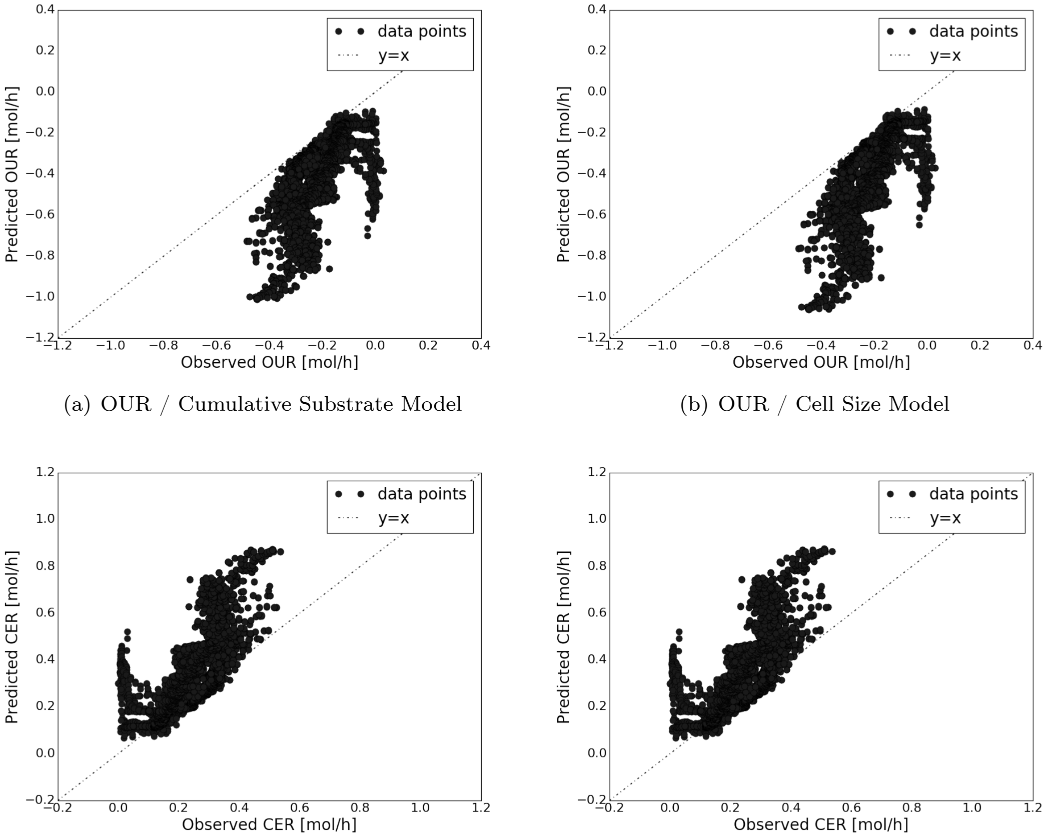
Figure 3.9: The observed values plotted against the predicted values of the product P, represented by dots in the plots, for all performed samples of all experiments for both models respectively The predicted values are on the y-axis and the observed values on the x-axis. The x = y line resembles a perfect prediction of the observed values, represented as dotted line in the plots.



(a) Galactose / Cumulative Substrate Model

(b) Galactose / Cell Size Model

Figure 3.10: The observed values plotted against the predicted values of the galactose *Gal*, represented by dots in the plots, for all performed samples of all experiments for both models respectively The predicted values are on the y-axis and the observed values on the x-axis. The x = y line resembles a perfect prediction of the observed values, represented as dotted line in the plots.



(a) OUR / Cumulative Substrate Model

(b) OUR / Cell Size Model

(c) CER / Cumulative Substrate Model

(d) CER / Cell Size Model

Figure 3.11: The observed values plotted against the predicted values of OUR and CER, represented by dots in the plots, for all performed samples of all experiments for both models respectively The predicted values are on the y-axis and the observed values on the x-axis. The x = y line resembles a perfect prediction of the observed values, represented as dotted line in the plots. Subplots a) and b) show the oxygen uptake rate OUR values for both models respectively. Subplots c) and d) show the carbon evolution rate CER values for both models respectively.

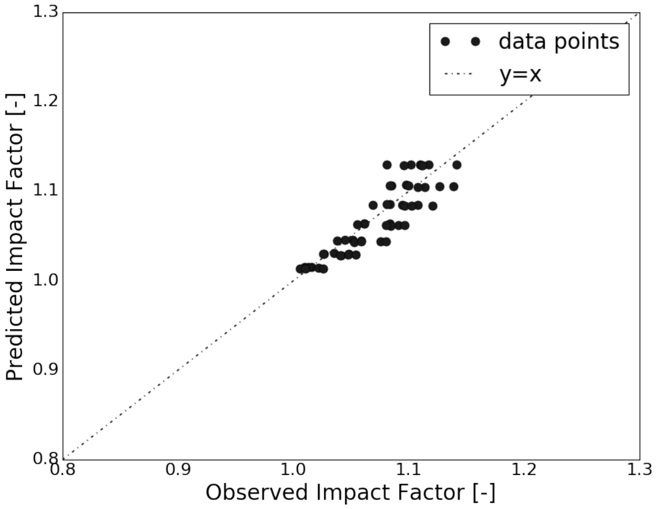


Figure 3.12: The observed values plotted against the predicted values of the impact factor IF (the relative cell size CS), represented by dots in the plots, for all performed samples of all experiments for both models respectively The predicted values are on the y-axis and the observed values on the x-axis. The x = y line resembles a perfect prediction of the observed values, represented as dotted line in the plots.

However, it is not skewed towards over- or under-estimation, showing that the decrease in fit is solely due to biological deviations and not due to a systematic miss-description of the system.

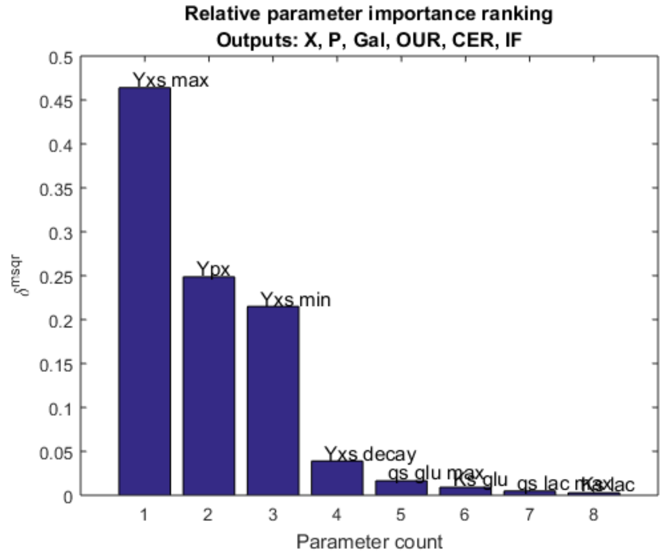
3.3 Parameter Sensitivity and Identifiability

The local sensitivity and structural identifiability are important criteria for model analysis. Sensitivity is thereby the basis for identifiability. In the following chapters the relative parameter importance ranking as well as the local sensitivity matrix of all outputs of the model is analysed for both models. Additionally, the specific local sensitivity time-course of the non-backbone parameters is analysed in detail, separately for all the outputs of the models individually. Furthermore, the structural identifiability of all parameters is calculated in order to asses, whether the local sensitivity of the parameters is sufficient to identify the parameters with the given model structure.

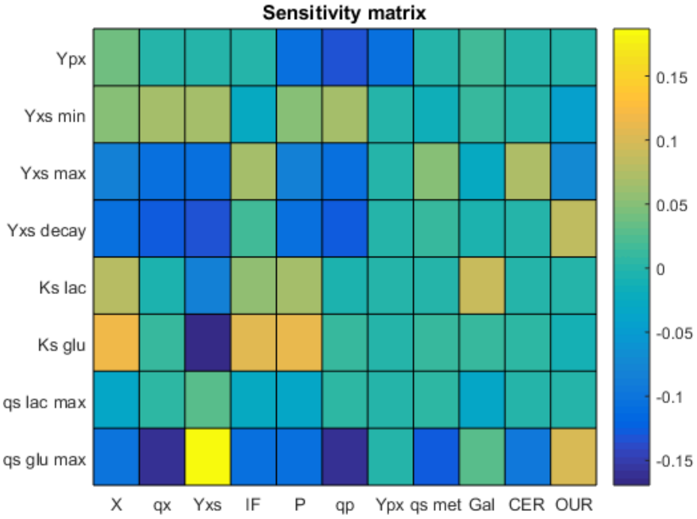
3.3.1 Cumulative Substrate Model

Local Sensitivity For the cumulative substrate model the parameters used in the model backbone for the sugar metabolism are generally less important than the parameters used for the biomass growth and product formation [Figure 3.13]. Since the numeric values of the product P are much higher compared to all the other used states, the relative importance of the $Y_{P/X}$ may be a bit overestimated in the parameter importance ranking (as discussed in Chapter 2.7). As for the sensitivity matrix, unsurprisingly the parameters used for the sugar metabolism show high sensitiveness for the model outputs which describe the sugar uptake of the strain. Apart form that, the results are relatively straight forward. The individual sensitivity of the $q_{S_{Glu},\max}$ on the $Y_{X/S}$ is the highest over all, since glucose is the main energy source for the cells, since everything else is dependent on the energy the cells have available. The parameters for the lactose uptake show a strong sensitivity for the Gal state, since they are directly linked by stoichiometry. The highest sensitivity of the parameter $Y_{P/X}$ is for the specific product formation rate q_P and the product P itself. For the biomass X associated parameters ($Y_{X/S,max}$, $Y_{X/S,crit}$ and $d_{Y_{X/S}}$), the highest sensitivity can be observed for the biomass growth rate q_X, the biomass per substrate yield $Y_{X/S}$ and the biomass X itself . Additionally, a correlation in sensitivity between the biomass X and product P associated parameters can be observed.

The detailed specific local sensitivity time-course of the model specific non-backbone parameters of the cumulative substrate model, shows the change in sensitivity of the parameters for the model outputs individually [Figure 3.14]. Unsurprisingly, the parameters $Y_{X/S,max}$ and $Y_{X/S,crit}$ show the highest sensitivity for the biomass X. Additionally, the parameter $Y_{X/S,max}$ shows a high sensitivity for gas rates OUR and CER. The parameter $d_{Y_{X/S}}$ shows a lower sensitivity for all the outputs than the other two biomass associated parameters. For the product P the parameters $Y_{X/S,max}$ and $Y_{X/S,cirt}$ are also very sensitive, surprisingly even more than the parameter $Y_{P/X}$.



(a) Relative Parameter Importance Ranking / Cumulative Substrate Model



(b) Sensitvity Matrix (5h after induction) / Cumulative Substrate Model

Figure 3.13: Parameter importance ranking for the states X, P, Gal, IF, CER and OUR as well as the local sensitivity matrix in the middle of the induction phase (5h after induction) for the cumulative substrate model.

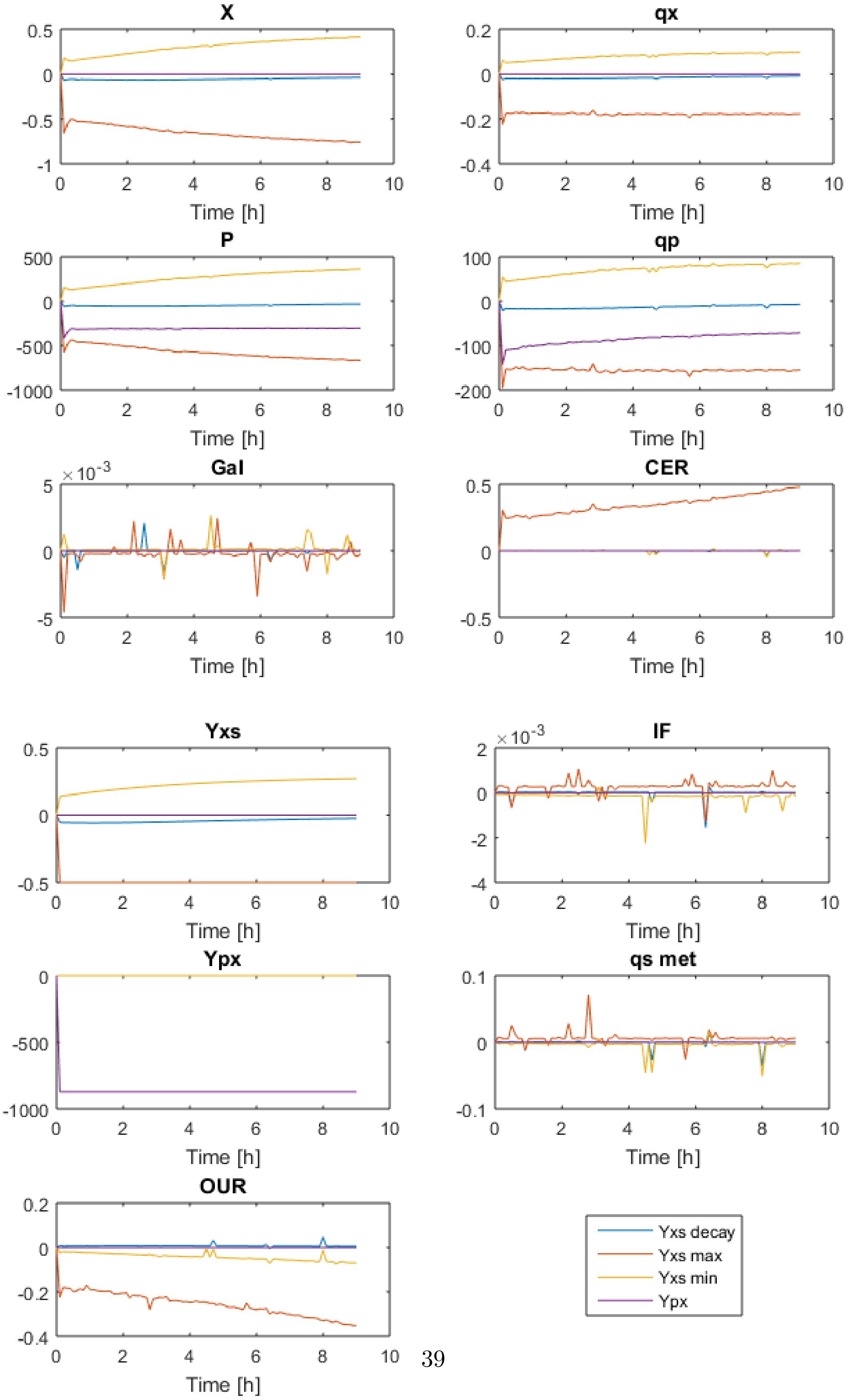


Figure 3.14: Specific local sensitivity time-coures of the 4 non-backbone parameters of the cumulative substrate model.

Table 3.4: Structural identifiability of the non-backbone parameter sets of the cumulative substrate model

Parameter Set	Number of Parameters	Determinant Measure	Collinearity Index (γ)
A			
$Y_{X/S,max}$ $Y_{X/S,crit}$ $Y_{P/X}$	3	3.6073	5.6244
B			
$d_{Y_{X/S}}$ $Y_{X/S,max}$ $Y_{P/X}$	3	0.14198	9.7658

This additionally supports the assumption that the importance of the parameter $Y_{P/X}$ is overestimated in the relative parameter importance ranking. For the impact factor IF and the galactose Gal none of the parameters have an especially high impact, since for the cumulative substrate model these states are mainly calculated with the backbone.

Structural Identifiability For the structural identifiability a threshold for the collinearity index (γ) of below 10 is applied according to [47]. The determinant measure can be calculated as measure for the amount of information obtainable by identification of that parameter set. The parameter sets are sorted first by the biggest number of parameters identifiable at the same time, and second by the highest determinant measure which are still below the collinearity threshold. Only the parameter sets with the highest determinant measure with previously unidentifiable parameters are shown. Sets with a lower determinant measure containing only parameters that can be identified in sets with a higher determinant measure are not shown. Furthermore, only the model specific non-backbone parameters are analysed. The structural identifiability of the cumulative substrate model shows that only three out of four parameters are structurally identifiable at the same time [Table 3.4]. Unfortunately, it is not possible to identify all four non-backbone parameters at the same time whilst being under the collinearity index (γ) threshold of 10.

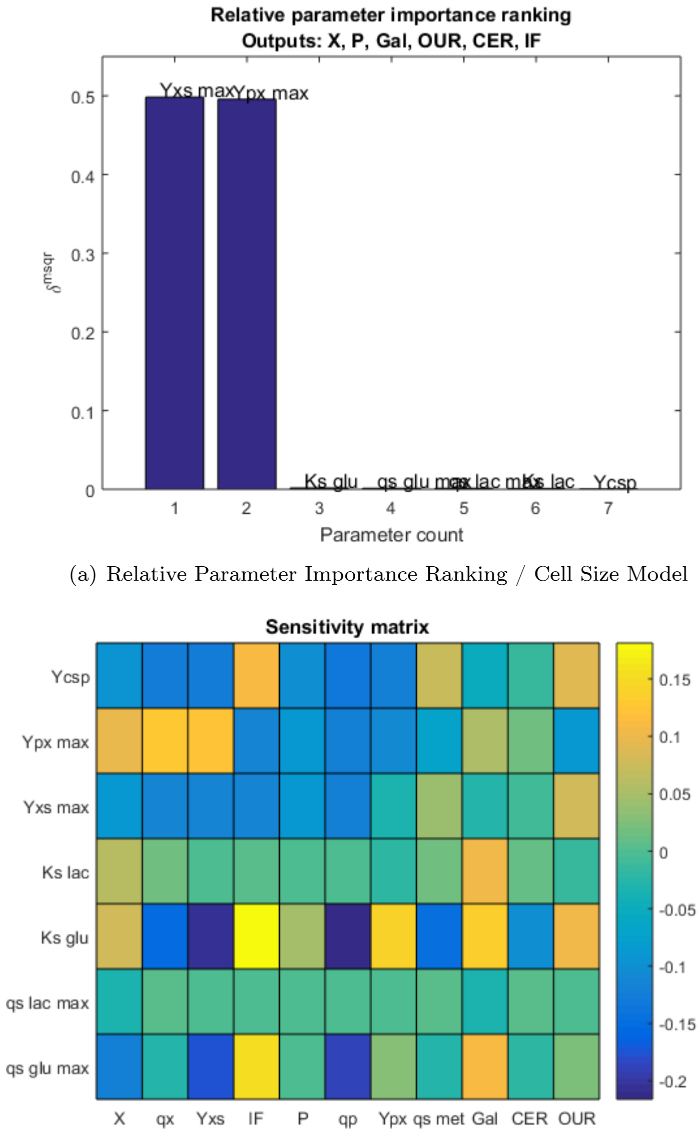
Table 3.5: Structural identifiability of the non-backbone paramter sets of the cell size model

Parameter Set	Number of Parameters	Determinant Measure	Collinearity Index (γ)
A			
$Y_{X/S,max}$ $Y_{CS/P}$ $Y_{P/X}$	3	0.17699	9.273

3.3.2 Cell Size Model

Local Sensitivity For the cell size model, the parameters used in the model backbone for the sugar metabolism again are generally less important than the parameters used for the biomass growth and product formation [Figure 3.15]. Of the three parameters used additionally to the backbone parameters, two ($Y_{X/S,max}$ and $Y_{P/X,max}$) have a very high relative importance. The third parameter, the cell size per product formation yield $Y_{CS/P}$, is the least important parameter in this ranking. That may be due to the small numeric value of the relative cell size CS combined with the small change in its value during the cultivation, leading to an underestimation in the unscaled relative parameter importance ranking. The high sensitivity of the of the parameter $Y_{CS/P}$ in the sensitivity matrix supports the assumption of underestimation in the relative parameter importance ranking. Furthermore, the parameters $Y_{X/S,max}$ and $Y_{P/X,max}$ have a high sensitivity for many model outputs, since they are interlinked to each other in the equations 3.21-3.23. The sensitivity matrix for the backbone parameters involved in the sugar metabolism give comparable results for both models.

The detailed specific local sensitivity time-course of the non-backbone parameters of the cell size model, shows the change of the sensitivity of the parameters for the model outputs individually [Figure 3.16]. Out of the three non-backbone parameters of the cell size model, only the parameter $Y_{X/S,max}$ is sensitive for the biomass and only the parameter $Y_{P/X,max}$ is sensitive for the product. The $Y_{P/X,max}$ and the $Y_{CS/P}$ have inverse sensitivities for the impact factor IF. Additionally, the parameter $Y_{X/S,max}$ is also sensitive for the OUR.



(a) Relative Parameter Importance Ranking / Cell Size Model

(b) Sensitvity Matrix (5h after induction) / Cell Size Model

Figure 3.15: Parameter importance ranking for the states X, P, Gal, IF, CER and OUR as well as the local sensitivity matrix in the middle of the induction phase (5h after induction) for the cell size model.

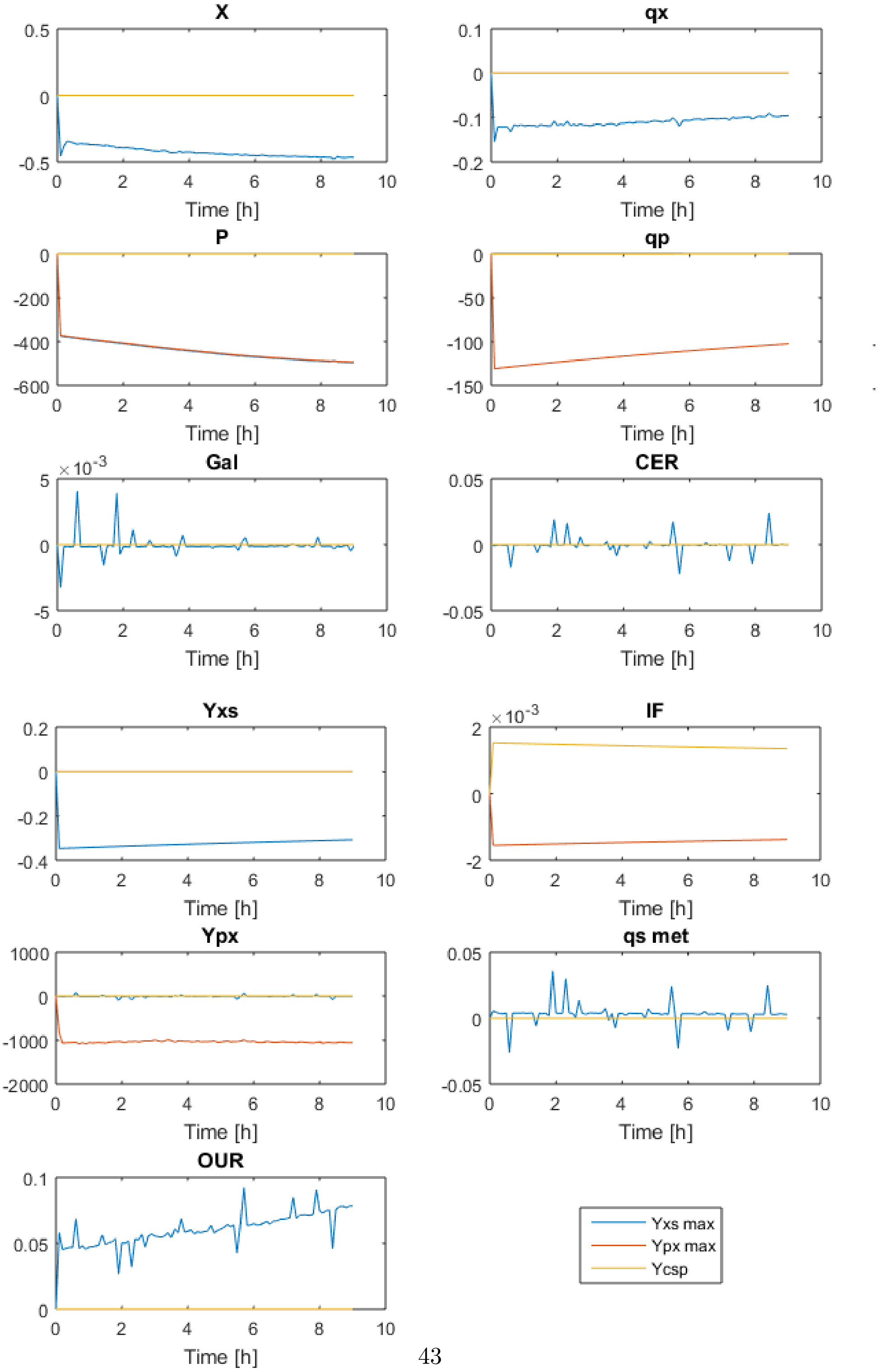


Figure 3.16: Specific local sensitivity time-coures of the 3 non-backbone parameters of the cell size model.

Structural Identifiability Concerning the structural identifiability, it is possible to identify all 3 non-backbone parameters at the same time while being below the collinearity threshold of 10 [47] for the cell size model [Table 3.5]. Additionally, the cell size model only needs three model specific non-backbone parameters for the description of the system whereas the cumulative substrate model relies on four [Table 3.1]. This allows for a greatly improved identifiability of the cell size model compared to the cumulative substrate model.

4 Discussion

For this study, a new model featuring the novel method of modelling the physiologic changes via the specific production rate q_P, combined with the monitoring of the physiologic changes (the relative cell size CS in *E. coli*), was compared to an existing approach of modelling the metabolic stress via the substrate consumption q_S. By using the specific production rate q_P instead of the specific substrate uptake rate q_S to model the metabolic stress, stress caused by the production of recombinant proteins [9] is getting explicitly accounted for. Furthermore, motoring the physiologic changes of the cells (the increase in relative cell size CS in *E. coli*), provides the ability to validate the predictions of the model, about the incurred metabolic stress, on-line during the cultivations. This offers a increased ability to refit model parameters during the cultivation, and thereby recalibrate the model during the process.

In the following, the physiological background of both models is discussed and compared. Furthermore, the performance of the models as well as their real-time applicability is assessed.

4.1 Evaluation of Model Performance

4.1.1 Quality of Fit

The quality of fit of the model was analysed by the normalised root mean square error (NRMSE). For the triplicate centre point, a significantly better description of the biomass growth could be derived with the cell size model, while the description of the other states gave comparable results. However, both models were describing the system very well ($<$ 11 % NMRSE). The standard deviation of the NMRSE values of the triplicate centre point shows an extremely good reproducibility of the description of the system ($<$ 3 % standard deviation of NMRSE). For all the DoE experiments, the NMRSE values for both models increases due to the broader set of conditions the models have to describe accurately. Hereby, the NMRSE values of the biomass X and product P are significantly lower for the cell size model, while the description of the other states gave comparable results. The cell size model shows an impressive capability to describe the system for a broad set of conditions simultaneously ($<$ 11 % NMRSE). The description of the system by the cumulative substrate model is still acceptable (< 20 % NMRSE) but cannot match the predictive qualities of the cell size model (< 9 % standard deviation of NMRSE compared to < 15 %). Overall the NMRSE and StDev values of the cell size model are lower compared to the cumulative substrate model. Thereby the accuracy of the description of the system is sufficient enough to be used for process control. For the cumulative substrate model, however, only the description of the triplicate centre point is accurate enough to consider it for process control. This shows that only the cell size model is able to describe the system accurately enough to facilitate process control for a broader set of conditions.

The time-resolved plots show a qualitatively better description of the behaviour of the measurements for the biomass X by the cell size model. However, the description of the cumulative substrate model gives a closer match for the product P. Those differences occur due to the different curve behaviours obtained by the different equations used for the biomass growth and product formation in both models [Equations (3.15) - (3.17) and (3.22) - (3.23) respectively]. The description of galactose Gal and the gas rates (OUR and CER) are very similar due to being mainly calculated via the shared identical backbone. The fitted values of the biomass per substrate yield $Y_{X/S}$ of the cell size model, povide a better description of the system than the fixed values of the cumulative substrate model, which assumes this values a physiological constants. The product by biomass yield $Y_{P/X}$ surprisingly is equally linear for both models. However, the cell size model does not require a constant $Y_{P/X}$, offering the possibility to also accurately describe systems where a time varying nature of the $Y_{P/X}$ is reported [58].

The observed-vs.-predicted plots offer a way to asses systematic missestimations of the system by the models. For the biomass X the cumulative substrate model systematically underestimates the low biomass concentrations at the beginning of the experiments slightly, and systematically overestimates the highest biomass concentrations at the end of the cultivations. For the cell size model no such effect is observable, further underscoring the strength of accuracy of the biomass prediction by the cell size model. For the other states the results of the observed-vs.-predicted plots are comparable. The only state where a systematic missestimation is observable is the OUR. In this case the systematic overestimation increases during the experiment, probably caused by the changes made in oxygen input into the reactors by the process control software in order to keep the dissolved oxygen (DO) above 30 % to ensure no oxygen limitation of the cells. Not having any significant missestimations of the system is required for the model to be considered for process control.

4.1.2 Quality of Model Structure

The ranking of the parameter importance (δ) generally shows that the backbone parameters for the sugar metabolism are of less importance for the models than the non-backbone biomass- and product-related parameters. Due to the use of an unscaled parameter importance ranking, the importance of the product related parameters might be a bit overestimated (due to their very high numeric values) while the cell size related parameters might be underestimated underestimated (due to their very low numeric values). The sensitivity matrix shows a high sensitivity of the glucose related parameters, due to glucose being the main energy source of the cells, while the lactose related parameters are of less importance. The non-backbone parameters show a high sensitivity for biomass X and product P in both models. The identifiability analysis shows that for the cell size model all model specific non-backbone parameters can be identified at once within the given collinearity index (γ) threshold. Therefore, a improved structural identifiability, due to the different model structure of the cell size model, can be achieved compared to the cumulative substrate model, were two parameters cannot be structurally identifiable with the given model structure in a single experiment.

4.1.3 Real-Time Applicability

Cumulative Substrate Model The cumulative substrate model only describes the system accurately enough to be applied in real-time for process control for the triplicate centre point. The main advantage of this method regarding its real-time applicability is the low amount of measurements (OUR and CER only), the model needs as feedback of the current state of the process. Both of these measurements are simple on-line measurements without any manual labour needed. Furthermore, OUR and CER are very commonly monitored process parameters were most bioreactors already are suitably equipped to monitor them [37, 61, 54, 8, 12]. Apart form that, only some strain specific parameters and the amount of sugar fed to the reactor are needed. The broad availability of the needed monitoring infrastructure makes a simple incorporation into new as well as existing bioprocesses possible. Additionally, only a few equations are needed to added to existing mechanistic models, in order to implement the cumulative substrate approach to model the strain performance decay, in existing mechanistic models [36, 35, 34, 33].

However, this method has it's physiological drawbacks (regarding problems of accurately describing biologically more complex metabolic stress causing factors). Nevertheless, due to its simplicity, the need of no hardware and only few equation changes, this method provides work in-intensive and easy to implement way of describing the performance decay, as a first step. If it is not able to provide the desired accuracy of the description of the performance decay of the cells, more specific methods can still be applied without much work being lost.

Cell Size Model The accuracy of the description of the cell size model is sufficient to be used for process control in real-time for the entire DoE. The performance decline of the strain is calculated using the specific product formation rate q_P instead of the specific substrate uptake rate q_S. Additionally, to OUR and CER, the cell size CS is measured with flow cytometry. By using the relative cell size to determine the physiologic state of the cells, automatised on-line flow cytometry devices can provide the possibility to quantitatively and qualitatively asses the biomass within the bioreactor in a single measurement without manual labour required [23, 24]. Hereby, additional valuable information about the current state of the bioprocess can be generated without additional manual labour required during the process.

This method allows for a much more psychophysically meaningful description of the performance decay, but a much deeper integration of the additionally needed equations into the existing set of equations is necessary. The expansion of this method for more special cases (like inclusion body formation or toxic products for example) is possible, though requries more experiments and modelling work to set up. Furthermore, the ability to asses the physiologic state of the cells offers greater possibilities for the prediction of the behaviour of the cells. However, these benefits are tied to the ability to monitor the physiologic state of the cells via their cell size, which may only work for *E. coli* [21, 22]. Nevertheless, even without the ability to monitor the physiologic state of the cells, the approach to model the performance decay via the actual product formation and not via the sugar uptake, can be valuable. An accurate description of the product formation q_P is a prerequisite though, however, that is a commonly expected feature for mechanistic models for bioprocesses anyway.

4.2 Comparison of the two Models

4.2.1 Cumulative Substrate Model

The first modelling approach is based upon the assumption that all stress causing factors during the production of recombinant proteins, that lead to the performance decay, can be described solely by the amount of sugar fed during producing conditions. The thereby employed physiological background is, that the fed sugar is the primary (and almost exclusive) source of energy for the cells. Naturally, the cells need that energy for all cell functions, from cell growth to the production of the recombinant protein. Yields, such as the biomass per substrate yield $Y_{X/S}$, are then used to break down, how much energy is used for what [58, 32]. Thereby, the amount of energy used for metabolic stress causing pathways, such as the production of the recombinant protein, is circumscribed by the total available energy to the cells [48]. That approach is valid given a constant product per substrate yield $Y_{P/S}$ (or as in this case $Y_{X/P}$), to be able to correlate the amount of sugar fed to the amount of metabolic stress incurred (each gram of sugar stands for a certain amount of metabolic stress caused in cells under producing conditions). Here the specific amount of sugar metabolised per gram of cells $q_{S_{met}}$ is used to quantify the amount of energy available for the cells, and thereby how much metabolic stress the cells are exposed to. Since the performance decay is a result of the load of metabolic stress over time, $q_{S_{met}}$ is cumulated up during the production of the recombinant protein ($S_{met,cum}$) to model the performance decay fo the cells [48]. The metabolic stress then causes an increase in energy (sugar) needed for the production of new cells (as stressed cells need more energy to grow and divide into new cells), described with a decrease in $Y_{X/S}$ [10].

The simplification of the factors causing the metabolic stress by the amount of sugar (energy) consumed provides several benefits and drawbacks. First of all, the biggest benefit of that method is its simplification of the problem; not all stress causing factors need to be exactly known and quantifiable, since all the effects are summed up by the energy that was utilised during producing conditions [10]. This is leading to a high transferability of this method between different processes, strains and organisms, since only a low amount of knowledge is needed, due to the very broad nature of the description.

However, the biggest drawbacks of that method also are caused by its simplification of the problem; the substitution of all stress causing factors by the amount of sugar consumed lacks in depth of the physiological description of the system. The performance decay of the cells under producing conditions, compared to non-producing cultivations, arises due to the production of the recombinant proteins, and not directly from the sugar uptake [9]. The substitution of all stress causing factors by the amount of sugar consumed, is only valid as long as a constant product per substrate yield $Y_{P/S}$ is given (however, a decrease in performance is also commonly observable in the production and not only the growth of the cells [58, 9, 10]). As mentioned above, for this method each gram of sugar stands for a certain amount of metabolic stress its consumption causes in cells under producing conditions. Hereby, 'producing conditions' suggest that the conditions for the cells are always the same during the whole production phase of the recombinant proteins. However, as the performance decay shows, there are physiological changes over the course of the cultivation [11, 12, 13, 21, 22]. Furthermore, other factors may also vary the amount of stress incurred by the cells per gram of sugar.

First of all, the feed rate (among other factors) has an direct impact on the growth and productivity of the cells (independent of the total amount fed) [59, 60, 58]. Also the production of inclusion bodies causes additional stress for the cells [14, 15, 17]. IB production in cells occurs for example when:

i) the cells are unable to fold the protein correctly
ii) the protein folding machinery of the cells gets overwhelmed (due to the strong artificial withdrawal of cell resources towards the production of the recombinant protein)
iii) or when the concentration of the product is surpassing its solubility in the cells.

For the last case, sugar consumed before and after the solubility of the product is exceeded, will not result in an equal amount of metabolic stress incurred by the cells. Another example, would be products, which have a certain toxicity towards the cells; lower concentrations are probably less harmful for the cells than higher ones, resulting in a higher amount of metabolic stress during their production, when their concentration is already greater.

In summary, the principal methodology applied may be very simple and therefore easy transferable to other processes, strains and organisms, but the method itself has several strong limitations [Table 4.1]. Furthermore, it's predictive power suffers from broad set of different process conditions, since they are not explicitly accounted for in the equations. It works well as long as time changing effects, as well as the effects caused be changed process conditions within the used design space on which the model is applied to, are minimal and can be neglected. Still, this method is used for various processes in literature [32, 48].

4.2.2 Cell Size Model

For the second approach, the cell size is measured as parameter to characterise the physiological state of the cells. For *E. coli* it is reported that high metabolic stress leads physiologic changes related to cell growth & division [21]. Hereby, an increase of the average cell size, cells have when they undergo cell division is reported [21]. Since this effect is attributed to metabolic stress [21], the cell size can be used to monitor the incurred metabolic stress, providing a way to qualitatively and not only quantitatively characterise the biomass. Here, the specific product formation rate q_P is used to predict the amount of metabolic stress the cells incur, since the production of recombinant proteins is causing a lot of stress for the cells [9]. Thereby, the predicted relative cell size increase CS, calculated with q_P, can be verified by monitoring the actual cell size with flow cytometry [23, 24].

By using the specific product formation rate q_P instead of the sugar uptake q_S during producing conditions, the stress caused by the production of recombinant products is explicitly accounted for. The increased depth of the description comes with its own benefits and drawbacks. First of all, the $Y_{P/S}$ (or as in this case the $Y_{P/X}$) does not need to be constant, which is commonly more realistic for *E. coli* cultivations [58, 9, 10]. Furthermore, because the specific product formation rate q_P itself is taken to quantify the amount of incurred metabolic stress, differences in production conditions are accounted for with this method, since q_P is influenced by the exact conditions used [18, 20, 17]. As a result, significantly less limitations on the size of the desired design space need to be made in order to keep the method valid. Furthermore, it offers the possibility to incorporate different producing conditions into the used design space, to specifically analyse their effects on the metabolic stress levels of the cells, which may be of high interest to researchers.

Additionally, when using q_P instead of q_S, product concentration dependent effects (regarding IB formation and toxic products for example) can also be accounted for, by adaptations of the current model, when needed. However, the method relies on the product formation being the main (the only changing) source of metabolic stress. Metabolic stress caused for example by the formation of by-products or due to high cell densities, are not specifically accounted for. In the current method they need to be treated as constant or neglectable and would require additional work to implement.

Table 4.1: Model comparison

	Cumulative Substrate Model	Cell Size Model
Feedback Variable	q_S	q_P
Description of the System	Simple Unspecific Description	Detailed Product Related Descpription
Monitorable Metabolic Stress State	No	Yes
Adaptability to Special Cases	Limited	High
Transferability to Different Organisms	High	Limited
Centre Point Fit	< 11 % NMRSE < 2 % StDev of NMRSE	< 10 % NMRSE < 3 % StDev of NMRSE
Full DoE Fit	< 20 % NMRSE < 15 % StDev of NMRSE	< 11 % NMRSE < 9 % StDev of NMRSE
Parameter Identifiability	2 Sets of 3 Parameters	All Parameters at Once

Transferring the methodology to different organisms would be more complicated than for the cumulative substrate method; it is based on the cell size increase as physiologic reaction to the incurred metabolic stress of *E. coli* cells [21], which will not be the case for all commonly used production host. The cell size state per se could be replaced by a generic metabolic stress state [9], but the benefit of being a able to monitor that state would be lost. Monitoring the stress reactions of different organisms may be difficult, cost and/or effort intensive or simply impossible with available methods depending on the organism used [21, 23, 50, 60].

Summed up, the used methodology provides a more physiological depth, with the possibility of further expansion of the model for more specific cases [14, 15, 17] [Table 4.1]. Additionally, it provides the ability to monitor and analyse different producing conditions in regards to metabolic stress. However the transfer to different organisms apart from *E. coli* could be very difficult and dependent on a good monitorable parameter for the physiological state of the cells [21, 23, 50, 60]. In literature mainly kinetic models that model the changes of the cell physiology [49, 13] are used. However, those models also rely on the specific substrate uptake rate q_S to model the physiologic changes. Additionally, also kinetic models modelling the impact of the production of recombinant proteins on the cell growth are published [9]. However, there is a lack of mechanistic models combing the production of recombinant proteins with the physiologic changes of the cells, and the monitoring opportunities that come with them, to facilitate process control.

4.2.3 Differences Compared to Literature

The decrease in metabolic performance during cultivations is commonly reported [48, 32, 10, 9, 38, 60, 58]. Thereby the optimal performance of the cells is reported to be decreased by the formation of by-products such as acetate [59, 62], but also the production of recombinant proteins itself [9, 60, 58]. In order to address the performance decay some work has been done to track the maximal culture capacity and the change of physiologic parameters, such as q_S, over the course of the cultivation [48, 10, 38, 63]. Furthermore, the impact on the yields, such as the biomass per substrate yield $Y_{X/S}$, has been assessed in detail [32, 58, 59, 38, 64]. Hereby, the cumulative metabolised substrate $S_{met,cum}$ is a regularly used variable to model metabolic stress [48, 32]. Kinetic models using the specific substrate uptake rate q_S in order to model the physiologic changes are published [49, 13]. Furthermore, kinetic models that model the impact of the production of recombinant proteins on the cell growth are reported [9].

However, there is a lack of mechanistic models using the specific product formation rate q_P to model physiologic changes caused by the production of recombinant proteins. None of the mentioned models combine the modelling of the impact of the production of recombinant proteins with the modelling of the physiologic changes of the cells, as the cell size model does. The novelty of this work thereby is, the combination of these modelling goals (impact of the production of recombinant proteins, physiologic changes of the cells) with a suitable monitoring strategy (exhaust gas rates, flow cytometry), which can provide new possibilities for model predictive control (MPC).

5 Conclusion

The goal of this contribution was to develop and compare different methodological approaches to mechanistically model the performance decay of *Escherichia coli* cells during cultivations. The two compared approaches were:

i) The cumulative substrate model: a simple, easy to implement method, with needs only a minimal monitoring effort; abstracting the metabolic stress incurred by the utilised energy during producing conditions via the respective sugar consumption $q_{S_{met}}$
ii) The cell size model: a more detailed method, with a sightly expanded monitoring strategy; using the product formation rate q_P to model the incurred amount of metabolic stress by the cells, due to the production of recombinant proteins, instead of an abstracted value

5.1 Model Development

The adapted version of the model described in [32], as well as the newly developed model both only relied on:
i) strain specific constants, which can be measured (like the sugar uptake rates)
ii) the initial values of the states at the start of the bioprocess (like the initial biomass in the reactor)
iii) the input values of the feed-rates for sugars, base and oxygen which they control
iv) as few as possible measured outputs (like the gas rates or the cell size) as feedback of the current state of the bioprocess, in order to keep the monitoring effort low

The newly developed model could achieve:
i) to take the physiological state of the cells into account
ii) to model the cellular metabolic stress via the specific product formation rate q_P
iii) to expand the monitoring strategy only by measuring the cell size with flow cytometry to not inflate the overall monitoring effort
iv) a better description of the process with lower NMRSE and StDev values
v) a better sensitivity and identifiability of the involved model parameters

5.2 Model Comparison

5.2.1 Quality of Fit

The cumulative substrate model is able to derive a valid description the system, despite its simplifications, as shown in [48] and this study. Thereby, the simplicity of this method leads to a high transferability of the method to other processes, strains and organisms. However, its limitations regarding changing physiologic and process conditions, are restricting the size of the design space it can accurately describe. This could be observed in this study, by a stronger increase in NMRSE values as well as their standard deviation when looking at the entire DoE instead of just the centre point. Furthermore, due to its assumptions that each gram of sugar corresponds to a certain amount of incurred metabolic stress, it might fail in metabolically more complex cases.

With the cell size model it could be successfully shown that previously reported changes in the physiology of *E. coli* [23, 22], which can be attributed to metabolic stress [21], can be used for process control with mechanistic models. The approach to predict the incurred metabolic stress via the product formation rate q_P, instead of the sugar uptake rate q_S, led to an increased capability of the model to accurately describe the system for broader design spaces. Additionally, it provides the ability to asses process conditions in regards to metabolic stress specifically, and the possibility for an expansion/adoption to cope with physiologically more complex cases.

5.2.2 Model Structure

With local sensitivity and structural identifiability analysis it could be proven, that for both approaches, valid models can be generated with the respective monitoring strategy used. A reduction in parameters for the description of biomass growth and product formation (from 4 in the cumulative substrate model to 3 in the cell size model) could be achieved. Thereby, the structure of the cell size model, and the local sensitivity of its parameters, made the simultaneous structural identifiability of all the model-specific non-backbone parameters possible. By the expansion of the monitoring strategy for the cell size model, the physiologic state of the cells, could be added as qualitative parameter of the biomass instead of just being able to quantify it. Published automatised on-line flow cytometry devices [24], could thereby monitor the quantity and the physiologic state of the biomass, without manual labour within a single measurement. Combined with modelling approaches, powerful tools or bioprocess control for *E. coli* can be derived using this new methodology.

5.3 Summary and Outlook

It can be concluded that the cell size model is superior to the cumulative substrate model for *E. coli*, with the trade-off of a single additional state that needs to be monitored, due to:

i) the more accurate description of the system with lower NMRSE and standard deviation of NMRSE values for a broad set of conditions (< 11 % NMRSE instead of < 20 % and < 9 % StDev of NMRSE instead of < 15 % for the full DoE fit)
ii) the increased structural identifiability of its parameters (all parameters are structurally identifiable at once instead of only in 2 sets)

The trade-off of having an additional monitored state is more than worth while, since additionally, the measurement of the physiological state of the cells used, and thereby the decline in specific performance, offers additional insight and new possibilities for model development and process control. However, the simplicity and transferability of the cumulative substrate method and the low amount of effort needed to implement it into existing models, makes it ideal as first try in modelling the performance decay in mechanistic models.

More accurate predictions by mechanistic models of industrial relevant bioprocesses, due to an increased insight in cell physiological reactions, can be used to decrease the formation of by-products, to more reliably reach the needed product quality as well as to increase the overall space-time-yields of production plants. Furthermore, the provided methods for process control as well as the increased amount of knowledge generateable with then, can help implementing the regulatory authorities initiatives such as PAT and QbD into industrial bioprocesses. Nervertheless, there is still a lot of work to do, to apply mechanistic models which take the physiologic state of the cells into account, like the cell size model, for bioprocess control. Hereby, the developed cell size model needs to be tested with model predictive control, in order to compare its control capabilities to commonly applied standard control strategies.

References

[1] Christoph Herwig. “Prozess analytische technologie in der biotechnologie”. In: *Chemie Ingenieur Technik* 82.4 (2010), pp. 405–414.

[2] Food, Drug Administration, et al. “Guidance for industry: PAT—A framework for innovative pharmaceutical development, manufacturing, and quality assurance”. In: *DHHS, Rockville, MD* (2004).

[3] US Food, Drug Administration, et al. “Guidance for industry: Q8 (R2) pharmaceutical development”. In: *Center for Drug Evaluation and Research* (2009).

[4] US Food, Drug Administration, et al. “Guidance for industry: Q9 Quality risk management”. In: *Bethesda, MD* (2006).

[5] ICH Harmonised Tripartite Guideline. “Q10: Pharmaceutical Quality System”. In: *International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use*. 2008.

[6] Burkhard Wilms et al. “High-cell-density fermentation for production of L-N-carbamoylase using an expression system based on the Escherichia coli rhaBAD promoter”. In: *Biotechnology and Bioengineering* 73.2 (2001), pp. 95–103.

[7] Xin Chen et al. “Optimization of glucose feeding approaches for enhanced glucosamine and N-acetylglucosamine production by an engineered Escherichia coli”. In: *Journal of industrial microbiology & biotechnology* 39.2 (2012), pp. 359–365.

[8] Patrick Wechselberger et al. “Efficient feeding profile optimization for recombinant protein production using physiological information”. In: *Bioprocess and biosystems engineering* 35.9 (2012), pp. 1637–1649.

[9] William E Bentley et al. “Plasmid-encoded protein: the principal factor in the “metabolic burden” associated with recombinant bacteria”. In: *Biotechnology and bioengineering* 35.7 (1990), pp. 668–681.

[10] Wieland N Reichelt et al. “Physiological capacities decline during induced bioprocesses leading to substrate accumulation”. In: *Biotechnology journal* 12.7 (2017).

[11] Frank Delvigne and Philippe Goffin. “Microbial heterogeneity affects bioprocess robustness: Dynamic single-cell analysis contributes to understanding of microbial populations”. In: *Biotechnology journal* 9.1 (2014), pp. 61–72.

[12] Paul Kroll, Ines V Stelzer, and Christoph Herwig. “Soft sensor for monitoring biomass subpopulations in mammalian cell culture processes”. In: *Biotechnology letters* 39.11 (2017), pp. 1667–1673.

[13] Covadonga Quirós et al. “Quantitative approach to determining the contribution of viable-but-nonculturable subpopulations to malolactic fermentation processes”. In: *Applied and environmental microbiology* 75.9 (2009), pp. 2977–2981.

[14] Fredrik Wållberg et al. “Monitoring and quantification of inclusion body formation in Escherichia coli by multi-parameter flow cytometry”. In: *Biotechnology letters* 27.13 (2005), pp. 919–926.

[15] Gareth Lewis et al. “The application of multi-parameter flow cytometry to the study of recombinant Escherichia coli batch fermentation processes”. In: *Journal of Industrial Microbiology and Biotechnology* 31.7 (2004), pp. 311–322.

[16] Hauke Lilie, Elisabeth Schwarz, and Rainer Rudolph. “Advances in refolding of proteins produced in E. coli”. In: *Current opinion in biotechnology* 9.5 (1998), pp. 497–501.

[17] David J Wurm et al. “Teaching an old pET new tricks: tuning of inclusion body formation and properties by a mixed feed system in E. coli”. In: *Applied microbiology and biotechnology* 102.2 (2018), pp. 667–676.

[18] David Johannes Wurm et al. “The E. coli pET expression system revisited—mechanistic correlation between glucose and lactose uptake”. In: *Applied microbiology and biotechnology* 100.20 (2016), pp. 8721–8729.

[19] David J Wurm et al. “Mechanistic platform knowledge of concomitant sugar uptake in Escherichia coli BL21 (DE3) strains”. In: *Scientific Reports* 7 (2017), p. 45072.

[20] Julian Kopp et al. “Impact of Glycerol as Carbon Source onto Specific Sugar and Inducer Uptake Rates and Inclusion Body Productivity in E. coli BL21 (DE3)”. In: *Bioengineering* 5.1 (2017), p. 1.

[21] Heléne Sundström et al. “Segregation to non-dividing cells in recombinant Escherichia coli fed-batch fermentation processes”. In: *Biotechnology letters* 26.19 (2004), pp. 1533–1539.

[22] Stefan Junne et al. “Electrooptical monitoring of cell polarizability and cell size in aerobic Escherichia coli batch cultivations”. In: *Journal of industrial microbiology & biotechnology* 37.9 (2010), pp. 935–942.

[23] Rui Zhao, Arvind Natarajan, and Friedrich Srienc. “A flow injection flow cytometry system for on-line monitoring of bioreactors”. In: *Biotechnology and bioengineering* 62.5 (1999), pp. 609–617.

[24] Tobias Broger et al. “Real-time on-line flow cytometry for bioprocess monitoring”. In: *Journal of biotechnology* 154.4 (2011), pp. 240–247.

[25] Elena Soriano et al. “Optimization of recombinant protein expression level in Escherichia coli by flow cytometry and cell sorting”. In: *Biotechnology and bioengineering* 80.1 (2002), pp. 93–99.

[26] Sara Alfasi et al. “Use of GFP fusions for the isolation of Escherichia coli strains for improved production of different target recombinant proteins”. In: *Journal of biotechnology* 156.1 (2011), pp. 11–21.

[27] Karl Schügerl. “Progress in monitoring, modeling and control of bioprocesses during the last 20 years”. In: *Journal of Biotechnology* 85.2 (2001), pp. 149–173.

[28] Paul Kroll et al. “Workflow to set up substantial target-oriented mechanistic process models in bioprocess engineering”. In: *Process Biochemistry* 62 (2017), pp. 24–36.

[29] Christoph Herwig et al. “Knowledge management in the QbD paradigm: manufacturing of biotech therapeutics”. In: *Trends in biotechnology* 33.7 (2015), pp. 381–387.

[30] World Health Organization. “Weinstein MC, O’Brien B, Hornberger J, Jackson J, Johannesson M, McCabe C, and Luce BR. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices—Modeling Studies”. In: *Value Health* 6.1 (2003), pp. 9–17.

[31] Paul Kroll et al. “Model-Based Methods in the Biopharmaceutical Process Lifecycle”. In: *Pharmaceutical research* 34.12 (2017), pp. 2596–2613.

[32] Sophia Ulonska et al. “Model predictive control in comparison to elemental balance control in an E coli fedbatch”. In: *CURRENTLY UNDER REVIEW* (2018).

[33] O Grigs et al. “Model predictive feeding rate control in conventional and single-use lab-scale bioreactors: a study on practical application”. In: *Chemical and biochemical engineering quarterly* 30.1 (2016), pp. 47–60.

[34] Mathias Aehle et al. “Increasing batch-to-batch reproducibility of CHO-cell cultures using a model predictive control approach”. In: *Cytotechnology* 64.6 (2012), pp. 623–634.

[35] M Kawohl, T Heine, and R King. “Model based estimation and optimal control of fed-batch fermentation processes for the production of antibiotics”. In: *Chemical Engineering and Processing: Process Intensification* 46.11 (2007), pp. 1223–1241.

[36] Stephen Craven, Jessica Whelan, and Brian Glennon. “Glucose concentration control of a fed-batch mammalian cell bioprocess using a nonlinear model predictive controller”. In: *Journal of Process Control* 24.4 (2014), pp. 344–357.

[37] Lisa Mears et al. “A review of control strategies for manipulating the feed rate in fed-batch fermentation processes”. In: *Journal of biotechnology* 245 (2017), pp. 34–46.

[38] Bernhard Henes and Bernhard Sonnleitner. “Controlled fed-batch by tracking the maximal culture capacity”. In: *Journal of biotechnology* 132.2 (2007), pp. 118–126.

[39] S Schaepe et al. “Avoiding overfeeding in high cell density fed-batch cultures of E. coli during the production of heterologous proteins”. In: *Journal of biotechnology* 192 (2014), pp. 146–153.

[40] Lisa Mears et al. “Mechanistic Fermentation Models for Process Design, Monitoring, and Control”. In: *Trends in biotechnology* (2017).

[41] H Scholten et al. *Good Modelling Practice Handbook*. 2000.

[42] Tobias Neymann, Lukas Hebing, and Sebastian Engell. *Computer-implemented method for creating a fermentation model*. US Patent App. 15/009,903. 2016.

[43] Lukas Hebing et al. “Efficient generation of models of fed-batch fermentations for process design and control”. In: *IFAC-PapersOnLine* 49.7 (2016), pp. 621–626.

[44] Sebastian Herold, Thomas Heine, and Rudibert King. “An automated approach to build process models by detecting biological phenomena in (fed-) batch experiments”. In: *IFAC Proceedings Volumes* 43.6 (2010), pp. 138–143.

[45] Sebastian Herold and Rudibert King. “Automatic identification of structured process models based on biological phenomena detected in (fed-) batch experiments”. In: *Bioprocess and biosystems engineering* 37.7 (2014), pp. 1289–1304.

[46] Rita Lencastre Fernandes et al. “Applying mechanistic models in bioprocess development”. In: *Measurement, Monitoring, Modelling and Control of Bioprocesses*. Springer, 2012, pp. 137–166.

[47] Roland Brun et al. “Practical identifiability of ASM2d parameters—systematic selection and tuning of parameter subsets”. In: *Water research* 36.16 (2002), pp. 4113–4127.

[48] Wieland N Reichelt et al. “Generic biomass estimation methods targeting physiologic process control in induced bacterial cultures”. In: *Engineering in Life Sciences* 16.8 (2016), pp. 720–730.

[49] Covadonga Quirós et al. “Application of flow cytometry to segregated kinetic modeling based on the physiological states of microorganisms”. In: *Applied and environmental microbiology* 73.12 (2007), pp. 3993–4000.

[50] N Borth et al. “Flow cytometric analysis of bacterial physiology during induction of foreign protein synthesis in recombinant Escherichia coli cells”. In: *Cytometry Part A* 31.2 (1998), pp. 125–129.

[51] Christopher J Hewitt et al. “Use of multi-staining flow cytometry to characterise the physiological state of Escherichia coli W3110 in high cell density fed-batch cultures”. In: *Biotechnology and bioengineering* 63.6 (1999), pp. 705–711.

[52] V Looser et al. “Flow-cytometric detection of changes in the physiological state of E. coli expressing a heterologous membrane protein during carbon-limited fedbatch cultivation”. In: *Biotechnology and bioengineering* 92.1 (2005), pp. 69–78.

[53] Matthew P DeLisa et al. “Monitoring GFP-operon fusion protein expression during high cell density cultivation of Escherichia coli using an on-line optical sensor”. In: *Biotechnology and bioengineering* 65.1 (1999), pp. 54–64.

[54] Patrick Wechselberger, Patrick Sagmeister, and Christoph Herwig. “Real-time estimation of biomass and specific growth rate in physiologically variable recombinant fed-batch processes”. In: *Bioprocess and biosystems engineering* 36.9 (2013), pp. 1205–1218.

[55] John A Nelder and Roger Mead. “A simplex method for function minimization”. In: *The computer journal* 7.4 (1965), pp. 308–313.

[56] Margaret H Wright. “Direct search methods: Once scorned, now respectable”. In: *Pitman Research Notes in Mathematics Series* (1996), pp. 191–208.

[57] Heinrich Senn et al. “The growth of Escherichia coli in glucose-limited chemostat cultures: a re-examination of the kinetics”. In: *Biochimica et Biophysica Acta (BBA)-General Subjects* 1201.3 (1994), pp. 424–436.

[58] Dai D Fan et al. “Characteristics of fed-batch cultures of recombinant Escherichia coli containing human-like collagen cDNA at different specific growth rates”. In: *Biotechnology letters* 27.12 (2005), pp. 865–870.

[59] Gregory W Luli and WILLIAM R Strohl. “Comparison of growth, acetate production, and acetate inhibition of Escherichia coli strains in batch and fed-batch fermentations.” In: *Applied and environmental microbiology* 56.4 (1990), pp. 1004–1011.

[60] Stefan Gnoth et al. “Control of cultivation processes for recombinant protein production: a review”. In: *Bioprocess and biosystems engineering* 31.1 (2008), pp. 21–39.

[61] Moira Monika Schuler and Ian William Marison. “Real-time monitoring and control of microbial bioprocesses with focus on the specific growth rate: current state and perspectives”. In: *Applied microbiology and biotechnology* 94.6 (2012), pp. 1469–1482.

[62] E Bech Jensen and S Carlsen. “Production of recombinant human growth hormone in Escherichia coli: expression of different precursors and physiological effects of glucose, acetate, and salts”. In: *Biotechnology and bioengineering* 36.1 (1990), pp. 1–11.

[63] HY Lin et al. “Determination of the maximum specific uptake capacities for glucose and oxygen in glucose-limited fed-batch cultivations of Escherichia coli”. In: *Biotechnology and bioengineering* 73.5 (2001), pp. 347–357.

[64] Luis Vidal et al. “Influence of induction and operation mode on recombinant rhamnulose 1-phosphate aldolase production by Escherichia coli using the T5 promoter”. In: *Journal of Biotechnology* 118.1 (2005), pp. 75–87.

List of Figures

List of Tables