TECHNISCHE UNIVERSITÄT MÜNCHEN

TUM School of Life Sciences

Soft sensors for *Pichia pastoris* bioprocesses: filling the gaps between uncertain process data and knowledge

Vincent Brunner

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzende: Prof. Dr.-Ing. Petra Först

Prüfer der Dissertation:
1. Prof. Dr.-Ing. Thomas Becker
2. Prof. Dr.-Ing. Dirk Weuster-Botz

Die Dissertation wurde am 02.05.2022 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 02.11.2022 angenommen.

Abstract

Bioprocesses pose challenges for process monitoring and quality control that are not present in typical chemical processes. For the monitoring of key quantities such as biomass, product, and substrate concentration, direct measurement methods only exist in very few cases. Further, analyte concentrations are typically very low. Against the background of often not completely defined bioprocess media, these low concentrations are difficult to measure online with sufficient accuracy due to potential cross-sensitivities. Soft sensors provide a remedy to this situation. "Software sensors" use existing process data (e.g., readings of different sensors) as inputs to a predictive model in order to indirectly determine a target quantity (e.g., biomass concentration). With regard to the data-information-knowledge hierarchy, the soft sensor model is in most cases used to compress process data into information (target quantity). This type of model is referred to as data-driven model. Two problems stand out here: First, process knowledge is neglected in this scenario. Second, the model inputs are assumed to be true so that faulty input data may corrupt the derived information.

This thesis aims to develop approaches beyond this core function of soft sensors (compression of data to information). For this purpose, three methodical building blocks are presented, which can be used modularly in the development of soft sensors. The first methodical building block comprises an approach for deriving process knowledge from the analysis of a soft sensor. The second building block serves to implement process knowledge in a hybrid model for a bioprocess with multiple process phases. The third building block provides validation of uncertain model inputs (e.g., due to sensor faults). This in turn allows to validate the prediction of a soft sensor. Swarm intelligence is used here to make the fault detection algorithm work more efficiently.

The soft sensors presented in this thesis are developed based on a standard biotechnological process: the cultivation of the methylotrophic yeast *Pichia pastoris*. This bioprocess is well established, but still offers potential for optimization in the areas of monitoring, control, and automation. Further, this process shows time-variant and non-linear behavior. This must be taken into account when developing soft sensors.

The presented approaches represent important building blocks to fill the gaps between uncertain process data and knowledge in the development of soft sensors. Each of them contributes to the process analytical technology toolbox and should promote the acceptance of soft sensors for quality control in the biotechnology industry.

Zusammenfassung

Bei Bioprozessen treten spezifische Herausforderungen an die Prozessüberwachung und Qualitätskontrolle auf, die bei typischen chemischen Prozessen nicht vorliegen. Für die Überwachung von Schlüsselgrößen wie Biomasse-, Produkt- und Substratkonzentration gibt es nur in wenigen Fällen direkte Messmethoden. Zudem sind die Analytkonzentrationen häufig sehr niedrig. Vor dem Hintergrund der oft nicht vollständig definierten Bioprozessmedien sind diese niedrigen Konzentrationen aufgrund von möglichen Querempfindlichkeiten nur schwer online mit ausreichender Genauigkeit zu messen. Softsensoren schaffen hier Abhilfe. “Software-Sensoren” nutzen vorhandene Prozessdaten (z. B. Messwerte verschiedener Sensoren) als Eingangsgrößen für ein Vorhersagemodell, um damit eine Zielgröße (z. B. Biomassekonzentration) indirekt zu bestimmen. Bezogen auf die Daten-Information-Wissen-Hierarchie wird das Softsensor-Modell in den meisten Fällen verwendet, um Prozessdaten zu Information (Zielgröße) zu komprimieren. Diese Art von Modell wird als datengetriebenes Modell bezeichnet. Dabei fallen zwei Probleme auf: Erstens bleibt in diesem Szenario vorhandenes Prozesswissen ungenutzt. Zweitens werden die Modelleingänge als wahr angenommen, sodass fehlerhafte Eingangsdaten die abgeleitete Information verfälschen können.
Ziel dieser Arbeit ist es, Ansätze zu entwickeln, die über diese zentrale Funktion von Softsensoren (Kompression von Daten zu Information) hinausgehen. Hierfür werden drei methodische Bausteine vorgestellt, die bei der Entwicklung von Softsensoren modular eingesetzt werden können. Der erste methodische Baustein beschreibt einen Ansatz zur Ableitung von Prozesswissen aus der Analyse eines Softsensors. Der zweite Baustein dient der Implementierung von Prozesswissen in ein hybrides Modell für einen Bioprozess mit mehreren Prozessphasen. Der dritte Baustein erlaubt die Validierung unsicherer Modelleingänge (z. B. durch Sensorfehler). Dies ermöglicht es wiederum, die Vorhersage eines Softsensors zu validieren. Hierbei wird Schwarmintelligenz genutzt, um den Algorithmus zur Fehlererkennung effizienter zu gestalten.
Die in dieser Arbeit vorgestellten Softsensoren werden auf Basis eines biotechnologischen Standardprozesses entwickelt: der Kultivierung der methylotrophen Hefe *Pichia pastoris*. Dieser Bioprozess ist gut etabliert, bietet aber ausreichend Optimierungspotenzial in den Bereichen der Überwachung, Regelung und Automatisierung. Weiter zeigt dieser Bioprozess zeitvariantes und nichtlineares Verhalten. Dies muss bei der Entwicklung von Softsensoren berücksichtigt werden.
Die vorgestellten Ansätze stellen wichtige Bausteine dar, um die Lücken zwischen unsicheren Prozessdaten und Wissen bei der Entwicklung von Softsensoren zu schließen. Jeder von ihnen trägt zum Werkzeugkasten der *Process Analytical Technology* bei und ist geeignet, die Akzeptanz von Softsensoren für die Qualitätskontrolle in der biotechnologischen Industrie zu fördern.

Publications

The following list contains **peer-reviewed publications** and non-peer-reviewed publications resulting from this work:

- **Brunner, V., Siegl, M., Geier, D., and Becker, T. (2021). Challenges in the development of soft sensors for bioprocesses: a critical review. *Frontiers in bioengineering and biotechnology* 9, 722202. doi: 10.3389/fbioe.2021.722202**
- **Brunner, V., Siegl, M., Geier, D., and Becker, T. (2020). Biomass soft sensor for a *Pichia pastoris* fed-batch process based on phase detection and hybrid modeling. *Biotechnology and bioengineering* 117, 2749–2759. doi: 10.1002/bit.27454**
- Brunner, V., Geier, D., Becker, T. (2020). Das „Unmessbare“ messbar machen: Softsensoren helfen, die Prozessführung in der Biotechnologie zu verbessern. *CITplus* 23, 28–29. doi: 10.1002/citp.202000415
- **Brunner, V., Klöckner, L., Kerpes, R., Geier, D. U., and Becker, T. (2019). Online sensor validation in sensor networks for bioprocess monitoring using swarm intelligence. *Analytical and bioanalytical chemistry* 412, 2165–2175. doi: 10.1007/s00216-019-01927-7**
- **Brunner, V., Hussein, M., and Becker, T. (2016). Biomass estimation in *Pichia pastoris* cultures by combined single-wavelength fluorescence measurements. *Biotechnology and bioengineering* 113, 2394–2402. doi: 10.1002/bit.26003**

Conference contributions

The following list contains **oral presentations** and poster presentations resulting from this work:

- **Brunner, V., Klöckner, L., Geier, D., Becker, T. (2018). Online sensor validation for a *Pichia pastoris* bioprocess using swarm intelligence based dynamic modeling. Smart Sensors – Mechanistic and Data Driven Modeling. Frankfurt, Germany, 01.–02.10.2018.**
- **Brunner, V., Klöckner, L., Geier, D., Becker, T. (2018). How can we detect sensor faults in sensor networks for bioprocesses?. ProcessNet-Jahrestagung und 33. DECHEMA-Jahrestagung der Biotechnologen 2018. Aachen, Germany, 10.–13.09.2018.**
- **Brunner, V., Geier, D., Becker, T. (2017). Use of chemometrics and single wavelength fluorescence measurements for biomass estimation in *Pichia pastoris* cultures. Himmelfahrtstagung 2017: Models for Developing and Optimising Biotech Production. Neu-Ulm, Germany, 22.–24.05.2017.**
- Brunner, V., Metzenmacher, M., Birle, S., Geier, D., Becker, T. (2017). Innovative methods of knowledge-driven methanol feedback control and state estimation for recombinant *Pichia pastoris* fed-batch cultivation. 4th European Conference on Process Analytics and Control Technology (EuroPACT 2017). Potsdam, Germany, 10.–12.05.2017.

Contents

1 Introduction

Quality by design (QbD) principles and process analytical technology (PAT) tools form the framework for a sustainable, risk-minimized, and automated manufacturing process for biological products.
The term QbD describes the building in of quality into the product by a thorough understanding of the relationships between the clinical properties of the product, the critical quality attributes (CQAs), the process, and the variability in raw materials (Rathore, 2009; Rathore and Winkle, 2009). By understanding these relationships, a (multidimensional) design space can be defined in which critical process parameters (CPPs) have been demonstrated to assure quality (Streefland et al., 2013). QbD can be seen as an "umbrella encompassing […] concepts including creation of a manufacturing knowledge base, risk-management principles, process design spaces, and PAT" (Read et al., 2010). PAT tools can in this context be used for online monitoring and control of CPPs and CQAs so that the process performance and product quality, respectively, can be assessed during manufacturing. The US Food and Drug Administration (FDA) distinguishes four main PAT tools in their 2004 PAT initiative (FDA, 2004): (1) multivariate tools for design, data acquisition, and analysis; (2) process analyzers; (3) process control tools; (4) continuous improvement and knowledge management tools. These tools form a system that allows control of measurable processes toward a desired endpoint and enables improvement of final product quality by reducing variability. The basic idea is to determine the variability of the process inputs (e.g., media components) and of the process itself in a timely manner; this allows a dynamic response to this variability, e.g., via closed-loop control. The term analytical in PAT includes "chemical, physical, microbiological, mathematical, and risk analysis conducted in an integrated manner" (FDA, 2004).
Within this framework, soft sensors are becoming increasingly important as PAT tool. A soft sensor is based on a mathematical model that allows determining a target process quantity indirectly. This is especially important when the direct measurement via process analyzers is not economically or technically feasible. For monitoring CQAs such as biomass or product concentration, soft sensors are in some cases the only solution to determine the target value online at all.

In the following sections, it is first shown how soft sensors integrate into the data-information-knowledge hierarchy of bioprocess monitoring. Furthermore, it is shown how soft sensors, as a PAT tool, allow to make the non-measurable measurable—even with uncertain input data. In addition, it is shown how soft sensors can help to build up a manufacturing knowledge base according to the QbD principles and how to use this knowledge for quality control of biotechnological processes. Finally, the characteristics of the cultivation of *Pichia pastoris* (now reclassified as *Komagataella phaffii*), which is chosen as use case in this thesis, are described.

1.1 The data-information-knowledge hierarchy in bioprocess monitoring

Within bioprocess monitoring, process data are initially generated by process analyzers and laboratory analyses. These *data* are products of observations (Rowley, 2007), and thus primarily raw signals without meaning. Once the data are processed in a way to be useful, they are referred to as *information* (Ackoff, 1989). Here, a condensation process of a wealth of data to little information takes place, which in the optimal case leads to a reduction of meaningless noise. The following two real-world scenarios illustrate this condensation process from unorganized and meaningless data to compact information: First, substrate concentration (e.g., glucose) in a bioprocess is often determined using chromatograms and other raw data such as bioreactor volume, sample volume, and dilution factors; second, the specific growth rate is typically determined using raw data such as cell mass, sample volume, and time. In such cases, the substrate concentration and the specific growth rate represent the meaningful information that can be used for monitoring, control, and fault diagnosis; whereas raw process data alone are of no particular use unless contextualized. This example shows that although data and information may have the same structure (value and unit), they differ in their function and their degree of condensation (Ackoff, 1989; Davenport and Prusak, 1998).

Information can further be used to generate knowledge. *Knowledge* is obtained by interconnecting meaningful information in the right way. Sticking to the above example, the relationship between the limiting substrate concentration S and the specific growth rate μ can be revealed via a suitable experimental design. In the form of the Monod equation, $\mu = \mu_{max} S/(K_S + S)$, with the Monod constant K_S and the maximum specific growth rate μ_{max} (Monod, 1949), this knowledge is quantifiable as well as transferable. These two properties are essential for knowledge management.

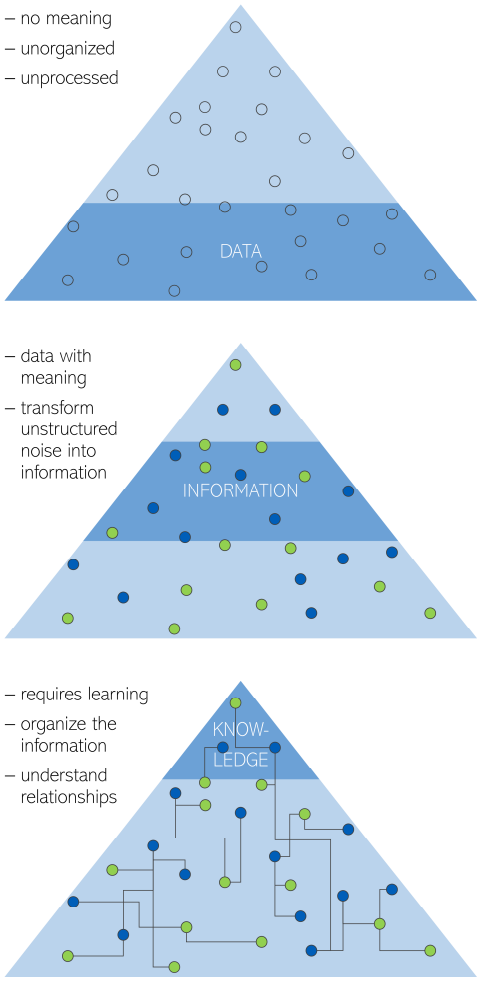


Figure 1: Pyramidal representation of the data-information-knowledge hierarchy. Initially, data are just a set of signals without meaning (shown as *empty dots*). When transforming data into information, the data get a meaning (*colored dots*). Finally, knowledge is obtained by inter-connecting meaningful information in the right way (*connected colored dots*). Illustration of the DIK context using connected colored dots originally by Hugh MacLeod (gapingvoid.com/semiotic-management-systems).

From this example, it can be seen that a large amount of data (raw data matrices of experiments) is condensed to a smaller amount of relevant information (S and μ). The knowledge about the relationships between S and μ resulting from several experiments can again be condensed to only one equation with two constants (K_S and μ_{max}). This condensation process that takes place when a wealth of data ultimately results in the literal tip of the iceberg—knowledge—is often illustrated with the help of the pyramidal data-information-knowledge (DIK) hierarchy (Figure 1). This representation is widely used in the literature on knowledge management and in some cases includes even higher levels, such as understanding, wisdom, or insight (Rowley, 2007; Jennex, 2017). However, entities higher than knowledge are not within the scope of this thesis.

The purpose of this short excursus on the DIK hierarchy was to show how bioprocess data condenses into usable information and ultimately into knowledge. How do soft sensors fit into the DIK hierarchy? The short answer first: The core function of soft sensors is the derivation of information from process data (Mandenius and Gustavsson, 2015). This information can be, for example, the prediction of a target value (e.g., S or μ) or the occurrence of a process deviation or sensor fault (Kadlec et al., 2009). However, existing knowledge is rarely left unused. Especially in the field of bioprocesses, existing knowledge in the form of mechanistic relationships has ever since the emergence of modern biotechnology assisted in the development of soft sensors (see exemplarily Yousefi-Darani et al. (2020) for Kalman filters). Thus, in order to answer the question of how soft sensors fit into the DIK hierarchy in more depth, a distinction must first be made between data-driven, mechanistic, and hybrid soft sensors.

1.2 Soft sensors: development and utilization

With regard to the DIK hierarchy, this section focuses primarily on the role of the tip (knowledge) as well as the composition and quality of the base of the pyramid (data). First, the role of process knowledge in the context of soft sensors is explored. Here it is shown how process knowledge can be implemented in soft sensors and, on the other hand, how process knowledge can be generated via soft sensors. Second, the challenge of uncertain input data to the soft sensor is overviewed and discussed. According to the "garbage in, garbage out" principle, the more erroneous the inputs, the less accurate the developed soft sensor and its predictions become.

1.2.1 The role of process knowledge

Dependent on the degree of process knowledge that is implemented, the soft sensor model can be classified as data-driven, mechanistic, or hybrid. Data-driven approaches use modeling methods from the spectrum of data science, such as variants of multiple linear regression (MLR; Jenzsch et al. (2006)), principal component regression (PCR; Zhu et al. (2018)), partial least squares regression (PLSR; Sokolov et al. (2015); Zheng and Song (2018)), support vector regression (SVR; Meng et al. (2019)), and artificial neural networks (ANN; Paquet-Durand et al. (2017)). Mechanistic

approaches make use of the available process knowledge in the form of first principle models, such as mass or energy balances (Sagmeister et al., 2013; Ohadi et al., 2015; Tahir et al., 2019). For example, Ohadi et al. (2015) made use of mass balances in the form of stoichiometric relationships between substrate uptake, growth, and product formation rates to predict the key process variables of a mammalian cell culture process (viable and dead cells, recombinant protein, glucose, and ammonia concentrations). In fed-batch processes, mass balances play a particularly important role because the mass of the system varies during the process (Sagmeister et al., 2013). In general, it can be assumed that purely data-driven approaches demand a larger amount of training data than purely mechanistic approaches (Stosch et al., 2014). Hybrid approaches combine both data-driven and mechanistic model parts. For example, a mechanistic model can use the output of a data-driven model as input if certain model terms are missing (no online measurements available) or cannot be predicted mechanistically with sufficient accuracy (Stosch et al., 2014; Solle et al., 2017). In addition, mechanistic model parts can be combined with other process data (e.g., fluorescence spectra) via (extended) Kalman filtering for a more accurate prediction compared to a purely mechanistic model (Ohadi et al., 2015).
As mentioned at the very beginning, knowledge management is an important element of QbD and PAT. Most of the knowledge about the product and process is accumulated through targeted experimentation during product and process development (Herwig et al., 2015). Using the mechanistic and hybrid approaches described above, this knowledge can be implemented during soft sensor development to make soft sensors more accurate and robust. However, little attention has been paid so far to the generation of process knowledge *by means of* soft sensor development.

The statistical methods used within soft sensor development can reveal complex relationships in data and assist in extracting information, which finally can improve process understanding (Matero et al., 2013). For example, variable selection and correlation analysis can as parts of chemometric methods serve for improving process understanding. Variable selection is typically used iteratively during soft sensor development to select the input variables that bear information for predicting the target variable. This preselection serves to reduce (multi)collinearity within the input data, which in turn can improve the accuracy of model coefficient estimates (Ma et al., 2009). Furthermore, computational costs as well as overfitting can be reduced by variable selection (Hawkins, 2004; Kaneko and Funatsu, 2012). An overview of variable selection methods for soft sensors is provided in Wang et al. (2015) and Souza et al. (2016). In most cases, the underlying concept of these methods is a correlation analysis between input data and target variable (Sokolov et al., 2017; Bidar et al., 2018).
How can these statistical methods concretely help improve process understanding and contribute to the knowledge base? The number of potential model inputs can often range into the hundreds or thousands (Souza and Araujo, 2011), especially when spectroscopic data are used (Ranzan et al., 2014; Tahir et al., 2019). In these cases, the statistical methods can draw the attention of the process expert to informative variables or interrelations that either confirm a priori knowledge or were previously

undiscovered. Moreover, the degrees of correlation become quantifiable during the variable selection process. Finally, the resulting soft sensor model can represent the mathematical relationship between process variables and the target quantity. This digitalized knowledge—in the form of model coefficients and structure—cannot only be used for monitoring and control purposes but is also part of the manufacturing knowledge base.

In summary, knowledge has two possible roles in the context of soft sensors: first, it can be generated during soft sensor development as a quasi by-product in addition to the actual prediction model (main outcome of soft sensor development); second, existing knowledge can be implemented into the soft sensor to enable accurate and robust prediction in the first place or to improve it.

1.2.2 Uncertain input data

The input data of soft sensors consist of available process measurements, which can be generated via laboratory analyses, sensors, and actuators. The measurements are distinguished, on the one hand, according to their *temporal availability* (information-theoretical distinction) and, on the other hand, according to their *measuring location* in relation to the bioreactor system or product (technological distinction).
According to their temporal availability, these measurements are categorized into *offline* (time-delayed availability) and *online* (without or with only a short time delay; “capable of just-in-time monitoring” (Luttmann et al., 2012)). While offline data are essential for soft sensor calibration, it is only in individual cases (Wu and Luo, 2010; Shardt et al., 2015) used as input to the soft sensor during the prediction step. In the vast majority of cases, only the data available online are used as inputs in the prediction step.
According to their measuring location in relation to the bioreactor system or product, the measurements are usually categorized into *in-line* (sample not removed from the process stream; invasive or noninvasive), *on-line* (sample analyzed in bypass), or *at-line* (sample removed, isolated from, and analyzed in close proximity to the process stream) (FDA, 2004). In this thesis, the notation with hyphen (e.g., off-line) is used for the description of the measuring location, whereas the notation without hyphen (e.g., offline) is used for the description of the temporal availability.
The typical online data for bioprocesses include readings for stirrer, gas flow, flow rates (pH correction agent(s), substrate(s)), temperature, pressure, pH, and pO_2 (Harms et al., 2002). More advanced measurement concepts, whose data can be used as input for soft sensors, include off-gas CO_2 and O_2, turbidity (transmission, transflexion, reflexion), impedance, flow cytometry, high performance liquid chromatography, spectroscopy (ultraviolet–visible, near- or mid-infrared, 2D fluorescence, Raman), ultrasound, in-situ microscope, and biosensors (Luttmann et al., 2012; Biechele et al., 2015; Simon et al., 2015). In many application areas, off-gas analysis and turbidity measurement are already considered standard.

Uncertainty in online measurement can result from the process level, the instrumentation level (sensors and actuators), and the communication level. At the process level, changes in media components (batch variability), external environment (weather or seasons), the production organism (intended or unintended genetic or physiologic changes), or process fouling can lead to uncertain input data. At the instrumentation level, incorrect calibration, abrasion of mechanical components, or unknown cross-sensitivities can lead to faults. At the communication level, loose contacts or malfunctions in the software for data acquisition can occur.
These measurement uncertainties of sensors and actuators, when used as input to soft sensors, lead to uncertain predictions. When developing and applying soft sensors, it is therefore necessary to evaluate the raw input data with respect to outliers (Adikaram et al., 2015). An univariate method for the identification of deviant data points is the moving window implementation of the Hampel identifier (Davies and Gather, 1993), which uses the median $\tilde{x}$ and median absolute deviation from the median, MAD. All data points outside the moving frame of $[\tilde{x} - n\,MAD, \tilde{x} + n\,MAD]$ are classified as outliers. n is a multiplier for tuning the sensitivity of the Hampel identifier.
Multivariate methods for outlier detection are mainly based on distance metrics such as Euclidean, Mahalanobis, or Canberra distance in the principal component space (Shyu et al., 2006). Principal component analysis (PCA) occupies a dominant position among the methods of raw data evaluation because it cannot only be used for multivariate outlier detection (Shyu et al., 2006; Thomassen et al., 2010), but also for correlation analysis (Sokolov et al., 2015), and grouping of datasets (Gunther et al., 2009; Sokolov et al., 2017).
The aforementioned methods represent the basic tools in the detection of outliers in the input data to soft sensors. However, additional challenges arise for bioprocesses with variable process lengths and, in some cases, multiple process phases (e.g., batch and fed-batch phase). The cultivation of *P. pastoris* is a suitable bioprocess to investigate these challenges, as will be shown in the following.

1.3 Monitoring and control of *Pichia pastoris* bioprocesses

1.3.1 Process strategies

P. pastoris is a widely used host for recombinant protein expression in academia and industry. Both inducible and constitutive promoters can be used for protein expression in this host system (Yang and Zhang, 2018). The most commonly used inducible promoter is pAOX1 (alcohol oxidase 1), whose expression is strongly induced by methanol and repressed by other carbon sources such as glucose and glycerol (Liang et al., 2012).
Methanol-induced *P. pastoris*-bioprocesses are typically separated into two main phases. In the first phase, the biomass generation phase, cells grow relatively fast up to the desired biomass concentration. This phase can be carried out in batch or fed-batch mode. Glycerol or another carbon source that represses expression is used as substrate. In the second phase, the product formation phase, methanol is used as inducer for protein expression. This phase is typically carried out as fed-batch with

methanol fed to the bioreactor system. The division of the bioprocess into these two phases makes it possible to separate biomass generation and product formation as far as possible, thus optimally directing the energy and anabolism of cells to the corresponding objective of the respective phase (highest biomass and product yield, respectively). Optionally, a transition phase between these two main phases can be inserted. This transition phase can be carried out either without any carbon source (goal: guaranteeing the complete derepression of the AOX promoter), with a feed of a carbon source other than methanol (goal: higher biomass concentration before induction), or with a mixed feed of methanol and another carbon source (goal: smoother adaption of cells to methanol) (Yang and Zhang, 2018; Liu et al., 2019).
In the product formation phase, the methanol concentration is of crucial importance to a reproducibly high product yield. Too high methanol concentrations are toxic and thus lead to lower cell viability. Cell lysis causes the release of intracellular proteases into the fermentation medium, thus causing increased proteolytic degradation of secreted target proteins (Yamashita et al., 2009). Wu et al. (2011) showed that high methanol concentrations (3.5 g L^{-1}) lead to increased degradation of the extracellular target protein (*Rhizopus chinenisis* prolipase) compared to no degradation at lower methanol concentrations (0.5–1.0 g L^{-1}). On the other hand, depending on the expressed target protein, a minimum methanol concentration is required to induce the AOX promoter. For these reasons, methanol concentration is in the fed-batch phase often controlled to a certain setpoint (Pla et al., 2006; Wu et al., 2011). However, as shown by Pla et al. (2006), the stability of the methanol controller can be disturbed by an increasing biomass concentration during the fed-batch phase.

Which process variables are typically monitored in *P. pastoris* bioprocesses? First of all, the product titer or activity (for enzymes) must be determined. Next, the parameters that have the greatest impact on protein production must be monitored. These include pH, temperature, and dissolved oxygen (Harms et al., 2008). For pAOX1-induced host systems, methanol concentration should also be determined. Biomass concentration is critical at several stages of the process (e.g., for adjusting inoculation volume and feed rates). Due to the importance of biomass concentration in *P. pastoris* bioprocesses, a separate section is devoted to biomass monitoring in the following.

1.3.2 Monitoring of a key variable of *Pichia pastoris* bioprocesses: biomass concentration

Biomass concentration represents the most important process variable in *P. pastoris* bioprocesses, along with product-related quality attributes such as product titer (Harms et al., 2008). The biomass concentration at the end of the batch phase has a direct effect on the expression level of recombinant protein (Wu et al., 2011). Biomass concentration also indicates process progress and is required as a starting point for almost all further mechanistic model calculations (Surribas et al., 2006b). The measurement of this key variable is therefore of utmost importance for process monitoring.

Biomass concentration is typically measured offline as either volumetric cell count (Broger et al., 2011), wet cell weight (Pla et al., 2006), or dry cell weight (Wu et al., 2011). Surrogate measurements such as optical density are also applied offline to determine biomass concentration (Harms et al., 2008). Offline measurements, as described above, are neither suitable for just-in-time monitoring nor as input to closed-loop control systems.

For the online determination of biomass concentration, many process analyzers based on different measurement principles have been developed. These measurement principles include turbidity, impedance, fluorescence, Raman, imaging, and ultrasound (Krause et al., 2011). Many authors have addressed the advantages and disadvantages of these measurement principles for bioprocesses in the past (Harms et al., 2002; Kiviharju et al., 2008; Simon et al., 2015; Grigs et al., 2021). The raw data of these process analyzers can be used together with other process data for the online determination of biomass concentration. However, from a technological point of view, problems can arise with these sensors, especially in industrial environments. The study by Kiviharju et al. (2008) deserves special mention here, as they compared not only various in-line sensors (turbidity and dielectric, infrared, and fluorescence spectroscopy) but also soft sensors regarding their performance and limitations for bioprocess monitoring. They conclude that the influence of the cultivation medium (presence of solid particles, fluorescence characteristics, etc.) and aeration (gas bubbles) on the accuracy and robustness of in-line sensors is considerably greater than for soft sensors. Further, soft sensors are often the most economical alternative, set the case that the process is repeated frequently, and a minimum of process knowledge is available. However, it must also be noted that a soft sensor can, of course, only be as cost-effective as the hardware sensors that are used as input variables. Additional studies compared possible model inputs and different modeling approaches for bioprocesses. Grigs et al. (2021) investigated which input data provides the best biomass predictions for two recombinant *P. pastoris* strains under different process conditions. Evaluated by the relative prediction error, turbidity (8 %) and consumption of pH correction agent (base, 8 %) were slightly superior to the variables oxygen uptake (10 %), permittivity (11 %), and carbon dioxide emission (13 %) as model inputs.

Studies comparing different modeling approaches have been performed for organisms other than *P. pastoris*. However, the transferability of the results to *P. pastoris* must be at least critically questioned, since the results of soft sensor development may depend on the process strategy or organism used. Jenzsch et al. (2006) compared various modeling approaches for the prediction of biomass concentration for *Escherichia coli* processes. In every case, the inputs consisted of carbon dioxide emission rate, oxygen uptake rate, and base consumption. With regard to the prediction error, feed forward ANN and polynomial regression with cumulative inputs outperformed other modeling techniques such as auto-associative ANN, Luedeking–Piret-based model, PCA model, and MLR with cumulative inputs. However, the authors point out that the choice of modeling method is often also a choice between accuracy and robustness. For example, the process knowledge implemented in the Luedeking–Piret-based model increased the robustness to faults as compared to the ANN-based models. Another

comparison of modeling approaches was made by Hocalar et al. (2011) for *Saccharomyces cerevisiae* fermentations with different operating conditions. The different approaches for biomass prediction were based on a kinetic model of overflow metabolism, a metabolic black-box model, an observer, differential evolution, and an ANN. The predictions based on the metabolic black-box model and differential evolution generally resulted in lower prediction errors than the other three approaches. However, the study by Hocalar et al. (2011) shows that the prediction errors are nearly as strongly influenced by the operating conditions as by the modeling approach.

For monitoring *P. pastoris* bioprocesses, the following can be concluded: Soft sensors offer a cost-effective alternative to hardware sensors. The implementation of existing process knowledge can increase the accuracy and robustness of a soft sensor and, if necessary, compensate for the lack of a large process database for model training. The choice of the modeling approach is thus not only dependent on the size of the process database but also on existing process knowledge.

1.4 Motivation and thesis outline

In the previous sections, the DIK hierarchy in bioprocess monitoring was described, the role of process knowledge and uncertain input data in the development of soft sensors were discussed, and the *P. pastoris* bioprocess was characterized. Based on this, the **initial situation and motivation** of this thesis can be summarized as follows:

- The cultivation of *P. pastoris* is suitable as a use case for soft sensor development: On the one hand, this cultivation system is commonly used in biotechnological production and is well described; on the other hand, it shows enough optimization potential regarding the monitoring of one of its key variables: biomass concentration.
- The core function of soft sensors with regard to the DIK hierarchy is the derivation of information (e.g., prediction of target value) from process data. Within mechanistic or hybrid modeling, existing process knowledge can additionally be implemented into the soft sensor to assist the data-to-information compression. However, especially for multiphase *P. pastoris* bioprocesses, there is a need for research in hybrid model-based soft sensors.
- Although useful knowledge can be drawn from the analysis of a soft sensor, the generation of process knowledge *by means of* soft sensor development has so far received comparatively little attention. Process knowledge can be derived during the development of soft sensors as a quasi by-product in addition to the actual prediction model (main outcome of soft sensor development).
- If the input to a soft sensor is faulty, there is a high probability that the output is faulty as well. Therefore, when dealing with uncertain process data, methods must be found to validate model inputs prior to their use in soft sensors.

The objectives of this thesis are, on the one hand, to generate process knowledge during soft sensor development; on the other hand, to develop robust soft sensors through the integration of existing process knowledge and validation of model inputs. The **main objective** is thus formulated as follows:

> The provision of novel concepts to fill the gaps between uncertain process data and knowledge within soft sensor development

In order to achieve this objective, solution approaches for the following **research questions** are to be found:

1. What are the remaining key challenges in the development of soft sensors for bioprocesses?
2. How can a soft sensor model be utilized to generate process knowledge?
3. How can process knowledge be implemented to develop a soft sensor model?
4. How can uncertain model inputs be validated prior to their use in a soft sensor?

The embedding of these research questions into the context of the DIK hierarchy is illustrated in the **graphical abstract of this thesis** (Figure 2). The following chapters are based on this structure. To answer the first research question, a critical review is conducted. The remaining research questions are addressed in the form of research articles.

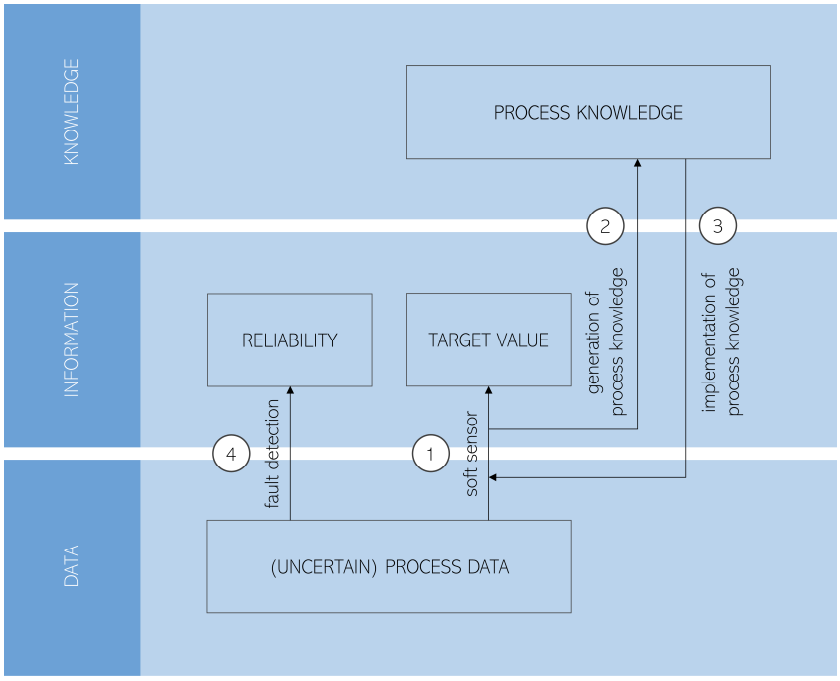


Figure 2: Graphical abstract of this thesis. The core function of soft sensors in the DIK hierarchy is the derivation of information (e.g., prediction of target value) from process data. The solution approaches to research questions 1 to 4 help to fill the gaps between uncertain process data and knowledge within soft sensor development.

2 Methods

2.1 Strain and culture conditions

The *P. pastoris* cultivation was chosen as use case in this work. The inocula were prepared in shake flasks containing the mineral medium FM22 (Stratton et al., 1998) supplemented with 2 mL trace element stock solution (PTM4) per L culture volume. The main culture medium was also FM22 supplemented with 2 mL PTM4 per L culture volume. Glycerol was used as carbon source in the batch phase. In cases where a fed-batch phase was additionally performed (Brunner et al., 2020), methanol supplemented with 12 mL PTM4 per L methanol was fed to the bioreactor.

For the small-scale cultivations, the *P. pastoris* wildtype strain DSMZ 70382 was used. Here, the main culture took place in the microplate reader Synergy™ H4 (BioTek Instruments, Inc., Winooski, VT, USA) with agitated black 96-well plates (200 µL working volume; Greiner Bio-One International GmbH, Kremsmuenster, Germany). Breathe-Easy® sealing membranes (Sigma-Aldrich Corporation, St. Louis, MO, USA) enabled gas exchange between the environment and the culture. Temperature was controlled to 30 °C. Further details can be found in the corresponding publication (Brunner et al., 2016).

For the bioreactor cultivations, a recombinant *P. pastoris* strain based on wildtype strain DSMZ 70382 was used. Here, the main culture took place in the bioreactor system Biostat® Cplus (15 L working volume, 42 L total volume; Sartorius AG, Goettingen, Germany). Temperature, pH, pressure, and dissolved oxygen were controlled to 30 °C, 5, 500 mbar, and 40 %, respectively. Further details can be found in the corresponding publications (Brunner et al., 2019; Brunner et al., 2020).

2.2 Laboratory analyses

Dry cell weight was determined in triplicate by centrifugation of either 200 µL (Brunner et al., 2016) or 2 mL (Brunner et al., 2020) cell suspension in previously weighed centrifuge tubes. Subsequently, the supernatant was discarded, and the cell pellet was dried to a constant weight at 80–90 °C. Bioreactor samples were taken using the BaychroMAT® autosampler (Bayer AG, Leverkusen, Germany) with a minimum sampling interval of 2 h. Further details can be found in the corresponding publications (Brunner et al., 2016; Brunner et al., 2020).

2.3 Sensors and actuators

The sensor and actuator data for the bioreactor cultivations consisted of sensor readings for temperature, pH, pressure, dissolved oxygen, O_2 and CO_2 in the off-gas (BlueInOne Cell sensor; BlueSens gas sensors GmbH, Herten, Germany), methanol (Alcosens; Heinrich Frings GmbH & Co. KG, Rheinbach, Germany), and turbidity (InPro 8100; Mettler-Toledo GmbH, Giessen, Germany) as well as actuator values for stirrer speed, air flow, and feed pump speed (base and methanol). The corresponding publications describe which data were used for modeling in each case (Brunner et al., 2019; Brunner et al., 2020).

2.4 Data management

Primary process control (temperature, pH, pressure, DO) and signal recording for the bioreactor cultivations were realized via the integrated digital control unit (DCU) of the bioreactor system Biostat® Cplus (Sartorius AG). The laboratory (offline) and process (online) data were stored in a central database with a recording interval of 30 s for the online data using the data management system SIMATIC SIPAT (Siemens AG, Munich, Germany). An OPC DA (open platform communications data access) server (Sartorius AG) was used as the real-time communication interface between the DCU, the data management system (SIMATIC SIPAT), and the online modeling software (SIMULINK, version R2019b; The MathWorks, Inc.).

2.5 Algorithm development

Offline data preprocessing, data analysis, and algorithm development were performed in MATLAB (versions 2016a-2019b; The MathWorks, Inc., Natick, MA, USA). SIMULINK (The MathWorks, Inc.) was used for the development and for running the soft sensor for biomass and the fuzzy controller for methanol that was used in Brunner et al. (2020). Further details on the algorithm development can be found in the corresponding publications (Brunner et al., 2016; Brunner et al., 2019; Brunner et al., 2020).

3 Summary of results

In this chapter, the results are summarized and the contributions to the thesis publications are listed. These summaries are followed by copies of the thesis publications. All copies are used with permission of the corresponding journals and authors.

Part 1: What are the remaining key challenges in the development of soft sensors for bioprocesses?

Title: Challenges in the development of soft sensors for bioprocesses: a critical review
Summary: A review study is conducted to identify the key challenges in the development of soft sensors for bioprocesses. The challenges are assigned to either the data, information, or knowledge domain. The key challenges being considered in this study are (1) variable process lengths, (2) multiple process phases, and (3) sensor faults. These challenges often occur synchronously, so that solution approaches are becoming increasingly complex. The corresponding solution approaches originate for their most part from areas other than biotechnology. Therefore, in addition to the practicability, the general applicability to bioprocesses is critically discussed. The main conclusions of this review are, first, that the level of implementable process knowledge is decisive for the choice of methods for handling variable process lengths and multiple process phases. Second, soft sensor predictions are in the presence of uncertain input data (potential sensor faults) only reliable if the input data are validated prior to their use in the soft sensor model. Since there is still a research gap regarding the validation of the input data to soft sensors for bioprocesses, sensor faults remain one of the key challenges in the development of soft sensors in this application area.
Contributions: The doctoral candidate created the structure of the review article, reviewed the literature, and drafted the manuscript. The co-authors critically reviewed and edited the manuscript. All authors have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Part 2: How can a soft sensor model be utilized to generate process knowledge?

Title: Biomass estimation in *Pichia pastoris* cultures by combined single-wavelength fluorescence measurements

Summary: The first soft sensor family shown here uses four single-wavelength fluorescence measurements as model inputs to predict the optical density of a *P. pastoris* culture (surrogate measurement for biomass concentration). The used wavelength pairs correspond to excitation-emission maxima for the biogenic fluorophores tryptophan, NAD(P)H (nicotinamide adenine dinucleotide (phosphate)), and riboflavin. The modeling techniques of MLR, PCR, and PLSR showed comparable prediction accuracy. The PLSR model was further analyzed via variable importance in the projection (VIP) scores to rate the information content of the used wavelength pairs. This analysis resulted in the highest weighting of the wavelength pair corresponding to tryptophan. The main conclusion of this study is that useful knowledge can be drawn from the analysis of a soft sensor. It was confirmed that tryptophan is more coupled to cell growth than NAD(P)H and riboflavin. This knowledge could be used for the development of a low-cost alternative to 2D fluorescence spectroscopy for biomass monitoring.

Contributions: The doctoral candidate reviewed the literature, designed the study, created, analyzed, and interpreted the data, developed the algorithms, and drafted the manuscript. The co-authors critically reviewed and edited the manuscript. All authors have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Part 3: How can process knowledge be implemented to develop a soft sensor model?

Title: Biomass soft sensor for a *Pichia pastoris* fed-batch process based on phase detection and hybrid modeling

Summary: A second soft sensor for the biomass prediction was developed to automatically adapt to the process phases (batch, transition, and fed-batch phase) of a *P. pastoris* bioprocess. The model parameters dynamically adapt according to the current process phase using a multilevel phase detection algorithm. A hybrid approach combining mechanistic (carbon balance) and data-driven modeling (MLR) is used for finally predicting biomass concentration. The main conclusion of this study is that the challenge of multiple process phases in the presence of time-variant behavior can be tackled without exponentiation of model complexity if the soft sensor algorithm works on two distinct but interconnected levels: the phase detection and the prediction step.

Contributions: The doctoral candidate reviewed the literature, designed the study, created, analyzed, and interpreted the data, developed the algorithms, and drafted the manuscript. The co-authors supported in the interpretation of the modeling results and critically reviewed and edited the manuscript. All authors have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Part 4: How can uncertain model inputs be validated prior to their use in a soft sensor?

Title: Online sensor validation in sensor networks for bioprocess monitoring using swarm intelligence

Summary: The third research study reports on the validation of readings from a turbidity sensor used to monitor a *P. pastoris*-batch process. Soft sensors for biomass concentration that are based on turbidity measurements would lose their predictive performance in the case of sensor faults. To detect sensor faults and thus to validate the turbidity sensor, process-time-dependent predictions of the turbidity sensor reading were established. Sensor faults are indicated by the deviation of these predictions from original sensor readings. Swarm intelligence is in this context used to determine the best prediction models according to model fit and overfitting (regularization approach). The main conclusion of this study is that even in a bioprocess with time-variant and non-linear behavior as well as variable process length, one of the remaining key challenges—sensor faults—can be tackled.

Contributions: The doctoral candidate reviewed the literature, designed the study, created, analyzed, and interpreted the data, developed parts of the algorithms, and drafted the manuscript. The co-authors developed parts of the algorithms and critically reviewed and edited the manuscript. All authors have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

3.1 Challenges in the development of soft sensors for bioprocesses: a critical review

frontiers in Bioengineering and Biotechnology

REVIEW
published: 20 August 2021
doi: 10.3389/fbioe.2021.722202

Challenges in the Development of Soft Sensors for Bioprocesses: A Critical Review

Vincent Brunner, Manuel Siegl, Dominik Geier and Thomas Becker*

Chair of Brewing and Beverage Technology, Technical University of Munich, Freising, Germany

Among the greatest challenges in soft sensor development for bioprocesses are variable process lengths, multiple process phases, and erroneous model inputs due to sensor faults. This review article describes these three challenges and critically discusses the corresponding solution approaches from a data scientist's perspective. This main part of the article is preceded by an overview of the status quo in the development and application of soft sensors. The scope of this article is mainly the upstream part of bioprocesses, although the solution approaches are in most cases also applicable to the downstream part. Variable process lengths are accounted for by data synchronization techniques such as indicator variables, curve registration, and dynamic time warping. Multiple process phases are partitioned by trajectory or correlation-based phase detection, enabling phase-adaptive modeling. Sensor faults are detected by symptom signals, pattern recognition, or by changing contributions of the corresponding sensor to a process model. According to the current state of the literature, tolerance to sensor faults remains the greatest challenge in soft sensor development, especially in the presence of variable process lengths and multiple process phases.

Keywords: soft sensor, online prediction, bioprocess, multiphase process, data synchronization, sensor fault, fault tolerance

OPEN ACCESS

Edited by:
Johannes Felix Buyel,
Fraunhofer Society (FHG), Germany

Reviewed by:
Karl Bayer,
University of Natural Resources and Life Sciences Vienna, Austria
Astrid Duerauer,
University of Natural Resources and Life Sciences Vienna, Austria

****Correspondence:***
Dominik Geier
dominik.geier@tum.de

Specialty section:
This article was submitted to Bioprocess Engineering, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *08 June 2021*
Accepted: *03 August 2021*
Published: *20 August 2021*

Citation:
Brunner V, Siegl M, Geier D and Becker T (2021) Challenges in the Development of Soft Sensors for Bioprocesses: A Critical Review. Front. Bioeng. Biotechnol. 9:722202. doi: 10.3389/fbioe.2021.722202

INTRODUCTION

The biologization of the manufacturing industry is leading to more and more processes that were previously based on chemical synthesis being replaced by biotechnological processes (Buyel et al., 2017). At the same time, the digitalization of these processes is leading to more transparent, lower-risk, and more efficient biological manufacturing (Scheper et al., 2021). At the intersection of these two trends—biologization and digitalization—a multitude of new technologies and approaches have emerged in recent decades. These include, in particular, advances in the fields of data science as well as monitoring and control technology for bioprocesses (Steinwandter et al., 2019). With the introduction of the quality by design (QbD) and process analytical technology (PAT) initiatives, this development has received institutional support (FDA, 2004; Rathore and Winkle, 2009).

Despite advances in bioprocess monitoring, many relevant process variables are still determined offline using laboratory analyses. On this basis, a prediction is made about the expected future behavior of the process. However, this procedure is often not sufficient to effectively react to process changes, for example, through closed-loop control. The development of soft sensors is a remedy to this situation.

A soft sensor ("software sensor") is a combination of process data (input) and a model that uses these input data to predict a target quantity (output). It is therefore an indirect measurement. The

input data used for the prediction are typically composed of signals from hardware sensors and actuators. Dependent on the degree of process knowledge that is implemented, the prediction model can be classified as data-driven, knowledge-based, or hybrid.

The application fields of soft sensors can be distinguished by the nature of the target quantity (Kadlec et al., 2009). The largest application field of soft sensors is the online prediction of physical quantities such as, for example, concentrations of biomass, substrate, intermediate, or product. These types of soft sensors are used when online analyzers are not available or economically feasible for process variables of interest. Further, soft sensors can be used within supervisory control applications to monitor the state of the process on a higher level and detect process faults (Liu et al., 2017; Besenhard et al., 2018; Dumarey et al., 2019). Soft sensors for process monitoring and process fault detection use historical process data to derive higher-level, non-physical process quantities such as latent variables (Kourti, 2005) that indicate deviations from the normal process conditions. Finally, soft sensors can be used to detect sensor faults. The soft sensor here is used to predict the reading of a hardware sensor. A deviation of the prediction and the hardware sensor reading indicates a sensor fault (Brunner et al., 2019). The falsified hardware sensor reading can be reconstructed using the soft sensor's prediction.

The development of soft sensors poses several challenges to the data scientist. These challenges can be assigned to either the data, information, or knowledge domain. **Table 1** lists the most important challenges together with corresponding solution approaches. Most of these solution approaches have been reviewed for the process industry, including phase division (Yao and Gao, 2009), adaption mechanisms for soft sensors (Kadlec et al., 2011), JIT learning (Kano and Fujiwara, 2012; Saptoro, 2014), data synchronization (Ündey et al., 2002), process fault detection (Venkatasubramanian et al., 2003a; Venkatasubramanian et al., 2003b; Venkatasubramanian et al., 2003c), dimension reduction (Pani and Mohanta, 2011), variable selection (Cawley and Talbot, 2010; Souza et al., 2016; Heinze et al., 2018), sensor fault detection and fault tolerance (Isermann, 2006; Isermann, 2011; Das et al., 2012), identification of overfitting (Hawkins, 2004), model maintenance (Wise and Roginski, 2015), digitalization of expert knowledge (Birle et al., 2013), and hybrid modeling (Stosch et al., 2014; Solle et al., 2017).

A small number of these reviews address bioprocesses, but in their majority, they play only a tangential role. Several of the above approaches are equally applicable to bioprocesses (e.g., variable selection, dimensional reduction). However, what needs an updated review or has not yet been reviewed at all in the context of bioprocesses are the following three challenges:

- variable process lengths,
- multiple process phases, and
- sensor faults.

Especially for bioprocesses, these challenges often occur in combination, so that solution approaches are becoming increasingly complex: Sensor faults, which impede the reliability of soft sensors, are more difficult to detect or compensate for in processes with variable lengths and dynamic behavior (Brunner et al., 2019); data synchronization (for processes of variable lengths) is more complex for multiphase processes (Doan and Srinivasan, 2008). The focus of this review is thus on the synchronous consideration of these three challenges of soft sensor development. This review aims to critically evaluate the corresponding solution approaches regarding their practicality and applicability to bioprocesses. The following applies here: As simple as possible, as complex as necessary.

This review article is structured as follows. First, an overview of the status quo in the development and online application of soft sensors is provided. Here, the typical steps of soft sensor development and the state of the art in online implementation are described. The following chapter concerns the challenges in soft sensor development for bioprocesses from a data scientist's perspective, namely, variable process lengths, multiple process phases, and sensor faults. The corresponding solution approaches are critically discussed. This chapter is followed by a conclusion that reveals the greatest remaining research gaps in soft sensor development for bioprocesses.

SOFT SENSORS: THE STATUS QUO

Soft sensors have become an important tool within the QbD/PAT framework, as reviewed by Mandenius and Gustavsson (2015), Randek and Mandenius (2018), and Rathore et al. (2021). One reason is that they are often the only means of determining critical process parameters (CPP) or critical quality attributes (CQA) online at all (Capito et al., 2015; Melcher et al., 2015; Sauer et al., 2019; Spann et al., 2019; Walch et al., 2019; Pais et al., 2020; Wasalathanthri et al., 2020a). Making these quantities measurable by means of soft sensors, in turn, allows CPPs or CQAs to be closed-loop controlled (Birle et al., 2015; Matthews et al., 2016; Voss et al., 2017; Brunner et al., 2020; Gomis-Fons et al., 2020). This type of control, also called inferential control, plays an important role in the automation of bioprocesses, since by far not all process quantities to be closed-loop controlled can be measured directly (Rathore et al., 2021).

As mentioned at the beginning, soft sensors are used to indirectly measure a target variable by combining a predictive model with corresponding input data. Process data used as input to soft sensors can compose differently depending on the organism (bacteria, yeast, filamentous fungi, mammalian or insect cells, etc.) used in upstream processing (USP) and the techniques used in downstream processing (DSP). Instrumentation of bioprocesses and thus possible input data for soft sensors have recently been reviewed by several authors (with varying emphases): Simon et al., 2015 (industrial application); Biechele et al., 2015 (USP, disposable technology); Mandenius and Gustavsson, 2015 (price, utility, and relevance of online analyzers for soft sensor development); Claßen et al., 2017 (spectroscopic sensors); Wasalathanthri et al., 2020b (spectroscopic sensors, chromatography, and mass spectrometry); Gargalo et al., 2020 (spectroscopic sensors, biosensors, and free-floating wireless sensors). Therefore, only

TABLE 1 | Overview of the most important challenges and corresponding solution approaches in the development of soft sensors. The challenges are herein broadly assigned to either the data, information, or knowledge domain.

Domain	Challenges	Solution approaches	Details and most important methods
Data	Multiple process phases	Phase detection and division	Algorithms for phase detection can be based on the shape of process trajectories (e.g., sharp peak in specific process variable) or the correlation structure of process variables (e.g., change in loading matrices of latent variable submodels) (Yao and Gao, 2009; Luo et al., 2016)
		Adaption mechanisms	The adaption of the prediction model to multiple process phases can be realized by moving window, recursive adaption, or ensemble-based methods (Kadlec et al., 2011). Just-in-time (JIT) learning is a special case of adaptive modeling, because the local JIT models are built during the online application (Kano and Fujiwara, 2012; Saptoro, 2014)
	Variable process lengths	Data synchronization	Datasets with variable process lengths can be aligned based on: indicator variable techniques, where a measured or computed variable (e.g., maturity index) indicates the progress of the process instead of time; curve registration techniques, where batch trajectories are aligned with respect to process landmarks (Ündey et al., 2002); and dynamic time warping (DTW), where the data patterns are compressed and expanded so that similar features are aligned
	Time-variant and nonlinear behavior	Adaption mechanisms	A prediction model for time-variant data with nonlinear behavior needs to be adaptive rather than static. Adaptive modeling approaches include moving window, recursive adaption, and ensemble-based methods (Kadlec et al., 2011) as well as JIT learning (Kano and Fujiwara, 2012; Saptoro, 2014)
	(Multi)collinearity	Dimension reduction	Latent variable methods (principal component analysis (PCA) or partial least squares (PLS) variants) intrinsically lead to a dimension reduction and thus eliminate (multi)collinearity (Pani and Mohanta, 2011)
		Variable selection	A sound variable selection can reduce (multi)collinearity. Approaches for variable selection include stepwise regression (e.g., backward elimination, forward selection), penalization of model complexity (e.g., based on least absolute shrinkage and selection operator), and through expert knowledge (e.g., a variables' variance is known to be just due to noise or control error and is thus excluded from model inputs) (Cawley and Talbot, 2010; Souza et al., 2016; Heinze et al., 2018)
Information	Process deviations or faults	Enlarge training data pool	The training data pool can be enlarged by the inclusion of datasets of various fault scenarios and the whole design space instead of only the operating space. Cases that are not covered in the training data pool will lead to unreliable extrapolation of the prediction model
		Process fault detection	Methods of process fault detection can be classified as based on quantitative models, qualitative models and search strategies, and on process history (Venkatasubramanian et al., 2003a; Venkatasubramanian et al., 2003b; Venkatasubramanian et al., 2003c)
	Sensor faults	Sensor fault detection	Sensor faults can be detected via various approaches (Das et al., 2012): symptom signal estimation, where the residual between the original and calculated (predicted) sensor reading indicates a sensor fault (Isermann, 2006; Isermann, 2011); multivariate statistical process control (MSPC), where faults are detected by the contribution of each input variable to underlying statistics of an empirical process model (e.g., PCA or PLS variants); and pattern recognition, where supervised or unsupervised learning algorithms are used to differentiate between faulty and non-faulty sensor data
		Fault tolerance	Fault tolerant soft sensors compensate for faults of inputs to the prediction model by a reconstruction of those inputs (Isermann, 2006; Isermann, 2011). Ensemble-based methods can potentially be used to discard or underweight sub-models with faulty model inputs
	Overfitting	Identification of overfitting	Overfitting can be determined during model evaluation via internal cross-validation (e.g., leave-one-out, *k*-fold, stratified, or time-series cross validation) and external (holdout) validation (Hawkins, 2004)
		Controlling model complexity	Model complexity can be controlled and thus overfitting can be reduced by a sound variable selection (see above)
	Deterioration of model performance	Model maintenance	In cases where the performance of the prediction model deteriorates due to unseen events (not yet included in the training data pool, e.g., changes in the production strain or seasonal changes in media components), the training data pool and sometimes also the model structure need to be updated (Wise and Roginski, 2015). In all other cases (similar events already included in the training data pool), adaptive modeling approaches such as recursive adaption and ensemble-based methods (Kadlec et al., 2011) as well as JIT learning (Kano and Fujiwara, 2012; Saptoro, 2014) can be used to maintain the prediction model
Knowledge	Implementation of expert knowledge	Digitalization of expert knowledge	Expert knowledge can be digitalized via fuzzy-logic-based approaches in the form of a rule base (Birle et al., 2013) or via first-principle models (Ohadi et al., 2015; Tahir et al., 2019)
		Hybrid modeling	Data-driven modeling can be combined with knowledge-based approaches to make use of available expert knowledge (Stosch et al., 2014; Solle et al., 2017). Hybrid modeling often results in a combination of the advantages and compensation of disadvantages of the two approaches

a compact selection of the most important process variables and analyzers, respectively, is given in this article. Typical online process data are composed of at least the following readings: flow rates, (differential) pressure (Krippl et al., 2021), temperature, pH, stirrer speed, pO_2, off-gas CO_2/O_2, and conductivity. Often, this standard instrumentation is supplemented by advanced measurement principles, such as turbidity (transmission, transflexion, reflection), impedance, pCO_2, high performance liquid chromatography (Dumarey et al., 2019), flow cytometry, *in-situ* microscopy, ultrasound, biosensors, proton-transfer-reaction mass spectrometry (Berbegal et al., 2020), and, last but not least, various spectroscopic techniques, such as ultraviolet–visible, near- or mid-infrared (Capito et al., 2015; Sauer et al., 2019; Walch et al., 2019; Wasalathanthri et al., 2020a; Cabaneros Lopez et al., 2021), 2D fluorescence (Melcher et al., 2015; Bayer et al., 2020), Raman (Matthews et al., 2016; Voss et al., 2017), and nuclear magnetic resonance (Kern et al., 2019).

As mentioned, the choice of analyzers used for monitoring and control depends on the used production organism. In mammalian bioprocesses (e.g., Chinese hamster ovary cells), for example, the cell concentration is in most cases significantly lower than in microbial bioprocesses (e.g., *Pichia pastoris*, *Saccharomyces cerevisiae*, *Escherichia coli*). Further, metabolite concentrations, which are particularly relevant in mammalian bioprocesses such as ammonium and lactate (Matthews et al., 2016), are relatively low. Due to higher growth rates, the cultivation time is typically shorter for microbial than for mammalian bioprocesses. For the development of soft sensors, special challenges may therefore arise for the respective expression system: First, the accuracy of the reference and online measurements limits the accuracy of the resulting soft sensors, which can take effect when analyte concentrations are low. Second, faster processes require higher measurement frequency according to the Nyquist–Shannon sampling theorem (microbial: ca. 20–120 h^{-1} (Voss et al., 2017; Cabaneros Lopez et al., 2021); mammalian: ca. 0.5–12 h^{-1} (Ohadi et al., 2015; Matthews et al., 2016)). This must be considered when specifying the prediction frequency of the soft sensor. Especially with the complex preprocessing necessary for spectroscopic data (see next section), the computational power can limit the prediction frequency of the soft sensor (Afseth et al., 2006).

Following this description of possible input data to a soft sensor, the subsequent section shows step by step how to develop a soft sensor. Afterwards, the state of the art in online implementation of soft sensors is shown, i.e., how the soft sensor is concretely used for online prediction.

Workflow of Soft Sensor Development

The development of soft sensors has been reviewed by several authors. Systematic approaches to soft sensor development have been presented by Fortuna et al. (2007), Kadlec et al. (2009), and Souza et al. (2016) for the process industry and by Haimi et al. (2013) for wastewater treatment plants. They all show a similar workflow. However, the focus of these review articles is on data-driven modeling approaches, and knowledge-based modeling approaches are for the most part neglected. Khatibisepehr et al. (2013) present a systematic workflow for soft sensor development based on Bayesian methods, which inherently combine knowledge-based and data-driven modeling.

The basic workflow used as a framework in this review article generally assumes a hybrid use of knowledge-based and data-driven approaches (**Figure 1**). The core of soft sensor development is setting up and evaluating the prediction model. Besides these mandatory steps, the workflow is nonrigid: It depends on the individual case (degree of process knowledge, noisiness of inputs, need for model maintenance, etc.) whether all steps are conducted to the full extent.

The first step in soft sensor development is to evaluate the available raw data in terms of outliers and patterns in the datasets. Outlier analysis is important to identify samples or measurements that distinctly stand out from the rest of the data. An initial correlation analysis between model input and output can provide a matrix of correlation coefficients (e.g., Pearson's), which helps to assess relationships among the data. When interpreting the results of correlation analysis, however, one must keep in mind that correlation is not equivalent to causality. The correlation analysis can, in combination with available process knowledge, already be employed to preselect information-bearing model inputs (Melcher et al., 2015; Bidar et al., 2018). These analyses provide the basis for the selection of suitable data preprocessing and modeling methods.

The purpose of data preprocessing is to transform the raw input data into a form that minimizes the effect of noise and outliers while preserving the information content. Methods of data preprocessing include formatting, centering, scaling (e.g., to variance), and—specifically for spectroscopic data—baseline correction and peak alignment (Afseth et al., 2006; Matthews et al., 2016; Voss et al., 2017). Signal processing by smoothing and filtering (e.g., Hampel filter (Pearson et al., 2016)) can help to reduce noise and eliminate outliers. However, it is important to note that all preprocessing measures applied during model establishment must also be executable online.

Process knowledge can be implemented into the soft sensor model. Knowledge-based model parts such as first-principle models (Ohadi et al., 2015; Steinwandter et al., 2017; Pappenreiter et al., 2019; Tahir et al., 2019; Krippl et al., 2021) can be employed to develop a more accurate and robust model. Process knowledge in the form of linguistic expressions can be digitalized using approaches based on fuzzy logic, as reviewed by Birle et al. (2013).

After these preceding steps, the actual correlation—the core of the soft sensor algorithm—is established. This correlation model maps the process data $\boldsymbol{X}$ (input) to the target quantity $\boldsymbol{y}$ (output) using model coefficients $\boldsymbol{b}$. In its simplest, linear form, this model can be formulated as:

$$\boldsymbol{y} = \boldsymbol{bX}. \tag{1}$$

If more than one target quantity is predicted with the same model, the vector $\boldsymbol{y}$ in **Eq. (1)** is replaced by the matrix $\boldsymbol{Y}$.

Taking into account the application fields of soft sensors described above, $\boldsymbol{y}$ can be a physical quantity that can be measured only offline (online prediction), a higher level, non-physical quantity (process monitoring and process fault

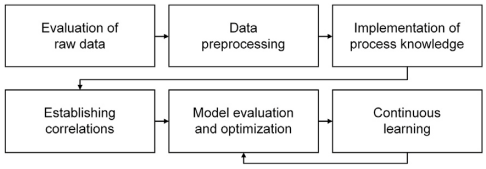


FIGURE 1 | Basic workflow of soft sensor development. A loop exists between model evaluation and optimization and continuous learning; however, revisions of the first four steps will in many cases be necessary to develop a sufficiently accurate and robust soft sensor.

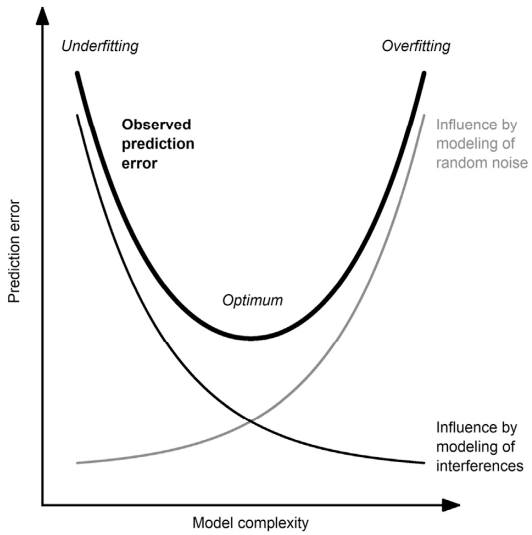


FIGURE 2 | Between the poles of underfitting and overfitting. The observed error (thick, black line) of a predictive model is influenced by the modeling of random noise (undesired; thin, gray line) and interference (desired; thin, black line). The optimal model complexity is a trade-off between these two competing effects and is case-dependent.

detection), or the reading of a hardware sensor (sensor fault detection). Various modeling techniques have so far been applied for soft sensor development, including variants of multiple linear regression (MLR; Jenzsch et al., 2006), partial least squares regression (PLSR; Sokolov et al., 2015; Voss et al., 2017; Zheng and Song, 2018; Walch et al., 2019; Cabaneros Lopez et al., 2021), principal component regression (PCR; Zhu et al., 2018), artificial neural networks (ANN; Paquet-Durand et al., 2017; Zhang et al., 2020), and support vector regression (SVR; Voss et al., 2017; Meng et al., 2019). The choice of the right modeling method depends on the degree of (multi)collinearity, nonlinearity, and the availability of process knowledge.

The model is typically trained, i.e., $\boldsymbol{b}$ is determined, using historical data $\boldsymbol{X_{hist}}$ and $\boldsymbol{y_{hist}}$ (exception: just-in-time (JIT) learning), so that $\boldsymbol{y_{hist}} = \boldsymbol{bX_{hist}}$. Subsequently, the resulting model needs to be evaluated in terms of goodness-of-fit, predictivity, and robustness (OECD, 2014). Here, the received model is used to predict the target quantity $\hat{\boldsymbol{y}}_{\boldsymbol{hist}}$, so that $\hat{\boldsymbol{y}}_{\boldsymbol{hist}} = \boldsymbol{bX_{hist}}$. After training and model evaluation, $\boldsymbol{b}$ can be used together with online process data $\boldsymbol{X_{on}}$ to predict the target quantity $\hat{\boldsymbol{y}}_{\boldsymbol{on}}$, so that $\hat{\boldsymbol{y}}_{\boldsymbol{on}} = \boldsymbol{bX_{on}}$.

For the robustness of the developed model, it is crucial that the model has neither too many nor too few model inputs nor too high nor too low model complexity, respectively (**Figure 2**). Methods to determine the optimal model complexity have been reviewed by several authors (Cawley and Talbot, 2010; Souza et al., 2016; Heinze et al., 2018).

Even with a robust and sufficiently accurate soft sensor, model quality or prediction performance, respectively, usually deteriorates if the process characteristics change (Kano and Fujiwara, 2012). Therefore, the maintenance or recalibration of soft sensors—just as for hardware sensors—is necessary in practice to preserve the quality of their prediction performance. In this context, model maintenance refers to the (automatic) adaptation of models in the event of changing system conditions. For the prediction models of a soft sensor, this means that the model parameters and, if necessary, the entire model structure (e.g., number and type of input variables) must be adapted over time.

Which programming environment or software solution is used to develop soft sensors in practice? Soft sensor development in the academic environment typically takes place in a programming language of choice such as Matlab (The MathWorks Inc.), Python, or R. The corresponding programming environments provide steadily growing libraries of functions or toolboxes for signal processing, data preprocessing, and model calibration and validation. Especially in the industrial environment, software specially developed for chemometrics is often used for soft sensor development (e.g., SIMCA by Sartorius AG; Unscrambler by Aspen Technology Inc.). Here, the full flexibility of development via program code is exchanged for a relatively straightforward and guided development process. Also, many vendors of online analyzers offer software modules for soft sensor development. In particular, vendors of spectroscopic sensors should be mentioned here (e.g., OPUS suite by Bruker Corp., iC suite by Mettler Toledo Inc., GRAMS suite by Thermo Fisher Scientific Inc.), but also vendors of other multivariate sensors (e.g., BlueVis by BlueSens gas sensors GmbH) offer corresponding software modules. Some software tools (chemometric and analyzer software) also offer the option to embed scripts generated via the above-mentioned programming languages into the soft sensor algorithm. This allows adding customized functions for signal processing and data preprocessing as well as developing prediction models that might not be included in the commercial software tool. Finally, soft sensors can also be developed on cloud-based platforms (e.g., MindSphere by Siemens AG, Predix by General Electric Co.) to have access to a wide variety of data processing and modeling tools and to be able to share the developed soft sensors across plant or company boundaries (Chen et al., 2020; Kabugo et al., 2020).

Online Implementation of Soft Sensors

How is a soft sensor used in practice for online prediction? In theory, the soft sensor is merely a combination of input data and a prediction model (see definition above). In practice, however, several additional aspects must be considered if a soft sensor is to be used for online prediction, i.e., implemented online.

First of all, online implementation of soft sensors requires at least communication between field (sensors and actuators) and control level (programmable logic controller and/or process control system) and in most cases also supervisory level (supervisory control and data acquisition, SCADA, and/or other data management system). The data used as inputs to the soft sensor can originate from various sources (Steinwandter et al., 2019). Therefore, a standardized communication between these sources and the software instance in which the soft sensor is implemented is essential. While a variety of standard communication protocols exist for communication between field and control level (4–20 mA, Modbus, Profibus, etc.), it is communication via OPC UA (open platform communications unified architecture) that seems to become the predominant standard for communication in the control and supervisory level (Chen et al., 2020; Biermann et al., 2021). Recent efforts even aim at field-level communication using OPA UA (Veichtlbauer et al., 2017). OPC UA, unlike its predecessors of OPC classic (data access, alarms and events, historical data access), allows hardware- and platform-independent communication.

Once the communication and thus the data flow between field, control, and supervisory level has been established, the question arises on which level of the automation pyramid the soft sensor is implemented. Technically, it is possible to implement soft sensors directly in the control level. However, the implementation of scripts directly in the control system is intended for end users only in exceptional cases and the proprietary language must be used (Nair et al., 2020). Systems above the control level, on the other hand, commonly offer the possibility to implement soft sensors directly or indirectly. In the direct variant, soft sensors are implemented in the SCADA (e.g., MFCS by Sartorius AG, Eve by Infors AG, BioXpert by Applikon Biotechnology BV) or other data management system (e.g., SIMATIC SIPAT by Siemens AG, synTQ by Optimal Industrial Technologies Ltd., xPAT by ABB Ltd., Lucullus PIMS by Securecell AG, LabVIEW by National Instruments Corp.). Here, preprocessing steps and model calculations can be implemented directly to a certain extent. More importantly, these software tools often offer the possibility to communicate with external chemometric or analyzer software (Matthews et al., 2016; Voss et al., 2017; Dumarey et al., 2019) or to integrate customized scripts that are executed online (Besenhard et al., 2018). In this indirect variant, soft sensors are implemented in real-time capable chemometric (e.g., SIMCA-online by Sartorius AG (Voss et al., 2017), Process Pulse by Aspen Technology Inc.) or analyzer software (e.g., CMET by Bruker Corp. (Wasalathanthri et al., 2020a), iC Quant by Mettler Toledo Inc. (Wu et al., 2015)) that communicates with the SCADA or data management system. Here, communication often already takes place via OPC UA (Kern et al., 2019). In this indirect implementation, the chemometric or analyzer software preferentially communicates information (e.g., the predicted value) rather than data back to the SCADA or data management system (Luttmann et al., 2012).

The PAT software products mentioned in this section are only a selection and should not be seen as a recommendation. For a more comprehensive overview of PAT software, the reader is referred to Chew and Sharratt (2010). The authors also list whether the respective software is compliant with regulatory requirements for electronic records and signatures according to 21 CFR Part 11 (FDA, 2003).

When soft sensors are implemented in an industrial environment, they must first undergo an intensive functional and risk assessment (qualification). A step-by-step guidance for structured development and implementation has been proposed by Randek and Mandenius (2018). This guidance considers the regulatory validation requirements for software including recommended protocols for installation, operational, and performance qualification. The validation of software, especially in the pharmaceutical environment, commonly follows guidelines such as GAMP 5 (ISPE, 2008), 21 CFR Part 11 (FDA, 2003), or EU GMP Annex 11 (EC, 2010).

CHALLENGES IN SOFT SENSOR DEVELOPMENT FOR BIOPROCESSES

This chapter concerns the challenges in soft sensor development for bioprocesses from a data scientist's perspective, namely, variable process lengths, multiple process phases, and sensor faults. For each of these three challenges, the problem statement is initially outlined. Subsequently, the solution approaches are critically discussed, linking them to the other two challenges, wherever possible. Each solution approach is summarized at the end particularly regarding its practicality and applicability to bioprocesses.

Variable Process Lengths

Problem Statement

From a process engineering perspective, the end of a bioprocess is defined either by the expiration of a certain process time or by the occurrence of a certain process event. Such termination events can be, for example, the reaching of a target value for the biomass or product concentration or a specific pattern in the process data (e.g., a CO_2 peak indicating the consumption of a carbon source).

In the case of an event-driven process end, the process length can vary from batch to batch due to multiple sources of variance. Besides the typical variance of biological reactions, variance can be introduced by raw materials (e.g., media or feed), by preceding processing units (e.g., preculture), or by deviations in the current process itself.

The variable length of process runs can lead to the following problems. First of all, it can distort the equal weighting of the individual datasets during model development and evaluation if the reference data $\boldsymbol{y}_{hist}$ are generated at a constant frequency: More reference data points for longer processes lead to an overweighting of longer processes compared to shorter ones. Secondly, in case a dynamic soft sensor model incorporates

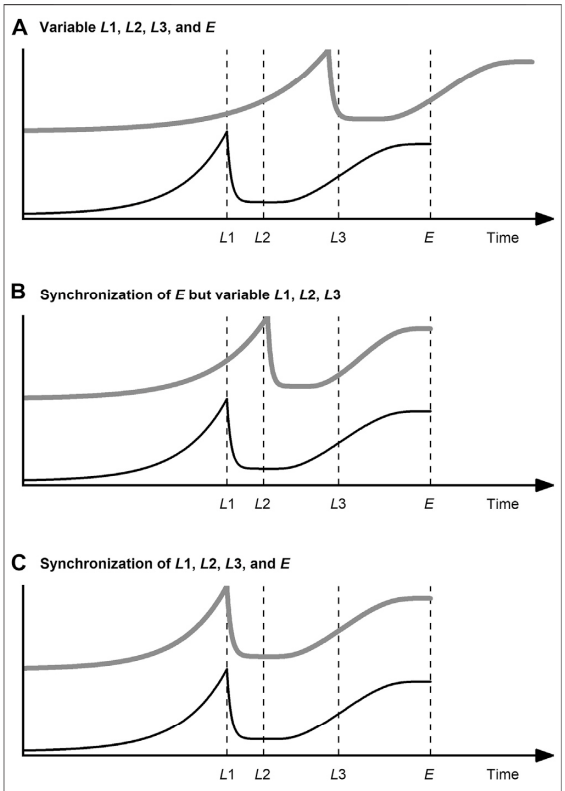


FIGURE 3 | Data synchronization of a query curve (gray, thick line) related to a reference curve (black, thin line) of a fictional process variable with landmarks *L*1, *L*2, and *L*3 and process end *E* (dashed line): **(A)** Initial situation with variable *E* and variable *L*1, *L*2, and *L*3; **(B)** Synchronization of *E* but no coincidence of *L*1, *L*2, and *L*3; **(C)** Synchronization of *E*, *L*1, *L*2, and *L*3. The landmarks *L*1, *L*2, and *L*3 represent typical curve features, namely a local maximum, a local minimum, and a trend reversal, respectively. The trigger event for this fictional process to end (*E*) is the reaching of a plateau after *L*1, *L*2, and *L*3 have been reached.

time as an input variable to compensate for time-variant behavior, variable process lengths lead to another problem: Model performance may deteriorate to the extent that the rate of process progress deviates (i.e., the process is too fast or too slow) from the historical data that were used to train the model. Thirdly, in multiphase processes, the sources of process variance described above can lead to deviations in the time of occurrence of process events (Ündey et al., 2002). If this issue is not accounted for, the adaptation of a soft sensor to process phases could be impeded.

Various methods of data synchronization have been developed to address the challenge of variable process lengths. Data synchronization has two goals, as illustrated in **Figure 3** for a fictional process variable: on the one hand, to bring all process datasets to the same lengths (**Figure 3B**); on the other hand, to ensure that the relevant process events (landmarks) coincide (**Figure 3C**).

The three techniques used most commonly for data synchronization are discussed in the following: indicator variable, curve registration, and dynamic time warping (DTW). The goal of all these methods is to find a warping function h that replaces the time t on the abscissa and thus to obtain synchronized process data $\boldsymbol{X}_{sync}$ (Ramsay and Silverman, 2005):

$$\boldsymbol{X}_{sync} = \boldsymbol{X}[h(t)]. \tag{2}$$

As part of soft sensors that are adaptable to variable process lengths, the synchronization algorithm needs to be executable both offline during model development (for $\boldsymbol{X}_{hist}$ and $\boldsymbol{y}_{hist}$) as well as during the online application (for $\boldsymbol{X}_{on}$).

As with all, the choice of the data synchronization method is highly dependent on the process being monitored (Rato et al., 2016; Rato et al., 2018). It should also be noted that, regardless of the method used for data synchronization, all subsequent levels of the monitoring algorithm (soft sensor prediction, fault detection, etc.) depend for better or worse on the robustness and accuracy of the synchronization method used.

Indicator Variable Techniques

In this method, the time scale is replaced by an alternative scale, the indicator variable. The indicator variable can be either a real (physical) process variable or an estimated process progress, often referred to as maturity index or percent completion. Process variables that are used as termination criteria for the process or as trigger variables for an automation system are particularly suitable as indicator variables (Ündey et al., 2003; García-Muñoz et al., 2011). Examples of process variables suitable as indicator variable are decrease of substrate concentration (Ündey et al., 2002), cumulative feed volume (Ündey et al., 2003), bioreactor volume, and biomass concentration (Rato et al., 2016). Regardless whether a real process variable or a maturity index is used, the indicator variable should ideally progress strictly monotonically, continuously, and smoothly and have the same start and end value (e.g., 0 and 100 % maturity) for all process runs (Nomikos and MacGregor, 1995; Ündey et al., 2002; Ündey et al., 2003).

When developing a prediction model for the maturity index, the percentage of process progress is calculated for the training data, e.g., by a simple linear transformation. The model requires monotonically progressing variables that correlate with process progress. Examples of the use of a maturity index for data synchronization in bioprocesses can be found in Krause et al. (2015) and Brunner et al. (2019). Both studies demonstrate how a maturity index based on a PLS model can be used to determine process progress online and thus enable adaption to the time-variant behavior of biological batch processes. Only through information about the process maturity was it possible to detect sensor faults in the respective bioprocesses.

Ündey et al. (2003) addressed the challenge of variable process length for a multiphase process, namely a simulated fed-batch penicillin fermentation with two phases (batch and fed-batch phase). They proposed using separate indicator variables for each process phase to compensate for the variable lengths of the

phases. As a result, the authors were able to construct tighter control limits for an MSPC model, which in turn enabled faster fault detection. A similar approach was presented by García-Muñoz et al. (2003) for an industrial drying process with three process stages. It was shown that incorporating warping information—i.e., "the information that comes out of an alignment" (García-Muñoz et al., 2003)—resulting from the stage-by-stage alignment can improve a quality prediction model.

In summary, indicator variables are suited for data synchronization of bioprocess data under the condition that there is a minimum understanding of the temporal behavior of the process variables. If this knowledge is available and especially if the process variable used as the indicator variable is used as termination criterion for the process, there is no more robust and simple method than this. Problems with the prediction of the maturity index can occur if the input variables of the model change fast in certain process phases and slowly in others. This is not uncommon, especially in USP (lag vs. exponential phase). Even if this is considered by using a non-linear model, the resolution of the input variables restricts the relative accuracy in "slow" process phases. This resolution is determined by the sensors and actuators used. In cases where it is difficult or impossible to find or calculate an indicator variable that comes close to the above-mentioned requirements (strictly monotonically progressing, etc.), curve registration techniques or DTW should be considered. Finally, it must be stressed that indicator variable techniques are per se designed to be independent of any landmarks. These structural features, which are especially helpful for multiphase processes, are ignored during data synchronization and thus cannot be exploited. Data synchronization with indicator variable techniques is therefore limited to the scenario shown in **Figure 3B**.

Curve Registration Techniques

Within functional data analysis, curve registration is referred to as the process of aligning one function curve to another (Ramsay and Silverman, 2005). In this sense, the term curve registration does not differ from the term data synchronization, only that it refers specifically to functional data. The process data are seen as observations of an underlying continuous function (Ündey et al., 2002). The curves are aligned with respect to their structural features, referred to as landmarks. These landmarks can be certain levels, extrema (minima, maxima), or trend reversals (see *L*1, *L*2, *L*3, and *E* in **Figure 3**). The relevant landmarks are identified using process knowledge and/or numerical computations, such as first and second derivative, respectively, and zero crossing (Ündey et al., 2002; Ramsay and Silverman, 2005). After matching the landmarks between reference and query, the sections between the landmarks are warped, which in the simplest case means that they are resampled linearly.

Williams et al. (2001) and Ündey et al. (2002) used curve registration to align the process data of a simulated fed-batch penicillin fermentation. For the alignment of multivariate data, the authors suggest first aligning all process data with respect to the landmarks of the most important variable (determined, e.g., via process knowledge). In the second step, a principal component analysis (PCA) is carried out and the process variables are aligned with respect to the landmarks of the first principal component. The second step is repeated until the landmarks converge. In these studies, it was shown that curve registration provides relatively smooth variable trajectories after the alignment compared to DTW; in this way, fewer false alarms occurred with MSPC-based fault detection. In one other of the few examples from the bioprocess field, Andersen and Runger (2012) used landmarks of a pharmaceutical batch fermentation process for data synchronization. The significant landmarks were automatically identified as the zero crossings of a continuous Gaussian wavelet transformation (Bigot, 2006). Afterwards, the resulting curve segments were warped linearly and piecewise for each segment.

In summary, curve registration techniques allow not only the alignment of variable lengths—as with an indicator variable—but also the alignment of curve features. Scenario C in **Figure 3** can therefore be achieved. Since the features of many process variables occur simultaneously at phase transitions, curve registration techniques are particularly suitable for multiphase processes (Ündey and Çınar, 2002). However, applications of curve registration for bioprocesses are rare. The existence of this niche in the field of bioprocesses can at most be explained by the circumstance that the indicator variable technique is more intuitive and comparatively easy to implement and DTW can be used with less fine tuning.

Dynamic Time Warping

DTW, initially developed for speech recognition (Sakoe and Chiba, 1978), was proposed for the synchronization of process data by Kassidas et al. (1998). Since then, it has become one of the most widely used methods for this purpose. Reasons for this are that not only variable process lengths but also landmarks can be aligned using DTW. Scenario C in **Figure 3** can therefore be achieved, just as with curve registration. DTW expands, contracts, or translates the time axis of the datasets in such a way that the shape of the variable trajectory is largely preserved, landmarks coincide in time and all datasets have a uniform number of measuring points. The basic sequence of DTW algorithms is as follows: First, the distance matrix (e.g., Euclidean) between the instants of the reference and the query time series is calculated. Then the warping path is searched for that minimizes the sum of distances and at the same time considers several boundary conditions (local, global, endpoint). Using this warping path, the query time series is aligned to the reference time series by expanding, contracting, and translating.

Since its introduction for data synchronization, the original DTW algorithm has been varied in several ways to address issues such as singularities. Singularity in this context refers to the mapping of a single point of the reference time series to multiple points of the query time series or vice versa. Derivative DTW (DDTW) uses local derivatives of the time series instead of raw data and was proposed for overcoming singularities (Keogh and Pazzani, 2001). DDTW compared to DTW tends to align more based on shape rather than magnitude (Spooner et al., 2018). Since numerical derivation often leads to an amplification of noise, a Savitzky-Golay filtering step can be implemented in the

DDTW algorithm to make the alignment more robust (Zhang et al., 2013).

For process data that show sections with many successive landmarks (feature-rich) and then again sections with few landmarks (feature-poor), a fixed warping resolution is often not sufficient. Therefore, a dynamic warping resolution was proposed by Gins et al. (2012). This is achieved by a combination of correlation optimized warping (COW; Nielsen et al., 1998; Fransson and Folestad, 2006) for feature-poor and DDTW for feature-rich sections (hybrid DDTW).

The difficulty with the online application of DTW is that a *partially complete* dataset (query) needs to be aligned with a *complete* dataset (reference). This issue was first addressed by Kassidas et al. (1998) and later further elaborated by González-Martínez et al. (2011). In both studies, the endpoint constraint, i.e., that the endpoint of the query must equal the endpoint of the reference, was omitted. This means, however, that the alignment has to be calculated at each sampling point and each time the recent history of the trajectory has to be considered. For this reason, a computationally efficient way of finding the optimal warping path within a moving window was proposed by González-Martínez et al. (2011), referred to as relaxed-greedy time warping (RGTW). Another online application of DTW was presented by Srinivasan and Qian (2007). They used dynamic locus analysis (Srinivasan and Qian, 2006) to identify the best matching signal segment from a reference library by making use of singular points (landmarks) and thus to determine the state of the process. For the actual online warping, a greedy version of the DTW algorithm, referred to as extrapolative time warping (XTW; Srinivasan and Qian, 2005), was used.

González-Martínez et al. (2014) extended the concept of DTW (offline application) and RGTW (online application) to the problem of multiple asynchronisms for a simulated *S. cerevisiae* fermentation. Multiple asynchronism in this context refers to a combination of at least two of the following asynchronism scenarios: variable process length; no coincidence or overlapping of key process events; initial delay or premature termination of a process. The authors proposed a two-step approach in which the asynchronism pattern is firstly detected based on the warping information and secondly batch synchronization is performed based on the detected pattern.

In the standard DTW procedure (univariate DTW), a single representative process variable is used as a reference to align all other process variables. In certain cases, however, univariate DTW can lead to misleading results; this includes, for example, a delayed measurement in a bypass (on-line) or the bioreactor periphery (at-line) compared to the remaining measurements in the bioreactor (in-line). In these cases, multivariate DTW (MDTW) should be considered. Two fundamental variants are distinguished in MDTW (Shokoohi-Yekta et al., 2015): Either DTW is performed separately for each of the process variables j, resulting in j potentially different alignments ("independent" MDTW); or the warping path is determined via a multidimensional p-norm as cost function, whereby multiple process variables are included in the calculation of the distance ("dependent" MDTW). For a review of MDTW, the reader is referred to Moser and Schramm (2019).

In summary, DTW and its variants have—at least for simulated data—proven to be well suited for synchronizing bioprocess data, both offline and online. No process knowledge is necessary to develop this preprocessing method. When dealing with multiple process phases, DTW can be used in two different ways: first, it can be used to detect process phases (Gollmer and Posten, 1996); second, it can be used to align data within a process phase (Doan and Srinivasan, 2008; Spooner et al., 2018). The use of DTW for these purposes is further described in the following section. Finally, it should be noted that the warping information can be used for the classification of deviations from normal operating conditions, such as sensor faults (González-Martínez et al., 2013). However, in order to identify the deviating sensor, each fault scenario of interest must explicitly be included in the training data pool.

Multiple Process Phases

Problem Statement

From a monitoring perspective, industrial processes can take place either in multiple processing units (multistage) or in a single one. A process with a single processing unit (e.g., USP in a bioreactor) can have multiple operational regimes, such as a batch and fed-batch phase, and is referred to as multiphase process. Multiphase processes are often treated analogously to multistage processes (Yao and Gao, 2009), i.e., different process phases are treated as if they took place in separate processing units.

The necessity of considering multiple process phases when developing a soft sensor is obvious: The relationships within the input data $\boldsymbol{X}$ (multicollinearity) and between the input data and the target variable $\boldsymbol{y}$ (correlation) can vary substantially in the individual phases. The challenges discussed in this section refer to changes in the relationships from $\boldsymbol{X}$ to $\boldsymbol{y}$ that are related to the process strategy. These include, for example, an induced change in media composition due to feeding, the start or end of a starvation phase, long-term changes between oxygen-limited and non-limited process conditions, or changes in the temperature setpoint. Changes in the relationships of $\boldsymbol{X}$ to $\boldsymbol{y}$ that are associated with time-variant and nonlinear behavior are not within the scope of this review article, although the respective adaption mechanisms partly overlap. These adaption mechanisms have been excellently reviewed by Kadlec et al. (2011).

Only with much greater development effort or available process knowledge will a global process model attain the same accuracy and robustness as several submodels for each process phase. Graphically expressed, the required model complexity (cf. **Figure 2**) of a global model is allocated to several less complex local models. This in turn can make it easier to optimize the model (Jin et al., 2015), for example, in terms of avoidance of overfitting.

The main difference between datasets of multistage and multiphase processes is that in multiphase processes the individual phase segments must first be identified and often cannot be precisely separated. The actual modeling step is therefore often preceded by a phase detection and division

step. The detection and division is based either on trajectories of phase-sensitive process variables or on the changing correlation structure among the process variables (Luo et al., 2016).

Trajectory-Based Phase Detection and Division

The sequence of the most biotechnological processes is not given by nature, but by process experts. Therefore, if knowledge about the process sequence is available, it is reasonable to use it for phase detection and division. The definition of landmarks by process experts leads to a solution that is both robust and comprehensible. An example of this can be found in Spooner et al. (2018), who, in contrast to their previous study (Spooner et al., 2017), first divided a bacterial fermentation process into two phases and then aligned the process data of these phases separately via DTW and DDTW, respectively. The pH and pH correction agent (flow and cumulative amount) signals were used to distinguish the phases. Brunner et al. (2020) used the off-gas CO_2 signal to detect the consumption of the carbon source and thus the end of the batch phase in a *P. pastoris* fed-batch bioprocess. To make the detection of this landmark (CO_2 peak) more robust, a threshold for the cumulative amount of pH correction agent was additionally implemented in the phase detection algorithm.

The role of DTW for data synchronization has already been described in the previous section. As a means of detecting process phases, it was first proposed by Gollmer and Posten (1996) for time-varying fed-batch bioprocesses (*E. coli* and *S. cerevisiae*). With the use of historical time trajectories of CO_2 and O_2 together with available process knowledge, six different process phases were classified. This reference (prototype) was used in the online application by the DTW algorithm to assign unknown process data to this pattern and thus detect the previously defined process phases. Doan and Srinivasan (2008) proposed a variant of DTW augmented by singular points (landmarks) for the combined detection and synchronization of process phases. They used substrate feed and pH as key variables for the detection of process phases of a simulated fed-batch penicillin fermentation. Phase changes were considered equivalent to the occurrence of singular points and were identified using the methods described in the previous section (Srinivasan and Qian, 2005; Srinivasan and Qian, 2007). DTW (offline) and XTW (online), respectively, were then used for data synchronization within the phase segments.

Especially in USP, phase transitions must be considered, as biological systems involve living cells, which do not react instantaneously to environmental changes. Luo et al. (2016) proposed a framework for adapting process models to a sequence of multiple process phases while explicitly considering phase transitions in a simulated fed-batch penicillin fermentation. They used fuzzy *c*-means (FCM) clustering for phase detection and division to account for the gradual transition from one steady phase to another. The FCM clustering algorithm was constrained by the temporal sequence of the dataset. Phase-based multiway PLS models were used for prediction in the steady phases, and JIT-PLS models were used during the transition phases.

In summary, trajectory-based algorithms are suitable for phase detection and division in cases where a minimum of process knowledge is available. This knowledge is necessary to select the phase-sensitive process variables. Provided that suitable phase-sensitive process variables can be identified, this approach is more comprehensible than the correlation-based approach. This is especially due to the fact that the identified phases usually correspond to operational phases (Yao and Gao, 2009). Finally, it must be emphasized that DTW is suitable not only for data synchronization in the case of variable process lengths, as described above, but also for the detection of multiple process phases.

Correlation-Based Phase Detection and Division

A difficulty with the methods mentioned so far is to find variables that are measurable and sensitive to the individual phases (Luo et al., 2016) and whose trajectories are reproducible and as noise-free as possible. In the following, methods are presented that accomplish phase detection and division without the need for process knowledge. These methods are based on changes in the correlation structure among process variables.

Camacho et al. (2008) proposed an algorithm based on latent variable models (PCA or PLS) for the detection and division of process phases for a *S. cerevisiae* and a wastewater treatment process. The whole process dataset is iteratively divided into incrementally smaller phases. At each iteration step, the separation point that leads to the maximum improvement of the explained variance of the PCA or PLS submodel, respectively, compared to the undivided dataset is identified (Camacho and Picó, 2006).

Another method to make use of changes in the correlation structure is to first determine the loading matrices of PCA or PLS submodels following a moving window approach and then to find groups in which the underlying variable correlation remains similar (Lu et al., 2004; Lu and Gao, 2005). These groups can, for example, be determined by *k*-means clustering (Lu et al., 2004).

However, if different operation modes with variable phase and process lengths are to be considered, the classical moving window approach leads to misleading results. The reason for this is that in the online application it is not clear whether the current moving window coincides with that of the reference (historical data). Therefore, the described method for phase detection and division (Lu et al., 2004) was extended by the ability to identify the current mode of operation (resulting in variable lengths) online. Zhang et al. (2018) generated a series of moving windows within a constrained searching range around the current sample. They then used the *k*-nearest neighbor rule to identify the most similar time slices. The time slices found in this way are then used as described above (loading matrices of submodels, *k*-means clustering) to enable online phase detection despite variable phase and process lengths.

Finally, Gaussian mixture models (GMM) have proven to be suitable for phase detection and division for a simulated fed-batch penicillin fermentation (Yu and Qin, 2009; Yu et al., 2013). Here, each phase is represented by a Gaussian component with distinct mean and covariance. The posterior probability is used to group the process data into separate process phases. This concept was later adopted for a real industrial bioprocess (Mei et al., 2017).

In summary, for correlation-based phase division, the reduced effort in implementing process knowledge is compensated with an increased effort in modeling. Depending on the modeling method, however, an entirely different division of the process phases may result and fine-tuning of the latent variable models is necessary (Luo et al., 2016). Because the correlation structure is of multivariate nature, the interpretability of the results of the phase division is limited in contrast to most trajectory-based methods.

Sensor Faults

Problem Statement

Sensor faults are defined as deviations of the observed sensor reading from the true value (Balaban et al., 2009; Sharma et al., 2010). They are distinguished according to the type of occurrence as abrupt (stepwise) or incipient (driftwise) faults (Isermann, 2006) and according to the shape of the deviation as bias, precision degradation, and complete failure.

During the training phase of soft sensor development, sensor faults can severely affect the resulting goodness-of-fit and predictivity. If sensor faults are present in the training data $\boldsymbol{X_{hist}}$ and $\boldsymbol{y_{hist}}$, these deviations may be reflected in the model coefficient $\boldsymbol{b}$ and the prediction $\hat{\boldsymbol{y}}_{\boldsymbol{hist}}$. In this case, evaluation criteria for goodness-of-fit and predictivity (e.g., R^2, root mean squared error) are affected. During the online application of the soft sensor, i.e., the prediction phase, sensor faults in $\boldsymbol{X_{on}}$ may directly affect the prediction of the target quantity $\hat{\boldsymbol{y}}_{\boldsymbol{on}}$.

The validation of sensor readings prior to their use for quality control, e.g., via soft sensors, is therefore of crucial importance, as outlined by Feital and Pinto (2015). A sensor reading is valid if there are no sensor faults or unconsidered influences on the measurement, which can occur due to cross-sensitivity to matrix compounds (matrix effects). Deviations between the observed sensor reading and the true value thus need to be detected, and a decision logic needs to classify the observed sensor reading as reliable (valid) or faulty (invalid). Valid sensor readings can be used for quality control by means of soft sensors, while invalid ones can lead to misleading results.

The fault tolerance of soft sensors, or, in other words, a reliable soft sensor prediction in the presence of sensor faults represents one of the remaining core challenges in the development of soft sensors. The reason for this is that the detection and subsequent compensation of sensor faults alone are difficult to realize, but they become even more complex when the conditions described above (variable process lengths and multiple process phases) occur simultaneously.

This section first discusses various methods of detecting sensor faults. Afterwards, the approaches for tolerance of soft sensors towards sensor faults are discussed.

Sensor Fault Detection

When a sensor i that is used to monitor a bioprocess gives faulty readings, its reliability r_i decreases. The aim of sensor fault detection is to detect these faulty readings and thus indirectly determine r_i. **Figure 4** shows four fundamental approaches to sensor fault detection.

Hardware redundancy uses multiple identical sensors to derive the occurrence and size of sensor faults in case of a significant discrepancy among these sensors (as, for example, in airplanes). Voter structures can be implemented into the fault detection algorithm to allow a "democratic" decision on which of the individual sensor values is faulty. If, for example, two of three sensors give a similar reading and the third reading deviates significantly, the third sensor is considered to be faulty. For hardware redundancy, the spatial distribution of the sensors must be considered, and the costs of the sensors limit this approach (Stork and Kowalski, 1999). Hardware redundancy can also be used to determine the type of fault as bias, gain, precision degradation, complete failure, and noise (Kullaa, 2013).

The other three approaches are based on analytical redundancy and are described in the following.

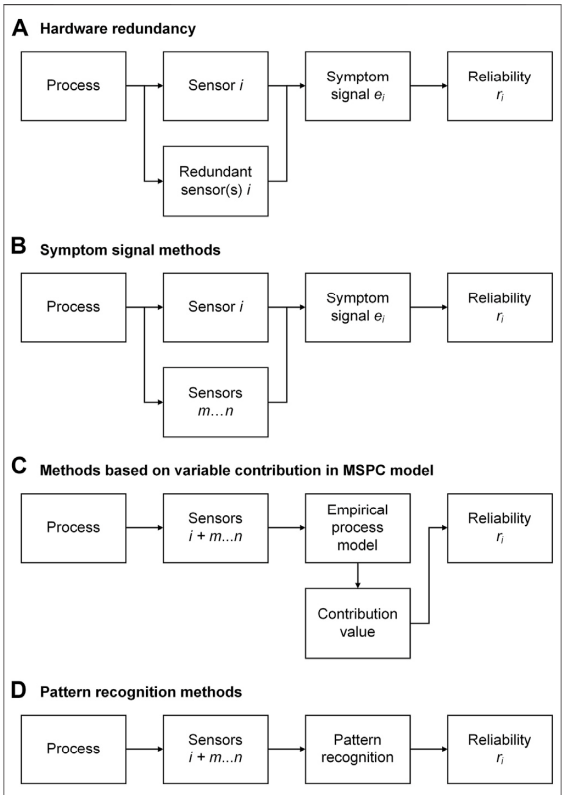


FIGURE 4 | Approaches for sensor fault detection: **(A)** hardware redundancy; **(B)** symptom signal methods; **(C)** methods based on variable contribution in MSPC model; **(D)** pattern recognition methods. The index *i* refers to *one* type of sensor for *one* physical or chemical variable, whereas *m . . . n* refers to *any other* types of sensors. For example, in **(A)**, the reliability of one turbidity sensor (sensor *i*) is determined using other turbidity sensors in the same bioreactor (redundant sensors *i*). In **(B)**, the reliability of a turbidity sensor (sensor *i*) is determined using other sensors such as exhaust gas, dissolved oxygen, and pH (sensors *m . . . n*). In **(C,D)**, the entire process data matrix consisting of the turbidity sensor to be monitored (sensor *i*) and the other sensors (*m . . . n*) is used as input to the empirical process model **(C)** and pattern recognition **(D)**, respectively. Approaches **(B–D)** are based on analytical redundancy.

Symptom Signal Methods

As mentioned in the introduction, soft sensors themselves can be used to assist in sensor fault detection. Here, the target quantity $\boldsymbol{y}$ of the soft sensor is the reading of the hardware sensor to be monitored. A deviation of the prediction $\widehat{\boldsymbol{y}}_{\boldsymbol{on}}$ from the original reading $\boldsymbol{y}_{\boldsymbol{on}}$ beyond a defined threshold value indicates a sensor fault. The residual between $\widehat{\boldsymbol{y}}_{\boldsymbol{on}}$ and $\boldsymbol{y}_{\boldsymbol{on}}$ is referred to as symptom signal $\boldsymbol{e}$ (Zarei and Shokri, 2014) and can in its simplest form be formulated as:

$$\boldsymbol{e} = \boldsymbol{y}_{\boldsymbol{on}} - \widehat{\boldsymbol{y}}_{\boldsymbol{on}}. \tag{3}$$

Several authors make use of state-space models for the generation of the symptom signal, as described in the following. Zarei and Shokri (2014) used a nonlinear unknown input observer to generate symptom signals and thereby to detect sensor faults in a simulated continuous stirred-tank reactor (CSTR) process. Alag et al. (2001) proposed a framework for sensor fault detection based on the symptom signal method exemplarily for a gas turbine power plant. They proposed a multi-step algorithm for determining $\widehat{\boldsymbol{y}}_{\boldsymbol{on}}$ followed by the generation and evaluation of the symptom signal. First, a redundant prediction for each sensor in the network is generated based on regression methods such as neural networks (redundancy creation). Subsequently, these predictions are fused with original sensor readings into a state-space model based on a Kalman filter approach (state prediction and fusion). The statistical properties of the symptom signal are in combination with probabilistic reasoning finally used to identify both abrupt and incipient sensor faults. Since a symptom signal is created for each sensor, the proposed methodology is capable of detecting multiple sensor faults simultaneously.

Autoassociative neural networks (AANN) were first introduced by Kramer (1991) for sensor fault detection and reconstruction in a simulated chemical batch process. They have proven to be effective in detecting sensor faults in a fermentation process (*Streptomyces virginiae*) with variable process length and multiple process phases (Huang et al., 2002). AANN are feed-forward neural networks consisting of an input, an output, and three hidden layers (mapping, bottleneck, and demapping layer). The outputs of the bottleneck layer are considered equivalent to the principal components of a nonlinear PCA (Kramer, 1991). The key concept of AANN is that the model is trained with fault-free process data $\boldsymbol{X}_{\boldsymbol{hist}}$ both as input and output, so that $\boldsymbol{X}_{\boldsymbol{hist}} = \boldsymbol{Y}_{\boldsymbol{hist}}$. The resulting nonlinear model is used for determining online predictions of the process data, $\widehat{\boldsymbol{Y}}_{\boldsymbol{on}}$, based on online measured process data, $\boldsymbol{X}_{\boldsymbol{on}} = \boldsymbol{Y}_{\boldsymbol{on}}$; then, analogous to **Eq. (3)**, the residual is calculated for each variable and used for the detection of sensor faults. This concept was extended to a complex nonlinear system with time-delays, namely a multicomponent distillation column (Perla et al., 2004).

The symptom signal method was used by Brunner et al. (2019) to detect sensor faults in a *P. pastoris* batch process. Due to the time-variant behavior and variable lengths of the batch processes, an indicator variable (maturity index) is introduced to predict the process progress online. For each process section, a set of prediction models for $\widehat{\boldsymbol{y}}_{\boldsymbol{on}}$ is generated. A regularization approach based on binary particle swarm optimization (PSO) is used to select the 25 best prediction models. The distribution of the predictions $\widehat{\boldsymbol{y}}_{\boldsymbol{on}}$ is compared to a moving window distribution of $\boldsymbol{y}_{\boldsymbol{on}}$ using the Kullback–Leibler divergence (Kullback and Leibler, 1951). The divergence between $\widehat{\boldsymbol{y}}_{\boldsymbol{on}}$ and $\boldsymbol{y}_{\boldsymbol{on}}$ indicates a sensor fault and is used to quantify the sensor reliability r_i.

Most studies use a fixed threshold for fault detection based on symptom signals. This can lead to false alarms when unforeseen events or noise occur in the sensor network data. In these cases, a time-varying as opposed to a fixed threshold can increase robustness and minimize the fault detection time, as shown by Armaou and Demetriou (2008) for simulated chemical processes. However, if process lengths vary, the threshold needs to adapt dynamically to process progress and not just to process time. For this reason, Brunner et al. (2019) proposed a dynamic threshold, which is calculated by means of the confidence width of $\widehat{\boldsymbol{y}}_{\boldsymbol{on}}$, which in turn is dependent on the process progress.

In addition to the mere detection of a sensor fault, information about the type of fault may also be necessary for the potential subsequent compensation (fault tolerance). To determine the type of fault as either bias, complete failure, drifting, or precision degradation, Dunia et al. (1996) developed a concept in which the symptom signal is generated using a PCA prediction model.

In summary, symptom signal methods are well suited for the detection of sensor faults and they are relatively intuitive due to their similarity to hardware redundancy. The main bottleneck of this approach is the model for the prediction of $\widehat{\boldsymbol{y}}_{\boldsymbol{on}}$, which is used for generating the symptom signal. For most bioprocesses, the model needs to consider time-variant behavior and variable process lengths. With the exception of AANNs, this model must be developed separately for each sensor to be monitored. The main advantage of the symptom signal method is that there is a direct reconstruction for the faulty sensor value available. Soft sensors or control systems, which depend on a reliable sensor input, can fall back on the reconstructed value and thus be designed fault tolerant.

Methods Based on Variable Contribution in Multivariate Statistical Process Control Model

Multivariate statistical process control (MSPC) and its corresponding empirical process models and control charts are another method to detect sensor faults. The original idea of MSPC is to map a wealth of process data $\boldsymbol{X}$ to one or a few higher-level, non-physical process quantities $\boldsymbol{y}$ or $\boldsymbol{Y}$, respectively, such as latent variables (Kourti, 2005). Deviations between historical, $\boldsymbol{X}_{\boldsymbol{hist}}$, and online process data $\boldsymbol{X}_{\boldsymbol{on}}$ are detected using control charts based on the complementary *SPE* (squared prediction error; sometimes denoted as *Q*) and Hotelling's T^2 statistics (Nomikos and MacGregor, 1995; Liu et al., 2017; Sánchez-Fernández et al., 2018). Once these test statistics indicate a significant deviation, the contribution of each input variable in $\boldsymbol{X}_{\boldsymbol{on}}$ to the test statistic(s) is calculated. Sensors or variables, respectively, with a significantly high contribution to the test statistic(s) are associated with a sensor fault. A general analysis of the

variable contribution approach is given by Qin (2003). Various methods for decomposing the test statistics to contributions, such as complete, partial, or reconstruction-based decomposition, were analyzed by Alcala and Qin (2011).

Sánchez-Fernández et al. (2018) combined the symptom signal and the contribution-based method for the detection of process and sensor faults. Residuals between predictions and observations for each variable in $\boldsymbol{X_{on}}$ are used as inputs to a PCA-based MSPC model. Residuals that are calculated with faultless training data are used for calculating the thresholds (control chart limits) for fault detection based on T^2 and *SPE* statistics. Multivariate and univariate exponentially weighted moving average control charts are used for the detection of process and sensor faults, respectively. Two simulated benchmark processes (Tennessee Eastman process and wastewater treatment plant) were used for validating the concept.

Another combination of the symptom signal and the contribution-based method was proposed by Yoo and Lee (2006) for sensor fault detection. Here, contribution plots assist in identifying faulty variables. In case the contribution plots indicate a fault, the original measurement is compared with a prediction based on a fuzzy PLS model of the corresponding variable. However, no algorithm was presented on how to derive the sensor reliability or the fault magnitude and type, respectively. The concept was evaluated on a real and a simulated wastewater treatment plant.

Reconstruction-based contributions (RBC) were proposed for sensor fault detection by Yue and Qin (2001). Here, T^2 and *SPE* are combined in a fault detection index φ. This combined index proved to have better detectability both for single and multiple sensor faults than if the contributions to T^2 and *SPE* are considered separately (Yue and Qin, 2001; Alcala and Qin, 2009). This concept was adopted by Tôrres et al. (2018) for pharmaceutical tablet manufacturing. The RBC approach was extended by Mnassri and Ouladsine (2015) to handle multiple and more complex sensor faults.

A contribution-based approach to sensor fault detection and tolerance was developed by Krause et al. (2015) for the monitoring of a yeast fermentation process. This approach does not consider the contribution to the test statistics as described above, but the direct contribution to the model $\boldsymbol{b}$ and the prediction $\widehat{\boldsymbol{y}}_{\boldsymbol{on}}$, respectively, for fault detection. They developed a PLS-based MSPC model using an indicator variable to compensate for variable process lengths. For each process section, a set of MSPC models is generated and PSO is used for finding the best models with respect to historical process data (with normal process behavior). Variable importance in the projection (VIP) scores (Chong and Jun, 2005) were used to evaluate the input variables for their contribution (information content) to the MSPC model. A reduction of the VIP score of a variable is assigned to a fault of the corresponding sensor.

A problem not to be underestimated in contribution-based fault detection is the smearing effect (Alcala and Qin, 2011; van den Kerkhof et al., 2013). Smearing here refers to the "influence of faulty variables on the contributions of non-faulty variables" (van den Kerkhof et al., 2013). Faulty variables (i.e., soft sensor inputs) can thus be concealed, and non-faulty variables can be incorrectly associated with faults. In contribution-based fault detection, groups of correlating variables are often displayed as faulty due to the smearing effect (van den Kerkhof et al., 2013); this is an obstacle especially for the often multicollinear data of bioprocesses.

To account for the nonlinearity of CSTR processes, several authors introduced the kernel PCA (Schölkopf et al., 1998) as a nonlinear extension of the PCA and adapted the calculation of the contributions to the T^2 and *SPE* statistics accordingly (Cho et al., 2005; Choi et al., 2005; Alcala and Qin, 2010).

The functional principle of AANN has already been described above for fault detection using symptom signals. Ren et al. (2018) proposed a reconstruction-based AANN to detect faults in nonlinear processes (simulated gas turbine). Both single and multiple faults could be detected despite the occurrence of smearing effects. It was further shown that in this case reconstruction-based AANN is superior to the other investigated methods (contribution plots-based PCA, contribution plots-based AANN, and reconstruction-based PCA) in terms of detection rate.

In summary, methods based on variable contribution currently represent the largest share among studies on sensor fault detection in the process industry. The main advantage of these methods is that the MSPC model can be used both for process and sensor fault detection. With only one MSPC model it is theoretically possible to monitor all input variables or sensors, respectively. To the best of our knowledge, however, there is only one study (Krause et al., 2015) that shows that, for highly multicollinear bioprocess data, smearing effects do not prevent successful sensor fault detection. For multiphase processes with variable process lengths, the MSPC models used for defect detection can be developed separately for each phase and a phase-specific indicator variable can be used for time synchronization (Ündey et al., 2003).

Pattern Recognition Methods

Unsupervised (clustering) and supervised (classification and regression) pattern recognition has been applied extensively for bioprocess monitoring (Lourenço et al., 2012; Rodríguez-Méndez et al., 2016). Also, in the detection of sensor faults by pattern recognition, a distinction is made between unsupervised and supervised methods.

In the case of unsupervised pattern recognition, the training data consist of fault-free process data. The relationships within the process variables are learned as patterns. A specific deviation from the fault-free pattern can then be assigned to a specific sensor fault (Barbariol et al., 2020). In this context, unsupervised pattern recognition is comparable to the aforementioned methods based on variable contribution in MSPC models: First, a deviation from the fault-free standard process is detected and then it is examined to determine to which variable the fault can be traced. These two approaches (MSPC vs. unsupervised pattern recognition) differ less by this underlying principle than by the modeling methods used (empirical process model vs. clustering algorithm).

Barbariol et al. (2020) used unsupervised anomaly detection algorithms to detect faults of a multiphase flow meter. Artificial faults were added to data of normal operating conditions. The

type of fault was identified as either complete failure, bias, precision degradation, or drift by a root cause analysis algorithm.

In the case of supervised pattern recognition, the training data contain sensor faults. These sensor faults can be artificial or real, but in any case, they must be labeled according to their reliability r_i or—conversely—their degree of faultiness $(1 - r_i)$. Faultiness is indicated either binarily (fault = true/false) or on a discrete (Mehranbod et al., 2003) or continuous scale (Guo and Nurre, 1991). Detecting sensor faults becomes a classification problem in case of binary or discrete faultiness and a regression problem in case of continuous faultiness. In both cases, labeled faulty data $\boldsymbol{X_{hist}}$ represent the inputs and the degree of faultiness (or the converse: r_i) represents the output. These input and output data are used for training the classification or regression model.

Guo and Nurre (1991) used supervised pattern recognition to detect and reconstruct sensor faults in a space shuttle main engine. Artificial random Gaussian noise was added to parts of fault-free data from normal operation. If the resulting artificial sensor readings are within the valid range, they are assigned a reliability of 0.9; if they are outside the valid range, they are assigned a reliability of 0.1. A feedforward ANN is trained with the manipulated sensor readings as inputs and the corresponding labeled reliability as outputs using a backpropagation algorithm to adjust the weights. In this way, even with a very small amount of original data, sensors whose readings do not match the rest of the sensor network can be identified. It was further shown that supervised pattern recognition is also suitable for the detection of multiple simultaneous sensor faults (Palmé et al., 2011) even in the presence of system failures (Romesis and Mathioudakis, 2003; Mathioudakis and Romessis, 2004). In this case, the training data must cover each of these cases (deviation from normal operation and multiple sensor faults), which causes the number of training patterns to increase rapidly.

In addition to the mere detection of faults, Mehranbod et al. (2003) distinguished between three different fault types (bias, drift, or noise) by identifying fault patterns in a moving window. They trained a Bayesian belief network to detect both single and multiple sensor faults in a polymerization reactor. This concept was later extended for the time-variant behavior of transient processes (Mehranbod et al., 2005).

In summary, pattern recognition methods are particularly attractive because ready-to-use—and in many toolboxes also auto-tuned—algorithms of machine learning can be applied to the problem of sensor fault detection without extensive statistical knowledge. At least in mechanical or chemical processes, efficient sensor fault detection can be realized with only little original data from normal operation together with artificial faults (Guo and Nurre, 1991). Despite this high potential, there are, to our knowledge, no studies that have explicitly used the previously trained pattern of sensor faults for their subsequent detection in bioprocesses. This lack of studies is all the more remarkable as pattern recognition methods are particularly efficient with such a high degree of multicollinearity as in bioprocesses.

Sensor Fault Tolerance

In the last sections, three different approaches to sensor fault detection were described. In the absence of sensor faults, the model input to soft sensors is considered reliable (online validation). But what is the use of online validation if the fault detected in the input data makes the soft sensor prediction unreliable? We know that something is going wrong, but we cannot change anything (upper branch in **Figure 5**). This is where the fault tolerance of soft sensors comes into play.

In general, modules for fault tolerance can be implemented at two layers of a soft sensor: at the inputs or in the actual soft sensor model.

The first variant of fault-tolerant soft sensors is shown in **Figure 5**. Here, sensor faults are first detected and, after a decision logic, they are compensated for by a reconstruction of the faulty sensor reading. This reconstruction is equivalent to missing data imputation (Dunia et al., 1996). The outputs of the fault tolerance module in **Figure 5** are the inputs to the soft sensor. The inputs and outputs of the described fault tolerant soft sensor are hereinafter referred to as $\boldsymbol{X_{on,FT}}$ and $\hat{\boldsymbol{y}}_{\boldsymbol{on,FT}}$. For bioprocesses, there are, to our knowledge, no studies available that explicitly address the development a fault-tolerant soft sensor based on fault detection and reconstruction. However, some authors have separately described the fault tolerance module shown in **Figure 5**, as described in the following.

In the already mentioned study by Huang et al. (2002), an AANN was used for sensor fault detection by means of a symptom signal and fault reconstruction in a fermentation process with variable process length and time-variant behavior. Examples of sensor fault reconstruction using AANN in applications other than biotechnology are given in Kramer (1991), Kramer (1992), and Hamidreza et al. (2014). Variable contribution statistics (T^2 and SPE) in a MSPC model were used by Lawal and Zhang (2017) to

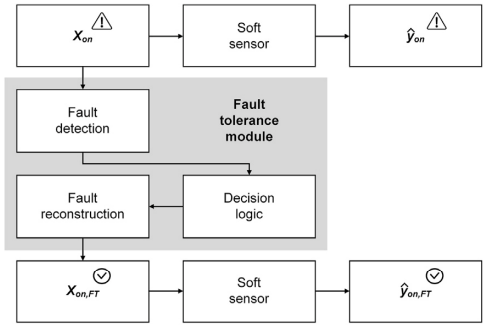


FIGURE 5 | Concept of a fault tolerance module for creating a fault-tolerant soft sensor. When a fault occurs in one or multiple variables of the soft sensor inputs $\boldsymbol{X_{on}}$, usually the soft sensor prediction $\hat{\boldsymbol{y}}_{\boldsymbol{on}}$ is also faulty dependent on the degree of influence of the variable(s) in the soft sensor model (upper branch). Fault-tolerant soft sensors are capable of compensating for faults in the inputs by the following procedure: Initially, the sensor fault is detected. Then, a decision logic determines whether or not to reconstruct the faulty sensor reading. This reconstruction results in a fault-free substitute for $\boldsymbol{X_{on}}$, namely, the fault-tolerant inputs $\boldsymbol{X_{on,FT}}$. The soft sensor uses these inputs for the determination of a fault-tolerant output (prediction) $\hat{\boldsymbol{y}}_{\boldsymbol{on,FT}}$.

detect and subsequently reconstruct faulty sensor reading within a crude distillation unit. In the above-mentioned study by Guo and Nurre (1991) an ANN was used to learn the patterns of both fault-free and faulty sensor readings to detect sensor faults. A separate ANN was trained to reconstruct the faulty readings.

In the second variant of fault-tolerant soft sensors, the faulty inputs are not reconstructed but the soft sensor algorithm itself is responsible for the fault management. For bioprocesses, there is, to our knowledge, only one study available that explicitly addresses fault tolerance by adapting the soft sensor models (Krause et al., 2015). The already described MSPC model developed by Krause et al. (2015) is capable of giving reliable predictions $\hat{\boldsymbol{y}}_{\boldsymbol{on,FT}}$ (here: higher-level process quantity) in the presence of sensor faults. As mentioned, PSO is used to find the best models with respect to historical process data. When sensor readings in $\boldsymbol{X}_{\boldsymbol{on}}$ differ significantly from historical data $\boldsymbol{X}_{\boldsymbol{hist}}$, they are penalized by the PSO cost function. This in turn results in a drastically decreased contribution of the faulty sensor reading and thus a fault-tolerant prediction of $\hat{\boldsymbol{y}}_{\boldsymbol{on}}$. With this approach to sensor fault tolerance the same has to be considered as with the entire MSPC concept: Both are ultimately based purely on a statistically significant deviation from $\boldsymbol{X}_{\boldsymbol{on}}$ to $\boldsymbol{X}_{\boldsymbol{hist}}$ and are thus strongly dependent on the size and quality of the process data pool for $\boldsymbol{X}_{\boldsymbol{hist}}$.

In summary, it must be noted that, with very few exceptions, there are no studies on fault-tolerant soft sensors for the process industry. With regard to fault detection before the subsequent reconstruction, all three methods described above are applicable. However, for the methods of variable contribution in a MSPC model and pattern recognition methods, a separate model must be developed to reconstruct the faulty sensor reading. Symptom signal methods offer the advantage that the reconstructed sensor reading is directly available.

CONCLUSION

Based on an overview of the status quo of soft sensor development and online implementation, this review article describes the challenges of variable process lengths, multiple phases, and sensor faults, and critically discusses the corresponding solution approaches. The challenges are considered both individually and synchronously, and the solution approaches are evaluated in terms of their practicality and applicability to bioprocesses.

Variable process lengths: Data synchronization techniques are employed to ensure that soft sensors provide correct predictions despite variable process lengths. For data synchronization, indicator variable techniques and particularly DTW dominate the bioprocess literature compared to curve registration techniques. Indicator variables alone can only be used for the alignment of the entire process lengths. In contrast, DTW and curve registration techniques can additionally be used for the alignment of landmarks. Indicator variable techniques require a higher degree of process knowledge (selection of appropriate process variables etc.) compared to DTW and curve registration techniques. DTW is the technique of choice when a solution is sought that does not require much process knowledge (compared to indicator variable techniques) and fine-tuning (compared to curve registration techniques).

Multiple process phases: The basic strategy for coping with multiple process phases is to divide the process datasets into individual phase segments and develop separate models for these segments. For the detection and division of process phases, trajectory-based and correlation-based methods have been proposed in the literature. Methods based on the progression of process trajectories, most notably via DTW, have to date been proposed more frequently in the bioprocess literature compared to correlation-based methods. Reasons for this include better comprehensibility of algorithms, easier interpretability of results, and coincidence with actual operational process phases in trajectory-based methods (Luo et al., 2016). On the other hand, correlation-based methods offer the advantage that they can be developed almost entirely without process knowledge. The consideration of phase transitions has so far been described only for trajectory-based methods (via FCM; Luo et al., 2016); for correlation-based methods, the consideration of phase transitions is still lacking.

Sensor faults: If the input to a soft sensor is faulty, there is a high probability that the output is faulty as well. Despite this obvious relation, studies on the detection of or even tolerance to sensor faults in bioprocesses are rare. Methods based on variable contributions in MSPC models are well established in the process industry for the identification of sensor faults. Further research is required to evaluate the applicability of these methods to highly collinear bioprocesses, as groups of correlating variables are often displayed as faulty due to smearing effects (van den Kerkhof et al., 2013). Symptom signal methods have been used to detect sensor faults and to reconstruct faults in bioprocesses. These methods, especially AANN, seem to be promising tools for the fault tolerance of soft sensors. The recognition of previously trained fault patterns has been used in mechanical engineering for fault detection, but to our knowledge has not yet been addressed in the bioprocess field. However, it can be assumed that this branch of machine learning will also increase in popularity in the field of bioprocesses due to steadily growing libraries of ready-to-use algorithms. For all three approaches presented for the detection of sensor faults (symptom signal, MSPC, pattern recognition) it could be shown that they are also capable of detecting simultaneously occurring sensor faults.

Synchronous consideration of the three challenges: The development of soft sensors for bioprocesses with multiple phases and variable process lengths has been investigated in several studies (e.g., Ündey et al., 2003; Luo et al., 2016). As described above, landmark-based data synchronization is particularly suitable for multiphase processes. For sensor fault detection for bioprocesses with variable lengths but without multiple phases individual studies exist (Krause et al., 2015; Brunner et al., 2019). Regarding sensor fault detection for multiphase bioprocesses with variable lengths, the question remains open as to which of the three methods presented is most suitable. This is because there is to the best of our knowledge only one study that provides a solution for the synchronous occurrence of all three challenges for bioprocesses (Huang et al., 2002).

The core conclusions of this review article are as follows:

- The choice of methods to handle variable process lengths and multiple process phases is dependent on the level of implementable process knowledge.
- The dilemma with sensor fault detection via soft sensors is that the input to the soft sensor can itself be erroneous.
- There is a clear research gap regarding the validation of the input data to soft sensors.
- Specifically, approaches to the tolerance of soft sensors to sensor faults need to be found.

Closing these gaps not only will allow existing sensor networks to be used more efficiently to monitor bioprocesses but will also strengthen confidence in soft sensors and PAT.

AUTHOR CONTRIBUTIONS

VB reviewed the literature and drafted the manuscript. MS, DG, and TB edited the manuscript. All authors have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

We appreciate the support from the German Federal Ministry of Education and Research (project 031B0475E), the German Research Foundation (project BE 2245/17-1), and the Open Access Publishing Fund of the Technical University of Munich (TUM).

REFERENCES

Afseth, N. K., Segtnan, V. H., and Wold, J. P. (2006). Raman Spectra of Biological Samples: A Study of Preprocessing Methods. *Appl. Spectrosc.* 60, 1358–1367. doi:10.1366/000370206779321454

Alag, S., Agogino, A. M., and Morjaria, M. (2001). A Methodology for Intelligent Sensor Measurement, Validation, Fusion, and Fault Detection for Equipment Monitoring and Diagnostics. *Artif. Intell. Eng. Des. Anal. Manuf.* 15, 307–320. doi:10.1017/s0890060401154053

Alcala, C. F., and Joe Qin, S. (2011). Analysis and Generalization of Fault Diagnosis Methods for Process Monitoring. *J. Process Control* 21, 322–330. doi:10.1016/j.jprocont.2010.10.005

Alcala, C. F., and Qin, S. J. (2009). Reconstruction-Based Contribution for Process Monitoring. *Automatica* 45, 1593–1600. doi:10.1016/j.automatica.2009.02.027

Alcala, C. F., and Qin, S. J. (2010). Reconstruction-based Contribution for Process Monitoring with Kernel Principal Component Analysis. *Ind. Eng. Chem. Res.* 49, 7849–7857. doi:10.1021/ie9018947

Andersen, S. W., and Runger, G. C. (2012). Automated Feature Extraction from Profiles with Application to a Batch Fermentation Process. *J. R. Stat. Soc. Ser. C* 61, 327–344. doi:10.1111/j.1467-9876.2011.01032.x

Armaou, A., and Demetriou, M. A. (2008). Robust Detection and Accommodation of Incipient Component and Actuator Faults in Nonlinear Distributed Processes. *AIChE J.* 54, 2651–2662. doi:10.1002/aic.11539

Balaban, E., Saxena, A., Bansal, P., Goebel, K. F., and Curran, S. (2009). Modeling, Detection, and Disambiguation of Sensor Faults for Aerospace Applications. *IEEE Sens. J.* 9, 1907–1917. doi:10.1109/jsen.2009.2030284

Barbariol, T., Feltresi, E., and Susto, G. A. (2020). Self-Diagnosis of Multiphase Flow Meters Through Machine Learning-Based Anomaly Detection. *Energies* 13, 3136. doi:10.3390/en13123136

Bayer, B., Stosch, M., Melcher, M., Duerkop, M., and Striedner, G. (2020). Soft Sensor Based on 2D-Fluorescence and Process Data Enabling Real-time Estimation of Biomass in *Escherichia coli* Cultivations. *Eng. Life Sci.* 20, 26–35. doi:10.1002/elsc.201900076

Berbegal, C., Khomenko, I., Russo, P., Spano, G., Fragasso, M., Biasioli, F., et al. (2020). PTR-ToF-MS for the Online Monitoring of Alcoholic Fermentation in Wine: Assessment of VOCs Variability Associated with Different Combinations of Saccharomyces/Non-Saccharomyces as a Case-Study. *Fermentation* 6, 55. doi:10.3390/fermentation6020055

Besenhard, M. O., Scheibelhofer, O., François, K., Joksch, M., and Kavsek, B. (2018). A Multivariate Process Monitoring Strategy and Control Concept for a Small-Scale Fermenter in a PAT Environment. *J. Intell. Manuf* 29, 1501–1514. doi:10.1007/s10845-015-1192-8

Bidar, B., Shahraki, F., Sadeghi, J., and Khalilipour, M. M. (2018). Soft Sensor Modeling Based on Multi-State-Dependent Parameter Models and Application for Quality Monitoring in Industrial Sulfur Recovery Process. *IEEE Sens. J.* 18, 4583–4591. doi:10.1109/jsen.2018.2818886

Biechele, P., Busse, C., Solle, D., Scheper, T., and Reardon, K. (2015). Sensor Systems for Bioprocess Monitoring. *Eng. Life Sci.* 15, 469–488. doi:10.1002/elsc.201500014

Biermann, F., Mathews, J., Nießing, B., König, N., and Schmitt, R. H. (2021). Automating Laboratory Processes by Connecting Biotech and Robotic Devices-An Overview of the Current Challenges, Existing Solutions and Ongoing Developments. *Processes* 9, 966. doi:10.3390/pr9060966

Bigot, J. (2006). Landmark-Based Registration of Curves via the Continuous Wavelet Transform. *J. Comput. Graph. Stat.* 15, 542–564. doi:10.1198/106186006x133023

Birle, S., Hussein, M. A., and Becker, T. (2013). Fuzzy Logic Control and Soft Sensing Applications in Food and Beverage Processes. *Food Control* 29, 254–269. doi:10.1016/j.foodcont.2012.06.011

Birle, S., Hussein, M. A., and Becker, T. (2015). On-Line Yeast Propagation Process Monitoring and Control Using an Intelligent Automatic Control System. *Eng. Life Sci.* 15, 83–95. doi:10.1002/elsc.201400058

Brunner, V., Klöckner, L., Kerpes, R., Geier, D. U., and Becker, T. (2019). Online Sensor Validation in Sensor Networks for Bioprocess Monitoring Using Swarm Intelligence. *Anal. Bioanal. Chem.* 412, 2165–2175. doi:10.1007/s00216-019-01927-7

Brunner, V., Siegl, M., Geier, D., and Becker, T. (2020). Biomass Soft Sensor for aPichia Pastorisfed-batch Process Based on Phase Detection and Hybrid Modeling. *Biotechnol. Bioeng.* 117, 2749–2759. doi:10.1002/bit.27454

Buyel, J. F., Twyman, R. M., and Fischer, R. (2017). Very-Large-Scale Production of Antibodies in Plants: The Biologization of Manufacturing. *Biotechnol. Adv.* 35, 458–465. doi:10.1016/j.biotechadv.2017.03.011

Cabaneros Lopez, P., Udugama, I. A., Thomsen, S. T., Roslander, C., Junicke, H., Iglesias, M. M., et al. (2021). Transforming Data to Information: A Parallel Hybrid Model for Real-time State Estimation in Lignocellulosic Ethanol Fermentation. *Biotechnol. Bioeng.* 118, 579–591. doi:10.1002/bit.27586

Camacho, J., Picó, J., and Ferrer, A. (2008). Multi-phase Analysis Framework for Handling Batch Process Data. *J. Chemom.* 22, 632–643. doi:10.1002/cem.1151

Camacho, J., and Picó, J. (2006). Multi-Phase Principal Component Analysis for Batch Processes Modelling. *Chemom. Intell. Lab. Syst.* 81, 127–136. doi:10.1016/j.chemolab.2005.11.003

Capito, F., Skudas, R., Kolmar, H., and Hunzinger, C. (2015). At-Line Mid Infrared Spectroscopy for Monitoring Downstream Processing Unit Operations. *Process Biochem.* 50, 997–1005. doi:10.1016/j.procbio.2015.03.005

Cawley, G. C., and Talbot, N. L. C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J. Machine Learn. Res.* 11, 2079–2107.

Chen, Y., Yang, O., Sampat, C., Bhalode, P., Ramachandran, R., and Ierapetritou, M. (2020). Digital Twins in Pharmaceutical and Biopharmaceutical Manufacturing: A Literature Review. *Processes* 8, 1088. doi:10.3390/pr8091088

Chew, W., and Sharratt, P. (2010). Trends in Process Analytical Technology. *Anal. Methods* 2, 1412. doi:10.1039/c0ay00257g

Cho, J.-H., Lee, J.-M., Wook Choi, S., Lee, D., and Lee, I.-B. (2005). Fault Identification for Process Monitoring Using Kernel Principal Component Analysis. *Chem. Eng. Sci.* 60, 279–288. doi:10.1016/j.ces.2004.08.007

Choi, S. W., Lee, C., Lee, J.-M., Park, J. H., and Lee, I.-B. (2005). Fault Detection and Identification of Nonlinear Processes Based on Kernel PCA. *Chemom. Intell. Lab. Syst.* 75, 55–67. doi:10.1016/j.chemolab.2004.05.001

Chong, I. G., and Jun, C. H. (2005). Performance of Some Variable Selection Methods when Multicollinearity Is Present. *Chemom. Intell. Lab. Syst.* 78, 103–112. doi:10.1016/j.chemolab.2004.12.011

Claßen, J., Aupert, F., Reardon, K. F., Solle, D., and Scheper, T. (2017). Spectroscopic Sensors for In-Line Bioprocess Monitoring in Research and Pharmaceutical Industrial Application. *Anal. Bioanal. Chem.* 409, 651–666. doi:10.1007/s00216-016-0068-x

Das, A., Maiti, J., and Banerjee, R. N. (2012). Process Monitoring and Fault Detection Strategies: A Review. *Int. J. Qual. Reliabil. Manage.* 29, 720–752. doi:10.1108/02656711211258508

Doan, X.-T., and Srinivasan, R. (2008). Online Monitoring of Multi-phase Batch Processes Using Phase-Based Multivariate Statistical Process Control. *Comput. Chem. Eng.* 32, 230–243. doi:10.1016/j.compchemeng.2007.05.010

Dumarey, M., Hermanto, M., Airiau, C., Shapland, P., Robinson, H., Hamilton, P., et al. (2019). Advances in Continuous Active Pharmaceutical Ingredient (API) Manufacturing: Real-Time Monitoring Using Multivariate Tools. *J. Pharm. Innov.* 14, 359–372. doi:10.1007/s12247-018-9348-7

Dunia, R., Qin, S. J., Edgar, T. F., and McAvoy, T. J. (1996). Identification of Faulty Sensors Using Principal Component Analysis. *AIChE J.* 42, 2797–2812. doi:10.1002/aic.690421011

EC (2010). Good Manufacturing Practice Medicinal Products for Human and Veterinary Use – Annex 11: Computerised Systems. Available at: https://ec.europa.eu/health/sites/default/files/files/eudralex/vol-4/annex11_01-2011_en.pdf.

FDA (2004). Guidance for Industry, PAT-A Framework for Innovative Pharmaceutical Development, Manufacturing and Quality Assurance. Available at: http://www.fda.gov/cder/guidance/published.html.

FDA (2003). Guidance for Industry: Part 11, Electronic Records; Electronic Signatures—Scope and Application. Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/part-11-electronic-records-electronic-signatures-scope-and-application.

Feital, T., and Pinto, J. C. (2015). Use of Variance Spectra for In-Line Validation of Process Measurements in Continuous Processes. *Can. J. Chem. Eng.* 93, 1426–1437. doi:10.1002/cjce.22219

Fortuna, L., Graziani, S., Rizzo, A., and Xibilia, M. G. (2007). *Soft Sensors for Monitoring and Control of Industrial Processes*. Berlin/Heidelberg: Springer Science & Business Media.

Fransson, M., and Folestad, S. (2006). Real-time Alignment of Batch Process Data Using COW for On-Line Process Monitoring. *Chemom. Intell. Lab. Syst.* 84, 56–61. doi:10.1016/j.chemolab.2006.04.020

García-Muñoz, S., Kourti, T., MacGregor, J. F., Mateos, A. G., and Murphy, G. (2003). Troubleshooting of an Industrial Batch Process Using Multivariate Methods. *Ind. Eng. Chem. Res.* 42, 3592–3601. doi:10.1021/ie0300023

García-Muñoz, S., Polizzi, M., Prpich, A., Strain, C., Lalonde, A., and Negron, V. (2011). Experiences in Batch Trajectory Alignment for Pharmaceutical Process Improvement Through Multivariate Latent Variable Modelling. *J. Process Control.* 21, 1370–1377. doi:10.1016/j.jprocont.2011.07.013

Gargalo, C. L., Udugama, I., Pontius, K., Lopez, P. C., Nielsen, R. F., Hasanzadeh, A., et al. (2020). Towards Smart Biomanufacturing: a Perspective on Recent Developments in Industrial Measurement and Monitoring Technologies for Bio-Based Production Processes. *J. Ind. Microbiol. Biotechnol.* 47, 947–964. doi:10.1007/s10295-020-02308-1

Gins, G., van den Kerkhof, P., and van Impe, J. F. M. (2012). Hybrid Derivative Dynamic Time Warping for Online Industrial Batch-End Quality Estimation. *Ind. Eng. Chem. Res.* 51, 6071–6084. doi:10.1021/ie2019068

Gollmer, K., and Posten, C. (1996). Supervision of Bioprocesses Using a Dynamic Time Warping Algorithm. *Control Eng. Pract.* 4, 1287–1295. doi:10.1016/0967-0661(96)00136-0

Gomis-Fons, J., Schwarz, H., Zhang, L., Andersson, N., Nilsson, B., Castan, A., et al. (2020). Model-based Design and Control of a Small-Scale Integrated Continuous End-to-End mAb Platform. *Biotechnol. Prog.* 36, e2995. doi:10.1002/btpr.2995

González-Martínez, J. M., Noord, O. E. de., and Ferrer, A. (2014). Multisynchro: a Novel Approach for Batch Synchronization in Scenarios of Multiple Asynchronisms. *J. chemometrics* 28, 462–475.

González-Martínez, J. M., Ferrer, A., and Westerhuis, J. A. (2011). Real-time Synchronization of Batch Trajectories for On-Line Multivariate Statistical Process Control Using Dynamic Time Warping. *Chemom. Intell. Lab. Syst.* 105, 195 206. doi:10.1016/j.chemolab.2011.01.003

González-Martínez, J. M., Westerhuis, J. A., and Ferrer, A. (2013). Using Warping Information for Batch Process Monitoring and Fault Classification. *Chemom. Intell. Lab. Syst.* 127, 210–217. doi:10.1016/j.chemolab.2013.07.003

Guo, T.-H., and Nurre, J. (1991). "Sensor Failure Detection and Recovery by Neural Networks," in IJCNN-91-Seattle International Joint Conference on Neural Networks, Seattle, WA, July 8–12, 1991.

Haimi, H., Mulas, M., Corona, F., and Vahala, R. (2013). Data-Derived Soft-Sensors for Biological Wastewater Treatment Plants: An Overview. *Environ. Model. Softw.* 47, 88–107. doi:10.1016/j.envsoft.2013.05.009

Hamidreza, M., Mehdi, S., Hooshang, J.-R., and Aliakbar, N. (2014). Reconstruction Based Approach to Sensor Fault Diagnosis Using Auto-Associative Neural Networks. *J. Cent. South. Univ.* 21, 2273–2281. doi:10.1007/s11771-014-2178-y

Hawkins, D. M. (2004). The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* 44, 1–12. doi:10.1021/ci0342472

Heinze, G., Wallisch, C., and Dunkler, D. (2018). Variable Selection - A Review and Recommendations for the Practicing Statistician. *Biom. J.* 60, 431–449. doi:10.1002/bimj.201700067

Huang, J., Shimizu, H., and Shioya, S. (2002). Data Preprocessing and Output Evaluation of an Autoassociative Neural Network Model for Online Fault Detection in Virginiamycin Production. *J. Biosci. Bioeng.* 94, 70–77. doi:10.1016/s1389-1723(02)80119-0

Isermann, R. (2011). *Fault-Diagnosis Applications: Model-Based Condition Monitoring: Actuators, Drives, Machinery, Plants, Sensors, and Fault-Tolerant Systems*. Berlin/Heidelberg: Springer Science & Business Media.

Isermann, R. (2006). *Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance*. Berlin/Heidelberg: Springer Science & Business Media.

ISPE (2008). GAMP 5 Guide: Compliant GxP Computerized Systems. Available at: https://ispe.org/publications/guidance-documents/gamp-5.

Jenzsch, M., Simutis, R., Eisbrenner, G., Stückrath, I., and Lübbert, A. (2006). Estimation of Biomass Concentrations in Fermentation Processes for Recombinant Protein Production. *Bioproc. Biosyst. Eng.* 29, 19–27. doi:10.1007/s00449-006-0051-6

Jin, H., Chen, X., Yang, J., Zhang, H., Wang, L., and Wu, L. (2015). Multi-Model Adaptive Soft Sensor Modeling Method Using Local Learning and Online Support Vector Regression for Nonlinear Time-Variant Batch Processes. *Chem. Eng. Sci.* 131, 282–303. doi:10.1016/j.ces.2015.03.038

Kabugo, J. C., Jämsä-Jounela, S.-L., Schiemann, R., and Binder, C. (2020). Industry 4.0 Based Process Data Analytics Platform: A Waste-To-Energy Plant Case Study. *Int. J. Electr. Power Energ. Syst.* 115, 105508. doi:10.1016/j.ijepes.2019.105508

Kadlec, P., Gabrys, B., and Strandt, S. (2009). Data-driven Soft Sensors in the Process Industry. *Comput. Chem. Eng.* 33, 795–814. doi:10.1016/j.compchemeng.2008.12.012

Kadlec, P., Grbić, R., and Gabrys, B. (2011). Review of Adaptation Mechanisms for Data-Driven Soft Sensors. *Comput. Chem. Eng.* 35, 1–24. doi:10.1016/j.compchemeng.2010.07.034

Kano, M., and Fujiwara, K. (2012). Virtual Sensing Technology in Process Industries: Trends and Challenges Revealed by Recent Industrial Applications. *J. Chem. Eng. Jpn.* 46, 1–17. doi:10.1252/jcej.12we167

Kassidas, A., MacGregor, J. F., and Taylor, P. A. (1998). Synchronization of Batch Trajectories Using Dynamic Time Warping. *AIChE J.* 44, 864–875. doi:10.1002/aic.690440412

Keogh, E. J., and Pazzani, M. J. (2001). "Derivative Dynamic Time Warping," in Proceedings of the First SIAM International Conference on Data Mining, Chicago, IL, April 5–7, 2001. Editor R. Grossman (Philadelphia, PA: SIAM), 1–11.

Kern, S., Wander, L., Meyer, K., Guhl, S., Mukkula, A. R. G., Holtkamp, M., et al. (2019). Flexible Automation with Compact NMR Spectroscopy for Continuous Production of Pharmaceuticals. *Anal. Bioanal. Chem.* 411, 3037–3046. doi:10.1007/s00216-019-01752-y

Khatibisepehr, S., Huang, B., and Khare, S. (2013). Design of Inferential Sensors in the Process Industry: A Review of Bayesian Methods. *J. Process Control.* 23, 1575–1596. doi:10.1016/j.jprocont.2013.05.007

Kourti, T. (2005). Application of Latent Variable Methods to Process Control and Multivariate Statistical Process Control in Industry. *Int. J. Adapt. Control. Signal. Process.* 19, 213–246. doi:10.1002/acs.859

Kramer, M. A. (1992). Autoassociative Neural Networks. *Comput. Chem. Eng.* 16, 313–328. doi:10.1016/0098-1354(92)80051-a

Kramer, M. A. (1991). Nonlinear Principal Component Analysis Using Autoassociative Neural Networks. *AIChE J.* 37, 233–243. doi:10.1002/aic.690370209

Krause, D., Hussein, M. A., and Becker, T. (2015). Online Monitoring of Bioprocesses via Multivariate Sensor Prediction within Swarm Intelligence Decision Making. *Chemom. Intell. Lab. Syst.* 145, 48–59. doi:10.1016/j.chemolab.2015.04.012

Krippl, M., Kargl, T., Duerkop, M., and Dürauer, A. (2021). Hybrid Modeling Reduces Experimental Effort to Predict Performance of Serial and Parallel Single-Pass Tangential Flow Filtration. *Separat. Purif. Technol.* 276, 119277. doi:10.1016/j.seppur.2021.119277

Kullaa, J. (2013). Detection, Identification, and Quantification of Sensor Fault in a Sensor Network. *Mech. Syst. Signal Process.* 40, 208–221. doi:10.1016/j.ymssp.2013.05.007

Kullback, S., and Leibler, R. A. (1951). On Information and Sufficiency. *Ann. Math. Statist.* 22, 79–86. doi:10.1214/aoms/1177729694

Lawal, S. A., and Zhang, J. (2017). "Actuator and Sensor Fault Tolerant Control of a Crude Distillation Unit," in *27th European Symposium on Computer Aided Process Engineering*. Editors A. Espuña, M. Graells, and L. Puigjaner (Amsterdam, Boston, Heidelberg: Elsevier), 1705–1710. doi:10.1016/b978-0-444-63965-3.50286-5

Liu, Y.-J., André, S., Saint Cristau, L., Lagresle, S., Hannas, Z., Calvosa, É., et al. (2017). Multivariate Statistical Process Control (MSPC) Using Raman Spectroscopy for In-Line Culture Cell Monitoring Considering Time-Varying Batches Synchronized with Correlation Optimized Warping (COW). *Analytica Chim. Acta* 952, 9–17. doi:10.1016/j.aca.2016.11.064

Lourenço, N. D., Lopes, J. A., Almeida, C. F., Sarraguça, M. C., and Pinheiro, H. M. (2012). Bioreactor Monitoring with Spectroscopy and Chemometrics: a Review. *Anal. Bioanal. Chem.* 404, 1211–1237. doi:10.1007/s00216-012-6073-9

Lu, N., and Gao, F. (2005). Stage-Based Process Analysis and Quality Prediction for Batch Processes. *Ind. Eng. Chem. Res.* 44, 3547–3555. doi:10.1021/ie048852l

Lu, N., Gao, F., and Wang, F. (2004). Sub-PCA Modeling and On-Line Monitoring Strategy for Batch Processes. *AIChE J.* 50, 255–259. doi:10.1002/aic.10024

Luo, L., Bao, S., Mao, J., Tang, D., and Gao, Z. (2016). Fuzzy Phase Partition and Hybrid Modeling Based Quality Prediction and Process Monitoring Methods for Multiphase Batch Processes. *Ind. Eng. Chem. Res.* 55, 4045–4058. doi:10.1021/acs.iecr.5b04252

Luttmann, R., Bracewell, D. G., Cornelissen, G., Gernaey, K. V., Glassey, J., Hass, V. C., et al. (2012). Soft Sensors in Bioprocessing: A Status Report and Recommendations. *Biotechnol. J.* 7, 1040–1048. doi:10.1002/biot.201100506

Mandenius, C.-F., and Gustavsson, R. (2015). Mini-Review: Soft Sensors as Means for PAT in the Manufacture of Bio-Therapeutics. *J. Chem. Technol. Biotechnol.* 90, 215–227. doi:10.1002/jctb.4477

Mathioudakis, K., and Romessis, C. (2004). Probabilistic Neural Networks for Validation of On-Board Jet Engine Data. *Proc. Inst. Mech. Eng. G: J. Aerospace Eng.* 218, 59–72. doi:10.1177/095441000421800105

Matthews, T. E., Berry, B. N., Smelko, J., Moretto, J., Moore, B., and Wiltberger, K. (2016). Closed Loop Control of Lactate Concentration in Mammalian Cell Culture by Raman Spectroscopy Leads to Improved Cell Density, Viability, and Biopharmaceutical Protein Production. *Biotechnol. Bioeng.* 113, 2416–2424. doi:10.1002/bit.26018

Mehranbod, N., Soroush, M., and Panjapornpon, C. (2005). A Method of Sensor Fault Detection and Identification. *J. Process Control.* 15, 321–339. doi:10.1016/j.jprocont.2004.06.009

Mehranbod, N., Soroush, M., Piovoso, M., and Ogunnaike, B. A. (2003). Probabilistic Model for Sensor Fault Detection and Identification. *AIChE J.* 49, 1787–1802. doi:10.1002/aic.690490716

Mei, C., Su, Y., Liu, G., Ding, Y., and Liao, Z. (2017). Dynamic Soft Sensor Development Based on Gaussian Mixture Regression for Fermentation Processes. *Chin. J. Chem. Eng.* 25, 116–122. doi:10.1016/j.cjche.2016.07.005

Melcher, M., Scharl, T., Spangl, B., Luchner, M., Cserjan, M., Bayer, K., et al. (2015). The Potential of Random forest and Neural Networks for Biomass and Recombinant Protein Modeling inEscherichia Colifed-Batch Fermentations. *Biotechnol. J.* 10, 1770–1782. doi:10.1002/biot.201400790

Meng, Y., Lan, Q., Qin, J., Yu, S., Pang, H., and Zheng, K. (2019). Data-Driven Soft Sensor Modeling Based on Twin Support Vector Regression for Cane Sugar Crystallization. *J. Food Eng.* 241, 159–165. doi:10.1016/j.jfoodeng.2018.07.035

Mnassri, B., El Adel, E. M., and Ouladsine, M. (2015). Reconstruction-based Contribution Approaches for Improved Fault Diagnosis Using Principal Component Analysis. *J. Process Control.* 33, 60–76. doi:10.1016/j.jprocont.2015.06.004

Moser, U., and Schramm, D. (2019). Multivariate Dynamic Time Warping in Automotive Applications: A Review. *Intell. Data Anal.* 23, 535–553. doi:10.3233/ida-184130

Nair, A. M., Hykkerud, A., and Ratnaweera, H. (2020). A Cost-Effective IoT Strategy for Remote Deployment of Soft Sensors – A Case Study on Implementing a Soft Sensor in a Multistage MBBR Plant. *Water Sci. Technol.* 81, 1733–1739. doi:10.2166/wst.2020.067

Nielsen, N.-P. V., Carstensen, J. M., and Smedsgaard, J. (1998). Aligning of Single and Multiple Wavelength Chromatographic Profiles for Chemometric Data Analysis Using Correlation Optimised Warping. *J. Chromatogr. A* 805, 17–35. doi:10.1016/s0021-9673(98)00021-1

Nomikos, P., and MacGregor, J. F. (1995). Multivariate SPC Charts for Monitoring Batch Processes. *Technometrics* 37, 41–59. doi:10.1080/00401706.1995.10485888

OECD (2014). *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*. Paris.

Ohadi, K., Legge, R. L., and Budman, H. M. (2015). Development of a Soft-Sensor Based on Multi-Wavelength Fluorescence Spectroscopy and a Dynamic Metabolic Model for Monitoring Mammalian Cell Cultures. *Biotechnol. Bioeng.* 112, 197–208. doi:10.1002/bit.25339

Pais, D. A. M., Galrão, P. R. S., Kryzhanska, A., Barbau, J., Isidro, I. A., and Alves, P. M. (2020). Holographic Imaging of Insect Cell Cultures: Online Non-invasive Monitoring of Adeno-Associated Virus Production and Cell Concentration. *Processes* 8, 487. doi:10.3390/pr8040487

Palmé, T., Fast, M., and Thern, M. (2011). Gas Turbine Sensor Validation Through Classification with Artificial Neural Networks. *Appl. Energ.* 88, 3898–3904. doi:10.1016/j.apenergy.2011.03.047

Pani, A. K., and Mohanta, H. K. (2011). A Survey of Data Treatment Techniques for Soft Sensor Design. *Chem. Product. Process Model.* 6. doi:10.2202/1934-2659.1536

Pappenreiter, M., Sissolak, B., Sommeregger, W., and Striedner, G. (2019). Oxygen Uptake Rate Soft-Sensing via Dynamic kLa Computation: Cell Volume and Metabolic Transition Prediction in Mammalian Bioprocesses. *Front. Bioeng. Biotechnol.* 7, 195. doi:10.3389/fbioe.2019.00195

Paquet-Durand, O., Assawarajuwan, S., and Hitzmann, B. (2017). Artificial Neural Network for Bioprocess Monitoring Based on Fluorescence Measurements: Training without Offline Measurements. *Eng. Life Sci.* 17, 874–880.

Pearson, R. K., Neuvo, Y., Astola, J., and Gabbouj, M. (2016). Generalized Hampel Filters. *EURASIP J. Adv. Signal Process.* 2016, 1–18.

Perla, R., Mukhopadhyay, S., and Samanta, A. N. (2004). "Sensor Fault Detection and Isolation Using Artificial Neural Networks," in 2004 IEEE Region 10 Conference TENCON, Chiang Mai, Thailand, November 24, 2004.

Qin, S. J. (2003). Statistical Process Monitoring: Basics and Beyond. *J. Chemom.* 17, 480–502. doi:10.1002/cem.800

Ramsay, J. O., and Silverman, B. W. (2005). *Functional Data Analysis*. New York, NY: Springer Science+Business Media, Inc.

Randek, J., and Mandenius, C.-F. (2018). On-line Soft Sensing in Upstream Bioprocessing. *Crit. Rev. Biotechnol.* 38, 106–121. doi:10.1080/07388551.2017.1312271

Rathore, A. S., Mishra, S., Nikita, S., and Priyanka, P. (2021). Bioprocess Control: Current Progress and Future Perspectives. *Life* 11, 557. doi:10.3390/life11060557

Rathore, A. S., and Winkle, H. (2009). Quality by Design for Biopharmaceuticals. *Nat. Biotechnol.* 27, 26–34. doi:10.1038/nbt0109-26

Rato, T. J., Rendall, R., Gomes, V., Chin, S.-T., Chiang, L. H., Saraiva, P. M., et al. (2016). A Systematic Methodology for Comparing Batch Process Monitoring Methods: Part I-Assessing Detection Strength. *Ind. Eng. Chem. Res.* 55, 5342–5358. doi:10.1021/acs.iecr.5b04851

Rato, T. J., Rendall, R., Gomes, V., Saraiva, P. M., and Reis, M. S. (2018). A Systematic Methodology for Comparing Batch Process Monitoring Methods:

Part II-Assessing Detection Speed. *Ind. Eng. Chem. Res.* 57, 5338–5350. doi:10.1021/acs.iecr.7b04911

Ren, S., Si, F., Zhou, J., Qiao, Z., and Cheng, Y. (2018). A New Reconstruction-Based Auto Associative Neural Network for Fault Diagnosis in Nonlinear Systems. *Chemom. Intell. Lab. Syst.* 172, 118–128. doi:10.1016/j.chemolab.2017.12.005

Rodríguez-Méndez, M. L., Saja, J. A. de., González-Antón, R., García-Hernández, C., Medina-Plaza, C., García-Cabezón, C., et al. (2016). Electronic Noses and Tongues in Wine Industry. *Front. Bioeng. Biotechnol.* 4, 81. doi:10.3389/fbioe.2016.00081

Romesis, C., and Mathioudakis, K. (2003). Setting up of a Probabilistic Neural Network for Sensor Fault Detection Including Operation with Component Faults. *J. Eng. Gas Turbines Power* 125, 634–641. doi:10.1115/1.1582493

Sakoe, H., and Chiba, S. (1978). Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Trans. Acoust. Speech, Signal. Process.* 26, 43–49. doi:10.1109/tassp.1978.1163055

Sánchez-Fernández, A., Baldán, F. J., Sainz-Palmero, G. I., Benítez, J. M., and Fuente, M. J. (2018). Fault Detection Based on Time Series Modeling and Multivariate Statistical Process Control. *Chemometrics Intell. Lab. Syst.* 182, 57–69. doi:10.1016/j.chemolab.2018.08.003

Saptoro, A. (2014). State of the Art in the Development of Adaptive Soft Sensors Based on Just-In-Time Models. *Proced. Chem.* 9, 226–234. doi:10.1016/j.proche.2014.05.027

Sauer, D. G., Melcher, M., Mosor, M., Walch, N., Berkemeyer, M., Scharl-Hirsch, T., et al. (2019). Real-time Monitoring and Model-based Prediction of Purity and Quantity during a Chromatographic Capture of Fibroblast Growth Factor 2. *Biotechnol. Bioeng.* 116, 1999–2009. doi:10.1002/bit.26984

Scheper, T., Beutel, S., McGuinness, N., Heiden, S., Oldiges, M., Lammers, F., et al. (2021). Digitalization and Bioprocessing: Promises and Challenges. *Adv. Biochem. Eng. Biotechnol.* 176, 57–69. doi:10.1007/10_2020_139

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Comput.* 10, 1299–1319. doi:10.1162/089976698300017467

Sharma, A. B., Golubchik, L., and Govindan, R. (2010). Sensor Faults. *ACM Trans. Sen. Netw.* 6, 1–39. doi:10.1145/1754414.1754419

Shokoohi-Yekta, M., Wang, J., and Keogh, E. (2015). "On the Non-trivial Generalization of Dynamic Time Warping to the Multi-Dimensional Case," in *Proceedings of the 2015 SIAM International Conference on Data Mining*. Editors S. Venkatasubramanian and J. Ye (Philadelphia, PA: Society for Industrial and Applied Mathematics).

Simon, L. L., Pataki, H., Marosi, G., Meemken, F., Hungerbühler, K., Baiker, A., et al. (2015). Assessment of Recent Process Analytical Technology (PAT) Trends: A Multiauthor Review. *Org. Process. Res. Dev.* 19, 3–62. doi:10.1021/op500261y

Sokolov, M., Soos, M., Neunstoecklin, B., Morbidelli, M., Butté, A., Leardi, R., et al. (2015). Fingerprint Detection and Process Prediction by Multivariate Analysis of Fed-Batch Monoclonal Antibody Cell Culture Data. *Biotechnol. Prog.* 31, 1633–1644. doi:10.1002/btpr.2174

Solle, D., Hitzmann, B., Herwig, C., Pereira Remelhe, M., Ulonska, S., Wuerth, L., et al. (2017). Between the Poles of Data-Driven and Mechanistic Modeling for Process Operation. *Chem. Ingenieur Technik* 89, 542–561. doi:10.1002/cite.201600175

Souza, F. A. A., Araújo, R., and Mendes, J. (2016). Review of Soft Sensor Methods for Regression Applications. *Chemom. Intell. Lab. Syst.* 152, 69–79. doi:10.1016/j.chemolab.2015.12.011

Spann, R., Gernaey, K. V., and Sin, G. (2019). A Compartment Model for Risk-Based Monitoring of Lactic Acid Bacteria Cultivations. *Biochem. Eng. J.* 151, 107293. doi:10.1016/j.bej.2019.107293

Spooner, M., Kold, D., and Kulahci, M. (2018). Harvest Time Prediction for Batch Processes. *Comput. Chem. Eng.* 117, 32–41. doi:10.1016/j.compchemeng.2018.05.019

Spooner, M., Kold, D., and Kulahci, M. (2017). Selecting Local Constraint for Alignment of Batch Process Data with Dynamic Time Warping. *Chemom. Intell. Lab. Syst.* 167, 161–170. doi:10.1016/j.chemolab.2017.05.019

Srinivasan, R., and Qian, M. (2007). Online Temporal Signal Comparison Using Singular Points Augmented Time Warping. *Ind. Eng. Chem. Res.* 46, 4531–4548. doi:10.1021/ie060111s

Srinivasan, R., and Qian, M. S. (2005). Off-Line Temporal Signal Comparison Using Singular Points Augmented Time Warping. *Ind. Eng. Chem. Res.* 44, 4697–4716. doi:10.1021/ie049528t

Srinivasan, R., and Sheng Qian, M. (2006). Online Fault Diagnosis and State Identification during Process Transitions Using Dynamic Locus Analysis. *Chem. Eng. Sci.* 61, 6109–6132. doi:10.1016/j.ces.2006.05.037

Steinwandter, V., Borchert, D., and Herwig, C. (2019). Data Science Tools and Applications on the Way to Pharma 4.0. *Drug Discov. Today* 24, 1795–1805. doi:10.1016/j.drudis.2019.06.005

Steinwandter, V., Zahel, T., Sagmeister, P., and Herwig, C. (2017). Propagation of Measurement Accuracy to Biomass Soft-Sensor Estimation and Control Quality. *Anal. Bioanal. Chem.* 409, 693–706. doi:10.1007/s00216-016-9711-9

Stork, C. L., and Kowalski, B. R. (1999). Distinguishing Between Process Upsets and Sensor Malfunctions Using Sensor Redundancy. *Chemom. Intell. Lab. Syst.* 46, 117–131. doi:10.1016/s0169-7439(98)00180-4

Tahir, F., Islam, M. T., Mack, J., Robertson, J., and Lovett, D. (2019). Process Monitoring and Fault Detection on a Hot-Melt Extrusion Process Using In-Line Raman Spectroscopy and a Hybrid Soft Sensor. *Comput. Chem. Eng.* 125, 400–414. doi:10.1016/j.compchemeng.2019.03.019

Tôrres, A. R., de Oliveira, A. D. P., Grangeiro, S., and Fragoso, W. D. (2018). Multivariate Statistical Process Control in Annual Pharmaceutical Product Review. *J. Process Control.* 69, 97–102. doi:10.1016/j.jprocont.2018.06.001

Ündey, C., and Çınar, A. (2002). Statistical Monitoring of Multistage, Multiphase Batch Processes. *IEEE Control. Syst.* 22, 40–52. doi:10.1109/MCS.2002.1035216

Ündey, C., Ertunç, S., and Çınar, A. (2003). Online Batch/Fed-Batch Process Performance Monitoring, Quality Prediction, and Variable-Contribution Analysis for Diagnosis. *Ind. Eng. Chem. Res.* 42, 4645–4658. doi:10.1021/ie0208218

Ündey, C., Williams, B. A., and Çınar, A. (2002). Monitoring of Batch Pharmaceutical Fermentations: Data Synchronization, Landmark Alignment, and Real-Time Monitoring. *IFAC Proc. Volumes* 35, 271–276. doi:10.3182/20020721-6-es-1901.01354

van den Kerkhof, P., Vanlaer, J., Gins, G., and van Impe, J. F. M. (2013). Analysis of Smearing-Out in Contribution Plot Based Fault Isolation for Statistical Process Control. *Chem. Eng. Sci.* 104, 285–293. doi:10.1016/j.ces.2013.08.007

Veichtlbauer, A., Ortmayer, M., and Heistracher, T. (2017). "OPC UA Integration for Field Devices," in 2017 IEEE 15th International Conference on Industrial Informatics (INDIN): University of Applied Science Emden/Leer, Emden, Germany, 24-26 July 2017 (Piscataway, NJ: IEEE), 419–424.

Venkatasubramanian, V., Rengaswamy, R., and Kavuri, S. N. (2003a). A Review of Process Fault Detection and Diagnosis. *Comput. Chem. Eng.* 27, 313–326. doi:10.1016/s0098-1354(02)00161-8

Venkatasubramanian, V., Rengaswamy, R., Kavuri, S. N., and Yin, K. (2003b). A Review of Process Fault Detection and Diagnosis. *Comput. Chem. Eng.* 27, 327–346. doi:10.1016/s0098-1354(02)00162-x

Venkatasubramanian, V., Rengaswamy, R., Yin, K., and Kavuri, S. N. (2003c). A Review of Process Fault Detection and Diagnosis. *Comput. Chem. Eng.* 27, 293–311. doi:10.1016/s0098-1354(02)00160-6

von Stosch, M., Davy, S., Francois, K., Galvanauskas, V., Hamelink, J. M., Luebbert, A., et al. (2014). Hybrid Modeling for Quality by Design and PAT-Benefits and Challenges of Applications in Biopharmaceutical Industry. *Biotechnol. J.* 9, 719–726. doi:10.1002/biot.201300385

Voss, J.-P., Mittelheuser, N. E., Lemke, R., and Luttmann, R. (2017). Advanced Monitoring and Control of Pharmaceutical Production Processes with Pichia pastoris by Using Raman Spectroscopy and Multivariate Calibration Methods. *Eng. Life Sci.* 17, 1281–1294. doi:10.1002/elsc.201600229

Walch, N., Scharl, T., Felföldi, E., Sauer, D. G., Melcher, M., Leisch, F., et al. (2019). Prediction of the Quantity and Purity of an Antibody Capture Process in Real Time. *Biotechnol. J.* 14, e1800521. doi:10.1002/biot.201800521

Wasalathanthri, D. P., Feroz, H., Puri, N., Hung, J., Lane, G., Holstein, M., et al. (2020a). Real-time Monitoring of Quality Attributes by In-line Fourier Transform Infrared Spectroscopic Sensors at Ultrafiltration and Diafiltration of Bioprocess. *Biotechnol. Bioeng.* 117, 3766–3774. doi:10.1002/bit.27532

Wasalathanthri, D. P., Rehmann, M. S., Song, Y., Gu, Y., Mi, L., Shao, C., et al. (2020b). Technology Outlook for Real-time Quality Attribute and Process Parameter Monitoring in Biopharmaceutical Development-A Review. *Biotechnol. Bioeng.* 117, 3182–3198. doi:10.1002/bit.27461

Williams, B. A., Ündey, C., and Cinar, A. (2001). Detection of Process Landmarks Using Registration for On-Line Monitoring. *IFAC Proc. Volumes* 34, 221–226. doi:10.1016/s1474-6670(17)33827-2

Wise, B. M., and Roginski, R. T. (2015). A Calibration Model Maintenance Roadmap. *IFAC-PapersOnLine* 48, 260–265. doi:10.1016/j.ifacol.2015.08.191

Wu, H., Read, E., White, M., Chavez, B., Brorson, K., Agarabi, C., et al. (2015). Real Time Monitoring of Bioreactor mAb IgG3 Cell Culture Process Dynamics via Fourier Transform Infrared Spectroscopy: Implications for Enabling Cell Culture Process Analytical Technology. *Front. Chem. Sci. Eng.* 9, 386–406. doi:10.1007/s11705-015-1533-3

Yao, Y., and Gao, F. (2009). A Survey on Multistage/Multiphase Statistical Modeling Methods for Batch Processes. *Annu. Rev. Control.* 33, 172–183. doi:10.1016/j.arcontrol.2009.08.001

Yoo, C. K., and Lee, I.-B. (2006). Integrated Framework of Nonlinear Prediction and Process Monitoring for Complex Biological Processes. *Bioproc. Biosyst. Eng.* 29, 213–228. doi:10.1007/s00449-006-0063-2

Yu, J., Chen, J., and Rashid, M. M. (2013). Multiway Independent Component Analysis Mixture Model and Mutual Information Based Fault Detection and Diagnosis Approach of Multiphase Batch Processes. *AIChE J.* 59, 2761–2779. doi:10.1002/aic.14051

Yu, J., and Qin, S. J. (2009). Multiway Gaussian Mixture Model Based Multiphase Batch Process Monitoring. *Ind. Eng. Chem. Res.* 48, 8585–8594. doi:10.1021/ie900479g

Yue, H. H., and Qin, S. J. (2001). Reconstruction-based Fault Identification Using a Combined index. *Ind. Eng. Chem. Res.* 40, 4403–4414. doi:10.1021/ie000141+

Zarei, J., and Shokri, E. (2014). Robust Sensor Fault Detection Based on Nonlinear Unknown Input Observer. *Measurement* 48, 355–367. doi:10.1016/j.measurement.2013.11.015

Zhang, A. H., Zhu, K. Y., Zhuang, X. Y., Liao, L. X., Huang, S. Y., Yao, C. Y., et al. (2020). A Robust Soft Sensor to Monitor 1,3-Propanediol Fermentation Process by Clostridium Butyricum Based on Artificial Neural Network. *Biotechnol. Bioeng.* 117, 3345–3355. doi:10.1002/bit.27507

Zhang, S., Zhao, C., and Gao, F. (2018). Two-directional Concurrent Strategy of Mode Identification and Sequential Phase Division for Multimode and Multiphase Batch Process Monitoring with Uneven Lengths. *Chem. Eng. Sci.* 178, 104 117. doi:10.1016/j.ces.2017.12.025

Zhang, Y., Lu, B., and Edgar, T. F. (2013). Batch Trajectory Synchronization with Robust Derivative Dynamic Time Warping. *Ind. Eng. Chem. Res.* 52, 12319–12328. doi:10.1021/ie303310c

Zheng, J., and Song, Z. (2018). Semisupervised Learning for Probabilistic Partial Least Squares Regression Model and Soft Sensor Application. *J. Process Control.* 64, 123–131. doi:10.1016/j.jprocont.2018.01.008

Zhu, P., Liu, X., Wang, Y., and Yang, X. (2018). Mixture Semisupervised Bayesian Principal Component Regression for Soft Sensor Modeling. *IEEE Access* 6, 40909–40919. doi:10.1109/access.2018.2859366

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

GLOSSARY

$\boldsymbol{b}$ model coefficients

E process end

$\boldsymbol{e}$ residual between prediction and original sensor reading

h warping function

i index for sensor of interest

j number of process variables

L landmark

$m \ldots n$ indices for sensors other than sensor i

r_i reliability of sensor i

SPE squared prediction error (sometimes denoted as Q)

t time

$\boldsymbol{X}$ process data = input to soft sensor

$\boldsymbol{X_{hist}}$ historical process data

$\boldsymbol{X_{on}}$ online process data

$\boldsymbol{X_{on,FT}}$ fault-tolerant online process data

$\boldsymbol{X_{sync}}$ synchronized process data

$\boldsymbol{y}$ or $\boldsymbol{Y}$ target quantity (vector or matrix) = output of soft sensor

$\boldsymbol{y_{on}}$ online data of target quantity

$\boldsymbol{y_{hist}}$ historical data of target quantity

$\tilde{\boldsymbol{y}}_{\boldsymbol{hist}}$ prediction for historical data of target quantity

$\hat{\boldsymbol{y}}_{\boldsymbol{on}}$ or $\hat{\boldsymbol{Y}}_{\boldsymbol{on}}$ prediction for target quantity (vector or matrix)

$\tilde{\boldsymbol{y}}_{\boldsymbol{on,FT}}$ fault-tolerant prediction for target quantity

φ combined fault detection index

AANN autoassociative neural network

ANN artificial neural network

COW correlation optimized warping

CPP critical process parameter

CQA critical quality attribute

CSTR continuous stirred-tank reactor

(D/M)DTW (derivative/multivariate) dynamic time warping

DSP downstream processing

EC European Commission

FCM fuzzy *c*-means

FDA Food and Drug Administration

GMM Gaussian mixture model

ISPE International Society for Pharmaceutical Engineering

JIT just-in-time

MLR multiple linear regression

MSPC multivariate statistical process control

OECD Organisation for Economic Co-operation and Development

OPC (UA) open platform communications (unified architecture)

PSO particle swarm optimization

RBC reconstruction-based contributions

RGTW relaxed-greedy time warping

PAT process analytical technology

PCA principal component analysis

PCR principal component regression

PLS(R) partial least squares (regression)

QbD quality by design

SCADA supervisory control and data acquisition

SVR support vector regression

USP upstream processing

VIP variable importance in the projection

XTW extrapolative time warping

3.2 Biomass estimation in *Pichia pastoris* cultures by combined single-wavelength fluorescence measurements

ARTICLE

BIOTECHNOLOGY and BIOENGINEERING

Biomass Estimation in *Pichia pastoris* Cultures by Combined Single-Wavelength Fluorescence Measurements

Vincent Brunner,[1] Mohamed Hussein,[1] Thomas Becker[2]

[1]Bio-PAT (Bio-Process Analysis Technology), Technische Universität München, Weihenstephaner Steig 20, Freising 85354, Germany; telephone: 49 (0)8161-71-3277; fax: 49 (0)8161-71-3883; e-mail: mohamed.hussein@tum.de

[2]Chair of Brewing and Beverage Technology, Technische Universität München, Freising, Germany

ABSTRACT: In this work, the evolution of different biogenic fluorophores involved in the metabolism of *Pichia pastoris* was determined at four different single-wavelength pairs (excitation/emission) during batch culture in microwell plates and used for effective and reliable biomass estimation by means of chemometric tools. The chemometric tools for biomass estimation were multiple linear regression (MLR), partial least squares regression (PLSR), and principal component regression (PCR). Variable importance in the projection (VIP) scores were used to rate the importance of model input variables, indicating tryptophan as the most important variable for biomass estimation. A direct correlation between the single fluorescence signals of tryptophan and biomass was additionally set up. Results indicate a successful fitting of the MLR, PLSR, PCR, and direct tryptophan correlation models for the present case and confirm the relevance of biogenic fluorophores for bioprocess state variables monitoring. The root mean squared error of prediction (RMSEP) between the predicted and measured values for the validation batches was 0.017, 0.023, 0.025, and 0.049 g L^{-1} dry cell weight for MLR, PLSR, PCR, and direct tryptophan correlation, respectively. The presented approach of indirectly measuring biomass based on combined single-wavelength fluorescence measurements can be used for the development of a low-cost alternative to 2D fluorescence spectroscopy.

Biotechnol. Bioeng. 2016;113: 2394–2402.

KEYWORDS: *Pichia pastoris*; bioprocess monitoring; biomass estimation; fluorescence spectroscopy; partial least squares regression; principal component regression

Correspondence to: M. Hussein
Contract grant sponsor: German Federal Ministry of Education and Research
Contract grant number: 031A616D
Received 9 February 2016; Revision received 8 April 2016; Accepted 5 May 2016
Accepted manuscript online 9 May 2016;
Article first published online 3 June 2016 in Wiley Online Library (http://onlinelibrary.wiley.com/doi/10.1002/bit.26003/abstract).
DOI 10.1002/bit.26003

Introduction

The methylotrophic yeast *Pichia pastoris* is frequently used in industry and research as a host for expression of heterologous genes with over 500 recombinant proteins being expressed in this system (Cereghino et al., 2002; Cregg et al., 2000). Despite the decades of experience with *P. pastoris* in industry and academia, optimization potential within the production of recombinant proteins in *P. pastoris* can still be found (Cos et al., 2006; Potvin et al., 2012). A common problem in bioprocess monitoring and control is the online determination of the key variables biomass, substrate, and product. Biomass is the central process variable because it directly indicates process performance and is included in all mathematical models describing cell growth (Surribas et al., 2006c).

For the online or at-line determination of *P. pastoris* biomass several techniques have been applied such as optical density measurements (Holmes et al., 2009; Jahic et al., 2002), near infrared spectroscopy (Crowley et al., 2005; Goldfeld et al., 2014), dielectric spectroscopy (Ehgartner et al., 2015; Fehrenbach et al., 1992), fluorescence spectroscopy (Surribas et al., 2006a,b,c), flow cytometry (Broger et al., 2011), as well as various soft sensors (Barrigón et al., 2012; Khatri and Hoffmann, 2006; Liang and Yuan, 2007; Sagmeister et al., 2013; Wechselberger et al., 2013).

One of the most promising techniques for *P. pastoris* biomass estimation is fluorometry, since many biogenic fluorophores are strongly linked to the progress of biomass development. This interconnection was used in several studies to correlate single excitation–emission wavelengths or whole 2D fluorescence spectra to yeast biomass (Horvath et al., 1993; Li and Humphrey, 1991; Scheper et al., 1984; Surribas et al., 2006a,b,c). The coenzyme nicotinamide adenine dinucleotide and its phosphorylated form, respectively, NAD(P)H (Armiger et al., 1986; Li and Humphrey, 1991; Scheper et al., 1984; Zabriskie and Humphrey, 1978) as well as the amino acid tryptophan (Horvath et al., 1993; Surribas et al., 2006c) were used for this purpose. In addition, riboflavin was used for the estimation of biomass and substrate in *P. pastoris* bioprocesses (Hisiger and Jolicoeur, 2005; Li and Humphrey, 1991; Surribas et al., 2006b). In all this studies, the potential to

predict biomass was either tested for whole 2D fluorescence spectra or single wavelengths. No study investigated the use of combined single-wavelength fluorescence measurements for biomass estimation in *P. pastoris* bioprocesses by means of chemometric tools, although this would be of great interest due to its simplicity and cost efficiency.

Differently to using the whole 2D spectrum for information extraction, biomass is in this study estimated based only on the predetermined set (Surribas et al., 2006b) of single-wavelength fluorescence measurements of tryptophan, NAD(P)H, and riboflavin. By using a small variable subset compared to a whole 2D fluorescence spectrum, data reduction and lower computational cost can be achieved. On the other hand, it is supposed that the biomass estimation based on tryptophan, NAD(P)H, and riboflavin is more reliable than a correlation between a single fluorophore's intensity alone (Hisiger and Jolicoeur, 2005) because single fluorophores can be subject to variations during the process (e.g., NAD(P)H variations due to oxygen limitation (Siano and Mutharasan, 1989)). For this reason, it is highly recommended to use multiple metabolites linked to cell growth for a reliable biomass estimation model. In this context, the proposed approach combines the high information content of 2D fluorescence spectroscopy for biomass estimation with the low model complexity and acquisition costs for single wavelength measurements.

The chemometric tools used for biomass estimation based on the fluorescence measurements were multiple linear regression (MLR), partial least squares regression (PLSR), and principal component regression (PCR). The modeling results were used to rate the wavelength pairs for the fluorophores tryptophan, NAD(P)H, and riboflavin based on their importance for biomass estimation. The share of each fluorophore in the used estimation model was quantified by means of variable importance in the projection (VIP) scores (Chong and Jun, 2005). Based on the rating of VIP scores, a simple linear correlation between the single tryptophan fluorescence and biomass was established. Finally, the prediction performance for the four different modeling approaches (MLR, PLSR, PCR, and direct tryptophan correlation model) was comparatively evaluated.

The major intention of this study was not to develop an advanced modeling approach for biomass estimation—MLR, PLSR, and PCR are rather standard methods—but to establish a scientific basis for the development of a low-cost fluorescence sensor with the same information content as 2D fluorescence spectroscopy for biomass monitoring. In this context, the applied chemometric methods give insight into the information content of each wavelength pair and allow a densification of the fluorescence data to reliable biomass information.

Materials and Methods

Strain and Culture Conditions

A single colony of *P. pastoris* type strain DSMZ 70382 grown on a YPD plate (yeast extract, $10\,g\,L^{-1}$; peptone, $20\,g\,L^{-1}$; glucose, $20\,g\,L^{-1}$; bacteriological agar, $15\,g\,L^{-1}$) was used to inoculate a 250 mL shake flask with 100 mL of the mineral medium FM22 with glycerol as carbon source: $(NH_4)_2SO_4$, $5\,g\,L^{-1}$; $CaSO_4 \cdot 2H_2O$, $1\,g\,L^{-1}$; K_2SO_4, $14.5\,g\,L^{-1}$; KH_2PO_4, $42.9\,g\,L^{-1}$; $MgSO_4 \cdot 7H_2O$, $11.7\,g\,L^{-1}$; glycerol, $40\,g\,L^{-1}$ (Stratton et al., 1998); and trace element solution, $2.0\,mL\,L^{-1}$ of the culture medium. The trace element stock solution contained: $CuSO_4 \cdot 5H_2O$, $2\,g\,L^{-1}$; KI, $0.08\,g\,L^{-1}$; $MnSO_4 \cdot H_2O$, $3\,g\,L^{-1}$; $Na_2MoO_4 \cdot 2H_2O$, $0.2\,g\,L^{-1}$; H_3BO_3, $0.02\,g\,L^{-1}$; $CaSO_4 \cdot 2H_2O$, $0.5\,g\,L^{-1}$; $CoCl_2$, $0.5\,g\,L^{-1}$; $ZnCl_2$, $7\,g\,L^{-1}$; $FeSO_4 \cdot H_2O$, $22\,g\,L^{-1}$; biotin, $0.2\,g\,L^{-1}$; conc. H_2SO_4, 1 mL. Although only a part of the supplied glycerol is necessary to reach the maximum biomass concentration in this cultivations, the standard composition of FM22 medium ($40\,g\,L^{-1}$ glycerol) was used in order to guarantee the transferability of the proposed approach to other *P. pastoris* cultivations with FM22 medium.

The shake flask culture was used to inoculate 200 μL FM22 medium in a black 96 well plate (Greiner Bio-One International GmbH, Kremsmünster, Germany). The plate was incubated with agitation at 30 °C in a Synergy™ H4 Hybrid Multi-Mode Microplate Reader (BioTek Instruments, Inc., Winooski, VT) for 50 h. Gas exchange was enabled using the gas-permeable Breathe-Easy® sealing membrane (Sigma–Aldrich Corporation, St. Louis, MO).

In this study, 36 of the 96 wells were used as blanks for turbidity and fluorescence measurements and contained only FM22 medium without cells. From the remaining 60 wells 35 were used completely for sampling (200 μL samples) resulting in a data set of 25 batches.

Biomass Determination

The goal of this study is to estimate biomass concentrations in microwell plates based on online fluorescence measurements. For creating the regression models, a reliable online reference value for biomass concentration is necessary. This online value, however, needed to be correlated to the offline biomass concentration determined as dry cell weight c_X to have a comparable value for biomass concentration that is independent of the used photometer.

Offline biomass concentration was determined as dry cell weight c_X by centrifugation of 200 μL (content of one well) of cell broth in a preweighed centrifuge tube, followed by washing the cells with phosphate buffered saline three times and drying to constant weight at 90 °C in a drying cabinet. However, due to high-relative standard deviations for dry cell weight determinations at small biomass concentrations (up to 29.2 % in the lag phase; data not shown) this data could not be used as reference value for modeling. It is for this reason that a correlation between c_X and optical density at 600 nm in offline mode ($OD_{600/off}$) was established for concentrations up to $c_X = 3\,g\,L^{-1}$ in threefold determination (1) (results see the Data Pre-Processing section).

$$c_X = d_1 \cdot OD_{600/off} \tag{1}$$

The relation (1) between c_X and $OD_{600/off}$ is linear because samples for optical density determination were diluted accordingly ($OD_{600/off}$ in the range of 0.1–0.3).

Online biomass concentration was determined by measuring optical density at 600 nm ($OD_{600/on}$) in the Synergy™ H4 Hybrid Multi-Mode Microplate Reader (BioTek Instruments, Inc., Winooski, VT). Measurement frequency was $4\,h^{-1}$. Correlations between $OD_{600/on}$ and $OD_{600/off}$ were established for optical density values up

to $OD_{600/off} = 8$ in fivefold determination (2) (results see the Data Pre-Processing section).

$$OD_{600/off} = c_2 \cdot OD^2_{600/on} + d_2 \cdot OD_{600/on} \quad (2)$$

The relation (2) between $OD_{600/off}$ and $OD_{600/on}$ is typically of higher order when diluting is not possible in online mode. The $OD_{600/on}$ measurements are in this work referred to as reference measurements for biomass concentration.

Fluorescence Measurements

Fluorescence intensities were measured online as relative fluorescence units ($RFU_{\lambda_{ex}/\lambda_{em}}$) in the Synergy™ H4 Hybrid Multi-Mode Microplate Reader (BioTek Instruments, Inc., Winooski, VT) at the excitation and emission wavelengths pairs $\lambda_{ex}/\lambda_{em}$ listed in Table I. Riboflavin has characteristic excitation maxima at $\lambda_{ex} = 370$ and 450 nm that can change with environmental conditions (Duggan et al., 1957; Li and Humphrey, 1991; Surribas et al., 2006b), which is why riboflavin was excited at this two wavelengths. The Synergy™ H4 Hybrid Multi-Mode Microplate Reader uses a xenon flash as light source for excitation and detects emission from the bottom of the microplates. Measurement frequency was $4\,h^{-1}$. A correlation (3) between $OD_{600/on}$ and $RFU_{290/350}$ was established for optical density values up to $OD_{600/off} = 8$ in fivefold determination (results see the PLSR and PCR Modeling section).

$$OD_{600/on} = d_3 \cdot RFU_{290/350} \quad (3)$$

Chemometrics

Data pre-processing, modeling, and post-processing were performed in MATLAB R2016a (The MathWorks, Inc., Natick, MA). A detailed description of the chemometric tools PLSR, PCR, and MLR as well as the underlying concepts can be found in various literature, for example, Geladi and Kowalski (1986), Marbach and Heise (1990), Rajalahti and Kvalheim (2011), Wold et al. (1987).

The raw batch data for MLR, PLSR, and PCR model generation consisted of two three-dimensional matrices for fluorescence and biomass data gathered online ($\boldsymbol{I} \times \boldsymbol{J} \times \boldsymbol{K}$, corresponding to number of batches $\boldsymbol{I}$, process variables $\boldsymbol{J}$, and sampling times $\boldsymbol{K}$). This 3D matrices were unfold lining up the batches into a 2D matrix $\boldsymbol{X}$ and the vector $\boldsymbol{y}$. $\boldsymbol{X}$ and $\boldsymbol{y}$ were mean-centered and used as independent and dependent variables for model calibration, respectively. During PLSR and PCR model calibration the preliminary models were internally validated by means of the mean squared error of prediction (MSEP) with size of test set N, predicted value $\boldsymbol{y_i}$, and reference value $\boldsymbol{y_{i/ref}}$ (4). The MSEP, sometimes denoted as mean squared error of cross-validation (MSECV), is obtained by a 10-fold cross validation (Mevik and Cederkvist, 2004) and is used to choose an appropriate number of principal components for PLSR and PCR modeling.

Table I. Excitation and emission wavelengths for the biogenic fluorophores tryptophan, NAD(P)H, and riboflavin predetermined by Surribas et al. (2006b).

Variable number	Biogenic fluorophor	Excitation wavelength λ_{ex} (nm)	Emission wavelength λ_{em} (nm)
1	Tryptophan	290	350
2	NAD(P)H	350	450
3	Riboflavin	370	530
4	Riboflavin	450	530

$$MSEP = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{y_i} - \boldsymbol{y_{i/ref}})^2 \quad (4)$$

The final models were validated by predicting the biomass of 10 equivalent cultivations. For evaluating the models, the root mean squared error of prediction (RMSEP) was calculated between measurement and prediction values (5):

$$RMSEP = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{y_i} - \boldsymbol{y_{i/ref}})^2} \quad (5)$$

In order to quantify the contribution of each of the four wavelength pairs to the information content with regard to biomass estimation, the variable importance in the projection (VIP) method was chosen. The resulting VIP scores are a prediction of the importance of a variable j in the PLS model. In contrast to variable selection by just considering the PLS weights alone, the VIP score gives a weighted sum of squares of the PLS weights by taking into account the explained variance of each principal component of the PLS model (Chong and Jun, 2005; Krause et al., 2015; Wold et al., 2001). The VIP score of the j-th variable is calculated by (6) with $\boldsymbol{b_k}$ representing the k-th element of the regression coefficients vector $\boldsymbol{b}$. The vectors $\boldsymbol{t_k}$ and $\boldsymbol{w_k}$ represent the k-th column vector of the score matrix $\boldsymbol{T}$ and weight matrix $\boldsymbol{W}$, respectively. Further, $\|\boldsymbol{w_k}\|$ is the Euclidean norm of $\boldsymbol{w_k}$, $SS(\boldsymbol{b_k t_k}) = \boldsymbol{b_k^2 t_k' t_k}$ is the percentage of $\boldsymbol{y_{i/ref}}$ explained by the k-th principal component of the PLS model, and h is the number of retained principal components, in this case $h = 2$ (Chong and Jun, 2005).

$$VIP_j = \sqrt{p\frac{\sum_{k=1}^{h}(SS(\boldsymbol{b_k t_k})(\boldsymbol{w_{jk}}/\|\boldsymbol{w_k}\|)^2)}{\sum_{k=1}^{h}SS(\boldsymbol{b_k t_k})}} \quad (6)$$

Variables with VIP scores below the predefined threshold of 1 are rated as less important because the average of squared VIP scores generally equals 1.

Results and Discussion

Small-Scale Batch Process Progression

As described in the Materials and Methods section, cultivations were conducted for 50 h and fluorescence and optical density were measured online with a frequency of $4\,h^{-1}$. The time course of the normalized mean values for fluorescence and biomass is shown in

Figure 1. The mean valued were normalized to a range of 0–1 for illustration purposes. The mean relative standard deviation for the biomass reference measurements is 4.20 % (calculated from 25 $OD_{600/on}$ data sets). The lag phase in the 200 µL scale lasts for 7 h followed by exponential growth (maximal growth rate $\mu_{max} = 0.250\,h^{-1}$) until 11–12 h. In a subsequent transition phase, which lasts until 36 h, the biomass increases to its maximum $c_{X/max} = 0.633 \pm 0.06\,g\,L^{-1}$ followed by the short stationary phase in which biomass reaches a plateau. After 42 h the biomass declines with an average decay rate of $\mu = -0.015\,h^{-1}$. Only low biomass concentrations can be obtained in this bioreactor system due to limited homogenization of the liquid phase (no stirrer), resulting in partial sinking of the cells, which gave rise to limited oxygen supply of the cells. This sinking of the cells was most likely also the reason for the sudden shifts in $OD_{600/on}$ and fluorescence measurements at cultivation time 8 h because the shaker function of the microplate reader was off for 0.5 h and thus the cells were not homogenized sufficiently in this time period. Thereby, the cells sank below the focal point for $OD_{600/on}$ and fluorescence measurement which resulted in erroneous values in this time period. This measurement error was nevertheless included in the data sets for modeling because both $OD_{600/on}$ and fluorescence measurements showed this shift and model robustness would increase by including this process disturbance.

During the time course, the biogenic fluorophores tryptophan, NAD(P)H, and riboflavin measured at their characteristic wavelengths (Table I) show a characteristic curve progression. Both NAD(P)H (ex/em 350/450 nm) and riboflavin (ex/em 370/530 nm) fluorescence intensities reach a high value during the lag phase and decline to a minimum shortly before the exponential phase ends (11 h). After 11 h the intensities for these fluorophores increase until the maximum biomass is reached, followed by a plateau phase. The intensity for riboflavin (ex/em 450/530 nm) fluorescence increases during the cultivation with a change of slope at about 11 h. The time course for tryptophan (ex/em 290/350 nm) follows the one for optical density closely, indicating a tight correlation between this amino acid's development and biomass generation. The mean relative standard deviations for the fluorescence measurements (calculated from 25 $RFU_{\lambda_{ex}/\lambda_{em}}$ data sets) are as follows: 3.31 % for tryptophan (ex/em 290/350 nm), 7.73 % for NAD(P)H (ex/em 350/450 nm), 8.26 %, and 5.56 % for riboflavin (ex/em 370/530 and 450/530 nm, respectively). Referring to the measurement deviation for the $OD_{600/on}$ and $RFU_{\lambda_{ex}/\lambda_{em}}$ data sets, the cultivations as well as the measurements are reproducible but still contain enough variability for model calibration and validation.

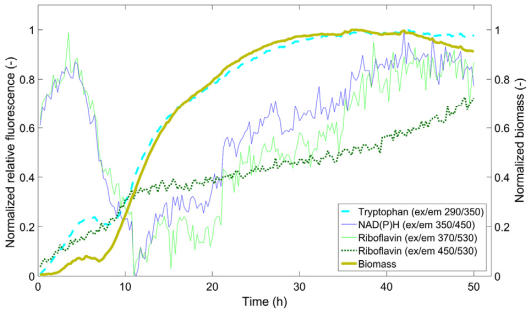


Figure 1. Time course of normalized fluorescence signals (tryptophan, NAD(P)H, and riboflavin) and biomass. Lines represent mean values for 25 batches.

Data Pre-Processing

As described in the Biomass Determination section, due to the high-relative standard deviation for dry cell weight determinations at small biomass concentrations (up to 29.2 % in the lag phase) the online biomass $OD_{600/on}$ was correlated to the optical density in offline mode $OD_{600/off}$ and the offline biomass concentration c_X, respectively, via (1) and (2). The correlation factors $d_1 = 0.280\,g\,L^{-1}$ (threefold determination; $R^2 = 0.982$; data not shown), $c_2 = 3.106$, and $d_2 = 2.204$ (fivefold determination; $R^2 = 0.999$; data not shown) were obtained.

For creating the regression models the 25 batches were separated randomly into 15 batches for model calibration and 10 batches for model validation. The data were further processed by mean-centering and matrix unfolding.

MLR Modeling

After data pre-processing, the training set for model calibration consisted of the independent variable matrix $\boldsymbol{X}$ (fluorescence intensity of the four fluorophores) and dependent vector $\boldsymbol{y}$ (optical density in online mode). The dependent vector $\boldsymbol{y}$ was correlated to the reference value c_X via (1) and (2), as described above.

Interactions between the variables in $\boldsymbol{X}$ can be considered or not during model generation (Preacher et al., 2006). In this study, both MLR approaches were evaluated, resulting in $\boldsymbol{X}$ and an extension of $\boldsymbol{X}$ by multiplicative combinations of the fluorescence intensities.

The final MLR models were validated externally by predicting the biomass for 10 validation cultivations. Determination coefficients $R^2 = 0.971$ for the MLR model with interactions and $R^2 = 0.967$ without interactions were calculated. The RMSEP (5) with and without interaction terms, respectively, between the predicted and measured values for the validation batches was 0.022 and 0.014 for the optical density measurements in online mode, corresponding to 0.025 and 0.017 $g\,L^{-1}$ dry cell weight. Compared to the maximum biomass concentration $c_{X/max} = 0.633\,g\,L^{-1}$ this resulted in a relative estimation error of 3.9 % and 2.6 %, respectively.

In the present study, the MLR without interaction terms provides better results than the MLR with interaction terms. It is supposed that including the interactive relations in the MLR model leads to overfitting, that is, new data cannot be predicted sufficiently. This statement is substantiated by the higher determination coefficient when using interaction terms.

The MLR model without interactions was further investigated for the relevance of each input variable on the estimation accuracy by the backward elimination method based on the correlation coefficients (Pires et al., 2008). A low-determination coefficient of a MLR model without a certain input variable indicates the importance of this variable. The resulting determination coefficient R^2 for the MLR model without tryptophan (ex/em 290/350 nm) was

0.506, while the determination coefficients of models without NAD(P)H (ex/em 350/450 nm), riboflavin (ex/em 370/530 nm), and riboflavin (ex/em 450/530 nm) were 0.967, 0.960, and 0.966, respectively. This indicates tryptophan as the most important variable for biomass estimation and an equal distribution of the other three variables.

Biomass could not be estimated accurately when not all four wavelength pairs were used for MLR modeling. The resulting relative estimation errors compared to the maximum biomass concentration $c_{X/\max} = 0.633\ \mathrm{g\,L^{-1}}$ were significantly higher than for a MLR model with all four variables: 15.3 %, 53.7 %, 48.9 %, and 54.6 % for omitted tryptophan (ex/em 290/350 nm), NAD(P)H (ex/em 350/450 nm), riboflavin (ex/em 370/530 nm), and riboflavin (ex/em 450/530 nm), respectively.

PLSR and PCR Modeling

PLSR and PCR were used as alternative to MLR to model the relationships between $\boldsymbol{y}$ and $\boldsymbol{X}$. The data were pre-processed and the regression models were validated as described before.

As a first step of PLSR and PCR model creation the number of principal components was determined because using all available components may be more than will be needed to adequately fit the data. A quick way to choose the number of components is to plot the percent of variance explained in the independent variables $\boldsymbol{X}$ as a function of the number of components, as done in Figure 2A for the PLSR. The first PLS component explains 48.05 % of the variance in $\boldsymbol{X}$. Adding a second PLS component to the model raises the explained variance to 96.96 %. For the PCR model the percent variance explained in X is 63.00 % and 96.99 % for two and three principal components, respectively. The uniformly higher curve for PCR compared to PLSR is due to the different ways of these methods in constructing the components. PLSR and PCR construct components to best explain variation in $\boldsymbol{y}$ and in $\boldsymbol{X}$, respectively. Two components were chosen as adequately because the addition of a third component only gives a minor increase to 99.23 % for both modeling approaches at the cost of model complexity. Another reason to not use a too large number of components is that this strategy leads to overfitting, that is, the model fits the data too well and does not generalize well to other data.

The chosen number of components was further confirmed by plotting the estimated mean squared prediction error (4) obtained by a 10-fold cross-validation (Mevik and Cederkvist, 2004) against the number of components used for model creation, as done in Figure 2B. The mean squared prediction error with using two components was ~0.001 both for PLSR and PCR, corresponding to a relative error of 0.8 % with respect to the maximum optical density in online mode $\mathrm{OD}_{600/\mathrm{on}} = 0.1220$. This internal validation of the model by cross-validation showed that the prediction error does not significantly decrease by using three components, whereas for using only one component it is over double the value as for two components for PLSR and even 13-fold for PCR.

The next step was to detect outliers and remove them from the training set by plotting the fitted response against the observed response, as done in Figure 2C. The figure further provides an evaluation of model accuracy for both modeling approaches. Both

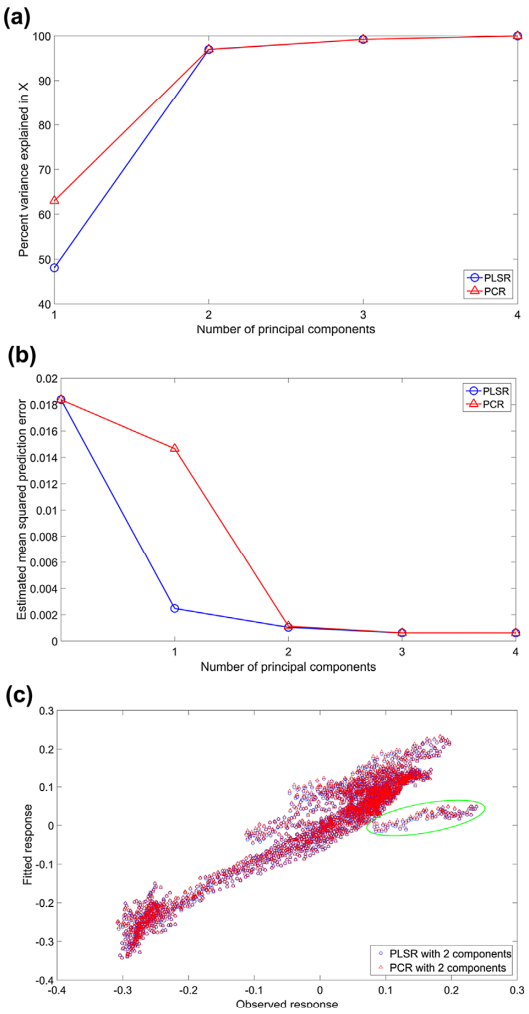


Figure 2. (**a**) Variance explained in X and (**b**) estimated mean squared prediction error ($\mathrm{OD}_{600/\mathrm{on}}$ data) for number of principal components. (**c**) Fitted versus observed response (calibration results) for PLSR and PCR with two components (outliers marked green).

PLSR and PCR show a fairly accurate fitting of the response variable $\boldsymbol{y}$.

Finally, the PLS weights were plotted against the dependent variables contained in $\boldsymbol{X}$ in order to describe the dependency of each component in the PLSR model on the original variable. Figure 3A shows the PLS weight for the variables that refer to the fluorescence intensities of tryptophan (ex/em 290/350 nm), NAD(P)H (ex/em 350/450 nm), and riboflavin (ex/em 370/530 and 450/530 nm). Analogically, the PCA loadings show the dependency of the components generated by the PCR model on the variables contained in $\boldsymbol{X}$, as shown in Figure 3B.

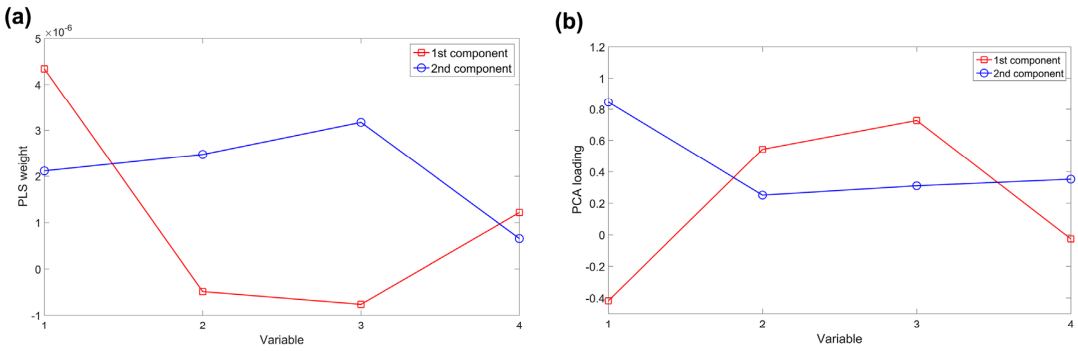


Figure 3. (**a**) PLS weights and (**b**) PCA loadings for fluorescence variables (1) tryptophan (ex/em 290/350 nm), (2) NAD(P)H (ex/em 350/450 nm) as well as (3) and (4) riboflavin (ex/em 370/530 and 450/530 nm, respectively).

The final PLSR and PCR models were validated externally by predicting the biomass for 10 validation cultivations. Determination coefficients $R^2 = 0.967$ for the PLSR model and $R^2 = 0.938$ for the PCR model were calculated. The RMSEP (5) for PLSR and PCR, respectively, between the predicted and measured values for the validation batches was 0.018 and 0.019 for the optical density measurements in online mode, corresponding to 0.023 and 0.025 g L^{-1} dry cell weight. Compared to the maximum biomass concentration $c_{X/\max} = 0.633$ g L^{-1} this resulted in a relative estimation error of 3.6 % and 3.9 %, respectively.

In the present study, the PLSR provides a slightly better prediction of biomass than the PCR. PLSR is supposed to have better prediction performance than PCR, since PLSR uses $\boldsymbol{y}$ in addition to $\mathbf{X}$ to determine the principal components. In the following, the PLSR model was further investigated on the contribution of each variable to the model in order to investigate if a model with only the most important variable fits well.

Variable Selection via Variable Importance in the Projection Scores

One major goal of this study is to obtain a reliable model with only a few fluorescence variables compared to, for example, 2D fluorescence spectroscopy with hundreds of variables. In this context, it is furthermore advantageous to quantify the contribution of each of the four wavelength pairs to the information content with regard to biomass estimation. This contribution to the model was quantified by means of the variable importance in the projection (VIP) scores, as described in the Chemometrics section. The resulting VIP scores are shown in Figure 4. The VIP score for tryptophan (ex/em 290/350 nm) is by far the highest with 1.828, followed by 0.513, 0.509, and 0.370 for riboflavin (ex/em 450/530 nm), riboflavin (ex/em 370/530 nm), and NAD(P)H (ex/em 350/450 nm), respectively. This distribution of VIP scores is supported by the statement of Horvath et al. (1993) that tryptophan outweighs NADH in the importance for the estimation of biomass in yeast culture processes. Further, the high importance of tryptophan in the PLSR model indicated by its VIP score is in accordance with the results of the backward elimination method for MLR modeling (MLR Modeling section) and the similar progressions of biomass and fluorescence signal of tryptophan shown in Figure 1.

It was further investigated whether the single tryptophan fluorescence is sufficient for reliably predicting the biomass in the present small-scale batch process. For this purpose, the correlation (3) was set up between $OD_{600/on}$ and the relative fluorescence units for tryptophan (ex/em 290/350 nm) $RFU_{290/350}$ with the correlation factor d_3 (fivefold determination; $R^2 = 0.993$; data not shown). A determination coefficient $R^2 = 0.993$ was calculated.

The RMSEP between the predicted and measured values for the validation batches was 0.036 for the optical density measurements in online mode, corresponding to 0.049 g L^{-1} dry cell weight. Compared to the maximum biomass concentration ($c_{X/\max} = 0.633$ g L^{-1}) this resulted in a relative estimation error of 7.7 %.

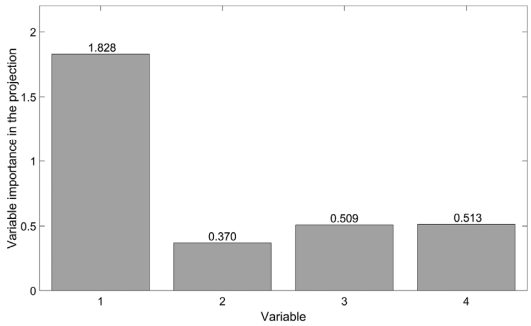


Figure 4. Variable importance in the projection (VIP) for variables (1) tryptophan (ex/em 290/350 nm), (2) NAD(P)H (ex/em 350/450 nm) as well as (3) and (4) riboflavin (ex/em 370/530 and 450/530 nm, respectively).

Comparative Evaluation of Prediction Performance

The results of biomass predictions with the four different modeling approaches (MLR, PLSR, PCR, and the single tryptophan fluorescence model) are summarized in Table II and Figure 5. Figure 5 shows the fluorescence model predictions and the estimations based on optical density measurements (referred to as reference measurements) for biomass concentration for a representative batch. As can be seen, the estimation of MLR fits the reference measurement well for the whole cultivation time, whereas PLSR and PCR fits the reference measurement well until the beginning of the decline phase at 42 h cultivation time. In the decline phase, the predicted value for biomass concentration falsely increases further for PLSR and PCR estimates. The MLR approach, thus, provides the best estimates in this case because even the decrease of cell concentration in the decline phase is modeled and the estimation error is lowest (Table II).

The stationary and decline phases are characterized by an increase in cell lysis rate, which is believed to cause release of proteases (Sinha et al., 2004). According to Surribas et al. (2006b), the resulting proteolysis is one reason for the discrepancy between model estimation and measurements in the stationary and decline phase for PLSR. A more specific reason for the low model accuracy (PLSR and PCR) after the exponential phase could be that the riboflavin (ex/em 450/530 nm) intensity seems to rise independently of the beginning of the stationary and decline phase, while the other three variables used for model generation reach a plateau value (see Fig. 1). Even a slight increase in the rate of riboflavin (ex/em 450/530 nm) development can be observed with the beginning of the decline phase that could be ascribed to the release of riboflavin during cell lysis.

Although MLR results in the best estimates according to the lowest RMSEP and the best fit over the whole cultivation time, it could have some disadvantages compared to PLSR and PCR. When multicollinearity is present in the data, the MLR model becomes highly sensitive to outliers and noise. This would be a drawback for the use in online monitoring applications, because data have to be filtered for outliers before modeling. PLSR and PCR, however, reduce collinearity before the actual regression step and are less sensitive to small errors in the $\boldsymbol{X}$ and $\boldsymbol{y}$ data (Rajalahti and Kvalheim, 2011).

The modeling accuracy of the single tryptophan fluorescence model is lower compared to the MLR, PLSR, and PCR models, as can be seen in Table II. The predicted biomass differs from the reference measured biomass by twice the estimation error (RMSEP)

Table II. Comparison of different modeling approaches by their regression coefficients R^2, root mean squared error of prediction (RMSEP), and the relative estimation error with respect to the maximum biomass concentration ($c_{X/\max} = 0.633\ \mathrm{g\,L^{-1}}$).

Modeling approach	R^2	RMSEP ($\mathrm{g\,L^{-1}}$)	Relative estimation error (%)
MLR (without interactions)	0.967	0.017	2.6
PLSR (2 PCs)	0.967	0.023	3.6
PCR (2 PCs)	0.938	0.025	3.9
$RFU_{290/350}$	0.993	0.049	7.7

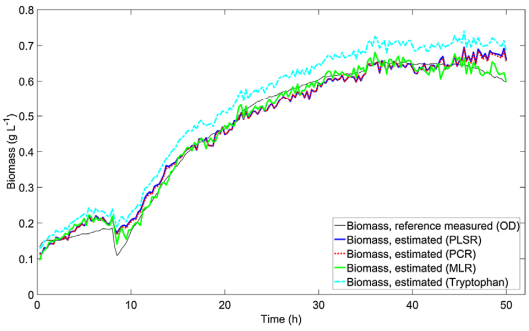


Figure 5. Time course of reference measured (OD estimation) and predicted biomass for one representative batch with different estimator models.

when NAD(P)H and riboflavin are not taken into account in the model (see also Fig. 5). It thus stands to reason that biomass estimation via combined fluorometric measurements and chemometric modeling is more reliable than a correlation between a single fluorophore's intensity alone.

Conclusions

In this work, the key variable biomass was successfully predicted for small-scale *P. pastoris* batch cultures based on the measurements of fluorescence intensity at four characteristic single-wavelength pairs (excitation/emission) and their combination by means of chemometric tools. These wavelength pairs corresponded to the biogenic fluorophores tryptophan, NAD(P)H, and riboflavin, with detection wavelengths identified before by Surribas et al. (2006b). Differently to these authors' approach of taking into account a whole 2D fluorescence spectrum for modeling, biomass in this work was estimated based on the combined measurement of three fluorophores at four single-wavelength pairs.

The three modeling approaches MLR, PLSR, and PCR were comparatively evaluated on their prediction performance with a slightly better model accuracy for MLR compared to PLSR and PCR. The importance of the four model input variables was rated by means of the VIP scores. Tryptophan (ex/em 290/350 nm) was found to be the most important fluorophore for biomass estimation in this study. Further, biomass was successfully predicted via a correlation between single tryptophan fluorescence and optical density measurements in online mode; however, model accuracy was lower compared to MLR, PLSR, and PCR.

In this study, 96 well plates were chosen as scale for cultivations in order to obtain a large data set under identical conditions (temperature, mixing properties, and oxygen supply) and enable simultaneous online monitoring of optical density and fluorescence intensity in a microplate reader. Only low biomass concentrations can be obtained in this bioreactor system due to limited oxygen supply and homogenization of the liquid phase. However, the presented approach seems to be promising for higher biomass concentrations ($>20\ \mathrm{g\,L^{-1}}$) because the underlying principle is still the same as for the established 2D fluorescence spectroscopy (Surribas et al., 2006a).

The approach's applicability for higher biomass concentrations and under large-scale conditions has to be examined in future with a 2D fluorescence spectrometer or a fluorescence sensor measuring only at the specified wavelength pairs.

Also, the effects of the fluorescent properties of other *P. pastoris* media than FM22 (e.g., BSM, BMGY, and YPD media) have to be investigated. It is supposed that the presented measurement principle performs well with almost colorless and defined media due to the absence of fluorescence interferences in the region of interest (Marose et al., 1998; Rhee et al., 2006) and that the effort to calibrate the model to such media is minor. In contrast, complex and colorful media are likely to lead to spectral overlaps (Hisiger and Jolicoeur, 2005) that would increase the effort for model calibration and potentially hinder successful biomass estimation. Hisiger and Jolicoeur (2005) reported on spectral overlaps of serum and phenol-red with NAD(P)H, riboflavin, and tryptophan when cultivating a mouse myeloma cell line (NS0). Their results indicate that biomass estimation based on the progress of this fluorophores is possible; however, the culture medium optimization is required. Contrary to yeast cells, mammalian cells cannot synthesize tryptophan and riboflavin de novo and thus are added to the medium, which is why their progress during cultivation will differ from that of yeasts. It is supposed that also fungal cell biomass can be monitored because 2D fluorescence spectroscopy was successfully applied for *Claviceps purpurea* biomass estimation (Boehl et al., 2003). In general, the proposed approach is more sensitive to spectral overlaps from media compounds than methods that use whole spectra (2D fluorescence, UV–VIS, NIR, MIR, or Raman spectroscopy) because the fluorescence signal cannot be compensated by signals at other wavelengths but the four specified ones.

Compared to turbidity measurements for biomass monitoring, the proposed approach simple sensor designs choose between range and precision.

The results show that biomass can successfully be predicted based on combined single-wavelength fluorescence measurements. The use of multiple metabolites linked to cell growth results in a more reliable biomass estimation model compared to using a single fluorophores' intensity alone. Based on this finding, the development of an alternative measurement system for data-intensive and expensive 2D fluorescence spectroscopy seems promising. The proposed measurement system would use LEDs (light emitting diodes) to excite only at a small set of wavelengths and the resulting spectrum would contain sufficient information for successful biomass estimation. The measurement system would consist of a fluorescence measurement device in addition to a turbidity measurement device (OD_{600} or near-infrared region). The biomass estimation would this way become more reliable because two different measurement principles are used. By comparing the results of biomass estimation via fluorometric measurements and PLSR and PCR modeling with the turbidity signal the beginning of cell lysis in the decline phase and the culture's metabolic activity could potentially be predicted. This information could then be used for process state estimation (end of exponential phase, beginning of decline phase) or the determination of optimal harvesting time for protease sensitive recombinant proteins.

To sum up, this study shows the high potential of biogenic fluorophores for biomass estimation. The proposed approach for online biomass estimation can be used as basis for the development of a fluorometric biomass sensor that offers similar information content for bioprocess monitoring and control purposes compared to expensive 2D fluorescence spectroscopy.

This work was supported by the German Federal Ministry of Education and Research (project 031A616D).

References

Armiger W, Forro J, Montalvo L, Lee J, Zabriskie D. 1986. The interpretation of on-line process measurements of intracellular NADH in fermentation processes. Chem Eng Commun 45:197–206.

Barrigón JM, Ramon R, Rocha I, Valero F, Ferreira EC, Montesinos JL. 2012. State and specific growth estimation in heterologous protein production by *Pichia pastoris*. AIChE J 58:2966–2979.

Boehl D, Solle D, Hitzmann B, Scheper T. 2003. Chemometric modelling with two-dimensional fluorescence data for *Claviceps purpurea* bioprocess characterization. J Biotechnol 105:179–188.

Broger T, Odermatt RP, Huber P, Sonnleitner B. 2011. Real-time on-line flow cytometry for bioprocess monitoring. J Biotechnol 154:240–247.

Cereghino GPL, Cereghino JL, Ilgen C, Cregg JM. 2002. Production of recombinant proteins in fermenter cultures of the yeast *Pichia pastoris*. Curr Opin Biotechnol 13:329–332.

Chong I-G, Jun C-H. 2005. Performance of some variable selection methods when multicollinearity is present. Chemom Intell Lab Sys 78:103–112.

Cos O, Ramón R, Montesinos JL, Valero F. 2006. Operational strategies, monitoring and control of heterologous protein production in the methylotrophic yeast *Pichia pastoris* under different promoters: A review. Microb Cell Fact 5:17.

Cregg JM, Cereghino JL, Shi J, Higgins DR. 2000. Recombinant protein expression in *Pichia pastoris*. Mol Biotechnol 16:23–52.

Crowley J, Arnold SA, Wood N, Harvey LM, McNeil B. 2005. Monitoring a high cell density recombinant *Pichia pastoris* fed-batch bioprocess using transmission and reflectance near infrared spectroscopy. Enzyme Microb Technol 36:621–628.

Duggan DE, Bowman RL, Brodie BB, Udenfriend S. 1957. A spectrophotofluorometric study of compounds of biological interest. Arch Biochem Biophys 68:1–14.

Ehgartner D, Sagmeister P, Herwig C, Wechselberger P. 2015. A novel real-time method to estimate volumetric mass biodensity based on the combination of dielectric spectroscopy and soft-sensors. J Chem Technol Biotechnol 90:262–272.

Fehrenbach R, Comberbach M, Petre J. 1992. On-line biomass monitoring by capacitance measurement. J Biotechnol 23:303–314.

Geladi P, Kowalski BR. 1986. Partial least-squares regression: A tutorial. Anal Chim Acta 185:1–17.

Goldfeld M, Christensen J, Pollard D, Gibson ER, Olesberg JT, Koerperick EJ, Lanz K, Small GW, Arnold MA, Evans CE. 2014. Advanced near-infrared monitor for stable real-time measurement and control of *Pichia pastoris* bioprocesses. Biotechnol Progr 30:749–759.

Hisiger S, Jolicoeur M. 2005. A multiwavelength fluorescence probe: Is one probe capable for on-line monitoring of recombinant protein production and biomass activity? J Biotechnol 117:325–336.

Holmes WJ, Darby RA, Wilks MD, Smith R, Bill RM. 2009. Developing a scalable model of recombinant protein yield from *Pichia pastoris*: The influence of culture conditions, biomass and induction regime. Microb Cell Fact 8:35.

Horvath JJ, Glazier SA, Spangler CJ. 1993. In situ fluorescence cell mass measurements of *Saccharomyces cerevisiae* using cellular tryptophan. Biotechnol Progr 9:666–670.

Jahic M, Rotticci-Mulder J, Martinelle M, Hult K, Enfors S-O. 2002. Modeling of growth and energy metabolism of *Pichia pastoris* producing a fusion protein. Bioprocess Biosys Eng 24:385–393.

Khatri NK, Hoffmann F. 2006. Impact of methanol concentration on secreted protein production in oxygen-limited cultures of recombinant *Pichia pastoris*. Biotechnol Bioeng 93:871–879.

Krause D, Hussein M, Becker T. 2015. Online monitoring of bioprocesses via multivariate sensor prediction within swarm intelligence decision making. Chemom Intell Lab Syst 145:48–59.

Li JK, Humphrey AE. 1991. Use of fluorometry for monitoring and control of a bioreactor. Biotechnol Bioeng 37:1043–1049.

Liang J, Yuan J. 2007. Oxygen transfer model in recombinant *Pichia pastoris* and its application in biomass estimation. Biotechnol Lett 29:27–35.

Marbach R, Heise H. 1990. Calibration modeling by partial least-squares and principal component regression and its optimization using an improved leverage correction for prediction testing. Chemom Intell Lab Sys 9:45–63.

Marose S, Lindemann C, Scheper T. 1998. Two-dimensional fluorescence spectroscopy: A new tool for on-line bioprocess monitoring. Biotechnol Progr 14:63–74.

Mevik BH, Cederkvist HR. 2004. Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). J Chemom 18:422–429.

Pires J, Martins F, Sousa S, Alvim-Ferraz M, Pereira M. 2008. Selection and validation of parameters in multiple linear and principal component regressions. Environ Model Software 23:50–55.

Potvin G, Ahmad A, Zhang Z. 2012. Bioprocess engineering aspects of heterologous protein production in *Pichia pastoris*: A review. Biochem Eng J 64:91–105.

Preacher KJ, Curran PJ, Bauer DJ. 2006. Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. J Educ Behav Stat 31:437–448.

Rajalahti T, Kvalheim OM. 2011. Multivariate data analysis in pharmaceutics: A tutorial review. Int J Pharm 417:280–290.

Rhee JI, Kang T-H, Lee K-I, Sohn O-J, Kim S-Y, Chung S-W. 2006. Application of principal component analysis and self-organizing map to the analysis of 2D fluorescence spectra and the monitoring of fermentation processes. Biotechnol Bioprocess Eng 11:432–441.

Sagmeister P, Wechselberger P, Jazini M, Meitz A, Langemann T, Herwig C. 2013. Soft sensor assisted dynamic bioprocess control: Efficient tools for bioprocess development. Chem Eng Sci 96:190–198.

Scheper T, Gebauer A, Sauerbrei A, Niehoff A, Schügerl K. 1984. Measurement of biological parameters during fermentation processes. Anal Chim Acta 163:111–118.

Siano S, Mutharasan R. 1989. NADH and flavin fluorescence responses of starved yeast cultures to substrate additions. Biotechnol Bioeng 34:660–670.

Sinha J, Plantz BA, Inan M, Meagher MM. 2005. Causes of proteolytic degradation of secreted recombinant proteins produced in methylotrophic yeast *Pichia pastoris*: Case study with recombinant ovine interferon-τ. Biotechnol Bioeng 89:102–112.

Stratton J, Chiruvolu V, Meagher M. 1998. High cell-density fermentation. In: Higgins DR, Cregg JM, editors. Pichia protocols. Totowa, NJ: Humana Press. p 107–120.

Surribas A, Amigo JM, Coello J, Montesinos JL, Valero F, Maspoch S. 2006a. Parallel factor analysis combined with PLS regression applied to the on-line monitoring of *Pichia pastoris* cultures. Anal Bioanal Chem 385:1281–1288.

Surribas A, Geissler D, Gierse A, Scheper T, Hitzmann B, Montesinos JL, Valero F. 2006b. State variables monitoring by in situ multi-wavelength fluorescence spectroscopy in heterologous protein production by *Pichia pastoris*. J Biotechnol 124:412–419.

Surribas A, Montesinos JL, Valero F. 2006c. Biomass estimation using fluorescence measurements in *Pichia pastoris* bioprocess. J Chem Technol Biotechnol 81:23–28.

Wechselberger P, Sagmeister P, Herwig C. 2013. Real-time estimation of biomass and specific growth rate in physiologically variable recombinant fed-batch processes. Bioprocess Biosys Eng 36:1205–1218.

Wold S, Esbensen K, Geladi P. 1987. Principal component analysis. Chemom Intell Lab Sys 2:37–52.

Wold S, Sjöström M, Eriksson L. 2001. PLS-regression: A basic tool of chemometrics. Chemom Intell Lab Sys 58:109–130.

Zabriskie D, Humphrey A. 1978. Estimation of fermentation biomass concentration by measuring culture fluorescence. Appl Environ Microbiol 35:337–343.

3.3 Biomass soft sensor for a *Pichia pastoris* fed-batch process based on phase detection and hybrid modeling

Received: 15 January 2020 | Revised: 5 May 2020 | Accepted: 5 June 2020

DOI: 10.1002/bit.27454

ARTICLE

BIOTECHNOLOGY AND BIOENGINEERING WILEY

Biomass soft sensor for a *Pichia pastoris* fed-batch process based on phase detection and hybrid modeling

Vincent Brunner | **Manuel Siegl** | **Dominik Geier** | **Thomas Becker**

Chair of Brewing and Beverage Technology, Technical University of Munich, Freising, Germany

Correspondence
Dominik Geier, Chair of Brewing and Beverage Technology, Technical University of Munich, Weihenstephaner Steig 20, 85354 Freising, Germany.
Email: dominik.geier@tum.de

Funding information
Bundesministerium für Bildung und Forschung, Grant/Award Number: 031B0475E

Abstract

A common control strategy for the production of recombinant proteins in *Pichia pastoris* using the alcohol oxidase 1 (AOX1) promotor is to separate the bioprocess into two main phases: biomass generation on glycerol and protein production via methanol induction. This study reports the establishment of a soft sensor for the prediction of biomass concentration that adapts automatically to these distinct phases. A hybrid approach combining mechanistic (carbon balance) and data-driven modeling (multiple linear regression) is used for this purpose. The model parameters are dynamically adapted according to the current process phase using a multilevel phase detection algorithm. This algorithm is based on the online data of CO_2 in the off-gas (absolute value and first derivative) and cumulative base feed. The evaluation of the model resulted in a mean relative prediction error of 5.52% and R^2 of .96 for the entire process. The resulting model was implemented as a soft sensor for the online monitoring of the *P. pastoris* bioprocess. The soft sensor can be used for quality control and as input to process control systems, for example, for methanol control.

KEYWORDS

biomass, hybrid model, phase detection, *Pichia pastoris*, soft sensor

1 | INTRODUCTION

The methylotrophic yeast *Pichia pastoris* (now reclassified as *Komagataella phaffii*) is frequently used as a host for expressing heterologous proteins for both basic research and industrial production (Cereghino, Cereghino, Ilgen, & Cregg, 2002). When the methanol-inducible alcohol oxidase 1 (AOX1) promotor is used for controlling protein expression, the process is typically separated into two main phases with different objectives. In the first phase, the carbon source—typically glycerol—is converted to biomass. It aims to produce large amounts of biomass before methanol induction. This phase is often referred to as the glycerol or biomass phase and can optionally be extended by a glycerol feed to accumulate more biomass before methanol induction (Gao et al., 2012; Jahic, Veide, Charoenrat, Teeri, & Enfors, 2006). The second phase, also referred to as the induction or methanol phase, starts when methanol is added to the medium to induce protein expression via the genetically modified AOX1 promotor. This phase aims to reproducibly generate the highest product titers.

Besides product titer, biomass concentration can be seen as one of the most critical quality attributes in upstream bioprocessing due to its effect on all other quality attributes, which holds true for *P. pastoris* bioprocesses in both the glycerol and the methanol phases (J. Harms, Wang, Kim, Yang, & Rathore, 2008). Several techniques such as turbidimetry, infrared or fluorescence spectroscopy, and flow cytometry are available for monitoring biomass, as reviewed by P. Harms, Kostov, and Rao (2002), Luttmann et al. (2012), and Schügerl (2001). However, the use of online measurement systems for monitoring biomass in a technical context is still often problematic. The reasons for this include lack of reliability, the considerable dependence on the process and product matrix (isolated solutions), and high standards of operation and

maintenance (Kano & Fujiwara, 2012). For these reasons, biomass is in many cases not measured online at all.

Because the direct measurement of biomass is often not feasible, soft sensors can be used for predicting it. Soft sensors consist of computational models or algorithms that allow the prediction of target values, such as biomass concentration, via continuously measured secondary variables, such as exhaust gas concentrations, dissolved oxygen (DO), and flow rates (Luttmann et al., 2012).

Various modeling techniques have been proposed for developing soft sensors, the majority of which are based on mechanistic or data-driven approaches. An overview of soft sensors and the selection of appropriate modeling techniques for online bioreactor state estimation has been presented elsewhere (Zhang, 2009). Mechanistic modeling approaches include, for example, differential balancing systems, which describe the material and energy conversions at the cellular level, as well as mass and energy balances (Jenzsch, Gnoth, Kleinschmidt, Simutis, & Lübbert, 2007). Data-driven approaches include, among others, artificial neural networks (ANN; Gonzaga, Meleiro, Kiang, & Maciel Filho, 2009) and methods from the field of multivariate statistical process control (Kadlec, Gabrys, & Strandt, 2009), such as principal component regression and partial least squares regression. In hybrid modeling, mechanistic and data-driven modeling approaches are combined, as reviewed by Kalos, Kordon, Smits, and Werkmeister (2003) and Solle et al. (2017).

The main challenges in the development of soft sensors are as follows: control of model complexity (overfitting vs. underfitting) (Kordon, Smits, Kalos, & Jordaan, 2003); limited amount of data sets or data points (Fortuna, Graziani, & Xibilia, 2009); outliers resulting from, for example, sensor faults (Zhang, 2009); adaption mechanisms for model maintenance (Bakirov, Gabrys, & Fay, 2017); input variable selection; reliability of soft sensors; and changes in process characteristics and operating conditions (Kano & Fujiwara, 2012). In addition, a specific challenge arises in soft sensor development for *P. pastoris* bioprocesses given its distinct process phases, as described previously: The underlying principles of prediction models for biomass are related to the inherent biological relationships between online measured variables and biomass (Chen, Nguang, Li, & Chen, 2004); thus, the soft sensor needs to be adaptive to the current process phase to give accurate prediction results throughout the entire process.

In this study, an adaptive soft sensor for biomass concentration was developed. The novelty of this study is that the soft sensor changes its model coefficients regarding the current process phase (batch, transition, or fed-batch phase) of the *P. pastoris* bioprocess. The soft sensor's underlying prediction model is based on a hybrid of mechanistic and data-driven approaches. The mechanistic part comprises mass balancing of carbon using methanol and CO_2 fluxes. The outcome of this mechanistic model—the generation rate of total organic carbon inside the bioreactor—is fed into a data-driven model that in turn leads to an online prediction of biomass concentration. The adaptability of the soft sensor to the distinct process phases is guaranteed by automatic and reliable detection of glycerol depletion based on online process variables, namely, CO_2 in the off-gas (absolute value and first derivative) and cumulative base feed. The soft sensor's model coefficients switch automatically depending on the current process phase and thus give accurate biomass predictions throughout the entire process. Finally, the soft sensor was implemented in a real-time capable system to enable online biomass monitoring.

2 | MATERIALS AND METHODS

2.1 | Strain and preculture conditions

The inoculum of a recombinant *P. pastoris* strain based on type strain DSMZ 70382 was prepared in three 150 ml shake flasks containing 50 ml of the mineral medium FM22 with glycerol as the carbon source: $(NH_4)_2SO_4$, 5 g · L^{-1}; $CaSO_4 \cdot 2H_2O$, 1 g · L^{-1}; K_2SO_4, 14.3 g · L^{-1}; KH_2PO_4, 42.9 g · L^{-1}; $MgSO_4 \cdot 7H_2O$, 11.7 g · L^{-1}; glycerol, 40 g · L^{-1} (Stratton, Chiruvolu, & Meagher, 1998); and trace element stock solution (PTM4), 2.0 ml · L^{-1} of the culture medium. The PTM4 stock solution contained $CuSO_4 \cdot 5H_2O$, 2 g · L^{-1}; KI, 0.08 g · L^{-1}; $MnSO_4 \cdot H_2O$, 3 g · L^{-1}; $Na_2MoO_4 \cdot 2H_2O$, 0.2 g · L^{-1}; H_3BO_3, 0.02 g · L^{-1}; $CaSO_4 \cdot 2H_2O$, 0.5 g · L^{-1}; $CoCl_2$, 0.5 g · L^{-1}; $ZnCl_2$, 7 g · L^{-1}; $FeSO_4 \cdot H_2O$, 22 g · L^{-1}; biotin, 0.2 g · L^{-1}; and conc. H_2SO_4, 1 ml. Cells were grown for 70 hr at 30°C on a shaker at 150 min^{-1}.

2.2 | Fed-batch cultivation in bioreactor

The shake flask culture was used to inoculate the main culture in the bioreactor Biostat® Cplus (Sartorius AG, Goettingen, Germany) with working and total volumes of 15 and 42 L, respectively. The main culture medium was FM22. Pressure, pH, temperature, and dissolved oxygen were controlled to 500 mbar, 5, 30°C, and 40%, respectively. NH_4OH was used as nitrogen source and to set and maintain a pH of 5. A dissolved oxygen minimum of 40% was controlled by a cascade control using variable stirrer speed (300–600 min^{-1}) and air flow rate (20–40 L · min^{-1}).

The end of the batch phase, that is, the depletion of glycerol, was indicated online by a characteristic peak in the off-gas CO_2 concentration. The complete depletion of glycerol was verified offline via HPLC analysis (data not shown). After a short transition phase, which prevents the potential repression of the AOX1 promotor by glycerol residues from the preceding batch phase, the culture was induced with methanol. The methanol feed was supplemented with 12 ml · L^{-1} PTM4 stock solution. Methanol concentration was controlled via a fuzzy logic controller to 4.5 g · L^{-1}. This controller uses methanol concentration as the main input and the feed rate of methanol as output. The general concept of fuzzy logic controllers is described, for example, in Birle, Hussein, and Becker (2013).

The off-gas CO_2 concentration was measured with a BlueInOne Cell sensor (BlueSens gas sensors GmbH, Herten, Germany). Methanol concentration was measured with an inline Alcosens sensor (Heinrich Frings GmbH & Co. KG, Rheinbach, Germany).

2.3 | Determination of dry cell weight

Dry cell weight was determined in triplicate by centrifugation of 2 ml cell suspension in previously weighed centrifuge tubes, followed by discarding the supernatant and drying the cell pellet to a constant weight at 80°C. Samples for the determination of dry cell weight were taken using the BaychroMAT® autosampler (Bayer AG, Leverkusen, Germany) with a minimum sampling interval of 2 hr.

2.4 | Data management

The digital control unit (DCU) of the Biostat® bioreactor (Sartorius AG) was used for primary process control (pressure, pH, temperature, and dissolved oxygen) and signal recording. SIMATIC SIPAT (Siemens AG, Munich, Germany) was used for data management and to store the process (online) and laboratory (offline) data in a central database with a recording interval of 30 s. Offline data preprocessing and modeling were performed in MATLAB R2019b (The MathWorks, Inc., Natick, MA); signal processing, real-time prediction of the target quantity, biomass concentration, by means of the developed soft sensor as well as model-based control via a fuzzy logic controller were performed in SIMULINK R2019b (The MathWorks, Inc.). An interface capable of real-time communication between the DCU, the data management system (SIMATIC SIPAT), and the online modeling software (SIMULINK) was realized via a Sartorius OPC DA server (Sartorius AG).

3 | RESULTS AND DISCUSSION

This study aims to develop a soft sensor for the prediction of biomass concentration that provides accurate online predictions for a multiphase process (batch, transition, and fed-batch phase) with two different carbon sources (glycerol and methanol). The general concept of the hybrid-model-based soft sensor presented here consists of two main levels: The first level comprises a phase detection algorithm to differentiate online among batch, transition, and fed-batch phase; the second level consists of a hybrid-model-based prediction equation that automatically adjusts the model parameters based on the current process phase (batch, transition, or fed-batch phase). For the development of the first and second levels, nine and six data sets, respectively, were used. Only the latter six data sets had a fed-batch phase with control of methanol concentration and therefore can be compared with each other.

The hybrid model uses a carbon balance as the mechanistic part. The result of the carbon balance is fed into a data-driven part to provide accurate prediction of the biomass concentration. The information-bearing model inputs that were used in this study to predict biomass concentration are cumulative methanol and base feed as well as concentrations of off-gas CO_2 and methanol.

Figure 1 shows the time course of the relevant model inputs of the soft sensor for an exemplary process run. This process run is used as an illustrative example throughout the following sections. In this case, the batch phase ends at 39.6 hr, followed by a transition phase that lasts for 6.9 hr, and a fed-batch phase that starts at 46.5 hr. In the batch phase, glycerol is metabolized and biomass is generated. The presence of the transition phase prevents the potential repression of the AOX1 promotor by glycerol residues from the preceding batch phase. In the transition phase, no significant increase (due to the absence of carbon sources) or decrease of biomass concentration was observable. In the fed-batch phase, methanol is fed into the bioreactor via a pump for the first time. Subsequently, methanol concentration is controlled to a setpoint of 4.5 $g \cdot L^{-1}$ via a fuzzy logic controller. This process run shows control errors such as high initial overshoot and an increasing deviation of the measured methanol concentration to the setpoint in the subsequent time course. Base (NH_4OH, 5 M) is fed into the bioreactor via a pump and is used to maintain pH at 5.0. The cumulative base feed represents the degree of metabolic activity, that is, substrate depletion. This variable shows

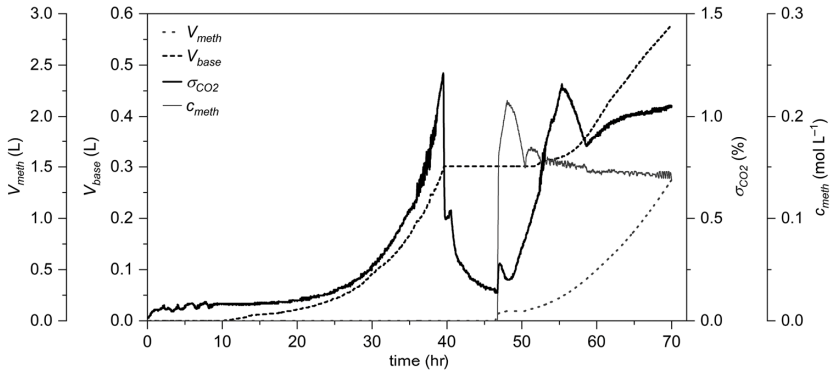


FIGURE 1 Time course of the relevant model inputs of the soft sensor for an exemplary process run, namely, cumulative feed volume of methanol, V_{meth}, and base, V_{base}, as well as concentrations of CO_2 in the off-gas, σ_{CO2}, and methanol, c_{meth}. For this exemplary process, the batch phase ends at 39.6 hr and the fed-batch phase starts at 46.5 hr

high collinearity to the biomass concentration (see later in Figure 6). In the batch phase, the off-gas CO_2 signal almost continuously increases until the end of this phase. Here, the signal drops abruptly and, except for minor fluctuations, begins to rise again only upon methanol induction. After methanol induction, the cells need to adapt to the metabolization of methanol.

3.1 | Multilevel process phase detection

3.1.1 | General concept of process phase detection

This algorithm step aims to differentiate among the three distinct process phases, which are listed in Table 1 together with its process data characteristics regarding process phase detection. The detection of the end of the batch phase is primarily based on the off-gas CO_2 signal. The metabolization of glycerol together with an increasing cell concentration leads to an almost continuous increase in CO_2 emission during the batch phase. When glycerol is depleted, the off-gas CO_2 signal drops abruptly, as shown in Figure 1 (here at 39.6 hr). The relationship between the CO_2 drop and substrate consumption is shown and discussed in detail in Munch et al. (2020). This abrupt drop is the main sign of the end of the batch phase and is hereinafter referred to as trigger 3. To increase robustness of the phase detection algorithm, two additional trigger conditions upstream of trigger 3 were implemented, namely the exceeding of absolute values for cumulative base feed (trigger 1) and off-gas CO_2 concentration (trigger 2).

The output of the algorithm for process phase detection is a binary value indicating whether the end of the batch phase has been reached (1 = true) or not (0 = false) together with the corresponding timestamp. Variable inputs to the algorithm consist of the signals for cumulative base feed (V_{base}) for trigger 1, the absolute off-gas CO_2 concentration (σ_{CO2}) for trigger 2, and the timewise derivative of the off-gas CO_2 concentration ($d\sigma_{CO2}/dt$) for trigger 3. Only when triggers 1 and 2 are initiated, that is, they are "true", trigger 3 is active and can be initiated. The end of the batch phase is indicated when all three triggers are "true."

The process variable V_{base} represents the cumulative metabolic activity regarding the consumption of the carbon source. Because the batch process starts with a glycerol concentration of 40 $g \cdot L^{-1}$, the total volume of base fed into the bioreactor at the end of the batch phase is restricted to the stoichiometry of glycerol metabolization. Trigger 1 is therefore initiated when a defined threshold for V_{base} is exceeded. In the transition phase, the variable V_{base} remains constant because cells do not grow. Similar to V_{base}, the process variable σ_{CO2} is strongly related to biomass growth and substrate consumption. During exponential growth, σ_{CO2} increases almost continuously until the end of the batch phase. Trigger 2 is therefore initiated when a defined threshold for σ_{CO2} is exceeded. This trigger is implemented to guarantee that natural fluctuations in σ_{CO2}, which can statistically occur in biological systems (see Figure 1), and sensor faults impede the functionality of the process phase detection as little as possible. Trigger 2 thus slightly increases robustness of the phase detection algorithm. Figure 2 shows the functioning of trigger 3 in terms of the time course of $d\sigma_{CO2}/dt$ for an exemplary process run. The value of $d\sigma_{CO2}/dt$ falls below the threshold uniquely at the end of the batch phase (here at 39.6 hr). A median filtering step was implemented before and after the derivation step to decrease noise of the variables σ_{CO2} and $d\sigma_{CO2}/dt$, respectively.

3.1.2 | Threshold definition

The thresholds for triggers 1, 2, and 3 were calculated as shown in (1), where $threshold_i$ is the threshold for the trigger variable used for process phase detection with $i = \{V_{base}, \sigma_{CO2}, d\sigma_{CO2}/dt\}$; $mean_i$ and std_i is the arithmetic mean and standard deviation, respectively, of the variable i at the end of the batch phase. The end of the batch phase was for this purpose defined as the time at the minimum of $d\sigma_{CO2}/dt$. SF is a constant safety factor of 3 that is implemented to avoid false positive detections of the end of the batch phase of the phase detection and thus to increase the robustness of the multilevel detection algorithm.

$$threshold_i = mean_i - std_i\, SF. \tag{1}$$

For illustration and comparison of the three triggers, Figure 3 shows the results (mean ± standard deviation) for the trigger variables normalized to the corresponding threshold. The resulting threshold values together with the mean and standard deviation are summarized in Table 2. These threshold values were implemented in SIMULINK to automatically detect the end of the batch phase and therefore to select the right model coefficients for the biomass soft sensor shown in the following.

3.2 | Mass balance for carbon

The underlying principle of the mechanistic modeling part is mass balancing of carbon. The boundary for the balancing is the bioreactor system: Carbon is fed into the bioreactor in the form of methanol (fed-batch phase) and leaves the boundary in the form of CO_2.

TABLE 1 Main characteristics of the three distinct process phases (batch, transition, and fed-batch phase) regarding process phase detection

Process phase	Main process objective	Carbon source	Main process data characteristic
Batch phase	Biomass generation	Glycerol	Abrupt drop in off-gas CO_2 signal at the end of batch phase
Transition phase	Derepression of the AOX1 promotor	None	No base feed due to absent cell growth
Fed-batch phase	Product formation	Methanol	Starting with methanol feed

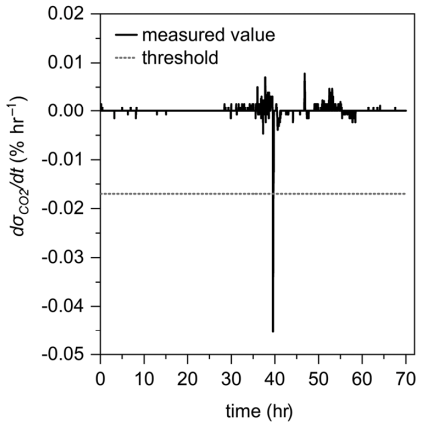


FIGURE 2 Timewise derivative of the off-gas CO_2 sensor reading, $d\sigma_{CO2}/dt$, for an exemplary process run. A median filter (window size = ten sensor readings) is implemented before and after the derivation step to handle noisy sensor readings. The characteristic negative peak (here at 39.6 hr) is the main indicator for the depletion of the batch phase substrate (glycerol) and thus the end of the batch phase. This landmark is used to initiate the start of the transition and fed-batch phase, respectively

The remaining carbon is in the form of either glycerol or methanol or is bound in cells as well as extracellular organic acids and proteins. The following sections show how the timewise rates of off-gas CO_2 and methanol are calculated. These rates are then balanced to enable

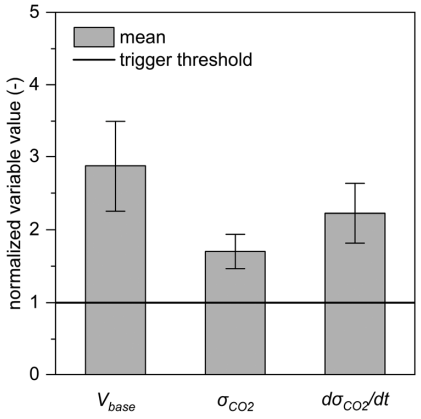


FIGURE 3 Triggers for the multilevel detection of the end of the batch phase (i.e., depletion of glycerol). Only when defined values for base, V_{base}, and absolute off-gas CO_2 concentration, σ_{CO2}, are reached for the first time, the last trigger—the timewise derivative of the off-gas CO_2 concentration, $d\sigma_{CO2}/dt$—is active. The three thresholds are defined based on the calculation of the mean and standard deviation for each of the three variables at the end of the batch phase as well as a safety factor. The diagram shows normalized absolute variable values; error bars correspond to the normalized standard deviation (n = 9)

TABLE 2 Threshold, mean, and standard deviation for the three trigger variables V_{base}, σ_{CO2}, and $d\sigma_{CO2}/dt$ at the end of the batch phase (n = 9)

Trigger number	Variable	Mean	Standard deviation	Threshold
1	V_{base}	540 ml	117 ml	188 ml
2	σ_{CO2}	1.284%	0.177%	0.755%
3	$d\sigma_{CO2}/dt$	−0.038% · hr^{-1}	0.007% · hr^{-1}	−0.017% · hr^{-1}

calculation of the formation rate of total organic carbon (TOC) that remains bound in cells as well as extracellular organic acids and proteins. To determine the cumulative amount of TOC online, this rate needs to be multiplied by the total liquid volume and numerically integrated. This cumulative amount of TOC is used in the subsequent data-driven modeling part to predict biomass concentration (Figure 4).

3.2.1 | Calculation of liquid volume

To calculate the total liquid volume, all feeds and removals (sampling) need to be considered. The total reactor volume V_{total} is calculated as in (2), where V_{start} is the start volume after inoculation; V_{base}, V_{meth}, and V_{afoam} are the cumulative volumes of base, methanol, and antifoam, respectively, fed into the bioreactor; $V_{samples}$ is the cumulative volume of samples automatically taken via the BaychroMAT® autosampler:

$$V_{total} = V_{start} + V_{base} + V_{meth} + V_{afoam} - V_{samples}. \quad (2)$$

3.2.2 | Calculation of carbon dioxide emission rate

The calculation of the carbon dioxide emission rate r_{CO2} in (3) is adapted from Takors (2013), where Q_{air} is the air flow rate, p is the pressure, R is the universal gas constant (8.314×10^{-2} L · bar · mol^{-1} · K^{-1}), T is the temperature, σ_{CO2} and σ_{O2} are the concentrations of carbon dioxide and oxygen, respectively, and the indices α and ω represent the gas inlet and outlet of the bioreactor, respectively:

$$r_{CO2} = \frac{Q_{air}\, p}{V_{total}\, RT}\left(\frac{1 - \sigma_{O2_\alpha} - \sigma_{CO2_\alpha}}{1 - \sigma_{O2_\omega} - \sigma_{CO2_\omega}}\sigma_{CO2_\omega} - \sigma_{CO2_\alpha}\right). \quad (3)$$

3.2.3 | Calculation of methanol reaction rate

As described above, errors in the methanol control, such as an initial overshoot or a deviation of the measured methanol concentration to the setpoint (Figure 1), can occur. The carbon balance is designed to compensate for disturbances of methanol control by incorporating the methanol accumulation rate $r_{meth,acc}$. Changes in $r_{meth,acc}$ result from the uptake of methanol by cells and methanol feeding (especially at the feed start when the methanol setpoint is reached for the first time). $r_{meth,acc}$ is determined by the timewise derivative of the

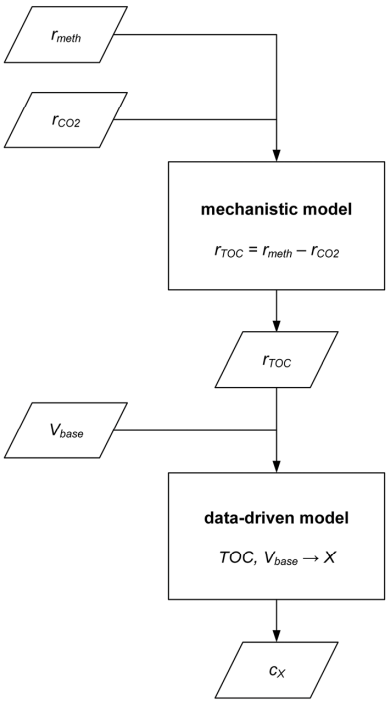


FIGURE 4 Simplified representation of the hybrid-model-based soft sensor for biomass concentration c_X. The methanol reaction rate r_{meth} and the carbon dioxide emission rate r_{CO2} are fed to the mechanistic model; carbon balancing is here used to calculate the formation rate of total organic carbon, r_{TOC}. The subsequent data-driven model uses the numerical integration of r_{TOC}, namely, *TOC*, together with the cumulative base feed, V_{base}, as inputs to calculate the amount of biomass *X*. Finally, *X* is divided by the total liquid volume inside the bioreactor, V_{total}, to calculate the biomass concentration c_X. Both the data-driven and the mechanistic parts can be carried out online

methanol concentration c_{meth} that is measured in the bioreactor (in-line), as follows:

$$r_{meth,acc} = \frac{dc_{meth}}{dt}. \tag{4}$$

The methanol reaction rate r_{meth} describes the net rate at which methanol accumulates in or is withdrawn from the liquid phase of the bioreactor and is calculated as follows, where $r_{meth,in}$ is the feed rate of methanol into the bioreactor related to the total liquid volume V_{total}:

$$r_{meth} = r_{meth,in} - r_{meth,acc}. \tag{5}$$

3.2.4 | Calculation of formation rate of total organic carbon

TOC refers to all carbon inside the bioreactor system that is bound in the substrate (glycerol or methanol) and cells as well as extracellular organic acids and proteins. The formation rate of TOC, r_{TOC}, is not directly measured by reference analysis but calculated as follows by balancing the methanol reaction rate r_{meth} and the carbon dioxide emission rate r_{CO2}:

$$r_{TOC} = r_{meth} - r_{CO2}. \tag{6}$$

In the batch phase $r_{meth} = 0$ and no glycerol is fed into the bioreactor; therefore, the carbon balance in this phase is $r_{TOC} = -r_{CO2}$.

3.3 | Development of hybrid-model-based soft sensor

3.3.1 | Combination of mechanistic and data-driven parts in a hybrid model

Figure 4 shows the soft sensor algorithm and how process variables are passed through the mechanistic and data-driven modeling parts to finally result in the online prediction of biomass concentration c_X. The output of the mechanistic part (mass balance for carbon), r_{TOC}, is together with V_{base} fed into the data-driven part. The data-driven part comprises a numerical integration step for r_{TOC} to obtain the cumulative amount of total organic carbon, *TOC*, and a multiple linear regression (MLR) step. MLR was chosen as regression method because the prediction model uses only the two inputs *TOC* and V_{base}.

Using *TOC* only for biomass prediction leads to acceptable prediction results (data not shown). However, the concentrations of dissolved carbon dioxide (H_2CO_3) as well as extracellular proteins (c_P) and organic acids, which can in most cases not be measured online, distort the biomass prediction. The prediction model for biomass is therefore complemented by adding information about acids in the medium. The process variable with most information about acids in the medium is the cumulative base feed, V_{base}. Because $c_P \ll c_X$, the extracellular protein concentration is neglected for biomass prediction.

TOC is calculated as follows by multiplication with V_{total} and numeric integration from the beginning of the process run (t_0) to the current time (t):

$$TOC = \int_{t_0}^{t} r_{TOC}\, V_{total}\, dt. \tag{7}$$

When in sum (up to t) more carbon passed the bioreactor boundary to the outside than to the inside, *TOC* has a negative value. The time course of *TOC* is together with r_{meth} and r_{CO2} illustrated in Figure 5 for an exemplary process run. In the batch phase (Figure 5a), the only carbon passing through the bioreactor boundary is CO_2. Therefore, *TOC* has a negative gradient. In the fed-batch phase (Figure 5b), methanol is fed to the bioreactor, resulting in a net positive gradient for *TOC*.

The soft sensor uses three distinct sets of model coefficients for each the batch, transition, and fed-batch phase. For model calibration

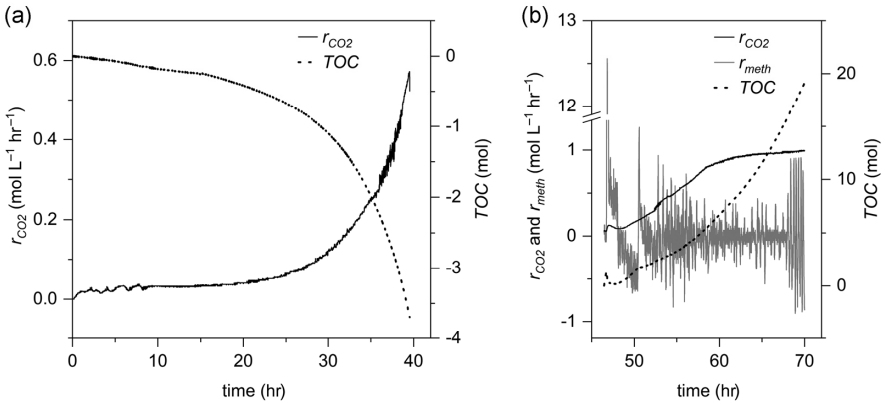


FIGURE 5 Illustration of the carbon balance for (a) the batch and (b) fed-batch phase for an exemplary process run. The carbon dioxide emission rate, r_{CO2}, and—in the fed-batch phase, additionally—the methanol reaction rate, r_{meth}, are used to calculate the formation rate of total organic carbon, r_{TOC}, as in (6). Multiplication of r_{TOC} with V_{total} and numeric integration as in (7) result in the cumulative amount of total organic carbon, *TOC*

via MLR in the batch phase, *TOC* and V_{base} are used as inputs and the biomass amount X (determined offline as dry cell weight) as output. The prediction equation is formulated as follows, where b_0, b_1, and b_2 are the model coefficients:

$$X = b_0 + b_1\,TOC + b_2\,V_{base}. \tag{8}$$

In the transition phase, no significant cell growth or decline was observed, so b_0 was set to the value of X at the end of the batch phase ($X_{batchend}$) and b_1 and b_2 were set to 0. In the fed-batch phase, b_0 was set to $X_{batchend}$ and b_1 and b_2 were determined analogously to the methods used in the batch phase.

The regression step in (8) is related to the total liquid volume inside the bioreactor, V_{total}. To determine the biomass concentration c_X, the biomass amount X is divided by V_{total}, as in the following equation:

$$c_X = \frac{X}{V_{total}}. \tag{9}$$

3.3.2 | Cross-validation approach for model calibration and validation

The model was calibrated and validated by a batch-wise cross-validation approach. The six data sets used for developing the biomass soft sensor were iteratively partitioned into two-thirds of calibration and one-third of validation data sets. This resulted in a total of 6!/(2!4!) = 15 different combinations of complementary subsets for cross-validation. For each iteration step, R^2, root mean squared error (*RMSE*), and normalized root mean squared error (*NRMSE*) of cross-validation were calculated separately for the batch and fed-batch phase as well as for the entire process (including the transition phase). R^2 is calculated for the four calibration data sets. The *NRMSE* in the following equation is the normalized version of *RMSE* and is calculated from reference measurements y and predictions $\hat{y}$ of the two validation data sets. y_{max} and y_{min} are the maximum and minimum values of y, respectively, and m is the number of data points in y:

$$NRMSE = \frac{1}{y_{max} - y_{min}}\sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2}. \tag{10}$$

The use of separate subsets for internal and external (holdout) validation (OECD, 2014) does not appear to be practicable because

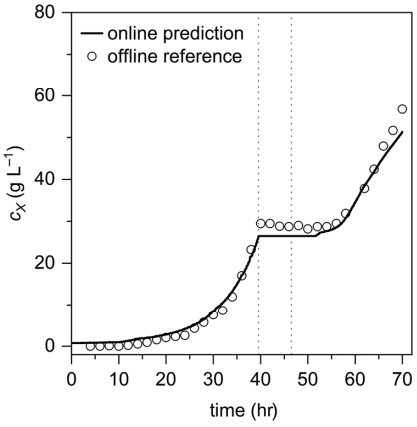


FIGURE 6 Online prediction of biomass concentration c_X during batch, transition, and fed-batch phase for an exemplary process run using the hybrid-model-based soft sensor. Both the batch and the fed-batch phase start with a lag phase after which cells grow exponentially (batch phase) or linearly (fed-batch phase). The two dashed, gray lines indicate the switches from batch to transition phase (39.6 hr) and from transition to fed-batch phase (46.5 hr), respectively

Process phase	$b_0 \pm CI_{.95,b_0}$ (g)	$b_1 \pm CI_{.95,b_1}$ (g·mol^{-1})	$b_2 \pm CI_{.95,b_2}$ (g·L^{-1})
Batch phase	4.60 ± 2.54	−13.69 ± 9.81	701.96 ± 79.62
Transition phase	Replaced by $X_{batchend}$	0	0
Fed-batch phase	Replaced by $X_{batchend}$	2.63 ± 1.57	1074.05 ± 94.28

TABLE 3 Results for model coefficients b_0, b_1, and b_2 in (8) and the corresponding 95% confidence intervals, $CI_{.95}$. In the transition and fed-batch phase, b_0 is set to the value of X at the end of the batch phase ($X_{batchend}$)

the total number of data sets that are available for model calibration and validation is too small (n = 6).

3.4 | Online prediction of biomass using the multilevel phase detection

The multilevel phase detection algorithm resulted in a 100% correct hit rate for the detection of the end of the batch phase. On average, the phase end was detected 2.56 measurements (corresponding to 77 s) before the minimum $d\sigma_{CO2}/dt$ was reached—which was defined as the end of the batch phase.

The arithmetic means for R^2, *RMSE*, and *NRMSE* are calculated using the abovementioned 15 combinations of n = 6 data sets. The mean R^2 for the batch and fed-batch phases is .97 and .95, respectively; the mean R^2 for the entire process is .96. The mean *RMSE* for the batch and fed-batch phase is 1.14 and 5.05 g · L^{-1}, respectively; the mean *RMSE* for the entire process is 3.57 g · L^{-1}, which results in a mean *NRMSE* of 5.52%.

Figure 6 shows the results for the online prediction of biomass concentration based on the hybrid-model-based soft sensor. The figure shows validation data of one iteration of the cross-validation for an exemplary process run. The underestimation of the online prediction at 40–52 hr and after 64 hr is due to an error in biomass prediction at the end of the batch phase that entails prediction errors in the transition phase.

The results for the model coefficients b_0, b_1, and b_2 in (8) are listed in Table 3. As described above, these model coefficients are used to determine the biomass amount X, which needs to be divided by V_{total} to calculate the biomass concentration c_X. V_{total} varies between $V_{total} = V_{start}$ = 10.00 L and on average V_{total} = 12.56 L (n = 6) at the end of the cultivation. In the batch phase, the intercept b_0 describes the initial biomass from inoculation. As mentioned above, b_0 was in the transition and fed-batch phase replaced by $X_{batchend}$, which has a mean of 253.47 g (n = 6). The model coefficient for *TOC*, b_1, is negative in the batch phase because here the carbon balance in (6) is simplified to $r_{TOC} = -r_{CO2}$ (boundary for the balancing is the bioreactor system) and thus *TOC* in (7) decreases with increasing CO_2 emission and biomass, respectively. In the fed-batch phase, in which methanol is fed to the bioreactor, *TOC* correlates positively with X. The model coefficient for V_{base}, b_2, is positive for both the batch and fed-batch phase. In the fed-batch phase, b_2 is more than 50% higher than in the batch phase, which means that more than 50% base is necessary to maintain the pH setpoint on glycerol compared to methanol. The soft sensor's model coefficients switch automatically depending on the current process phase. The differences in the model coefficients b_1 and b_2 between the individual process phases indicate the necessity for the adaption of model coefficients with changing process phases.

The accuracy of the estimates of the model coefficients is given by the corresponding 95% confidence intervals, $CI_{.95}$ (Table 3). None of the $CI_{.95}$ contains the value zero, which is considered to be a primary indication that the model inputs are to a certain degree significant to the model output, biomass. The width of $CI_{.95}$ relative to the absolute value of the model coefficient is a further indicator for the quality of the regression and hence for the uncertainty of the soft sensor model (Fernandes et al., 2012). For b_0, the ratio of the width of $CI_{.95}$ to the absolute value of the model coefficient is 55%; for b_1, the ratio is 72% and 60% for the batch and fed-batch phase, respectively; for b_2, the ratio is 11% and 9% for the batch and fed-batch phase, respectively.

The contribution of the model coefficients b_0, b_1, and b_2 to the prediction of X is illustrated in Figure 7 for an exemplary process run. Here, each model coefficient's contribution was determined by disassembling the linear combination in (8) and dividing each model coefficient's prediction by the total model prediction. As expected, the contribution of b_0 starts with an initial value of 100% at the process start and decreases relative to the contribution increases of b_1 and b_2. Until the end of the batch and fed-batch phase, the

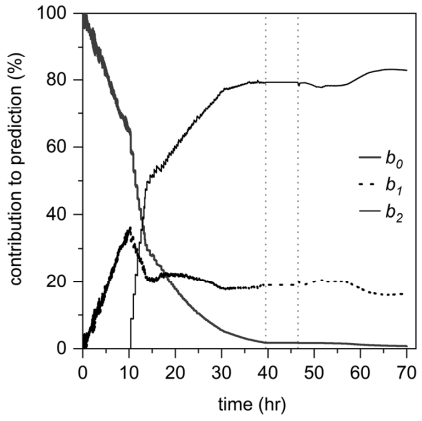


FIGURE 7 Contribution of model coefficients b_0, b_1, and b_2 to the prediction of biomass amount X during batch, transition, and fed-batch phase for an exemplary process run. The two dashed, gray lines indicate the switches from batch to transition phase (39.6 hr) and from transition to fed-batch phase (46.5 hr), respectively. The soft sensor updates its model coefficients automatically for the three distinct process phases

contribution of b_0 falls to values of 1.72% and 0.57%, respectively. Since b_0 is in the transition and fed-batch phase replaced by $X_{batchend}$, the contributions to $X_{batchend}$ (18.99% for b_1 and 79.29% for b_2) are used as offset for the contributions of b_1 and b_2 throughout the latter process phases. The contribution of b_1 initially rises to a maximum of 36.41% approximately at the end of the lag phase and reaches contributions of 18.99% and 18.26%, respectively, at the end of the batch and fed-batch phase. The contribution of b_2 starts to rise when base is first fed to the bioreactor (see Figure 1) and reaches values of 79.29% and 81.18%, respectively, at the end of the batch and fed-batch phase. It can be concluded from these results that, approximately after the end of the lag phase, V_{base} has a higher impact on biomass prediction than *TOC*. This result is consistent with the apparent high collinearity of V_{base} (see Figure 1) and c_X (see Figure 6).

4 | CONCLUSIONS

As mentioned at the beginning of this paper, several challenges can arise when attempting to develop soft sensors. One of these is specific to *P. pastoris* bioprocesses with distinct process phases such as batch, transition, and fed-batch phase. The underlying principles of prediction models for biomass are related to the inherent biological relations (Chen et al., 2004), which differ depending on the substrate used in the specific process phase. The fundamental differences in the metabolism of different carbon sources have a visible impact on CO_2 emission and the consumption of pH correction agent (see Figure 1), which are two of the main model inputs used in this study. For multiphase processes with more than one substrate, this means that the probability of finding a single model that captures the information necessary for prediction of biomass is rather low.

This study demonstrates the application of a multilevel phase detection algorithm to determine the end of the batch phase (glycerol depletion) online. In every tested case, the algorithm provided the correct end time of the batch phase. The detection of this end time was used to trigger the transition phase and the subsequent methanol induction. The knowledge about the significantly reduced CO_2 emission that comes with glycerol depletion was effectively utilized. Specifically, the stoichiometric restrictions concerning the cumulative amount of supplied base (trigger 1) and emitted CO_2 (trigger 2) were used to increase robustness of the third trigger (timewise derivative of the off-gas CO_2 signal). The usage of purely data-driven approaches for process phase detection (e.g., Abonyi, Feil, Nemeth, & Arva, 2005; Ye, Wang, & Yang, 2017) did not appear practicable in this case because only a relatively small number of data sets (nine) were available for the development of the phase detection algorithm.

The output of the phase detection algorithm was used to switch the parameters of the prediction model online. The prediction model was calibrated offline using a hybrid-model-based approach. The output of the mechanistic part (carbon balance) is fed to the data-driven part (MLR) to provide an accurate prediction of the biomass concentration. The process runs were conducted under the same operating conditions (initial glycerol concentration, constant setpoints for methanol, pH, dissolved oxygen, temperature, and pressure). However, the process runs and corresponding data sets used in this study were subject to variance of initial biomass concentration, which in turn resulted from the variability of the preculture. Further, errors in the methanol control, such as an initial overshoot or a deviation of the measured methanol concentration to the setpoint (Figure 1), occurred and additionally increased the variance between the data sets. Despite this variance between the used data sets, model evaluation results in a mean relative prediction error of 5.52% and R^2 of .96 for the entire process. These two evaluation criteria are of similar magnitude to those of other biomass soft sensors for *P. pastoris* fed-batch processes (Beiroti, Aghasadeghi, Hosseini, & Norouzian, 2019; Crowley, Arnold, Wood, Harvey, & McNeil, 2005; Fazenda et al., 2013; Surribas, Geissler, et al., 2006; Surribas, Montesinos, & Valero, 2006). In the approach presented here, however, the soft sensor is adaptable online to the different process phases, and no cost-intensive spectroscopic measurement system is necessary. The robustness of the soft sensor with regard to different process conditions (e.g., variation of methanol, pH, and temperature setpoint) was not in the scope of this study. These investigations are subject of future research.

The main constraint of the presented soft sensor is that the prediction in the transition and fed-batch phase is directly dependent on the prediction result in the batch phase. This is due to the passing on of the biomass prediction at the end of the batch phase ($X_{batchend}$) as a start value for the prediction models of the subsequent phases. The effect of error propagation can be visualized by considering the slight decrease of R^2 and increase of prediction error between batch and fed-batch phase. It should further be noted that the carbon balance in the individual phases depends on constant ratios of biomass formation, CO_2 emission, and—in the fed-batch phase—methanol metabolization. Longer periods of substrate limitation or metabolite inhibition would impede an accurate biomass prediction if these scenarios are not included in the data sets used for model calibration.

Knowledge-based relationships were combined with data-driven methodology in this study. No general statement can be made here about whether mechanistic, data-driven, or hybrid approaches are superior because the choice is strongly dependent on the available process knowledge and measurement systems (offline/online) as well as the number of data sets and data points (Solle et al., 2017). However, in this study, the usage of a hybrid approach appears to be suitable because of the benefits from both components of it. This is due to the availability of the necessary online measurement systems for capturing the information relevant for modeling biomass (off-gas CO_2, methanol, cumulative base feed) and, on the other hand, the relatively small number of data sets (six) for model calibration and validation.

The transferability of the developed phase-dependent soft sensor to other fed-batch cultivations with different *P. pastoris* strains, control strategies, media, and process parameters must be investigated in future research. It is supposed that the presented

approaches for process phase detection and hybrid-model-based prediction are transferable to any methanol-induced *P. pastoris* process provided that the carbon source used for initially generating biomass (glycerol, glucose, etc.) is not co-fed to methanol.

The developed algorithm for process phase detection and the prediction model were implemented as a soft sensor for the online monitoring of biomass. The soft sensor can be used for quality control and as input to the process control system, for example, for methanol control.

NOMENCLATURE

b_0, b_1, b_2	model coefficients (g, $g \cdot mol^{-1}$, $g \cdot L^{-1}$)
c	molar or mass concentration
$CI_{.95}$	95% confidence interval for model coefficients (g, $g \cdot mol^{-1}$, $g \cdot L^{-1}$)
c_{meth}	methanol concentration ($mol \cdot L^{-1}$)
c_P	extracellular protein concentration ($g \cdot L^{-1}$)
c_X	biomass concentration ($g \cdot L^{-1}$)
m	number of data points in y (-)
mean	arithmetic mean of trigger variable (L or % or $\% \cdot hr^{-1}$)
n	number of data sets (-)
NRMSE	normalized root mean squared error (%)
p	pressure (bar)
Q_{air}	air flow rate ($L \cdot hr^{-1}$)
r	timewise rate
R	universal gas constant ($L \cdot bar \cdot mol^{-1} \cdot K^{-1}$)
R^2	coefficient of determination (-)
RMSE	root mean squared error ($g \cdot L^{-1}$)
r_{CO2}	carbon dioxide emission rate ($mol \cdot L^{-1} \cdot hr^{-1}$)
r_{meth}	methanol reaction rate ($mol \cdot L^{-1} \cdot hr^{-1}$)
$r_{meth,acc}$	methanol accumulation rate ($mol \cdot L^{-1} \cdot hr^{-1}$)
$r_{meth,in}$	methanol feed rate ($mol \cdot L^{-1} \cdot hr^{-1}$)
r_{TOC}	formation rate of total organic carbon ($mol \cdot L^{-1} \cdot hr^{-1}$)
SF	constant safety factor (-)
std	standard deviation of trigger variable (L or % or $\% \cdot hr^{-1}$)
T	temperature (K)
t	time (hr)
threshold	threshold for trigger variable (L or % or $\% \cdot hr^{-1}$)
TOC	cumulative amount of total organic carbon (mol)
V_{afoam}	cumulative volume of antifoam (L)
V_{base}	cumulative volume of base (L)
V_{meth}	cumulative volume of methanol (L)
$V_{samples}$	cumulative volume of samples (L)
V_{start}	start liquid volume inside bioreactor after inoculation (L)
V_{total}	total liquid volume inside bioreactor (L)
X	biomass amount (g)
$X_{batchend}$	biomass amount at the end of the batch phase (g)
y	reference measurement ($g \cdot L^{-1}$)
$\hat{y}$	prediction ($g \cdot L^{-1}$)
y_{max}/y_{min}	maximum/minimum values of reference measurements y ($g \cdot L^{-1}$)
α	index for gas inlet of the bioreactor (-)
σ	volume concentration
σ_{CO2}	off-gas CO_2 concentration (%)
$d\sigma_{CO2}/dt$	timewise derivative of the off-gas CO_2 concentration ($\% \cdot hr^{-1}$)
σ_{O2}	off-gas O_2 concentration (%)
ω	index for gas outlet of the bioreactor (-)

ACKNOWLEDGMENTS

This study was supported by the German Federal Ministry of Education and Research (project 031B0475E). The authors would like to thank Jens Traenkle's group (Bayer AG) for providing the Baychromat® autosampler.

ORCID

Vincent Brunner http://orcid.org/0000-0002-3310-2236

REFERENCES

Abonyi, J., Feil, B., Nemeth, S., & Arva, P. (2005). Modified Gath–Geva clustering for fuzzy segmentation of multivariate time-series. *Fuzzy Sets and Systems*, *149*, 39–56.

Bakirov, R., Gabrys, B., & Fay, D. (2017). Multiple adaptive mechanisms for data-driven soft sensors. *Computers & Chemical Engineering*, *96*, 42–54.

Beiroti, A., Aghasadeghi, M. R., Hosseini, S. N., & Norouzian, D. (2019). Application of recurrent neural network for online prediction of cell density of recombinant *Pichia pastoris* producing HBsAg. *Preparative Biochemistry and Biotechnology*, *49*, 352–359.

Birle, S., Hussein, M., & Becker, T. (2013). Fuzzy logic control and soft sensing applications in food and beverage processes. *Food Control*, *29*, 254–269.

Cereghino, G. P. L., Cereghino, J. L., Ilgen, C., & Cregg, J. M. (2002). Production of recombinant proteins in fermenter cultures of the yeast *Pichia pastoris*. *Current Opinion in Biotechnology*, *13*, 329–332.

Chen, L. Z., Nguang, S. K., Li, X. M., & Chen, X. D. (2004). Soft sensors for on-line biomass measurements. *Bioprocess and Biosystems Engineering*, *26*, 191–195.

Crowley, J., Arnold, S. A., Wood, N., Harvey, L. M., & McNeil, B. (2005). Monitoring a high cell density recombinant *Pichia pastoris* fed-batch bioprocess using transmission and reflectance near infrared spectroscopy. *Enzyme and Microbial Technology*, *36*, 621–628.

Fazenda, M. L., Dias, J. M., Harvey, L. M., Nordon, A., Edrada-Ebel, R., LittleJohn, D., & McNeil, B. (2013). Towards better understanding of an industrial cell factory: Investigating the feasibility of real-time metabolic flux analysis in *Pichia pastoris*. *Microbial Cell Factories*, *12*, 51.

Fernandes, R. L., Bodla, V. K., Carlquist, M., Heins, A. L., Lantz, A. E., Sin, G., & Gernaey, K. V. (2012). Applying mechanistic models in bioprocess development. *Measurement, monitoring, modelling, and control of bioprocesses* (pp. 137–166). Berlin: Springer.

Fortuna, L., Graziani, S., & Xibilia, M. G. (2009). Comparison of soft-sensor design methods for industrial plants using small data sets. *IEEE Transactions on Instrumentation and Measurement*, *58*, 2444–2451.

Gao, M.-J., Zheng, Z.-Y., Wu, J.-R., Dong, S.-J., Li, Z., Jin, H., … Lin, C.-C. (2012). Improvement of specific growth rate of *Pichia pastoris* for effective porcine interferon-α production with an on-line model-based glycerol feeding strategy. *Applied Microbiology and Biotechnology*, *93*, 1437–1445.

Gonzaga, J., Meleiro, L. A. C., Kiang, C., & Maciel Filho, R. (2009). ANN-based soft-sensor for real-time process monitoring and control of an industrial polymerization process. *Computers & Chemical Engineering*, *33*, 43–49.

Harms, J., Wang, X., Kim, T., Yang, X., & Rathore, A. S. (2008). Defining process design space for biotech products: Case study of *Pichia pastoris* fermentation. *Biotechnology Progress*, *24*, 655–662.

Harms, P., Kostov, Y., & Rao, G. (2002). Bioprocess monitoring. *Current Opinion in Biotechnology*, *13*, 124–127.

Jahic, M., Veide, A., Charoenrat, T., Teeri, T., & Enfors, S. O. (2006). Process technology for production and recovery of heterologous proteins with *Pichia pastoris*. *Biotechnology Progress*, *22*, 1465–1473.

Jenzsch, M., Gnoth, S., Kleinschmidt, M., Simutis, R., & Lübbert, A. (2007). Improving the batch-to-batch reproducibility of microbial cultures during recombinant protein production by regulation of the total carbon dioxide production. *Journal of Biotechnology*, *128*, 858–867.

Kadlec, P., Gabrys, B., & Strandt, S. (2009). Data-driven soft sensors in the process industry. *Computers & Chemical Engineering*, *33*, 795–814.

Kalos, A., Kordon, A., Smits, G., & Werkmeister, S. (2003). Hybrid model development methodology for industrial soft sensors. In *Proceedings of the 2003 American control conference* (pp. 5417-5422). IEEE.

Kano, M., & Fujiwara, K. (2012). Virtual sensing technology in process industries: Trends and challenges revealed by recent industrial applications. *Journal of Chemical Engineering of Japan*, *46*(1), 1–17.

Kordon, A., Smits, G., Kalos, A., & Jordaan, E. (2003). Robust soft sensor development using genetic programming. *Nature-Inspired Methods in Chemometrics*, *23*, 69–108.

Luttmann, R., Bracewell, D. G., Cornelissen, G., Gernaey, K. V., Glassey, J., Hass, V. C., … Mandenius, C. F. (2012). Soft sensors in bioprocessing: A status report and recommendations. *Biotechnology Journal*, *7*, 1040–1048.

Munch, G., Schulte, A., Mann, M., Dinger, R., Regestein, L., Rehmann, L., & Büchs, J. (2020). Online measurement of CO2 and total gas production in parallel anaerobic shake flask cultivations. *Biochemical Engineering Journal*, *153*, 107418.

OECD. (2014). *Guidance document on the validation of (quantitative) structure-activity relationship [(Q) SAR] models*. OECD Publishing.

Schügerl, K. (2001). Progress in monitoring, modeling and control of bioprocesses during the last 20 years. *Journal of Biotechnology*, *85*, 149–173.

Solle, D., Hitzmann, B., Herwig, C., Pereira Remelhe, M., Ulonska, S., Wuerth, L., & Steckenreiter, T. (2017). Between the poles of data-driven and mechanistic modeling for process operation. *Chemie Ingenieur Technik*, *89*, 542–561.

Stratton, J., Chiruvolu, V., & Meagher, M. (1998). High cell-density fermentation. In D. R. Higgins & J. M. Cregg (Eds.), *Pichia protocols* (pp. 107–120). Totowa, NJ: Humana Press.

Surribas, A., Geissler, D., Gierse, A., Scheper, T., Hitzmann, B., Montesinos, J. L., & Valero, F. (2006). State variables monitoring by in situ multi-wavelength fluorescence spectroscopy in heterologous protein production by *Pichia pastoris*. *Journal of Biotechnology*, *124*, 412–419.

Surribas, A., Montesinos, J. L., & Valero, F. F. (2006). Biomass estimation using fluorescence measurements in *Pichia pastoris* bioprocess. *Journal of Chemical Technology & Biotechnology: International Research in Process, Environmental & Clean Technology*, *81*, 23–28.

Takors, R. (2013). Industrielle Mikrobiologie, *Bioverfahrenstechnik. Industrielle mikrobiologie* (pp. 19–41). Berlin: Springer.

Ye, A. X., Wang, B. P., & Yang, C. Z. (2017). Time sequential phase partition and modeling method for fault detection of batch processes. *IEEE Access*, *6*, 1249–1260.

Zhang, H. (2009). Software sensors and their applications in bioprocess. In *Computational intelligence techniques for bioprocess modelling, supervision and control* (pp. 25–56). Berlin: Springer.

How to cite this article: Brunner V, Siegl M, Geier D, Becker T. Biomass soft sensor for a *Pichia pastoris* fed-batch process based on phase detection and hybrid modeling. *Biotechnology and Bioengineering*. 2020;117:2749–2759. https://doi.org/10.1002/bit.27454

3.4 Online sensor validation in sensor networks for bioprocess monitoring using swarm intelligence

Analytical and Bioanalytical Chemistry (2020) 412:2165–2175
https://doi.org/10.1007/s00216-019-01927-7

RESEARCH PAPER

Online sensor validation in sensor networks for bioprocess monitoring using swarm intelligence

Vincent Brunner[1] · Lukas Klöckner[1] · Roland Kerpes[1] · Dominik Ulrich Geier[1] · Thomas Becker[1]

Received: 8 March 2019 / Revised: 29 April 2019 / Accepted: 16 May 2019 / Published online: 8 July 2019

Abstract
Sensor faults can impede the functionality of monitoring and control systems for bioprocesses. Hence, suitable systems need to be developed to validate the sensor readings prior to their use in monitoring and control systems. This study presents a novel approach for online validation of sensor readings. The basic idea is to compare the original sensor reading with predictions for this sensor reading based on the remaining sensor network's information. Deviations between original and predicted sensor readings are used to indicate sensor faults. Since especially batch processes show varying lengths and different phases (e.g., lag and exponential phase), prediction models that are dependent on process time are necessary. The binary particle swarm optimization algorithm is used to select the best prediction models for each time step. A regularization approach is utilized to avoid overfitting. Models with high complexity and prediction errors are penalized, resulting in optimal predictions for the sensor reading at each time step (5% mean relative prediction error). The sensor reliability is calculated by the Kullback–Leibler divergence between the distribution of model-based predictions and the distribution of a moving window of original sensor readings (moving window size = 10 readings). The developed system allows for the online detection of sensor faults. This is especially important when sensor data are used as input to soft sensors for critical quality attributes or the process control system. The proof-of-concept is exemplarily shown for a turbidity sensor that is used to monitor a *Pichia pastoris*-batch process.

Keywords Sensor network · Sensor fault · Fault detection · Online validation · Particle swarm optimization

Introduction

The use of multimodal sensor networks has become an integral part of bioprocess monitoring. Multimodal sensor networks consist of different types of sensor devices measuring various process variables simultaneously and online. The sensor's measurement quality is restricted to a calibration prior to the process or experiment. Once this calibration is successfully performed and a sensor provides data, the reliability of the measurement system is usually not investigated further. However, besides the uncertainties typical for biological reactions [1], bioprocess data generated by means of multimodal sensor networks can be afflicted by errors resulting from sensor faults.

A sensor fault is generally defined as the deviation of the observed sensor reading from the true value [2, 3]. Faults in sensor networks can be classified as bias (intermittent, stepwise, drift-wise, or cyclic deviation), precision degradation (increase in sensor reading variance), and complete failure (severe malfunction, e.g., permanent zero value due to disconnection) of one or multiple sensors [4–6]. Possible causes of sensor faults include damaged sensors, short-circuited connections, and calibration errors [2]. Furthermore, measurement can be affected by unconsidered cross-sensitivity to matrix compounds (matrix effects).

Previous work on online sensor validation

The ability to detect sensor faults online and thus to validate the measurement system during the running process requires a supervision system. This system reads sensor network data and includes a decision logic with sensor reliability as an outcome, i.e., deciding whether a sensor reading can be

Published in the topical collection *Advances in Process Analytics and Control Technology* with guest editor Christoph Herwig.

✉ Dominik Ulrich Geier
dominik.geier@tum.de

[1] Chair of Brewing and Beverage Technology, Technical University of Munich, Weihenstephaner Steig 20, 85354 Freising, Germany

regarded as faulty or reliable. As outlined by Feital and Pinto [7], online sensor validation is an integral part of quality control via process analytical technology (PAT), especially when sensor data are used as input to soft sensors for the determination of critical quality attributes and process parameters or to the process control system. In the latter cases, false control responses can be avoided and productivity losses can potentially be reduced by the early detection of faults while the process is still running in a controllable state.

Several concepts for detecting sensor faults exist. In most cases where it is aspired to verify that the online measurement system provides reliable data for the target quantity, it is typically validated offline via lab reference measurements. For applications where high process safety must be guaranteed, one or a number of identical sensors in easy-to-compare positions can be installed and checked for equality [8]. However, lab reference measurements and redundant sensors incur additional costs. Furthermore, in the first case, it is not possible to replace the falsified sensor reading online and thus use the replaced value for control purposes. The latter case leaves open the question of which sensor reading to trust when only a small number of redundant sensors are used.

For these reasons, a series of approaches to the online validation of sensor readings has been developed, as reviewed by Das et al. [9] and Isermann [10]. These approaches can be classified as statistical, artificial intelligence, and model-based techniques, as well as their hybrid variants [9]. All these methods are, to a certain degree, based upon the information redundancy inherent in sensor network data. However, only a minority of these approaches have proven their transferability to highly non-linear and time-variant processes, such as biological batch processes. Below, previous work on online sensor validation is reviewed briefly, focusing on non-linear processes.

The multivariate statistical process control (MSPC) approach uses principal component analysis (PCA) and partial least squares (PLS) or their variants to build an empirical model based on measurements of a reference database describing the normal operation of the process. Deviations from normal operation, such as process or sensor faults, are then indicated in control charts [11]. The contribution of each input variable to the underlying statistics of the latent variable models allows one to investigate which input variable(s) caused the deviation [12, 13], thus allowing sensor fault detection and identification.

The pattern recognition approach uses artificial intelligence methodology to detect the patterns of sensor faults and can therefore be considered as a pattern recognition problem. For example, Bayesian networks were trained with artificially added sensor faults and subsequently used for fault identification [4, 14]. For this approach, however, separate prediction models are necessary to replace the falsified sensor reading. Guo and Nurre [15] trained two different artificial neural networks for a sensor validation problem: the first was trained to detect and isolate sensor faults and the second to recover the values of critical variables when their measurements fail.

Another promising approach to sensor fault detection is to predict the sensor reading of interest using the data stream of the remaining sensor network and evaluating the residual between the predicted and original sensor reading, which is referred to as a *symptom signal* [16]. This approach contains two major steps—*prediction* and *symptom signal evaluation*—for which suitable methods must be found depending on the application. Symptom signals generated by PCA model predictions were used by Dunia et al. [17] to identify sensor faults and determine the type of fault as bias, complete failure, drifting, or precision degradation. Alag et al. [18] proposed a method of adaptive time-series modeling based upon an artificial neural network for sensor reading prediction and used the statistical properties of the residuals in combination with probabilistic reasoning to identify sensor faults. Ibargüengoytia et al. [19] used probabilistic propagation based on Bayesian networks for residual generation and probabilistic reasoning for fault detection inside the sensor network. Zarei and Shokri [16] proposed a non-linear unknown input observer based upon a Bayesian filter to generate the residuals. The advantage of these approaches is that the replacement for falsified sensor readings is directly accessible. However, this approach has a bottleneck, namely the accuracy of the underlying model that gives the prediction to be compared with the original sensor reading. For fault detection and recovery in batch processes with varying lengths and different phases (e.g., lag, exponential, and stationary phase or process phases resulting from multi-substrate media), prediction models must be capable of operating in real time and depend upon the process time, i.e., they must be dynamic [20]. However, most studies on fault detection and identification use a static prediction model for the whole process. Furthermore, the information redundancy inherent in the sensor network is not utilized efficiently when only one model input combination is used for prediction, i.e., for residual generation. It is supposed that sensor fault detection is more robust when several model combinations are used for residual generation, resulting in a distribution of predictions. Finally, a fixed threshold for fault detection, as used in most studies, can yield false alarms when noise or unforeseen events occur in the sensor network data.

Scope and outline of this study

In the present study, a novel approach for online sensor validation using dynamic modeling based on swarm intelligence was developed. The approach can be summarized in four steps: (I) the process progress (maturity) is predicted via PLS regression (PLSR), making it possible to divide the process into overlapping process segments; (II) for each process segment, a pool of possible models for the prediction of the

sensor reading to be validated is defined; the inputs of these models consist of various combinations of the remaining sensor network's readings, as in Krause et al. [13]; (III) an optimal subset of models for the prediction of the sensor reading is determined for each process segment using the binary particle swarm optimization (BPSO) algorithm [21]; and (IV) the BPSO-based best models are used to create a distribution of predictions for each process time step. The deviation of the original and predicted sensor readings, evaluated by means of the Kullback–Leibler divergence, indicates a sensor fault and is used to quantify the sensor reliability. The threshold between a faulty and reliable sensor reading is dynamic and based upon the total accuracy of the prediction models selected via the BPSO algorithm.

This study shows the proof-of-concept of this approach for a turbidity sensor that is used to monitor a *Pichia pastoris*-batch process. Turbidity measurements are frequently used as the main input to soft sensors for the prediction of biomass concentration. Biomass concentration is associated with the majority of critical quality attributes (CQA) in upstream bioprocessing [22] and its online monitoring is therefore pivotal for quality control. Soft sensors for biomass concentration based on turbidity measurements would lose their predictive performance in the case of sensor faults or unexpected deviations. Unexpected deviations of turbidity readings are hereinafter defined as deviations of the correlation of biomass with the turbidity reading that cannot be explained by any other available online or offline process analysis. The developed system makes it possible to detect faults as well as unexpected deviations of the turbidity sensors used for the prediction of biomass concentration online and, if necessary, to replace the falsified original value with that predicted via the sensor network. The advantages of this approach are that the replacement for the falsified sensor reading is directly accessible and sensor validation is reliable due to taking the whole sensor network's information into account in a dynamic manner.

Materials and methods

Strain and preculture conditions

A single colony of a recombinant *P. pastoris* strain based on type strain DSMZ 70382 was grown on a YPD plate (yeast extract, 10 g L^{-1}; peptone, 20 g L^{-1}; glucose, 20 g L^{-1}; bacteriological agar, 15 g L^{-1}). This working culture was used to inoculate the preculture in three 150-mL shake flasks containing 50 mL of the mineral medium FM22 with glycerol as the carbon source: $(NH_4)_2SO_4$, 5 g L^{-1}; $CaSO_4 \cdot 2H_2O$, 1 g L^{-1}; K_2SO_4, 14.3 g L^{-1}; KH_2PO_4, 42.9 g L^{-1}; $MgSO_4 \cdot 7H_2O$, 11.7 g L^{-1}; glycerol, 40 g L^{-1} [23]; and trace element solution, 2 mL L^{-1} of the culture medium. The trace element stock solution contained: $CuSO_4 \cdot 5H_2O$, 2 g L^{-1}; KI, 0.08 g L^{-1}; $MnSO_4 \cdot H_2O$, 3 g L^{-1}; $Na_2MoO_4 \cdot 2H_2O$, 0.2 g L^{-1}; H_3BO_3, 0.02 g L^{-1}; $CaSO_4 \cdot 2H_2O$, 0.5 g L^{-1}; $CoCl_2$, 0.5 g L^{-1}; $ZnCl_2$, 7 g L^{-1}; $FeSO_4 \cdot H_2O$, 22 g L^{-1}; biotin, 0.2 g L^{-1}; conc. H_2SO_4, 1 mL.

Bioreactor and sensor network

A total of 150 mL of preculture was used to inoculate the main culture in a Biostat® Cplus bioreactor (42 L total volume; Sartorius AG, Goettingen) with a working volume of 15 L and control of pressure, pH, temperature, and dissolved oxygen (DO); the controller setpoints were 500 mbar, 5, 30 °C, and 40%, respectively. A DO minimum of 40% was controlled in a sequence cascade by agitation with three Rushton impellers (300–600 min^{-1}) followed by air flow (20–40 L min^{-1}) via a ring sparger. NH_4OH was used as nitrogen source and to set and maintain a pH of 5. The end of the batch process was defined as the time at which the carbon source glycerol was completely depleted, as indicated by a characteristic peak in the off-gas CO_2 signal. To measure the O_2 and CO_2 concentrations in the off-gas, a BlueInOne Cell sensor (BlueSens gas sensors GmbH, Herten) was used. Turbidity was measured with an optical fiber sensor InPro 8100 (Mettler-Toledo GmbH, Giessen), measuring the backscattered light of the cell suspension.

The portion of sensor network data used for modeling consisted of sensor readings for O_2 and CO_2 in the off-gas, DO, and turbidity, as well as actuator values for air flow, stirrer speed, and base feed (see Fig. 1). The process variables pressure, pH, and temperature are robustly controlled to a specific setpoint and therefore contain no information that would be relevant for modeling.

Data management

The Biostat® digital control unit (Sartorius AG, Goettingen) was used for primary process control. The Sartorius OPC server and SIMATIC SIPAT (Siemens AG, Munich) were used for communication and data management, respectively. Data were collected in a central SIPAT process database with a recording rate of 30 s. Data pre-processing and modeling were performed in MATLAB R2019a (The MathWorks, Inc., Natick, MA).

Results and discussion

This study presents a novel approach for online validation of sensor readings. The basic idea is to compare the original sensor reading with predictions for this sensor reading based on the remaining sensor network's information. Deviations between original and predicted sensor readings are used to indicate sensor faults. The algorithm is structured into two branches (see Fig. 2). The left branch is used for automatic

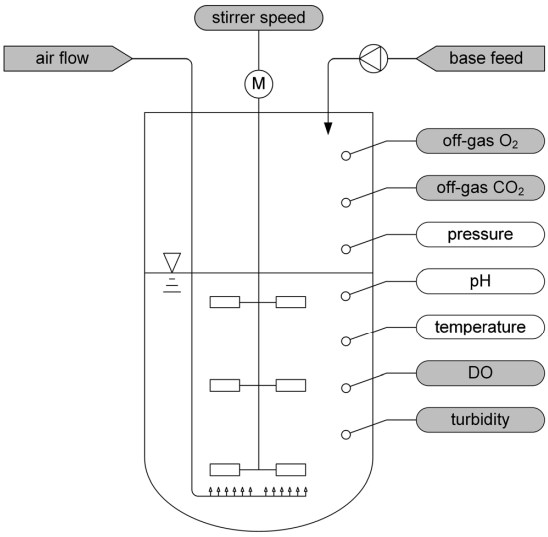


Fig. 1 Bioreactor and sensor network with all sensors and actuators used for the monitoring and control of the *P. pastoris*-batch process. Signals used in the prediction models are marked in gray. The online data stream comprises sensor readings for O_2 and CO_2 in the off-gas, pressure, pH, dissolved oxygen (DO), temperature, and turbidity, as well as actuator values for air flow, stirrer speed, and base feed. Since pressure, pH, and temperature are robustly controlled to a specific setpoint, they are not used in the prediction models

model selection and calibration based on historical process data; the right branch uses these models for the validation of sensor readings based on the online data stream. Fig. 2 shows how data and model parameters are passed through the algorithm. Each step of the algorithm is described below in detail.

A proof-of-concept for this algorithm is exemplarily shown for a turbidity sensor that is used to monitor a *P. pastoris*-batch process. Automatic model selection and calibration (see Fig. 2, left branch) were performed with four data sets of batch processes, which did not show significant sensor faults. Three further data sets were used to simulate the online application of the algorithm, i.e., to perform online sensor validation (see Fig. 2, right branch). The batch process that was chosen as an illustrative example in the following sections showed multiple significant faults of the turbidity sensor. The structure of this chapter is oriented according to the four algorithm steps described previously (see the "Introduction" section).

Determination of process progress

The interrelations between the sensor of interest and the remaining sensor network are assumed to be similar for a certain process phase (e.g., lag or exponential phase). However, the duration and progress of batch processes can vary due to, e.g., variable cell viability and vitality in the preculture. This results in sometimes shorter, sometimes longer batch processes. Thus, the first step of the algorithm is to develop a model for the process progress, hereinafter referred to as maturity $\boldsymbol{m}$. When a batch process is finished, $\boldsymbol{m}$ can be calculated via (1), where $\boldsymbol{t}$ is the time vector and t_{end} corresponds to the total batch duration.

$$\boldsymbol{m} = \frac{\boldsymbol{t}}{t_{end}} \tag{1}$$

The model in (2) is used to predict the maturity $\boldsymbol{m}$ for given sensor network data $\boldsymbol{X_m}$ at time step t with $\boldsymbol{b_m}$ corresponding to the model coefficients.

$$\boldsymbol{m} = \boldsymbol{X_m}(t)\,\boldsymbol{b_m} \tag{2}$$

The main prerequisite for the determination of process progress via the proposed approach (indicator variable technique) is that model inputs should be monotonically increasing or decreasing [24]. Thus, a subgroup $\boldsymbol{X_m}$ of the whole sensor network data $\boldsymbol{X}$ was used for the determination of the process progress. This subgroup comprised all process variables showing almost monotonical behavior, namely base feed, off-gas CO_2 and O_2, and DO, as well as the logarithm of each of these variables. Logarithmic transformation was used to account for the exponential behavior of process variables related to biomass growth in biological batch processes. This resulted in a total of eight predictor variables (four original, four logarithmic) used for the maturity model. Even though DO is a controlled variable, it is a suitable indicator for maturity due to its reproducible decline from 100% at process start to 40% at about three fourths of the total batch duration. After DO reaches the setpoint of 40%, it is reliably controlled to that value (see the "Materials and methods" section).

The model coefficients, $\boldsymbol{b_m}$, were determined by PLSR using the ln- and non-transformed historical sensor network data $\boldsymbol{X_{hist,m}}$ as the predictor and the maturity as the predicted variable. Three latent variables were used for the PLSR. The contributions of the ln- and non-transformed variables to the PLSR model were calculated via the variable importance in the projection (VIP) method [25]. This method is used to quantify the information content of each variable used in the resulting PLSR model. The ln- and non-transformed variables base feed, off-gas CO_2, off-gas O_2, and DO contributed with 62%, 6%, 5%, and 27%, respectively, to the maturity model. The high impact of base feed to the maturity model is in line with the quasi-linear increase of base feed during the batch process. Validation with the three test data sets resulted in $NRMSE = 11\%$ (normalized root mean squared error) and $R^2 = 0.85$. Fig. 3 shows the result of the maturity model for an exemplary validation batch process with a total batch duration $t_{end} = 35.8$ h. The bisector depicts the calculated maturity as in (1). For this batch process, the predicted maturity starts with 8% due to the high amount of base fed to the bioreactor at the very beginning in this case and the high

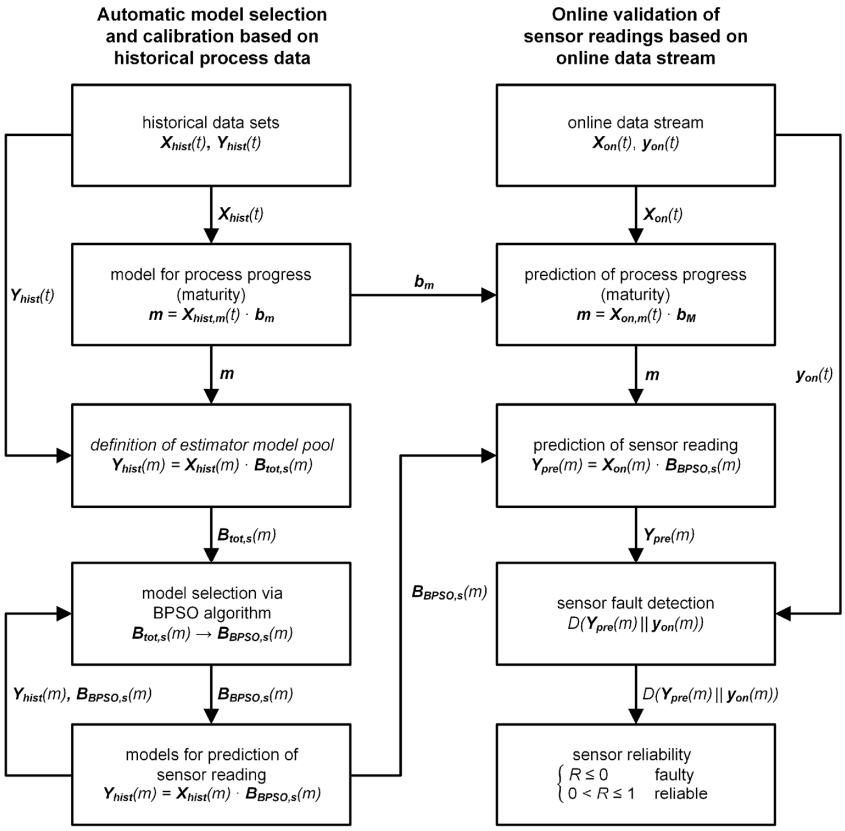


Fig. 2 Algorithmic structure for sensor fault detection using the binary particle swam optimization (BPSO) algorithm. X_{hist}, Y_{hist}, X_{on}, Y_{on}, and Y_{pre} correspond to the historical, online, and predicted data for the model inputs X and the variable to be predicted Y or y, respectively. $X_{hist,m}$ and $X_{on,m}$ correspond to a subgroup of the whole sensor network data X_{hist} and X_{on}, respectively, and are used for the determination of the process progress (maturity). $B_{tot,s}$ and $B_{BPSO,s}$ correspond to the model coefficients of all models included in the model pools and the ones determined via the BPSO algorithm, respectively, where s is the used process segment. $D(Y_{pre}(m)||y_{on}(m))$ is the Kullback–Leibler divergence from $y_{on}(m)$ to $Y_{pre}(m)$ as in (13). R corresponds to the sensor reliability calculated by the Kullback–Leibler divergences from $Y_{pre}(m)$ to the 95% confidence limits of $Y_{pre}(m)$ and $Y_{on}(m)$, respectively, as in (15). Sensor readings with $R \leqq 0$ are regarded as significantly faulty, while readings in the range of $0 < R \leqq 1$ are regarded as significantly reliable. Furthermore, t = time and m = maturity (process progress)

influence of base feed in the maturity model. Leaps in the maturity domain (e.g., at 7 h) result from slight inaccuracies of the underlying PLSR model and the not completely monotonous behavior of the model inputs. Considering the difficulty of finding a reliable measure for process progress before the process is actually finished, the shown maturity model is accurate enough to use the predicted maturity in the following algorithm steps.

The model for maturity is necessary for selecting the right model pool independent of process time, as described in the next section.

Definition of prediction model pools

The maturity model (2) is used to disassemble the whole batch process into b overlapping process segments. Each segment s_a = $s_1 \ldots s_b$ has an identical maturity size of $m = 30\%$ corresponding to a time span of approximately 10 h. Given a recording rate of 30 s, each segment contains about 1200 sensor readings for model calibration. The size of this time span represents a trade-off between insufficient amount of data points for reliably calibrating the prediction models (too small time span) and not capturing different process phases, such as the lag and exponential phase (too large time span). The start of segment s_{a+1} is shifted by $m = +1\%$ compared to segment s_a, resulting in overlapping segments. Although m is for the calibration of the maturity model (2) limited to 0% and 100%, the segmentation was conducted for $m = -5\%$ to 105%, resulting in $b = 81$ segments. By permitting this bidirectional extrapolation of $\Delta m = 5\%$ for the segmentation, the segments at process start ($m = 0\%$) and end ($m = 100\%$) are overlapping. This overlapping is essential for having smooth predictions

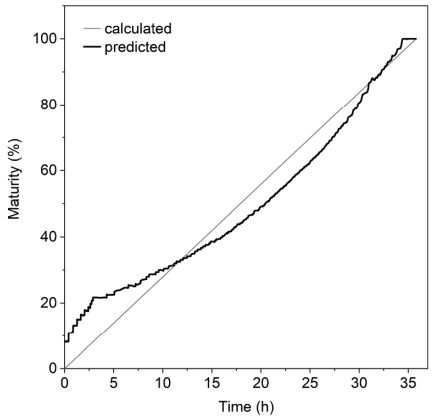


Fig. 3 Calculated and predicted maturity (i.e., process progress) for an exemplary validation batch process with a batch duration t_{end} = 35.8 h. For this batch process, the predicted maturity starts at 8% due to the high amount of base fed to the bioreactor at the very beginning in this case and the high influence of base feed upon the maturity model

for the sensor reading of interest. A segmentation without overlaps would result in leaps of the sensor prediction. Additionally, slight inaccuracies of maturity predictions at the process end are covered by permitting segments up to $m = 105\%$.

For each segment s_a, a pool of possible prediction models (3) is set up to map the historical sensor network data $\boldsymbol{X_{hist}}$ to the historical sensor readings of interest $\boldsymbol{Y_{hist}}$. The maturity model (2) is used for transformation so that the prediction models (3) are a function of maturity m rather than of time t. $\boldsymbol{B_{tot,s}}$ is the matrix of model coefficients.

$$\boldsymbol{Y_{hist}}(m) = \boldsymbol{X_{hist}}(m)\,\boldsymbol{B_{tot,s}}(m) \tag{3}$$

The matrix of independent variables, $\boldsymbol{X_{hist}}$, contains, besides raw signals (data for CO_2 and O_2 in the off-gas, DO, air flow, stirrer speed, and base feed), polynomial combinations. With these extensions, interrelations between the sensor network readings and the quadratic and cubic characteristics of the prediction models are accounted for. These extensions result in a total of 78 different model inputs comprising the following structure of terms:

- Six single raw, quadratic, and cubic terms (e.g., DO, $(DO)^2$, and $(DO)^3$, respectively)
- 15 combinations of raw terms (e.g., DO*(base feed))
- 15 combinations of quadratic terms (e.g., $(DO)^2$*(base feed)^2)
- 30 combinations of mixed raw and quadratic terms (e.g., DO*(base feed)^2)

By allowing all possible model structures except for the case where all coefficients in $\boldsymbol{B_{tot,s}}$ are zero, the maximum number of possible combinations for each process segment s_a is $2^{77} \approx 1.51 \times 10^{23}$. Due to this high number of possible combinations, an intelligent search is necessary for finding the best prediction models.

A prediction model pool is herein referred to as a unique combination of model inputs for (3) and can be different for each process segment s_a to account for the time variance of batch processes.

To find the optimal prediction models from the b model pools regarding prediction error and the risk of overfitting, the BPSO algorithm is used, as described in the following section. PLSR is used as regression method for solving (3) due to its superior behavior with highly collinear data compared to, for example, multiple linear regression (MLR) [26].

Model selection via the BPSO algorithm

The BPSO algorithm is used to determine a subset of the best prediction models out of the previously described model pool for each of the $b = 81$ process segment. In contrast to the continuous PSO, each particle in the BPSO represents its position with binary values, as first described by Kennedy and Eberhart [27]. The position vector is thus a bit vector with the dimension of the search space. In the present optimization problem, each bit of this position vector describes whether the according input term out of the prediction-model pool for segment s is used in the model structure (1 or "true") or not (0 or "false"). For the number of iterations, $n_{iterations}$, the optimization algorithm evaluates the cost function for each particle out of the total number of particles, $n_{particles}$, (population size) to determine the optimal subset of prediction models. The BPSO algorithm was adapted from Khanesar et al. [21], as described below.

The BPSO changes the particle positions, i.e., the model selection, by updating the position bits by means of a change rate v_i (corresponding to the velocity in a continuous PSO). Each particle's bit has one of the two change rates v_{ij}^1 or v_{ij}^0, as in (4), where v_{ij}^1 and v_{ij}^0 describe the rate at which the jth bit of the ith particle changes its value from 0 to 1 or from 1 to 0, respectively, at iteration step z.

$$v_{ij}(z) = \begin{cases} v_{ij}^1 & \text{if} \quad x_{ij} = 0 \\ v_{ij}^0 & \text{if} \quad x_{ij} = 1 \end{cases} \tag{4}$$

Updating the change rates v_{ij}^1 and v_{ij}^0 is conducted according to the local (p_{ibest}) and global best position (p_{gbest}), as in (5), (6), and (7), where d_{ij}^1 and d_{ij}^0 are temporary variables that depend on p_{ibest} and p_{gbest} and w is the inertia weight. The local and global best positions in (7) are integers in {0, 1} and are determined by evaluating the cost function (11), which is described later on. p_{ibest} is the best position of the ith particle, whereas p_{gbest} is the best position of all particles up to the current iteration step. Both p_{ibest} and p_{gbest} are updated in each iteration step. c_1 and c_2 are positive constants that can be

considered as the impacts of the local and global bests, respectively. r_1 and r_2 are uniformly distributed random numbers (data type "double") in [0, 1] generated for each iteration step.

$$v_{ij}^1 = w\, v_{ij}^1 + d_{ij,1}^1 + d_{ij,2}^1 \tag{5}$$

$$v_{ij}^0 = w\, v_{ij}^0 + d_{ij,1}^0 + d_{ij,2}^0 \tag{6}$$

The temporary variables d_{ij}^1 and d_{ij}^0 are determined following the rules in (7).

$$\begin{aligned}
&\text{if } p_{ibest,j} = 1 \quad \text{then } d_{ij,1}^1 = r_1\, c_1 \quad \text{and } d_{ij,1}^0 = -r_1\, c_1\\
&\text{if } p_{ibest,j} = 0 \quad \text{then } d_{ij,1}^0 = r_1\, c_1 \quad \text{and } d_{ij,1}^0 = -r_1\, c_1\\
&\text{if } p_{gbest,j} = 1 \quad \text{then } d_{ij,2}^1 = r_2\, c_2 \quad \text{and } d_{ij,1}^0 = -r_2\, c_2\\
&\text{if } p_{gbest,j} = 0 \quad \text{then } d_{ij,2}^1 = r_2\, c_2 \quad \text{and } d_{ij,1}^0 = -r_2\, c_2
\end{aligned} \tag{7}$$

As shown in (8), the thus selected change rate is processed by means of a uniform sigmoid function sig(·) to transform the change rate v_{ij} to v_{ij}', where v_{ij}' is the probability that the bit changes its value; v_{ij}' is constrained to the interval [0, 1].

$$v_{ij}'(z) = \mathrm{sig}\left(v_{ij}(z)\right) = \frac{1}{1 + e^{-v_{ij}(z)}} \tag{8}$$

The position of each particle is updated iteratively by (9), where $\bar{x}_{ij}$ is the complement of x_{ij} and r_{ij} is a uniformly distributed random number in [0, 1].

$$x_{ij}(z+1) = \begin{cases} \bar{x}_{ij}(z) & \text{if} \quad r_{ij} < v_{ij}'(z+1) \\ x_{ij}(z) & \text{else} \end{cases} \tag{9}$$

The position of each swarm particle represents one model input combination, as described previously. A PLSR is conducted for each particle i, i.e., a certain model input combination, and iteration step z with training data $\boldsymbol{X_{hist}}$ as predictor and $\boldsymbol{y_{hist}}$ as target variable, resulting in predictions for the sensor reading of interest, $\hat{\boldsymbol{y}}_{\boldsymbol{hist}}$.

The cost function is designed such that the normalized root mean squared error (*NRMSE*) of the prediction models resulting from a given model input combination is minimized. The *NRMSE* in (10) is calculated from training data $\boldsymbol{y_{hist}}$ and $\hat{\boldsymbol{y}}_{\boldsymbol{hist}}$ by a 10-fold cross validation, where y_{min} and y_{max} is the minimum and maximum value of $\boldsymbol{y_{hist}}$, respectively, and l is the number of data points in $\boldsymbol{y_{hist}}$.

$$NRMSE = \frac{1}{y_{max} - y_{min}} \sqrt{\frac{1}{l} \sum_{k=1}^{l} \left(\hat{y}_{k,hist} - y_{k,hist}\right)^2} \tag{10}$$

Furthermore, a regularization approach is used to avoid the risk of overfitting. Models with high complexity, as quantified by means of the total number of model terms ($n_{i,terms}$) and the number of latent variables of the PLSR ($n_{i,lat}$) are penalized. The cost function of the BPSO algorithm can thus be written as in (11), where $cost_i$ is the variable to be minimized and ω_1, ω_2, and ω_3 are the weights for the three cost function terms.

$$cost_i = \omega_1\, NRMSE_i + \omega_2\, n_{i,terms} + \omega_3\, n_{i,lat} \tag{11}$$

The BPSO algorithm is used to determine a subset of the 25 best prediction models from each of the b model pools described above. Since the process segments are overlapping, the number of predictions for each process time step varies (Fig. 4): there are fewer prediction models at the beginning and end of the batch process and a maximum number of prediction models in the middle (plateau). As shown in Fig. 4, the number of predictions per time step is 750 between 5.7 h and 28.9 h, resulting from 30 overlapping segments and 25 models selected via the BPSO algorithm. The number of available models before and after 5.7 h and 28.9 h, respectively, is influenced by segmentation based on the predicted maturity. Thus, leaps in the time domain can occur, resulting in deviations of the trapezoidal shape.

The parameter settings of the BPSO algorithm are summarized in Table 1. The parameterization of the inertia weight w is a compromise between exploratory (high w) and exploitative (low w) search behavior and can be a fixed or dynamically changing value [28]. The choice of $w = 0.8$ is close to the upper limit of the range suggested by Del Valle et al. [28] as $w = [0.4, 0.9]$. A fixed w was preferred over a linearly decreasing w (as for example described by Shi and Eberhart [29]), because the swarm would lose its exploration mode when w is lowered during iterations. The best acceleration coefficients c_1 and c_2 regarding convergence were 0.45 and 0.6, respectively

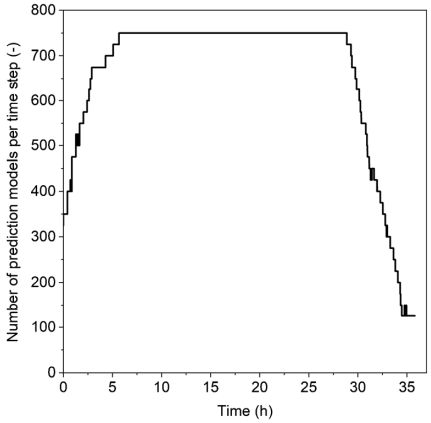


Fig. 4 Number of prediction models for an exemplary batch process with a batch duration $t_{end} = 35.8$ h. Between process times of 5.7 h and 28.9 h, 750 different BPSO predictions are available. This plateau results from a maximum number of 30 overlapping segments and 25 models per segment. Before and after 5.7 h and 28.9 h, respectively, fewer segments overlap; thus, the number of available models is influenced by segmentation based on the predicted maturity (i.e., process progress)

Table 1 Parameters for the BPSO algorithm used in this study

Parameter	w	c_1	c_2	ω_1	ω_2	ω_3	$n_{iterations}$	$n_{particles}$
Value	0.8	0.45	0.6	500	1	1	100	50

w is the inertia weight; c_1 and c_2 are positive constants that can be considered as the impacts of the local and global bests, respectively; ω_1, ω_2, and ω_3 are weights for the three cost function terms in (11); $n_{particles}$ is the total number of particles (i.e., the population size); $n_{iterations}$ is the total number of iterations

(data not shown). The cost function weights ω_1, ω_2, and ω_3 were parameterized so that the effect of the cost function terms $NRMSE_i$, $n_{i,terms}$, and $n_{i,lat}$ described in (11) have an impact ratio of approximately 70, 25, and 5%, respectively, with respect to the total cost. This weighting of impact was implemented due to the different importance of each of the three quality criteria for regression models and resulted in global mean values of 0.05, 9.43, and 2.87, respectively, for $NRMSE_i$, $n_{i,terms}$, and $n_{i,lat}$. The number of iterations, $n_{iterations}$, was set to 100 because convergence was reached after at least 80 iterations (data not shown). The number of swarm particles, $n_{particles}$, was set to 50. An increase of $n_{particles}$ did not improve convergence (data not shown).

The results of the model selection via the BPSO algorithm are illustrated by means of the model inclusion. Model inclusion is referred to as the percentage of all 50 swarm particles that include the specific input term in the model structure. As swarm particles are randomly generated inside the search space, the model inclusion is randomly distributed for all 78 possible model inputs at the iteration start. At the iteration end, the BPSO converges to an optimal set of prediction models, to which a few model inputs contribute more than others. As can be seen in Fig. 5, which shows the model inclusion at the iteration start and end for a process segment in the late exponential phase, the main contributing model inputs (> 50%) are DO, (base feed)^3, (air flow)*(off-gas CO_2)^2, and (stirrer speed)*(off-gas O_2)^2. In this example, the model inputs (base feed)^3 and (air flow)*(off-gas CO_2)^2 are even contained in 100% of the optimal prediction models. This result is in line with the observation of high collinearity between the turbidity reading and the data of these parameters in this process state.

The result of the BPSO algorithm is a set of 50 optimal prediction models for each process segment. The 25 best prediction models in terms of cost function evaluation ($cost_i$) were selected and handed over to the next step of the algorithm.

Sensor fault detection

The main idea behind sensor fault detection in this study is to compare the distribution of predictions with that of original online sensor readings. For discrimination between two discrete

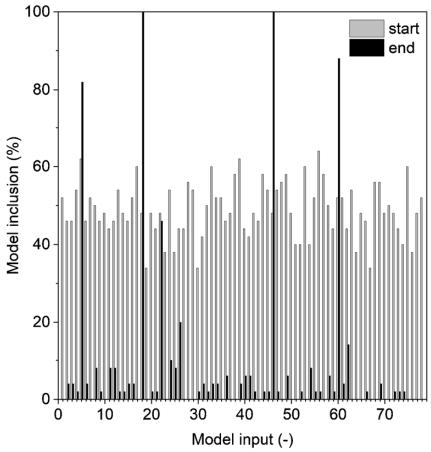


Fig. 5 Representation of the model inclusion calculated via the BPSO algorithm at the iteration start and end for a process segment in the late exponential phase. At the iteration start, the model inclusion (i.e., the portion of prediction models that includes the specific model input) is randomly distributed around 50% for all 78 possible model inputs. At the iteration end, the model pool is narrowed down to a smaller subset of optimal prediction models, including main contributions (> 50%) of the model inputs with numbers 5, 18, 46, and 60, corresponding to DO, (base feed)^3, (air flow)*(off-gas CO_2)^2, and (stirrer speed)*(off-gas O_2)^2

probability distributions $\boldsymbol{Q}$ and $\boldsymbol{P}$, the directed Kullback–Leibler divergence D [30] from $\boldsymbol{Q}$ to $\boldsymbol{P}$ is formulated as

$$D\left(\boldsymbol{P}\|\boldsymbol{Q}\right) = \sum_{i=1}\left(\boldsymbol{P}_i \log_2 \frac{\boldsymbol{P}(i)}{\boldsymbol{Q}(i)}\right) \tag{12}$$

where the logarithms are taken to base 2 if information is measured in bits. D is zero if the distributions $\boldsymbol{Q}$ and $\boldsymbol{P}$ are equal and increases logarithmically with increasing discrepancy.

The best prediction models, as selected by the BPSO and ranked by means of the cost function (11), are used to generate the distribution of online predictions for the sensor reading of interest for each process segment. This results in a vector $\boldsymbol{y_{pre}}$ containing the prediction results for each measurement step. The distribution of original sensor readings is generated using a moving window approach. Moving window regression with first degree polynomial regression (window size = 10 readings) is used to transform the original readings, as in (13), where $\boldsymbol{e}$ is the residual vector between the original readings and the moving window predictions, $\bar{\boldsymbol{y}}_{\boldsymbol{on}}$ is the arithmetic mean of the last ten original readings, and $\boldsymbol{y_{on}}$ is a vector containing the transformed readings for each measurement step. Every single turbidity reading is transformed this way using the sensor readings of the last 5 min (= 10 readings).

$$\boldsymbol{y_{on}} = \boldsymbol{e} + \bar{\boldsymbol{y}}_{\boldsymbol{on}} \tag{13}$$

By using first order polynomial regression, the quasi-linear increase of turbidity readings inside the time frames is accounted for. The moving window regression thus transforms the quasi-

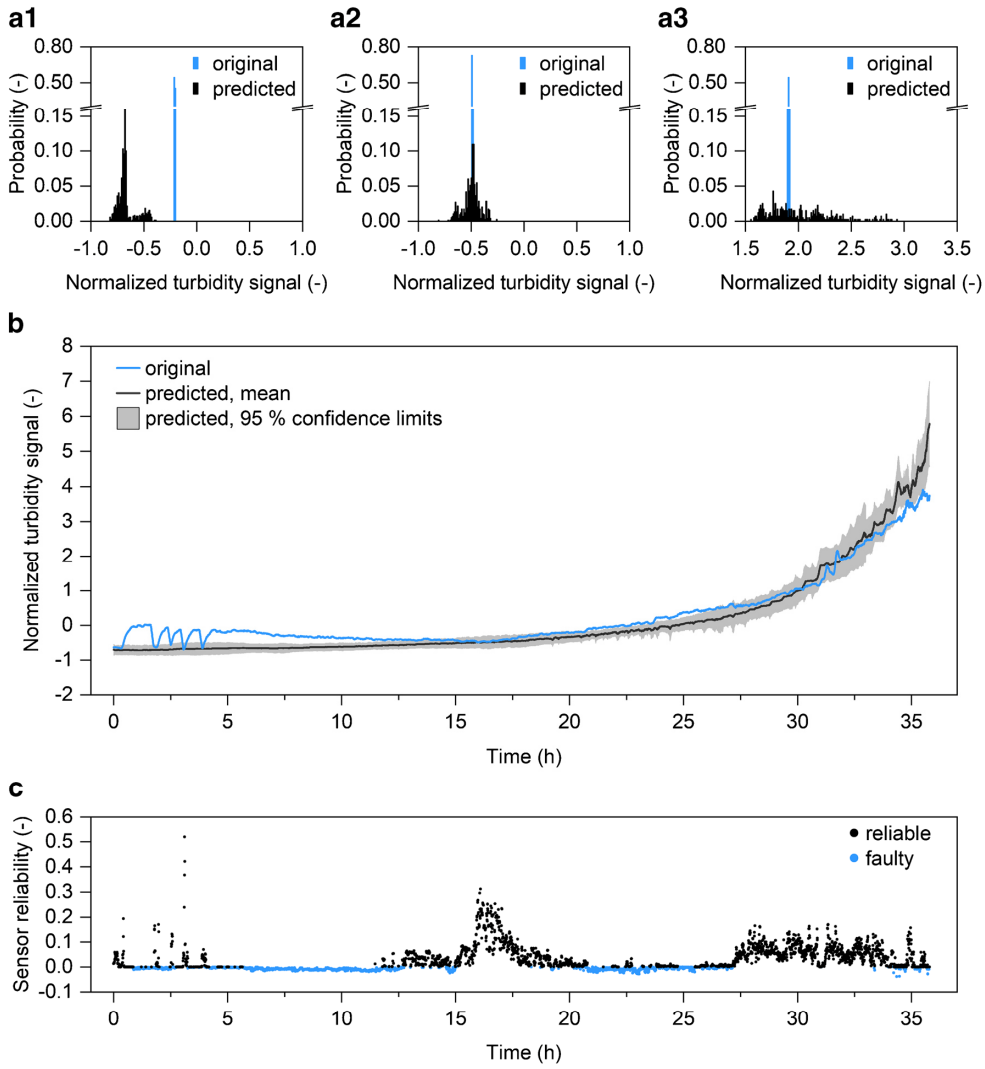


Fig. 6 (a1–a3) Histograms of the prediction models at different process times (8, 16, and 32 h) with distributions of original (blue bars) and BPSO prediction data (black bars) of the normalized turbidity signal. The probability indicates how often one of the 750 prediction models for this time step predicted a value in the corresponding bar range. (b) Exemplary batch process with original (blue line) and mean BPSO prediction (black line) of the normalized turbidity signal. The confidence interval (gray corridor) corresponds to the interval in which the model predictions lie with a probability of 95%. Significant faults of the turbidity sensor occur at 0–12 h, 23–26 h, and after 34 h. (c) Reliability of the turbidity sensor R calculated with a dynamic threshold. Turbidity readings with $R \leq 0$ are regarded as significantly faulty (blue dots), while readings in the range of $0 < R \leq 1$ are regarded as significantly reliable (black dots)

linear original time frame (polynomial degree = 1) to a quasi-constant transformed time frame (polynomial degree = 0). This quasi-constant time frame of transformed sensor readings $\boldsymbol{y_{on}}$ is subsequently used for fault detection.

Since the decision as to whether a sensor reading is regarded as reliable or faulty should depend upon the consistency of the predictions $\boldsymbol{y_{pre}}$, a dynamic threshold is introduced to quantify the sensor reliability R, as in (14). Here, $\boldsymbol{y_{thresh}}$ corresponds to the 95% confidence limits of $\boldsymbol{y_{pre}}$ and is thus dependent on the consistency of the predictions $\boldsymbol{y_{pre}}$.

$$R = \frac{D\left(\boldsymbol{y_{pre}} \| \boldsymbol{y_{thresh}}\right) - D\left(\boldsymbol{y_{pre}} \| \boldsymbol{y_{on}}\right)}{D\left(\boldsymbol{y_{pre}} \| \boldsymbol{y_{thresh}}\right)} \tag{14}$$

The formula for R (14) is designed such that $R_{max} = 1$ if $\boldsymbol{y_{pre}}$ and $\boldsymbol{y_{on}}$ are equal, and $R \leq 0$ for sensor readings outside of the 95% confidence interval of $\boldsymbol{y_{pre}}$. As in (15), sensor readings with $R \leq 0$ are regarded as significantly faulty, whereas readings in the range of $0 < R \leq 1$ (i.e., sensor readings are inside the 95% confidence interval of $\boldsymbol{y_{pre}}$) are regarded as significantly reliable.

$$R = \begin{cases} \text{faulty} & \text{if} \quad R \leq 0 \\ \text{reliable} & \text{if} \quad 0 < R \leq 1 \end{cases} \tag{15}$$

The results of sensor fault detection are summarized in Fig. 6. At each time step, a distribution of predicted ($\boldsymbol{y_{pre}}$) and original sensor readings ($\boldsymbol{y_{on}}$) is generated using the prediction models selected via the BPSO algorithm and the moving window data, respectively. Graphs a1–a3 show the histograms of these distributions exemplarily at process times of 8, 16, and 32 h. A divergence between the original data and those predicted via the whole sensor network indicates a sensor fault. Graph b depicts the time course of the original turbidity signal, together with the mean and 95% confidence interval of the BPSO predictions; the graph can be used to indicate significant sensor faults at 0–12 h, 23–26 h, and after 34 h.

Using the graphs a and b, sensor faults can be identified qualitatively. To quantify sensor faults, the Kullback–Leibler divergence between $\boldsymbol{y_{pre}}$ and $\boldsymbol{y_{on}}$ is calculated. Together with a dynamic threshold that depends upon the distribution range of $\boldsymbol{y_{pre}}$, the sensor reliability is calculated. Graph c shows the results of this quantification approach.

The batch process chosen for illustration shows many deviations from the expected turbidity reading. These deviations could result from foaming and/or coalescing bubbles on the probe tip, as described by Gregory and Thornhill [31], and lead to significant discrepancies between expected and observed turbidity readings, especially at 0–12 h, 23–26 h, and after 34 h. Biological reasons for the oscillating signal at the beginning of the process (0–4 h) can be excluded due to the significantly higher time constants for biomass growth and decline, as well as offline reference measurements for dry cell weight (data not shown).

Conclusions

This novel approach to online sensor validation uses process time-dependent predictions of a sensor reading based on the whole sensor network's information to calculate a symptom signal and thus indicate sensor faults. The underlying principle is based upon the information redundancy inherent in the sensor network's data. It is supposed that the readings of the whole sensor network are more trustworthy than those of the single sensor to be supervised and validated.

The proof-of-concept is exemplarily shown for a turbidity sensor that is used to monitor a *P. pastoris*-batch process. The novel approach allows sensor faults such as bias, precision degradation, and complete failure to be detected. Besides that, it also allows unexpected deviations of turbidity readings to be detected. Such deviations were in this context defined as deviations of the correlation of biomass to turbidity that could not be explained by any other available online or offline process analysis. Soft sensors for biomass concentration that are based on turbidity measurements would lose their predictive performance in the case of sensor faults or unexpected deviations. The turbidity sensor was successfully supervised, and significant deviations from the expected turbidity readings were indicated.

The direct usage of the BPSO prediction instead of the original reading might appear obvious. The presented approach for the prediction of a certain target value based upon sensor network data can be used directly for the development of soft sensors for, e.g., biomass concentration. In many cases, however, the hardware turbidity sensor used for quality control (e.g., via a biomass soft sensor) is part of the accepted process validation and thus cannot be replaced by a prediction. The developed online validation system can in these cases be used to verify the results of a soft sensor.

The transferability of this online validation system to batch processes with other production organisms such as *Saccharomyces cerevisiae*, *Escherichia coli*, or mammalian cells, as well as different process parameters, must be investigated in future research. It is supposed that this approach is transferable to sensors for the monitoring of any kind of biological batch process provided that a minimal degree of redundancy (mathematically expressed: collinearity) exists in the sensor network data. Defining this minimum will be the main challenge for future studies.

The algorithmic structure is designed in such a way that computational power is mainly needed offline (Fig. 2, left branch) for model selection via the BPSO algorithm, whereas the actual online validation (Fig. 2, right branch) is not computationally intensive. This would, for instance, allow the online validation system to be run on embedded systems with low computational power.

The presented approach is subject to two major constraints. First, an experienced process expert is essential to exclude sensor faults in the historical process data used for model selection. Second, major process faults (e.g., contamination) or varying process settings could cause misleading predictions and thus malfunction of the sensor validation system. If, however, these constraints are considered, the presented approach can be used for reliable online sensor validation. This is especially important when the sensor data are used as input to a process control system or to soft sensors for CQA, as with a turbidity sensor used to predict biomass concentration. The present approach thus contributes to the PAT toolbox and will

help drive the acceptance of online sensors for quality control in the biotechnology industry.

Acknowledgments This study was funded by the German Federal Ministry of Education and Research (grant number 031B0475E) and the German Research Foundation (grant number BE 2245/17-1).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Nicoletti M, Jain L, Giordano R. Computational intelligence techniques as tools for bioprocess modelling, optimization, supervision and control. Computational intelligence techniques for bioprocess modelling, supervision and control. Berlin: Springer; 2009. p. 1–23.
2. Sharma AB, Golubchik L, Govindan R. Sensor faults: detection methods and prevalence in real-world datasets. ACM TOSN. 2010;6(3):23.
3. Balaban E, Saxena A, Bansal P, Goebel KF, Curran S. Modeling, detection, and disambiguation of sensor faults for aerospace applications. IEEE Sensors J. 2009;9(12):1907–17.
4. Mehranbod N, Soroush M, Piovoso M, Ogunnaike BA. Probabilistic model for sensor fault detection and identification. AIChE J. 2003;49(7):1787–802.
5. Cha Y-J, Agrawal AK. Robustness studies of sensor faults and noises for semi-active control strategies using large-scale magnetorheological dampers. J Vib Control. 2016;22(5):1228–43.
6. Zhang HQ, Yan Y. A wavelet-based approach to abrupt fault detection and diagnosis of sensors. IEEE Trans Instrum Meas. 2001;50(5):1389–96.
7. Feital T, Pinto JC. Use of variance spectra for in-line validation of process measurements in continuous processes. Can J Chem Eng. 2015;93(8):1426–37.
8. Becker T, Breithaupt D, Doelle HW, Fiechter A, Schlegel G, Shimizu S, et al. Biotechnology, 5. Monitoring and modeling of bioprocesses. Ullmann's encyclopedia of industrial chemistry. 2009.
9. Das A, Maiti J, Banerjee R. Process monitoring and fault detection strategies: a review. Int J Qual Reliab Manag. 2012;29(7):720–52.
10. Isermann R. Fault-diagnosis systems: an introduction from fault detection to fault tolerance: Springer Science & Business Media; 2006.
11. Kourti T. Application of latent variable methods to process control and multivariate statistical process control in industry. Int J Adapt Control Signal Process. 2005;19(4):213–46.
12. Goulding PR, Lennox B, Sandoz DJ, Smith KJ, Marjanovic O. Fault detection in continuous processes using multivariate statistical methods. Int J Syst Sci. 2000;31(11):1459–71.
13. Krause D, Hussein M, Becker T. Online monitoring of bioprocesses via multivariate sensor prediction within swarm intelligence decision making. Chemom Intell Lab Syst. 2015;145:48–59.
14. Mehranbod N, Soroush M, Panjapornpon C. A method of sensor fault detection and identification. J Process Control. 2005;15(3):321–39.
15. Guo T-H, Nurre J, editors. Sensor failure detection and recovery by neural networks. Neural Networks, 1991., IJCNN-91-Seattle international joint conference on; 1991: IEEE.
16. Zarei J, Shokri E. Robust sensor fault detection based on nonlinear unknown input observer. Measurement. 2014;48:355–67.
17. Dunia R, Qin SJ, Edgar TF, McAvoy TJ. Identification of faulty sensors using principal component analysis. AIChE J. 1996;42(10):2797–812.
18. Alag S, Agogino AM, Morjaria M. A methodology for intelligent sensor measurement, validation, fusion, and fault detection for equipment monitoring and diagnostics. AI EDAM. 2001;15(4):307–20.
19. Ibargüengoytia PH, Vadera S, Sucar LE. A probabilistic model for information and sensor validation. Comput J. 2005;49(1):113–26.
20. Frank PM. Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy: a survey and some new results. Automatica. 1990;26(3):459–74.
21. Khanesar MA, Teshnehlab M, Shoorehdeli MA, editors. A novel binary particle swarm optimization. Control & Automation, 2007. MED'07. Mediterranean conference on; 2007: IEEE.
22. Harms J, Wang X, Kim T, Yang X, Rathore AS. Defining process design space for biotech products: case study of *Pichia pastoris* fermentation. Biotechnol Prog. 2008;24(3):655–62.
23. Stratton J, Chiruvolu V, Meagher M. High cell-density fermentation. *Pichia* protocols. Berlin: Springer; 1998. p. 107–20.
24. Ündey C, Williams BA, Cınar A. Monitoring of batch pharmaceutical fermentations: data synchronization, landmark alignment, and real-time monitoring. IFAC Proc Vol. 2002;35(1):271–6.
25. Chong I-G, Jun C-H. Performance of some variable selection methods when multicollinearity is present. Chemom Intell Lab Syst. 2005;78(1):103–12.
26. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst. 2001;58(2):109–30.
27. Kennedy J, Eberhart RC, editors. A discrete binary version of the particle swarm algorithm. Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE international conference on; 1997: IEEE.
28. Del Valle Y, Venayagamoorthy GK, Mohagheghi S, Hernandez J-C, Harley RG. Particle swarm optimization: basic concepts, variants and applications in power systems. IEEE Trans Evol Comput. 2008;12(2):171–95.
29. Shi Y, Eberhart RC, editors. Parameter selection in particle swarm optimization. International conference on evolutionary programming. Berlin: Springer; 1998.
30. Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat. 1951;22(1):79–86.
31. Gregory M, Thornhill N. The effects of aeration and agitation on the measurement of yeast biomass using a laser turbidity probe. Bioprocess Eng. 1997;16(6):339–44.

4 Discussion

The core function of soft sensors is to derive utilizable information (output) from process data (input). Soft sensors based purely on data-driven prediction models provide exactly this core function. Soft sensors based on mechanistic or hybrid prediction models additionally use process knowledge to assist in the compression of process data to information. The process data are products of observations made in the real process environment and thus subject to uncertainties such as sensor faults. However, faulty inputs to a soft sensor will so far most likely lead to faulty outputs.
Within the scope of this thesis, the key challenges in the development of soft sensors for bioprocesses were first identified (first thesis publication, section 3.1, Brunner et al. (2021)). The challenges of variable process lengths, multiple process phases, and sensor faults were discussed along with corresponding solution approaches. In conclusion, the availability of process knowledge plays a crucial role in selecting the appropriate approaches for handling variable process lengths and multiple process phases. Moreover, there is still a research gap regarding the validation of the input data to soft sensors for the application area of bioprocesses. Another identified key challenge in this context is the tolerance of soft sensors to sensor faults. Fault-tolerant soft sensors are addressed at the end of this discussion.
With this in mind, the following three research questions were investigated for the use case of a *P. pastoris* bioprocess: How can a soft sensor model be utilized to generate process knowledge? How can process knowledge be implemented to develop a soft sensor model? How can uncertain model inputs be validated prior to their use in a soft sensor? The solution approaches to these questions form building blocks with which the gaps between uncertain process data and knowledge can be filled.

The **first building block** provided here comprises an approach to derive process knowledge as a quasi by-product of soft sensor development (second thesis publication, section 3.2, Brunner et al. (2016)). For this purpose, the model inputs of a biomass soft sensor based on a PLSR model were analyzed with respect to their weighting. The model inputs comprised four single-wavelength fluorescence measurements corresponding to the biogenic fluorophores tryptophan (ex/em 290/350 nm), NAD(P)H (ex/em 350/450 nm), and riboflavin (ex/em 370/530 and 450/530 nm, respectively). The selection of these wavelength pairs was based on the study by Surribas et al. (2006a) on monitoring a *P. pastoris* bioprocess using 2D fluorescence spectroscopy. The weighting in the present study was quantified by means of VIP scores, which can give an indication of the importance of the variables in the model (Wold et al., 2001; Chong and Jun, 2005). Evaluation of VIP scores shows that tryptophan has, of all fluorophores investigated, the highest weighting in the PLSR model. In addition, the importance of the four aforementioned input variables was evaluated in an MLR-based soft sensor for biomass using the backward elimination

method based on correlation coefficients (Pires et al., 2008). Here, the results also indicate tryptophan as the most relevant input variable. The importance of tryptophan for biomass prediction in *P. pastoris* bioprocesses is consistent with the above mentioned study of Surribas et al. (2006a). These authors used complete 2D fluorescence spectra (ex/em 270–550/310–590 nm), whereas only four wavelength pairs were used in the present study. One goal of the present study was to show that biomass concentration can also be predicted using only these four wavelength pairs instead of the whole 2D fluorescence spectrum.

Regarding the use of MLR as modeling method in this study, the following should be noted: Only four variables were used as inputs to the biomass soft sensor. This small number of inputs argues for the use of MLR. However, overfitting is likely when using MLR, especially when (multi)collinearity is present in the data. The MLR soft sensor is therefore relatively sensitive to outliers and noise in the input data, which can lead to drawbacks in online applications. Thus, for online applications, the soft sensors based on PLSR or PCR would be more suitable, as these methods can better handle (multi)collinearity.

The broader goal of the present study was to quantify the knowledge of the relationships between the selected fluorophores and biomass concentration. It was shown that VIP scores and backward elimination are suitable methods to reveal and quantify this knowledge during the development process of soft sensors. In the present study, a modest number of input variables (four) were ranked according to their importance in the prediction model. In cases where it is necessary to determine the relevant input variables from a larger number of input variables, swarm intelligence methods such as the ant colony optimization algorithm can be used to accelerate the selection process (Ranzan et al., 2014).

The knowledge generated in this way can be used to select relevant wavelength pairs for biomass monitoring in low-cost alternatives to 2D fluorescence spectroscopy. In these fluorometers (e.g., spectromex® ATFM200, Aquasant Messtechnik AG, Bubendorf, Switzerland), the number of LEDs (light emitting diodes) or excitation wavelengths, respectively, is technically and thus also economically limited. Furthermore, the relevant wavelength pairs for biomass prediction may vary between different cultivation strategies and organisms (Faassen and Hitzmann, 2015). The quantified knowledge, i.e., the rating on the importance of excitation wavelengths for biomass prediction, can help here in selecting the relevant LEDs. This in turn leads to more efficient process monitoring, as only what is really important is actually measured.

In summary, the first building block contributes to the generation of knowledge by data-driven tools. It further contributes to the challenge of variable selection, which is an important step of soft sensor development. Both together allow to reduce overfitting in model training and to develop more robust soft sensors. In addition, the knowledge obtained in this way is available in quantified form. Together with existing expert knowledge, it can become part of the manufacturing knowledge base required within the QbD framework.

The **second building block** serves to implement process knowledge in a hybrid model for a bioprocess with multiple process phases (third thesis publication, section 3.3, Brunner et al. (2020)). Process knowledge was implemented here at two levels of the soft sensor: first, in the phase detection algorithm; second, in the hybrid prediction model for biomass concentration. The transitions between the phases were determined automatically and the corresponding set of model coefficients was selected for the prediction model. The biomass soft sensor thus adapts to the individual process phases.

The *P. pastoris* bioprocess studied consisted of three process phases with different substrate and thus different metabolism: batch phase on glycerol (biomass generation), transition phase without substrate, and fed-batch phase on methanol (product formation). The main task of the phase detection algorithm was to detect the end of the batch phase, i.e., the complete consumption of glycerol. The end of the transition phase, i.e., the beginning of the fed-batch phase, was predefined in the automation system and therefore did not have to be detected first.

Two of the three triggers for detecting the complete consumption of glycerol were based on the trajectory of the off-gas CO_2 concentration (absolute value σ_{CO2} and first derivative $d\sigma_{CO2}/dt$). However, since sensor faults or minor process deviations can have a serious effect on the correct functioning of this trigger—especially in the case of $d\sigma_{CO2}/dt$—a knowledge-based safeguard measure was implemented in the phase detection algorithm. For this purpose, the amount of base consumed V_{base} was implemented upstream of the two triggers mentioned. This trigger is based on knowledge of the stoichiometric relationship between the constant starting concentration of glycerol (S_0 = 40 g L^{-1}) and the maximum amount of V_{base} when glycerol is fully consumed. While it was not possible to determine a precise value for the final amount of base consumed (V_{base} = 540 ± 117 mL) due to the buffering effect of the medium, no extreme outliers were expected for V_{base} based on process experience. Implementing this knowledge makes the phase detection algorithm more robust to faults of the CO_2 off-gas sensor. Alternatives to this hybrid of knowledge-based and trajectory-based phase detection would be, for example, purely trajectory-based or correlation-based methods of phase detection and division (Brunner et al., 2021). Of the purely trajectory-based methods, for example, online variants of dynamic time warping (DTW), such as extrapolative time warping (Srinivasan and Qian, 2005, 2007) or relaxed-greedy time warping (González-Martínez et al., 2011) might be suitable in this case. The aforementioned methods not only allow the phases to be detected and separated, but also compensate for variable process lengths. The same is true for correlation-based methods as presented by Lu et al. (2004). In this study, the correlation structure was represented by loading matrices of moving-window PCA or PLS submodels; the process phases were then determined via k-means clustering. Future research needs to investigate whether these methods can achieve similar robustness of phase detection as the described approach.

Depending on the current process phase, different prediction models were trained (offline) and used to predict biomass concentration (online). Process knowledge was implemented into the soft sensor models via a carbon balance. In this carbon balance, the system boundary was the bioreactor system: carbon left the bioreactor only in the

form of CO_2. In the fed-batch phase, the only carbon influx was in the form of methanol. The carbon in the bioreactor was either in the form of substrate (glycerol or methanol) or bound in the form of cell mass, extracellular protein (negligible), and acids (e.g., H_2CO_3). The latter could not be measured online. Since the information about the proportion of carbon bound in acids was missing (and not constant throughout the cultivation), the biomass concentration could not be directly inferred from the carbon balance. However, information about the formation of acids was available at least indirectly via V_{base}, since the added base serves to correct the pH. Mechanistically linking the carbon balance to V_{base} was impracticable due to the aforementioned buffering effect of the medium. Therefore, the model output of the carbon balance was linked to the indirect information about the acid formation rate (V_{base}) via MLR. The present hybrid model thus shows a serial structure of mechanistic (carbon balance) and data-driven (MLR) parts (Stosch et al., 2014; Solle et al., 2017).
In summary, the second building block contributes to the implementation of process knowledge in an adaptive soft sensor. Process knowledge here not only enables biomass prediction via a serial hybrid model, but also makes the phase detection algorithm more robust. In addition, it was confirmed that bioprocess data that are often standard, such as off-gas CO_2 and base consumption (Jenzsch et al., 2006; Grigs et al., 2021), have a value for biomass monitoring that should not be underestimated compared to more expensive spectrometer data. As with the first building block, the aim was again to develop a robust soft sensor with as little input data as possible.

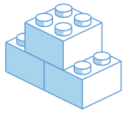
The **third building block** aims to detect sensor faults in bioprocesses with variable process lengths (fourth thesis publication, section 3.4, Brunner et al. (2019)). As mentioned, the validation of soft sensor inputs is a key challenge, as there are comparatively few approaches in literature, but the problem is ubiquitous. In bioprocesses, the validation of inputs, i.e., the detection of sensor faults, is further impeded because bioprocesses often vary in length and show time-variant behavior. In these cases, constant or purely time-dependent thresholds for sensor fault detection (Armaou and Demetriou, 2008) can lead to false-positive fault detections (false alarms) or concealment of faults. This means that variable process lengths add complexity to the problem of fault detection. In the presented study, the threshold as well as the whole fault detection algorithm were therefore designed to be dynamic, i.e., dependent on the current process section. The turbidity readings obtained during the batch phase of *P. pastoris* bioprocesses were used for the proof-of-concept.
In addition to physical redundancy, there are generally three different approaches for the detection of sensor faults, all of which are based on information redundancy in the sensor network (Brunner et al., 2021): symptom signal methods, methods based on variable contribution in a MSPC model, and pattern recognition methods. In the here used symptom signal method, residuals (symptom signals) are formed between the original sensor reading and a prediction of the sensor reading. To increase resistance to outliers in the predictors and make fault detection more robust, distributions of original and predicted sensor readings were compared instead of single values. The distributions of the original readings were obtained using a moving-window approach. The distributions of the predictions were determined via the binary particle swarm

optimization (BPSO) algorithm. This algorithm selected the 25 best prediction models out of a pool of various PLSR-models for each process section (maximum number of possible model combinations: 1.51×10^{23}) according to a cost function that was to be minimized. The cost function included the prediction error, the number of model inputs, and the number of latent variables of the PLSR models. This regularization approach penalized complex models—and thus the tendency to overfitting. The Kullback–Leibler divergence (Kullback and Leibler, 1951) between the distributions of the original and the predicted sensor readings indicated a sensor fault and was used to classify the sensor reading as reliable or faulty. The threshold for this classification was computed based on the changing confidence interval of the 25 predictions. A less accurate prediction resulted in a wider confidence interval and thus an increased threshold for fault detection. This dynamic design of the threshold allows false alarms to be prevented.

As indicated, the distributions of the predictions for the turbidity sensor were determined for each process section. The current process sections, and thus the correct model pool for the BPSO search, were determined online using a separate maturity index model. This approach was taken to cope with the variable process lengths and the time-variant behavior of the batch process. Alternatives to compensate for variable process lengths via the here used indicator variable (maturity index) are DTW and curve registration techniques (Brunner et al., 2021). These alternatives offer advantages over indicator variable in case of multiphase processes, since the information about landmarks is utilized during data synchronization (Bigot, 2006; González-Martínez et al., 2014). However, in the investigated batch process without prominent landmarks (peaks, etc.), these advantages would not have come into effect.

Although the presented algorithm contains some strongly data-driven parts (e.g., automatic model selection via BPSO), it must be noted that a process expert still has to ensure that there are no significant sensor faults in all training datasets. In contrast, there are approaches where only one fault-free dataset is required to develop fault detection algorithms. It was shown by Guo and Nurre (1991) for a space shuttle engine that an ANN can be trained to detect sensor faults by using one process dataset to which artificial random Gaussian noise was applied. This pattern recognition method for detecting sensor faults can in principle also be applied to bioprocesses (data not shown), but further research is needed here.

In summary, the third building block contributes to the reliability of, and thus confidence in, soft sensors for bioprocesses. Studies on the related topics of validation of soft sensor inputs and detection of sensor faults in bioprocesses are rare. Even fewer studies exist on the additional challenges of variable process lengths or time-variant process behavior (Huang et al., 2002; Krause et al., 2015). This study thus contributes to filling this gap in the field of soft sensor development. Moreover, it can be used as a starting point for another research objective: fault tolerance of soft sensors.



The three building blocks serve the main objective stated in the thesis outline (section 1.4): the provision of novel concepts to fill the gaps between uncertain process data and knowledge within soft sensor development. The broader view on these novel concepts allows for the following **main conclusions**.
In the field of soft sensor development, there is no “Swiss Army knife” method that combines all the desired functions. The development of soft sensors tends to be more like the selection of suitable building blocks from a construction set. A wealth of different methods exists for each step of the soft sensor development workflow—some of which have been evaluated for bioprocesses, some of which have not. The first thesis publication (section 3.1, Brunner et al. (2021)) discusses the extent to which methods that address the challenges of variable process length, multiple process phases, and sensor faults can be applied to bioprocesses. From both this review and the third (section 3.3, Brunner et al. (2020)) and fourth (section 3.4, Brunner et al. (2019)) thesis publications, the following can be concluded: The availability of process knowledge influences the choice of the preprocessing and modeling method; the choice of the preprocessing and modeling method in turn influences the robustness and accuracy of the soft sensor. Against this background, it seems all the more important to have methods for generating process knowledge as a quasi by-product of soft sensor development (first thesis publication, section 3.2, Brunner et al. (2016)).
The second main conclusion relates to the base of the DIK pyramid presented at the beginning of this thesis. The data used for process monitoring and as input to soft sensors are subject to uncertainties. Ignoring these uncertainties causes the accuracy of soft sensors to decrease if the inputs are erroneous ("garbage in, garbage out" principle). Besides ignoring, there are two possible strategies to deal with uncertainties in the input data.
The first strategy is to detect the faults in the input data in order to generate an alarm in the process control or data management system. For this strategy, the building block described above was provided (fourth thesis publication, section 3.4, Brunner et al. (2019)). This alarm information or quality rating (e.g., reliable/faulty), respectively, can be attached to the sensor reading as metadata during signal transmission (e.g., via OPC UA, open platform communications unified architecture). In case an examination of these quality tags of the input data is implemented in the soft sensor algorithm, the output of the soft sensor can be provided with another quality tag. This can assist in deciding whether the soft sensor output is to be used as input to an inferential controller or for real time release (Mandenius and Gustavsson, 2015).

The second strategy to deal with uncertainties in soft sensor input data is fault tolerance. A **brief outlook** on this so far almost unresearched area of soft sensor development is given in the following.
Fault-tolerant soft sensors *compensate* for faults, as opposed to just *detecting* them. Thus, soft sensors as PAT tool would remain operational in case of faulty inputs. In the first thesis publication (section 3.1, Brunner et al. (2021)), two different approaches are described on how fault tolerance can be realized for soft sensors. In the first variant, sensor faults are first detected and then compensated for by reconstructing the faulty sensor reading (Huang et al., 2002). The soft sensor model remains unchanged, since

the fault is already corrected at the input layer. To implement this variant, the approach presented in the fourth thesis publication (section 3.4, Brunner et al. (2019)) would have to be supplemented by an algorithmic solution that manages the reconstruction of the faulty reading. In the second variant, the fault is compensated by the soft sensor model itself. Here, the faulty input layer remains unchanged, but the model adapts in a way that compensates for the fault. In Krause et al. (2015), fault tolerance was integrated into a soft sensor model via a regularization approach: model inputs that deviated significantly from the historical data were penalized during model building. The contribution of faulty inputs was drastically reduced; the resulting soft sensor prediction can thus be considered fault tolerant.
Both approaches to fault tolerance of soft sensors have advantages and disadvantages. With the first variant (fault tolerance module at the inputs), the fault-compensated process data can be used for any monitoring or control purposes, not just as an input to a soft sensor. With the second variant (fault tolerance integrated in the model), this positive side effect does not exist. However, it is assumed that the regularization of faulty model inputs is algorithmically less complex to implement than taking the detour via fault-compensated inputs. To provide a sound comparison of these two variants, much more studies are needed on fault-tolerant soft sensors—especially in the field of bioprocesses with the aforementioned challenges. Although this thesis provides one building block for achieving the goal of fault-tolerant soft sensors, there is still a considerable need for research in this area.

In summary, soft sensors offer many possibilities to link the domains of uncertain process data, information, and knowledge. By extending the core function of deriving information from data, soft sensors can be designed to be more reliable (validated model inputs and knowledge-based predictions) and even assist in knowledge generation. Soft sensors thus not only serve the goals of PAT in numerous ways (Mandenius and Gustavsson, 2015), but also assist in building up and utilizing the manufacturing knowledge base required in QbD environments.

References

Ackoff, R. L. (1989). From data to wisdom. *Journal of applied systems analysis* 16, 3–9.

Adikaram, K. K. L. B., Hussein, M. A., Effenberger, M., and Becker, T. (2015). Data Transformation Technique to Improve the Outlier Detection Power of Grubbs' Test for Data Expected to Follow Linear Relation. *Journal of Applied Mathematics* 2015, 1–9.

Armaou, A., and Demetriou, M. A. (2008). Robust detection and accommodation of incipient component and actuator faults in nonlinear distributed processes. *AIChE J.* 54, 2651–2662.

Bidar, B., Shahraki, F., Sadeghi, J., and Khalilipour, M. M. (2018). Soft sensor modeling based on multi-state-dependent parameter models and application for quality monitoring in industrial sulfur recovery process. *IEEE Sensors Journal* 18, 4583–4591.

Biechele, P., Busse, C., Solle, D., Scheper, T., and Reardon, K. (2015). Sensor systems for bioprocess monitoring. *Eng. Life Sci.* 15, 469–488. doi: 10.1002/elsc.201500014

Bigot, J. (2006). Landmark-Based Registration of Curves via the Continuous Wavelet Transform. *Journal of Computational and Graphical Statistics* 15, 542–564.

Broger, T., Odermatt, R. P., Huber, P., and Sonnleitner, B. (2011). Real-time on-line flow cytometry for bioprocess monitoring. *Journal of Biotechnology* 154, 240–247.

Brunner, V., Hussein, M., and Becker, T. (2016). Biomass estimation in Pichia pastoris cultures by combined single-wavelength fluorescence measurements. *Biotechnology and bioengineering* 113, 2394–2402.

Brunner, V., Klöckner, L., Kerpes, R., Geier, D. U., and Becker, T. (2019). Online sensor validation in sensor networks for bioprocess monitoring using swarm intelligence. *Analytical and bioanalytical chemistry*, 2165–2175.

Brunner, V., Siegl, M., Geier, D., and Becker, T. (2020). Biomass soft sensor for a Pichia pastoris fed-batch process based on phase detection and hybrid modeling. *Biotechnology and bioengineering* 117, 2749–2759.

Brunner, V., Siegl, M., Geier, D., and Becker, T. (2021). Challenges in the Development of Soft Sensors for Bioprocesses: A Critical Review. *Frontiers in bioengineering and biotechnology* 9, 722202.

Chong, I.-G., and Jun, C.-H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems* 78, 103–112.

Davenport, T. H., and Prusak, L. (1998). *Working knowledge: How organizations manage what they know*. Harvard Business Press.

Davies, L., and Gather, U. (1993). The identification of multiple outliers. *Journal of the American Statistical Association* 88, 782–792.

Faassen, S. M., and Hitzmann, B. (2015). Fluorescence spectroscopy and chemometric modeling for bioprocess monitoring. *Sensors (Basel)* 15, 10271–10291.

FDA (2004). Guidance for industry, PAT-A Framework for Innovative Pharmaceutical Development, Manufacturing and Quality Assurance. *http://www.fda.gov/cder/guidance/published.html*.

González-Martínez, J. M., Ferrer, A., and Westerhuis, J. A. (2011). Real-time synchronization of batch trajectories for on-line multivariate statistical process control using Dynamic Time Warping. *Chemometrics and Intelligent Laboratory Systems* 105, 195–206.

González-Martínez, J. M., Noord, O. E. de, and Ferrer, A. (2014). Multisynchro: a novel approach for batch synchronization in scenarios of multiple asynchronisms. *Journal of chemometrics* 28, 462–475.

Grigs, O., Bolmanis, E., and Galvanauskas, V. (2021). Application of In-Situ and Soft-Sensors for Estimation of Recombinant P. pastoris GS115 Biomass Concentration: A Case Analysis of HBcAg (Mut+) and HBsAg (MutS) Production Processes under Varying Conditions. *Sensors (Basel)* 21.

Gunther, J. C., Baclaski, J., Seborg, D. E., and Conner, J. S. (2009). Pattern matching in batch bioprocesses—comparisons across multiple products and operating conditions. *Computers & Chemical Engineering* 33, 88–96.

Guo, T.-H., and Nurre, J. (1991). Sensor failure detection and recovery by neural networks.

Harms, J., Wang, X., Kim, T., Yang, X., and Rathore, A. S. (2008). Defining process design space for biotech products: case study of Pichia pastoris fermentation. *Biotechnology progress* 24, 655–662.

Harms, P., Kostov, Y., and Rao, G. (2002). Bioprocess monitoring. *Current Opinion in Biotechnology* 13, 124–127.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences* 44, 1–12.

Herwig, C., Garcia-Aponte, O. F., Golabgir, A., and Rathore, A. S. (2015). Knowledge management in the QbD paradigm: manufacturing of biotech therapeutics. *Trends Biotechnol* 33, 381–387. doi: 10.1016/j.tibtech.2015.04.004

Hocalar, A., Türker, M., Karakuzu, C., and Yüzgeç, U. (2011). Comparison of different estimation techniques for biomass concentration in large scale yeast fermentation. *ISA transactions* 50, 303–314.

Huang, J., Shimizu, H., and Shioya, S. (2002). Data preprocessing and output evaluation of an autoassociative neural network model for online fault detection in virginiamycin production. *Journal of bioscience and bioengineering* 94, 70–77.

Jennex, M. E. (2017). Big data, the internet of things, and the revised knowledge pyramid. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems* 48, 69–79.

Jenzsch, M., Simutis, R., Eisbrenner, G., Stückrath, I., and Lübbert, A. (2006). Estimation of biomass concentrations in fermentation processes for recombinant protein production. *Bioprocess and biosystems engineering* 29, 19–27.

Kadlec, P., Gabrys, B., and Strandt, S. (2009). Data-driven soft sensors in the process industry. *Computers & Chemical Engineering* 33, 795–814.

Kaneko, H., and Funatsu, K. (2012). A new process variable and dynamics selection method based on a genetic algorithm-based wavelength selection method. *AIChE J.* 58, 1829–1840.

Kiviharju, K., Salonen, K., Moilanen, U., and Eerikäinen, T. (2008). Biomass measurement online: the performance of in situ measurements and software sensors. *J Ind Microbiol Biotechnol* 35, 657–665.

Krause, D., Birle, S., Hussein, M. A., and Becker, T. (2011). Bioprocess monitoring and control via adaptive sensor calibration. *Eng. Life Sci.* 11, 402–416.

Krause, D., Hussein, M. A., and Becker, T. (2015). Online monitoring of bioprocesses via multivariate sensor prediction within swarm intelligence decision making. *Chemometrics and Intelligent Laboratory Systems* 145, 48–59.

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics* 22, 79–86.

Liang, S., Wang, B., Pan, L., Ye, Y., He, M., Han, S., et al. (2012). Comprehensive structural annotation of Pichia pastoris transcriptome and the response to various carbon sources using deep paired-end RNA sequencing. *BMC genomics* 13, 738.

Liu, W.-C., Inwood, S., Gong, T., Sharma, A., Yu, L.-Y., and Zhu, P. (2019). Fed-batch high-cell-density fermentation strategies for Pichia pastoris growth and production. *Crit Rev Biotechnol* 39, 258–271.

Lu, N., Gao, F., and Wang, F. (2004). Sub-PCA modeling and on-line monitoring strategy for batch processes. *AIChE J.* 50, 255–259.

Luttmann, R., Bracewell, D. G., Cornelissen, G., Gernaey, K. V., Glassey, J., Hass, V. C., et al. (2012). Soft sensors in bioprocessing: a status report and recommendations. *Biotechnol J* 7, 1040–1048. doi: 10.1002/biot.201100506

Ma, M.-D., Ko, J.-W., Wang, S.-J., Wu, M.-F., Jang, S.-S., Shieh, S.-S., et al. (2009). Development of adaptive soft sensor based on statistical identification of key variables. *Control Engineering Practice* 17, 1026–1034. doi: 10.1016/j.conengprac.2009.03.004

Mandenius, C.-F., and Gustavsson, R. (2015). Mini-review: soft sensors as means for PAT in the manufacture of bio-therapeutics. *J. Chem. Technol. Biotechnol.* 90, 215–227. doi: 10.1002/jctb.4477

Matero, S., van den Berg, F., Poutiainen, S., Rantanen, J., and Pajander, J. (2013). Towards better process understanding: chemometrics and multivariate measurements in manufacturing of solid dosage forms. *Journal of pharmaceutical sciences* 102, 1385–1403.

Meng, Y., Lan, Q., Qin, J., Yu, S., Pang, H., and Zheng, K. (2019). Data-driven soft sensor modeling based on twin support vector regression for cane sugar crystallization. *Journal of Food Engineering* 241, 159–165.

Monod, J. (1949). The Growth of Bacterial Cultures. *Annu. Rev. Microbiol.* 3, 371–394. doi: 10.1146/annurev.mi.03.100149.002103

Ohadi, K., Legge, R. L., and Budman, H. M. (2015). Development of a soft-sensor based on multi-wavelength fluorescence spectroscopy and a dynamic metabolic

model for monitoring mammalian cell cultures. *Biotechnology and bioengineering* 112, 197–208.

Paquet-Durand, O., Assawarajuwan, S., and Hitzmann, B. (2017). Artificial neural network for bioprocess monitoring based on fluorescence measurements: Training without offline measurements. *Engineering in Life Sciences* 17, 874–880.

Pires, J., Martins, F. G., Sousa, S., Alvim-Ferraz, M., and Pereira, M. C. (2008). Selection and validation of parameters in multiple linear and principal component regressions. *Environmental Modelling & Software* 23, 50–55.

Pla, I. A., Damasceno, L. M., Vannelli, T., Ritter, G., Batt, C. A., and Shuler, M. L. (2006). Evaluation of Mut+ and MutS Pichia pastoris phenotypes for high level extracellular scFv expression under feedback control of the methanol concentration. *Biotechnology progress* 22, 881–888.

Ranzan, C., Strohm, A., Ranzan, L., Trierweiler, L. F., Hitzmann, B., and Trierweiler, J. O. (2014). Wheat flour characterization using NIR and spectral filter based on Ant Colony Optimization. *Chemometrics and Intelligent Laboratory Systems* 132, 133–140.

Rathore, A. S. (2009). Roadmap for implementation of quality by design (QbD) for biotechnology products. *Trends Biotechnol* 27, 546–553.

Rathore, A. S., and Winkle, H. (2009). Quality by design for biopharmaceuticals. *Nature biotechnology* 27, 26–34.

Read, E. K., Shah, R. B., Riley, B. S., Park, J. T., Brorson, K. A., and Rathore, A. S. (2010). Process analytical technology (PAT) for biopharmaceutical products: Part II. Concepts and applications. *Biotechnology and bioengineering* 105, 285–295.

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of information science* 33, 163–180.

Sagmeister, P., Wechselberger, P., Jazini, M., Meitz, A., Langemann, T., and Herwig, C. (2013). Soft sensor assisted dynamic bioprocess control: Efficient tools for bioprocess development. *Chemical Engineering Science* 96, 190–198.

Shardt, Y. A. W., Hao, H., and Ding, S. X. (2015). A New Soft-Sensor-Based Process Monitoring Scheme Incorporating Infrequent KPI Measurements. *IEEE Trans. Ind. Electron.* 62, 3843–3851. doi: 10.1109/TIE.2014.2364561

Shyu, M.-L., Chen, S.-C., Sarinnapakorn, K., and Chang, L. (2006). “Principal component-based anomaly detection scheme,” in *Foundations and novel approaches in data mining* (Springer), 311–329.

Simon, L. L., Pataki, H., Marosi, G., Meemken, F., Hungerbühler, K., Baiker, A., et al. (2015). Assessment of Recent Process Analytical Technology (PAT) Trends: A Multiauthor Review. *Org. Process Res. Dev.* 19, 3–62.

Sokolov, M., Ritscher, J., MacKinnon, N., Bielser, J.-M., Brühlmann, D., Rothenhäusler, D., et al. (2017). Robust factor selection in early cell culture process development for the production of a biosimilar monoclonal antibody. *Biotechnology progress* 33, 181–191.

Sokolov, M., Soos, M., Neunstoecklin, B., Morbidelli, M., Butté, A., Leardi, R., et al. (2015). Fingerprint detection and process prediction by multivariate analysis of fed-batch monoclonal antibody cell culture data. *Biotechnology progress* 31, 1633–1644.

Solle, D., Hitzmann, B., Herwig, C., Pereira Remelhe, M., Ulonska, S., Wuerth, L., et al. (2017). Between the poles of data-driven and mechanistic modeling for process operation. *Chemie Ingenieur Technik* 89, 542–561.

Souza, F., and Araujo, R. (2011). "Variable and time-lag selection using empirical data," in *2011 IEEE 16th Conference on Emerging Technologies & Factory Automation: (ETFA 2010) ; Toulouse, France, 5 - 9 September 2011*, ed. Z. Mammeri (Piscataway, NJ: IEEE).

Souza, F. A. A., Araújo, R., and Mendes, J. (2016). Review of soft sensor methods for regression applications. *Chemometrics and Intelligent Laboratory Systems* 152, 69–79.

Srinivasan, R., and Qian, M. (2005). Off-line Temporal Signal Comparison Using Singular Points Augmented Time Warping. *Ind. Eng. Chem. Res.* 44, 4697–4716.

Srinivasan, R., and Qian, M. (2007). Online Temporal Signal Comparison Using Singular Points Augmented Time Warping. *Ind. Eng. Chem. Res.* 46, 4531–4548.

Stosch, M. von, Davy, S., Francois, K., Galvanauskas, V., Hamelink, J.-M., Luebbert, A., et al. (2014). Hybrid modeling for quality by design and PAT-benefits and challenges of applications in biopharmaceutical industry. *Biotechnol J* 9, 719–726.

Stratton, J., Chiruvolu, V., and Meagher, M. (1998). "High cell-density fermentation," in *Pichia protocols* (Springer), 107–120.

Streefland, M., Martens, D. E., Beuvery, E. C., and Wijffels, R. H. (2013). Process analytical technology (PAT) tools for the cultivation step in biopharmaceutical production. *Eng. Life Sci.* 13, 212–223.

Surribas, A., Geissler, D., Gierse, A., Scheper, T., Hitzmann, B., Montesinos, J. L., et al. (2006a). State variables monitoring by in situ multi-wavelength fluorescence spectroscopy in heterologous protein production by Pichia pastoris. *Journal of Biotechnology* 124, 412–419.

Surribas, A., Montesinos, J. L., and Valero, F. F. (2006b). Biomass estimation using fluorescence measurements in Pichia pastoris bioprocess. *J. Chem. Technol. Biotechnol.* 81, 23–28.

Tahir, F., Islam, M. T., Mack, J., Robertson, J., and Lovett, D. (2019). Process monitoring and fault detection on a hot-melt extrusion process using in-line Raman spectroscopy and a hybrid soft sensor. *Computers & Chemical Engineering* 125, 400–414.

Thomassen, Y. E., van Sprang, E. N. M., van der Pol, L. A., and Bakker, W. am (2010). Multivariate data analysis on historical IPV production data for better process understanding and future improvements. *Biotechnology and bioengineering* 107, 96–104.

Wang, Z. X., He, Q. P., and Wang, J. (2015). Comparison of variable selection methods for PLS-based soft sensor modeling. *Journal of Process Control* 26, 56–72. doi: 10.1016/j.jprocont.2015.01.003

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* 58, 109–130.

Wu, D., Yu, X. W., Wang, T. C., Wang, R., and Xu, Y. (2011). High yield Rhizopus chinenisis prolipase production in Pichia pastoris: impact of methanol concentration. *Biotechnology and Bioprocess Engineering* 16, 305–311.

Wu, Y., and Luo, X. (2010). A novel calibration approach of soft sensor based on multirate data fusion technology. *Journal of Process Control* 20, 1252–1260.

Yamashita, S., Yurimoto, H., Murakami, D., Yoshikawa, M., Oku, M., and Sakai, Y. (2009). Lag-phase autophagy in the methylotrophic yeast Pichia pastoris. *Genes to Cells* 14, 861–870.

Yang, Z., and Zhang, Z. (2018). Engineering strategies for enhanced production of protein and bio-products in Pichia pastoris: a review. *Biotechnol Adv* 36, 182–195.

Yousefi-Darani, A., Paquet-Durand, O., and Hitzmann, B. (2020). The Kalman Filter for the Supervision of Cultivation Processes. *Adv Biochem Eng Biotechnol*.

Zheng, J., and Song, Z. (2018). Semisupervised learning for probabilistic partial least squares regression model and soft sensor application. *Journal of Process Control* 64, 123–131.

Zhu, P., Liu, X., Wang, Y., and Yang, X. (2018). Mixture Semisupervised Bayesian Principal Component Regression for Soft Sensor Modeling. *IEEE Access* 6, 40909–40919.