Die approbierte Originalversion dieser Diplom-/Masterarbeit ist in der Hauptbibliothek der Technischen Universität Wien aufgestellt und zugänglich.
http://www.ub.tuwien.ac.at

The approved original version of this diploma or master thesis is available at the main library of the Vienna University of Technology.
http://www.ub.tuwien.ac.at/eng

TECHNISCHE UNIVERSITÄT WIEN

D I P L O M A R B E I T

Comparison of different model-based observers for monitoring substrate and product concentrations in a *Penicillium chrysogenum* Fed-batch process

Thema

Ausgeführt am Institut für

Chemical, Environmental and Bioscience Engineering

der Technischen Universität
Wien

unter der Anleitung von **Univ.Prof. Dipl.-Ing. Dr.techn. Christoph Herwig** und **Dipl.-Ing. Julian Kager** als verantwortlich mitwirkenden Universitätsassistenten

durch

Vladimir Berezhinskiy, BSc

Name

Datum

Unterschrift(Student)

Abstract

Real-time process monitoring and control strategies are needed to guarantee the required product quality and the improvement of the manufacturing process. Monitoring in the fermentation industry often relies on the analysis of offline samples. This time-consuming measurement technique hinders the application of automated feedback control. Novel developments in state estimation and the availability of chemometrics makes it nowadays possible to use online collected information in order to estimate non-measured system states. In this work, an industrial strain of *Penicillium chrysogenum* was used as a model organism in a fed-batch fermentation process with penicillin as a main product. Two vibrational spectroscopy methods (near- and mid-infrared) were introduced. Spectral data was used for the construction of black-box PLS models, which were able to predict the concentrations of soluble components (with prediction errors below 30% for most of the processes). However, it was shown, that constructed PLS models lack transferability. Kinetic modeling was also applied. Nevertheless, in order to be able to react to the unforeseen process changes, kinetic models need real-time process information. Therefore, automated feedback regulation was supported by a model-based observer – particle filter. Combination of off-gas measurements, near- and mid-infrared based PLS models, as well as the kinetic model via particle filter led to the establishment of a good, real-time process monitoring strategy. Finally, the established system was validated by a real process and proper process control was successfully reached.

List of abbreviations

PAT	Process Analytical Technology
PLS or PLSR	Partial Least Squares Regression
MLR	Multi-Linear Regression
PC	Principal Component
PCA	Principal Component Analysis
PCR	Principal Component Regression
CER	Carbon dioxide Evolution Rate
OUR	Oxygen Uptake Rate
PF	Particle Filter
CDW	Cell Dry Weight
FDA	US Food and Drugs Administration
HPLC	High-Pressure Liquid Chromatography
OPC	Open Platform Communication
PEN	Penicillin (product)
POX	Phenoxyacetate (precursor)
X	Biomass
PI	Proportional Integral control
FT-IR	Fourier-Transform Infrared spectroscopy
MIR	Mid Infrared spectroscopy
NIR	Near Infrared spectroscopy
UV	Ultraviolet
QbD	Quality by Design
CPP	Critical Process Parameter
CQA	Critical Quality Attribute
DoE	Design of Experiment
Sago	Savitzky-Golay smoothing/differentiation algorithm
pdf	Probability Density Function

List of variables

C_i	The concentration of component i $[g/l]$
q_i	Biomass-specific rate of component i $[g(i)/g(X)/h]$
r_i	Uptake/production rate of component i $[g(i)/l/h]$
V	Volume $[L]$
$\dot{V}_i$	The volumetric flow rate of component i $[l/h]$
V_R	Reactor volume $[L]$
$Y_{i/j}$	The yield of component i per component j $[g(i)/g(j)]$
γ_k	Collinearity index $[-]$
ρ_K	Determinant value $[-]$
$\Delta\theta$	Parameter error $[-]$
CER	Carbon dioxide Evolution Rate $[mol(CO_2)/h]$
OUR	Oxygen Uptake Rate $[mol(O2)/h]$
n	Mole $[mol]$
σ^2	Standard deviation $[units\ of\ measurement]$

Table of contents

Introduction

Penicillium

A strain of *Penicillium ssp.* was firstly isolated by Alexander Flemming in 1928 and a paper about its antibacterial properties was published in 1929[i]. This discovery resulted in a huge impact on modern medicine.

Penicillium is a genus which belongs to the phyla Ascomycota of the subkingdom Dikarya. Fungi of the phyla Ascomycota can reproduce both sexually and asexually, either by fusion of opposite mating types or by conidia spores. The compound penicillin is produced by several species of *Penicillium* and *Aspergillus*. Today's industrial strains are highly mutated derivatives of *P.chrysogenum* species. A strain Wis Q-176 produces more than 2.5 mM penicillin and it is the ancestor of the most industrial strains used today[ii].

When *Penicillium* spore starts to germinate, its growth rapidly becomes polarized and goes only in one direction, forming a *hypha*. A hypha grows only at the tip and during the growth of a hypha new tips are formed, which results in newly branched hyphae. When there are lots of hyphal-elements present, they form a *mycelium*. Thus, spore germination results in a branched network (Figure 1[ii] and Figure 2).

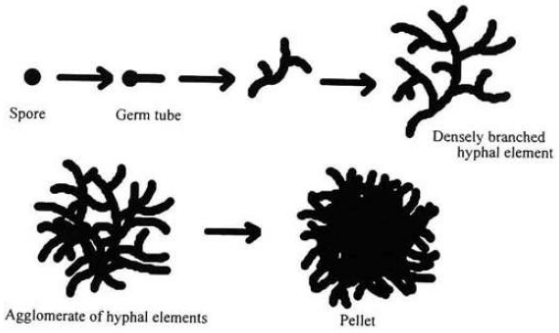

Figure 1[ii]: Hyphal and pellet formation

Therefore, cultures of *Penicillium* can be present as a mycelium, i.e. dispersed, or they form pellets – spherical agglomerates of several hyphae (Figure 1[ii] and Figure 2).

These morphological differences have important consequences for penicillin production. A key role here is played by the substrate and oxygen uptake ability of the fungus. If pellets are too thick, oxygen and sugars cannot reach cells in the middle of the pellet, and therefore these regions die and cells lyse. This means that even despite high biomass, which can be measured here by using the cell dried weight method, the real amount of living and producing cells is lower. Lots of factors, such as agitation, for example[iii], were determined to have an effect on the fungus morphology[iv].

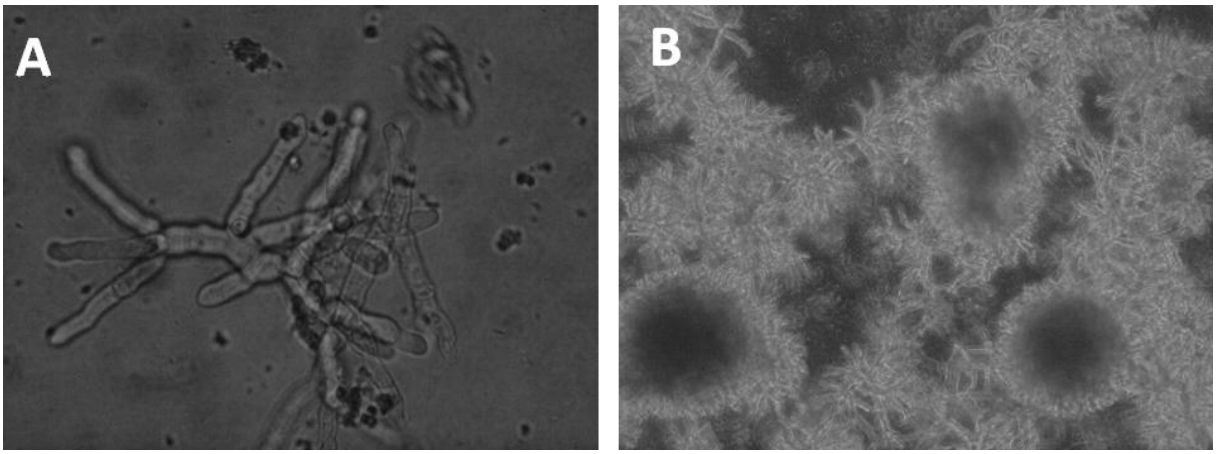

Figure 2: Microscopic pictures of *P.chrysogenum* hyphae (A, 100x magnification) and pellets (B, 60x magnification)

Penicillin V

Penicillin is a β-lactam antibiotic. In general, it acts more efficient against gram-positive bacteria, rather than gram-negative ones. Its antibiotic function is characterized by inhibition of bacterial cell-wall synthesis. Experiments have proved that penicillin blocks the formation of peptide cross-links, inhibiting transpeptidation[v].

Biochemical pathway of penicillin V production is shown in Figure 3. It can be seen that the synthesis consists of three enzymatic steps. The sulfur in penicillin V stems from cysteine and the first step is the condensation of three amino acids[ii]. After the second reaction, where isopenicillin N is formed, the side chain of isopenicillin N is exchanged with phenoxyacetate. This can occur either by one step or by two-step mechanism via 6-aminopenicillanic acid. Before that, phenoxyacetate has to be activated in the form of a CoA-ester via acetyl-CoA synthetase, which is repressed by glucose and induced by acetate[ii].

Described pathway makes it clear that addition of phenoxyacetate feed as a precursor is necessary for the production of penicillin V. Furthermore experiments have shown that additional L-cysteine, L-valine and L-α-aminoadipate feeds can increase penicillin V production rates significantly[ii].

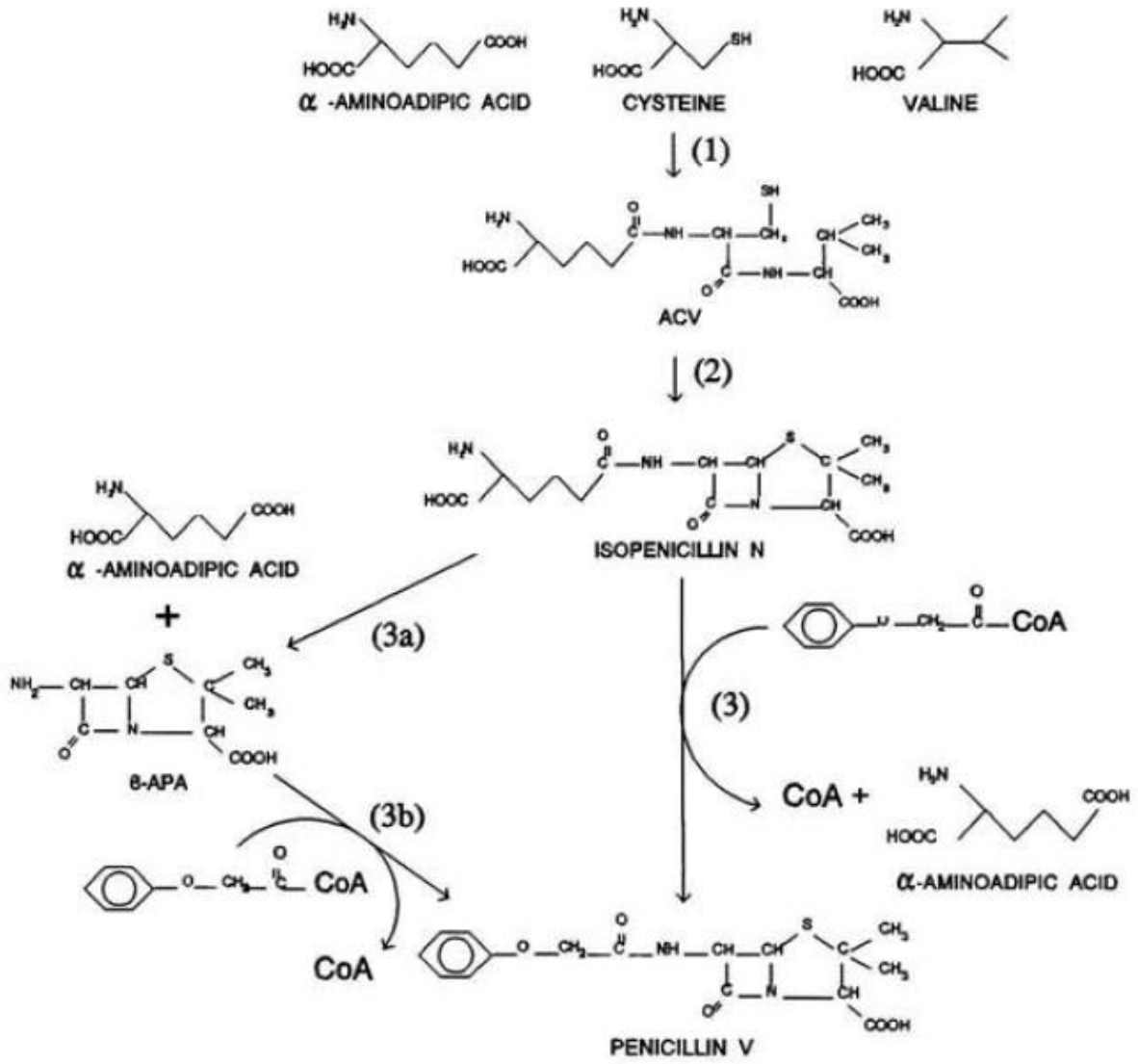

Figure 3[ii]: Biochemical pathway of penicillin production. ACV denotes α-aminoadipyl-cysteinyl-valine and 6-APA denotes 6-aminopenicillanic acid. The enzymes are: (1) LLD-ACV synthetase; (2) Isopenicillin N synthetase; (3) Acyl-CoA: isopenicillin N acyltransferase; (3a) Isopenicillin N amidohydrolase; (3b) Acetyl-CoA: 6-APA acyltransferase

Penicillin production in fed-batch

Penicillium is able to utilize a wide range of carbon and nitrogen sources, including different sugars and polysaccharides, amino acids and lipids[ii]. Fermentation of *Penicillium* in stirrer tank reactor is the standard penicillin production process[vi],[vii],[viii].

Every bioreactor, independent on its form, has to be able to do up to five basic things[ix]:

- Homogenizing
- Suspending
- Dispersion
- Energy and mass transfer
- Provide sterile conditions

The early penicillin production was developed in a stirred tank reactor[ii] and it still remains the preferred one, as the costs for modification of existing technologies are lower than for introduction of a completely new system.

Stirred tank reactors are usually cylindrical, with a typical ratio between height and diameter of 1:1 till 3:1. Heat is controlled with the heating/cooling mantle or external exchanger loops. Gas supply is provided through the gas-sparger and a reactor normally has baffles for breaking the laminar flow. Oxygen flow pro reactor volume (q_{O2}, $[\mathrm{mol}/(\mathrm{l} * \mathrm{h})$) can be expressed as follows[ix]:

$$q_{O2} = k_1 \frac{A}{V}(c_1^* - c_1)$$

And c_1^* is a saturation concentration of O_2 in the liquid phase, corresponding to the gas phase. A is the surface area of gas bubbles, V is the volume of liquid and k_1 is a mass transport coefficient. Therefore, increasing stirrer velocity (increasing A), aeration or additional pressure can increase oxygen transport to the cells. Higher temperatures and high viscosity of fungal cultures have a negative impact on the oxygen solubility.

Standard fermentation production of penicillin V is a fed-batch process, which takes from 120 to 200 hours[vi]. The inoculum for a fed-batch process is produced in one or more successive batch cultivations and contains normally rather much biomass (1-3 g/l)[ii]. The first batch is inoculated with spores with a concentration of 10^{-8}-10^{-9} spores per liter[ii]. The typical batch medium composition can be seen in Table 1[ii].

Table 1[ii]: Typical medium composition for batch cultivations of *P.chrysogenum*

Component	Batch
Corn steep liquor	50 g/l
Sucrose	30 g/l
Phenoxyacetic acid	-
$(NH_4)_2SO_4$	10 g/l
KH_2PO_4	2 g/l
$CaCl_2$ $2H_2O$	60 mg/l
Antifoam agent (pleuronic can be used)	0.2 ml/l

In order to achieve higher biomass concentrations, batch media has a higher concentration of a carbon source which is typically sucrose or glucose. Corn steep liquor, which is a by-product of corn wet-

milling, is used as a complex nitrogen source. It contains large amounts of lactate (224 g/kg[ii]) and aminoacids with a content of about 40% (w/w)[ii].

Fed-batch media composition differs in its amount of salts and does not contain corn steep liquor or sugars. Therefore, ammonia and glucose feeds are required. In order to induce penicillin production, glucose has to remain limiting during the fed-batch stage.

Normally achieved yield of penicillin to phenoxyacetate is about 1 C-mole per C-mole, which is significantly lower than the theoretical C-molar yield of 2. This can be explained through the oxidation of phenoxyacetate, and therefore its concentration has to be kept at a low level, especially during the fed-batch phase[ii].

Dissolved oxygen concentration is demanding, and is normally kept at high levels (equal or more than 40% oxygen saturation), during the whole process. Optimal fermentation temperature is between 25 and 30 ˚C. Regulation of pH is not needed during the batch phase as an increase in its value at 0.5 from the minimum point indicates the end of the batch phase. Fed-batch pH value is normally kept constant between 6.4 and 6.8[vi]. The typical specific grown rate is found to be about 0.2 h^{-1} for a batch phase and kept at lower levels (below 0.1 h^{-1}) for a fed-batch phase[ii].

A summary of crucial factors for *P.chrysogenum* fermentation process is presented in Table 2, and a schematic representation of a standard fermentation can be seen in Figure 4.

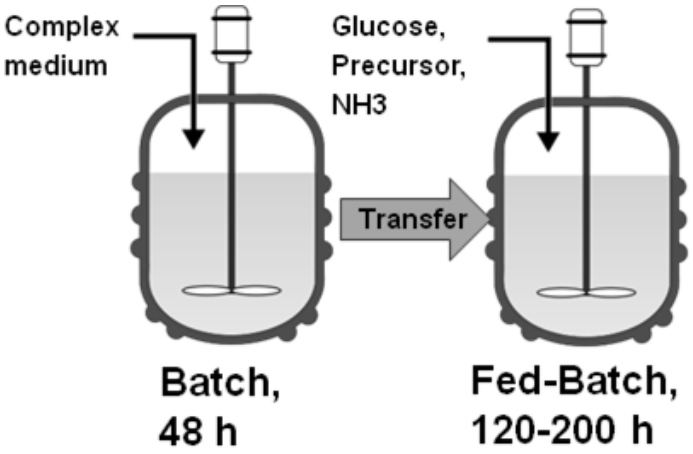

Figure 4: Schematic representation of a standard *P.chrysogenum* fermentation

Table 2[ii]: Crucial factors of *P.chrysogenum* fermentation process

	Batch	**Fed-batch**
Temperature [˚C]	25-30	
pH [-]	Not regulated	6.4-6.8
Dissolved oxygen [% oxygen saturation]	Over 40	
Medium	Complex	Defined
Glucose feed	Excess amount	Limiting concentration
Precursor concentration [g/l]		2-5
NH_3 concentration [g/l]		Over 1.5

Process models for penicillin production in *P.chrysogenum*

A mechanistic model is a mathematical description of a dynamic process. Typically, the mechanistic model is a combination of fundamental first-principles models of the physical processes with empirical models for metabolic rates and growth kinetics[x]. Model parameters are representing a connection between the mathematical basis and empirical process data.

The reliability of a mechanistic model can be evaluated through standard engineering tools such as identifiability, uncertainty and sensitivity analyses[xxxii]. When the reliability of a model is documented, it can be applied at different stages of process development: planning, design, monitoring, and control[x]. Implementation of mechanistic models allows a better understanding of changes in critical process parameters (CPPs) and critical quality attributes (CQAs) caused by process changes, which is an important trend of quality by design (QbD) framework[x].

Mechanistic models can be used for offline and online process development as well as for online model-based control. During the offline process development, estimation of optimal process operation conditions is usually done through a series of scale-up experiments. Mechanistic models can be applied at this stage in order to understand equipment limitations at different scales as well as for assessing process sensitivity to the changes in process conditions[x]. Model-based estimation strategies allow estimation of states which cannot be measured directly.

The main limitation of the mechanistic model is dependent on the current understanding of a dynamic system. Therefore, significant time and resources have to be invested during the mechanistic model development in comparison to black-box models, which can be constructed more rapidly. Nevertheless, in contrast to mechanistic models, black box models are not applicable outside the conditions used to develop the model[x].

One of the most-used kinetic models, that describes hyphal grow and penicillin production in *P.chrysogenum* was developed by *Paul and Thomas*[xi,xxix].

Efforts to develop industrial-scale fed-batch fermentation were also made by *Goldrick et al*[xii]: a simulator for a fed-batch fermentation process of *P.chrysogenum* was developed there using MATLAB. The developed simulator is able to take into account such factors as viscosity of the liquid phase, temperature, pH and dissolved CO_2 and O_2. Moreover, concentrations of ammonia and precursor were also included in the developed model.

The underlying model was the model developed by *Paul et al.*[xxix], which describes the following kinetics:

1. Fungal hyphae can be divided into the following parts:
   1.1. A_0 – active growing hyphal regions
   1.2. A_1 – non-growing hyphal regions
   1.3. A_2 – vacuoles
   1.4. A_3 – lysing regions
2. Total biomass ($X, [g]$) is calculated as:

$$X = \sum_{i=0}^{3} A_i$$

3. The main grown and penicillin production limiting carbon sources are glucose and lactose

4. Influences of POX, oxygen, and nitrogen on the fungal growth and penicillin production are not considered in this model
5. Branching, differentiation, extension, penicillin production and vacuolation rates undergo *Monod kinetic*
   5.1. High concentrations of substrate inhibit differentiation and penicillin production
   5.2. Penicillin is produced by non-growing regions
   5.3. Penicillin is hydrolyzed constantly
   5.4. Lactose consumption is inhibited by high glucose concentrations due to catabolite repression

Schematic representation of different hyphal regions in *P.chrysogenum* can be seen in Figure 5[xxix] and a schematic representation of the overall model reactions are presented in Figure 6[xxix].

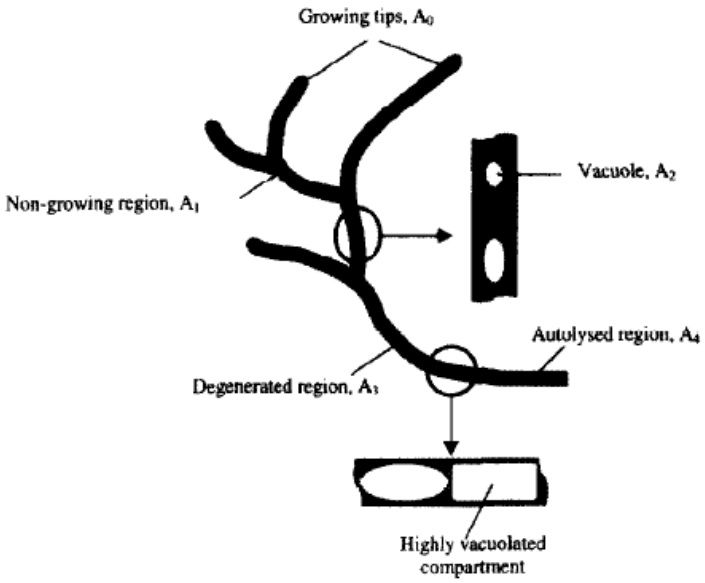

Figure 5[xxix]: Different hyphal regions in a *P.chrysogenum* species

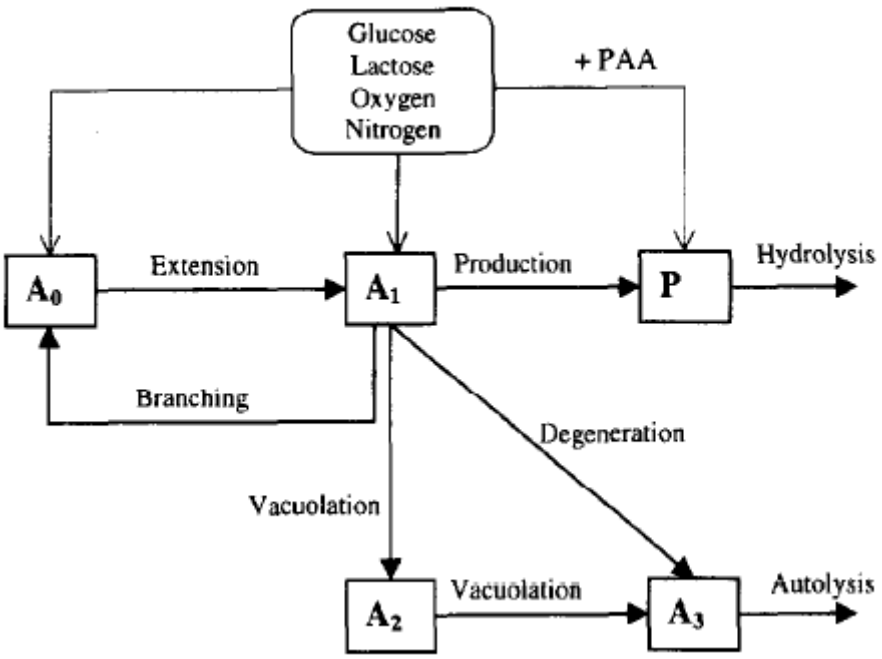

Figure 6[xxix]: Schematic representation of reactions which are used by *Paul et al.* model construction

Process Analytical Technology (PAT)

The process analytical technology initiative resulted in the development of many different tools and software packages dedicated to bioprocess monitoring and control. The complexity of biological processes, the need to operate in sterile environments in order to achieve required operation safety and the relatively few real-time direct measurements available led to a challenge of creative solutions and identification of new topics to be investigated[xiii].

The quality of industrially produced pharmaceutical products is dependent on how well and robust they are and how the manufacturing process is designed[xiv]. The main goal of the PAT framework is, therefore, to design, control and analyze manufacturing processes. A QbD initiative deals with the design of the processes during the development phase of a new product in order to provide the required product quality. Understanding of relationships between CQAs and process parameters is needed in order to identify CPPs and is an important part of QbD framework[xiv]. Application of PAT and QbD results in the desired product quality, which is not reliant on the end-of-line tests anymore[xiv].

Media composition is one of the most important things which have to be determined during pre-inoculation procedures. For that goal, a statistical "Design of Experiment" (DoE) approach is normally applied in small batch equipment[xiii].

Post-inoculation operations define the result of the overall fermentation cycle. Direct and indirect measurements can be applied during this step in order to estimate CQAs.

Most of the direct routine monitoring methods of CQAs applied in the industry are based on off-line sampling. Biomass is usually measured gravimetrically. Concentrations of nutrients and metabolites are normally estimated by HPLC or enzymatic assays. However, time-consuming offline methods are not optimal for a dynamic process as they do not afford real-time process monitoring and control. Recent developments in biosensor and biochip technology propose promising results, but their application field is still quite restricted.

Some of the direct measurements can be done online or real-time. Online HPLC allows a better process monitoring, but introduces additional contamination risks. Real-time biomass estimations can be done using at-line dielectric spectroscopy instruments[xiii]. However, this method still lacks such parameters as robustness and transferability and needs additional calibration steps.

Indirect estimations, based on measurements of parameters, which do correlate with CQAs, are usually done more easily than direct measurements. Off-gas measures, which can be done real-time, do, usually, provide enough information for estimation of biomass amount with the help of material balances. As lots of organic substances absorb in the mid-infrared range, another possibility is mid-infrared spectroscopy, which can be applied for real-time estimation of concentrations of crucial substances. Near-infrared spectroscopy can also be applied for measuring process fingerprints[xiv]. Additional multivariate data analysis for spectra evaluation, which has to be performed, is based on standard chemometrical regression tools. High precision and stability of infrared instruments provide improved process control strategy. Some of the infrared measuring instruments allow non-invasive real-time process monitoring, which is prevailing due to higher operation safety. Combination of IR measurements with chemometrics methods for qualitative and quantitative analysis and process control is being more popular in the pharmaceutical industry and biotechnology[xv],[xvi],[xxxv],[xxvii],[xvii].

State estimation

State estimation is applicable to all areas of engineering and life sciences. State estimation can be done via a wide set of mathematical algorithms. All of these mathematical algorithms are solving a filtering problem – they have to provide the most probable estimates of states based on the underlying modeled relation and (secondary) measurements. More generalized filtering problem can be formulated as finding an estimate of the current state based on the past inputs, past outputs, the model and initial conditions[xviii].

In the case of kinetic models, states are concentrations of different substances (biomass, substrate, precursor, etc.). Having some states (measurements) available, state estimator allows calculating the most probable values for unmeasured states.

Some estimators are designed and work properly only with linear models.

The linear model can be described as[xxxiii]:

$$\dot{x} = Ax + Bu$$

$$y = Cx$$

Where x is the state vector, u is the control vector and y is the output. A, B and C are system, input and output matrices correspondingly, which can depend on time as well. A linear system is observable when the initial state is uniquely defined by the input and the time derivatives of the output[xviii].

Non-linear systems can be described as[xxxiii]:

$$\dot{x} = f(x, u, w)$$

$$y = h(x, v)$$

Where w and v indicate process and measurement noise respectively. And f and h are arbitrary vector valued functions, which can be time dependent. In order to apply linear tools, non-linear systems must be linearized by the expansion of f in a Taylor series around the nominal operating point $(\bar{x}, \bar{u}, \bar{w})$ so, that:

$$\dot{\tilde{x}} = A\tilde{x} + B\tilde{u} + Lw$$

Expansion in a similar way for y gives:

$$\tilde{y} = C\tilde{x} + Dw$$

Wide used *Kalman Filter* is the best *linear* filter for any kind of system[xxxiii]. However, in the case of non-linear system *non-linear* filters can provide a better solution[xxxiii]. *Kager et al.*[xxx] combined the model of *Paul et al.*[xxix] with online (off-gas) and offline measurements with such a Bayesian state estimation algorithm (*Particle Filter*) for monitoring biomass growth. This strategy allowed carrying out better-controlled fermentations, by correcting a kinetic model with the real values. Exact workflow description of the non-linear filter used in this work – *Particle Filter* is given in **Materials and Methods.**

Spectroscopy

Analysis of radiation which is emitted, absorbed or scattered by molecules is called spectroscopy. Atomic spectra can be measured in order to get detailed information about the electronic structure of an element. Information about the molecules can be obtained from photons in the radiation range of radio waves to the ultraviolet (Figure 7). Molecular spectra contain more information and are more complicated as they carry not only the information about electronic transitions, but also the information about transitions between its rotational and vibrational states[xxi] .

Recently, such methods as infrared, RAMAN, photoacoustic, fluorescence, and UV spectroscopy, as well as mass spectroscopy, are applied in bioprocess engineering[xiv]. Most of these measurements can be done in-situ, which allows real-time estimations and process monitoring. Fluorescence instruments are often used for pH, dissolved oxygen and carbon dioxide estimation[xix]. Photoacoustic instruments are widely applied for off-gas analysis. Mass-spectroscopy is performed off-line and allows estimation of many factors, which includes proteome and metabolome analysis, evaluation of expression levels and growth characteristics of an organism[xiv]. Infrared instruments can provide detailed chemical information about the compound and are mostly used for measurements of nutrient and metabolite concentrations during cultivations.

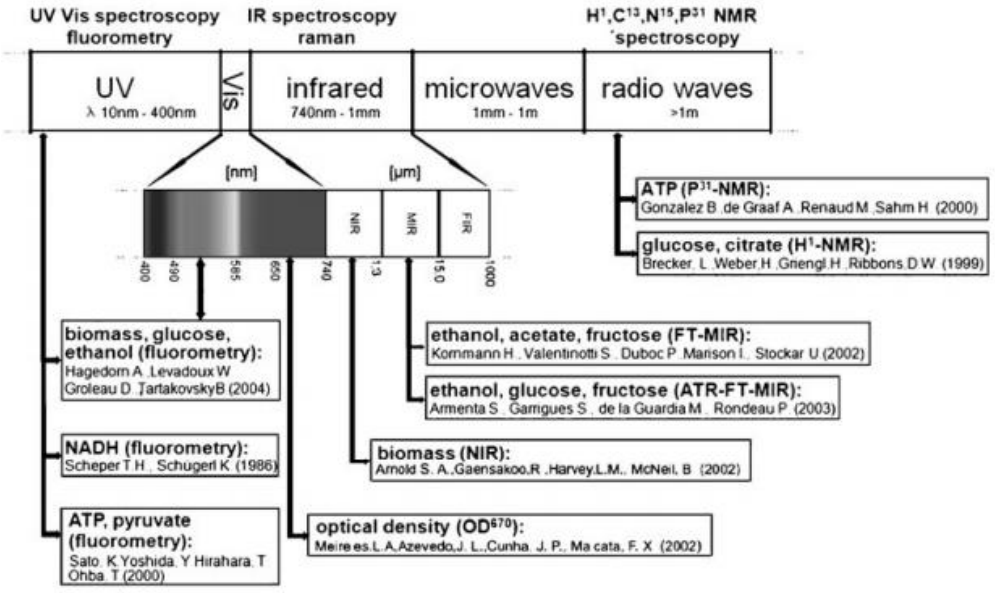

Figure 7[xx]: Recent applications of spectroscopic methods in biotechnology

IR Spectroscopy

Infrared spectroscopy is based on the ability of molecules to absorb energy at wavelengths which are specific for their structure, dependent on their vibrational energy. The vibration of a molecule combines stretching and bending motions of individual bonds. A quantum theoretical explanation is given by anharmonic oscillator[xx] and different types of energy levels are illustrated in Figure 8.

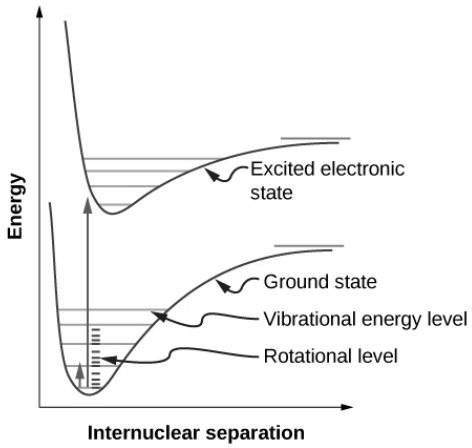

Figure 8: Electronic, rotational and vibrational energy levels in a diatomic molecule

In order to be detected via infrared spectroscopy, the dipole moment of a molecule must be able to change due to the absorption of IR radiation[xxi]. Specific absorption energies are dependent on the number of vibrational modes. If a molecule has N atoms, then for nonlinear molecules there are 3N-6 vibrational modes, and for a linear molecule 3N-5. For example, CO_2 is a linear molecule and therefore it has 3N-5 modes. An infrared spectrum of a CO_2 molecule is shown in Figure 9.

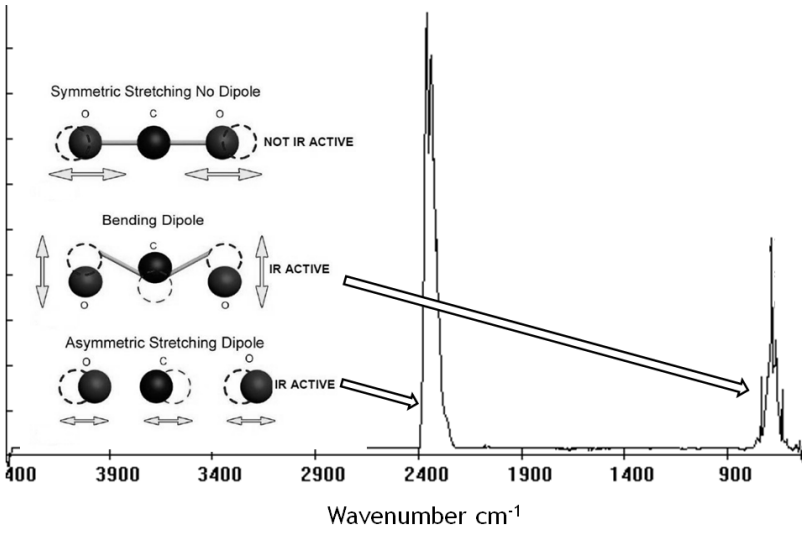

Figure 9: Vibrational modes of CO_2 molecule and its infrared spectrum

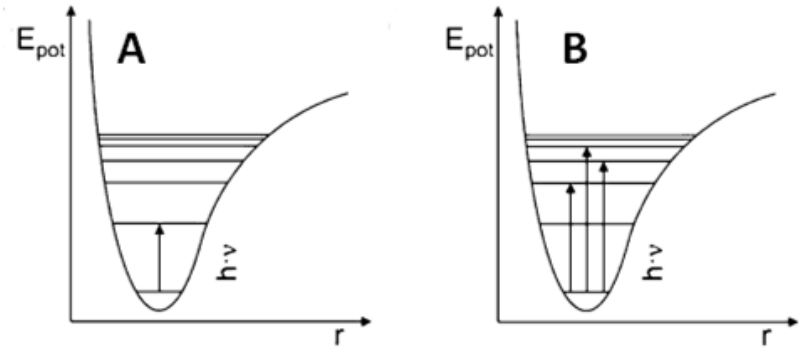

Figure 10[xx]: Illustration of fundamental vibrations by MIR (A) and overtones and combinations of vibrations by NIR (B)

Mid-infrared absorption corresponds to wavenumbers in the range of 200-4000 cm^{-1} and NIR corresponds to the range of 4000-12500 cm^{-1} (Figure 7). MIR spectroscopy reflects the fundamental vibrations of molecular bonds, whereas NIR spectroscopy reflects overtones and combinations of vibrations, thereby making MIR spectra more informative concerning the biomolecular composition of the sample[xvi]. A schematic illustration is presented in Figure 10[xx].

NIR spectroscopy can be applied non-invasive, by measuring the spectra through the reactor glass wall, which is beneficial for bioindustry (higher operation safety, lower sterilization costs, etc.). On the other hand, MIR spectra may be more informative as not all the substances absorb (or have enough high absorption) in the NIR spectral range.

Modern spectrometers use a Fourier transformation technique. This allows separating a total absorption signal into single signals which correspond to each wavenumber. Thus, the whole spectrum can be observed continuously and higher analysis sensitivity can be achieved.

Evaluation procedure

A simultaneous presence of a big number of components during the fermentation process and overlapping absorption bands makes the evaluation of IR spectra a quite complex task. Fingerprint region (1500 – 500 cm^{-1}), which is typically used for identification of single substances due to their high specific patterns in this part, is also difficult to interpret if the number of components is high. Therefore, in a bid to extract information about concentrations of substances, multivariate data analysis is required.

There are three main parts of spectral data processing:

- Mathematical pre-treatment
- Variable (wavenumber) selection
- Regression model construction

Data pre-processing is usually necessary to normalize the data, reduce noise impact or increase the sensitivity of the method to a certain substance. Most used mathematical data pre-processing methods are *Standard Normal Variate*, *Multiplicative Scatter Correction*, *Mean Centring*, *Orthogonal Signal Correction*. Derivatives of the spectra can be considered as well, therefore *Savitzky-Golay* differentiation is often applied[xv]. The exact description of the methods used in this work is given in the section **Materials and Methods**.

Variable selection is an important part of the evaluation of spectral data. Meaningless parts of the spectrum have to be removed and should not be used for the construction of the regression model. This step is crucial regarding the robustness and transferability of the further constructed regression model. Choosing component-specific wavenumbers is the simplest strategy, however, it is hardly applicable if the number of components is high and their spectra overlap. Complex strategies, such as *Genetic Algorithms*[xxii] or *Artificial Neuronal Networks*[xxiii] can be applied, but their implementation requires additional computational power and training data. The exact description of wavenumber selection procedure performed in this work as well as selected wavenumbers are given in the section **Results and Discussion**.

There are lots of possible ways to construct a regression model. Following steps should be done for a regression model construction:

- Model definition
- Model calibration
- Model validation
    - Validation via external data-set (preferred)
    - Cross-validation via internal set: a part of the used data set is defined as a test set and used for validation
- Model update when the new data is coming

Each step is crucial for the outcome. Wrong evaluation or bad calibrated models will not provide satisfying results. Validation gives a hint about model transferability. Data used for the model update should be consistent.

Linear regression method is the simplest way of data processing. *Multi-Linear Regression* is an expanded form of a typical linear regression. This method allows computing regression models for a multivariate dataset. However, MLR is quite limited, because it cannot be applied if collinearities are present in the analyzed dataset. In the case of IR spectral data, wavenumbers are considered as

variables. This means that lots of variables will be dependent on each other, there will be collinearities, because typical IR absorption bands are pretty broad. Therefore, another regression method has to be applied here.

Principal Component Regression overcomes the collinearity problem as the initial coordinate system of the data is being rotated, and new independent variables are formed. Therefore, initial variables are multiplied with "weights", forming a new vector as a linear combination of the original variables. The so-called *Principal Components,* or *scores,* are representing the new coordinates of the data and are uncorrelated.

The method applied in this work – *Partial Least Squares* is similar to PCR (mentioned before). However, not only the *X* data, but also the dependent variables, *Y* data, are proceeded through PCA in this case (the PCA algorithm itself is also applied in a slightly different way). In the case of PLS, least squares algorithm computes the regression between *X* and *Y*. Moreover PLS procedure can be applied for an analysis of multivariate *Y* data (such an algorithm is called PLS2, more detailed description of an applied PLS method can be seen in the section **Materials and Methods**).

The principle workflow, how to establish a proper quantitative IR based model can be seen in Figure 11. As the IR spectral data cannot be considered as a primary quantification method itself in case of biotechnological fermentations, it needs the presence of a reference method, which is used then for model construction.

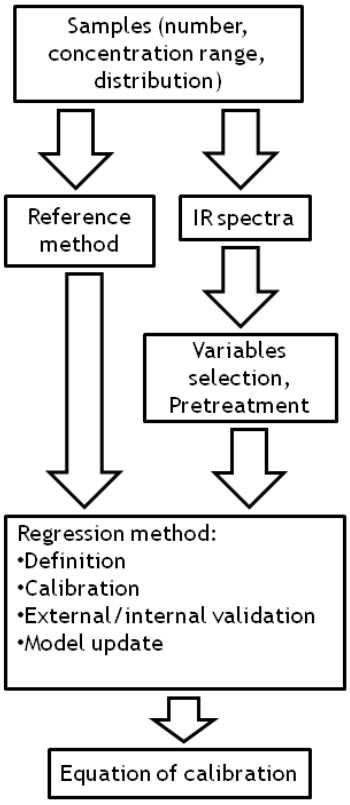

Figure 11: Establishment of a proper, quantitative IR spectral-based regression model

The performance of a model can be evaluated via different criteria, such as *R-square*, *Standard Errors*, *Root Mean Square Errors*, etc. The detailed description of the evaluation methods used in this work is described in **Materials and Methods**

Comparing MIR and NIR spectral data, it can be said that both measurement methods can be applied for PLS model construction, however, the performance of the obtained regression models may be different and depends on the problem statement[xvi].

Challenges in monitoring and control of biopharmaceutical processes

Establishment of appropriate bioprocess monitoring and control strategies is still challenging. Multianalyte nature of complex biological systems and simultaneous course of hundreds of chemical reactions inside the living cells and outside the cells in the liquid and gaseous phases requires convoluted solutions.

Current penicillin V production is a well-established and optimized process. Nonetheless, such crucial states as concentrations of penicillin, ammonia, precursor, biomass, and substrate are mostly measured with common offline sampling techniques. This approach is time-consuming and does not afford any possibility for real-time monitoring and control.

Off-gas measures can be evaluated in real-time. They produce good biomass estimations, based on mass balances, but are not sufficient for calculation of other central states.

Understanding of grown and production pathways lead to the development of a great number of kinetic models, describing *P.chrysogenum* fermentation process. Kinetic models have good transferability and prediction strength. However, they require good starting values and are not always capable to react to the unforeseen process changes.

Novel instruments, such as mid and near-infrared spectrometers, can be implemented for process monitoring as well. In comparison to usual offline sampling, their signal can be obtained and interpreted in real-time. A capability to make some IR measurements non-invasive reduces the contamination risks. Modern regression methods, such as PLS, make it possible to construct data-driven models, based on spectral data, for any process, independent on its nature. However, as there is no biochemical background in such a data-driven model, it is hard to control and interpret.

The possible addition of described real-time measurements to the kinetic model should shift the model to correct values, providing lacking information and resulting in better monitoring and control strategies. However, biological systems are often non-linear and are therefore difficult to observe. Thus, a proper state-estimator, which can be applied to a non-linear system, such as *Particle Filter*, has to be preferred.

Goals of the thesis

The main goal of this work was to combine estimates made by PLS models based on IR spectral data with a model-based state estimator in order to achieve an improved monitoring method for the biotechnological penicillin production process.

The following was needed to achieve the goal:

- Performing fermentation processes in order to get training datasets for establishment and configuration of
    - the model-based state estimator
    - PLS models, based on near- and mid-infrared measurements
- Combination of different measurements and setup configurations of a state observer
- Validation of the developed solutions

Work plan

Before the main goal could have been achieved, experimental data should have been generated. Therefore, the AJ8 fermentation process was carried out. Obtained data was used for the establishment of PLS models and different model-based observer configurations. Thus, the established system was used for monitoring of the next experiment (JL1). Data, generated in JL1 process, was used for PLS models update and estimation of kinetic model parameters. Therefore, the established system was prepared for the control of the validation experiment (JL2). After the validation, performed control strategies were evaluated and compared.

Therefore, the following steps were necessary:

I. Literature search and study of available methodologies
II. Experiment (AJ8)
    a. Evaluation of experimental data
    b. Establishment and configuration of NIR and MIR based PLS models and model-based estimator
    c. Performing quasi-real-time simulations with different observer configurations and comparison of results
III. Experiment (JL1)
    a. Application of established systems for process monitoring
    b. Evaluation of experimental data
    c. Update of PLS models
    d. Parameter estimation for a kinetic model
IV. Validation experiment (JL2)
    a. Application of established systems for process control
    b. Evaluation of experimental data
    c. Comparison of applied control strategies

Materials and Methods

Fermentation process

Strain

An industrial strain of *P.chrysogenum* was used in this work.

Process description

Six fermentations with the names: AJ8A, AJ8B, AJ8C and JL1A, JL1B, JL1C were carried out during the main phase of this work. Additional fermentations with the names JL2A, JL2B were performed as a validation experiment.

All the fermentations were performed in a DASGIP (Eppendorf AG, Germany) multi-bioreactor system. This system consists of four parallel glass bioreactors, 2.7 L volume each. Each reactor has five ports for feeding and one for sampling, four baffles for breaking the laminar flow, heating blanket, gas-sparger and a stirrer with three Rushton turbines. The picture of the whole setup can be seen in Figure 12.

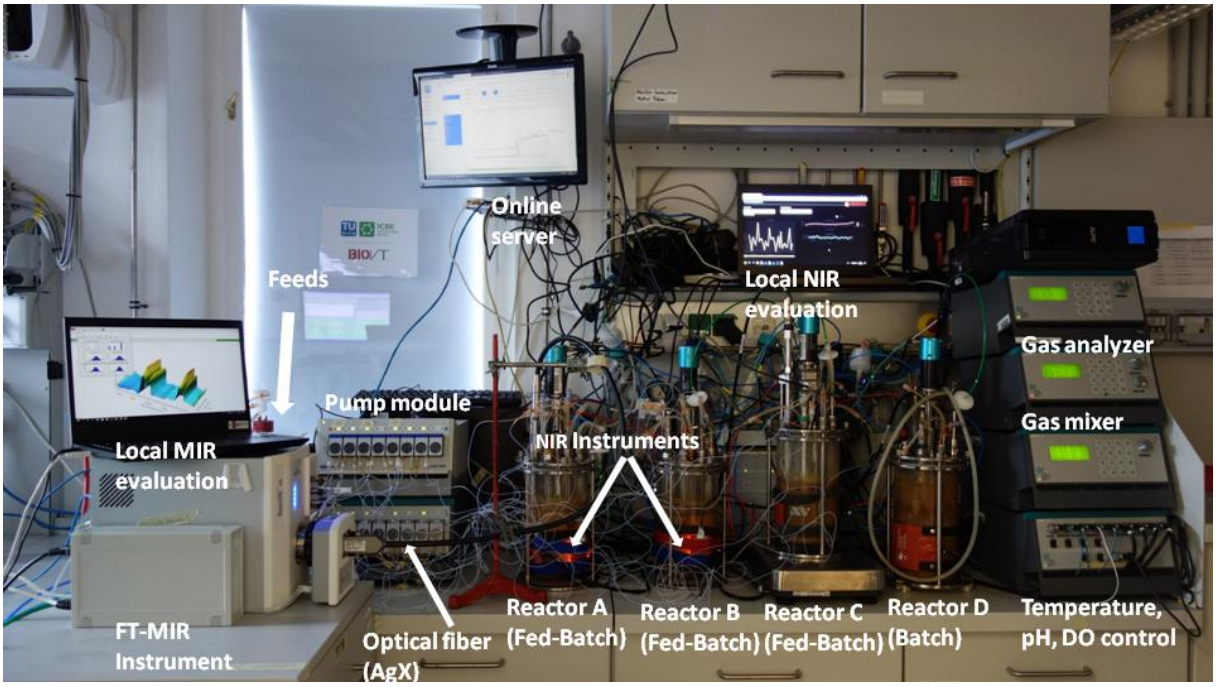

Figure 12: Set-up used in the current work

Applied control loops are illustrated in Figure 13. Regulation of pH was made with a PID controller based on acid and base addition. Dissolved oxygen was controlled with a PID controller via stirrer velocity and oxygen content in the inlet gas flow, and was measured via optical DO sensor (Hamilton, USA). Feeds (POX, ammonia, and substrate) were controlled via computer, and the correct feeding rates set-points were calculated based on the model predictions (processes AJ8 and JL1) or state estimations (process JL2).

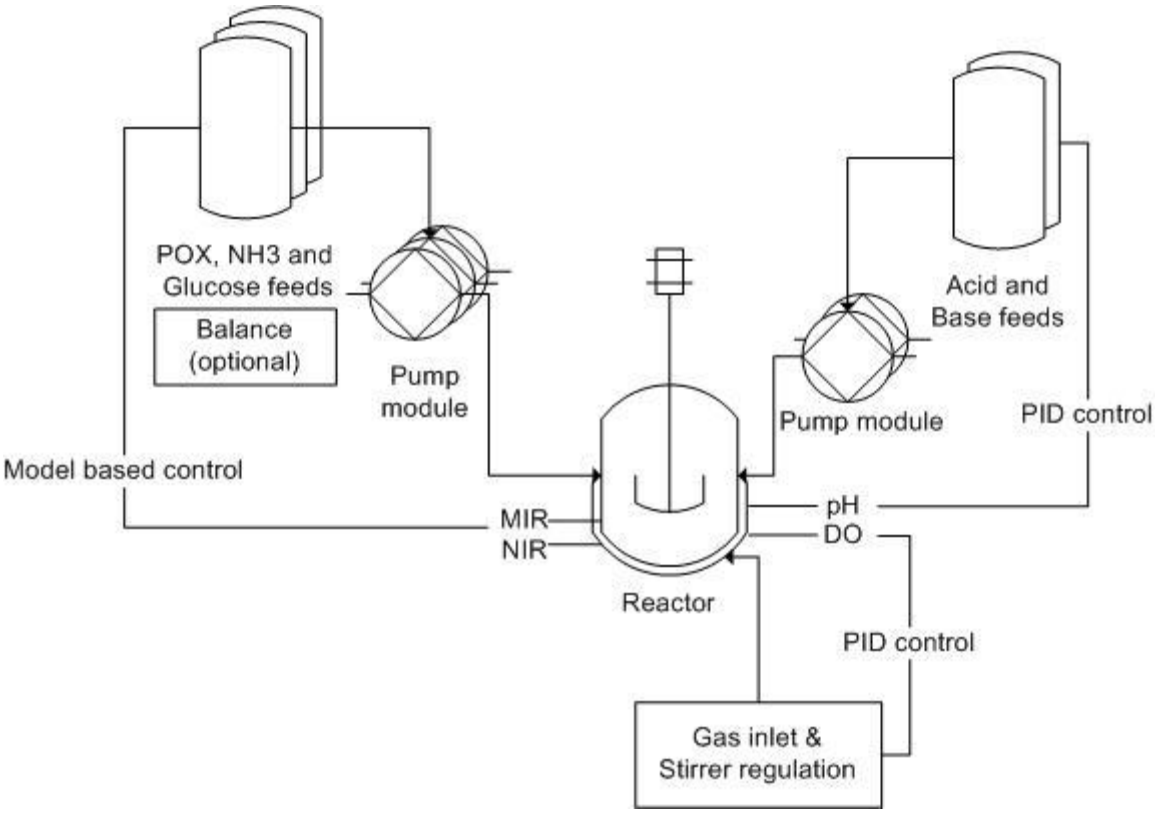

Figure 13: Principle schema of the set-up

Gas content was kept according to the needs within the usage of air (21% O_2), pure O_2 (99,99%) and N_2 (99,99%) gas inlet from flasks. Several DASGIP controlling modules (TC4SC4 for stirrer and temperature; MX4/4 for aeration; PH4PO4 for pH and pO2; MP8 for feeds addition) were connected to a PC and controlled via DASware control software (Eppendorf AG). One of the four reactors was always used for a batch phase (with a 2 L medium inside). Before the fermentation, fully equipped reactors with the medium inside were sterilized at 122 ˚C for 20 min. Batch process was performed under a temperature of 25 ˚C with 40% pO2 and uncontrolled pH.

After the batch phase, which usually took about 2 days and its end was indicated by the increase of pH, 300 ml of cell broth was transferred in each of three reactors (which has already contained 1.7 L of defined fed-batch medium). Typical fed-batch fermentation usually took about 140-160 hours. Dissolved oxygen was kept at different levels within the gas mix supply. Stirrer velocity was controlled to keep a certain power input. Feeds were 500 g/l, 80 g/l and 100 g/l for glucose, POX and $(NH_4)_2SO_4$ correspondingly.

The biomass specific uptake rate of the limiting substrate (q_s) was kept at different levels for different experiments based on model predictions by manipulating the glucose feed.

Different dissolved oxygen, power input and q_s set points were applied during the processes in order to expand the model validity space and reach different growing conditions.

Online measurements

Online measures in the reactor were performed via temperature measuring sensor (Pt element), pH electrode (Hamilton, USA), pO2 electrode (Mettler-Toledo, Switzerland), permittivity measurement sensor (Incyte, Hamilton, USA), invasive FTIR (middle IR range) measurement (Mettler-Toledo) with *AgX* optical fiber, and non-invasive NIR microspectrometers (NIRONE, Spectral Engines, Finland).

Off-gas measurements of CO_2 and O_2 were made using DASGIP (GA4) gas analyzer using infrared and paramagnetic principle, respectively.

To calculate CER [mol/h] and OUR [mol/h] from the measured CO_2 and O_2 signals, the equilibrium between liquid and gas-phase was assumed. The values themselves were calculated via[xxiv]:

$$OUR = \frac{G}{W} * (O_2^{in} - O_2^{our}) \text{ and } CER = \frac{G}{W} * (CO_2^{out} - CO_2^{in})$$

Where G $[\frac{mol}{s}]$ is the total molar gas flow, W $[kg]$ is the culture weight, O_2 $[-]$ and CO_2 $[-]$ are the gas volume fractions.

Offline measurements

Offline samples were taken every 8 hours and the sample volume was logged. Analysis of penicillin V and phenoxyacetate was done via HPLC using a ZORBAX C-18 column (Agilent Technologies, USA) and 28% acetonitrile, 6 mM phosphoric acid, 5 mM KH_2PO_4 as elution buffer. Before the HPLC analysis was performed, probes were centrifuged for 5 min at 4800 rpm and then at 15000 rpm for 15 min. Sugar concentrations were also measured with HPLC (Agilent) with a Supelco gel C-610 H ion exchange column (Sigma-Aldrich, USA) with refractive index detector (Agilent), 0.1% phosphoric acid was used as an eluent and was supplied isocratically at 4 ˚C with 0.5 ml/min flow rate. Glucose and ammonia were quantified after the 5 min centrifugation at 4800 rpm from the supernatant enzymatically with Cedex BioHT (Roche GmbH, Switzerland) with a detection limit of 0.02 g/l for glucose and 0.048 g/l for NH_3. Biomass concentrations were estimated gravimetrically after double centrifugation (separation and washing with distilled water) at 4800 rpm (10 min) of 5 ml cell broth. Cell pellets were dried at 105 ˚C for at least 3 days.

The network architecture of validation experiments (JL1, JL2)

All calculations and data analysis (data pre-processing, construction of PLS models, calculation of rates, etc.), if not mentioned others, were carried out using a MATLAB software (MathWorks, USA).

Complex network architecture was built for real-time process control and observation of the validation experiment (JL2) is displayed in Figure 14.

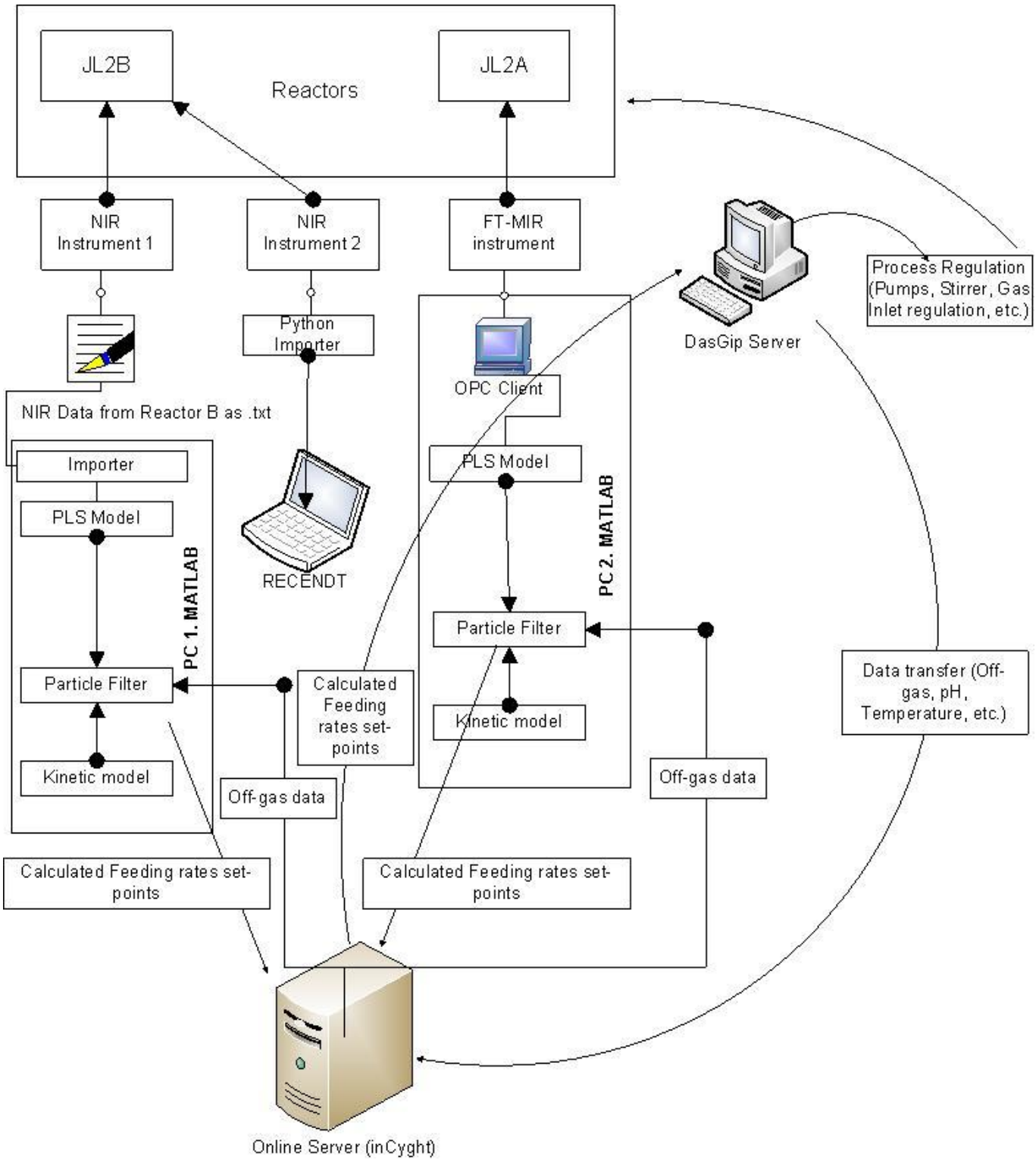

Figure 14: Network architecture, used in the validation experiment

Calculations of feeding rates set-points were performed on local PCs (denoted as PC1 and PC2 in Figure 14) in MATLAB using a developed script. DASGIP software was necessary, as it was controlling and collecting all the real-time process data (off-gas data, feeding rates, pH values, etc.) except spectral measurements. To transfer the calculated set-point to DASGIP, an online server (inCyght, Exputec GmbH, Austria) was used. The online server allowed it also to visualize any data of the three running fermentations by connecting to the corresponding website.

Reactor JL2A was controlled via off-gas, MIR measurements (Mettler-Toledo, ReactIR) and a kinetic model. Spectra were imported into a workspace of a running MATLAB process via build-in OPC client of FT-MIR instrument software. Spectral data was processed with the constructed PLS models, which were based on historical data (AJ7A, AJ8A, and JL1A experiments). Calculated states (concentrations of penicillin, precursor, and ammonia) were sent to particle filter together with off-gas data from DASGIP. This allowed estimating the most probable states which were used as a basis for process control.

Reactor JL2B was controlled via off-gas, NIR measurements (NIRONE, Spectral Engines), and a kinetic model. As there was no OPC server available for NIR devices, these results were saved as a text file on a local PC. MATLAB importer, which was written to work in online mode, imported these text files into the workspace of the running process. The programmed importer was constructed in a way, that it was importing the three last spectra and taking the mean. PLS models, based on historical data (JL1A, JL1B), were used for calculation of states (AJ8A and AJ8B were not used by the construction of PLS models for JL2B control as the process conditions were different). Further data processing was made in the same way as for Rector JL2A: combining of PLS results with off-gas data and a kinetic model by particle filter lead to estimating of the most probable states for process control.

NIR evaluation software (RECENDT, Austria), based on a web-based Python application, was additionally tested.

Calculations and data analysis

Material balance, rates, and yields calculation

A general material balance for component *i* can be written as[xxv]:

$$\dot{V}_{In}c_{i,In} - \dot{V}_{Out}c_{i,Out} + V_R r_i = V_R \frac{\partial c_i}{\partial t} + c_i \frac{\partial V_R}{\partial t}$$

Therefore,

$$r_i = \frac{\left(V_R \frac{\partial c_i}{\partial t} + c_i \frac{\partial V_R}{\partial t}\right) - (\dot{V}_{In}c_{i,In} - \dot{V}_{Out}c_{i,Out})}{V_R}$$

Depending on the component, rates could be calculated from this equation. For example for substrate:

$$r_S = \frac{\dot{V}_{In}(c_S - c_{S,In})}{V_R}$$

The biomass-specific rate of component i can be calculated per definition:

$$q_i = \frac{r_i}{c_X}$$

Yields are defined as:

$$Y_{i/j} = \frac{qi}{q_j} = \frac{ri}{r_j}$$

Error evaluation and data pre-processing

Errors

To qualify and compare different methods applied in this work it was necessary to calculate errors of predicted values ($\hat{y}_i$) relative to real values (y_i). Therefore, different types of errors and evaluating methods were calculated as described in the literature[xxxiv].

In the case of calculating the standard error, it is necessary to know the number of degree of freedom. In the case of multivariate regression, this is not a trivial question and is not so easy to estimate the real number. Thus, it is become common to calculate a *Root Mean Square Error (RMSE)*, which is calculated with the number of probes (*n*) instead of the degree of freedom:

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

RMSE can be normalized by the mean of the range in order to calculate *Normalized Root Mean Square Error (NRMSE):*

$$NRMSE = \frac{RMSE}{y_{i,max} - y_{i,min}}$$

NRMSE values below 30% were considered as "acceptable". NRMSE values below 20% were assumed as "low" and NRMSE values fewer than 10% were accounted for "very low". Prediction NRMSE, calculated for different methods which were applied in this work, can be seen in **Supplement**.

Another important value is the so-called *BIAS*. *BIAS* is a middle value of all residues. The more the value of *BIAS* is closer to zero the less systematic error is present in the model. *BIAS* is calculated as follows:

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

Standard Errors are calculated as follows:

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

To be able to evaluate the fit of the model, *R-square* values were calculated. *R-square* compares the fit of the chosen model with that of a horizontal straight line and is calculated as:

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

It is important to make a point that despite the name, *R-square* value can be negative when the chosen model fits worse than a horizontal line.

Confidence bands

For better visualization of constructed regression models, it was necessary to calculate confidence intervals for future observations. These are given as[xxvi]:

$$\hat{Y} \pm s_3 t_{n-2}$$

With

$$s_3^2 = s_{y,x}^2 \left[1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{[x^2]}\right]$$

$$s_{y,x}^2 = \sum \frac{(Y_i - \hat{Y}_i)^2}{n - 2}$$

$$[x^2] = \sum_i (x_i - \bar{x})^2$$

Smoothing algorithm

To smooth the predictions of noisy models, *Savitzky-Golay* filter was used. This is a so-called polynomic –smoothing algorithm. The algorithm fits a polynomial of a grade k (in this work $k = 1$) in the defined window (a size of 100 was used, if not mentioned otherwise) increasing the signal to noise ratio without greatly destroying the signal. *Savitzky-Golay* smoothing was performed in MATLAB with the function *smooth*.

Data pre-treatment methods

Additional data pre-processing was done according to the literature[xxxiv,xxxv,xxvii].

Mean-centering (MC) was done by subtraction of middle absorption value of all spectral values at a certain time point ($\bar{x}$) from each absorption value (of each wavenumber) at this time point (x_i).

This is made by applying the following formula:

$$x_{i,MC} = (x_i - \bar{x})$$

Standard Normal Variate (SNV) transformation allows correcting the scatter effects and decreasing noise impact. Therefore, middle absorption value of all spectral values at a certain time point ($\bar{x}$) is subtracted from each absorption value (of each wavenumber) at this time point (x_i), and then all the absorption values (of each wavenumber) are deviated through the standard deviation of the spectrum at this time point.

This is made by applying the following formula:

$$x_{i,SNV} = \frac{(x_i - \bar{x})}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}}}$$

Another important procedure is a derivation of spectra. As derivation increases the effect of noise, *Savitzky-Golay* derivation was used. *Savitzky-Golay* algorithm allowed it to fit a polynomial of 2nd degree through the signal values at each time point with a frame size of 11 (polynomial degree and window size were set according to the literature[xxxv]). After that, the 1st or 2nd order derivative of the fitted polynomial was calculated.

Linear regression

In the case of permittivity measurements, it was necessary to make linear regression models. Therefore, a *least-square fit* was applied. Linear regression models a relationship between the

dependent variable (y) and one or more independent variables (x) through the slope (β_1), intercept (β_0) and error term (ε) and can be described with the following equation:

$$y = \beta_0 + \beta_1 * x + \varepsilon$$

When there is a set of *n* observed values of x and y available, a simple system of linear equations can be solved:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & \cdots & x_1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Partial Least Squares (PLS) modeling[xxviii,xxxiv]

Model construction

Partial Least Squares Regression (PLSR or PLS) is a standard evaluation routine in chemometrics and is one of the most used approaches for relating two matrices **X** and **Y** by a linear multivariate model. The main advantage of PLS comparing to MLR is that it can analyze data with strong collinearities. This is an extremely important property, while analyzing spectral data.

There are two main PLS variants: PLS1 and PLS2. The difference between them is in the number of dependent variables (y_i). In the case of PLS1, there is only one dependent variable and **Y** is a vector. In the case of PLS2, **Y** is a matrix and the number of dependent variables is higher than one (Figure 15). The advantage of PLS1 procedure is that the wavenumbers selection can be made independent for each single component, resulting in a set of more robust single-component models comparing to the overall model which would be obtained via PLS2.

In this work, PLS1 was applied and therefore the abbreviation PLS further is written for denoting PLS1.

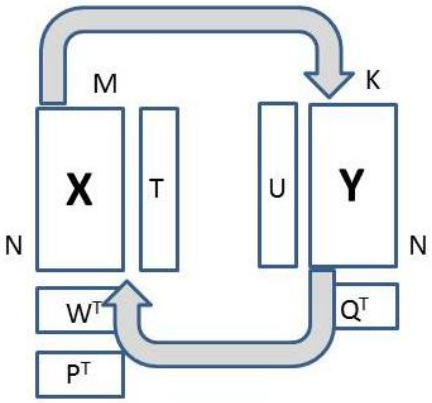

Figure 15: Graphical illustration of PLS

Loadings (P) and scores (T) are calculated for X (Figure 15). An intermediate step is calculating W-matrix, which connects X and Y data. Same as for X, Y scores (U) and loadings (Q) are also calculated. Loadings are describing the weights applied to rotate the original data matrix. Loading vector is in the linear combination of original variables. Scores are new coordinates of the data and are uncorrelated.

The procedure is described as follows[xxxiv]:

For A PLS components (can also be denoted as *Principal Components*, *PCs*, or *Latent Variables*, *LVs*), starting from $a = 1$,

$$X_a = X, y_a = y$$

In contrast to PCA, where the column of X with the highest variance would be selected for the first determination of scores (t_a), y-values are used for scores calculation in case of PLS. X-data is then regressed to the y-vector so that the error (E) is minimized:

$$X_a = y_a w_a^T + E$$

Where w_a are so-called weighted loadings which carry the relation between X and y. Weighted loadings are orthogonal and normed so that they have a length of 1:

$$w_a = \frac{X_a^T y_a}{\sqrt{(X_a^T y_a)(X_a^T y_a)^T}}$$

Within the weighted loadings, X-scores could be found in the same way – minimizing the error E:

$$X_a = t_a w_a^T + E, \text{ which gives } t_a = X_a w_a$$

Now, the normal PCA loadings can be calculated as:

$$X_a = t_a p_a^T + E, \text{ which gives } p_a = X_a^T t_a / (t_a^T t_a)$$

The calculated X-scores are then used for y-loadings computation:

$$y_a = t_a q_a + f, \text{ which gives } q_a = t_a^T y_a / (t_a^T t_a)$$

Now the first PLS component is calculated. By calculating further components, this information has to be eliminated from the dataset, so:

$$X_{a+1} = X_a - t_a p_a^T, \text{ and } y_{a+1} = y_a - q_a t_a$$

As the goal of PLS is a construction of a regression model, therefore, it can be written:

$$y = b_0 + Xb$$

With

$$b = W(P^T W)^{-1} q, \text{ and } b_0 = \bar{y} - \bar{x}^T b$$

The regression coefficients are also called the *beta* matrix. Having a test dataset, the *beta* matrix can be calculated and it is possible to apply it to unknown X data for calculation of y.

Additional interpretations

When PLS models are constructed, new variables, *PCs*, are used then instead of the original ones. These variables are linear combinations of the original ones (described by the loadings). Usually, the number of *PCs* needed for describing a spectral dataset is much lower than the number of original variables (wavenumbers). This is caused by the fact that new variables do carry only the information about how strong y is affected by X and the noise is eliminated.

The loadings show how the scores are influenced by the original variables. Choosing the variables (wavenumbers in this work) with the high weights values is a widespread approach of variable selection.

W-loadings are the connection between the dependent values and variables. They are not always (but can be) the same as the P-loadings, which are showing the relation between X-data und T-scores.

The part of the unexplained data, *residuals*, is very important for the detection of outliers and finding systematic errors. *Residuals* should be distributed normally around zero and should not be high (high values denote a poor model prediction ability).

Robustness of a model

As each additional PC can contain noise, the number of PCs should remain low, in order to increase model robustness and reduce noise impact. Therefore, PC number selection is an important step of PLS model construction.

The usual way to test the robustness of a model is cross-validation, where the random part of the dataset is used for a model building and another random part of the same dataset is used as a test set. Observing cross-validation errors, decisions about the correct PC number can be made.

To improve the outcome of this method, the so-called "Monte-Carlo repetitions" can be used. In this case, the test data set is chosen (randomly) more than once and each time cross-validation errors are being calculated. In case of stable, non-varying (with the increasing number of repetitions) errors it is supposed that the model with a current number of PCs is robust. This method allows for better PC number estimations and more realistic error predictions.

Kinetic modeling

Kinetic modeling is a mathematical representation of biochemical knowledge about the process (for a detailed description see **Introduction**).

Penicillin model

The kinetic model applied in this work is based on the one developed by *Paul et al.*[xxix], and described by *Kager et al.*[xxx]. The model has 7 states and 24 parameters. Equations and parameters which are describing the model can be seen in **Supplement**.

Sensitivity analysis

Sensitivity and identifiability analysis was performed in a similar way as described in the literature[xxxvii]. After the model definition and calculation of its outputs, sensitivity has to be computed. Sensitivity measure (δ_j^{msqr}) is defined as:

$$\delta_j^{msqr} = \sqrt{\frac{1}{n} * \sum_{i=1}^{n} s_{ij}^2}$$

Where

$$s_{ij} = \frac{\Delta\theta_j}{sc_i}\frac{\partial\eta_i}{\partial\theta_j}$$

With $\theta = (\theta_1, \dots, \theta_n)^T$ – denoting parameter vector, $\eta(\theta) = (\eta_1(\theta), \dots, \eta_n(\theta))^T$ – simulation results, and $S = \{s_{ij}\}$ and $\tilde{s}_{ij} = \frac{s_{ij}}{||s_{ij}||}$ – nondimensional sensitivity matrix and normalized matrix correspondingly, and sc_i is a scaling factor with the same units as the corresponding observation.

After sensitivity is computed and parameter importance ranking is produced, collinearity index γ_k is calculated as follows:

$$\gamma_k = \frac{1}{min_{||\beta=1||}||\tilde{S}_K\beta||} = \frac{1}{\sqrt{\tilde{\lambda}_k}}$$

Where $\tilde{S}_K$ is $n * k$ submatrix of $\tilde{S}$ with a columns which correspond to parameters in K, β – is a vector of coefficients of length k and $\tilde{\lambda}_k$ is the smallest eigenvalue of ${\tilde{S}_K}^T\tilde{S}_K$. When the value of collinearity index exceeds a threshold of 10, selected parameters subset is supposed to be poorly identifiable.

Determinant values are calculated as:

$$\rho_K = det\left({S_K}^T S_K\right)^{\frac{1}{2k}} = \left(\prod_{j=1}^{k}\lambda_j\right)^{1/2k}$$

Determinant combines information about sensitivity and collinearity. High value of ρ_K denoted 'good' identifiability of parameter set K (with low $\gamma_k$and high δ_j^{msqr}) and vice versa.

Determinant measures were used to select the best identifiable parameter set.

Calculation parameter set errors

Large parameter sets tend to contain collinearities and therefore have no unique solution.

Thus, it was necessary to calculate relative errors for estimated parameter sets, which was done as described in the literature[xxxi],[xxxii], where parameter error ($\Delta\theta$) was defined as:

$$|\Delta\theta| = \sqrt{diag(COV(\theta))}\,,$$

And, thus:

$$\theta = \mp\Delta\theta$$

And $COV(\theta)$ is an inverse of a so-called *Fischer Information Matrix (FIM)*, with

$$FIM = (dy_k/dp)^T * W * (dy_k/dp),$$

Where W denotes diagonal 'weights' matrix which is constructed based on absolute errors of the available measurements (σ):

$$W = \begin{bmatrix} 1/\sigma_{n1}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma_{nj}^2 \end{bmatrix}$$

And (dy_k/dp) is the sensitivity matrix, calculated in the previous paragraph **(Sensitivity analysis)**. The *FIM* is calculated only for states and time points which are available as measurements.

Model-based control for validation experiment

As it was already mentioned, an additional validation experiment (JL2) was performed at the end of the described work in order to validate and compare constructed PLS models when applied together with off-gas results for process control via particle filter.

The success of a validation experiment is indicated when the pre-defined q_s set-points are reached and kept constant during the whole (main phases of) process run. It was important to keep only the substrate (glucose) on a limiting level, while the concentrations of POX and NH_3 should have remained constantly non-limiting.

Therefore, a correct feed control, based on estimations of specific rates by particle filter was needed. As such, a model-based control was chosen, which means that the corresponding POX and NH_3 feeds (f $[ml/h]$) were calculated via:

$$f = \frac{(V * \hat{r_i})}{c_{i,feed}} * 1000$$

With V $[l]$– reactor volume, $\hat{r_i}$ $\left[\frac{g}{l*h}\right]$ – estimated volumetric rate of component i, and $c_{i,feed}$ $[\frac{g}{l}]$– concentration of component i in the corresponding feed.

The feed for glucose was calculated as follows:

$$f = \frac{(V * \hat{c_x} * q_s)}{c_{Glc,feed}} * 1000$$

With $\hat{c_x}$ $[\frac{g}{l}]$ - estimated biomass concentration.

Estimations of volumetric rates and biomass concentrations were done via state observer and available real-time measurements (MIR for JL2A, NIR for JL2B, and off-gas values for JL2A and JL2B).

Particle Filter (PF)

In order to solve the filtering problem a Bayesian approach, a so-called *Particle Filter* (*PF*) was applied in this work. The particle filter is a probability-based estimator which showed its good estimation ability for non-linear systems and is more flexible as the well-known *extended Kalman filter* (*EKF*)[xxxiii].

The main idea of *Particle Filtering* can be described as follows[xxxiii]:

Assuming that there is a non-linear system, where the state x_k and measurement y_k are following the equations:

$$x_{k+1} = f_k(x_k, w_k),\ y_k = h_k(x_k, v_k)$$

Where w_k and v_k are process and measurement noise, with known *pdf*'s, at each time step k, respectively.

Now, assuming that the *pdf* of the initial state $p(x_0)$ is known, a pool of *particles*, N, is created randomly. N particles correspond to N states and are denoted as $x_{0,i}^{+}$ $(i = 1, .., N)$.

For $k = 1,2,..$:

$$x_{k,i}^{-} = f_{k-1}(x_{k-1}^{+}, w_{k-1}^{i}), (i = 1, .., N),$$

Where noise vectors w_{k-1}^{i} are also generated randomly, based on known *pdf* of w_{k-1}.

When the measurement is received, conditional relative likelihood (q_i) that the measurement is equal to a specific measurement (y^*) can be calculated for each particle, as described by *Simon et al.*[xxxiii]:

$$q_i \sim \frac{1}{(2\pi)^{m/2}|R|^{1/2}} * exp\ (\frac{-\left[y^* - h\left(x_{k,i}^{-}\right)\right]^T R^{-1}\left[y^* - h\left(x_{k,i}^{-}\right)\right]}{2})$$

R is the measurement covariance matrix:

$$R = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_k^2 \end{bmatrix}$$

And, m-dimensional measurement equation is given as:

$$y_k = h(x_k) + v_k$$

And

$$v_k \sim N(0, R)$$

Relative likelihoods need to be normalized via:

$$q_i = \frac{q_i}{\sum_{j=1}^{N} q_i}$$

Next, particles are being resampled and a set of *a posteriori* particles $x_{k,i}^{+}$ is obtained and propagated to the next step.

Resampling can be done in many ways. In the current work, a multinomial resampling was used. Therefore, a random number r, distributed uniformly on [0,1] is created. After that, normalized likelihoods are summed until the sum is greater than r, which determines the selected particle. Illustration of the multinomial resampling procedure is given in Figure 16[xxxiii].

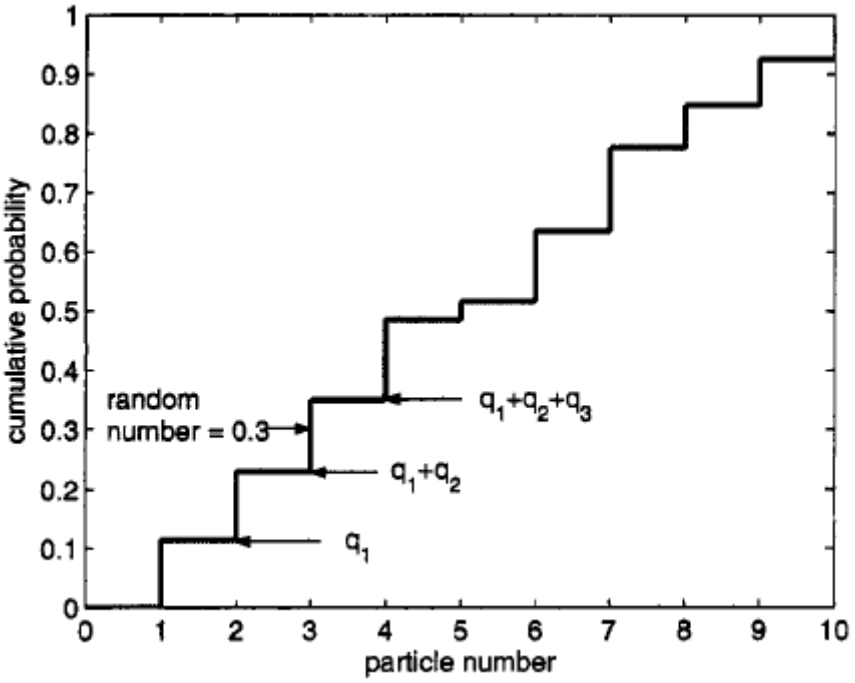

Figure 16[xxxiii]: Graphical interpretation of the resampling procedure used; q_i denotes particle of a number i

When $\sum_{l=1}^{j} q_l$ is greater than r, a new particle $x_{k,i}^{+}$ equal to $x_{k,j}^{-}$ is created.

In order to improve the resampling step and eliminate sample impoverishment (collapse of all particles into few or only one particle), roughening was used. Therefore, additional random Gaussian noise was added to each particle after it has been resampled. The noise was selected to be in the range of measurement accuracy.

Results and Discussion

Calibration experiments

Process overview

Typical profiles of feeding rates, temperature, pH, OUR, CER and measured oxygen concentrations, which were carried out during the fed-batch phase are presented in Figure 17 (AJ8B process is illustrated). Fed-batch flow-in, dissolved oxygen and agitator speed profiles can be seen in Figure 18 (AJ8B process is illustrated).

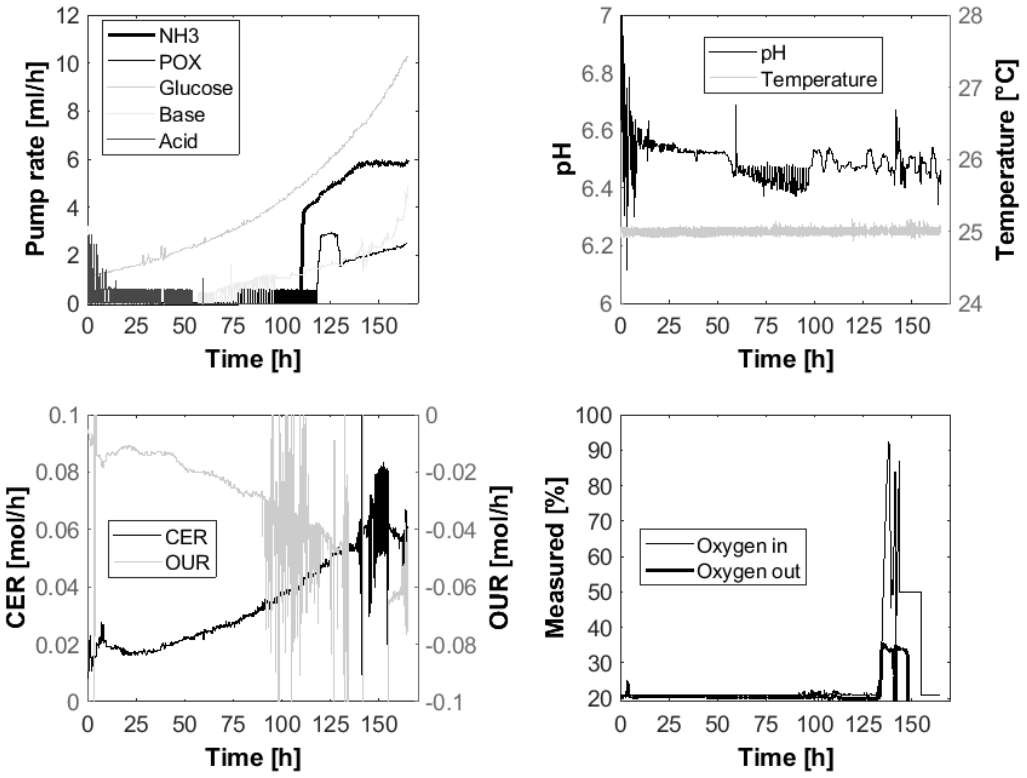

Figure 17: Typical profiles of feeding rates; pH; temperature; calculated CER and OUR; measured in/out oxygen concentration carried out during the fed-batch phase. AJ8B process is shown

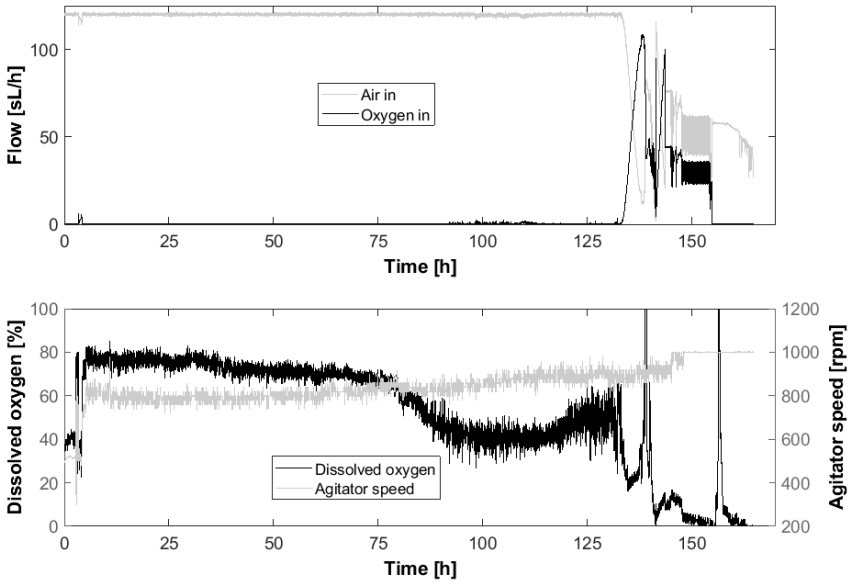

Figure 18: Typical profiles of air and oxygen flow in; dissolved oxygen and agitator speed carried out during the fed-batch phase. AJ8B process is shown

As can be seen, fed-batch fermentation processes, which were carried out during this work, included three substrate feeds (glucose, precursor and ammonia). These were controlled by the model or an expert. While the limiting substrate (glucose) was added during the whole fermentation, precursor and ammonia were first added after a certain time. This is explainable through the fact, that enough NH_3 and POX have already been present in the fed-batch medium at the beginning of the process. Temperature and pH were kept constant at 25±0.05 °C and 6.5 correspondingly.

The problem, which is often occurred during fermentations, can be seen in Figure 17 (measured in/out oxygen concentrations). When outlet oxygen concentration was higher than ca. 35-40%, the gas analyzer was not able to measure it. This led to wrong OUR estimation. Therefore, only CER could have been used for further simulations and kinetic model predictions.

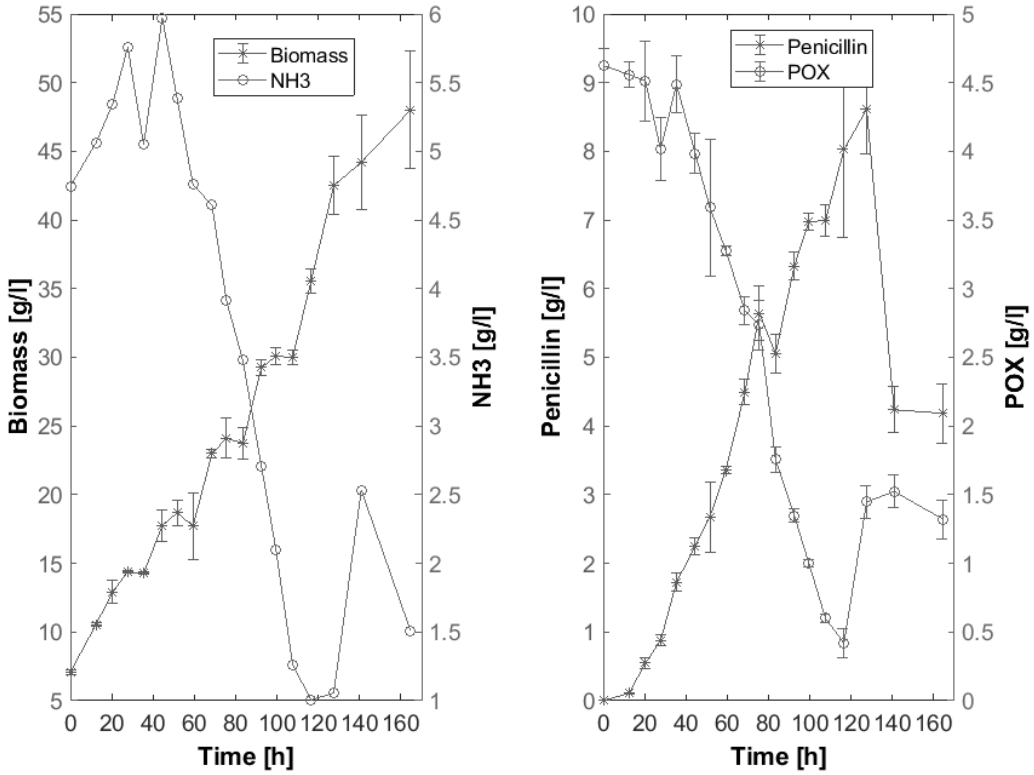

Figure 19: Typical penicillin, POX, ammonia and biomass profiles during the fed-batch phase; AJ8B process is shown

Typical concentration profiles of penicillin, POX, ammonia, and biomass during the fed-batch phase are presented in Figure 19 (AJ8B process is shown).

As it can be seen an increase of penicillin is followed by a decrease of the precursor. The concentration of ammonia decreases as well, while biomass is increasing.

Data consistency and quality

Due to the different feeding profiles, amounts of the fermented biomass and obtained product were not the same for different fermentations. Maximal values of measured penicillin and biomass concentrations, as well as minimal measured concentrations of precursor and ammonia, can be seen in Table 3.

Deviations from set-points in form of RMSE of such parameters as pH, temperature and dissolved oxygen concentration are presented in Table 4. Relatively high RMSE values for dissolved oxygen are stemming from the measurement outliers (when the oxygen concentration was increased very rapidly, the pO2 electrode has shown the values over 100%).

The generated experimental data was considered as consistent and applied for further evaluation procedures.

Table 3: Maximal values of measured penicillin and biomass, and minimal values of measured POX and NH_3

Experiment	Penicillin max [g/l]	Phenoxyacetate min [g/l]	NH_3 min [g/l]	Biomass max [g/l]
AJ7A	6.04	2.52	0.4469	38.1067
AJ7B	7.6533	2.6267	0.6138	36.6067
AJ7C	6.1467	3.12	0	42.38
AJ8A	7.7965	0.645	0.822	50.3267
AJ8B	8.613	0.4183	1.006	48.0133
AJ8C	3.6141	3.3067	4.71	14.7133
JL1A	9.206	0.096	1.485	24.25
JL1B	7.596	0.8973	1.9864	16.35
JL1C	9.5093	0	2.1042	16.35

Table 4: Deviation of measured pH, temperature and dissolved oxygen concentration from process set-points in form of RMSE

Experiment	pH RMSE [-]	Temperature RMSE [˚C]	Dissolved oxygen RMSE [% oxygen saturation]
AJ7A	0.0570	0.0376	36.6621
AJ7B	0.0509	0.0454	13.2835
AJ7C	0.0611	0.0400	12.5048
AJ8A	0.0894	0.0403	27.5370
AJ8B	0.0594	0.0438	27.6762
AJ8C	0.0898	0.0516	14.0993
JL1A	0.0997	0.0380	8.9440
JL1B	0.1054	0.0418	3.3026
JL1C	0.0963	0.0363	2.4653

PAT measurements

Table 5 demonstrates which measurements were available for the fermentation processes. Obviously, PLS models need training datasets and can be made only in case of the presence of IR measurements. Therefore, 3 processes were used as the training dataset for MIR PLS construction (AJ7A, AJ8A, JL1A), 4 processes were used for NIR based PLS (AJ8A, AJ8B, JL1A, JL1B), and 9 processes were used to estimate parameters of the kinetic model. Permittivity measurements were not taken into particle filter simulations as their transferability was too poor.

The verification experiments JL2A and JL2B were performed at the end of the work (see **Validation experiment**) and, therefore, were not used for the overall model constructions.

During AJ7A, NIR spectra were also measured, however other wavenumbers were recorded. Therefore, this process was not taken into consideration.

Table 5: Table of available data for each experiment

Experiment	CER	OUR	NIR	MIR	Offline (CDW, Penicillin, POX, NH_3)	Permittivity
AJ7A	+	+	-	+	+	-
AJ7B	+	+	+	-		-
AJ7C	+	+	-	-		-
AJ8A	+	+	+	+		-
AJ8B	+	+	+	-		+
AJ8C	+	+	-	-		+
JL1A	+	+	+	+		-
JL1B	+	+	+	-		+
JL1C	+	+	-	-		+
JL2A	+	+	+	+		-
JL2B	+	+	+	-		+

Permittivity measurements as a possibility of biomass estimation

As it was described before, permittivity measurements were performed for several processes (Table 5).

Measured permittivity allowed it to make a correlation based on a linear regression between permittivity values and biomass. However, this method works until a certain time point (ca. till 2/3 of the process), where permittivity starts to decrease (approximately 120 hours for AJ8B, see Figure 20) and does not correlate with biomass anymore.

Therefore, it was necessary to split the permittivity signal into two parts – increasing and decreasing part. After smoothing the raw permittivity signal via the *Savitzky-Golay* filter, the first derivative of the signal was calculated. A positive sign of the first derivative indicated an increasing part of the signal.

Calculating only with the increasing part of permittivity curve, linear regression based on the lowest sum of squares was performed (Figure 20).

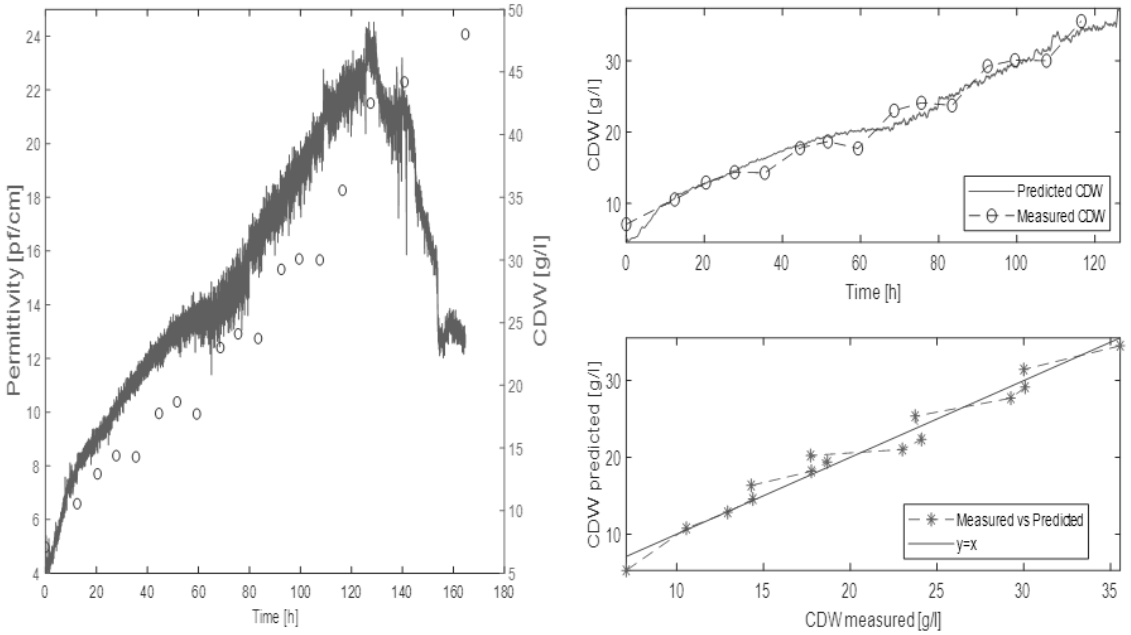

Figure 20: Permittivity vs. Biomass (left) and results of linear regression: y = k*1.6761 + (-2.4193) (right) (Process AJ8B, RMSE=1.43 g/l; BIAS = -6.8686e-15 g/l; SE = 2.1909; R2 = 0.96757)

Though the fact that such a model seems to have a strong prediction ability (low RMSE and high R-square value), it is absolutely not transferable to other experiments as the measurement is very sensitive to different operational conditions. Constructed equations in form of $y = k * x + b$, differ in slope (k), as well as in constant part (b) for different experiments. Model robustness is also very low, because, linear regression models are very sensitive to outliers.

Therefore, the permittivity measure shows its good workability only for relative control during a defined process. One can observe if the biomass is increasing, but any quantification, without additional calibration, does not seem to be possible. The only possibility here could be to use a possibly the same medium with same concentrations of feeds for the creation of a transferable and robust model for one well-defined process.

IR spectroscopy

As it was already mentioned in **Materials and Methods**, mid and near IR spectra were measured during the fermentation processes. Non-Invasive NIR measurements were measured in the range of 5100-7400 cm^{-1} and invasive MIR measurements were recorded in a range of 600-3000 cm^{-1}.

Liquid substances are assumed to be dissolved uniformly in the bioreactor, and their spectra can be measured directly. Solid substances could not be measured directly precise enough, in the described process, because of the process dynamics and applied instruments. Thus, soluble penicillin, NH_3, and phenoxyacetate were tried to be quantified via IR spectroscopy. Therefore, the evaluation procedure, described in **Introduction** and **Materials and Methods**, was carried out and PLS1 models were constructed in a similar way as described in the literature[xxvii], [xxxv].

Data pre-processing

Data pre-processing is necessary to align, normalize and smooth the data in order to prepare it for further evaluation.

PLS regression was made with a MATLAB *plsregress* function (which has a build-in mean-centering procedure). Therefore, mean-centering was always carried out when PLS models were constructed.

Such operation as SNV of online data was applied in several cases to improve PLS predictions. In Figure 46 and Figure 47 (**Supplement**), raw MIR and NIR spectra (of AJ8A and AJ7B consequently) can be seen. When SNV is applied it normalizes the spectra so that noise signal is reduced and background effects are eliminated. The effect of SNV on MIR and NIR spectra can be seen in Figure 43 and Figure 48 correspondingly **(Supplement**).

Another pre-processing method is Savitzky-Golay differentiation. The first and second derivative of a spectrum can carry important information about changes of concentrations of substances and their conversion rates changes which are not observable in a raw spectrum. In some cases, Savitzky-Golay differentiation can also reduce noise impact[xxxiv]. An example of spectra derivatives (first and second order) can be seen in Figure 44 and Figure 45 **(Supplement)**.

To verify whether these pre-processing procedures were necessary, combined datasets from all of the available experiments (each for MIR and for NIR spectra) were pre-treated and processed via PLS. Different combinations (SNV with 1st or 2nd derivative, SNV only, derivatives only) were tried. Errors and R^2 values were observed to determine the relevance of the procedure for each substance (penicillin, POX, and NH_3). Table 6 demonstrates these errors for MIR spectra based PLS, and Table 7 for NIR spectra based PLS. Not less important was to provide high model robustness, so RMSE was not the only criteria by the decision making (see **Robustness of a model**).

An example of how strong does data pre-processing can influence model results is shown on the NIR based PLS model for penicillin (Figure 21 and Figure 22).

Table 6: RMSE calculations for determining correct data pre-treatment method for MIR spectra PLS processing for a single experiment (mean-centering was always carried out due to MATLAB built-in procedure and is therefore not extra mentioned)

	RMSE [g/l]						Number of PCs
Substance	Sago 1st derivative	Sago 2nd derivative	SNV online data	Sago 1st derivative and SNV online data	Sago 2nd derivative and SNV online data	No data pre-treatment	
Penicillin	0.69323	0.68647	1.7588	1.7366	1.7698	0.6248	4
POX	0.80094	0.78998	0.79622	0.84871	0.84114	0.76076	3
NH_3	0.6655	0.63056	0.52677	0.51207	0.58862	0.71967	

Table 7: RMSE calculations for determining correct data pre-treatment method for NIR spectra PLS processing for a single experiment (mean-centering was always carried out due to MATLAB build-in procedure and is therefore not extra mentioned)

	RMSE [g/l]						Number of PCs
Substance	Sago 1st derivative	Sago 2nd derivative	SNV online data	Sago 1st derivative and SNV online data	Sago 2nd derivative and SNV online data	No data pre-treatment	
Penicillin	1.7041	1.6424	1.5163	1.5351	1.3308	1.718	4
POX	0.82361	0.78301	0.8373	0.84036	0.75373	0.7926	
NH_3	1.0142	1.0158	1.0344	1.0643	0.99659	1.0154	

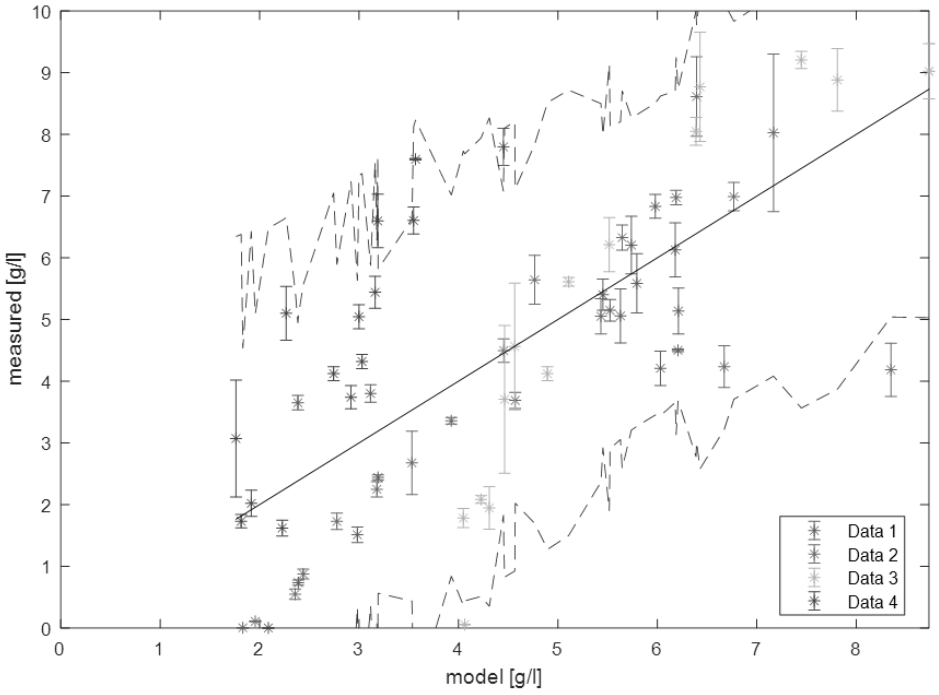

Figure 21: NIR based PLS model results for penicillin prediction (RMSE = 1.718 g/l). No data pre-treatment; 4 PCs; 4 datasets used (Data 1 -Data 4): AJ8(A,B), JL1(A,B)

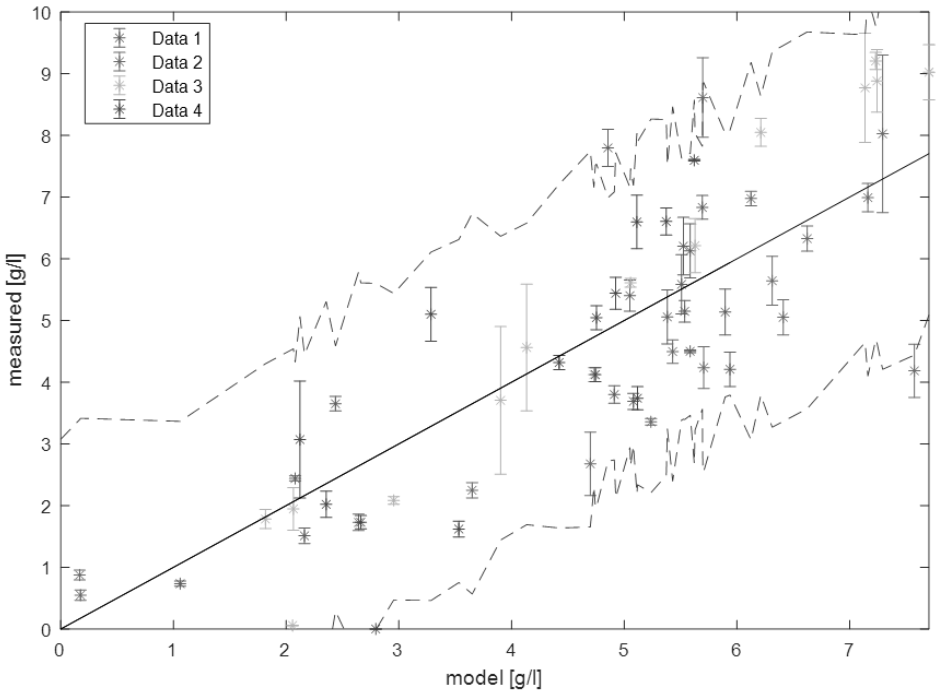

Figure 22: NIR based PLS model results for penicillin prediction (RMSE = 1.3308 g/l). Sago 2nd derivable and SNV of spectra; 4 PCs; 4 datasets used (Data 1 -Data 4): AJ8(A,B), JL1(A,B)

Based on calculated RMSE and error variations (see **Robustness of a model**), following data pre-processing methods were chosen (Table 8):

Table 8: Data pre-treatment procedures chosen

	Penicillin		**POX**		**NH_3**	
	MIR	**NIR**	**MIR**	**NIR**	**MIR**	**NIR**
SNV	-	+	-	+	-	
Savitzky-Golay differentiation (derivative order if applied)	1st order	2nd order	-	2nd order	1st order	
Mean-centering (build-in MATLAB function)	+					

Choosing the optimal wavenumber region

Variable (wavenumber) selection is a crucial step for building a PLS model based on spectral data. Not all of the variables carry important information and some of them can only represent noise in certain cases.

The simplest way of wavenumber selection is to use the fingerprint region for each measured substance. This possibility was also considered, based on the literature[xxxv, xxxvi].

Nevertheless, there are several reasons, why this method is not (or not always) the best one. First of all, some substances (as penicillin for example) not always have known and/or strong absorption in NIR spectral range. Secondly, in case of analyzing, only the known absorption range of a substance, information about possible intermediates or secondary products and its concentration changes over time, which can be carried out by PLS, can be lost. Finally, dealing with multianalyte systems, possible absorption shifts, caused by molecular interactions must not be excluded.

Therefore, optimal wavenumber selection was carried out based on PLS weights (only in case of MIR based PLS for penicillin, wavenumbers were selected as described in the literature[xxxv]). Moreover, not only the raw spectra, but also 1st and 2nd derivatives of each spectrum were taken into account while selecting the wavenumber range. The selection was carried out by choosing overlapping regions with the highest weight values of raw and derivatives of raw spectra.

Table 9: Wavenumber selection results

	Wavenumber [cm^{-1}]		
	Penicillin	**POX**	**NH_3**
MIR	1307-1352 & 1747-1817[xxxv]	1000-1150 & 1400-1750 & 1900-2600	850-1150 & 1400-1520 & 1950-2300
NIR	6050-7500 & 5200 -5700	5100-5500 & 6100-6400 & 6500-7500	6200-7500

Wavenumber selection results are presented in Table 9. As it can be seen some of the wavenumber regions are overlapping within different substances. This is also explainable through the fact that these substances and their concentration changes are dependent on each other in the current fermentation process.

PLS models, based on all experiments, available for the certain kind of measurement (NIR or MIR), as well as models based on single experiments, which were built with chosen wavenumbers, showed better performance in form of smaller RMSE and higher R^2 values then the models where the whole measured wavenumber range was used as variables.

This step was determinative for the robustness and transferability of further constructed PLS models. Despite described procedures, which were made to reduce the number of variables, models were still overfitted and derived noisy results. More complex algorithms could have been applied in order to reduce overfitting of the models. However, this would request additional training datasets which were not available during the current work.

Robustness of a model

A not less important part of building a proper PLS model is choosing the number of *Principal Components* which are used. If too many PCs are used, the model is overfitted and contains unnecessary information (such as noise for example).

In the current work, 10% of offline data was used as a test set with 10 Monte-Carlo repetitions by building each model. On the presented example (Figure 23) it can be seen that starting from the fifth principal component error values begin being unstable and varying with a number of repetitions. This indicates that NIR based PLS model (based on all available historical NIR data) with 5 or more principal components is unstable and is overfitted with noise and therefore carries unwanted information.

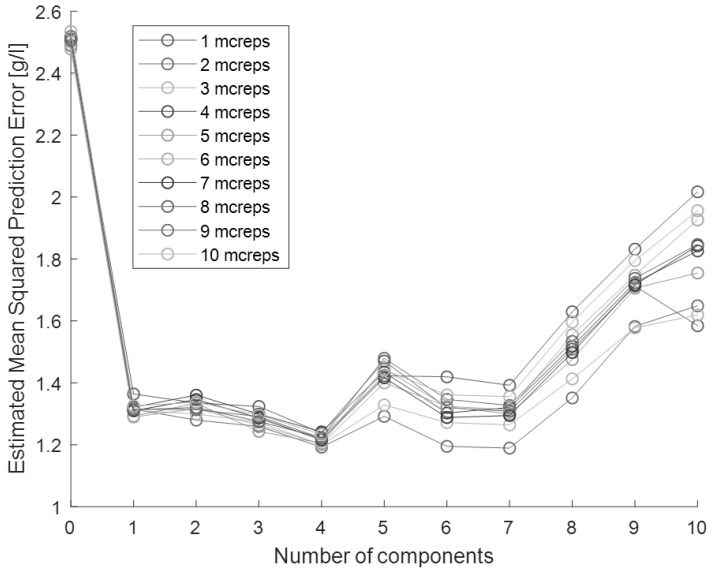

Figure 23: Predicted errors vs. PC number. NIR based PLS for NH3, based on all available NIR data, 'mcreps' denotes Monte-Carlo repetitions

Construction of a model

In order to be able to predict concentrations of different substances during the fermentation process, robust and possibly transferable PLS models are needed. These model properties are dependent on the training dataset which is used by model creation. Training data should be consistent and, ideally, must not contain outliers. Beyond that, the size of training dataset should be large enough. In order to increase the size of training dataset, results of all fermentation processes which were available for a certain type of measurement (MIR or NIR) were combined, and used for PLS model constructions.

The MATLAB script written, allows the addition of a new data as soon as it is available for model update

The original data matrix is ordered as it can be seen in Figure 24. For building a complete model, based on all available experiments, data matrices were ordered after each other so that all of them had the same variables (wavenumbers) with corresponding adsorption values.

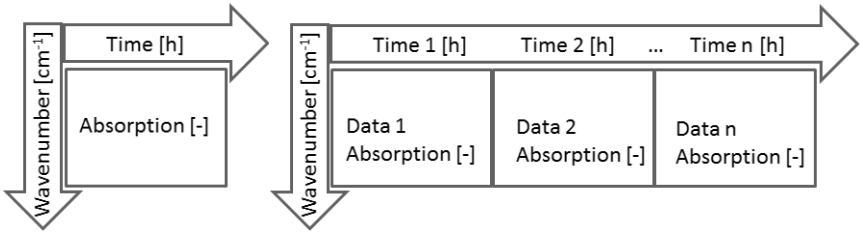

Figure 24: Data matrix orientation (left – original, right – used by the creation of a model based on multiple experiments)

To perform PLS, offline values were matched together with corresponding absorption values, as it is shown in Figure 25, giving the training dataset.

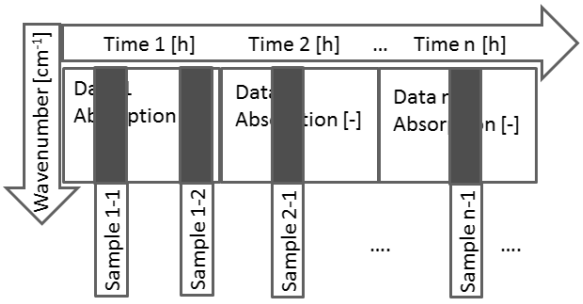

Figure 25: Matching offline values with the corresponding absorption for PLS construction

The main result of a constructed PLS model is a beta matrix. The beta matrix contains information about the relationship between a set of absorption values by different wavenumbers and concentrations of a certain substance. To apply a constructed PLS model on the unknown dataset, the corresponding beta matrix should be multiplied with the pre-proceeded absorption values. This results in the concentration values of the desired component in the same units as used by model creation. The schematic illustration of this procedure is presented in Figure 26.

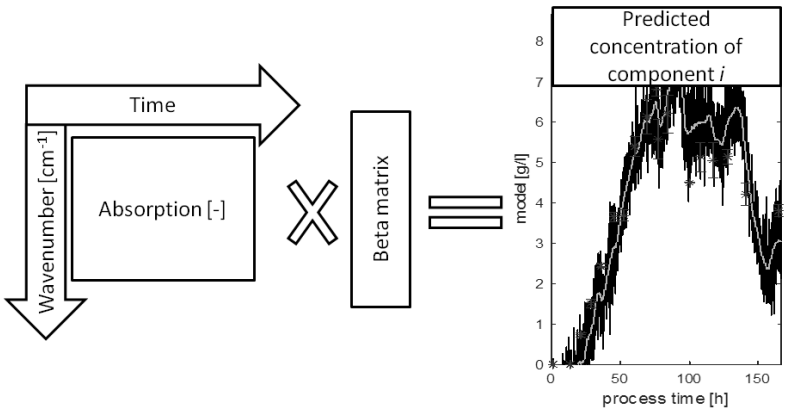

Figure 26: Schematic illustration of the application of a PLS model on spectral data

MIR based PLS

As described in section **Construction of a model**, the overall model, based on all MIR measurements (for processes AJ7A, AJ8A, JL1A) was constructed.

Plotted results of overall models can be seen in Figure 49, Figure 50 and Figure 51 in **Supplement**. Calculated errors and R-square values are presented in Table 10. Penicillin and ammonia NRMSE are lower 10%, and POX NRMSE is lower 20%. Low BIAS values denote low systematic error and high R-square highlights good prediction of the model.

The obtained beta-matrix was then applied again to every single data set.

As it can be seen from Table 11, PLS predictions for penicillin and POX, show high errors in JL1A experiment (NRMSE for penicillin is 11%, however, NRMSE for POX and ammonia are 26% and 23% correspondingly).

The media composition of JL1A experiment was different from the one which was used during AJ7A and AJ8A. Also, the POX addition during the JL1A experiment was problematic as feed-lines were blocked. This led to the fact that as the total number of samples carried out during AJ7A and AJ8A was larger than the one for JL1A, PLS model was overfitted with AJ7A and AJ8A data and shifted JL1A predictions in the same direction.

AJ8A predictions obtained are acceptable for all three substances measured (NRMSE for penicillin and ammonia are under 10% and NRMSE for POX is under 20%).

NRMSE for AJ7A experiment are lower 20% for penicillin and ammonia, and over 30% for precursor.

Table 10: Errors and R-square values of constructed PLS models based on all MIR data

Substance	RMSE [g/l]	BIAS [g/l]	SE [g/l]	R-square [-]	Number of PCs
Penicillin	0.69323	5.1112e-15	0.4898	0.92854	4
POX	0.76076	7.2898e-16	0.58989	0.62305	3
NH_3	0.40665	2.5137e-17	0.16854	0.92653	4

Table 11: Errors and R-square values when a constructed PLS model was applied to the historical dataset

Experiment	RMSE [g/l] / R-square [-]		
	Penicillin	POX	NH3
AJ7A	0.37323 / 0.9683	0.64054 / -0.62417	0.27433 / 0.92215
AJ8A	0.6024 / 0.92479	0.4996 / 0.85473	0.32339 / 0.97115
JL1A	1.0305 / 0.89689	1.0996 / 0.34888	0.59707 / 0.30271

NIR based PLS

Same computations as in case of MIR were made with the available NIR dataset (AJ8A, AJ8B, JL1A, JL1B processes). Table 12 illustrates the results of the overall model. Plotted results of overall models can be seen in Figure 52, Figure 53 and Figure 54 in **Supplement**. Low BIAS values denote low systematic error of the model.

Table 12: Errors and R-square values of constructed PLS models based on all NIR data

Substance	RMSE [g/l]	BIAS [g/l]	SE [g/l]	R-square [-]	Number of PCs
Penicillin	1.3308	4.2576e-15	1.7995	0.71305	4
POX	0.75373	-6.6261e-16	0.57728	0.68233	
NH_3	1.0142	2.8196e-16	1.0452	0.57616	

As it can be seen, prediction errors of NIR based PLS for penicillin and ammonia are higher than the ones for MIR based PLS. Penicillin NRMSE for NIR based PLS model (based on all MIR data) is lower 20%, but is still 7% higher than the one for MIR based PLS model (based on all MIR data). NRMSE for ammonia is 19.7% for NIR based PLS model (based on all NIR data), which is much higher than 7.6% for MIR based PLS model (based on all MIR data). NRMSE for POX are approximately the same for both (NIR and MIR based PLS models) and are lower 20%.

Higher prediction errors for NIR based PLS models can be caused by several reasons. First, MIR spectra contain more information about the structure of the molecule[xvi], comparing to NIR. Secondly, NIR instruments applied were measuring spectra through the reactor glass wall which could lead to lower sensitivity due to additional light scattering and reduction of the light signal.

The constructed overall model was applied to every single dataset, and calculated RMSE and R-square values can be seen in Table 13.

Table 13: Errors and R-square values when a constructed PLS model was applied to the historical dataset

Experiment	RMSE [g/l] / R-square [-]		
	Penicillin	POX	NH3
AJ8A	1.1073 / 0.70164	0.65066 / 0.73017	0.8756 / 0.78485
AJ8B	1.9508 / 0.41767	0.66596 / 0.77935	0.94823 / 0.697
JL1A	2.7527 / 0.17379	1.7333 / -0.58275	0.78422 / -0.12379
JL1B	3.4075 / -1.5993	1.6605 / -1.0801	1.3915 / -0.42385

NRMSE for AJ8A and AJ8B experiments are acceptable and are lower 30% for all substances.

However, less precise results of NIR based PLS predictions were achieved for JL1A and JL1B. NRMSE for JL1A and JL1B experiments are over 30% for all substances (except the penicillin prediction for JL1A).

This can be caused by different media composition used in AJ8 and JL1 experiments. Another reason could be different operational conditions during AJ8 and JL1 experiments. During the AJ8 experiment, there was a constant precursor feed (starting from a point where the POX concentration itself was low, Figure 17). In the case of JL1, no POX feed was applied due to technical reasons. As PLS is a purely data-driven approach, it does not take feed profiles into account (in contrast to the kinetic model), while calculating the resulting concentrations of substances. The total number of offline samples, which were taken during both AJ8 experiments is higher than the one for JL1. This

means that the PLS algorithm is overfitted with AJ8 data and tries to shift the results of JL1 in the same direction.

Thus, another two models, based only on AJ8 (A and B) and JL1 (A and B) were made. NRMSE for all compounds and processes became, therefore, lower than 30%.

Results of applying these models to historical datasets can be seen in Table 14.

Table 14: Errors and R-square values when two constructed PLS models (AJ8 and JL1 based) were applied to historical dataset

Experiment	RMSE [g/l] / R-square [-] / Number of PCs			Model
	Penicillin	POX	NH3	
AJ8A	0.98156 / 0.76554 / 4	0.63362 / 0.74412 / 4	0.85225 / 0.79617 / 2	AJ8
AJ8B	1.2648 / 0.75519 / 4	0.5559 / 0.84626 / 4	0.99684 / 0.66513 / 2	
JL1A	0.67352 / 0.95054 / 4	0.38996 / 0.91989 / 4	0.40048/ 0.70693 / 6	JL1
JL1B	2.6971 / -0.62852 / 4	0.72238 / 0.60635 / 4	0.71575 / 0.62328 / 6	

In spite of lower PC number for AJ8 based PLS used by constructing a model for NH_3, RMSE remained almost the same (AJ8B) or even became lower (AJ8A). In case of JL1 based models for NH_3 estimation, too many PCs were needed for proper predictions, which lead to noisy models and even low RMSE values should not be considered as the main criterion (see Figure 27).

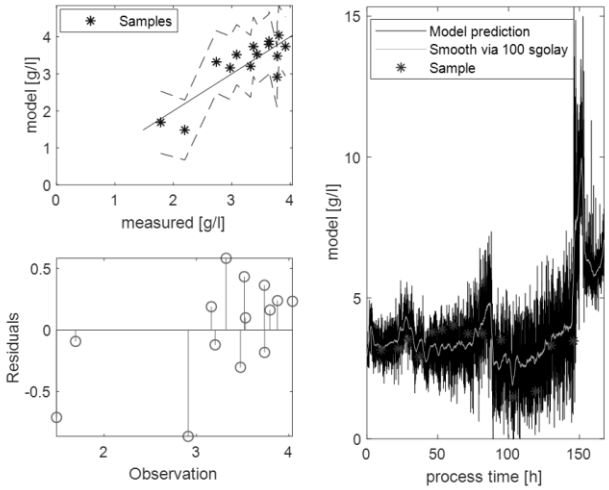

Figure 27: NH3 prediction for JL1A based on NIR (JL1) PLS model

Construction of separated PLS models for AJ8 and JL1 experiments resulted in good estimations, but the transferability of these models is questionable and can only be determined by a validation experiment.

PLS models cannot be applied outside the conditions used for their development. Therefore, additional fermentations with NIR measurements should be done and data has to be collected. Having more data, noise impact is going to be reduced, and more robust and transferable PLS models can be constructed and applied.

Kinetic modeling

Parameter estimation

Identification of parameters of a kinetic model is not a trivial problem. In the current work, empirically defined parameters[xxx] were used. However, some of the model parameters are non-identifiable having no single solution and therefore possessing a high variance. Therefore, identifiable parameter sets should have been determined.

Parameter estimation was done in a similar way as described in the literature[xxxvii]. After model definition, prior analysis and calculation of model outputs (see **Materials and Methods**), sensitivities were computed, and then parameters were ranked based on their influence on the model outputs. Results of both of these steps are shown in Figure 28.

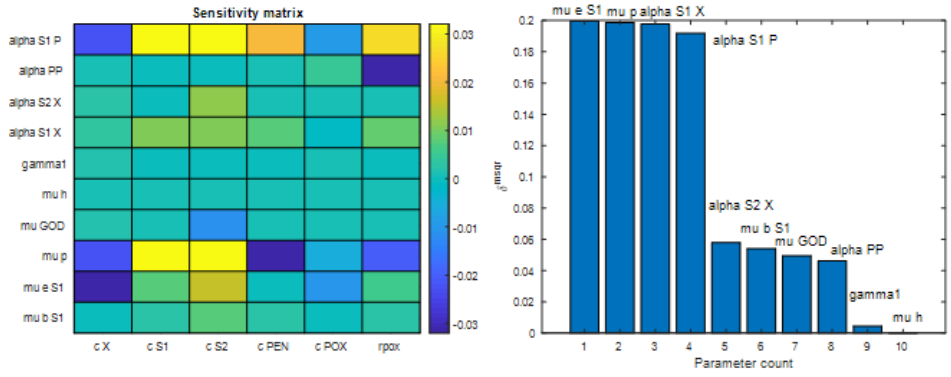

Figure 28: Results of sensitivity analysis (left): sensitivity vs. states (color indicates values); and computation of parameter importance (right): sensitivity measure vs. parameter; mu e S1, mu p, alpha S1 X, alpha S1 P, alpha S2 X, mu b S1, mu GOD, alpha PP, gamma 1, mu h denote μ_e, μ_p, α_0, α_p, α_{Gln}, μ_0 , μ_{GOD}, $\alpha_{POX/PEN}$, γ_1, μ_h respectively

Blue and yellow colors of the sensitivity matrix (Figure 28, left) denote high negative and positive correlation of the state to the corresponding parameter, respectively. According to the calculated δ_j^{msqr} values, the three most high-ranked parameters were μ_e, μ_p and α_0 (Figure 28, right). Afterwards, identifiable parameter sets, which should have been estimated, were chosen based on the calculated determinant values. Based on collinearity index calculations, maximal acceptable parameter set size was 3 (with a defined threshold of 10 for the collinearity index). Identifiable parameter sets are presented in Table 15.

Table 15: Identifiable parameter sets (decreasing rank order). Parameter sets with determinant values over 1.6 were considered as best identifiable

Parameter set (decreasing order of identifiability)	Determinant value [-]	Collinearity index [-]
μ_e, μ_p	**1.6326**	**1.7128**
$\alpha_{POX/PEN}$, α_p	**1.6209**	**1.2287**
μ_p, $\alpha_{POX/PEN}$	**1.6201**	**1.2948**
μ_{GOD}, α_0, $\alpha_{POX/PEN}$	1.1575	1.6899
α_0, α_{Gln}, $\alpha_{POX/PEN}$	1.1552	1.7168
μ_p, μ_{GOD}, α_0	1.152	1.8121

Results presented here, are based on average calculations during the whole process. In fact, dealing with a dynamic process (fed-batch profile), parameter sensitivities are varying over time (Figure 55, **Supplement**). Best identifiable parameter sets were defined as the ones with the lowest collinearity index (below 10) and a determinant value over 1.6 (the first three parameter sets in Table 15).

Calculation of errors for estimated parameter sets

Relative errors for the best identifiable parameter sets (Table 15, parameter sets with determinant values over 1.6), were calculated as described in **Materials and Methods** and can be seen in Table 16. Calculated errors were very low (under 1%) and therefore the estimation of the illustrated parameters was carried out (Table 16). Estimation of these parameters was done via a simplex algorithm (MATLAB 2018a: *fminsearch*), minimizing the weighted residual sum of squares. Parameter values are presented in the **Supplement**.

Table 16: Errors calculated via *FIM* for the best identifiable parameter sets. JL1A experiment was used as a data source

Parameter set (decreasing order of identifiability)	Relative parameter errors [%] (separated by semicolon)
μ_e, μ_p	0.25963; 0.17207
$\alpha_{POX/PEN}$, α_p	0.39516; 0.0098812
μ_p, $\alpha_{POX/PEN}$	0.44175; 0.021181

Thus, identification and estimation of the best identifiable parameters led to good kinetic model predictions. An example of model simulation for one selected process is presented in Figure 29.

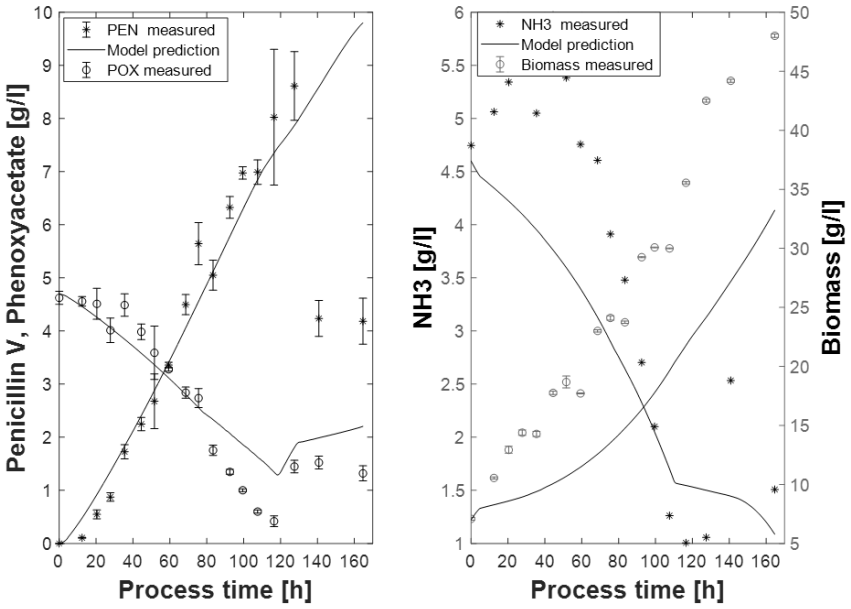

Figure 29: An example of kinetic model application to a selected dataset (AJ8B). Prediction NRMSE for biomass, NH3, POX and penicillin are 26%, 22%, 20% and 13% respectively

Observability index

Observability plot for an applied kinetic model (see **Materials and Methods** and **Introduction**) is shown in Figure 30. Underlying calculations were made according to the literature[xxxviii],[xxxix].

Presented figures illustrate which measurements are necessary for the estimation of different model states, which are dependent on each other through kinetic model equations as described in **Materials and Methods**. Plots are calculated for the following measurements: concentrations of biomass, penicillin V, phenoxyacetate and ammonia as well as CER and OUR. Higher observability index on the Y-axis denotes higher information content. Black and white colors denote an absence or a presence of the measurement correspondingly.

It can be seen that the penicillin measurement is always required.

Most of the measurements (concentrations of ammonia, penicillin, biomass, precursor) needed for states description are usually done offline. However offline sampling is not always applicable for process control as biomass estimation via cell-dried-weight methods needs at least 72 hours. Penicillin, POX and ammonia measurements can be done more quickly, but are still time-consuming. This could be crucial in case of a running fermentation process.

It can be seen that the system, described by the model, is not observable if the off-gas data only (CER and OUR) is available.

Thus, in order to carry out a proper model-based process monitoring, there was a need in (online) measurements of as many compounds as possible (the more measurements are present, the better observability index is reached).

As described before in **Process overview** off-gas measurements for oxygen, could not have been performed properly, because of the gas-analyzer limitations.

Therefore, a set of available measurements has to be used. PLS models based on NIR or MIR spectra are able to predict penicillin, precursor, and ammonia. Together with CER, this allows observing a model, with high observability index (Figure 30).

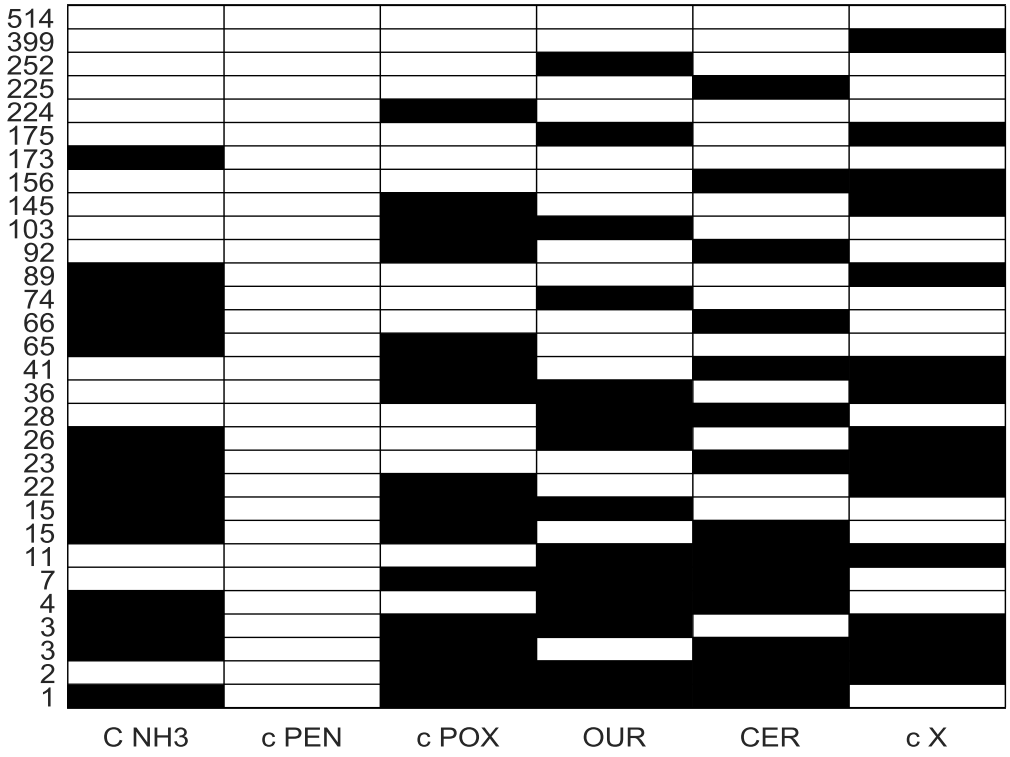

Figure 30: Observability plot for the applied kinetic model. Following states can be estimated by the model: S1 (glucose), S2 (gluconate), biomass, penicillin, POX, NH_3. White colored squares denote that the measurement (concentration of ammonia, penicillin, POX, biomass; CER and OUR) is present. Black colored squares denote the absence of the measurement. Y-axis represents the observability index

Comparison of different monitoring strategies

Observers based on the kinetic model and CER

Off-gas measurements (CO_2 and O_2) can improve model estimations when combined with a kinetic model and a filtering algorithm.

As it was already mentioned, oxygen off-gas measurements were not performed correctly as the used gas analyzer could not detect oxygen amounts above ca. 35-40%. Therefore, the simplest particle filter configuration was simulated based only on the kinetic model and CER. This combination does not provide a completely observable system (see **Observability index**). Nevertheless, concentrations of precursor, ammonia, and biomass could still be estimated.

Start concentrations (of penicillin, POX, glucose, and biomass) were set according to HPLC measurement results. Noise addition to particle filter can be seen in Table 17.

Table 17: Configuration of a particle filter for CER and kinetic model based simulations

Parameter	Process state noise (absolute)								Measurement noise (absolute)
	Volume [l]	A0 [g/l]	A1 [g/l]	Glucose [g/l]	Gluconate [g/l]	PenV [g/l]	POX [g/l]	NH_3 [g/l]	CER [mol/l*h]
Value	0.0001	0.1	0.2	0.05	0.05	0.001	0.001	0.001	0.005

Errors and R-square values of these simulations are presented in Table 18.

Table 18: Prediction errors of particle filter simulations, when combined with CER

Experiment	RMSE [g/l] / R-square [-]		
	POX	NH_3	Biomass
AJ8A	1.2033 / 0.15729	0.89808 / 0.7775	4.0133 / 0.86511
AJ8B	0.65459 / 0.79821	1.1689 / 0.52385	1.526 / 0.9828
AJ8C	2.37 / -14.2486	3.9834 / -43.5212	3.3078 / -2.013
AJ7A	1.3276 / -6.5957	0.86061 / 0.2413	4.2962 / 0.85061
AJ7B	1.091 / -7.9815	1.0286 / -0.50608	2.7674 / 0.92634
AJ7C	0.81942 / -2.8973	0.79002 / 0.51357	4.6497 / 0.87076
JL1A	0.56448 / 0.8284	1.6685 / 4.4451	4.7016 / 0.12336
JL1B	0.94359 / 0.2868	2.2021 / -2.6741	2.3613 / -0.22682
JL1C	1.2897 / -0.091632	1.6506 / -1.6472	2.2433 / -0.10723

Biomass prediction errors are acceptable for all processes (below 30%), however, there is an underestimation shift which appears due to the lack of real-time information (Figure 31).

Despite low R-square values and unacceptable errors for precursor (almost for all processes, except AJ8A, AJ8B, JL1A), it can be seen (Figure 31) that model fits quite well during the first 2/3 of the process, and starting from a point of approximately 100 hours, predicted concentrations are not precise enough anymore.

Ammonia predictions are also acceptable only for several processes (AJ7C, AJ8A, and AJ8B).

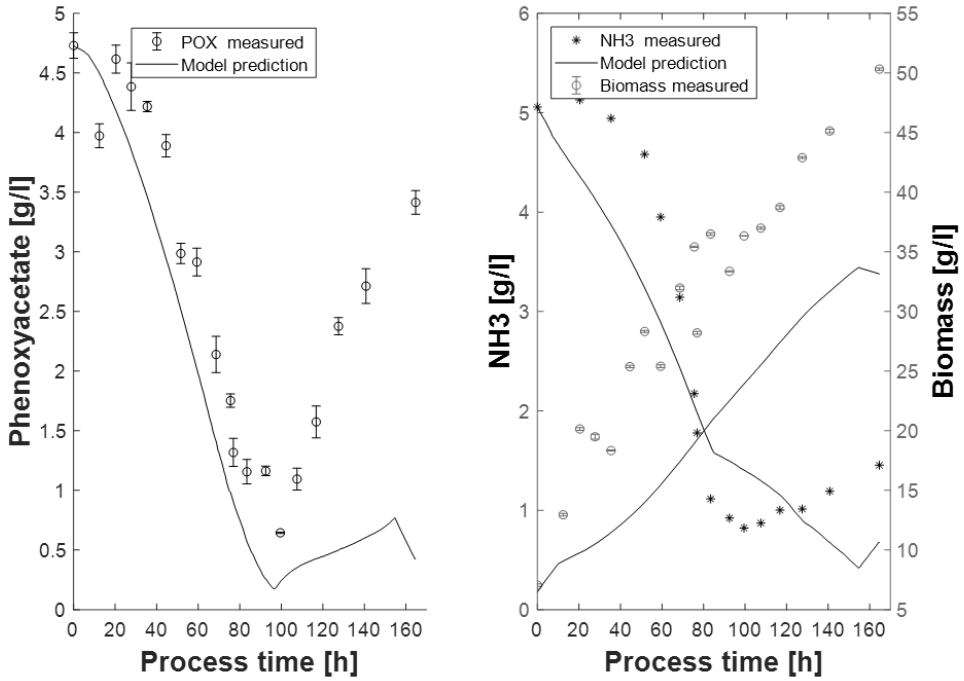

Figure 31: Predicted and measured concentrations of POX, ammonia, and biomass; State estimator with CER (process AJ8A is shown)

However, as it was already mentioned, it is impossible to estimate all main model states (penicillin concentration in particular) based on CO_2 off-gas measurements only. Thus, the main goal of these simulations was an acceptable estimation of biomass concentrations which have been actually achieved.

Observers based on spectral and off-gas measurements

In order to achieve estimation results, spectral data, which was transformed via PLS into concentrations of different substances (see **IR spectroscopy**) was combined with off-gas measurements and added to the particle filter during the model simulation. As it was mentioned already, this guarantees full observability.

In contrast to the previous paragraph (**Observers based on the kinetic model and CER**), where only one online measurement was taken into account, different measurements here were evaluated for their accuracy while included to the particle filter with measurement errors. Therefore, absolute errors calculated for the processes were used at the beginning as start points to create the particles. However, to improve the prediction ability particle filter was slightly tuned during several simulations and particle filter configuration can be seen in Table 19 and Table 20.

Table 19: Configuration of a particle filter when combined CER, spectral data and kinetic model, process state noise

Parameter	Process state noise (absolute)							
	Volume [l]	A0 [g/l]	A1 [g/l]	Glucose [g/l]	Gluconate [g/l]	PenV [g/l]	POX [g/l]	NH_3 [g/l]
Value	0.0001	0.1	0.2	0.005	0.005	0.4	0.2	0.3

Table 20: Configuration of a particle filter when combined CER, spectral data and kinetic model, measurement state noise

Parameter	Measurement noise (absolute)						
	CER [mol/l*h]	Penicillin FTIR [g/l]	POX FTIR [g/l]	NH_3 FTIR [g/l]	Penicillin NIR [g/l]	POX NIR [g/l]	NH_3 NIR [g/l]
Value	0.005	0.6	0.5	0.3	0.8	0.7	0.4

Before a simulation, based on all available spectral and online data, was performed, it was important to see, how precise would be the simulation, based on the NIR or MIR data only. This question is interesting, because the applied NIR measurements were made non-invasive (in contrast to MIR, where a light-conducting cable was put into the reactor through one of the ports) by fixing micro spectrometers on the reactor glass wall. Such an approach by itself (besides lower costs of the device) is increasing the operational safety and reduces costs for extra ports in the reactor.

Kinetic model, CER & NIR (non-invasive)

As mentioned above, in order to check whether only NIR measurements can significantly improve the results of the kinetic model, simulations with NIR and CER results were performed.

Prediction errors for these simulations are presented in Table 21.

Table 21: Prediction errors of particle filter simulations, when combined with non-invasive NIR measurements

Experiment	RMSE [g/l] / R-square [-]			
	Penicillin	**POX**	**NH_3**	**Biomass**
AJ8A	1.7254 / 0.066278	0.80205 / 0.58211	0.97172 / 0.72184	4.9198 / 0.69377
AJ8B	1.2953 / 0.76421	0.90796 / 0.61176	1.0358 / 0.62608	3.5864 / 0.90502
JL1A	2.466 / 0.4095	1.7213 / -0.5956	0.61237 / 0.2665	6.8326 / -0.85136
JL1B	3.099 / -0.84659	0.72887 / 0.57446	2.0364 / -2.1421	1.7161 / 0.35199

In comparison to simulations, based on the kinetic model and CER only, the addition of NIR data improved prediction results significantly. First, all of the model states became observable. Secondly, lower penicillin, precursor, and ammonia prediction errors were achieved. Finally, concentration profiles became more correct (see **Observers based on the kinetic model and CER**).

Prediction errors for all substances are acceptable for AJ8A and AJ8B processes. Process JL1A predictions were acceptable only for penicillin and ammonia, and JL1B process predictions were acceptable only for biomass and precursor. Better estimations for AJ8 processes are caused by better PLS predictions for this datasets. Unacceptable errors for JL1A and JL1B are also caused by the prediction deviations at the end of the processes (similar as described in **Observers based on the kinetic model and CER**).

Kinetic model, CER & MIR (invasive)

Secondly, particle filter simulations for the three processes, where MIR data was available (AJ7A, AJ8A, JL1A) were done.

Prediction errors for these simulations can be seen in Table 22.

Table 22: Prediction errors of particle filter simulations, when combined with MIR measurements

Experiment	RMSE [g/l] / R-square [-]			
	Penicillin	**POX**	**NH_3**	**Biomass**
AJ7A	0.52207 / 0.94219	0.52883 / - 0.20529	0.28312 / 0.91789	3.7103 / 0.88858
AJ8A	1.6237 / 0.17312	0.614 / 0.7551	0.70839 / 0.85217	4.8764 / 0.69915
JL1A	0.9317 / 0.91571	1.2161 / 0.20362	0.653 / 0.16596	6.6098 / -0.73259

It can be seen that compared to observer configurations discussed in the previous paragraphs (**Observers based on the kinetic model and CER, Kinetic model, CER & NIR (non-invasive)**), prediction accuracy has been improved significantly.

This is caused through more precise MIR based PLS estimations, compared to NIR based PLS as well as smaller absolute errors which were set by particle filter configuration. Thus, the state estimator gives more weight to PLS estimations and shifts the model to correct values. Predictions made for all substances are acceptable for processes AJ7A and AJ8A. In the case of JL1A process, biomass and ammonia prediction errors were over 30%, which was caused by bad PLS estimations.

Kinetic model, NIR (non-invasive) & MIR (invasive)

In order to achieve the best state estimation accuracy, all spectral data were processed via PLS and added to the state observer with off-gas measurements. This was done for the two experiments where MIR, NIR, off-gas and offline data were available: AJ8A and JL1A.

Prediction errors are presented in Table 23.

Table 23: Prediction errors of particle filter simulations, when combined with MIR and NIR

Experiment	RMSE [g/l] / R-square [-]			
	Penicillin	**POX**	**NH_3**	**Biomass**
AJ8A	1.1453 / 0.58861	0.51263 / 0.82929	0.62319 / 0.88559	4.262 / 0.77018
JL1A	1.8364 / 0.67254	1.4708 / -0.16502	1.0097 / -0.99414	4.9147 / 0.042118

As can be seen, a combination of MIR and NIR based PLS models have dramatically improved AJ8A simulation results. NRMSE for all substances became below 20%. Penicillin errors became more than 3 times lower, compared to state observer with CER. POX and ammonia prediction errors became ca. double and quarter less, respectively. The only less precise estimation for AJ8A process (compared to predictions of state estimator with CER) is biomass concentration. This is explained through the fact that as soon penicillin is present (and its high concentration is given to the model as a true state), the model assumes high biomass growth and overestimates its concentration. In order to improve this deviation, further kinetic model improvements should be made.

Simulation results for JL1A process became better or remain the same, compared to other observer combinations. Still, as already mentioned before, technical problems occurred during JL1 (different media composition and blocked precursor feeding line) have disturbed PLS predictions. Therefore, several things can be made here: PLS model improvement through further experiment under the same conditions and kinetic model adaptation (taking into account the loss of productivity over time in particular).

Discussion

Performance of different measuring methods

In the current work, lots of different measuring methods were introduced and applied during the fermentation processes. Also, the models based on these measures are different in their performance and quality. Table 24 illustrates the relative characteristics of different measuring instruments and methods.

The offline analysis is the most precise, stable and selective method, but it lacks on frequency.

In contrast to offline measurements, off-gas can be measured in real-time, but it is noisy and therefore is not so precise. Beyond that, off-gas measurements contain the information about the overall cell metabolism only, and, therefore, cannot be used for estimation of single metabolites, such as penicillin, precursor and ammonia, in the described process.

The kinetic model showed its good performance, however, as it was shown in **Observers based on the kinetic model and CER,** kinetic model lack real-time information. Process deviations are not included and a kinetic model does not take into account the loss of productivity at a certain time point.

PLS, based on spectral methods may be very precise for concentration predictions, but its transferability is strongly depended on the amount of data which is available for model building. It was shown that PLS models are often overfitted with noise. Data-driven nature of PLS does not allow it to take into account feeding rates and other dynamic process factors. Moreover, spectral data obtained in this work can also give no information (PLS models can be still constructed, but its prediction ability is very poor) about solid substances. Therefore, no biomass predictions could have been made here with the usage of spectral data only.

State-estimator implementation allows a combination of advantages of all of the approaches described and makes it possible to reach the result which none of these methods can show alone.

Table 24: Relative properties of different measuring methods/approaches

Measurement	Selectivity	Precision	Transferability	Real-time	Stability
CER	++	++	+	+	++
OUR	++	++	+	+	+
NIR based PLS	++	+	+	+	++
FTIR based PLS	++	++	++	+	++
Offline (HPLC)	+++	+++	-	-	+++
Permitivity	++	+	-	+	++
Kinetic model	++	+	+++	+	+++
Observer	**+++**	**++**	**+++**	**+**	**+++**

Comparison of different strategies applied

Figure 32 illustrates observed vs. predicted plots for all of the different models applied (kinetic and PLS models). As described before not all of the processes could have been used for PLS models construction. Presented PLS models are based on all of the spectral data available for the corresponding type of measurement (3 datasets for MIR and 4 datasets for NIR).

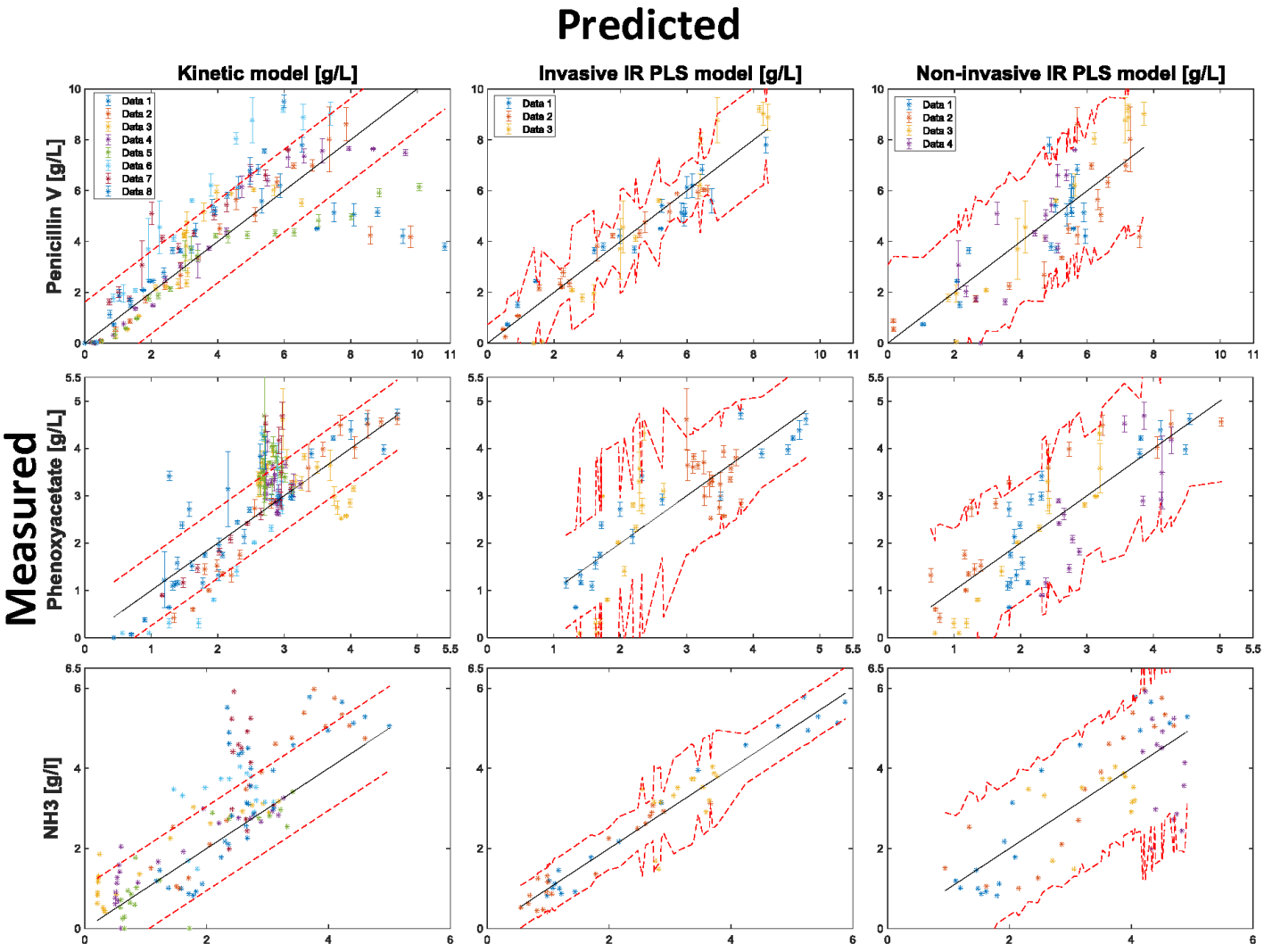

Figure 32: Observed vs. predicted for different estimation strategies (kinetic model, NIR based PLS, MIR based PLS). Dotted lines denote prediction confidence intervals for PLS and RMSE for the kinetic model. 8 datasets (Data 1 – Data 8) - AJ7(A,B,C), AJ8(A,B), JL1(A,B,C) were simulated for kinetic model; 3 datasets (Data 1 – Data 3) – AJ8(A,B), JL1(A) were simulated for MIR; 4 datasets (Data 1 – Data 4) – AJ8(A,B), JL1(A,B) were simulated for NIR

Prediction errors calculated for the processes from Figure 32 are presented in Table 25.

Table 25: Prediction errors of the three used approaches

Method / number of datasets	RMSE [g/l] / R-square [-]		
	Penicillin	**Phenoxyacetate**	**Ammonia**
Kinetic model / 8	1.62 / 0.61	0.74 / 0.59	1.05 / 0.55
MIR based PLS / 3	0.69 / 0.93	0.76 / 0.62	0.41 / 0.93
NIR based PLS / 4	1.33 / 0.71	0.75 / 0.68	1.01 / 0.58

As can be seen, NIR based PLS predictions and kinetic model estimations are approximately in the same error range. MIR based PLS gives better results for all compounds, except phenoxyacetate.

In order to be able to predict biomass, which cannot be predicted via spectral based PLS, the kinetic model must be applied. When combined with CER, it results in good biomass estimates, but lacks effective penicillin and POX predictions. Comparison of CER and kinetic model based observer with spectral based PLS predictions for one selected process is presented in Figure 33.

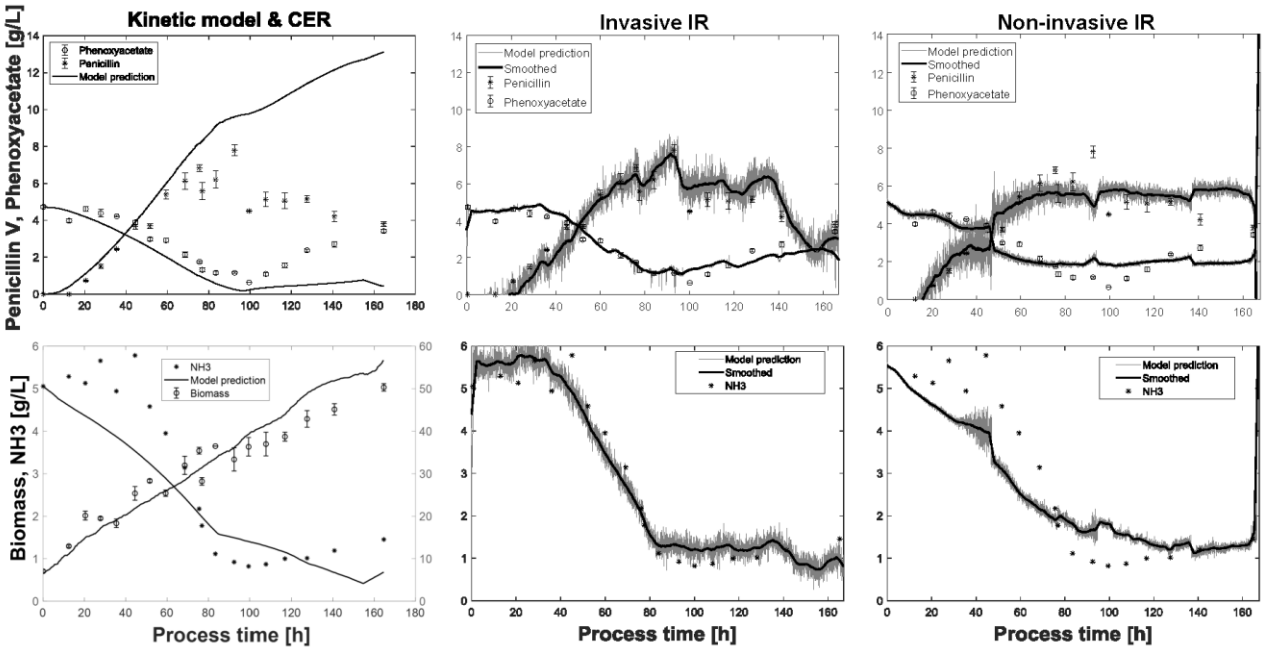

Figure 33: Penicillin, POX, NH_3, and biomass (can only be predicted by kinetic model) predictions of different models when applied to the AJ8A process (prediction NRMSE for penicillin, precursor, ammonia and biomass, correspondingly, are 51.0%; 29.5%; 18.1%; 9.28% for the kinetic model and CER; and 7.73%; 12.2%; 6.52% and 14.2%; 15.9%; 17.6% for penicillin, precursor and ammonia for MIR and NIR based PLS, correspondingly)

As there is no direct information about penicillin concentration (this state is not observable), when combining a kinetic model with CER, spectral-based PLS models still give better results.

Therefore, spectral measurements were also added to the observer. Figure 34 and Figure 35 illustrate results of addition of NIR and MIR based PLS predictions to the particle filter together with the kinetic model and CER, correspondingly.

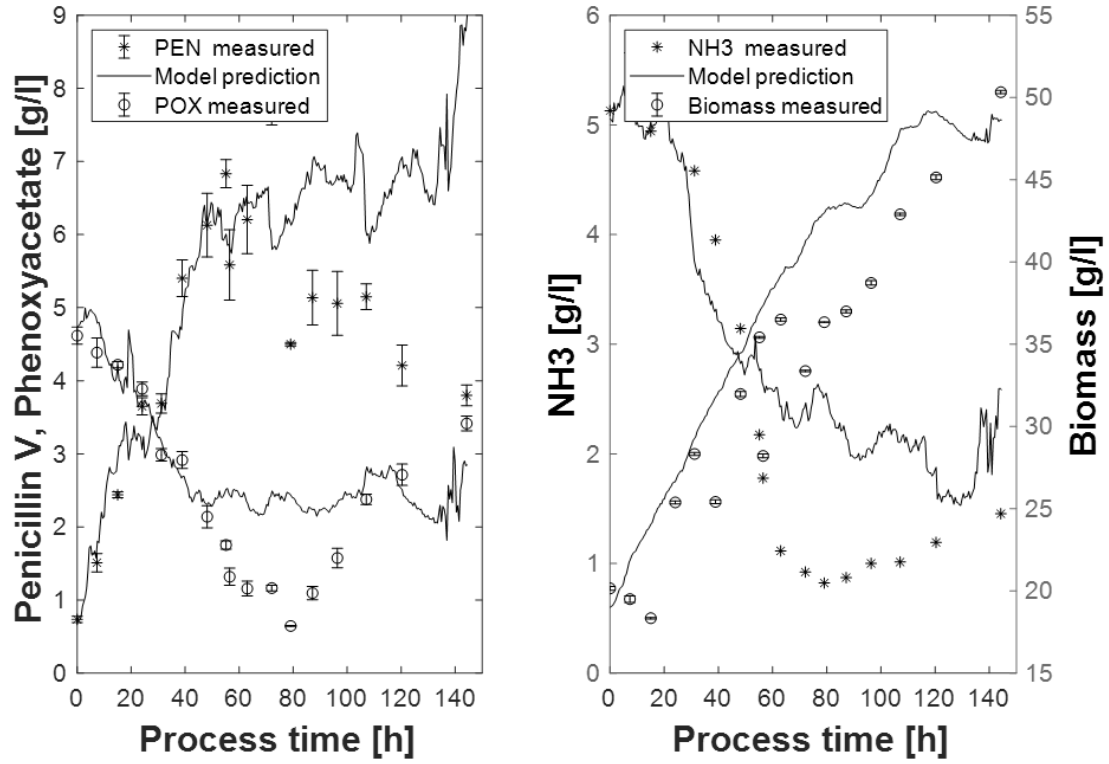

Figure 34: NIR based PLS predictions (of concentrations of penicillin, POX, and ammonia over time) combined with the kinetic model and CER via observer (AJ8A process). NRMSE for penicillin, POX, ammonia and biomass are 22.1%; 19.6%; 19.6%; 11.4% correspondingly

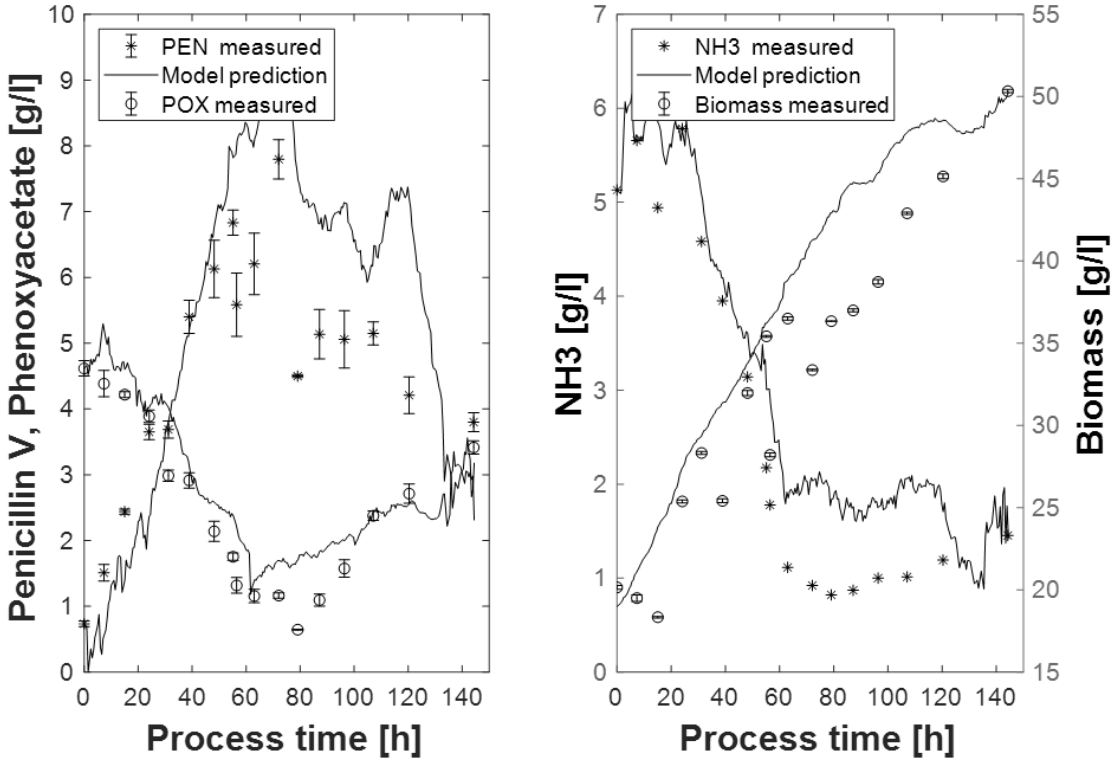

Figure 35: MIR based PLS predictions (of concentrations of penicillin, POX, and ammonia over time) combined with the kinetic model and CER via observer (AJ8A process). NRMSE for penicillin, POX, ammonia and biomass are 20.8%; 15.0%; 14.3%; 11.3% correspondingly

It can be seen that after PLS results were added to the state observer, biomass became overestimated. However, penicillin and phenoxyacetate predictions became more precise. The model was corrected and the particle filter shifted the penicillin concentrations to lower values. As NIR based PLS is more noisy and error-prone, it has less weight while combined with the kinetic model and penicillin shift is smaller as in MIR based PLS.

Figure 36 shows predictions based on all constructed models (kinetic model, NIR and MIR based PLS).

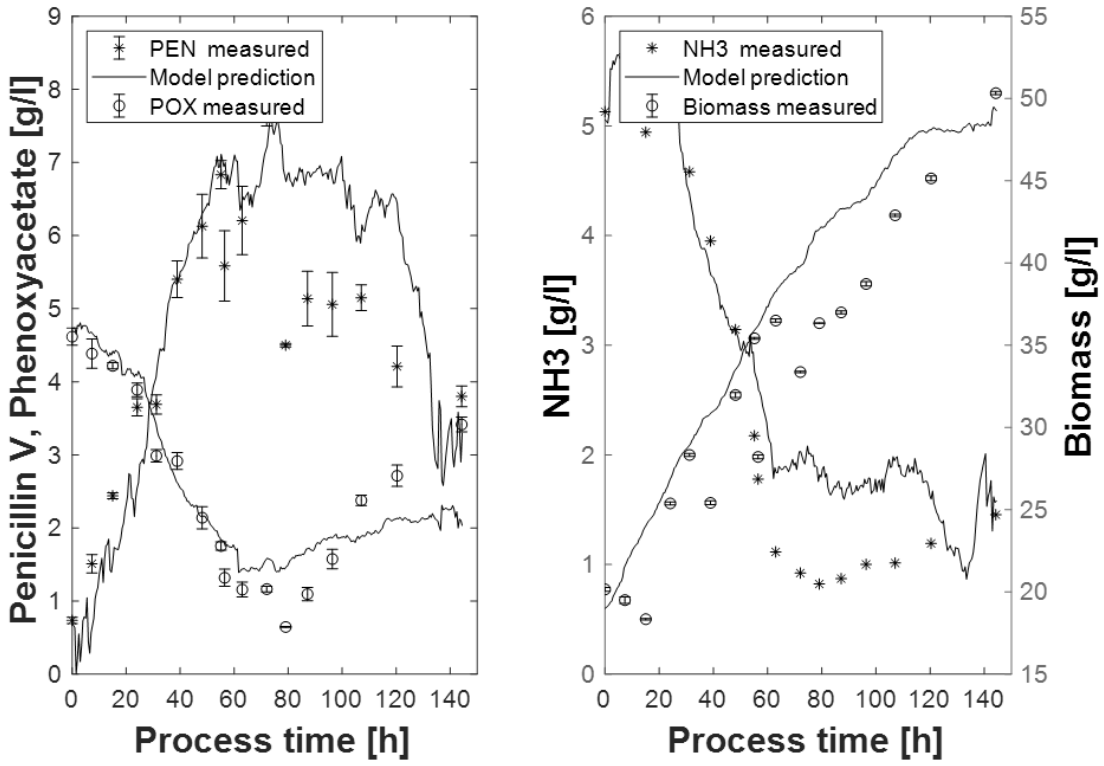

Figure 36: NIR and MIR based PLS predictions (of concentrations of penicillin, POX, and ammonia over time) combined with the kinetic model and CER via observer (AJ8A process). NRMSE for penicillin, POX, ammonia and biomass are 14.7%; 12.5%; 12.6%; 9.9% correspondingly

As it can be seen from Table 26 and Table 27, when NIR based PLS is combined with CER and kinetic model via state observer it can predict concentrations of all of the substances of interest with a prediction NRMSE below 30%. Ammonia prediction errors remain in the same range and biomass prediction is getting worse (compared to **Observers based on the kinetic model and CER**). This result is better than the one obtained from the spectral data only (no biomass prediction possible) or the one based only on state observer with CER – as penicillin errors there are too high.

When MIR measurements are added to an observer, it increases the accuracy of predictions of all the substances (besides biomass) compared to **Observers based on the kinetic model and CER**.

Table 26: Prediction errors for penicillin, POX, ammonia, and biomass of different methods when applied to AJ8A process

	RMSE [g/l]			
	Penicillin	**Phenoxyacetate**	**NH_3**	**Biomass**
State observer	3.97	1.14	0.90	4.01
MIR (invasive IR)	0.60	0.50	0.32	-
NIR (non-invasive IR)	1.11	0.65	0.88	-
State observer with NIR based PLS	1.73	0.80	0.97	4.9
State observer with MIR based PLS	1.62	0.61	0.71	4.88
State observer with MIR and NIR based PLS	**1.15**	**0.51**	**0.62**	**4.26**

It can be seen that a combination of all spectral measurements with CER and kinetic model leads to an optimal estimation of the most important states. Data-driven PLS model shifts a kinetic model to lower penicillin values and therefore corrects the model, which assumes a constant production rate. On the other side, the underlying kinetic model does not contain as much noise as PLS models and the obtained predictions are relatively smooth. Combining these two methods leads to an observable and robust model with prediction errors which are in the same range or lower than errors of any single method.

Table 27: Normalized prediction errors for penicillin, POX, ammonia, and biomass of different methods when applied to AJ8A process

	NRMSE [%]			
	Penicillin	**Phenoxyacetate**	**NH_3**	**Biomass**
State observer	50.9	29.5	18.1	9.28
MIR (invasive IR)	7.73	12.2	6.52	-
NIR (non-invasive IR)	14.2	15.9	17.6	-
State observer with NIR based PLS	22.1	19.6	19.6	11.4
State observer with MIR based PLS	20.8	15.0	14.3	11.3
State observer with MIR and NIR based PLS	**14.7**	**12.5**	**12.6**	**9.85**

Validation experiment

As it was already mentioned in **Materials and Methods**, in order to verify the ability of the developed system to control the process, a set of validation experiments was performed (JL2A and JL2B).

The first goal was to construct a fully automated set-up, which could be controlled with particle filter predictions based on the underlying kinetic model and real-time results of different measurements, such as spectral measurements and off-gas results. The underlying online architecture is illustrated in Figure 14 in section **Materials and Methods**.

Although the combination of all measurements showed the best results, experiments with NIR and MIR measurements only were conducted to compare their performances and the corresponding PLS models, when applied for process control.

The success of validation experiments could be submitted if pre-defined limiting substrate biomass specific rate set-points remained constant during the process. It was also important to keep the concentrations of ammonia and precursor at the constant, non-limiting level as these are affecting the penicillin production rate.

Results of the validation experiment are described in the following two paragraphs.

Validation of MIR based PLS combined with kinetic model and off-gas data

First, the system based on the MIR spectra and off-gas results is going to be described.

This was the case of JL1A fed-batch process. The feeding profiles, as well as the calculated specific rate and other fermentation results, are presented in Figure 37.

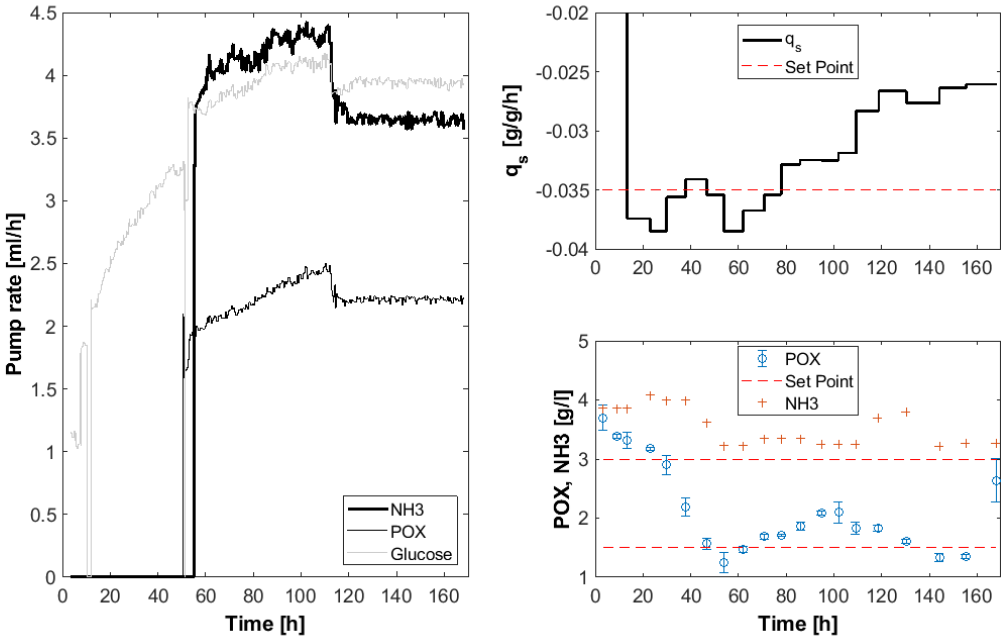

Figure 37: Fermentation results of JL1A experiment

As it can be seen in Figure 37, glucose was kept limiting and q_s has also remained constant between the values of -0.04 and -0.03 *g/g/h* near to set point of -0.035 *g/g/h* (with a slightly decreasing absolute value at the end of the process). Lower absolute q_s values were reached at the beginning of the process (not shown for better visualization of the main phase).

Concentrations of ammonia and precursor were kept non-limiting and constant (note that the feed control for NH_3 and POX was first turned on after ca. 55 hours).

However, PLS models, based on historical data from AJ7A, AJ8A, and JL1A have not resulted in good estimations for all substances, as illustrated in Figure 38.

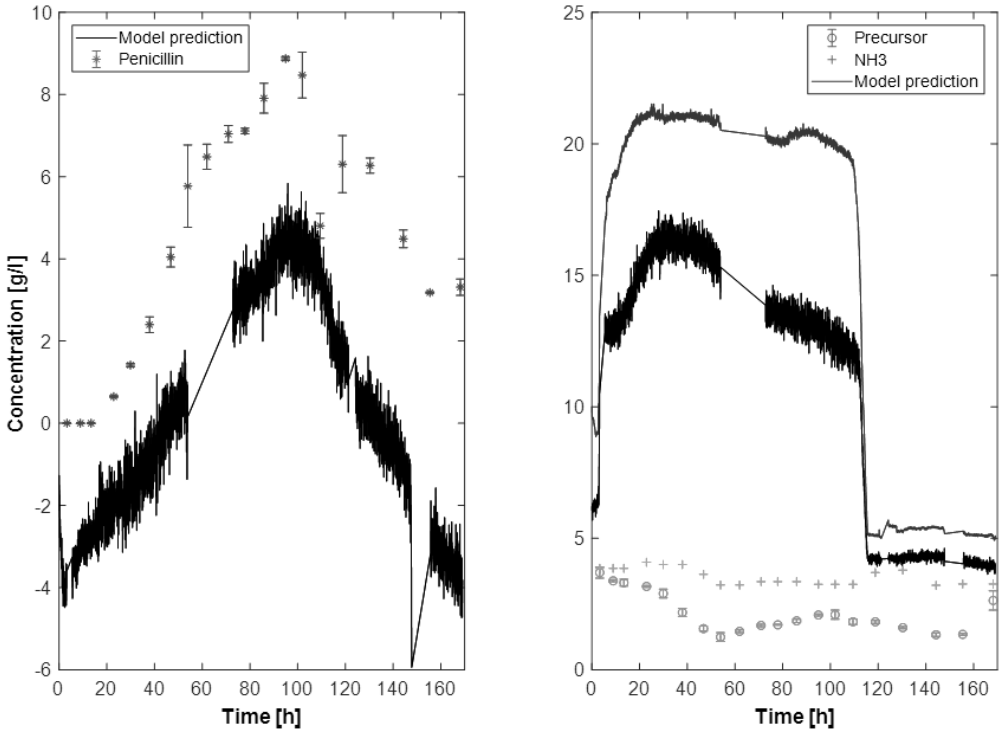

Figure 38: PLS predictions of penicillin, POX, and ammonia for JL2A based on historical data

Penicillin prediction of MIR based PLS was the best for all three substances (Figure 38) as POX and ammonia did not agree with the off-line data at all, even in their relative profiles. However, there is an obvious off-set even for penicillin, which should be eliminated while applying these results for process control.

Possible reasons for this result could be more correct chosen wavenumbers for penicillin and the fact that penicillin is presented in higher concentrations and therefore shows a higher absorption. In order to predict POX and ammonia with the current PLS algorithm two things could be done: generating more data through further fermentations and choosing other wavenumbers via more complex algorithms.

PLS calculations were started at the beginning of the process. After ca. 55 hours test run, it has been decided to change the particle filter setting by increasing the error of PLS prediction. This resulted in stable and good process control. Particle filter estimations are shown in Figure 39.

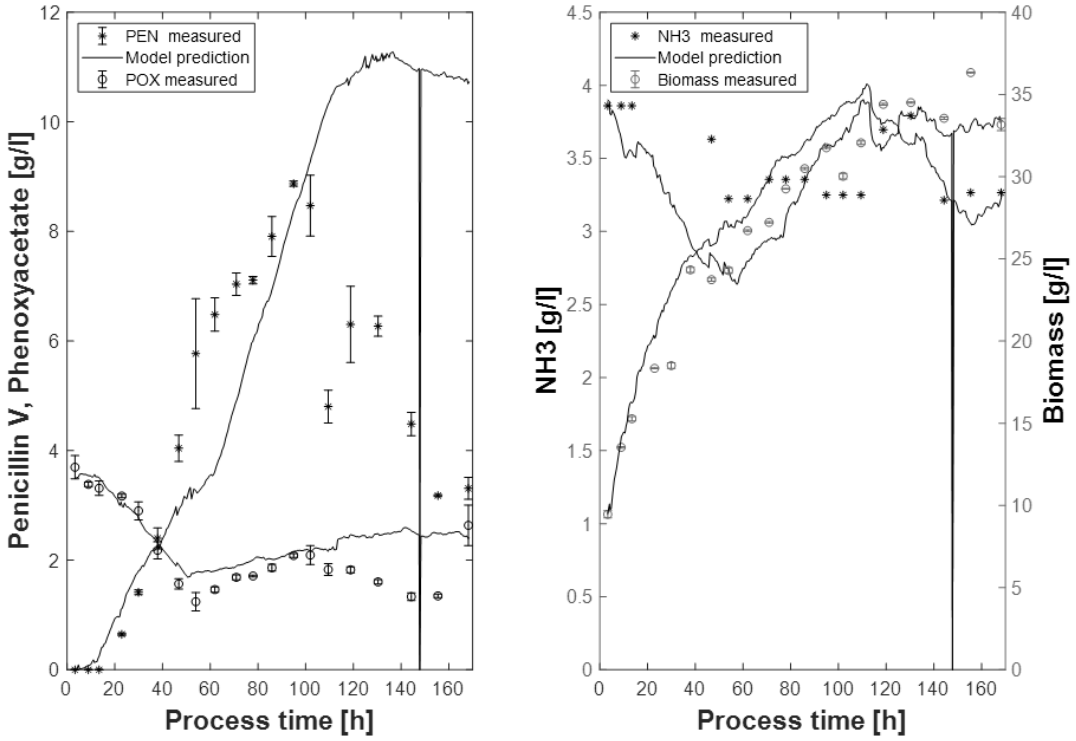

Figure 39: Particle filter estimations and off-line measured concentrations of penicillin, POX, NH3 and biomass for process JL2A

Calculated errors of particle filter predictions are shown in Table 28.

Table 28: Errors of particle filter estimations for JL2A process

Experiment	RMSE [g/l] / R-square [-]			
	Penicillin	**POX**	**NH_3**	**Biomass**
JL2A	3.2514 / 0.81522	0.49766 / 0.55192	0.50427 / -1.6128	3.2514 / 0.81522

Obtained results are comparable to the past ones for historical data (AJ7, AJ8, and JL1 process sets). Penicillin predictions are correct for at least 2/3 of the process. However, when fungi start to produce less, the model is still shifting particle filter estimations to higher values. This results in unacceptable prediction NRMSE. Biomass is properly predicted with low error (below 20%), due to real-time results of off-gas measurements. POX estimations are acceptable as the prediction NRMSE is below 30%. Prediction NRMSE for ammonia is over 30%, because of wrong PLS predictions.

Therefore, validation experiment JL2A was successful in a sense that constructed set-up was fully automatized and controlled via particle filter. Despite not-efficient PLS predictions for penicillin and ammonia, particle filter was stable enough to make proper estimations which were used for a good feeding control strategy.

Validation of NIR based PLS combined with kinetic model and off-gas data

Process JL2B was controlled with NIR based PLS and off-gas results via particle filter. As NIR based PLS models showed lower performance while applied to the historical data set, lower prediction accuracy was expected. Fermentation results of JL2B process can be seen in Figure 40.

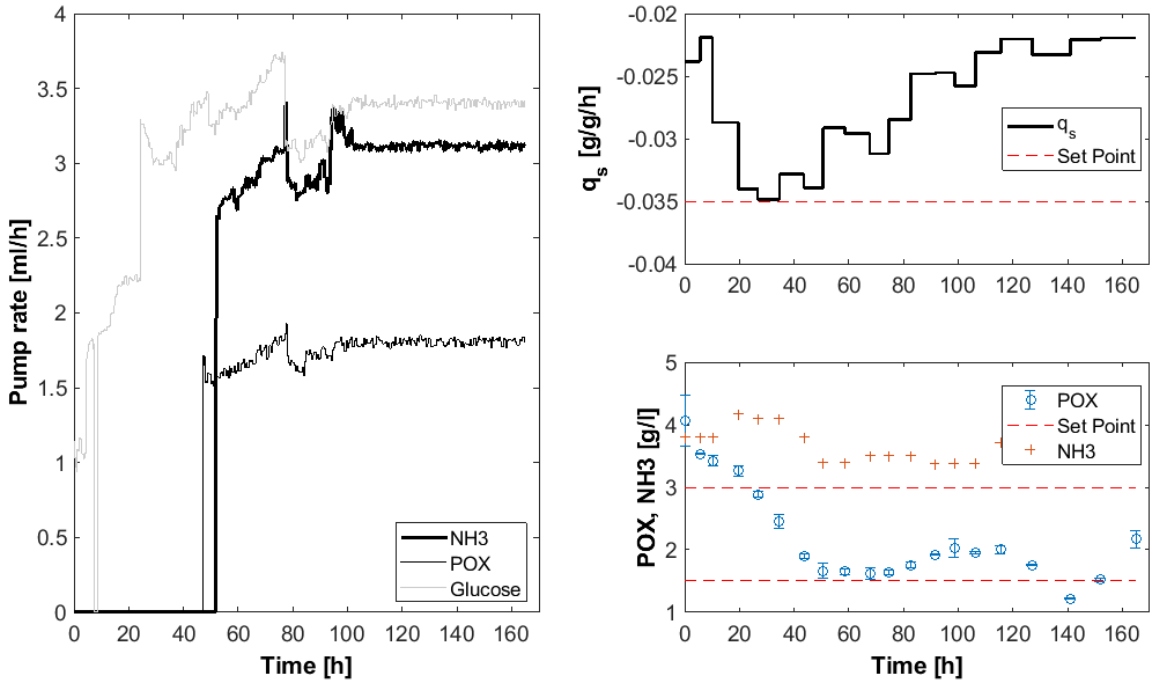

Figure 40: Fermentation results of JL2B experiment

As can be seen, concentration profiles of ammonia and precursor remained constant and non-limiting starting from the time point when the control was turned on (approximately 55 hours after the process start).

Lower absolute q_s values were reached, because of biomass underestimation. Nevertheless, specific uptake rate of glucose remained approximately constant near the set-point of -0.035 *g/g/h* at least for the half of the process (with a slightly decreasing absolute value at the end, similar to JL2A).

Thus, proper control during this fermentation run was achieved.

PLS prediction results based on historical data (JL1A and JL1B) are shown in Figure 41.

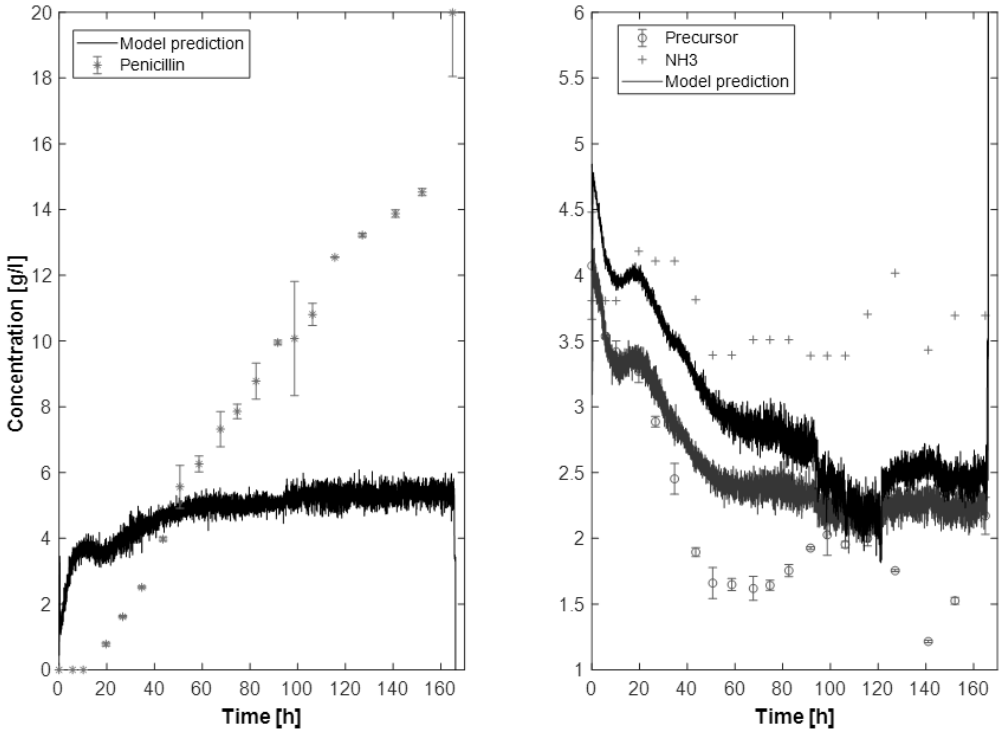

Figure 41: PLS predictions of penicillin, POX, and ammonia for JL2B based on historical data

In contrast to MIR based PLS predictions for JL2A (Figure 38), NIR based PLS did not lead to proper penicillin predictions. On the other hand, predicted concentration curves for POX and ammonia are decreasing, which matches with expectations and off-line results. However, NIR based PLS models seem to be too overfitted as their predictions are just repeating the results of historical data and a simple offset addition cannot lead to correct estimations in this case.

Therefore after approximately 55 hours of a test run, particle filter configuration was tuned in the same way as for JL2A process – PLS prediction errors were increased. Particle filter estimations can be seen in Figure 42.

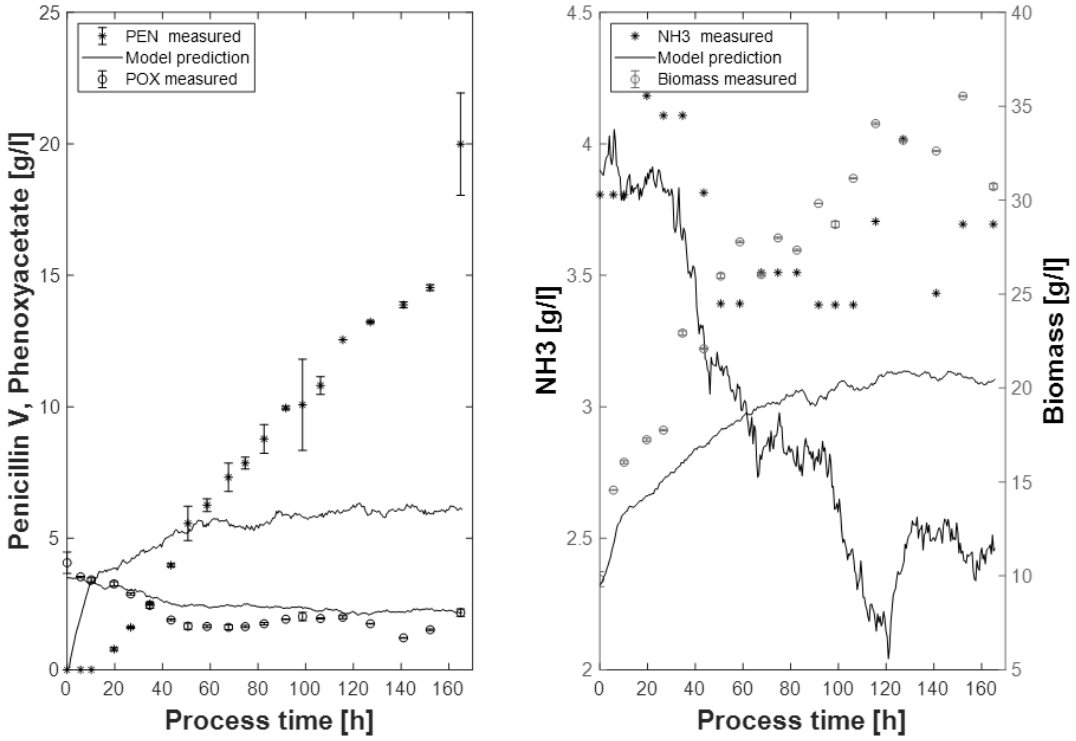

Figure 42: Particle filter estimations and off-line measured concentrations of penicillin, POX, NH3 and biomass for process JL2B

Calculated errors of particle filter estimations can be seen in Table 29.

Table 29: Errors of particle filter estimations for JL2B process

Experiment	RMSE [g/l] / R-square [-]			
	Penicillin	**POX**	**NH_3**	**Biomass**
JL2B	5.188 / 0.1329	0.54498 / 0.49707	0.80594 / -8.491	8.8684 / -0.59127

In contrast to JL2A, penicillin is underestimated at the end of the process, which has resulted due to wrong PLS predictions. This has also led to underestimated biomass. Therefore, as expected NIR based PLS combined with off-gas and observer showed lower prediction ability with prediction NRMSE below 30% for penicillin and POX only.

However as it can be seen from fermentation results (Figure 40), the process was still properly controlled and the experiment can also be considered as successful.

Validation experiment discussion

In section **Validation experiment** it was shown that the developed system was applied for real-time process control. Estimates made by particle filter were used for feeding rates calculations. Both of the processes were supported by spectral data (NIR or MIR) which was transformed into concentrations of POX, ammonia, and penicillin via PLS.

NIR based PLS showed lower prediction ability when compared to the MIR based one. However, this is quite questionable, if the overall performance of NIR is lower, as the constructed NIR based models were based only on JL1 process set - JL1A and JL1B (in contrast to MIR based PLS models which were based on three processes). This was made due to quite different media compositions and operational conditions in AJ8 and JL1.

Despite the low prediction ability of PLS models and high prediction errors (Table 30), particle filter was still able to produce proper estimations for process control. Constant biomass specific limiting substrate uptake rates were achieved for at least a half of the overall process time. It has also succeeded to keep precursor and ammonia at non-limiting, constant levels.

Despite the results of quasi-real-time simulations (simulations with the historical data), which proved, that spectral data can improve kinetic model predictions, the validation experiment has shown that this approach needs further development.

Due to the data-driven nature of PLS, more fermentations with different media compositions and feeding profiles have to be carried out for generating more data in order to improve PLS predictions through the reduction of model overfitting and the elimination of noise impact. Additional developments of such a PLS model construction step as variable selection should also be done.

Table 30: Errors of particle filter estimations for JL2A and JL2B processes

Experiment	NRMSE [%]			
	Penicillin	**POX**	**NH_3**	**Biomass**
State observer with MIR based PLS (JL2A)	36.6	20.3	57.6	12.1
State observer with NIR based PLS (JL2B)	25.9	19.1	101.4	34.5

Conclusion

The United States Food and Drugs Administration guidelines for process development recommend to study and control the product during the whole product lifecycle[xl]. Thus, an important production step, fermentation process, must efficiently provide the pre-defined product quality, which requires proper control.

Near- and mid-infrared spectroscopy instruments were introduced in this work and the obtained spectra, together with offline data, were used for PLS models construction. Afterward, constructed models were applied to historical data. Four datasets (AJ8A, AJ8B, JL1A, JL1B) were used for the construction of NIR based PLS models and three datasets (AJ7A, AJ8A, JL1A) were used for MIR based PLS models construction. This resulted in estimates of soluble fermentation products (penicillin, ammonia, and precursor), with different prediction errors. Mid-IR spectroscopy showed its high precision and transferability when applied to historical data (AJ7, AJ8, and JL1). The higher sensitivity of MIR was caused by the invasive way of measurement as well as the fact that a MIR spectrum contains more information about the biomolecular composition of the substance[xvi]. Near-IR measurements were done non-invasive, and, therefore, despite higher prediction errors (comparing to MIR based PLS models), have great potential and advantage in bioindustry due to high operational safety and low costs. The applied technology still needs improvements, as, despite acceptable accuracy (prediction NRMSE lower 30%), obtained for most of PLS models, they were prone to errors when applied on the external data. Devices with higher sensitivity and better wavenumber selection procedures could improve the results.

Another approach – kinetic modeling was also introduced in this work. It was shown that a well-known kinetic model, developed by *Paul et.al.*[xxix] has a strong prediction ability, but does lack in real-time information, and therefore is not able to react on process deviations. Thus, the kinetic model was combined with real-time off-gas data via particle filter, and acceptable biomass estimations for all historical datasets (AJ7, AJ8, JL1) were achieved. Nevertheless, it was shown, that to guarantee full observability, real-time information about other crucial components (penicillin, precursor and ammonia) is required (Figure 30).

Combinations of near- and mid-infrared based PLS models with a model-based observer and off-gas measurements led to a completely observable system and resulted in good estimations of concentrations of biomass, penicillin, precursor, and ammonia for most of the processes (except JL1). Prediction failures were caused by the fact that the corresponding PLS models were not completely transferable. Therefore further fermentations are required in order to eliminate noise impact and construct robust and transferable PLS models. Lowest prediction errors were achieved by the addition of both – NIR and MIR measurements to the state observer with off-gas data (Table 26 and Table 27).

Obtained model-based observer predictions are smooth and possible to be done real-time. It was shown that this method leads to transferable, stable, selective, precise and robust estimations which none of the applied measurement techniques can perform themselves.

Finally, established model-based observers were successfully applied for the control of real fermentation processes – JL2A and JL2B. Despite overfitted PLS models, particle filter was stable enough to produce good estimates, which resulted in constant non-limiting concentrations of precursor and ammonia as well as the constant biomass specific uptake rate of the limiting substrate which was kept at the desired set-point for at least a half of each process (Figure 37 and Figure 40).

Supplement

Kinetic model equations

Model equations are written as described in the literature[xxx].

Model terms are presented in Table 31.

Table 31: Model terms as described in [xxx]

Term	Description	Unit
V	Volume	*[l]*
F_{in}	Total inlet feed	*[l/h]*
u_{Glc}	Feed of glucose	*[l/h]*
u_{POX}	Feed of precursor	*[l/h]*
$F_{in,ammonia}$	Feed of ammonia	*[l/h]*
$F_{in,titration}$	Total feed of acid and base	*[l/h]*
F_{out}	Total outlet feed	*[l/h]*
c_{A0}	Concentration of A_0	*[g/l]*
c_{A1}	Concentration of A_1	*[g/l]*
c_{Glc}	Concentration of glucose	*[g/l]*
c_{Gln}	Concentration of gluconate	*[g/l]*
$c_{f,i}$	Feed concentration of component *i*	*[g/l]*
$r_{b,0}$	Rate of branch formation	*[g/(l*h)]*
$r_{d,1}$	Differentiation rate	*[g/(l*h)]*
$r_{e,1}$	Rate of extension of non-growing parts	*[g/(l*h)]*
$r'_{b,0}$	Rate of branch formation by gluconate	*[g/(l*h)]*
$r'_{e,1}$	Extension for gluconate	*[g/(l*h)]*
r_p	Rate of penicillin production	*[g/(l*h)]*

According to the material balance equation:

$$\frac{\partial V}{\partial t} = F_{in}(t) - F_{Out}(t) = u_{Glc}(t) + u_{POX}(t) + F_{in,ammonia}(t) + F_{in,titration}(t) - F_{out}(t)$$

Branching and differentiation rates can be written as:

$$r_{b,0} = \frac{\mu_0 * c_{A1} * c_{Glc}}{K_0 + c_{Glc}}, r_{d,1} = \frac{\gamma_1 * c_{A0}}{K_1 + c_{Glc} + c_{Gln}}, r'_{b,0} = \frac{\mu'_0 * c_{A1} * c_{Gln}}{(K'_0 + c_{Gln}) * (1 + \frac{c_{Glc}}{K'})}$$

Extension rates are calculated as follows:

$$r_{e,1} = \frac{\mu_e * c_{A,0} * c_{Glc}}{K_e + c_{Glc}}$$

and

$$r'_{e,1} = \frac{\mu'_e * c_{A,0} * c_{Gln}}{(K'_e + c_{Gln}) * (1 + \frac{c_{Glc}}{K''})}$$

Where for growing and non-parts is valid:

$$\frac{\partial c_{A0}}{\partial t} = r_{b,0} + r'_{b,0} - r_{d,1} - \frac{F_{in} c_{A0}}{V}$$

and

$$\frac{\partial c_{A1}}{\partial t} = r_{e,1} + r'_{e,1} + r_{d,1} - r_{b,0} - r'_{b,0} - \frac{F_{in} c_{A1}}{V}$$

The rate of penicillin formation is:

$$r_p = \frac{\mu_p * c_{A1} * (c_{Glc} + c_{Gln})}{K_p + (c_{Glc} + c_{Gln}) * (1 + \frac{c_{Glc} + c_{Gln}}{K_I})} * \frac{c_{POX}}{K_{POX} + c_{POX}}$$

And the changes of glucose and gluconate concentrations are given as:

$$\frac{\partial c_{Glc}}{\partial t} = -\alpha_0 * r_{b,0} - \alpha_e * r_{e,1} - \alpha_p * r_p - \frac{\mu_{GOD} * c_{Glc}^2}{K_{GOD} + c_{Glc}} + \frac{u_{Glc} * c_{f,Glc}}{V} - \frac{F_{in} * c_{Glc}}{V},$$

$$\frac{\partial c_{Gln}}{\partial t} = -\alpha_{Gln} * (r_{b,0} + r'_{e,1}) + \frac{\mu_{GOD} * c_{Glc}^2}{K_{GOD} + c_{Glc}} - \frac{F_{in} * c_{Gln}}{V}$$

For the changes of penicillin and phenoxyacetate concentration is valid:

$$\frac{\partial c_{PEN}}{\partial t} = r_p - \mu_h * c_{PEN} - \frac{F_{in} * c_{PEN}}{V},$$

and

$$\frac{\partial c_{POX}}{\partial t} = -r_p * \alpha_{\frac{POX}{PEN}} + \frac{u_{pox} * c_{f,POX}}{V} - \frac{F_{in} * c_{POX}}{V}$$

Thus, described model contains 7 states and 24 parameters.

Model parameters can be seen in Table 32.

Table 32: Model parameters as described in [xxx]

Parameter	Description	Value/Unit
α_0	Yield for conversion of glucose to A_0	*1.8061 g Glc g^{-1} A_0^{-1}*
α_e	Yield for conversion of glucose to A_1	*1.8061 g Glc g^{-1} A_1^{-1}*
α_{Gln}	Yield for conversion of gluconate to A_0 and A_1	*2.100 g Gln g^{-1} $(A_0+A_1)^{-1}$*
α_p	Yield for conversion of glucose to penicillin	*2.6896 g Glc g^{-1} PEN^{-1}*
$\alpha_{POX/PEN}$	Yield for conversion of precursor to penicillin	*0.3833 g POX g^{-1} PEN^{-1}*
γ_1	Maximum rate of glucose and gluconate consumption for transformation of A_0 to A_1	*0.0090 g (Glc+Gln) l^{-1}*
μ_0	Specific rate for branching in response to glucose	*0.005 h^{-1}*
μ_0'	Specific rate for branching in response to gluconate	*0.005 h^{-1}*
μ_e	Specific rate for extension in response to glucose	*0.5391 h^{-1}*
μ_e'	Specific rate for extension in response to gluconate	*0.5391 h^{-1}*
μ_{GOD}	Specific rate for glucose oxidation	*0.2063 h^{-1}*
μ_h	First order constant of penicillin hydrolysis	*0 h^{-1}*
μ_p	Specific rate for penicillin production	*0.0126 h^{-1}*
K'	Inhibition constant for branching of gluconate in response to glucose	*0.0300 g Glc l^{-1}*
K''	Inhibition constant for extension of gluconate in response to glucose	*0.0300 g Glc l^{-1}*
K_0	Saturation constant for branching in response to glucose	*0.040 g Glc l^{-1}*
K_0'	Saturation constant for branching in response to gluconate	*0.040 g Gln l^{-1}*
K_1	Inhibition constant for differentiation in response to glucose and gluconate	*0.0820 (Glc+Gln) l^{-1}*
K_e	Saturation constant for extension in response to glucose	*0.0820 g Glc l^{-1}*
K_e'	Saturation constant for extension in response to gluconate	*0.0820 g Gln l^{-1}*
K_{GOD}	Saturation constant for glucose oxidation	*0.010 g Glc l^{-1}*
K_I	Saturation constant for penicillin production in response to glucose and gluconate	*0.0131 g (Glc+Gln) l^{-1}*
K_p	Inhibition constant for penicillin production in response to glucose and gluconate	*0.2610 g (Glc+Gln) l^{-1}*
K_{POX}	Saturation constant of precursor conversion into penicillin	*0.300 g POX l^{-1}*

Figures

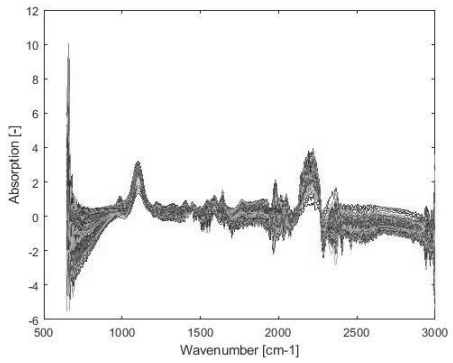

Figure 43: MIR spectra of the fermentation process after SNV (AJ8-A)

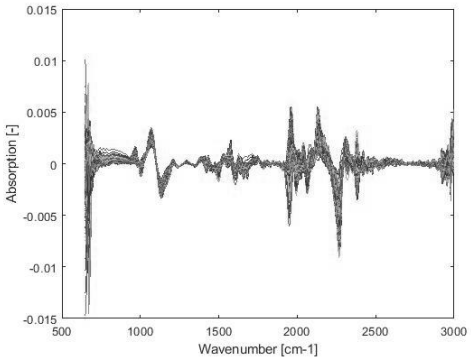

Figure 44: First derivative (Savitzky-Golay) of MIR spectra (AJ8-A)

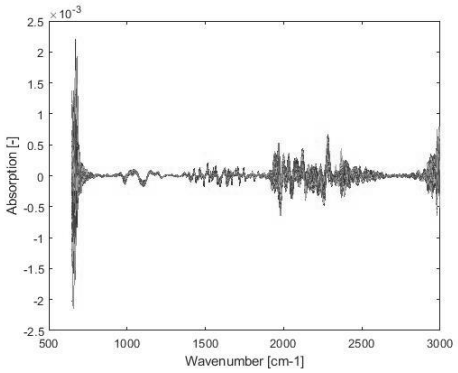

Figure 45: Second derivative (Savitzky-Golay) of MIR spectra (AJ8-A)

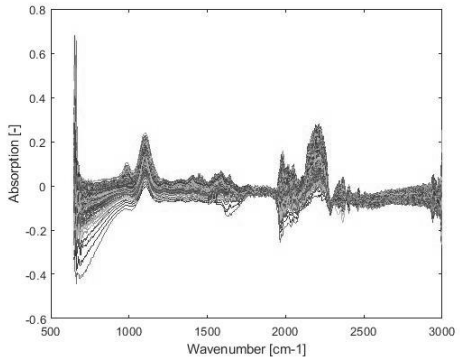

Figure 46: Raw MIR spectra of the fermentation process (AJ8-A)

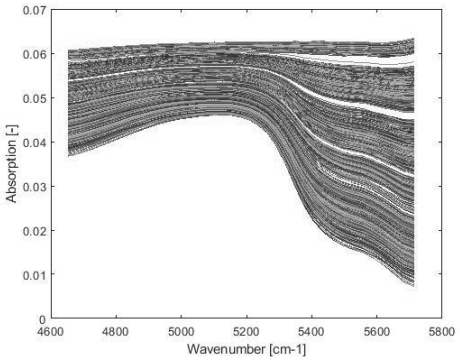

Figure 47: Raw MIR spectra of the fermentation process (AJ7-B)

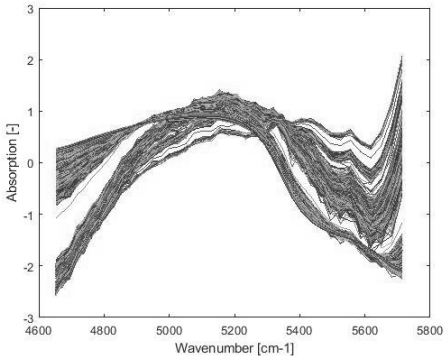

Figure 48: NIR spectra of the fermentation process after SNV (AJ7-B)

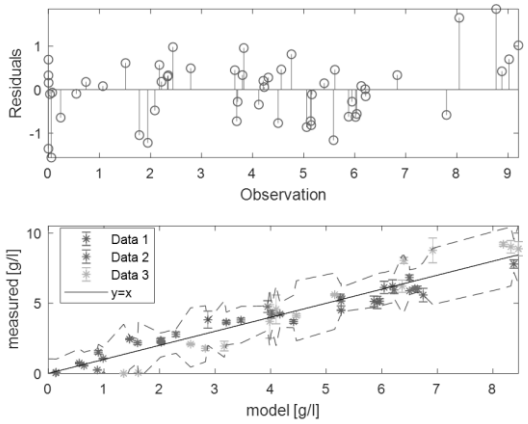

Figure 49: Model constructed for penicillin, based on all historical MIR data

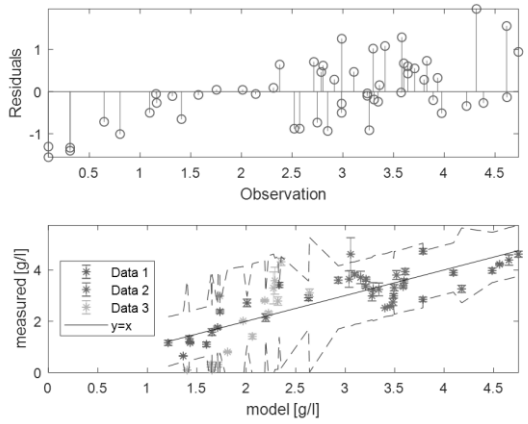

Figure 50: Model constructed for POX, based on all historical MIR data

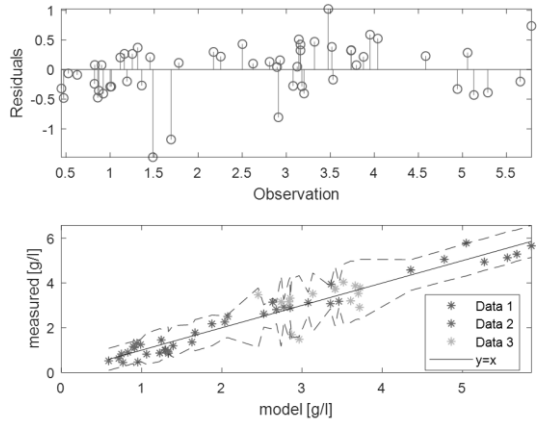

Figure 51: Model constructed for NH3, based on all historical MIR data

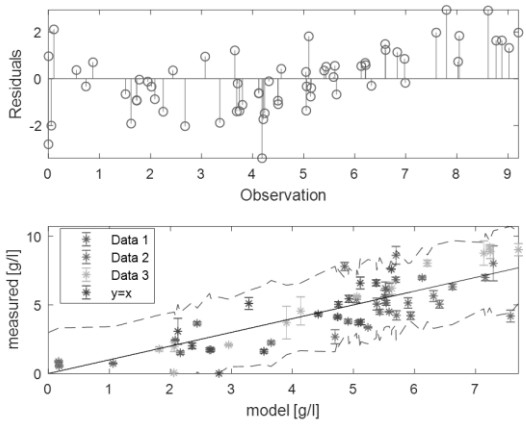

Figure 52: Model constructed for penicillin, based on all historical NIR data

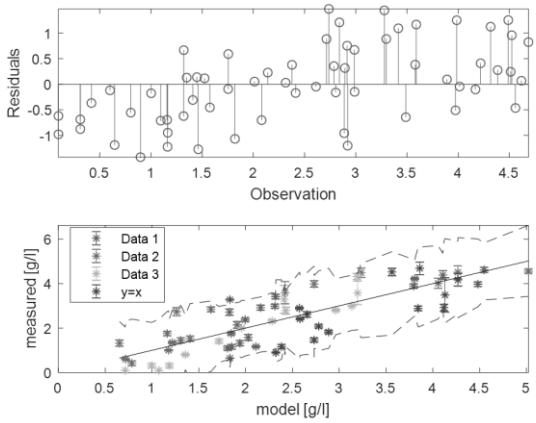

Figure 53: Model constructed for POX, based on all historical NIR data

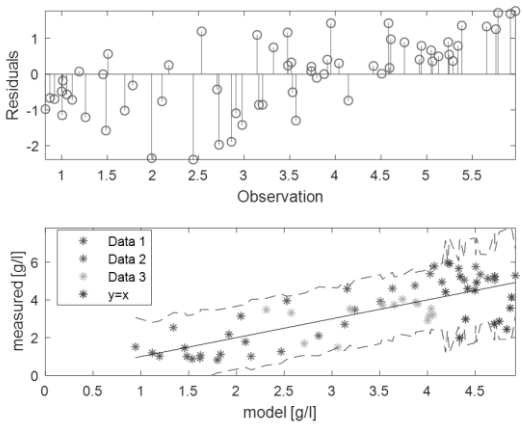

Figure 54: Model constructed for NH3, based on all historical NIR data

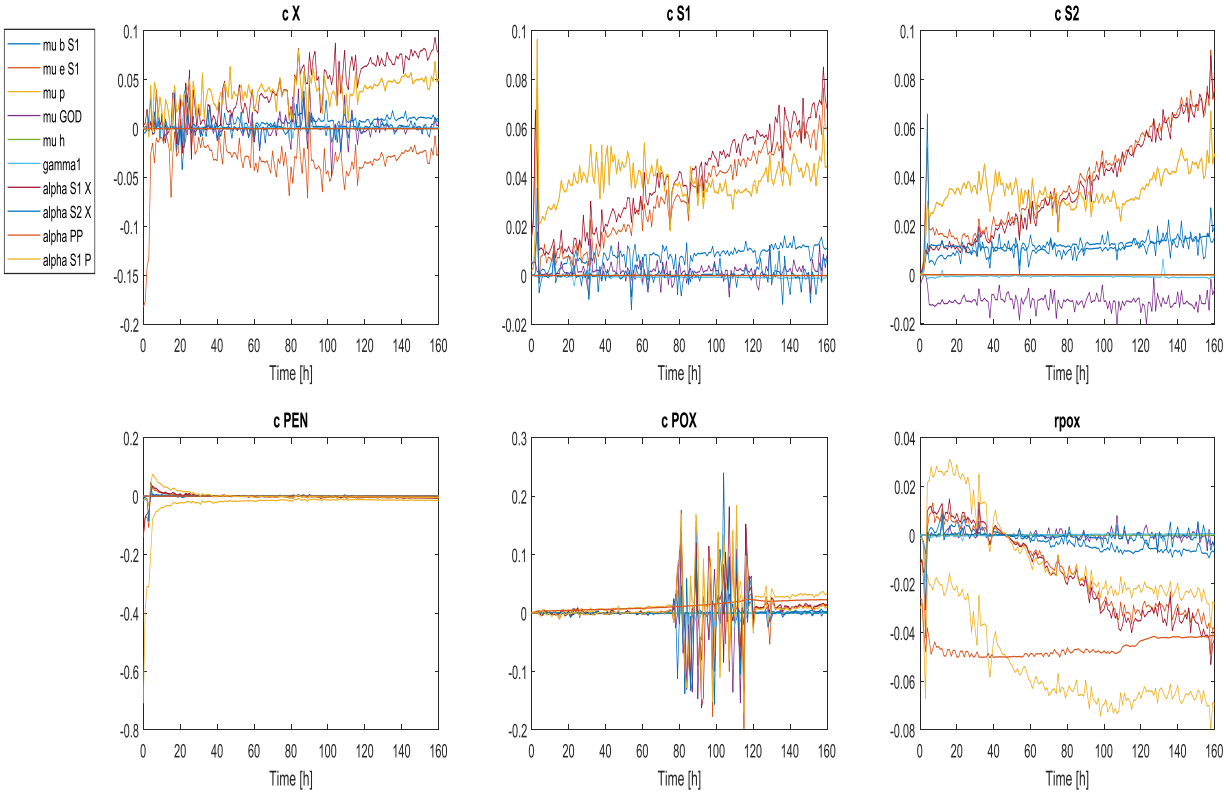

Figure 55: Calculated parameter sensitivity [-] over process time [h] (process AJ8B)

Prediction NRMSE of different methods applied

Table 33: Prediction NRMSE for different methods applied in this work

Experiment	NRMSE [%]			
	Penicillin	**Precursor**	**Ammonia**	**Biomass**
MIR based PLS				
AJ7A	6.179305	30.64785	10.04872	-
AJ8A	7.726544	12.23011	6.518645	-
JL1A	11.19379	26.05687	23.37967	-
NIR based PLS				
AJ8A	14.20253	15.92803	17.64967	-
AJ8B	22.64948	15.83094	19.0906	-
JL1A	29.90115	41.07346	30.70796	-
JL1B	44.85914	43.8358	35.38372	-
Validation experiment, PF predictions				-
JL2A	36.64995	20.28285	57.56507	12.09823
JL2B	25.95415	19.07459	101.3761	34.48498
MIR based PLS model (based on all MIR data)				
	7.530198	16.41692	7.620734	-
NIR based PLS model (based on all NIR data)				
	14.45579	16.26521	19.68938	-
NIR based PLS model (separate models for AJ8 and JL1)				
AJ8A	11.39626	14.69536	16.54533	-
AJ8B	14.68478	12.89283	19.35236	-
JL1A	7.316098	8.497156	9.032229	-
JL1B	29.2972	15.74053	16.14267	-
Historical data: PF, NIR and CER				
AJ8A	22.13044	19.63403	19.58718	11.37087
AJ8B	15.03889	21.58366	20.85363	8.763005
JL1A	26.78688	40.7891	23.97878	41.70466
JL1B	40.79779	19.24155	51.78385	19.8648
Historical data: PF and CER				
AJ7A	-	63.42139	31.52187	13.50157
AJ7B	-	71.77632	38.92821	9.181818
AJ7C	-	52.08288	23.14873	12.9446
AJ8A	-	29.45655	18.1028	9.275725
AJ8B	-	15.56065	23.53332	3.728626
AJ8C	-	116.3704	188.3404	43.52368
JL1A	-	13.3763	65.33401	28.69752
JL1B	-	24.90998	55.99746	27.33334
JL1C	-	33.01252	48.31684	25.96743

Experiment	NRMSE [%]			
	Penicillin	**Precursor**	**Ammonia**	**Biomass**
Historical data: PF, MIR and CER				
AJ7A	8.643543	25.26298	10.36994	11.66028
AJ8A	20.82601	15.0306	14.27918	11.27056
JL1A	10.12057	28.81754	41.50512	40.34474
Historical data: PF, MIR, NIR, CER				
AJ8A	14.68992	12.54908	12.56178	9.850532
JL1A	19.94786	34.85308	39.53716	29.99823

References

[i] Fleming, A. (1929). On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B. influenzae. *British journal of experimental pathology*, *10*(3), 226.
[ii] Nielsen, J. (1997). *Physiological engineering aspects of Penicillium chrysogenum*. World Scientific.
[iii] Smith, J. J., Lilly, M. D., & Fox, R. I. (1990). The effect of agitation on the morphology and penicillin production of Penicillium chrysogenum. *Biotechnology and bioengineering*, *35*(10), 1011-1023.

[iv] Nielsen, J., Johansen, C. L., Jacobsen, M., Krabben, P., & Villadsen, J. (1995). Pellet formation and fragmentation in submerged cultures of Penicillium chrysogenum and its relation to penicillin production. *Biotechnology Progress*, *11*(1), 93-98.
[v] Tipper, D. J., & Strominger, J. L. (1965). Mechanism of action of penicillins: a proposal based on their structural similarity to acyl-D-alanyl-D-alanine. *Proceedings of the National Academy of Sciences*, *54*(4), 1133-1141.
[vi] Elander, R. P. (2003). Industrial production of β-lactam antibiotics. *Applied microbiology and biotechnology*, *61*(5-6), 385-392.
[vii] Brown, W. E., & Peterson, W. H. (1950). Factors affecting production of penicillin in semi-pilot plant equipment. *Industrial & Engineering Chemistry*, *42*(9), 1769-1774.
[viii] Hobby, G. L. (1985). Penicillin: meeting the challenge.
[ix] Chmiel, H. (Ed.). (2011). *Bioprozesstechnik* (Vol. 3). Heidelberg: Spektrum Akademischer Verlag.
[x] Mears, L., Stocks, S. M., Albaek, M. O., Sin, G., & Gernaey, K. V. (2017). Mechanistic fermentation models for process design, monitoring, and control. Trends in biotechnology, 35(10), 914-924.
[xi] Paul, G. C., & Thomas, C. R. (1996). A structured model for hyphal differentiation and penicillin production using Penicillium chrysogenum. *Biotechnology and bioengineering*, *51*(5), 558-572.
[xii] Goldrick, S., Ştefan, A., Lovett, D., Montague, G., & Lennox, B. (2015). The development of an industrial-scale fed-batch fermentation simulation. *Journal of biotechnology*, *193*, 70-82.
[xiii] Alford, J. S. (2006). Bioprocess control: Advances and challenges. Computers & Chemical Engineering, 30(10-12), 1464-1475.
[xiv] Streefland, M., Martens, D. E., Beuvery, E. C., & Wijffels, R. H. (2013). Process analytical technology (PAT) tools for the cultivation step in biopharmaceutical production. Engineering in life sciences, 13(3), 212-223.
[xv] Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmond, A., & Jent, N. (2007). A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *Journal of pharmaceutical and biomedical analysis*, *44*(3), 683-700.
[xvi] Sales, K. C., Rosa, F., Sampaio, P. N., Fonseca, L. P., Lopes, M. B., & Calado, C. R. (2015). In situ near-infrared (NIR) versus high-throughput mid-infrared (MIR) spectroscopy to monitor biopharmaceutical production. *Applied spectroscopy*, *69*(6), 760-772.
[xvii] Krämer, D., & King, R. (2017). A hybrid approach for bioprocess state estimation using NIR spectroscopy and a sigma-point Kalman filter. *Journal of Process Control*.
[xviii] Kailath, T. (1980). *Linear systems* (Vol. 156). Englewood Cliffs, NJ: Prentice-Hall.
[xix] Whitford, W., & Julien, C. (2007). Analytical technology and PAT. *BioProcess International*, *5*(1).
[xx] Landgrebe, D., Haake, C., Höpfner, T., Beutel, S., Hitzmann, B., Scheper, T., ... & Reardon, K. F. (2010). On-line infrared spectroscopy for bioprocess monitoring. *Applied microbiology and biotechnology*, *88*(1), 11-22.
[xxi] Atkins, P., & De Paula, J. (2013). *Elements of physical chemistry*. Oxford University Press, USA.
[xxii] Goicoechea, H. C., & Olivieri, A. C. (2003). A new family of genetic algorithms for wavelength interval selection in multivariate analytical spectroscopy. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *17*(6), 338-345.
[xxiii] Udelhoven, T., & Schütt, B. (2000). Capability of feed-forward neural networks for a chemical evaluation of sediments with diffuse reflectance spectroscopy. *Chemometrics and intelligent laboratory systems*, *51*(1), 9-22.

[xxiv] Aehle, M., Kuprijanov, A., Schaepe, S., Simutis, R., & Lübbert, A. (2011). Simplified off-gas analyses in animal cell cultures for process monitoring and control purposes. *Biotechnology letters*, *33*(11), 2103.
[xxv] Chmiel, H. (Ed.). (2011). *Bioprozesstechnik* (Vol. 3). Heidelberg: Spektrum Akademischer Verlag.
[xxvi] Lohninger, H. (2010). Fundamentals of statistics. *Retrieved December*, *5*, 2010.
[xxvii] Luoma, P., Golabgir, A., Brandstetter, M., Kasberger, J., & Herwig, C. (2017). Workflow for multi-analyte bioprocess monitoring demonstrated on inline NIR spectroscopy of P. chrysogenum fermentation. *Analytical and bioanalytical chemistry*, *409*(3), 797-805.
[xxviii] Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, *58*(2), 109-130.
[xxix] Paul, G. C., Syddall, M. T., Kent, C. A., & Thomas, C. R. (1998). A structured model for penicillin production on mixed substrates. *Biochemical engineering journal*, *2*(1), 11-21.
[xxx] Kager, J., Herwig, C., & Stelzer, I. V. (2018). State estimation for a penicillin fed-batch process combining particle filtering methods with online and time delayed offline measurements. *Chemical Engineering Science*, *177*, 234-244.
[xxxi] DiStefano III, J. (2015). Dynamic systems biology modeling and simulation. Academic Press.
[xxxii] Mandenius, C. F., & Titchener-Hooker, N. J. (2013). Measurement, monitoring, modelling and control of bioprocesses.
[xxxiii] Simon, D. (2006). *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*. John Wiley & Sons.
[xxxiv] Kessler, W. (2007). Multivariate datenanalyse: für die pharma, bio-und Prozessanalytik. John Wiley & Sons.

[xxxv] Koch, C., Posch, A. E., Goicoechea, H. C., Herwig, C., & Lendl, B. (2014). Multi-analyte quantification in bioprocesses by Fourier-transform-infrared spectroscopy by partial least squares regression and multivariate curve resolution. Analytica chimica acta, 807, 103-110.

[xxxvi] National Institute of Standards and Technology (NIST) standard Reference Database 69: NIST Chemistry WebBook

[xxxvii] Brun, R., Kühni, M., Siegrist, H., Gujer, W., & Reichert, P. (2002). Practical identifiability of ASM2d parameters—systematic selection and tuning of parameter subsets. *Water research*, *36*(16), 4113-4127.
[xxxviii] Nakhaeinejad, M., & Bryant, M. D. (2011). Observability analysis for model-based fault detection and sensor selection in induction motors. *Measurement Science and Technology*, *22*(7), 075202.
[xxxix] Golabgir, A., Hoch, T., Zhariy, M., & Herwig, C. (2015). Observability analysis of biochemical process models as a valuable tool for the development of mechanistic soft sensors. *Biotechnology progress*, *31*(6), 1703-1715.
[xl] FDA, U. (2009). Guidance for Industry. Q8 (R2) Pharmaceutical Development.