Monitoring Parallel Robotic Cultivations with Online Multivariate Analysis

Authors:

Sebastian Hans, Christian Ulmer, Harini Narayanan, Trygve Brautaset, Niels Krausch, Peter Neubauer, Irmgard Schäffl, Michael Sokolov, Mariano Nicolas Cruz Bournazou

Abstract:

In conditional microbial screening, a limited number of candidate strains are tested at different conditions searching for the optimal operation strategy in production (e.g., temperature and pH shifts, media composition as well as feeding and induction strategies). To achieve this, cultivation volumes of >10 mL and advanced control schemes are required to allow appropriate sampling and analyses. Operations become even more complex when the analytical methods are integrated into the robot facility. Among other multivariate data analysis methods, principal component analysis (PCA) techniques have especially gained popularity in high throughput screening. However, an important issue specific to high throughput bioprocess development is the lack of so-called golden batches that could be used as a basis for multivariate analysis. In this study, we establish and present a program to monitor dynamic parallel cultivations in a high throughput facility. PCA was used for process monitoring and automated fault detection of 24 parallel running experiments using recombinant E. coli cells expressing three different fluorescence proteins as the model organism. This approach allowed for capturing events like stirrer failures and blockage of the aeration system and provided a good signal to noise ratio. The developed application can be easily integrated in existing data- and device-infrastructures, allowing automated and remote monitoring of parallel bioreactor systems.

Monitoring Parallel Robotic Cultivations with Online Multivariate Analysis

Sebastian Hans [1,*,†] 🆔, **Christian Ulmer** [1,†], **Harini Narayanan** [2] 🆔, **Trygve Brautaset** [3] 🆔, **Niels Krausch** [1] 🆔, **Peter Neubauer** [1] 🆔, **Irmgard Schäffl** [1] 🆔, **Michael Sokolov** [4] and **Mariano Nicolas Cruz Bournazou** [4,*] 🆔

[1] Chair of Bioprocess Engineering, Institute of Biotechnology, Technische Universität Berlin, Ackerstraße 76, 13355 Berlin, Germany; c.ulmer@campus.tu-berlin.de (C.U.); n.krausch@tu-berlin.de (N.K.); peter.neubauer@tu-berlin.de (P.N.); irmgard.schaeffl@campus.tu-berlin.de (I.S.)

[2] Department of Chemistry and Applied Biosciences, Institute of Chemical and Bioengineering, ETH Zürich, Vladimir-Prelog-Weg 1, CH-8093 Zurich, Switzerland; nharini@chem.ethz.ch

[3] Department of Biotechnology and Food Science, Norwegian University of Science and Technology, Sem Sælandsvei 6-8, 7491 Trondheim, Norway; trygve.brautaset@ntnu.no

[4] DataHow AG, c/o ETH Zürich, HCl, F137, Vladimir-Prelog-Weg 1, CH-8093 Zurich, Switzerland; m.sokolov@datahow.ch

[*] Correspondence: sebastian.hans@tu-berlin.de (S.H.); n.cruz@datahow.ch (M.N.C.B.)

[†] These authors contribute equally to this work.

Abstract: In conditional microbial screening, a limited number of candidate strains are tested at different conditions searching for the optimal operation strategy in production (e.g., temperature and pH shifts, media composition as well as feeding and induction strategies). To achieve this, cultivation volumes of >10 mL and advanced control schemes are required to allow appropriate sampling and analyses. Operations become even more complex when the analytical methods are integrated into the robot facility. Among other multivariate data analysis methods, principal component analysis (PCA) techniques have especially gained popularity in high throughput screening. However, an important issue specific to high throughput bioprocess development is the lack of so-called golden batches that could be used as a basis for multivariate analysis. In this study, we establish and present a program to monitor dynamic parallel cultivations in a high throughput facility. PCA was used for process monitoring and automated fault detection of 24 parallel running experiments using recombinant *E. coli* cells expressing three different fluorescence proteins as the model organism. This approach allowed for capturing events like stirrer failures and blockage of the aeration system and provided a good signal to noise ratio. The developed application can be easily integrated in existing data- and device-infrastructures, allowing automated and remote monitoring of parallel bioreactor systems.

Keywords: high throughput bioprocess development; online data analysis; multivariate analysis; principal component analysis; laboratory automation; SiLA; design of experiments; bioprocess monitoring

1. Introduction

Microbial cell factories are widely used for biotechnological production processes. The development of effective bioprocesses requires screening of many microbial strains under various cultivation conditions. State-of-the-art bioprocess development is known to be time consuming, laborious, and for having a low success rate [1]. Process parameters, such as microbial host, vector size and copy-number, feeding strategy, and media composition have a significant impact on the profitability, reliability, and sustainability of the final manufacturing process. Considering all these factors and evaluating them in relation to each other often calls for a high number of cultivations. To ensure reliability

of data and results, parallel cultivations providing a number of replicates to compensate for the variance of biological systems are additionally needed. Many bioprocess conditions are difficult to study due to the lack of suitable high throughput (HT) facilities to perform all these cultivations in a short time. Consequently, current process improvement strategies in high throughput bioprocess development (HTBD) are based on expert knowledge [2] and static design of experiments (DoE) [3–5]. In order to reduce the number of cultivations to an appropriate level, many factors are only partially and incompletely weighed against each other. On the other hand, emerging technologies such as automation and digitization enable a faster product development and shorter cycles from construction of a microbial clone to an optimal bioprocess [6–8] by increasing the possible number of parallel cultivations. Although model-based tools (e.g., for process control or computer aided design) are the standard in other industries [9–12], they are rarely used in bioprocess development despite their big potential [13,14]. A major challenge is the lack of suitable tractable mathematical models that are required but difficult to develop due to the complexity of biological systems [15]. This is especially the case for process development, where knowledge of the new microbial strains is limited and an exhaustive investigation of mutants likely to be discarded is considered unnecessary.

The difference between process control (batch vs. fed-batch), cultivation system (e.g., shake flask, multi well plate or lab scale bioreactor vs. production scale bioreactor) and the resulting different metabolic and stress conditions between screening and manufacturing makes scale-up between these stages challenging. An essential factor for scale up is a detailed knowledge about the process dynamics [16]. Changes in pH, oxygen limitation or the availability of media components should be considered during the conditional screening phase. The technical requirements are already met by modern HT robots with developments at the μL scale. Significant advantages are offered by mini-bioreactors (MBR) with working volumes between 10 and 250 mL [17–19], integrated online sensors for pH and DOT, integrated control of pH and substrate feeding [20–22] and automated at-line sampling and analysis [8,21,23]. Additionally, computer aided tools that enable advanced process control and feedback operation of the robots and the MBRs have been developed, but are rarely used [17,18,20,22,24–26].

However, the number of parallel cultivations made possible in this way is now hardly achievable in the present way, with manual corrections and interventions. Therefore, process monitoring of parallel cultivations is a major challenge in HTBD, especially if no models to describe the bioprocess are available to guide the operator. Fully automated solutions that include process and feeding control, online and at-line monitoring are very challenging in parallel MBR cultivations [20,21]. The main difficulties are the analysis of multi-dimensional and highly correlated data sets, monitoring and operating a large number of bioreactors and the intrinsic need to solve an optimal experimental design problem considering all MBRs simultaneously [24,27] in a period of time. Additional challenges arise when industrial conditions are investigated at the milliliter scale [22,25].

In production processes, the operation strategy is well defined, historical data is at hand, and the goal is typically to run the current cultivations within predefined critical quality attributes (CQAs) or the "golden batch" conditions. Under these conditions, multivariate analysis (MVA) tools have been widely applied to supervise the process and assure its reproducibility [28]. MVA tools such as principal component analysis (PCA) have become increasingly popular in the field of bioprocesses due to their capability to represent highly correlated multi-dimensional datasets in a reduced space, separating process noise while retaining maximum information. Some of the early applications of PCA include process monitoring, detection of failures or anomalies, and statistical process control [19,28–33]. However, for all these applications, a PCA model is usually trained based on an "in-control", i.e., "golden batch", basis to detect deviations from the targeted production run characteristics [34].

During screening, where the goal is to find new strains best suited for the bioprocess at hand, this training data is of course unavailable. The lack of a "golden batch" makes it very challenging to diagnose or even identify faults or disturbances in cultivations with no historical data. Although in principle historical data of similar processes can be used as reference points for development, one cannot

rely on pre-defined process behavior and constraints, as is typically the case in production. Due to various factors, the physiology and phenotype of the cells is known to vary during cultivation time [35] (e.g., population heterogeneities [36]). This makes choosing a setup and tuning of control strategies in advance very difficult. For this reason, we need to develop tools that exploit the information generated in parallel MBRs online to rapidly develop models for process monitoring and to project the large data sets into a low dimensional visual representation.

Our previous work showed that PCA can be used in parallel MBR experiments to identify and improve feeding strategies with a low number of experiments [27]. In this work we developed a program that enables the monitoring of parallel dynamic cultivations in real-time, supported with visual representations as well as automated event triggers (Figure 1). The added value of this method is enhanced process monitoring and automated fault detection. This is demonstrated in an experimental campaign with 24 parallel MBR operated by a fully automated robotic system. This program allows a compressed representation and visualization of the ongoing experiments, enabling a comparison between the states of parallel cultivations. The PCA method is applied in a moving horizon framework to allow a rapid detection of specific events and to track the dynamic evolution of the reactors. The two approaches together provide an informative overview of the bioreactor's performance and state. Thus, they enable the operator to determine whether all parallel cultivations are running within critical parameter limits and display a warning in case of deviations. Critical bioreactors can be easily identified and tended to.
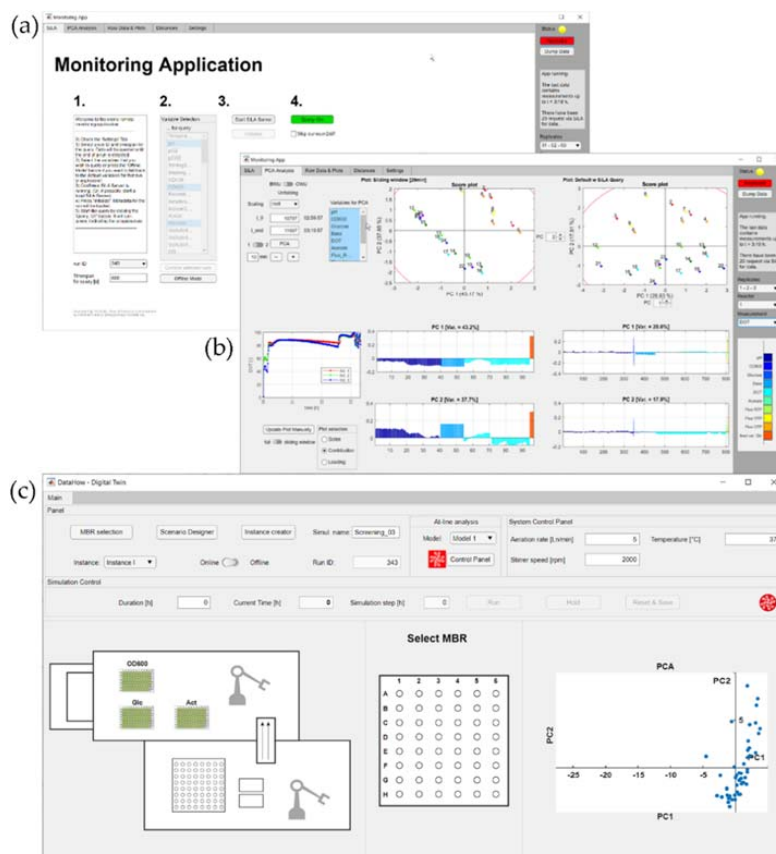

Figure 1. Screenshots of the monitoring application. Screenshot (**a**) depicts the landing page where the user selects an experiment, adapts settings, and starts the data query and subsequent online data analysis; (**b**) shows the main panel with plots for process data and results from the PCA; and (**c**) depicts the central control app of the digital twin (DTW) of the process. It allows for fast process monitoring via direct control and observation of single MBRs as well as fast identification of process failures based on PCA.

In addition, they enable an automated and secure transfer of the cultivation data during the runtime of the experiment. This allows the often computationally intensive online data evaluation to be performed in specialized laboratories, as a service or by project partners. Here we used an efficient protocol for communication, enabling the monitoring of the robotic experiments remotely, reducing the physical barrier separating theoretical work and practical wet lab. The implementation is based on a SILA2 protocol (preliminary standard as of January 2019).

2. Materials and Methods

2.1. Experimental Facility

The wet experiments were performed in a high throughput bioprocess development facility composed of two liquid handling stations (LHS) (Freedom EVO 200, Tecan Group Ltd., Männedorf, Switzerland; Microlab Star, Hamilton Company, Bonaduz, Switzerland) and one mini bioreactor system (bioREACTOR 48, 2mag AG, Munich, Germany) which is mounted on the first LHS (Tecan). The entire facility and its functionality is described in [21]. To trigger non-optimal microbial cultivation conditions the volume balance control described by Haby et al. [21] was switched off.

2.2. Microbial Model Strain and Cultures

All cultivations were performed with *E. coli* K-12 BW25113 (F-, DE(*araD-araB*)567, *lacZ4787(del)::rrnB-3*, LAM-, *rph-1*, DE(*rhaD-rhaB*)568, *hsdR514*) carrying plasmid pAG032. As described by Gawin et al. 2018 [37] the plasmid pAG032 contains three fluorescent protein (CFP, YFP, RFP) encoding genes, each under transcriptional control of different promotors. The CFP fluorescent signal gene is coupled to an σ^{32} related promoter and is constitutively expressed. The *rpsJ* constitutive promoter is responsible for the YFP expression and serves as an indicator for the number of ribosomes in the cell. As placeholder for a recombinant product RFP expression is under control of the XylS/Pm promoter system. Frozen cells (stored at $-80\,°C$) were transferred into TY medium and kept for 5 h at $37\,°C$. Afterwards an aliquot was added to EnPresso B medium (Enpresso GmbH, Berlin, Germany) with $6\ U\ L^{-1}$ Reagent A and stored at $37\,°C$ overnight. The main culture was adjusted to an OD_{600} of 1 in MS medium [38] with $5\ g\ L^{-1}$ glucose and distributed to the MBR culture vessels as 10 mL aliquots. Ampicillin ($0.1\ g\ L^{-1}$) was added to all cultures for plasmid maintenance. The main cultivation was started with 2000 rpm at $37\,°C$. The stirrer speed was increased stepwise by 200 rpm every 5 min to 3000 rpm. The maximum specific growth rate of *E. coli* BW25113 (pAG032) was calculated based on previous experiments without induction at $0.72\ h^{-1}$. The applied feed rates are summarized in Table 1, the calculations are based on the equations by Enfors 2019 [39]. After six hours of cultivation the cultures were induced by the addition of 50 μL 0.1 M m-toluic acid to a final concentration 0.5 mM.

Table 1. Summary of applied feed rates for cultivation of *E. coli* BW25113 (pAG032). The listing of the reactors in columns and rows represents the reactor allocation at the two liquid handling stations (LHS).

	Reactor		μ_{set}	% of μ_{max}
1	2	3	0.65	90
4	5	6	0.58	80
7	8	9	0.50	70
10	11	12	0.43	60
13	14	15	0.36	50
16	17	18	0.29	40
19	20	21	0.22	30
22	23	24	0.14	20

2.3. Sampling and Analytics

On-line measurements for pH and DOT were taken every 30 s. For at-line analysis, the MBRs were sampled column-wise every 15 min during the cultivation, and samples were directly transferred

into V-shaped 96-microwell plates. The sampling plates contained 15 μL dried 2M NaOH to ensure direct metabolic inactivation of the samples [8]. The samples were stored for five cycles on the LHS deck at 4 °C before being transferred to the second analytic LHS. The sample storage time on the LHS deck was between 2 and 75 min. For OD_{600} and fluorescence measurements the samples were diluted by the LHS [21] and measured in a Synergy^{TM} MX microwell plate reader (BioTek Instruments GmbH, Bad Friedrichshall, Germany). The undiluted samples were filtrated to isolate the cells, and glucose and acetic acid concentrations in the supernatant were measured. The detailed procedure of the automated workflow is described in Haby et al. 2019 [21]. Outliers based on traceable technical issues are marked as invalid and not included in the data analysis.

2.4. Software Framework

Online data transfer was enabled by a server–client architecture based on the SiLA 2 (Association Consortium Standardization in Lab Automation, Rapperswil-Jona, Switzerland, sila-standard.org) standard. The server is located at the Chair of Bioprocess Engineering at Technische Universität Berlin, while the Client is distributed with the application. The server–client framework is written in Java 8 (Oracle Corporation, Santa Clara, CA, USA). The client requests information about an (running or completed) experiment to which the server replies. The server is equipped with a driver that connects to the centralized MySQL (Oracle Corporation, Santa Clara, CA, USA) database, allowing access to all process and meta data of an experiment. Upon request by a client, the server pulls the data from the database via SQL queries and returns the information to the client. The client formats the information into a string complying to the XML standard and saves it on the local machine.

2.5. Monitoring Application

The monitoring application serves as graphical user interface (Supplementary Code S1). The application itself is written in MATLAB (2018a, MathWorks, Natick, MA, USA, 2019) using the MATLAB App Designer environment. A MATLAB script initiates the client and parses the information of the XML file into a data structure compatible to MATLAB. The parsing is based on a modified version of the xml2struct function from MATLAB File Exchange [40].

2.6. Data Processing for PCA

The input variables for the PCA consisted of on-line (pH, DOT) and at-line (OD_{600}, glucose and acetate concentrations, fluorescence for red, yellow, and cyan fluorescent proteins) measurements as well as the logged volumes for base and glucose addition. The time differences between the sampling of at-line measurements were interpolated to a reference time using piecewise cubic Hermite interpolating polynomials.

The PCA of the three-dimensional dataset (reactor x variable x time) is unfolded in a batch-wise manner [41]. This approach essentially converts time to a distinguishing factor of each variable, i.e., defining one variable per time instance where it was quantified. Following the unfolding the dataset was mean-centered and scaled to unit-variance. Additionally, to account for the different frequency of measurements, the data was block-scaled: the trajectories of one variable among all was scaled by dividing each column by the square root of available number of data points [42].

2.7. Principal Component Analysis

The PCA is computed by the built-in MATLAB function *pca*. Detailed mathematical representation of the algorithm can be found in [43]. The optimal number of principal components (PCs) are selected automatically based on the improvement in % variance explained. An empirical threshold was set at 5%.

Score plots were used to visually represent the replicates' run behavior. Given the unfolding choice, the traditional loading plots become too complex with many lines for direct interpretation.

Thus, contribution plots were employed to aid the operator in relating patterns in the score plots to actual occurrences in the process [44].

PCA was applied to series of dataset collected in time that are augmented in two ways. First, a moving horizon or sliding window mode of data augmentation is used to detect failures of sensors or faulty measurements rapidly. In this approach a window length is chosen (x mins) and an update time is chosen (y mins). All data available in the window length is used to build the PCA model and a new model is built by sliding the window by y minutes. Secondly, a full horizon mode of data augmentation is used to track and compare the dynamic evolution of the different MBRs. In this approach, all data available is used to build the PCA model. In this work we chose to build a full horizon PCA model for each time in the reference time set.

The Hotelling's T^2 distance measure was used to detect unexpected drifts of an MBR compared to its replicates. Automated triggers were set for reactors outside the 90% confidence ellipse. The variable causing the behavior was automatically detected using key properties of PCA [43] to have a preliminary diagnosis of the event. The latent variable contributing most to the run was identified using the formula stated below:

$$cos^2_{i,l} = \frac{f^2_{i,l}}{\sum_l f^2_{i,l}} \tag{1}$$

where $f^2_{i,l}$ is the squared score of observation i on latent variable l. Subsequently, squared loadings of all the variables on this principal component were analyzed to identify the key driver of the failure.

2.8. Automated Warnings

Robotic experiments typically run without automated supervision systems. Failures besides arm movements or device malfunction cannot be detected unless an operator is monitoring the process. To tackle this issue, a simple method to trigger alarms was developed. For this, the Euclidean distance of each point to the center of its cluster of replicates was calculated in the sliding window triggering a message if user defined constraints were violated. This information allows the operator to quickly grasp the current status of the overall cultivation and assess the similarity of the replicates. Additionally, this first step towards online automatic classification of outliers enables more robust data selection for online optimal experimental re-design [45], a process that requires fast and thorough data selection that can hardly be done manually.

2.9. Pipetting Accuracy

To assess the pipetting accuracy of the cultivation LHS, weighed 96 well plates are filled with coloured demineralised water. Absorption maximum of the used liquid was determined via a spectral scan by the plate reader at 445 nm. The pipetting scheme was set up so each needle pipettes three columns in one row on one aspiration cycle to mimic the experimental setup in the cultivation. The factor for absorption mL^{-1} was calculated as shown below and applied to the plate.

$$\text{Factor}\left[\frac{\text{absortion}}{\text{mL}}\right] = \frac{\text{avrg. absortion of all wells} * \text{no. of wells}}{\frac{\text{weight differnce}}{\text{density water}}} = \frac{\text{total absortion}}{\text{total volume in plate}} \tag{2}$$

3. Results

The aim of the study was to demonstrate the functionality of the application and its capability to remotely monitor parallel cultivations, detect failures and guide the operator through the experiment. To this end, 24 microbial cultivations were carried out with eight different feeding strategies in triplicate. The aim of the study was to force several failures to test the performance of the program under critical conditions.

3.1. Microbial Cultivation

As model system, we use previously constructed recombinant strain *E. coli* BW25113 (pAG032) that expresses three different fluorescence proteins under individual control of three different promoters [37]. The batch phase of the 24 parallel cultivations lasted 2.6 h ± 2.04 min. The feed for the first nine reactors was started at 3.2 h of cultivation after consumption of acetic acid. The feed for all other reactors was started at 4 h. Different feed rates were set for each reactor triplicate. The feed rates varied between 20% and 90% of the maximum growth rate (see Table 1). Due to the pulse-based nature of the feeding, DOT oscillations started together with the feed additions. Cultivation data are shown in Figure 2 and available in Supplementary Table S1. Depending on the feeding rate, the increase of reactor volumes over time differs. For reactors 1–3 (with the highest feed rate of 90% of μ_{max}) the critical volume was reached after 5.5 h, reactors 10–12 (60% of μ_{max}) reached the critical volume after 8.8 h. Reactors with a low feed rate never reached critical volume levels. The DOT profile of reactor 17 decreased between 8.7 h and 9.8 h, however, this was due a technical issue and was solved during the running cultivation. When a critical volume level was reached, the DOT dropped to zero and the glucose consumption rate decreased.
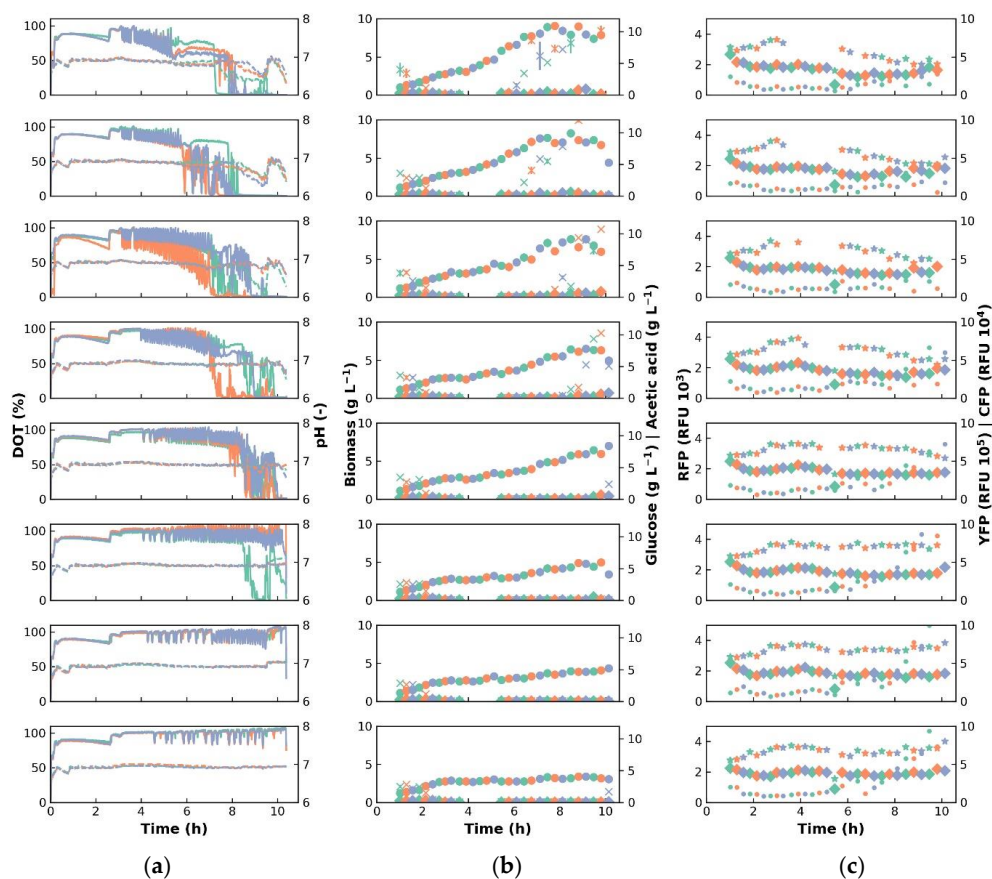

Figure 2. Cultivation data of the experiment. The rows represent a group of replicates with three mini-bioreactors (MBR). From top to bottom row, the applied feed decreased by 10% from 90% to 20% of μ_{max}. (**a**) Solid lines: DOT (%); dashed lines: pH (-); (**b**) dots: biomass (g L^{-1}); stars: glucose (g L^{-1}); diamonds: acetic acid (g L^{-1}); (**c**) dots: RFP (RFU $\times 10^3$) (*rpsJ* constitutive promoter); stars: YFP (RFU $\times 10^5$) (inducible XylS/Pm promoter); diamonds: CFP (RFU $\times 10^4$) (σ^{32} related constitutive promoter).

In all cultivations a decrease of CFP activity (σ^{32} related promoter) was observed during the batch phase. Furthermore, the CFP activity is on the same level and stays constant during the feeding phase for low and moderate feed rates. For higher feed rates (60–90% of μ_{max}) the CFP activity is lower

during the feeding phase and increases with the beginning of the oxygen limitation. The specific CFP signal increases during the batch phase and decreased during the feed phase at higher feed rates. In cultivations with a lower feed rate (20–40% of μ_{max}) the YFP signal (indicator for ribosomes per cell) stays constant during the feed rate. Between the cultivation time from 3–5 h, some YFP measurements were marked invalid because the detector limit of the used plate reader was exceeded. Later samples are analyzed in a higher dilution. For the two highest feed rates nearly no specific increase of RFP (inducible product) was observed. The highest specific RFP activity was reached with a feed rate of μ_{set} of 0.22 h^{-1} (30% of μ_{max}).

3.2. User Interface

The main features of the program developed here are (i) the operator support with a visual compression of the large number of bioreactors and variables that need to be supervised, (ii) the secure and reliable remote access via the framework, and (iii) developing an automated event trigger and fault detection tool. Additionally, a user-friendly interface was developed to demonstrate the added value of the tool and allow its test in real experiments with experienced operators.

The central program developed in MATLAB covers all Server-Client connections, data management and -analyses and offers a graphical user interface. The user may choose from different plots commonly used in PCA such as score, scree, contribution, and loading plots. Input variables for data analysis can be varied to explore different aspects. To monitor the cultivation, the application separates data for a full horizon and a moving window approach (see Figure 3).
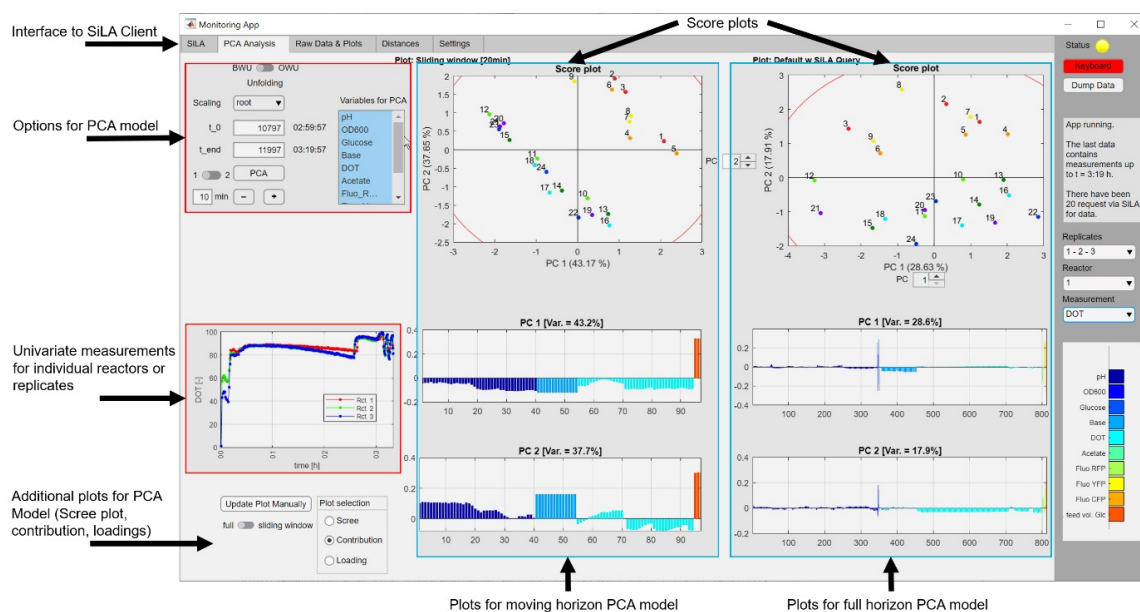

Figure 3. A screenshot of the graphical user interface of the application. The data shown is from the described experiment at 3:19 h. The left blue shows plots for the moving horizon PCA model, whereas the right blue rectangle depicts the same plots for a full horizon PCA model.

3.3. Moving and Full Horizon Setup

In the moving horizon setup, the window's timeframe was set to 20 min, a duration empirically determined based on experience and trials with historical data. Thus, the input variables for the sliding window PCA are the set points for pipetting volumes (base + feed) and the online measurements (pH + DOT).

Analysis of the loading vectors in the full horizon setup showed that the variables cumulated glucose feed and biomass correlate positively and are strongly pronounced on the first principal component. As the feed was set differently for each group of replicates, this finding is sound. However,

from 03:00 h onwards, a trend can be observed on the second PC where the scores for the reactors of the replicates have monotonous decreasing values on the y-axes (see Figure 4). The posteriori analysis of the pipetting system showed that the feedings were indeed following this trend.
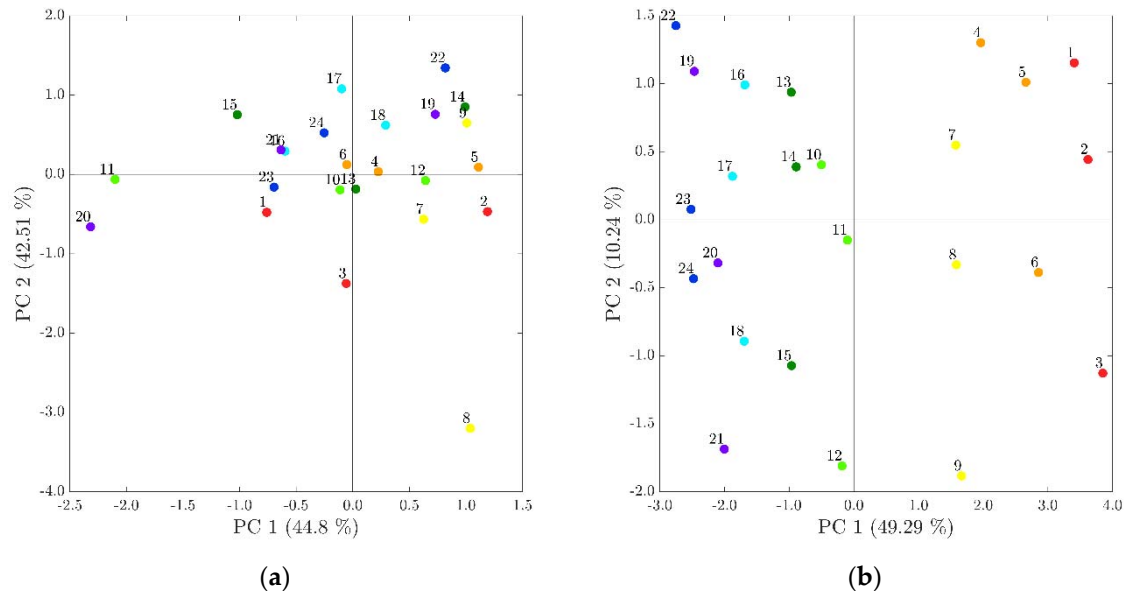

Figure 4. Full horizon PCA approach at different times. (**a**) Score plot for the full horizon principal component analysis (PCA) model at time point t = 30 min. (**b**) Score plot for the full horizon PCA model at t = 10:22 h (entire cultivation). The eight groups of replicates are indicated by color and the reactors are numbered consecutively. The variance explained by the PC is indicated in percent in parenthesis.

3.4. Event Monitoring Based on PCA

During this study, the program continuously observed the cultivations, updating its data every 10 min. Several incidents observed during the cultivation were detected properly by the program. We discuss three of these events: (1) stirrer failure in one bioreactor, caused by problems in the magnetic system, (2) overfilling of a bioreactor, caused by deactivation of the volume control, and (3) disturbance of air supply in a bioreactor, caused by, e.g., droplets in the inlet.

3.4.1. Stirring Failure

The moving horizon PCA model with eleven pH and DOT measurements revealed at least three reactors behaving differently after 20 min of batch phase. Reactors 3, 8, and 20 were identified as outliers by the automated program (see Figure 5b). DOT was identified as the causal variable for reactor 3, while pH was identified the causal variable for reactor 20. The first and second PC explain 44.4% and 43.1% variance, respectively. In the loading plots the orthogonal relation of the input variable pH and DOT is clearly visible, hence allowing to trace back the deviation in the score plot to the raw measurements (Figure 5c). While the variable from the DOT trajectory impacts the second PC almost exclusively, the pH trajectory has an impact on the first PC. Corresponding deviations were also observed in the on-line measurements. For reactors 3 and 8 lower DOT values were measured during the first 10 min of the cultivation (Figure 5a). In both reactors this was caused by a technical issue. The magnetic stirrer of reactor 8 did not start properly (this issue was detected by the operator and solved promptly). Reactor 20 is located the furthest away from the center point in respect to the first PC, indicating a lower pH but usual DOT. This finding is supported by the physiological state of the reactor.
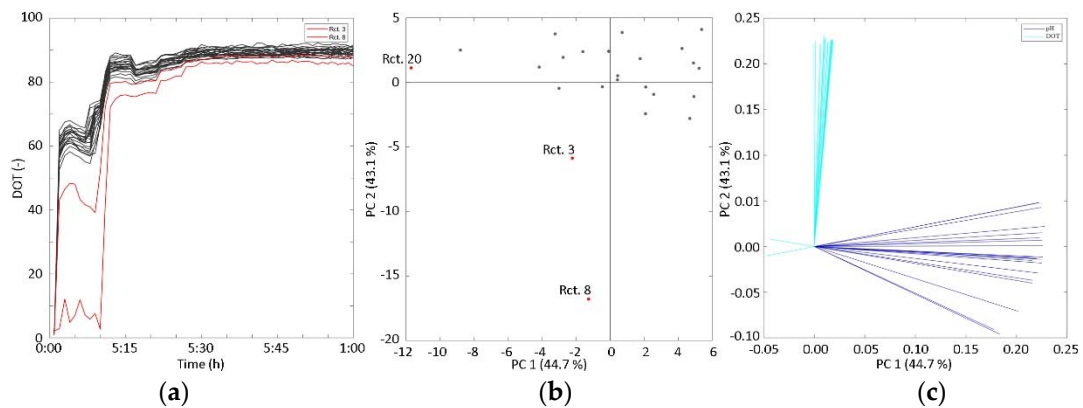
Figure 5. Detection of stirrer failures. (**a**) Trajectories for DOT and pH for all 24 reactors in the first hour of the cultivation. (**b**) Score plot for the sliding window PCA model with DOT and pH trajectories as input variables. The timeframe is $t_0 = 0$ to $t_{end} = 20$ min. Reactors 8 and 20 are distinctly separated from the main cluster. The variance explained by the PC is indicated in percent in parenthesis. (**c**) Loadings of the first two principal components for the same PCA model.

3.4.2. Reactor Overfill

At 05:40 h, the program detected reactor 3 to be an outlier. The contributing variable was identified to be the pH. Analyzing the score plot of the sliding window PCA at 05:40 h showed that reactors 2 and 3 did separate from their cluster of replicates (Figure 6c). Compared to the score plots at 05:20 h (Figure 6b), these two reactors where the only ones that did not move uniformly in one direction. Rather, the scores of reactors 2 and 3 shifted from the first to the fourth quadrant. Inspection of the first two PC shows that they explain more than 90% of the variance. The weights for the loading vector of the second component show negative correlation of pH and base addition to DOT (not shown). The DOT trajectory for these reactors did not feature the expecting oscillating pattern, indicating that the cultivation stopped reacting to the pulse-based feeding (Figure 6a).
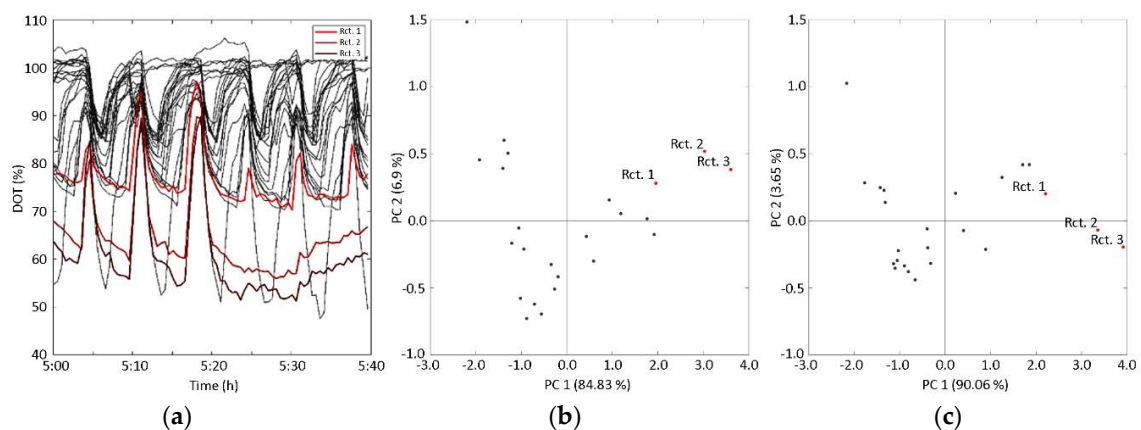

Figure 6. Reactor overfill. (**a**) Trajectories of the DOT profile from 05:00–05:40 h with highlighted reactors (1,2,3; highest feed rate). The pattern of peaks corresponds to the pulse-based feeding. (**b**,**c**) Score plot of the first two PC. The sliding window PCA model was built with data from 05:00–05:20 h (**b**) and 05:20–05:40 h (**c**) into the cultivation, respectively. The variance explained by the PC is indicated in percent in parenthesis.

While all three MBR were fed the same volume of glucose, the added volume of base differed. The total volume added decreases in reverse order of the reactors (3–1). Due to the missing volume control, this caused the reactors to exceed their upper volume limit, causing a blockage of the aeration

system. The at-line analysis of the glucose and acetate media concentration showed a drastic increase of glucose and slight increase of acetate (Figure 2).

3.4.3. Blockage of the Aeration System

Another instance causing an automated trigger was at 08:50 h when reactor 16 was identified to be an outlier with DOT as the causal variable. The score plots of the sliding window PCA models showed that the replicates (reactors 16, 17 and 18) behaved very similarly up to the time point 08:50 h (Figure 7b). At this point the score corresponding to reactor 16 abruptly digressed from the cluster (Figure 7c). The DOT profile of reactor 16 decreased unexpectedly in respect to its two replicates. This difference in sensor data was detected in the score plot for the first two PCs. The loadings for the DOT trajectories are negative in both sliding window PCA models (loading plot not shown). Comparison with the actual data implies a move of the score towards the positive axes, a trend that can be seen in the score plot (Figure 7c). The variance is explained to more than 95% by the first two principal components, indicating that the PCA model is well suited to describe the data and that the depicted score plot is sufficient for online monitoring. On the loading vector for the 1st PC, DOT and pH both correlates negatively to the base addition. Moreover, the weights of the input variable trajectories in the loadings do not differ significantly throughout one variable for the first loading vector. For the loading vector of the second PC, the weights for the most recent DOT measurement values increase (data not shown). Despite the fact that reactor 16 did not exceed its critical volume level, droplets of reactor medium may have covered the aeration hole and temporarily hinder oxygen transport into the reactor.
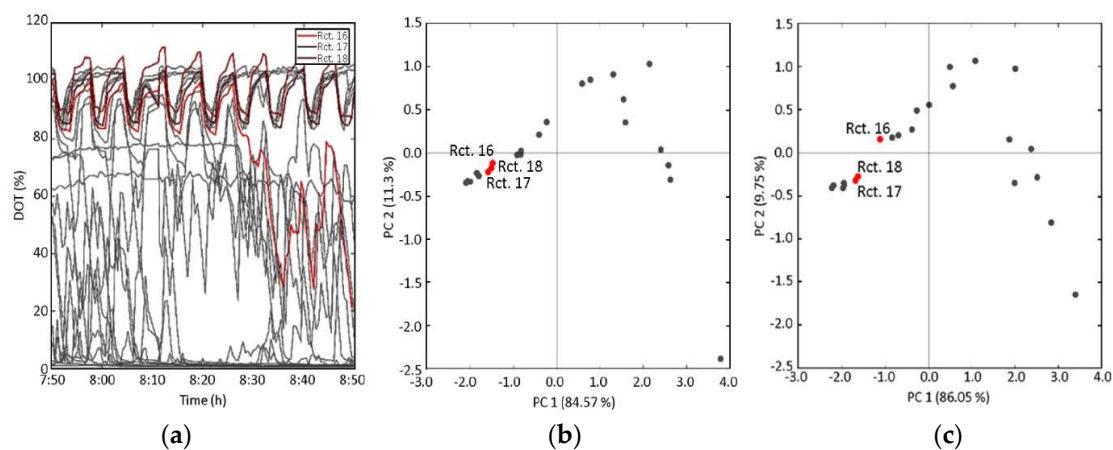

Figure 7. Aeration system blockage. (**a**) Trajectories of the DOT profile from 07:50–08:50 h into the cultivation with highlighted reactors (16,17,18; feed rate: μ_{set} = 40% of μ_{max}). The DOT profiles of Reactor 16 deviates from its group of replicates, indicating the system stops reacting to the pulse-based feeding. (**b**,**c**) Score plot based on the sliding window PCA model with input variables pH, DOT, base volume addition, and feed volume addition in the time range t_0 = 08:10 h to t_{end} = 08:30 h (**b**) and t_0 = 08:30 h to t_{end} 08:50 h (**c**).

3.4.4. Pipetting Volume Inaccuracy

Based on the information of the scalar representation of abnormality, an overall trend stood out. The score plots of the PCA after 03:00 h showed that the value of the second PC usually monotonously de- or increased (see Figure 4b), corresponding to the arrangement of the MBR in the 2mag system. The accuracy of pipetting volume on the LHS for the cultivation was investigated after conclusion of the experiment. For this the LHS was given an identical setpoint for pipetting a small volume of colored water. The absorption differed from the actual pipetting volume by 2% for each column of MBR compared to the setpoint (data not shown). This pipetting volume inaccuracy accumulated throughout the cultivation and caused the pattern of score scattering on the second PC, as their order corresponds to increasing column indices of the MBR setup.

4. Discussion

Here we present a program to monitor parallel bioreactor systems in high-throughput microbial cultivations. With this program, we can supervise if replicates follow a coherent pattern throughout the experiment or, in case of irregularities, identify the failure automatically. Furthermore, the most common handling errors like stirrer failure, overfilling, and blockage of the aeration system were properly identified during the cultivation by the program.

The major benefits of this tool are manifold: Firstly, a compression of the information originally in multivariate form into a few two-dimensional plots facilitates supervision of the process. Secondly, a basic analysis of the data using standard PCA is sufficient to detect a number of undesired events during cultivation as well as to monitor the reproducibility of the process and the operating robots. Finally, the automated event detection methods in the program are a step forward to a better and more intensive use of robotic experimental facilities, allowing longer campaigns and experiments with low supervision.

4.1. MVA for Monitoring Parallel Cultivations

With PCA as the mathematical method that is applied to the data, it is worth reflecting on its applicability in the herein described case. The PCA was used for data visualization and modest triggers such as outliers, but not for building a model of the process. This can only be done to a limited extent by the monitoring platform.

In future work, sophistication of the mathematical methods that are utilized for the tasks of data pre-processing, data analysis and model generation would improve the insights gained with the program. Some possibilities are PLS models for prediction, or the integration of existing mechanistic models [45–47] for the use of semi-parametric hybrid models [48,49]. In addition, alternative methods to the Euclidean distance that account for the correlation nature of the variable should be explored to ensure broader applicability of the classification of "unwanted" reactor behavior based on the scores. Possible methods would be a variation of the Mahalanobis distance with the benefit of considering the explained variance in the projected space [50].

Regarding the time delay of the at-line analytics, which on average accumulated to about 45 min in this cultivation, possible solutions would be to use different measurement methods or re-designing the workflow of the LHS that carried out the at-line analytics to free resources for an adapted sampling scheme. Alternative measurement methods additionally improve the timely availability of process data. As replacement of at-line measurements with online sensors, Raman spectroscopy [51], non-invasive biomass sensors [52] or fluorescence measurements of intracellular reporter systems [37] can further improve the approach. Another benefit of the presented program is the support in terms of calibration, operation, and selection of analytical devices and measurement frequency, i.e., the sampling and analytics workflow of the LHS could be adjusted dynamically based on the importance of input variables to the PCA as indicated in the loadings of the PCs [53].

4.2. Common Failure Events

The stirrer failure event demonstrated that quick reaction is crucial for a successful cultivation. A supervision of the cultivations' status by simple visual inspection of each critical process variable in each reactor is already difficult for 24 reactors. The undesired reactor disturbances and the variables with the largest contribution to it are automatically detected using modest approach such as the inherent properties of PCA. Additionally, the trends are visualized in a concise two-dimensional score and contribution plots which allows the operator to examine the status of reactors of interest in more depth. However, as with the stirrer failure, the timely availability of data remains a key issue.

Still, the results also show that the abstraction of PCA is not sufficient to derive concrete measures to steer the process back to the experimental constraints. However, in the case of the events that caused the sudden drop of DOT, PCA proved to be useful in the reduction of the data dimension.

As the feeding strategy in high throughput experiments usually varies for each group of replicates and the DOT trajectory is oscillating due to the pulse-based feeding, visually inspecting the profiles for all reactors throughout a 10 h cultivation would become very cumbersome or even impossible. Thus, as demonstrated in this paper, score plots are a great tool to help identify irregularities.

5. Conclusions

We here develop and demonstrate the benefits of a remote tool for online multivariate analysis of parallel cultivations and its applicability to monitor dynamic HT experiments—a task that has become hardly manageable by lab operators. The black box nature of the MVA makes it independent of the type of experiments or cultivation vessels used. This feature makes it particularly useful for early stage bioprocess development with little process understanding and highly varying experimental systems. We have shown that the program is able to provide data-based insight to guide operator decisions in a large multidimensional data set at the example of complex HT fed-batch cultivations. Despite high process variability and the large amount of datapoints, the tool simplified the task to derive decisions and allows conclusion to be drawn about the similarity of replicates and causes for deviations, improving the reproducibility of a bioprocess.

The freely available program (see below) presented in this work is a useful tool for the monitoring and operation of high throughput bioprocess development facilities. Due to the modular program structure and the fact that the software relies on an open server–client interface, only small code adjustments on the backend are necessary if the app is deployed in a different environment, such as a different working group.

This approach is in the scope of the different model variants [13]: A descriptive model allows for visualization and interpretation in a reduced dimension space along the central process information. Moreover, it also supports diagnosis of abnormalities and deviations. For the future, it would be important to further develop the tool towards predictive and prescriptive capabilities, i.e., to proactively foresee deviations and suggest on possible process control actions, so to most efficiently utilize the experimental data and potential of each ongoing experiment.

References

1. Neubauer, P.; Glauche, F.; Bournazou, M.N.C. Editorial: Bioprocess Development in the era of digitalization. *Eng. Life Sci.* **2017**, *17*, 1140–1141. [CrossRef]
2. Neubauer, P. Editorial: Towards faster bioprocess development. *Biotechnol. J.* **2011**, *6*, 902–903. [CrossRef] [PubMed]

3. Islam, R.S.; Tisi, D.; Levy, M.S.; Lye, G.J. Framework for the Rapid Optimization of Soluble Protein Expression in Escherichia coli Combining Microscale Experiments and Statistical Experimental Design. *Biotechnol. Prog.* **2007**, *23*, 785–793. [CrossRef] [PubMed]

4. Glauche, F.; Pilarek, M.; Bournazou, M.N.C.; Grunzel, P.; Neubauer, P. Design of experiments-based high-throughput strategy for development and optimization of efficient cell disruption protocols. *Eng. Life Sci.* **2016**, *17*, 1166–1172. [CrossRef]

5. Küppers, T.; Steffen, V.; Hellmuth, H.; O'Connell, T.; Bongaerts, J.; Maurer, K.-H.; Wiechert, W. Developing a new production host from a blueprint: Bacillus pumilus as an industrial enzyme producer. *Microb. Cell Factories* **2014**, *13*, 46. [CrossRef]

6. Jorgensen, J.T. A challenging drug development process in the era of personalized medicine. *Drug Discov. Today* **2011**, *16*, 891–897. [CrossRef]

7. Paritala, P.K.; Manchikatla, S.; Yarlagadda, P.K.D.V. Digital Manufacturing- Applications Past, Current, and Future Trends. *Procedia Eng.* **2017**, *174*, 982–991. [CrossRef]

8. Nickel, D.B.; Bournazou, M.N.C.; Wilms, T.; Neubauer, P.; Knepper, A. Online bioprocess data generation, analysis, and optimization for parallel fed-batch fermentations in milliliter scale. *Eng. Life Sci.* **2016**, *17*, 1195–1201. [CrossRef]

9. Kim, J.H. A Review of Cyber-Physical System Research Relevant to the Emerging IT Trends: Industry 4.0, IoT, Big Data, and Cloud Computing. *J. Ind. Integr. Manag.* **2017**, *2*, 1750011. [CrossRef]

10. Ghobakhloo, M. The future of manufacturing industry: A strategic roadmap toward Industry 4.0. *J. Manuf. Technol. Manag.* **2018**, *29*, 910–936. [CrossRef]

11. Gani, R. *Computer Aided Process and Product Engineering, Comput*; WILEY-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2006; pp. 647–666. [CrossRef]

12. Stephanopoulos, G.; Reklaitis, G.V. Process systems engineering: From Solvay to modern bio- and nanotechnology. *Chem. Eng. Sci.* **2011**, *66*, 4272–4306. [CrossRef]

13. Narayanan, H.; Luna, M.F.; Von Stosch, M.; Bournazou, M.N.C.; Polotti, G.; Morbidelli, M.; Butté, A.; Sokolov, M. Bioprocessing in the Digital Age: The Role of Process Models. *Biotechnol. J.* **2019**, *15*, e1900172. [CrossRef] [PubMed]

14. Abt, V.; Barz, T.; Cruz-Bournazou, M.N.; Herwig, C.; Kroll, P.; Möller, J.; Pörtner, R.; Schenkendorf, R.; Cruz, N. Model-based tools for optimal experiments in bioprocess engineering. *Curr. Opin. Chem. Eng.* **2018**, *22*, 244–252. [CrossRef]

15. Löffler, M.; Simen, J.D.; Jäger, G.; Schäferhoff, K.; Freund, A.; Takors, R. Engineering E. coli for large-scale production – Strategies considering ATP expenses and transcriptional responses. *Metab. Eng.* **2016**, *38*, 73–85. [CrossRef]

16. Sonnleitner, B. Bioprocess automation and bioprocess design. *J. Biotechnol.* **1997**, *52*, 175–179. [CrossRef]

17. Wilson, L.J.; Lewis, W.; Kucia-Tran, R.; Bracewell, D.G. Identification of upstream culture conditions and harvest time parameters that affect host cell protein clearance. *Biotechnol. Prog.* **2019**, *35*, e2805. [CrossRef]

18. Randek, J.; Mandenius, C.-F. On-line soft sensing in upstream bioprocessing. *Crit. Rev. Biotechnol.* **2017**, *38*, 106–121. [CrossRef]

19. Kourti, T. Process Analytical Technology Beyond Real-Time Analyzers: The Role of Multivariate Analysis. *Crit. Rev. Anal. Chem.* **2006**, *36*, 257–278. [CrossRef]

20. Hemmerich, J.; Noack, S.; Wiechert, W.; Oldiges, M. Microbioreactor Systems for Accelerated Bioprocess Development. *Biotechnol. J.* **2018**, *13*, 1700141. [CrossRef]

21. Haby, B.; Hans, S.; Anane, E.; Sawatzki, A.; Krausch, N.; Neubauer, P.; Bournazou, M.N.C. Integrated Robotic Mini Bioreactor Platform for Automated, Parallel Microbial Cultivation With Online Data Handling and Process Control. *SLAS Technol. Transl. Life Sci. Innov.* **2019**, *24*, 569–582. [CrossRef]

22. Anane, E.; García, Á.C.; Haby, B.; Hans, S.; Krausch, N.; Krewinkel, M.; Hauptmann, P.; Neubauer, P.; Bournazou, M.N.C. A model-based framework for parallel scale-down fed-batch cultivations in mini-bioreactors for accelerated phenotyping. *Biotechnol. Bioeng.* **2019**, *116*, 2906–2918. [CrossRef] [PubMed]

23. Janzen, N.H.; Striedner, G.; Jarmer, J.; Voigtmann, M.; Abad, S.; Reinisch, D. Implementation of a Fully Automated Microbial Cultivation Platform for Strain and Process Screening. *Biotechnol. J.* **2019**, *14*, e201800625. [CrossRef]

24. Bournazou, M.N.C.; Barz, T.; Nickel, D.B.; Cárdenas, D.L.; Glauche, F.; Knepper, A.; Neubauer, P. Online optimal experimental re-design in robotic parallel fed-batch cultivation facilities. *Biotechnol. Bioeng.* **2016**, *114*, 610–619. [CrossRef] [PubMed]

25. Barz, T.; Sommer, A.; Wilms, T.; Neubauer, P.; Bournazou, M.N.C.; Throughput, A. Adaptive optimal operation of a parallel robotic liquid handling station. In Proceedings of the 9th Vienna Internacional Conference Mathematical Model, Vienna, Austria, 21–23 February 2018; pp. 901–906. [CrossRef]

26. Hans, S.; Gimpel, M.; Glauche, F.; Neubauer, P.; Bournazou, M.N.C.; Gimpel, M. Automated Cell Treatment for Competence and Transformation of Escherichia coli in a High-Throughput Quasi-Turbidostat Using Microtiter Plates. *Microorganisms* **2018**, *6*, 60. [CrossRef] [PubMed]

27. Sawatzki, A.; Hans, S.; Narayanan, H.; Haby, B.; Krausch, N.; Sokolov, M.; Glauche, F.; Riedel, S.L.; Neubauer, P.; Bournazou, M.N.C. Accelerated Bioprocess Development of Endopolygalacturonase-Production with Saccharomyces cerevisiae Using Multivariate Prediction in a 48 Mini-Bioreactor Automated Platform. *Bioengineering* **2018**, *5*, 101. [CrossRef] [PubMed]

28. Nomikos, P.; MacGregor, J.F. Monitoring batch processes using multiway principal component analysis. *AIChE J.* **1994**, *40*, 1361–1375. [CrossRef]

29. Sokolov, M.; Ritscher, J.; MacKinnon, N.; Brühlmann, D.; Rothenhäusler, D.; Thanei, G.; Soos, M.; Stettler, M.; Souquet, J.; Broly, H.; et al. Robust factor selection in early cell culture process development for the production of a biosimilar monoclonal antibody. *Biotechnol. Prog.* **2016**, *33*, 181–191. [CrossRef] [PubMed]

30. Undey, C.; Ertunç, S.; Mistretta, T.; Looze, B. Applied advanced process analytics in biopharmaceutical manufacturing: Challenges and prospects in real-time monitoring and control. *J. Process. Control.* **2010**, *20*, 1009–1018. [CrossRef]

31. Gunther, J.; Baclaski, J.; Seborg, D.; Conner, J. Pattern matching in batch bioprocesses—Comparisons across multiple products and operating conditions. *Comput. Chem. Eng.* **2009**, *33*, 88–96. [CrossRef]

32. Thomassen, Y.E.; Van Sprang, E.N.; Van Der Pol, L.A.; Bakker, W.A. Multivariate data analysis on historical IPV production data for better process understanding and future improvements. *Biotechnol. Bioeng.* **2010**, *107*, 96–104. [CrossRef]

33. Kirdar, A.; Conner, J.; Baclaski, J.; Rathore, A.S. Application of Multivariate Analysis toward Biotech Processes: Case Study of a Cell-Culture Unit Operation. *Biotechnol. Prog.* **2007**, *23*, 61–67. [CrossRef] [PubMed]

34. Wang, X.; Kruger, U.; Irwin, G.W. Process Monitoring Approach Using Fast Moving Window PCA. *Ind. Eng. Chem. Res.* **2005**, *44*, 5691–5702. [CrossRef]

35. Brand, E.; Junne, S.; Anane, E.; Bournazou, M.N.C.; Neubauer, P. Importance of the cultivation history for the response of Escherichia coli to oscillations in scale-down experiments. *Bioprocess Biosyst. Eng.* **2018**, *41*, 1305–1313. [CrossRef] [PubMed]

36. Delvigne, F.; Goffin, P. Microbial heterogeneity affects bioprocess robustness: Dynamic single-cell analysis contributes to understanding of microbial populations. *Biotechnol. J.* **2013**, *9*, 61–72. [CrossRef]

37. Gawin, A.; Peebo, K.; Hans, S.; Ertesvåg, H.; Irla, M.; Neubauer, P.; Brautaset, T. Construction and characterization of broad-host-range reporter plasmid suitable for on-line analysis of bacterial host responses related to recombinant protein production. *Microb. Cell Factories* **2019**, *18*, 80. [CrossRef]

38. Anane, E.; Sawatzki, A.; Neubauer, P.; Bournazou, M.N.C. Modelling concentration gradients in fed-batch cultivations of E. coli - towards the flexible design of scale-down experiments. *J. Chem. Technol. Biotechnol.* **2018**, *94*, 516–526. [CrossRef]

39. Enfors, S.-O. *Fermentation Process Technology*; Technische Universität Berlin: Sven-Olof Enfors, Stockholm, Sweden, 2019.

40. Wouter Falkena, xml2struct, MATLAB Central File Exchange. Available online: https://www.mathworks.com/matlabcentral/fileexchange/28518-xml2struct (accessed on 1 January 2019).

41. MacGregor, J.; Kourti, T. Statistical process control of multivariate processes. *Control. Eng. Pr.* **1995**, *3*, 403–414. [CrossRef]

42. Jolliffe, I.T. Principal Component Analysis. *Springer Series in Statistics* **2002**, *98*, 487. [CrossRef]

43. Abdi, H.; Williams, L. Principal component analysis. Wiley Interdiscip. *Rev. Comput. Stat.* **2010**, *2*, 433–459. [CrossRef]

44. Anane, E.; Neubauer, P.; Bournazou, M.N.C. Modelling overflow metabolism in Escherichia coli by acetate cycling. *Biochem. Eng. J.* **2017**, *125*, 23–30. [CrossRef]

45. Sokolov, M.; Ritscher, J.; MacKinnon, N.; Souquet, J.; Broly, H.; Morbidelli, M.; Butté, A. Enhanced process understanding and multivariate prediction of the relationship between cell culture process and monoclonal antibody quality. *Biotechnol. Prog.* **2017**, *33*, 1368–1380. [CrossRef] [PubMed]
46. Duan, Z.; Wilms, T.; Neubauer, P.; Kravaris, C.; Bournazou, M.N.C. Model reduction of aerobic bioprocess models for efficient simulation. *Chem. Eng. Sci.* **2020**, *217*, 115512. [CrossRef]
47. Von Stosch, M.; Hamelink, J.-M.; Oliveira, R. Hybrid modeling as a QbD/PAT tool in process development: An industrial E. coli case study. *Bioprocess Biosyst. Eng.* **2016**, *39*, 773–784. [CrossRef] [PubMed]
48. Narayanan, H.; Sokolov, M.; Morbidelli, M.; Butté, A. A new generation of predictive models: The added value of hybrid models for manufacturing processes of therapeutic proteins. *Biotechnol. Bioeng.* **2019**, *116*, 2540–2549. [CrossRef] [PubMed]
49. Neubauer, P.; Cruz, N.; Glauche, F.; Junne, S.; Knepper, A.; Raven, M. Consistent development of bioprocesses from microliter cultures to the industrial scale. *Eng. Life Sci.* **2013**, *13*, 224–238. [CrossRef]
50. 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. [CrossRef] [PubMed]
51. Lee, H.L.; Boccazzi, P.; Gorret, N.; Ram, R.J.; Sinskey, A.J. In situ bioprocess monitoring of Escherichia coli bioreactions using Raman spectroscopy. *Vib. Spectrosc.* **2004**, *35*, 131–137. [CrossRef]
52. Buchenauer, A.; Hofmann, M.; Funke, M.; Büchs, J.; Mokwa, W.; Schnakenberg, U. Micro-bioreactors for fed-batch fermentations with integrated online monitoring and microfluidic devices. *Biosens. Bioelectron.* **2009**, *24*, 1411–1416. [CrossRef]
53. Kourti, T.; Nomikos, P.; MacGregor, J.F. Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS. *J. Process. Control.* **1995**, *5*, 277–284. [CrossRef]