Jan-Patrick Voss
Nina E. Mittelheuser
Roman Lemke
Reiner Luttmann

Research Center of Bioprocess
Engineering and Analytical
Techniques, Department of
Biotechnology, Hamburg
University of Applied Sciences,
Hamburg, Germany

Research Article

Advanced monitoring and control of pharmaceutical production processes with *Pichia pastoris* by using Raman spectroscopy and multivariate calibration methods

This contribution includes an investigation of the applicability of Raman spectroscopy as a PAT analyzer in cyclic production processes of a potential Malaria vaccine with *Pichia pastoris*. In a feasibility study, Partial Least Squares Regression (PLSR) models were created off-line for cell density and concentrations of glycerol, methanol, ammonia and total secreted protein. Relative cross validation errors $RMSE_{cvrel}$ range from 2.87% (glycerol) to 11.0% (ammonia). In the following, on-line bioprocess monitoring was tested for cell density and glycerol concentration. By using the nonlinear Support Vector Regression (SVR) method instead of PLSR, the error $RMSE_{prel}$ for cell density was reduced from 5.01 to 2.94%. The high potential of Raman spectroscopy in combination with multivariate calibration methods was demonstrated by the implementation of a closed loop control for glycerol concentration using PLSR. The strong nonlinear behavior of exponentially increasing control disturbances was met with a feed-forward control and adaptive correction of control parameters. In general the control procedure works very well for low cell densities. Unfortunately, PLSR models for glycerol concentration are strongly influenced by a correlation with the cell density. This leads to a failure in substrate prediction, which in turn prevents substrate control at cell densities above 16 g/L.

Keywords: Bioreactor monitoring and control / In-line Raman spectroscopy / Partial least squares regression / Process analytical technology / Support vector regression

Additional supporting information may be found in the online version of this article at the publisher's web-site

Received: November 8, 2016; revised: August 17, 2017; accepted: August 24, 2017

DOI: 10.1002/elsc.201600229

1 Introduction

The Process Analytical Technology (PAT) initiative, which was published in 2004 by the US Food and Drug Administration (FDA), proposes the application of advanced process analyzers

with the aim of gaining knowledge and control of pharmaceutical production processes [1]. In biotechnological production, the cultivation of microorganisms is one of the most important unit operations. Due to the complex nature of such processes, bioreactor monitoring is still a challenging task [2].

Spectroscopic sensors are very well suited for establishing PAT because they are capable of measuring multiple physical and chemical variables at the same time [3, 4]. Due to advances in spectrometer equipment, e.g. fiber-optic connected in-line probes, analyzers can be used under harsh production conditions [2, 3]. Near infrared (NIR) spectroscopy is widely used and accepted. Many successful applications of NIR spectroscopy for monitoring of biotechnological upstream processes are reported [5].

Raman spectroscopy is a promising alternative (or addition) to NIR spectroscopy, because substances are basically

Correspondence: Prof. Dr. Reiner Luttmann (reiner.luttmann@haw-hamburg.de), Department of Biotechnology, Hamburg University of Applied Sciences, Ulmenliet 20, 21033 Hamburg, Germany.

Abbreviations: ANN, Artificial Neural Networks; CDW, Cell Dry Weight; CHO, Chinese Hamster Ovary; FDA, Food and Drug Administration; MVDA, Multivariate Data Analysis; NIR, near infrared; PCA, Principle Component Analysis; PLSR, Partial Least Squares Regression; RBF, Radial Basis Function; SCADA, Supervisory Control and Data Acquisition; SNV, Standard Normal Variate; SVR, Support Vector Regression; VIP, Variable Importance on Projection

Raman-active when they are NIR-inactive [6]. As a consequence, water has only a weak signal in the Raman spectrum and therefore does not disturb analyzes in the culture broth. Furthermore, Raman spectra are more defined than NIR spectra, which allows for more specific analyzes regarding chemical substances [7].

Raman spectroscopy has already been used successfully as a PAT analyzer in pharmaceutical but non-biotechnological processes like powder blending and freeze-drying [8]. However, there are relatively few publications regarding Raman spectroscopy for on-line monitoring of bioprocesses. Lee et al. used Raman spectroscopy to quantify glucose, acetate, formate, lactate and phenylalanine in an *Escherichia coli* process without common multivariate calibration methods like PLSR but with simple least squares fitting [9]. Most publications regarding bioreactor monitoring are concerned with determination of e.g. glucose, lactate, ammonia as well as total and viable cell densities with PLSR models in Chinese Hamster Ovary (CHO) cell processes [7, 10–12]. These are much easier to monitor compared with microbial processes due to a smaller oxygen demand resulting in lower aeration rates and agitation speeds as well as a much lower biomass content.

Craven et al. demonstrated the applicability of Raman spectroscopy for feedback control of the glucose concentration in a CHO cell process using a nonlinear model predictive controller (NMPC) instead of a standard proportional/integral (PI) algorithm [13].

The application of multivariate methods is almost essential in order to extract relevant information from complex Raman spectra and they are an ideal tool to implement PAT with spectroscopic analyzers [2, 14]. The prevalent multivariate calibration method is Partial Least Squares Regression (PLSR), because it is efficient, relatively simple and because PLSR functionality is often included in spectrometer software packages [2].

PLSR is a multivariate linear regression method [15]. However, process measurements can contain nonlinearities originating from temperature shifts or other influences. Therefore it can be advantageous to use nonlinear regression methods in order to improve the performance of quantification [16]. Support Vector Regression (SVR) can handle high dimensional data. In combination with nonlinear Kernel functions (e.g. Gaussian Radial Basis Function, RBF) SVR is therefore very well suited for the use with spectroscopic data [16]. In that respect it has an advantage over other nonlinear regression methods like Artificial Neural Networks (ANN), which requires variable selection in order to reduce computing times [16, 17]. Barman et al. successfully quantified glucose levels in human blood samples with Raman spectroscopy and SVR [18]. However, no publication could be found regarding Raman spectroscopy in combination with SVR for monitoring of cultivation processes.

In this work, cyclic production processes of a potential Malaria vaccine with the methylotrophic yeast *Pichia pastoris* were under investigation. The aim was to develop multivariate calibration (PLSR) models for five important process variables. The cell density c_{XL} in the liquid phase L, substrate concentrations c_{sJM} of glycerol ($j = 1$) and methanol ($j = 2$) were examined as well as concentrations of ammonia (C_{AltotM}) and total secreted protein (c_{ptotM}), all in the media phase M. Optimal preprocessing methods as well as PLSR model settings were determined.

Another goal was to study negative influences on the performance of these models. The impact of cells in the measured matrix on predictions was investigated by comparing PLSR models from off-line Raman measurements in the culture supernatant (media phase M) with models from measurements in the culture broth including cells (liquid phase L). The effect of disturbances occurring in on-line processes (aeration, agitation) was studied by comparing off-line models with models from on-line Raman measurements. Furthermore, in on-line investigations the nonlinear regression method SVR was compared with the linear PLSR.

Process control is the next step in implementing spectroscopic PAT methods in biopharmaceutical processes. Therefore, the applicability of Raman spectroscopy in combination with PLSR modeling was tested via a feed-forward closed loop control of the glycerol concentration in the liquid phase.

2 Materials and methods

2.1 Production of therapeutic proteins with *Pichia pastoris*

2.1.1 Product, strains and media

The malaria vaccine candidate D1M1H was chosen as the target product. It is an artificial protein developed by Faber et al., following a diversity covering approach [19]. *Pichia pastoris* KM71H was used as host system for production of D1M1H. Some cultivations were performed with a *Pichia* wildtype strain (DSM 70382) without expression of any recombinant protein. Cultivation media was prepared as described by Loegering et al. [20].

2.1.2 Cultivation strategy

Cultivation of the methylotrophic yeast *Pichia pastoris* commonly comprises three phases. The first phase is an unlimited batch phase on the primary substrate glycerol (S1), which is conducted for cell breeding. The second phase is a limited glycerol fed-batch phase for further cell mass generation at a reduced oxygen demand. In phase three, cells are grown on their secondary substrate methanol (S2) in an unlimited fed-batch. This part is called the production phase, because methanol induces heterologous protein expression in *Pichia pastoris*.

In this work, the applicability of Raman spectroscopy for bioprocess monitoring was investigated in production processes following a cyclic strategy. Figure 1 shows a cultivation process consisting of five cycles. Basically, the three process phases described above are repeated successively. This is realized by harvesting the majority of cells at the end of a production phase and using the remaining cells as inoculum for the next cultivation cycle. The detailed bioprocessing was described by Martens et al. [21, 22].

2.1.3 Bioreactor equipment

Cyclic cultivations for the feasibility study presented in Section 3.1 were performed in a BIOSTAT® C (Sartorius Stedim Systems,

Figure 1. Off-line measurements of a cyclic production process with *Pichia pastoris*, c_{XL} (○) = cell density in the culture broth (liquid phase L), all following concentrations in the supernatant (media phase M): c_{S1M} (□) = glycerol, c_{S2M} (△) = methanol, c_{PtotM} (○) = total secreted protein.

Germany). Cultivations for on-line investigations (Sections 3.2 and 3.3) were carried out in a BIOSTAT® ED10 (Sartorius).

Both bioreactors have the same maximum working volume of 10 l and very similar control configurations.

They are equipped with the standard instrumentation for measurement and control of e.g. temperature, dissolved oxygen tension and pH value.

An extended instrumentation was available for both bioreactors in form of an in-line methanol measurement system (Biotechnologie Kempe, Germany) and exhaust gas analysis regarding oxygen and carbon dioxide (BlueSens, Germany). Methanol measurement was used for control of the methanol concentration c_{S2M} in the media phase. Exhaust gas balancing was used in this work for on-line determination of volumetric oxygen and glycerol uptake rates (see Section 3.3).

Furthermore, the bioreactors are equipped with an expanded infrastructure, which allowed for automatic harvest and refresh operations as needed for the intended cyclic cultivation strategy. This involves pneumatic valves (GEMÜ Apparatebau, Germany), peristaltic pumps (Watson-Marlow, UK) and balances (Sartorius).

A high degree of process automation was realized using recipes following the ISA S88 standard established in the SCADA-System MFCS/win 3.0 (Sartorius).

2.1.4 Cultivation conditions

Initial cell densities c_{XL0} , glycerol concentrations c_{S1M0} as well as liquid volumes V_{L0} varied between cycles of the different investigated processes and are dependent on respective experimental goals. Bioreactor aeration (F_{nG}) was maintained constantly at 10 L/min using air. The temperature of the liquid phase ϑ_L was

controlled at 30°C for cell growth on glycerol and at 22°C in production phases. The pH value was maintained at 5.6 by addition of 12.5% ammonia solution or 2 M phosphoric acid respectively. The dissolved oxygen tension pO_2 was controlled at 25% by cascading the agitation speed N_{St} .

For the feasibility study (Section 3.1) the methanol concentration c_{S2M} was controlled at different randomized set points c_{S2Mw} in the production phase in order to generate sufficient variance for PLSR modeling. In production phases of experiments with on-line spectroscopic measurements, c_{S2M} was controlled at 1 g/L.

2.2 Off-line reference measurements

2.2.1 Cell density

The cell density c_{XL} was calculated from the cell dry weight (cdw) in duplicates. For the determination of cell dry weight, 1 ml culture broth was separated by centrifugation at 20,160 g for three minutes in a weighed micro reaction tube. After decanting the supernatant and drying the cell pellet for 48 h at 80°C, the reaction tube was weighed again. The observed mass difference is called cell dry weight.

2.2.2 Substrate concentrations

The substrate concentrations c_{SjM} for glycerol ($j = 1$) and methanol ($j = 2$) in the supernatant M were determined by ion-exclusion chromatography. The used Rezex™ RHM-Monosaccharide column (Phenomenex®, USA) was tempered at 60°C. As mobile phase 10 mM H_2SO_4 was used at a flow rate of 0.6 mL/min. Analysis was performed using a LaChrom Elite®

HPLC-System (VWR-Hitachi, USA) with a refractive index detector (L-7490). Before injection of 40 μL sample, the culture supernatant was filtered through a 0.22 μm syringe filter.

2.2.3 Ammonia concentration

The flow injection analysis system TAS 2000 (Jüke Systemtechnik, Germany) was used to measure the ammonia concentration C_{AltotM} as described by Kaiser [23]. The measurement follows Berthelot's method and is based on reaction of ammonia with hypochlorite and salicylic acid to indophenol, which can be detected photometrically at a wavelength of 605 nm.

2.2.4 Total secreted protein concentration

The total secreted protein concentration c_{PtotM} was quantified using a Bradford Protein Assay Kit (Bio-Rad Laboratories, USA) following the standard procedure for microtiter plates as described in the manufacturer's manual.

2.2.5 Calculation of concentrations in the liquid phase L

Multivariate calibration models based on Raman measurements in the culture broth (liquid phase L) require reference measurements also referring to L. The liquid phase consists of the media phase M and the wet cell phase Z. However, most off-line measurements give results directly referring to the media phase M.

Therefore, a concentration c_{IL} of component I in phase L,

$$c_{\text{IL}}(t) = \left(1 - \frac{\alpha_{\text{Z/X}}}{\rho_{\text{Z}}} \cdot c_{\text{XL}}(t)\right) \cdot c_{\text{IM}}(t), \quad (1)$$

with

$\alpha_{\text{Z/X}} :=$ mass ratio of wet to dry cells [-],

$\rho_{\text{Z}} :=$ density of wet cells [g/L],

has to be calculated from the associated concentration c_{IM} in the media phase and the cell density c_{XL} . A value of 3.7 was used for the mass ratio $\alpha_{\text{Z/X}}$ and the density ρ_{Z} was determined as 1150 g/L.

2.3 Raman spectroscopy

The feasibility study presented in Section 3.1 was performed with a RamanRxn1TM spectrometer (Kaiser Optical Systems, USA). Fibre-optics (length: 5 m) were used to connect the device to a bIO-LAB immersion probe (Kaiser, length: 220 mm, diameter: 12 mm). The off-line Raman measurements were carried out by submerging the probe approximately 15 mm in the liquid sample. At least 30 mL of each sample were provided in dark 50 mL centrifuge tubes. The samples were measured with an integration time of 35 seconds and two accumulations. Each off-line measurement including the positioning of the bIO-LAB probe was performed three times.

For on-line experiments (Sections 3.2 - 3.3) a multichannel RamanRxn2TM analyzer (Kaiser) was used. One channel was operated with an immersion probe connected by fibre-optics to the spectrometer. The bIO-LAB probe was mounted on a 25 mm side port of the bioreactor using a Pg 13.5 adaptor. For on-line

measurements an integration time of 30 s and five accumulations were chosen.

In all cases, Raman spectra were recorded in the wavenumber range $\nu = 200 - 3200 \text{ cm}^{-1}$ with an increment of 1 cm^{-1} . Raman scattering was excited at a wavelength of 785 nm with a diode laser. The laser power was set to 400 mW.

For spectrometer control and spectra recording, the software iC Raman 3.0 (Kaiser) was used.

2.4 Multivariate calibration

2.4.1 Used MVDA software

PLSR model development was performed either with SIMCA 13.0.3 (MKS data analytics solutions (former MKS Umetrics), Sweden) or with PLS_Toolbox 8.1.1 (Eigenvector Research, USA) in combination with MATLAB[®] 8.5.0/R2015a (The Mathworks, USA).

2.4.2 Spectra preprocessing, variable selection and outlier detection

For PLSR modeling, different spectral preprocessing methods were tested and compared. However, the standard normal variate (SNV) filter and calculation of the first derivative are the only methods used in final PLSR models. First derivatives of Raman spectra were calculated with the convolution method by Savitzky and Golay with a first order polynomial and a frame size of 15 variables ($\Delta 15 \text{ cm}^{-1}$) [24]. For PLSR model generation, spectra were additionally mean centered.

Variable selection was performed after preprocessing using the Variable Influence on Projection (VIP) statistic by Wold et al. as implemented in the software SIMCA [25]. The VIP value gives a measure of the contribution of an X-variable on the explained Y-variance of a PLSR model [26]. Only single contiguous spectral regions were selected.

After variable selection, Raman spectra were checked for outliers. For that, a principle component analysis (PCA) model was built, which explains at least 95% variance (R^2_{Xcum}) of the used spectral data with a minimum number of principle components. Samples with Hotelling T^2 values exceeding the 95% confidence limit were excluded as strong outliers from further PLSR model development.

2.4.3 Evaluated model characteristics

As the main parameter for PLSR model performance evaluation, the common root mean square error RMSE_{cv} was used. The index cv indicates its calculation from a cross validation procedure. In addition, the prediction error RMSE_{p} was calculated from an external validation with an independent test set, if applicable.

The investigated Y-variables have different units and their measured values have different magnitudes. In order to compare model performance for different Y-variables, the relative error $\text{RMSE}_{\text{cvrel}}$,

$$\text{RMSE}_{\text{cvrel}} = \frac{\text{RMSE}_{\text{cv}}}{y_{\text{CSmax}} - y_{\text{CSmin}}} \cdot 100\% \quad (2)$$

with

$y_{C_{Smax}}$:= maximum measured value of variable y in the calibration set,

$y_{C_{Smin}}$:= minimum measured value of variable y in the calibration set,

was calculated from the root mean square error $RMSE_{cv}$ divided by the calibration range of the respective variable. The relative prediction error $RMSE_{Pre}$ was calculated analogously.

Furthermore, models were evaluated by examining the statistic R^2_{cv} . This parameter is the determination coefficient for a linear fit of predicted versus measured Y -values. The statistic was used as an indicator of the general suitability of the used spectral data for multivariate modeling regarding the Y -variable under investigation. Again, a respective statistic R^2_p was used for externally validated models.

2.4.4 Cross validation and model tuning

Cross validation (cv) was performed in eight steps, assuring that every observation was used exactly once for validation. For every cv step, approx. 12.5 % of the observations in the calibration set were excluded from model generation and used for prediction with the respective sub-model. Excluded observations were uniformly distributed over the chronologically sorted data set ("venetian blinds"). E.g. for the first step, samples 1, 9, 17 etc. were part of the same cv group. Replicates were also kept in the same cv group. The error $RMSE_{cv}$ was calculated from all observations after finalizing the eighth step.

The number r of PLS components used in a PLSR model was mainly determined by the value of r , for which the error $RMSE_{cv}$ (or $RMSE_p$) reached a minimum or a lower saturation value. In the case of doubt, one component less was chosen in order to prevent overfitting.

2.4.5 Support vector regression

SVR Models were built with PLS_Toolbox 8.1.1 (Eigenvector), which uses an implementation of the LIBSVM library [27]. In contrast to PLSR model development, spectral data were not mean centered before building SVR models. As Kernel function the nonlinear Gaussian radial basis function (RBF) was used.

Determination of the SVR model parameters γ , C and ε was carried out using a modified version of the method proposed by Devos et al. for classification purposes [28]. Model tuning involved grid searches in order to determine the possible predictive power of SVR models in terms of cross validation error $RMSE_{cv}$, which led to the selection of a local minimum of that error in the three dimensional parameter space. In a second step an optimization of the model parameters was performed in order to minimize both the $RMSE_{cv}$ as well as the number of support vectors n_{sv} . The used procedure resulted in good models with a reduced chance of overfitting.

Figure 2. Networking of the involved software systems for control of the glycerol concentration c_{SIL} with Raman spectroscopy and PLS regression.

2.5 On-line PLSR model application for closed loop control

In Section 3.3 a closed loop control of the glycerol concentration c_{SIL} , quantified with Raman spectroscopy, is presented. For that, c_{SIL} had to be predicted from newly recorded spectra on-line. Figure 2 illustrates the interaction of the software systems involved in this task.

Necessary PLSR model executions were performed by the calculation module SIMCA Q 13.0.3 (Umetrics) on-line. For this reason, the work station program SIMCA 13.0.3 had to be used in order to establish models in advance.

The software Kaiser Data Link 1.1 (Kaiser) was used to transfer spectra from the recording software iC Raman to SIMCA Q and to specify the PLSR model file for SIMCA Q to work with. Another important task of Kaiser Data Link was the distribution of predicted values on the laboratory network via OPC protocol.

SIMATIC WinCC 7.0 SP2 (Siemens) was used as a central data distribution management system. OPC variables were read from their origins and written to their destination systems by the OPC functionality of WinCC.

The PI-control algorithm was established in the bioreactor SCADA system MFCs/win, which reads process data and distributes these via OPC protocol.

Additional on-line calculations were needed for control of the glycerol concentration. These were programmed as MATLAB® scripts and implemented into the PAT software SIMATIC SIPAT 4.0.0.3 (Siemens, Germany).

SIPAT was configured in a way that these calculations were executed every three minutes and that results were distributed on the network via OPC protocol as well. The calculations themselves were carried out by MATLAB® 8.3.0/R2014a.

3 Results and discussion

3.1 Monitoring of bioprocess variables - a feasibility study

For the intended feasibility study, two cultivations were investigated. The first one consists of a single, frequently sampled cycle

and was performed with the *Pichia* wildtype strain DSM 70382. A total amount of 50 samples was collected and used for spectroscopic as well as reference measurements. The second cultivation consists of five cycles, out of which 65 samples were analyzed. Here the production strain *Pichia pastoris* KM71H D1M1H was used.

The first process led to a generation of four models for concentrations in the liquid phase L, including cell density (c_{XL}), substrates glycerol and methanol (c_{SIL}), and ammonia (C_{AltotL}). Furthermore, three models were built for substrate and ammonia concentrations (c_{SJM} and C_{AltotM}) in the media phase.

The second process was used for the generation of a model for the total secreted protein concentration C_{PtotM} . Unfortunately, only supernatants were available of this cultivation. Therefore, a comparison of model performance for the media and liquid phase was not possible. The second process was also used to externally validate models from the first process for substrate concentrations c_{SJM} . Some results of the feasibility study are shown in Fig. 3, whereas Table 1 gives an overview of results for all built models.

The best PLSR model for the cell density c_{XL} was obtained by using the complete Raman spectrum pretreated by calculating the first derivative. Three PLS components were used to describe 97.2% of the spectral variance R^2_{Xcum} and led to a cross validation error $RMSE_{cv}$ of 3.23 g/L, which corresponds to a relative error of 4.40 % of the calibration range.

Better predictive performance was obtained for the glycerol concentration c_{SIM} with a relative cross validation error of 2.87 %. Only one PLS component was needed to describe 99.1 % of the spectral variance. X-variables were restricted to the wave number range $\nu = 2750 - 3050 \text{ cm}^{-1}$, which contains a strong double peak for high glycerol concentrations. This signal originates from stretching vibrations of alkane C-H bonds present in glycerol molecules [29]. SNV filtering led to the best predictive ability. External validation of that model resulted in a prediction error $RMSE_p$ of 2.01 g/L, which is expectedly higher than the cross validation error $RMSE_{cv}$ of 1.56 g/L.

A similar model was built for the glycerol concentration c_{SIL} (in the liquid phase with cells) where two PLS components led to a cross validation error of 1.79 g/L. It can be concluded, that Raman measurements in the liquid phase lead to more complex models with less predictive ability.

For development of the PLSR model for c_{SIM} , samples with glycerol concentrations of 0 g/L were excluded in order to obtain an as uniformly as possible distribution of glycerol concentration values. This is not suitable for development of models for c_{SIL} because Raman spectra are strongly influenced by the present cells in this case. In combination with a high negative correlation of cell density and glycerol concentration (correlation coefficient $R = -0.977$), such a model will have a very poor predictive ability for samples with glycerol concentrations of 0 g/L.

The best model for the methanol concentration c_{S2M} was obtained from the first derivative of Raman spectra limited to the region $\nu = 2750 - 3050 \text{ cm}^{-1}$ as well. A relative cross validation error of 3.69 % was achieved by using two PLS components. External validation with samples from the second process led to a very similar result. The best model for prediction of the methanol concentration c_{S2L} has an absolute cross validation error of 0.226 g/L, which is almost twice the error of a model for

the supernatant. Therefore it can be concluded that in case of methanol concentration, cells present in the liquid phase have a significant negative influence on the predictive ability of PLSR models.

An opposite behavior was observed for the ammonia concentrations C_{AltotM} and C_{AltotL} . The PLSR model for the supernatant has an error of 11.0 % of the calibration range. A model from the liquid phase consisting of four instead of two PLS components led to a slightly lower absolute cross validation error, which translates into a relative error of 5.72 %. This is due to the higher calibration range for the ammonia concentration in the liquid phase. The unexpected better result for the culture broth model can be explained by a high correlation of ammonia concentration and cell density ($R = -0.806$).

On one hand, such a model would not be of practical use if the process changes. E.g. different initial substrate concentrations or process phase durations would lead to different cell densities, which ultimately would result in wrong ammonia predictions. On the other hand, in cyclic processes where these conditions do not change, such a model could be used with care and may serve as a qualitative monitoring tool e.g. for the indication of an ammonia limitation.

Prediction of the total secreted protein concentration C_{PtotM} with a PLSR model consisting of four PLS components has a relative cross validation error of 10.3 %. It could be observed that especially high protein concentrations were not predicted well.

It was shown that Raman spectra of protein solutions with relevant concentrations do not contain suitable signals for model creation [30]. Therefore, it is expected that the obtained model is based on some indirect correlation of cell growth with the total protein concentration, which provides a certain predictive ability.

Quantification based on indirect measurements could be used on purpose in order to develop an on-line monitoring tool for secreted proteins. Compared to current off-line or at-line alternatives this would be a strong improvement, even with relatively high errors.

3.2 On-line monitoring of the cell density and glycerol concentration

For on-line bioprocess monitoring with Raman spectroscopy, the cell density c_{XL} and the glycerol concentration c_{SIL} were selected as target variables. Models were built from a cultivation of *Pichia pastoris* KM71H D1M1H consisting of five cycles. Results for the on-line predictions are shown in Fig. 4.

A total of 164 samples were taken. Raman measurements were performed automatically on-line. Three cultivation cycles with 90 samples were used as calibration set. The validation set consisted of 74 observations belonging to the remaining two cycles.

In order to reduce the high negative correlation between cell density and glycerol concentration, glycerol spike experiments were performed. Calibration and validation data sets both contain one cycle with glycerol spikes. However, a high

Figure 3. Results of the feasibility off-line study. (A) Prediction of the cell density c_{XL} (B) Prediction of the glycerol concentration c_{S1M} (C) External validation of the model in B (D) Prediction of the methanol concentration c_{S2M} (E) External validation of the model in D (F) Prediction of the ammonia concentration c_{AltotM} (G) Prediction of the total secreted protein concentration c_{PtotM} .

correlation ($R = -0.813$) remains for the observations in the calibration set.

When choosing the model settings for on-line prediction, findings from the feasibility study were accounted for. Therefore, used spectral regions and preprocessing methods were identical. On-line models were optimized by adjusting the number of PLS components. Table 2 shows the performance for on-line

prediction in comparison with off-line models presented in the previous section.

The best on-line PLSR model for cell density consists of two PLS components and has a cross validation error of 2.54 g/L. The on-line model for glycerol concentration is composed of three instead of two PLS components compared to the off-line model and reaches slightly higher absolute and relative cross validation

Table 1. Comparison of off-line PLSR models for the process variables under investigation, upper part: calibration/cross validation, lower part: external validation

Analyte	C _{XL}	C _{S1M}	C _{S1L}	C _{S2M}	C _{S2L}	C _{AltotM}	C _{AltotL}	C _{PtotM}
Variable select. ^{a)}	no	yes	yes	yes	yes	no	no	no
Preprocessing ^{b)}	1 st deriv.	SNV	SNV	1 st deriv.	SNV	none	none	none
n _{CS} ^{c)} [-]	50 (150)	20 (60)	50 (150)	50 (150)	50 (150)	46 (138)	46 (138)	61 (183)
r [-]	3	1	2	2	3	2	4	4
R ² _{Xcum} [-]	0.972	0.991	0.951	0.977	0.959	0.981	0.999	0.999
R ² _{cv} [-]	0.971	0.995	0.998	0.984	0.882	0.784	0.93	0.887
RMSE _{cv} [g/L]	3.23	1.56	1.79	0.116	0.226	2.51 mM	2.11 mM	40.5 mg/L
RMSE _{cvrel} [%]	4.40	2.87	3.32	3.69	9.89	11.0	5.72	10.3
n _{VS} ^{c)} [-]	-	26 (78)	-	65 (195)	-	-	-	-
R ² _p [-]	-	0.997	-	0.972	-	-	-	-
RMSE _p [g/L]	-	2.01	-	0.115	-	-	-	-
RMSE _{prel} [%]	-	3.70	-	3.66	-	-	-	-

^{a)} Yes: variables limited to $\nu = 2750 - 3050 \text{ cm}^{-1}$.

^{b)} Mean Centering was always applied.

^{c)} Numbers in parentheses indicate the total number of observations including replicates.

errors. External validation led to better results for both variables under investigation than the cross validation procedure, for which reason the models were considered as valid.

However, some inconsistencies in performance of cell density models were observed. The main issue is visualized in Fig. 4A). The relationship between predicted and measured values shows some curvature. This led to the conclusion that the linear relationship between intensity values of Raman spectra and cell density, as it is required for PLSR modeling, is only true for a first approximation.

Because of that, the method SVR was tested for both variables. The Gaussian RBF was used as kernel function in order to model possible nonlinear relationships. For SVR, the same spectral regions were used as for PLSR. Preprocessing changed in a way, that mean centering was not applied to the spectra and that SNV filtering was found to be the best method for both variables.

Calibration information and cross validation results for the SVR models are also given in Table 2. For both variables, sparse models with a low number n_{sv} of support vectors were generated.

Sparse models are considered to have a good generalization ability and therefore minimizing the chance of overfitting [31]. In Table 2 external validation results of on-line PLSR and SVR models are shown. For cell density, the SVR model produced significantly better results than the PLSR model. The external validation error RMSE_p improved from 2.11 g/L to 1.23 g/L. When plotting predicted cell density values against reference values as shown in Fig. 4B) it can be observed, that for the nonlinear SVR model a linear relationship can be achieved, indicating that in this case nonlinear SVR modeling is more suitable than PLSR modeling.

In contrast to the cell density monitored with SVR, only minor improvements were achieved for the glycerol concentration c_{S1L} . A lower error is attributed to a better prediction of glycerol concentrations of 0 g/L by the SVR model. Fig. 4C) shows the prediction results for PLSR and SVR models in a time course of the validation set cycles.

3.3 Closed loop control with Raman spectroscopy and PLS regression

3.3.1 General remarks

Experiments for controlling the glycerol concentration c_{S1L} were performed in the same bioreactorsystem, which was used to develop on-line models. The cultivation strategy changed in a way that cycles only consisted of a glycerol fed-batch phase starting with an initial glycerol concentration c_{S1L0} equal to the set point of glycerol control c_{S1Lw} , which was chosen as 10 g/L.

The on-line PLSR model used for the control experiments was built from the data presented in Section 3.2. Unfortunately the control experiments had to be performed before model development process was completed. Due to that reason, the used PLSR model for prediction of glycerol concentration is not identical to that presented in Section 3.2. The main difference is that no variable selection was performed. Ultimately, the model was based on complete, SNV-filtered Raman spectra of 72 samples and consisted of four PLS components. External validation of the model was carried out with 56 samples and led to a prediction error RMSE_p of 2.91 g/L.

3.3.2 Description of the substrate control concept

This section describes the general glycerol control concept, which involves parameter adaptation and feed-forward control. A more detailed theoretical elaboration is given in the Supporting Information. Figure 5 shows a schematic of the glycerol control concept.

For substrate control, Raman spectra A_{Lram} are measured in-line and used for on-line glycerol prediction with a PLSR model. This creates the controlled variable c_{S1Lpls} , which is compared with the manual fixed set point c_{S1Lw} . A proportional/integral (PI) control algorithm is used to influence the set point F_{R1w} of a glycerol pump, which compensates for glycerol consumption by the cells.

Figure 4. External validation results for on-line PLSR and SVR models. (A) Predicted versus measured cell density c_{XL} for a PLSR model. (B) Predicted versus measured cell density c_{XL} for a SVR model. (C) On-line prediction time courses of c_{XL} and c_{S1L} , c_{S2M} := in-line methanol measurement.

The control loop design is based on the balance equation of glycerol concentration c_{S1L} ,

$$\dot{c}_{S1L}(t) = \frac{F_{R1}(t)}{V_L(t)} \cdot [c_{S1R1} - c_{S1L}(t)] - S_1 UR(t), \quad (3)$$

with the volumetric substrate uptake rate $S_1 UR$,

$$S_1 UR(t) = q_{S1/X}(t) \cdot c_{XL}(t) = \frac{y_{1X/O}(t)}{y_{X/S1}(t)} \cdot OUR(t) \quad (4)$$

$$= y_{S1/O} \cdot Q_{O2}(F_{nG}, V_L, x_{IGin}, x_{IG}, t),$$

which is proportional to the oxygen uptake rate OUR .

The glycerol/oxygen uptake equivalence $y_{S1/O}$ is assumed to be time-invariant for unlimited growth and was calculated as 2.78 from the yield coefficients given in Martens [22].

The oxygen uptake rate is observable via gas phase balancing with the volumetric oxygen supply rate Q_{O2} , which is calculated on-line from measurements of F_{nG} , V_L as well as O_2 - and CO_2 -mole fractions in the gas inlet and the exhaust line.

The controlled system shows a nonlinear behavior ($F_{R1} \cdot c_{S1L}$) and includes a strong exponential increase of the disturbance $S_1 UR$ during unlimited growth conditions. This creates a problem in control parameter adjustment, which has to be solved by an adaptive control procedure.

Calculation of the control parameters controller gain K_C and controller integral time constant T_I for a desired control behavior is carried out by linearization of the plant equation (Eq. (3) with Eq. (4)) at a certain operating point [c_{S1Lop} , F_{R1op} , V_{Lop} , Q_{O2op}] and its analysis in the Laplace-transformed domain (see

Table 2. Comparison of multivariate calibration models for prediction of the cell density and glycerol concentration, upper part: calibration/cross validation, lower part: external validation

Analyte	Cell density c_{XL}			Glycerol concentration c_{SIL}		
Origin	off-line	on-line	on-line	off-line	on-line	on-line
Method	PLSR	PLSR	SVR	PLSR	PLSR	SVR
Preprocessing ^{a)}	1st deriv.	SNV	SNV	SNV	SNV	SNV
$n_{CS}^{b)}$ [-]	50 (150)	90	90	50 (150)	90	90
r [-]	3	2	-	2	3	-
R^2_{Xcum} [-]	0.972	0.976	-	0.951	0.972	-
R^2_{cv} [-]	0.971	0.963	0.992	0.998	0.986	0.996
RMSE _{cv} [g/L]	3.23	2.54	1.23	1.79	2.07	1.22
RMSE _{cvrel} [%]	4.40	6.03	2.92	3.32	3.98	2.35
γ [-]	-	-	$3.27 \cdot 10^{-3}$	-	-	0.112
C [-]	-	-	$4.04 \cdot 10^2$	-	-	3.43
ε [-]	-	-	$8.42 \cdot 10^{-2}$	-	-	$4.70 \cdot 10^{-2}$
n_{sv} [-]	-	-	24	-	-	35
n_{vs} [-]	-	74	74	-	69	69
R^2_P [-]	-	0.97	0.992	-	0.978	0.993
RMSEP [g/L]	-	2.11	1.23	-	1.85	1.58
RMSEP _{rel} [%]	-	5.01	2.92	-	3.56	3.04

^{a)}For PLSR models, mean centering was additionally applied.

^{b)}Numbers in parentheses indicate the total number of observations including replicates.

Figure 5. Schematic of the subsystems used for adaptive closed loop control of the glycerol concentration c_{SIL} with Raman spectroscopy and PLS regression.

Supporting Information). Adaptation of the control parameters is then realized by updating the operating point at discrete time points t_k throughout the process.

The changing values of the volume $V_{Lop}(t_k)$ and the oxygen supply rate $Q_{O2op}(t_k)$ are directly measurable, whereas the operating point of the controlled variable $c_{SILop}(t_k)$ is equal to the set point c_{SILw} of the glycerol controller.

The operating point F_{R1op} of substrate feed rate,

$$F_{R1op}(t_k) = \frac{y_{S1/O} \cdot Q_{O2op}(t_k) \cdot V_{Lop}(t_k)}{c_{S1R1} - c_{S1Lw}} = F_{R1k}, \quad (5)$$

can be derived from Eq. (3) and Eq. (4) for steady state conditions and is also known at t_k .

The plant actuation via the feed pump set point F_{R1w} ,

$$F_{R1w}(t) = F_{R1k} + \Delta F_{R1w}(t), \quad (6)$$

is realized with a disturbance feed-forward control with the changing operating point F_{R1k} .

The manipulation variable ΔF_{R1w} of the PI controller is only correcting F_{R1k} . As a consequence, the integral part of the controller has to be reset at each adaptation time t_k .

3.3.3 Presentation of control results and discussion

In Fig. 6 the results for a closed loop control of the glycerol concentration c_{SIL} with constant controller gains K_C are displayed.

In experiment A, K_C is fixed to a high value of $0.04 \text{ L}^2/(\text{g h})$ and the glycerol feed concentration c_{S1R1} is adjusted to 500 g/L. The integral time constant T_{ikap} is adapted to an aperiodic behavior according to (Supporting Information eq. (27)).

Reaching the set point c_{SILw} , the control is switched on at $t = 25 \text{ h}$. In the next three hours the controlled variable c_{SILpls} follows the assumed critical damped behavior. When rising above the set point at $t = 28 \text{ h}$, the controller reduces the feed rate F_{R1} below the demand F_{R1k} . The off-line measurement $c_{S1Lhplc}$ shows the real process response in a strong decrease whereas the predicted variable c_{SILpls} is only slowly following this behavior.

At $t = 29.5 \text{ h}$ the control deviation becomes negative and the controller increases F_{R1} exponentially until it reaches its upper bound F_{R1max} .

The substrate concentration $c_{S1Lhplc}$ shows a strong increase following the control action whereas the substrate prediction c_{S1Lpls} decreases further on.

For the second experiment B in Fig. 6 a similar control behavior was observed. Here, the controller gain K_C was set to $0.012 \text{ L}^2/(\text{g h})$, much lower than in experiment A, while the controller integral time constant T_{ikosc} followed (Supporting Information eq. (28)). Therefore, an oscillating behavior with a constant oscillation period T_P of 2 h was maintained.

Figure 6. Results for closed loop control of the glycerol concentration c_{S1Lpl} , (A) Operation of non-oscillating control with critical damping, (B) Glycerol control with fixed oscillation period T_P , c_{XL} (o) := cell density in the liquid phase L, c_{S1Lpl} := glycerol concentration in L predicted with Raman spectroscopy and PLSR, $c_{S1Lhplc}$ (□) := HPLC off-line measurement F_{R1k} := required feed rate (operating point), T_{ik} := adapted controller integral time constant.

The glycerol feed concentration c_{S1R1} was increased to 600 g/L, because of which the feed rate F_{R1k} increases on a lower trajectory in comparison to experiment A.

During the first 4 h the control shown in Fig. 6B follows the expected behavior. Beginning at $t = 53 \text{ h}$ the decline of prediction c_{S1Lpl} leads to the same failure in the substrate control procedure as shown in Fig. 6A.

It was concluded that this behavior was caused by an influence of cell density on the used PLSR model. As stated above, cell density and glycerol concentration are highly negatively correlated. Raman spectra from the liquid phase contain information of cell density, which can be the basis of a PLSR model for glycerol concentration. Such a model would predict decreasing glycerol concentrations, when the cell density is increasing.

This corresponds to the observed behavior in the control experiments.

In order to cope with this phenomenon, spike experiments were performed in the cultivation used for PLSR model generation (see Section 3.2). Obviously, this measure was not sufficient. As stated above, the on-line PLSR model for glycerol control experiments is based on complete Raman spectra. In comparison, the results presented in Section 3.2 were obtained by a model based on the wave number range $\nu = 2750 - 3050 \text{ cm}^{-1}$ which contains a strong glycerol signal.

This variable selection minimizes the influence of the cell density on glycerol prediction.

It is therefore assumed that repetition of experiments with the final PLSR or even SVR models as presented in Section 3.2

would lead to more satisfactory results. Unfortunately, this was not possible within time frame of this work.

Observation of the adapted control parameter T_{Ikap} in Figure 6A may lead to the conclusion that parameter adaptation is not strictly necessary, because T_{Ikap} shows only a minor increase from 1.80 h to ~ 1.90 h. This is of course only true for the limited control duration, which is due to the described failure. T_{Ikap} is increasing exponentially, which means that for longer control periods and therefore higher cell densities, parameter adaption would be necessary in order to maintain the intended aperiodic control behavior. Furthermore, the applied feed-forward control, which is a result of the idea of parameter adaptation, already accounts for a large part of the nonlinear system dynamics and is therefore necessary. The additional control parameter adaptation does not add much to the complexity of the control procedure and is therefore recommended.

4 Concluding remarks

In an off-line feasibility study, Raman spectra were used to develop PLSR models for five important variables of pharmaceutical production processes with *Pichia pastoris*. Thereby, crucial settings for these models like spectral preprocessing, spectral region of interest and the necessary number of PLS components were determined.

Even though these settings are specific for the given process and investigated variables it can be stressed that simple preprocessing methods and a relatively low number of PLS components led to good PLSR models. Simple models are likely to have a good generalization ability, which was proven by validation of two models with an independent test set.

As previously described and explained in the literature, cells have a significant influence on quality of PLSR models, which is a major challenge for bioreactor monitoring [2,32]. In this work, PLSR models based on measurements from the cell free culture supernatant and the culture broth were compared for three process variables. In general, cells in the measurement matrix lead to more complex models in terms of higher numbers of needed PLS components.

However, the presence of cells does not necessarily decrease predictive performance of PLSR models and can lead to lower prediction errors (see Section 3.1, ammonia concentration) when a strong correlation exists between the variable under investigation and the cell density, which causes a major signal in the Raman spectrum. In order to cope with that, variable selection and spike experiments need to be performed.

On the other hand, indirect correlations may be used intentionally for quantification of process variables that do not show distinct signals in the Raman spectrum. One example is total secreted protein concentration [4,30].

As expected, performance of on-line PLSR models is lower than off-line models. This is caused by disturbance e.g. by process aeration and agitation. Furthermore, disturbances may be caused by nonlinear influences [16]. Multivariate calibration

of cell density was significantly improved by nonlinear modeling with Support Vector Regression (SVR) and Gaussian RBF kernel function. Therefore SVR could be a promising alternative to PLSR. A disadvantage of SVR is a high computational effort needed for model generation and a reduced model interpretability.

Raman spectroscopy in combination with PLSR modeling was successfully used for closed loop control of glycerol concentration in the liquid phase for low cell densities. Limitations arose due to the mentioned influence of the cell density on PLSR models from the liquid phase. This influence was reduced in models developed after conducting the control experiments, most probably by selecting a small spectral range. Application of these more advanced PLSR (and SVR) models should lead to more convincing results for glycerol control.

Raman spectroscopy in combination with multivariate calibration methods is very suitable for monitoring and control of *Pichia pastoris* production processes. Further effort is required in order to develop strategies that result in more reliable and capable calibration models. Raman spectroscopy has many advantages regarding bioprocess monitoring when compared to NIR spectroscopy, which is more widely used and accepted. Therefore, more publications involving Raman spectroscopy are needed in order to determine possibilities and to overcome challenges, which are currently being faced in bioreactor monitoring.

Practical application

This paper contains extensive studies about monitoring of pharmaceutical production processes with *Pichia pastoris* using Raman spectroscopy and multivariate data analysis. Furthermore a concrete application of that monitoring system i.e. closed loop control of glycerol concentration is presented. The submitted work thereby contributes to the Process Analytical Technology (PAT) initiative of the U.S. Food and Drug Administration (FDA) for pharmaceutical production processes and should be of interest to the pharmaceutical industry as well as to researchers involved in monitoring and control of bioprocesses with spectroscopic methods.

The authors thank the BMBF - German Federal Ministry of Education and Research for financial support (FKZ: 1756 $\times$ 09 and FKZ: 031567), MKS data analytics solutions (former MKS Umetrics) for loan SIMCA software licenses as well as Kaiser Optical Systems for providing loan spectrometer systems and technical support.

The authors have declared no conflict of interest.

Nomenclature

A_{Lram}	[IU]	Raman intensity measurement in liquid phase L
c_{IK}	[g/L]	mass concentration of component I in system K
C_{IK}	[mM]	molar concentration of component I in system K
F_{nG}	[L/h]	total aeration rate under normalized conditions
F_{R1}	[L/h]	feed rate from reservoir 1
G_j	[-]	complex transfer function of sub system J (Laplace transformed domain)
I_{UR}	[g/(L h)]	volumetric uptake rate of component I
K_0	[-]	open loop gain
K_C	[L ² /(g h)]	controller gain
K_{Sk}	[g h/L ²]	plant gain at time point t_k
m	[g]	mass
M_{O_2}	[g/mol]	molar mass of oxygen
OTR	[g/(L h)]	oxygen transfer rate
$q_{\text{I/X}}$	[h ⁻¹]	cell specific reaction rate of component I
Q_{O_2}	[g/(L h)]	volumetric oxygen supply rate
R	[-]	correlation coefficient
r	[-]	number of PLS components
R^2	[-]	coefficient of determination, explained variance
RMSE	[-]	root mean square error (multiple units)
T_{I}	[h]	controller integral time constant
T_{Ikap}	[h]	controller integral time constant at time point t_k for aperiodic control behavior
T_{Ikosc}	[h]	controller integral time constant at t_k for oscillating control behavior
T_{p}	[h]	oscillation period
T_{Sk}	[h]	plant lag time at time point k
V_K	[l]	volume of system K
V_{nM}	[l]	molar gas volume under normalized conditions
x_{IG}	[-]	molar fraction of component I in the gas phase G
y	[-]	target variable (MVDA)
$Y_{\text{S1/O}}$	[-]	glycerol/oxygen uptake equivalence
$Y_{\text{X/I}}$	[-]	cell yield coefficient for component I

Greek symbols

$\alpha_{\text{Z/X}}$	[-]	mass ratio of wet to dry cells
ν	[cm ⁻¹]	Raman shift
ρ_{Z}	[g/L]	density of wet cells

Indices

0	initial value
Al	alkali / ammonia
ap	aperiodic
C	controller
C/CO ₂	carbon dioxide
cdw	cell dry weight
CS	calibration set
cum	cumulative (over all PLS components)
G	gas phase
hplc	high performance liquid chromatography
i, j, k	counting indices for e.g. time points
in	inlet

L	liquid phase / culture broth
M	media phase / culture supernatant
max	maximum value
min	minimal value
n	under normalized conditions
O/O ₂	oxygen
off	off-line
op	operating point
osc	oscillation
P	product / protein, prediction
pls	obtained with partial least squares regression
R1	reservoir 1
rel	relative
S1	substrate 1 (glycerol)
S2	substrate 2 (methanol)
sv	support vectors
svr	obtained with support vector regression
tot	total
VS	validation set
w	set point
X	cells (cell dry weight), X data domain in MVDA
Y	Y data domain in MVDA
Z	wet cell phase

5 References

- [1] FDA, PAT - A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance. 2004.
- [2] Lourenço, N. D., Lopes, J. A., Almeida, C. F., Sarraguça, M. C. et al., Bioreactor monitoring with spectroscopy and chemometrics: a review. *Anal. Bioanal. Chem.* 2012, 404, 1211–1237.
- [3] Skibsted, E., Engelsen, S. B., Spectroscopy for Process Analytical Technology (PAT), in: *Encyclopedia of Spectroscopy and Spectrometry*, Academic Press, New York, 2010, pp. 2651–2661.
- [4] Streefland, M., Martens, D. E., Beuvery, E. C., Wijffels, R. H., Process analytical technology (PAT) tools for the cultivation step in biopharmaceutical production. *Eng. Life Sci.* 2013, 13, 212–223.
- [5] Hoehse, M., Alves-Rausch, J., Prediger, A., Roch, P. et al., Near-infrared spectroscopy in upstream bioprocesses. *Pharm. Bioprocess.* 2015, 3, 153–172.
- [6] Perkampus, H.-H., *Lexikon Spektroskopie*, VCH, Weinheim 1993.
- [7] Abu-Absi, N. R., Kenty, B. M., Cuellar, M. E., Borys, M. C. et al., Real time monitoring of multiple parameters in mammalian cell culture bioreactors using an in-line Raman spectroscopy probe. *Biotechnol. Bioeng.* 2011, 108, 1215–1221.
- [8] De Beer, T., Burggraave, A., Fonteyne, M., Saerens, L. et al., Near infrared and Raman spectroscopy for the in-process monitoring of pharmaceutical production processes. *Int. J. Pharm.* 2011, 417, 32–47.
- [9] Lee, H. L. T., Boccazzi, P., Gorret, N., Ram, R. J. et al., In situ bioprocess monitoring of Escherichia coli bioreactions using Raman spectroscopy. *Vib. Spectrosc.* 2004, 35, 131–137.
- [10] Berry, B., Moretto, J., Matthews, T., Smelko, J. et al., Cross-scale predictive modeling of CHO cell culture growth and metabolites using Raman spectroscopy and multivariate analysis. *Biotechnol. Prog.* 2015, 31, 566–577.

- [11] Mehdizadeh, H., Lauri, D., Karry, K. M., Moshgbar, M. et al., Generic Raman-based calibration models enabling real-time monitoring of cell culture bioreactors. *Biotechnol. Prog.* 2015, 31, 1004–1013.
- [12] Whelan, J., Craven, S., Glennon, B., In situ Raman spectroscopy for simultaneous monitoring of multiple process parameters in mammalian cell culture bioreactors. *Biotechnol. Prog.* 2012, 28, 1355–1362.
- [13] Craven, S., Whelan, J., Glennon, B., Glucose concentration control of a fed-batch mammalian cell bioprocess using a non-linear model predictive controller. *J. Process Control* 2014, 24, 344–357.
- [14] Mercier, S. M., Diepenbroek, B., Wijffels, R. H., Streefland, M., Multivariate PAT solutions for biopharmaceutical cultivation: current progress and limitations. *Trends Biotechnol.* 2014, 32, 329–336.
- [15] Wold, S., Sjöström, M., Eriksson, L., PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* 2001, 58, 109–130.
- [16] Thissen, U., Pepers, M., Üstün, B., Melssen, W. J. et al., Comparing support vector machines to PLS for spectral regression applications. *Chemom. Intell. Lab. Syst.* 2004, 73, 169–179.
- [17] Li, Y., Yuan, J., Prediction of Key State Variables using Support Vector Machines in Bioprocesses. *Chem. Eng. Technol.* 2006, 29, 313–319.
- [18] Barman, I., Kong, C.-R., Dingari, N. C., Dasari, R. R. et al., Development of Robust Calibration Models Using Support Vector Machines for Spectroscopic Monitoring of Blood Glucose. *Anal. Chem.* 2010, 82, 9719–9726.
- [19] Faber, B. W., Younis, S., Remarque, E. J., Rodriguez Garcia, R. et al., Diversity Covering AMA1-MSP119 Fusion Proteins as Malaria Vaccines. *Infect. Immun.* 2013, 81, 1479–1490.
- [20] Loegering, K., Mueller, C., Voss, J.-P., Wagenfuehrer, C. et al., An integrated scale-down plant for optimal recombinant enzyme production by *Pichia pastoris*. *Biotechnol. J.* 2011, 6, 428–436.
- [21] Martens, S., Borchert, S.-O., Faber, B. W., Cornelissen, G. et al., Fully automated production of potential Malaria vaccines with *Pichia pastoris* in integrated processing. *Eng. Life Sci.* 2011, 11, 429–435.
- [22] Martens, S., *Analyse und Synthese von zyklischen Herstellungsprozessen artifiziieller Diversity Covering (DiCo) Malariavakzinkandidaten*, doctoral dissertation, Gottfried Wilhelm Leibniz Universität Hannover, 2014, retrieved from <https://doi.org/edok01.tib.uni-hannover.de/edoks/e01dh14/787330795.pdf>
- [23] Kaiser, C., *Methoden zur Prozessüberwachung und Modellierung von Herstellungsprozessen rekombinanter Proteine mit Escherichia coli*, doctoral dissertation, Gottfried Wilhelm Leibniz Universität Hannover, 2010, retrieved from <https://doi.org/edok01.tib.uni-hannover.de/edoks/e01dh10/630528934.pdf>
- [24] Savitzky, A., Golay, M. J., Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* 1964, 36, 1627–1639.
- [25] Wold, S., Johansson, E., Cocchi, M., PLS - Partial Least Squares Projection to Latent Structures. In: Kubinyi, H.: *3D QSAR in Drug Design: Volume 1: Theory Methods and Applications*, Springer Science & Business Media, Dordrecht, 1993.
- [26] Farrés, M., Platikanov, S., Tsakovski, S., Tauler, R., Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. *J. Chemom.* 2015, 29, 528–536.
- [27] Chang, C.-C., Lin, C.-J., LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol. TIST* 2011, 2, 27.
- [28] Devos, O., Ruckebusch, C., Durand, A., Duponchel, L. et al., Support vector machines (SVM) in near infrared (NIR) spectroscopy: Focus on parameters optimization and model interpretation. *Chemom. Intell. Lab. Syst.* 2009, 96, 27–33.
- [29] Vandenabeele, P., *Practical Raman spectroscopy: an introduction*, Wiley, The Atrium, Southern Gate, Chichester, West Sussex, United Kingdom 2013.
- [30] Paul, A., Carl, P., Westad, E., Voss, J.-P. et al., Towards Process Spectroscopy in Complex Fermentation Samples and Mixtures. *Chem. Ing. Tech.* 2016, 88, 756–763.
- [31] Kecman, V., Support Vector Machines - An Introduction, in: *Support Vector Machines: Theory and Applications*, Springer, Berlin 2005.
- [32] Iversen, J. A., Berg, R. W., Ahring, B. K., Quantitative monitoring of yeast fermentation using Raman spectroscopy. *Anal. Bioanal. Chem.* 2014, 406, 4911–4919.