Pharm Res (2017) 34:2596–2613
DOI 10.1007/s11095-017-2308-y

EXPERT REVIEW

Model-Based Methods in the Biopharmaceutical Process Lifecycle

Paul Kroll [1,2] • Alexandra Hofer [1] • Sophia Ulonska [1] • Julian Kager [1] • Christoph Herwig [1,2]

Received: 5 July 2017 / Accepted: 21 September 2017 / Published online: 22 November 2017

ABSTRACT Model-based methods are increasingly used in all areas of biopharmaceutical process technology. They can be applied in the field of experimental design, process characterization, process design, monitoring and control. Benefits of these methods are lower experimental effort, process transparency, clear rationality behind decisions and increased process robustness. The possibility of applying methods adopted from different scientific domains accelerates this trend further. In addition, model-based methods can help to implement regulatory requirements as suggested by recent Quality by Design and validation initiatives. The aim of this review is to give an overview of the state of the art of model-based methods, their applications, further challenges and possible solutions in the biopharmaceutical process life cycle. Today, despite these advantages, the potential of model-based methods is still not fully exhausted in bioprocess technology. This is due to a lack of (i) acceptance of the users, (ii) user-friendly tools provided by existing methods, (iii) implementation in existing process control systems and (iv) clear workflows to set up specific process models. We propose that model-based methods be applied throughout the lifecycle of a biopharmaceutical process, starting with the set-up of a process model, which is used for monitoring and control of process parameters, and ending with continuous and iterative process improvement via data mining techniques.

KEY WORDS bioprocess · data mining · modelling · monitoring · optimization

✉ Christoph Herwig
christoph.herwig@tuwien.ac.at

[1] Research Area Biochemical Engineering, Institute of Chemical Environmental and Biological Engineering, Vienna University of Technology, Gumpendorfer Straße 1a – 166/4, A-1060 Vienna, Austria

[2] Christian Doppler Laboratory for Mechanistic and Physiological Methods for Improved Bioprocesses, TU Wien, Vienna, Austria

ABBREVIATIONS

ANN	Artificial neural networks
CHO	Chinese hamster ovary
CMA	Critical material attributes
CMB-DoE	Continuous model-based experimental design
CPP	Critical process parameters
CQA	Critical quality attributes
DoE	Design of experiments
E.coli	*Escherichia coli*
HPLC	High performance liquid chromatography
IC	Ion chromatography
ICH	International Conference on Harmonization
ICP	Induced coupled plasma
iPLS	Interval partial least squares
IR	Infrared
kPP	Key process parameters
M^3C	Measurement, monitoring, modeling and control
MB-DoE	Model-based design of experiments
MIR	Mid-infrared
MLR	Multiple linear regression
MPC	Model predictive control
NIR	Near-infrared
NRMSE	Normalized root mean square error
OPLS-DA	Orthogonal partial least squares-discriminate analysis
OSC	Orthogonal signal correction
PAT	Process analytical technology
PCA	Principle component analysis
PLS	Partial least squares
PLS-DA	Partial least squares-discriminate analysis
QbD	Quality by design
QTPP	Quality target product profile
SQP	Sequential quadratic programming
SSE	Sum of square errors
SVM	Supported vector machine

INTRODUCTION

A safe product is the target of every production process. In the field of pharmaceutical products, this is ensured by elaborate approvals and the continuous control of independent authorities such as the Food and Drug Administration or the European Medicines Agency. The International Conference on Harmonisation (ICH) has established quality guidelines which should be considered during process lifecycle (1). A process lifecycle includes process development, scale up and continuous optimization until product discontinuation. The basis of a production process is a definition of the product by a quality target product profile (QTPP) that includes critical quality attributes (CQA) (2), such as physicochemical properties, biological activity, immunochemical properties, purity and impurities (3). The aim of each industrial production process is to satisfy the predetermined CQAs with a maximum of productivity. According to the ICH Q8(R2) guidelines the quality by design (QbD) approach is a one way of engineering an adequate production process. QbD combines sound science and quality risk management in order to identify critical material attributes (CMA) and critical process parameters (CPP), which shows significant effects on CQAs. In addition, the functional relationships between CMAs and CPPs on CQAs should be investigated (2). This requires the use of mathematical models within the framework of QbD. The basic idea of the QbD approach is that a process with controlled CMAs and CPPs in a defined design space will lead to continuous CQAs and finally to a sufficient QTPP. In order to achieve this goal process analytical technology (PAT) – tools are used. PAT includes the tasks of designing, analyzing and controlling production processes based on real-time monitoring of critical parameters including them CMAs, CPPs and CQAs (2,4).

In addition to adequate product quality, each process aims for high productivity. This includes the thoughtful use of raw materials, technologies and human resources in addition to the reduction of unwanted by-products. In contrast to CPPs which only affect product quality, key process parameters (kPP) affect productivity and economical viability (5). During the whole process lifecycle, CPPs and kPPs have to be improved in order to react to changed boundary conditions such as fluctuations in raw materials, new production facilities and locations, new technologies and constantly fluctuating staff. In summary, the following four challenges arise during a process lifecycle: I) generation of process knowledge, II) process monitoring, III) process optimization and IV) continuous improvement of the process (Fig. 1). In order to fulfill these challenges during the entire process lifecycle, a lifecycle management is indispensable. ICH Q8(R2) and Q12 address this issue but don't give any concrete solution (2,6). The reason for this is the lack of practicable knowledge management systems (7).

In order to solve the four previously presented challenges during the process lifecycle an overview of available methods and technologies is given in the present manuscript. The focus lies on model-based methods which are characterized by the use of mathematical models. Basically, each process model can be described by an Eq. [1], which is defined by a model output/outputs y, a function f, the time t, model states x, and the design vector φ, including all necessary process parameters (CPPs & kPPs) such as feed rates, temperature, pH etc., and the model parameters θ.

$$y = f(t, x, \varphi, \theta) \tag{1}$$

In order to show the interaction between the four separate challenges (I-IV), the red line of this paper will be analogous to a simple control loop (Fig. 1). This approach allows a scientific discussion of interaction and a possible outlook with respect to the process lifecycle.

The first challenge (challenge I) investigated is the identification of CPPs, kPPs and the generation of process knowledge. Process development and improvement can only occur if relationships and interactions are understood. With respect to model-based methods process relevant (critical) knowledge is defined as the sum of relationships and interactions, which should be considered in a process model in order to predict a target value (CPP, kPP or CQA). Modelling is a tool for the identification and description of these relationships with mathematical equations verified by statistic tests. In chapter 2.1, different modelling workflows will be presented and compared with respect to the modelling goal, complexity as well as transferability to biopharmaceutical production processes. The basis of the parametrization and verification of each model are data. Especially in model development, data have a major impact on model structure and validity space. Therefore, there is a strong iteration between modelling and data collection. In the second part of chapter 2.1 methods are described as to which data should be collected to verify the process model. The main output of this chapter are methodologies in order to generate adequate model structures f and model parameters θ, which can be adapted during the process lifecycle based on additional data and knowledge. The models and their parameters are necessary for further monitoring and control applications.

Since every process is affected by certain disturbances which affect quality and productivity, monitoring is a need for biopharmaceutical production processes (challenge II) (2). Monitoring is defined as the supervision of process parameters and variables, which is needed for subsequent control actions. Monitoring hereby includes the collection of information by measurements and subsequent data processing, whereas in the latter model-based methods can be applied. These methods

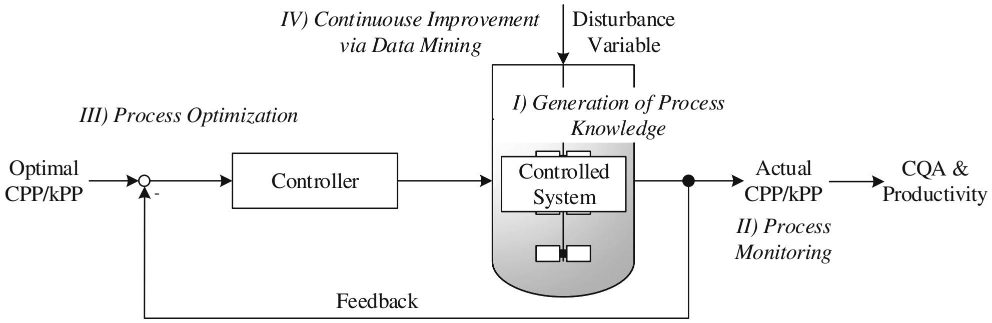

Fig. 1 A simple control loop with the related four challenges (I-IV) of process development and the process lifecycle. Challenge I is the generation and storage of knowledge within models. Challenge II is the process monitoring. Challenge III is the determination of optimal process conditions for different applications and IV the continuous improvement of a process by data mining tools.

and their application in monitoring will be discussed in chapter 2.2. It will focus on methods which help to define the needed measurements, allow the combination of multiple measurements to handle process noise and measurement uncertainty and finally allow the estimation of unmeasured states.

The final aim of process design is process control, discussed in chapter 2.3. Mainly two topics will be discussed within the challenge III) "Process Optimization". The first topic is a clear description of the control goal within certain boundaries that are based on product, technical, physiological and economic limits. Thereby various model-based methods for open-loop and closed-loop applications will be presented. With regard to model-based methods, methodologies for optimal and predictive control are presented. The second topic is estimating an "optimal" design vector and identifying critical process limitations, which provide an important input for further process optimization.

Certain disturbances that affect every process can be classified as a) known but neglected and b) unknown and neglected ones. Both can have a significant impact on process performance and should be continuously improved. This continuous improvement (challenge IV) is a key innovation motor for existing processes during their entire lifecycle. New analytical methods, measurement devices, automation, further data evaluation and others can lead to process relevant knowledge which should be taken into account. Within chapter 2.4 this continuous improvement of the process will be investigated. Regarding model-based methods the focus will be on data-mining tools, which allow researchers to set up hypotheses of potential correlations. These hypotheses are a necessary input for further process model-extensions and support the overall goal of an adequate product quality and high productivity throughout the entire process lifecycle.

Finally, an overall statement on further applications and perspectives of model-based methods within the biopharmaceutical process lifecycle is presented in the conclusion.

RESULTS & DISCUSSION

Generation of Process Knowledge

Modelling

Within the process lifecycle, knowledge is defined as the ability to describe relationships between (critical) process parameters and critical quality or performance attributes. This knowledge needs to be documented. The trend of the last years is clearly from a transfer approach, which is based on spoken and written words, towards a model approach (8). In the context of biopharmaceutical processes, this indicates the possible usage of process models as knowledge storage systems (9). The setup of these process models is still challenging. Contributions presenting workflows for modeling are increasing (10–15). According to good modelling practice, the single steps of modelling are always similar (14). These steps are: i) setup of a modelling project, ii) setup of a model, iii) analysis of the model. In addition, the documentation of the complete modelling project should be entire and transparent.

The basis of each modelling workflow is a clear definition of the model goal. This often resents a major challenge and cannot be achieved without iterations between modelers and project managers. The model goal should include the definition of target values, acceptance criteria and boundary conditions. Furthermore, the application of the model should be considered. Each process related model should be as simple as possible and as accurate as necessary. From this dogma, it follows that a model should only include necessary (critical) states, model parameters and process parameters. Depending on the goal of the model, different model types are suitable. Frequently used is the classification between data driven, mechanistic and hybrid models (16). In terms of applications of models, the classification between dynamic and static models is more appropriate. Dynamic models include differential equations, typically over time or location coordinates which allow prediction. Static models are correlations which cannot provide time-dependent simulation results. Hence, they are not applicable for prediction over time or location, which is commonly required in bioprocess development. Data driven, mechanistic as well as hybrid models can be both, static and dynamic.

The set up and analysis of a model are iterative steps within each modelling workflow (13,17), which are illustrated in Fig. 2. For the setup of models, different approaches are reported in literature. To date, experts are required to set up models, as they strongly depend on prior knowledge. This limited prior knowledge is a general gap for the application of all model-based methods. Only a few workflows for automated modelling are available. With regard to the process lifecycle, the focus of this review will be on these automated workflows for the setup of dynamic mechanistic models. A generic and strongly knowledge-driven approach is shown by the company Bayer AG (18,19): Based on an extensive dynamic metabolic flux model in combination with a generic algorithm, the initial complex model is reduced to the most necessary parts. The benefit of this top-down approach is the intense use of prior knowledge. The working group of King shows another approach, based on the detection of process events in combination with a model library (10,20,21). The benefit of this approach is that less prior knowledge is necessary and the transferability on other bioprocesses is given. As one of the drawbacks of model-based methods is the validation of models and there parameters an automated workflow for the generation of substantial target-oriented mechanistic process models was developed in our working group (22). This approach allows the generation and validation of process models with less prior knowledge and without model libraries.

The analysis of each model follows the same order. Based on collected data and an assumed model structure a parameter fit is performed. With the use of optimization algorithms, the model parameters are adapted in a way, that the previous defined descriptor is optimized (see chapter 2.3). Typically, this is a minimization of a model deviation, which can be described by different characteristics such as the sum of square errors (SSE), a normalized root mean square error (NRMSE), a profile likelihood or other descriptors. A comparison of the achieved descriptor with a previously defined acceptance criterion is the first analysis of each model. If this fails, the model structure is not suitable for the present issue. If the model passes, the model structure could be suitable to describe the relation. The next analysis is focused on the model parameters $\hat{\theta}$ and their deviations. Therefore, typically, an identifiability analysis is performed which follows two aims: The first aim is the structural identifiability of model parameters, which is necessary for process models with the aim of monitoring and control. If structural identifiability is not given, model parameters can compensate each other due to cross correlation. This results in multiple solutions and can lead to spurious results. There are several methods in order to evaluate structural identifiability (23,24). If structural identifiability is given, practical identifiability should be investigated in order to fulfil the second goal, which is a statement about confidence intervals of model parameters based on existing data (24,25). This is necessary in order to decide if parameters can be estimated with the data available. If practical identifiability is given, the model parameters are significant. If not, two statements can be made: i) the available data allows no determination of the model parameter or ii) the model structure allows cross correlations between model parameters and is therefore not as simple as possible.

In addition to the analysis of model and model parameter deviations, there is a variety of methods to characterize models with their focus on robustness. The first check should be a global behavior test with the goal of ensuring the right implementation of a model: here the model is tested with extreme input values. Additionally, if possible, certain redundancy should be implemented in the model (see chapter 2.2). Typical approaches are material balances, as they are typically used for yeast or microbial processes (13,26). Another frequently used method is a sensitivity analysis with the aim of

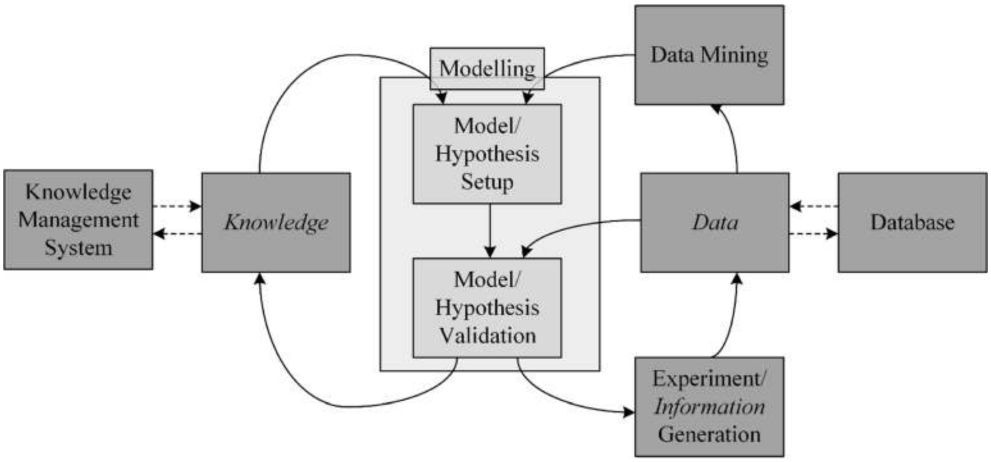
Modelling
Data Mining
Model/ Hypothesis Setup
Knowledge Management System
Knowledge
Data
Database
Model/ Hypothesis Validation
Experiment/ Information Generation

Fig. 2 Systematic overview of a model-development including interlinks between data, database and datamining, information and necessary experiments and knowledge.

showing the impact of deviations of model parameters, process parameters and model inputs on model outputs (27,28). The information obtained in this sensitivity analysis can be used to improve the model within the process lifecycle. With respect to the further usage of models certain causes for deviations must be considered. Deviations can be mainly obtained from two sources. The first source of deviations is the model structure in itself. Based on the principle that a model is always a sum of assumptions, there is always an accepted model deviation with a predefined validity space. In addition, models can always only explain a part of the process variance. Disturbances that are not considered in the model cannot be explained by it. Within the concept of process lifecycle this implies a continuous model improvement (see chapter 2.4). The second source of variance is the deviation of the model parameters $\hat{\theta}$ caused by changing and not explained sources of variance. This can be improved by adapting the model structure or parameters. For both additional information is necessary. It can be provided by additional data or additional hypotheses from data mining methodologies (see chapter 2.4) leading to new model structures (Fig. 2). With respect to real-time application, several methodologies for model adaption are shown in chapter 2.2.

Generation of Information

During the process development certain experiments must be performed in order to identify CPPs and an adequate design space and to verify process models. The most widely used strategy is the standard design of experiments (DoE) (29), which is given as an example in the guidelines ICH Q8(R2) (2). However, the applicability of standard DoEs for bioprocesses comprising a huge number of potential CPPs is not given to the full extent. The reason for this is mainly the model-based data evaluation, which typically only assumes linear or quadratic effects between process parameters and quality/product attributes. Known relationships describing physiological interactions are usually not taken into account in standard DoEs. Therefore other model-based methods are available which are based on information.

In order to verify certain process models, information is necessary. Within this context, information is defined as the possibility of estimating the model parameters θ of a model f with collected data. Mathematical statistics call this the Fisher information, which is described by the Fisher information matrix (H_θ). H_θ depends, besides the static model structure, on the model parameters θ and the design vector φ and can be estimated by Eq. [2] (30,31). The design vector includes all possible process parameters, which are considered in the model, and sampling points (t_k) where additional data are collected. H_θ^0 describes the initial fisher information matrix, n_{sp} the number of sampling points, N_y the number of model states y and N_θ the number of model parameters θ.

$$H_\theta(\theta,\varphi) = H_\theta^0 + \sum_{k=1}^{n_{sp}} \sum_{i=1}^{N_y} \sum_{j=1}^{N_y} s_{ij} \left[\frac{\partial \hat{y}_i(t_k)}{\partial \theta_l}^T \frac{\partial \hat{y}_j(t_k)}{\partial \theta_m} \right]_{l,m=1\ldots N_\theta} \qquad (2)$$

Applications for H_θ are mainly found in the model-based design of experiments (MB-DoE). An experiment has per definition the aim to prove, refute or confirm a hypothesis. In the case of MB-DoE the hypothesis is the process model in itself. Therefore, the information content of a planned experiment is maximized depending on φ. This information content should be a criterion extracted from H_θ. The D-optimal design which indicates a maximization of the determinant of H_θ is frequently used. Other descriptors include the maximization of the trace of H_θ (A-optimal) or the maximization of the smallest eigenvalue (E-optimal) (32). Telen *et al.* investigated additional criteria and showed the applicability of MB-DoE in order to estimate model parameters from a simulated fed-batch study (33). In addition, drawbacks of the single criteria are discussed and a novel multi-objective approach is investigated. This implies that MB-DoE strongly depends on the chosen information criteria. This must be transparent in order to ensure systematic and sound decisions. Table I shows some applications of MB-DoE and a summary of novel approaches to design criteria.

However, information is strongly coupled with the identifiability analysis of modelling workflows (see chapter 2.1.1) (44,47–49). As described there, available data are necessary in order to estimate identifiable parameters. MB-DoE is the model-based method solving this issue. Several publications show the application of MB-DoE in order to reduce the experimental effort with the goal of verifying process models. An issue for MB-DoE is the handling with uncertainties based on model and experimental deviations (50). One possibility is the real-time adaption of the experimental design, which is called continuous model-based experimental design (CMB-DoE) (39,40) or online optimal experimental re-design (41,42). This is finally a control issue and strongly related to process monitoring (see chapter 2.2) and optimization (see chapter 2.4).

Process Monitoring

Process monitoring is the description of the actual state of the process system in order to detect deflections of CPPs or key process parameters in time. With regard to the definition of PAT, monitoring without a feedback for process control is only measurement (51). Process monitoring can be seen in the context of measurement, monitoring, modeling and control (M^3C) (52). After describing the first tasks of monitoring

Table I Summary of Applications and Novel Publications with Respect to Model-Based Experimental Design

Method	Criteria	Application	Real-time	Reference
Application paper				
Signal to noise ratio	SNR = const	estimation of sampling points with respect to deviations on specific rates	at-line/ off-line	(34)
Sequential experimental design	D-criteria	experimental design within a model discrimination workflow	at-line/ off-line	(35,36)
Optimal dynamic experiments	–	MB-DoE in microbioreactor systems under use of FTIR spectroscopy as monitoring tool	at-line/ real-time	(37)
Simultaneous solution Approach for MB-DoE	A, D & E - criteria	design of feed rates and adaptive optimal sampling strategy	at-line/ off-line	(38)
CMB-DoE	A, D & E - criteria	adaption of a dynamic experiment under usage of real-time data control on information criteria	real-time	(39,40)
Online optimal experimental re-design	A-criteria	adaption of a dynamic experiment under usage of real-time data control on information criteria	real-time	(41,42)
Model discriminating experimental design	–	Model descrimination within an sequential workflow	at-line/ real-time	(43)
Design criteria paper				
D-optimal design		reduction of parameter interactions with MB-DoE under usage of a multi objective optimization criteria	at-line/ off-line	(44)
DMOO design (multi objective optimization)				
Multi objective approach		Multi-objective MB-DoE to descriminate between models and estimate kinetic parameters	at-line/ off-line	(45)
Anticorrelation criteria		anticorrelation criteria to estimate model parameters	at-line/ off-line	(46)

and the real-time data collection, model based methods for the subsequent data processing are presented, followed by the description of implemented examples in the field of biotechnology, which are also collected in Table II.

Measurements are a central part of monitoring as they provide the time resolved raw information of the ongoing process. Measurement methodologies and devices should be simple, robust and as accurate as necessary. Besides well-established measurements such as pH, dissolved oxygen and gas analysis, a vast amount of process analyzers is available nowadays; however the development of measurement techniques is still a field for extensive research in Biotechnology (53,54). These include - but are not limited to - chemical /biological measurements, which are characterized by a high sensitivity and by physical sensors mainly represented by spectroscopic technologies (UV/VIS-, IR-; dielectric-, RAMAN-spectroscopy) (55–58). In order to include process analyzers into monitoring it is not important whether measurements are performed in-line, on-line, at-line or off-line, but it is important that the data are available in time to detect deflections and to perform control actions.

Table II Monitoring Solutions within Biotechnology

Monitoring goal	Model scenario	Measurement scenario	Process system	Algorithm	Highlights	Ref.
Biomass growth	mass-balance with fixed stoichiometry	carbon in and outflow	*P. chrysogenum*	SQP (sequential quadratic programming)		(64)
Biomass growth	mass-balance with variable stoichiometry	carbon and electron in and outflow	*P. pastoris* and *E. coli*	–	use of system redundancy	(65)
Oxygen consumption	mass balance	offgas	*CHO*	–	simple and robust	(66)
CO2 production	mass balance	offgas	*CHO*	–	carbonate buffered media	(67,68)
Biomass concentration	kinetic model	sugar measurements	*Daucus carota*	extended kalman filter	field of plant cells	(69)
Substrates & biomass	kinetic model	CO2, sugars, product	*S. cerevisiae*	extended kalman filter	NIR based online measurement	(70)
Product & biomass	kinetic flux model	offgas analysis, product	*P. chrysogenum*	particle filter	Raman based online measurements	(58)
Biomass growth	kinetic model	offline & online offgas	*S. clavuligerus*	extended kalman filter	account for measurement delay	(71)
Biomass growth	kinetic model with energy balance	calorimetry	*E. coli*	–	robust growth determination	(72)

After data collection, the measured raw information needs to be converted into the desired monitoring outputs. This conversion is to be performed in real-time and includes data preprocessing e.g. outlier detection, data conversion and state and parameter estimation. For these purposes, mathematical models and model-based methods can be used. Hereby measurements provide real-time data of the ongoing process, whereas the deployed model contains prior knowledge, technical and biological relationships and boundaries of the system (16). This combination of measurements and mathematical models is referred to as soft-sensor (software sensor).

In Fig. 3 the working principle of a software sensor is shown: The process states are described by x and the monitoring outputs by y. In addition to measurements, the designed inputs (u) are included as time dependent variables. The software-implemented models and estimation algorithms can hereby be of any format and structure. As a result the soft-sensor provides an estimate ($\hat{x}$) of the current state.

Critical to the implementation of models in monitoring is the prediction and estimation ability of the model. Apart from the determination of reliable and significant model parameters (see chapter 2.1) the observability is important. An observability analysis can assess the structure of models in order to test whether the information contained in a set of measurements is sufficient for estimating model states (59). A simple approach is the numerical determination of initial values with a subset of known state trajectories, which fails in the unobservable and succeeds in the observable case. This can also be used to define the needed measurement accuracy and frequency in order to fulfil the monitoring goal. To guarantee observability the methodology can also be used to define suitable measurement combinations for specific model implementations, which has exemplarily been shown by our group for *P. pastoris* and *P. chrysogenum* processes (58).

Once the measurement scenario is defined, it needs to be interlinked with the model. Therefore, several algorithms are available, which can be summarized as observers or filters (60). The goal of the observer is to reconstruct current states of interest by real-time collected information and the given process model. Although the appropriate observer type is strongly dependent on the monitoring goal and the process model, the underlying principle is always similar. An additional model and state error $\boldsymbol{\epsilon}(t)$ is added to the model representation of the previous chapter 2.1 eq. [3]). In a second relation, the so-called monitoring scenario, the monitoring outputs y with error $v(t)$ are represented as a function of x (eq. [4]). Under the condition of observability, which means that the provided information in y is enough to reconstruct x, the current states can be estimated. Additionally, the measurement errors as well as process noise are considered as weightings (61,62).

$$x(t) = f(t, x, \varphi, \theta) + \boldsymbol{\epsilon}(t) \tag{3}$$

$$y(t) = h(t, x, \theta) + v(t) \tag{4}$$

Using this approach, multiple measurements can be combined or unmeasured states can be reconstructed. Additionally, this methodology can be used to provide a real-time estimate based on infrequent or very noisy measurements, which can exemplarily be seen in (63). Hereby Goffaux and Wouwer (2005) implemented different observer algorithms in a cell culture process and changed measurement noise and model uncertainty. In order to cope with nonlinearities and the complexity of biological systems suitable filtering algorithms need to be implemented, such as extended and unscented Kalman and particle filters (62). Kalman filters are especially suitable when the model is well-suited and only measurement and process noise occur. Particle filters allow a certain degree of model uncertainty and non-Gaussian noise distributions. In Table II examples of different monitoring implementations in biotechnology can be found.

Simple examples of successful model based monitoring are based on mass balancing (64,65,73). Thus elemental in- and out- fluxes of the reactor are measured. Considering the law of the conservation of mass, conversion rates can be determined. By applying multiple material balances, system redundancy can hereby increase the robustness of the methodology.

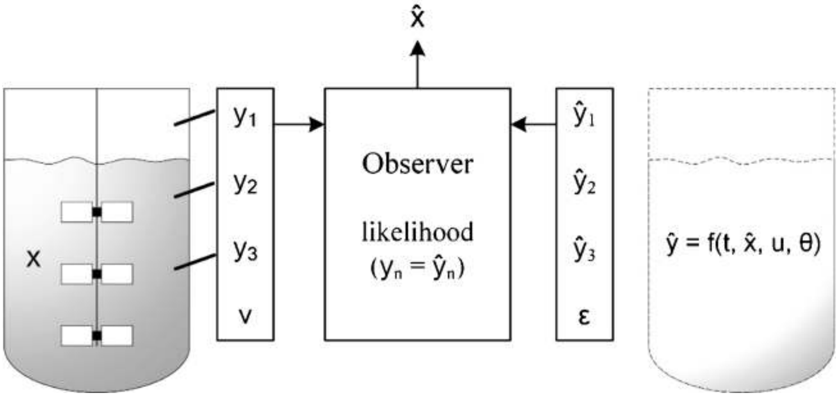

Fig. 3 Principle of model based monitoring with multiple measurements. Through the reconciliation of measured model outputs with current model simulations actual process states can be estimated by considering measurement and process uncertainty.

Kinetic models, which are more detailed and enable the description of cell internal behavior, are also well suited as soft-sensors. The limiting factor is often the system observability of complex kinetic models. Therefore, these models have to be simplified according to the monitoring goal. Aehle *et al.*, for example, showed that offgas-measurement in combination with a simple model can be used to increase the reproducibility and robustness of a mammalian cell culture process (66,74).

Recent implementations by Krämer *et al.* and Golabgir *et al.* have extended the monitoring scenario by spectroscopic NIR and RAMAN measurements in order to obtain system observability of more complex models (58,70). For this purpose, the spectral data were transformed by partial least square regression (PLS) into product and substrate concentrations, which were then used as observer input. Other approaches deal with the incorporation of delayed offline measurements for real time monitoring (75–77). The additional information can help to bring the observer on the right track until the next measurement is available.

In order to provide reliable and robust monitoring as a basis for control, the inclusion of all available process information and knowledge is needed. With this regard the presented model based methods enable i) the determination of needed measurements to guarantee system observability ii) the inclusion of process knowledge in form of a model iii) possible system redundancy with multiple measurements iv) the evaluation of process and measurement noise, which finally leads to v) most probable estimates of the current state of interest.

Process Optimization

Industrial processes aim to find process inputs (also denoted as design vector) to achieve the process goal (e.g. produce a certain product with defined specifications) and simultaneously an optimal process performance with respect to criteria like maximal profit. Additionally, those inputs have to respect physiological and technical constraints as well as product and system rationales. Optimal means getting to the best achievable results with respect to specified (might counteracting) objectives and conditions. If a reliable process model exists, it can be used to determine the optimal process inputs. In addition, the process should ideally be controlled to achieve an optimal process performance. Table III summarizes a selection of examples for model-based optimization and control from literature. In the following, typical optimization goals, variables and optimization spaces according to literature are described. Afterwards, an overview on methods and software of how to perform optimizations is presented. Finally, following a description of aspects of model based optimal control, typical challenges are presented.

Mathematically, optimization problems are typically interpreted as minimization problems of an objective function. In general, three types of optimization objectives typically arising in different stages of the process lifecycle can be distinguished. These are optimizing (i) information content, (ii) productivity and (iii) robustness and reproducibility: (i) Especially but not only during process development optimization algorithms are used to find the parameters of a process model by minimizing the model deviation from the given data (see chapter 2.1.1) or to maximize the information content of planned experiments (see chapter 2.1.2) to obtain adequate process models. (ii) When having a reliable process model, the optimization of the productivity of the process is typically aimed at, e.g. to achieve highest amounts of biomass or product at the end of the process (78–80). (iii) Finally, robustness and reproducibility of an optimized process are typical goals. In this case the objective is usually a minimal deviation from identified (optimal) set points during the whole process. Examples are dissolved oxygen or pH, but also variables like metabolite concentration (81), growth rate or a process variable related to it like the oxygen consumption rate (74). In these cases a dynamic model is needed (see chapter 2.1.1).

A fact to be considered during model development is that only inputs that are included in the process model can be optimized (see chapter 2.1.1). For bioprocesses those are usually feed-rates or initial values. The optimization space is frequently constrained, as shown in Fig. 4: on the one hand, physiological and technical constraints like maximal volume, feed rates or culture time (78,79) have to be considered - on the other hand, the optimization space has to be restricted to an area where the model can be trusted, a region the exact location of which is typically hard to define (see chapter 2.1.1). Because product quality is the priority aim of pharmaceutical production processes, the design space is limited by certain product rationales (e.g. pH and temperature area) too. In addition to that, reducing the size of the optimization space also can speed up the computation time which is needed for time-sensitive optimization tasks. The optimization space is strongly dependent on the process lifecycle. New models, monitoring methods, control strategies, regulatory requirements and changed costs can lead to an expansion of the optimization space and therefore to new optimal design vectors.

There are various methods to solve optimization problems. In some cases the optimization problem can be solved analytically, which means a solution function (for example in integral form) can be obtained. However, frequently (nonlinear) numerical algorithms have to be applied. Various optimization algorithms exist, detailed descriptions can be found in textbooks like (85) or the review of (86). For bioprocesses, frequent implementations of the Nelder-Mead simplex algorithm (fminsearch and its derivates) (87) or differential evolution (88) in MATLAB are used (74,78,83). Another powerful method for large-scale nonlinear optimization is the software package IPOPT (89) e.g. used by (79) for optimizing

Table III Summary

Optimization goal	Optimization space / Constraints	Optimization variable	Optimized process / System	Algorithm	Remarks	References
Information content						
Biomass concentration, conversion of PFAP	–	media components	Synechococcus	ANNSGA (artificial neural network supported genetic algorithm)	ANN	(92)
Productivity - Offline						
Maximal biomass productivity in minimum culture time	Constraints for feed, volume, culture time	Constant/ stair case / exponential feed rate parameter	Hybridoma cell fed-batch	fminsearchcon	Offline optimization	(78)
Maximize amount of cells	Constraints for feeds and volume	Feed	Bakers yeast	Heuristic, analytical and numerical (adaptation of Jacobsons's algorithm (93))		(94)
Productivity - Online						
Maximize productivity and yield in case of uncertainties	Volume, feed rate, operation time, amount of added substrate	Optimal feeding profile	Lysine production fed-batch	ACADO toolkit	Investigation of robust multi-objective optimal control	(91)
Process profitability (costs of product and inducer)	–	Glucose and inducer concentration	E. Coli	Pontryagin's maximum principle	Optimal control	(95)
Maximize biohydrogen production	Constraints for feed, terminal region, culture time	Nutrient flow	Cyanobacteria fed-batch	IPOPT (after converting optimal control problem to nonlinear optimization problem with orthogonal collocation)	Simulation MPC with parameter estimation	(79)
Maximize Productivity	Max volume	Feed	Steptomyces tendae		MPC	(80)
Robustness						
Control glucose to a setpoint	–	Glucose feed rate profile	CHO fed-batch	SQP (sequential quadratic programming)	MPC	(81)
Control consumed oxygen to a setpoint	–	Glutamine feed rate	CHO fed-batch	Simplex	MPC	(74)

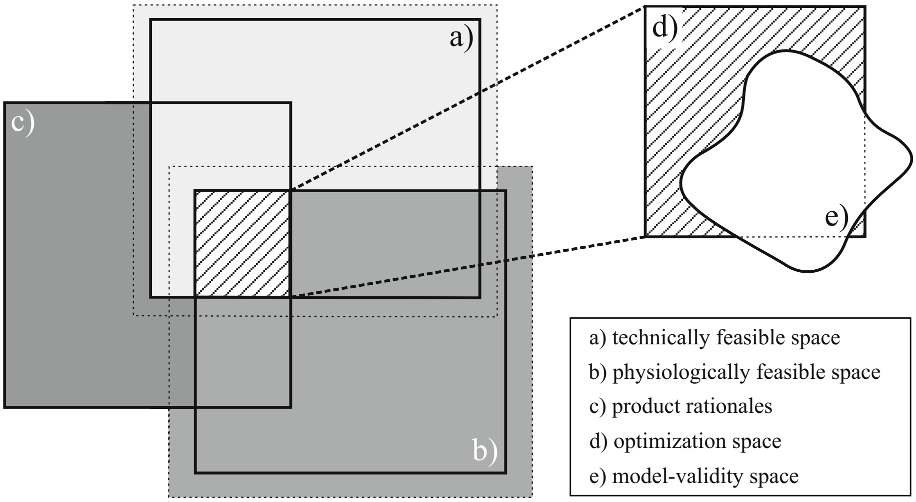

Fig. 4 Optimization space limited by technically and physiologically feasible space as well as by product and system rationales. The potential innovation space is the space where it can be increased e.g. by more knowledge about the system.

biohydrogen production. More applied algorithms are listed in Table III. When choosing the optimization algorithm, one has to ponder aspects like the number of variables to be optimized, the complexity of the model, the implementation environment or the acceptable duration of the optimization. The last point is of major importance when performing optimizations during the process. In case the optimal design vector is time-dependent it might has to be parametrized. This is frequently done by discretizing the input signal via partially constant, linear or parabolic functions (also termed as zero, first or second order hold). Simulations are a valuable tool to investigate configuration details e.g. how to parameterize the design vector. This can also help to ensure a fast computation (80).

When the optimal values of the process inputs are found, various possibilities for controlling the process to achieve the desired optimal performance exist: a simple method is to determine the optimal design vector once and control the process on those predefined set points. This approach is state of the art in most production processes.

However, this control method possibly fails when process deviations occur due to model uncertainties or unknown or neglected process disturbances which are not considered previously. The reason is that this strategy does not consider the real values of the process outputs (the controlled variables) during manipulating the inputs (the manipulated variables). This can lead to unwanted process behavior: (80) computed optimal profiles for three feeds (ammonium, phosphate, glucose) based on a mechanistic model. They studied the effect of model uncertainties by varying the model parameters and applying those feed profiles determined with the initial parameters. The results revealed a high dependency of end product (the optimization goal) on the model parameters: in 60% of the simulations less product than in the original case was achieved. They concluded that this can be avoided by applying closed-loop control. In this case the manipulated variables are adjusted based on the values of the controlled variables. Besides classic closed-loop controllers like PID controllers, a well-known and powerful representative method is model predictive control (MPC) (80–85): a dynamic model is used to find the optimal inputs with respect to a defined objective function as described above. However, instead of performing the computation only once in the beginning, the optimization is repeated after a defined control horizon to react towards process deviations. Therefore, the optimization problem has to be solved in real-time, which demands robust and fast optimization algorithms. In order to be able to discover process deviations information about the current process state is needed. Depending on the measurement environment monitoring strategies as described in chapter 2.2 have to be applied.

Dynamic optimization of bioprocesses is linked with several challenges. E.g., in case of multiple objectives it is difficult to choose an optimal solution: typically, there can be counteracting objectives in such that one objective can only be improved by worsening the other, which implies a trade-off is needed. This set of solutions is known as Pareto front. More theory on this topic can e.g. be found in the textbook by (90). Another aspect is robustness towards process deviations and model uncertainties. One way to deal with this is presented by (91), who investigated robust multi-objective optimal control in case of model uncertainties by interpreting robustness as additional objective. Another typically occurring phenomenon is, that the optimal design vector lies at the boundaries of the optimization space. One the one hand, this can be critical if the optimization space is not defined properly, for example due to limited knowledge about the validity space of the model. On the other hand this implies that the process might be optimized by increasing the optimization space e.g. by deriving more knowledge to increase the model validity space and improve the model or by technical innovations.

Summing up, optimization tasks occur during different stages of the process lifecycle, with the highly diverse goals of maximal information content, productivity and robustness and reproducibility, respectively. Methods for optimization and control are limited by the quality and the inputs of the model. In addition to that, closed loop optimal control is also limited by issues like process noise or uncertainties of the model and the system. Therefore a suitable monitoring strategy has to be established and suitable observers have to be applied. In addition to that, if optimization has to be performed online and probably unsupervised, fast and trustworthy algorithms are demanded. However, in those cases, where this is fulfilled, MPC is a valuable tool to achieve optimal processes.

Data Mining for Detection of Disturbance Variables

Although sophisticated control strategies are applied to modern processes (achieved using the above described methods with respect to determination, monitoring and optimization of CPPs), fluctuations in process performance inevitably occur. For that reason, continuous process improvement is necessary, which can be achieved by data mining techniques in order to detect disturbance variables.

Generally, every bioprocess includes known but neglected or tolerated disturbances, such as the control ranges of process parameters like pH, dissolved oxygen, feeding profiles etc. On the other side, there are unknown disturbances that might undermine process robustness and that should be identified in later stages of the process development or during process improvement. In the following, we want to focus on the upstream of biopharmaceutical processes as this is the major source of disturbances. According to the exemplification of the bioreactor as a disperse multiphase-system, these unknown disturbances can be grouped in the following classes as follows:

1) Biomass as disturbance variable, either due the genotype (e.g. repression or induction of certain genes) or phenotype (e.g. morphological changes)
2) The composition of or single substances in the fluid phases as disturbance variable (e.g. raw material variability, metabolites, process additives)
3) Physical and local characteristics such as inhomogeneities as disturbance variable (e.g. improper dispersal of base/acid or feeds, inhomogeneities in dissolved oxygen etc.)

The detection of disturbance variables aims at enhancing the understanding of process fluctuations, thereby increasing process robustness or process performance and can finally even lead to improvement of control strategies (see chapters 1 and 2.3). The ability of process intervention according to knowledge gained via an analysis of disturbance variables is strongly coupled to the optimization and especially to the innovation space. This means that a possible intervention is limited by the biological system itself, e.g. physiological parameters like maximal specific uptake rates, but also by external factors such as technical feasibility or logistical and organizational factors, e.g. time line for upstream to downstream processing, shift work etc. One the one hand, the development or implementation of new analytical methods or probes for the characterization of the system and its disturbances can lead to an extension of the innovation space of the investigated bioprocess and thereby enhance process control strategies. On the other hand, technical or organizational constraints can restrict process intervention – within the borders of the innovation space - although a disturbance was successfully detected (Fig. 4).

Generally, the detection of important disturbance variables follows a data-driven knowledge discovery approach, mainly focusing on data mining methods (Fig. 5), i.e. statistical methods to extract information from large data sets. Risk assessment tools are commonly used for process development (2) and can also facilitate the identification of possible disturbance classes (see definition above) within the design space of the process. A prominent example of these tools is the Ishikawa (or fishbone) diagram, which illustrates that this form of

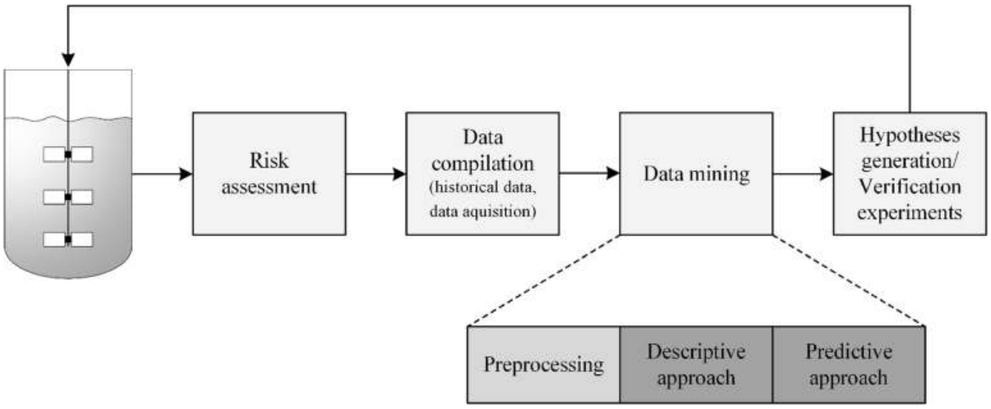
Risk assessment
Data compilation (historical data, data aquisition)
Data mining
Hypotheses generation/ Verification experiments
Preprocessing
Descriptive approach
Predictive approach

Fig. 5 Workflow showing the data-driven knowledge discovery approach for the detection and minimization of disturbance variables. After selection of the targeted disturbance class via risk assessment tools, data has to be generated and/or accumulated. Indications about disturbing variables/ descriptors can be generated by correlation analysis or – if possible – via mechanistic modelling. Obtained knowledge/ information has to be implemented in the design space to allow minimization of the identified disturbances.

process improvement is done at later stages of the process development as some prior knowledge about the process is necessary (i.e. QbD approach). According to the outcome of the risk assessment, data has to be generated or compiled. Every modern biotechnology production plant is equipped with systems that record and archive continuous and intermittent data of every process. These historical data can be used for data mining and the identification of disturbance variables - even including known but neglected disturbances. Examples of the assessment of historical data are given in (96–98). (99), for instance, used a three-step approach previously introduced by (100,101) in order to optimize an *E.coli* process for green fluorescent protein production.

Often, historical data do not represent the probable disturbance class well enough, which is why additional data are needed. These data are commonly generated via analytical measurements of specific components (e.g. HPLC, IC or ICP analysis), for instance of the raw material for media production. Examples of this approach are given by (102) and by our group (103), who focused on the detailed characterization of complex raw material. As this approach is very laborious and analytically challenging fingerprinting methods such as near infrared (NIR), mid infrared (MIR) or (2D)-fluorescence spectroscopy can be applied to complex matrices. These methods generate an overall but still specific description of the composition of a complex material or media (e.g. a spectra) without the identification of certain substances, i.e. a fingerprint of the material. Spectroscopic fingerprinting methods were applied by (104–108) in order to determine the variability and disturbances of applied raw material.

Before data mining techniques can be applied, it should be noted that the characteristics of bioprocess data is its heterogeneity with respect to time scale. As already mentioned in chapter 2.2, bioprocess data can be continuous measurements, intermittent measurements or even one-time measurements at the beginning or the end of the process, such as raw material attributes or process titer, respectively. Hence, before data analysis can be started, preprocessing techniques, feature selection or even dimensionality reduction has to be performed. Examples of these techniques applied for historical datasets are described in (96), such as filter and wrapper methods or principle component analysis (PCA) for dimensionality reduction. If additional analytical data are generated at one point of time, e.g. measurements of specific components or fingerprinting data of the used raw material, other preprocessing methods have to be applied. For fingerprinting spectra, first, second or third order derivatives are commonly used in order to reduce noise from the spectral data. Additionally, data can be mean-centered or normalized, depending on the statistical method that is used for further analysis (109–113). In the following step the actual data mining starts, which can be categorized in descriptive or predictive approaches (96) (Table IV).

In the descriptive approach methods for discriminant analysis are applied in order to identify patterns or clusters in the dataset. Common methods are PCA, e.g. applied by (102,103) and cluster analysis (98). For the predictive approach methods are applied that allow correlation analysis, i.e. the preprocessed data or selected features are correlated with process outcomes (i.e. CQAs and productivity) in order to identify possible relationships. Typical methods are multiple linear regression (MLR), partial least squares (PLS) regression and artificial neural networks (ANN). There are also modifications of these methods available that overcome certain drawbacks of the original method as well as relatively new methods such as support vector machines (SVM).

Jose *et al.* analyzed two raw materials via two fingerprinting techniques (105). In order to combine the spectra of these two materials PCA models for both raw materials were generated and the scores of these models were used for the generation of an interval partial least squares (iPLS) regression model which allowed a correlation between raw material quality and product yield and titer. iPLS is a graphical extension of regular PLS models. It divides spectral data into equidistant subintervals of which validated calibration models are developed. Hence, this method allows to depict relevant information in different spectral subdivisions and is able to remove interferences from other regions (114). Another method proposed by Gao *et al.* for the identification of raw material and process performance is the orthogonal partial least squares – discriminant analysis (OPLS-DA) (104). This method equals partial least squares – discriminant analysis (PLS-DA) which is a combination of canonical correlation analysis and linear discriminant analysis Thus, providing descriptive as well as predictive information (115,116). The integration of an orthogonal signal correction (OSC)-filter, which should allow the separation between predictive and non-predictive variation, should improve the interpretation of the model (117,118). Nevertheless, the superiority of OPLS-DA over PLS-DA is critically discussed among experts. Balabin *et al.* introduced an extension of ANN, namely support vector machines (SVM), for spectroscopic calibration and as data mining technique (119). It has the advantage of providing global models that are often unique, which is a benefit compared to normal ANN.

Descriptive as well as predictive methods result in the generation of hypotheses about disturbances, crucial parameters or interactions. These hypotheses have to be evaluated or experimentally verified by experts (e.g. via experimental design as mentioned in chapter 2.1.2) before they can be implemented in the control strategy. At this stage the control loop (Fig. 1) can be restarted by the integration of gained knowledge in the model or even by the introduction of new CPPs or kPPs. This approach can additionally result in the improvement of product quality and productivity.

In general there are three major challenges in process improvement via detection of disturbance variables: The

Table IV Various Methods are Available for the Data to Information Approach, which is applied for the Identification and Minimization of Disturbance Variables. The Most Common Ones are Stated here Including Information about Linearity, Advantages and Disadvantages as well as References to Literature

Approach	Method	Advantages	Disadvantages	Output	Method literature	Application literature
Descriptive	PCA	• Orthogonal • Dimensionality reduction • Easily applicable • Provides overview of input matrix • Classification of data	• Difficult to interpret if more PCs are significant • no correlations with process response possible • linear	• Loadings ➔ describes the correlation between variables in an orthogonal manner • Scores ➔ shows grouping/ clustering/ patterns/ trends ➔ facilitates interpretation due to additional dimensionality reduction	(120)	(102,103,107,108,121–123)
Descriptive	Cluster analysis (CA)	• Classification of data • Multiple algorithms are available ➔ adaption to problem statement possible	• No dimensionality reduction ➔ complicates the identification of trends • Linear • No Correlation with process response	• Dendrogramm ➔ clusters can be seen and especially the distance between clusters can be analyzed		(98,121,122)
Descriptive and predictive	PLS-DA	• Dimensionality reduction • Prediction of group membership • Classification of data • Easily applicable	• Linear • Y-variable (i.e. class) has to be declared before analysis • Knowledge about method necessary (choice of threshold, PLS1 or PLS2) • Overfitting	• Scores ➔ shows grouping/ clustering/ patterns/ trends ➔ facilitates interpretation due to additional dimensionality reduction • Weights/ loadings ➔ relates classifier to underlying variable	(115,116)	
Descriptive and predictive	OPLS-DA	• Orthogonal • see PLS-DA	• see PLS-DA	• see PLS-DA		(104,117)
Predictive	MLR	• Easily applicable • Correlation with process response	• not applicable for fingerprinting analysis (due to collinearities) • linear	• ANOVA validation • Coefficients with confidence intervals ➔ representing variables that correlate with process response		(124)
Predictive	PLS	• Dimensionality reduction • Correlation with process response • Variable ranking available • Easily applicable	• Not orthogonal • Correlations are assumed to be linear (only “quasi-nonlinear” algorithmic adaptations • available like Poly-PLS or Spline-PLS) • Small validity space • linear	• Observed *vs* predicted • Coefficients with confidence intervals ➔ representing variables that correlate with process response	(119,125,126)	(108,121,123,124)
Predictive	PCR	• Dimensionality reduction • Easily applicable • Orthogonal • Correlation with process response	• Difficult to interpret if more PCs are significant • Correlations are assumed to be linear	• see PCA and MLR	(127)	(128,129)
Predictive	ANN	• Correlation with process response • Adaptive learning • Self-organization • Fault tolerance via redundant coding • Real-time operating ability • Easy insertion into existing technologies • non linear	• Mathematically demanding • difficult to implement for process development • iterative workflow • dependence of final result on initial parameters • tendency to overfitting • high training time and computational resources • non-uniqueness of final result	• Observed *vs* predicted (cross validation)	(119,130)	(124)
Predictive	SVM	• see ANN • handling high dimensional input vectors	• see ANN	• see ANN	(119)	

identification of an adequate analytical method for in-depth investigation of disturbance variables, such as cell morphology, raw material or scale-up effects (e.g. inhomogeneties, biomass segregation), is demanding, especially with increasing complexity of the process. The knowledge about method errors and general deviations during the process is necessary in order to allow adequate conclusions from data mining. Additionally, the choice of the appropriate statistical method that is applied to the data compilation is crucial to achieving meaningful patterns, clusters and correlations and has also an impact on the interpretability of the results.

Summing up, for continuous process improvement, the evaluation of both historical data as well as the generation of new data with respect to probable disturbance variables is necessary. Data mining of these huge datasets allows the generation of hypotheses which can be verified by experiments. Gained knowledge can further on be implemented in existing models in order to improve process robustness and performance (Fig. 2).

CONCLUSIONS

During the biopharmaceutical process lifecycle, countless challenges arise: uncontrollable external conditions, fluctuations in raw material, inaccuracies in process control and continuous innovations - and they all affect the process performance over time. The trend of the last few years has clearly pointed towards a model approach in order to ensure knowledge transfer during the entire process lifecycle and, additionally, during different processes. Model-based methods allow the applicability of the stored knowledge. In the presented review the applicability of model-based methods in order to ensure control has been shown. To reach the goal of control four challenges were investigated: I) generation of process knowledge, II) process monitoring, III) process optimization and IV) continuous improvement of the process (Fig. 1).

The first challenge includes the identification of CPPs and kPPs, hence, the generation of process knowledge. If relations and interactions within the process are understood, the main challenge is the setup and the verification of process models in order to predict a target value (CPP, kPP or CQA). This is a critical step because the model quality has an impact on accuracy, precision, applicability and the validity area of all model-based methods. Main issues in the field of modelling are a lack of experts and tools for the model setup in biopharmaceutical production processes. In addition, process-models should be extended or adapted during the whole process lifecycle. Therefore, modelling is a typical bottleneck for the application of model-based methods in industrial processes. In order to overcome this problem, we presented modelling workflows for the setup of models. Additionally, methods for the generation of information during experiments by model-based experimental design are presented.

The second challenge is an adequate process monitoring. The combination of real-time measurements and model-based methods like observers allow an optimal usage of monitoring capacities. Model-based methods are already widespread and accepted in the area of process monitoring since they allow the estimation of hard or not measureable parameters and variables, which are necessary for subsequent control tasks. The bottleneck of monitoring methods is mainly the transferability between different processes and scales concerning measurement methods and software environment. During the process lifecycle new real-time measurement sensors, changing process models and new control tasks should be considered in the process monitoring concept.

The third challenge is process optimization and process control. First of all, a proper definition of the optimization objective is needed. Especially in case of multiple objectives an adequate weighting of the different goals is not easy but important. The second task is to find an optimal design vector for the process. Model-based methods are valuable tools to declare the optimum. Nevertheless, multidimensional optimization tasks are generally hard to implement as well as computationally demanding. Furthermore, successful optimization highly depends on the model quality as well as knowledge about the validity space of the model.

The fourth challenge is the continuous improvement of the process based on additional research and historical data assessment. Therefore, datamining tools are widespread and accepted as model-based methods in order to generate hypotheses, which can be experimentally evaluated and furthermore gained knowledge can be included in the process model. Bottleneck of these datamining tools are mainly the availability of adequate measurement methods for the generation of additional data and the interpretability of descriptive as well as predictive model-based methods.

Irrespective of the availability of model-based methods, a certain acceptance of these methods in the biotechnological community has to be generated. Hence, the benefits of the application of model-based methods on process development and production have to be demonstrated. Additionally, the training of the users is of great importance as well as the presentation of all methods in more user-friendly tools. In combination with continuous support and further development of the process model, model-based methods are powerful tools to ensure the overall goal of biopharmaceutical processes, i.e. the guarantee of high product quality.

ACKNOWLEDGMENTS AND DISCLOSURES

The authors would like to thank Jens Fricke for his support through instructive discussions and with the graphs and Katharina Oberhuber for the English proof reading. Financial support was provided by the Austrian research funding association (FFG) under the scope of the COMET program within the research project "Industrial Methods for Process Analytical Chemistry - From Measurement Technologies to Information Systems (imPACts)" (contract # 843546) and the Christian Doppler Forschungsgesellschaft (grant number 171).

REFERENCES

1. Guideline IHT. Development and manufacture of drug substances (chemical entities and biotechnological/biological entities) Q11. London: European medicines agency; 2011.
2. Guideline IHT. Pharmaceutical development Q8 (R2). 2009.
3. Guideline IHT. Specifications: test procedures and acceptance criteria for biotechnological/biological products Q6B. 1999.
4. FDA. Guidance for Industry PAT-A framework for innovative pharmaceutical development, Manufacturing, and Quality Assurance. wwwfdagov. 2004.
5. Rathore AS, Winkle H. Quality by design for biopharmaceuticals. Nat Biotech. 2009;27(1):26–34.
6. Guideline IHT. Q12: technical and regulatory considerations for pharmaceutical product lifecycle management endorsed by the ich steering committee on 9 September 2014. 2014;1.
7. Ragab MAF, Arisha A. Knowledge management and measurement: a critical review. J Knowl Manag. 2013;17(6):873–901.
8. Studer R, Benjamins VR, Fensel D. Knowledge engineering: principles and methods. Data Knowl Eng. 1998;25(1):161–97.
9. Herwig C, Garcia-Aponte OF, Golabgir A, Rathore AS. Knowledge management in the QbD paradigm: manufacturing of biotech therapeutics. Trend. Biotechnol. 2015;33(7):381–7.
10. Herold S, Heine T, King R. An automated approach to build process models by detecting biological phenomena in (fed-)batch experiments. IFAC P Vol. 2010;43(6):138–43.
11. Jakeman AJ, Letcher RA, Norton JP. Ten iterative steps in development and evaluation of environmental models. Environ Model Softw. 2006;21(5):602–14.
12. Refsgaard JC, Henriksen HJ. Modelling guidelines—terminology and guiding principles. Adv Wat Resour. 2004;27(1):71–82.
13. Waveren H, Groot S, Scholten H, Geer FCV, Wösten JHM, Koeze RD, *et al.* Good modelling practice Handbook. 2000.
14. Weinstein MC, O'Brien B, Hornberger J, Jackson J, Johannesson M, McCabe C, et al. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR task force on good research practices—modeling studies. Value Health. 2003;6(1):9–17.
15. Donoso-Bravo A, Mailier J, Martin C, Rodríguez J, Aceves-Lara CA, Wouwer AV. Model selection, identification and validation in anaerobic digestion: a review. Water Res. 2011;45(17):5347–64.
16. Mandenius C-F, Gustavsson R. Mini-review: soft sensors as means for PAT in the manufacture of bio-therapeutics. J Chem Technol Biotech. 2015;90(2):215–27.
17. Almquist J, Cvijovic M, Hatzimanikatis V, Nielsen J, Jirstrand M. Kinetic models in industrial biotechnology – improving cell factory performance. Metab Eng. 2014;24:38–60.
18. Neymann T, Helbing L, Engell S. Computer-implemented method for creating a fermentation model. United States Patents. 2016.
19. Hebing L, Neymann T, Thüte T, Jockwer A, Engell S. Efficient generation of models of fed-batch fermentations for process design and control. IFAC-PapersOnline. 2016;49(7):621–6.
20. Leifheit J, King R. Systematic structure and parameter identification for biological reaction systems supported by a software-tool. IFAC P Vol. 2005;38(1):1095–100.
21. Herold S, King R. Automatic identification of structured process models based on biological phenomena detected in (fed-)batch experiments. Bioprocess Biosyst Eng. 2014;37(7):1289–304.
22. Kroll P, Hofer A, Stelzer IV, Herwig C. Workflow to set up substantial target-oriented mechanistic process models in bioprocess engineering. Process Biochem. 2017;
23. Chis O-T, Banga JR, Balsa-Canto E. Structural identifiability of systems biology models: a critical comparison of methods. PLoS One. 2011;6(11):e27755.
24. Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, et al. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformat. 2009;25(15):1923–9.
25. Meeker WQ, Escobar LA. Teaching about Approximate Confidence Regions Based on Maximum Likelihood Estimation. American Statist. 1995;49(1):48–53.
26. Wechselberger P, Seifert A, Herwig CPAT. method to gather bioprocess parameters in real-time using simple input variables and first principle relationships. Chem Eng Sci. 2010;65(21): 5734–46.
27. Lemaire C, Schoefs O, Lamy E, Pauss A, Mottelet S. Modeling of an aerobic bioprocess based on gas exchange and dynamics: a novel approach. Bioprocess Biosyst Eng. 2014;37(9):1809–16.
28. King JMP, Titchener-Hooker NJ, Zhou Y. Ranking bioprocess variables using global sensitivity analysis: a case study in centrifugation. Bioprocess Biosyst Eng. 2007;30(2):123–34.
29. Mandenius C-F, Brundin A. Bioprocess optimization using design-of-experiments methodology. Biotechnol Prog. 2008;24(6):1191–203.
30. Galvanin F, Barolo M, Bezzo F. A framework for model-based design of experiments in the presence of continuous measurement systems. IFAC P Vol. 2010;43(5):571–6.
31. Zullo LC. Computer aided design of experiments: an engineering approach: Imperial College London (University of London). 1991.
32. Franceschini G, Macchietto S. Model-based design of experiments for parameter precision: State of the art. Chem Eng Sci. 2008;63(19):4846–72.
33. Telen D, Logist F, Van Derlinden E, Tack I, Van Impe J. Optimal experiment design for dynamic bioprocesses: a multi-objective approach. Chem Eng Sci. 2012;78:82–97.
34. Wechselberger P, Sagmeister P, Herwig C. Model-based analysis on the extractability of information from data in dynamic fed-batch experiments. Biotechnol Prog. 2013;29(1):285–96.
35. Schwaab M, Luiz Monteiro J, Carlos Pinto J. Sequential experimental design for model discrimination: Taking into account the posterior covariance matrix of differences between model predictions. Chem Eng Sci. 2008;63(9):2408–19.

36. Schwaab M, Silva FM, Queipo CA, Barreto AG Jr, Nele M, Pinto JCA. new approach for sequential experimental design for model discrimination. Chem Eng Sci. 2006;61(17):5791–806.
37. Schaber SD, Born SC, Jensen KF, Barton PI. Design, execution, and analysis of time-varying experiments for model discrimination and parameter estimation in microreactors. Org Process Res Dev. 2014;18(11):1461–7.
38. Hoang MD, Barz T, Merchan VA, Biegler LT, Arellano-Garcia H. Simultaneous solution approach to model-based experimental design. AICHE J. 2013;59(11):4169–83.
39. Barz T, López Cárdenas DC, Arellano-Garcia H, Wozny G. Experimental evaluation of an approach to online redesign of experiments for parameter determination. AICHE J. 2013;59(6): 1981–95.
40. Galvanin F, Boschiero A, Barolo M, Bezzo F. Model-based design of experiments in the presence of continuous measurement systems. Ind Eng Chem Res. 2011;50(4):2167–75.
41. Cruz Bournazou MN, Barz T, Nickel DB, Lopez Cárdenas DC, Glauche F, Knepper A, *et al.* Online optimal experimental redesign in robotic parallel fed-batch cultivation facilities. Biotechnol Bioeng. 2016:n/a-n/a.
42. Neddermeyer F, Marhold V, Menzel C, Krämer D, King R. Modelling the production of soluble hydrogenase in Ralstonia eutropha by on-line optimal experimental design. IFAC-PapersOnline. 2016;49(7):627–32.
43. Brik Ternbach M, Bollman C, Wandrey C, Takors R. Application of model discriminating experimental design for modeling and development of a fermentative fed-batch L-valine production process. Biotechnol Bioeng. 2005;91(3):356–68.
44. Maheshwari V, Rangaiah GP, Samavedham L. Multiobjective framework for model-based design of experiments to improve parameter precision and minimize parameter correlation. Ind Eng Chem Res. 2013;52(24):8289–304.
45. Galvanin F, Cao E, Al-Rifai N, Gavriilidis A, Dua V, editors. Model-based design of experiments for the identification of kinetic models in microreactor platforms. 12th international symposium on process systems engineering and 25th European symposium on computer aided process engineering. Elsevier; 2015.
46. Franceschini G, Macchietto S. Novel anticorrelation criteria for model-based experiment design: Theory and formulations. AICHE J. 2008;54(4):1009–24.
47. Banga JR, Balsa-Canto E. Parameter estimation and optimal experimental design. Essays Biochem. 2008;45:195–210.
48. López CDC, Barz T, Körkel S, Wozny G. Nonlinear ill-posed problem analysis in model-based parameter estimation and experimental design. Comput Chem Eng. 2015;77:24–42.
49. López CDC, Barz T, Peñuela M, Villegas A, Ochoa S, Wozny G. Model-based identifiable parameter determination applied to a simultaneous saccharification and fermentation process model for bio-ethanol production. Biotechnol Prog. 2013;29(4):1064–82.
50. Barz T, Arellano-Garcia H, Wozny G. Handling Uncertainty in Model-Based Optimal Experimental Design. Ind Eng Chem Res. 2010;49(12):5702–13.
51. Glassey J, Gernaey KV, Clemens C, Schulz TW, Oliveira R, Striedner G, et al. Process analytical technology (PAT) for biopharmaceuticals. Biotechnol J. 2011;6(4):369–77.
52. Mandenius C-F. Recent developments in the monitoring, modeling and control of biological production systems. Bioprocess Biosyst Eng. 2004;26(6):347–51.
53. Vojinović V, Cabral JMS, Fonseca LP. Real-time bioprocess monitoring: part I: In situ sensors. Sensors Actuators B Chem. 2006;114(2):1083–91.
54. Schügerl K. Progress in monitoring, modeling and control of bioprocesses during the last 20 years. J Biotechnol. 2001;85(2): 149–73.
55. Abu-Absi NR, Kenty BM, Cuellar ME, Borys MC, Sakhamuri S, Strachan DJ, et al. Real time monitoring of multiple parameters in mammalian cell culture bioreactors using an in-line Raman spectroscopy probe. Biotechnol Bioeng. 2011;108(5):1215–21.
56. Roychoudhury P, Harvey LM, McNeil B. The potential of mid infrared spectroscopy (MIRS) for real time bioprocess monitoring. Anal Chim Acta. 2006;571(2):159–66.
57. Striedner G, Bayer K. An advanced monitoring platform for rational design of recombinant processes. In: Mandenius C-F, Titchener-Hooker NJ, editors. Measurement, monitoring, modelling and control of bioprocesses. Berlin: Springer Berlin Heidelberg; 2013. p. 65–84.
58. Golabgir A, Herwig C. Combining mechanistic modeling and raman spectroscopy for real-time monitoring of fed-batch penicillin production. Chem Ing Tech. 2016;88(6):764–76.
59. Nakhaeinejad M, Bryant MD. Observability analysis for model-based fault detection and sensor selection in induction motors. Meas Sci Technol. 2011;22(7):075202.
60. Mohd Ali J, Ha Hoang N, Hussain MA, Dochain D. Review and classification of recent observers applied in chemical process systems. Comput Chem Eng. 2015;76:27–41.
61. Dochain D. State and parameter estimation in chemical and biochemical processes: a tutorial. J Process Contr. 2003;13(8):801–18.
62. Simon D. Optimal state estimation: Kalman, H infinity, and nonlinear approaches. New York: Wiley; 2006.
63. Goffaux G, Vande Wouwer A. Bioprocess state estimation: some classical and less classical approaches. In: Meurer T, Graichen K, Gilles ED, editors. Control and observer design for nonlinear finite and infinite dimensional systems. Berlin: Springer Berlin Heidelberg; 2005. p. 111–28.
64. Mou D-G, Cooney CL. Growth monitoring and control through computer-aided on-line mass balancing in a fed-batch penicillin fermentation. Biotechnol Bioeng. 1983;25(1):225–55.
65. Wechselberger P, Sagmeister P, Herwig C. Real-time estimation of biomass and specific growth rate in physiologically variable recombinant fed-batch processes. Bioprocess Biosyst Eng. 2013;36(9):1205–18.
66. Aehle M, Kuprijanov A, Schaepe S, Simutis R, Lubbert A. Simplified off-gas analyses in animal cell cultures for process monitoring and control purposes. Biotechnol Lett. 2011;33(11):2103–10.
67. Frahm B, Blank H-C, Cornand P, OelÃŸner W, Guth U, Lane P, et al. Determination of dissolved CO2 concentration and CO2 production rate of mammalian cell suspension culture based on off-gas measurement. J Biotechnol. 2002;99(2):133–48.
68. Bonarius HPJ, de Gooijer CD, Tramper J, Schmid G. Determination of the respiration quotient in mammalian cell culture in bicarbonate buffered media. Biotechnol Bioeng. 1995;45(6):524–35.
69. Albiol J, Robusté J, Casas C, Poch M. Biomass estimation in plant cell cultures using an extended Kalman filter. Biotechnol Prog. 1993;9(2):174–8.
70. Krämer D, King R. On-line monitoring of substrates and biomass using near-infrared spectroscopy and model-based state estimation for enzyme production by S. cerevisiae. IFAC-PapersOnLine. 2016;49(7):609–14.
71. Gudi RD, Shah SL, Gray MR. Adaptive multirate state and parameter estimation strategies with application to a bioreactor. AICHE J. 1995;41(11):2451–64.
72. Biener R, Steinkämper A, Hofmann J. Calorimetric control for high cell density cultivation of a recombinant Escherichia coli strain. J Biotechnol. 2010;146(1–2):45–53.
73. Jobé AM, Herwig C, Surzyn M, Walker B, Marison I, von Stockar U. Generally applicable fed-batch culture concept based on the

detection of metabolic state by on-line balancing. Biotechnol Bioeng. 2003;82(6):627–39.

74. Aehle M, Kuprijanov A, Schaepe S, Simutis R, Lubbert A. Increasing batch-to-batch reproducibility of CHO cultures by robust open-loop control. Cytotechnology. 2011;63(1):41–7.

75. Gopalakrishnan A, Kaisare NS, Narasimhan S. Incorporating delayed and infrequent measurements in Extended Kalman Filter based nonlinear state estimation. J Process Contr. 2011;21(1):119–29.

76. Guo Y, Huang B. State estimation incorporating infrequent, delayed and integral measurements. Automatica. 2015;58:32–8.

77. Soons ZITA, Shi J, van der Pol LA, van Straten G, van Boxtel AJB. Biomass growth and kLa estimation using online and offline measurements. IFAC P Vol. 2007;40(4):85–90.

78. Amribt Z, Niu H, Bogaerts P. Macroscopic modelling of overflow metabolism and model based optimization of hybridoma cell fed-batch cultures. Biochem Eng J. 2013;70:196–209.

79. del Rio-Chanona EA, Zhang D, Vassiliadis VS. Model-based real-time optimisation of a fed-batch cyanobacterial hydrogen production process using economic model predictive control strategy. Chem Eng Sci. 2016;142:289–98.

80. Kawohl M, Heine T, King R. Model based estimation and optimal control of fed-batch fermentation processes for the production of antibiotics. Chem Eng Process Process Intensif. 2007;46(11):1223–41.

81. Craven S, Whelan J, Glennon B. Glucose concentration control of a fed-batch mammalian cell bioprocess using a nonlinear model predictive controller. J Process Contr. 2014;24(4):344–57.

82. Mandenius C-F, Titchener-Hooker NJ. Measurement, monitoring, modelling and control of bioprocesses. Berlin: Springer; 2013.

83. Craven S, Shirsat N, Whelan J, Glennon B. Process model comparison and transferability across bioreactor scales and modes of operation for a mammalian cell bioprocess. Biotechnol Prog. 2013;29(1):186–96.

84. Dewasme L, Amribt Z, Santos LO, Hantson AL, Bogaerts P, Wouwer AV. Hybridoma cell culture optimization using nonlinear model predictive control. IFAC P Vol. 2013;46(31):60–5.

85. Nocedal J, Wright SJ. Numerical optimization 2nd. New York: Springer; 2006.

86. Biegler LT. An overview of simultaneous strategies for dynamic optimization. Chem Eng Process Process Intensif. 2007;46(11): 1043–53.

87. Lagarias JC, Reeds JA, Wright MH, Wright PE. Convergence properties of the nelder–mead simplex method in low dimensions. SIAM J Optimiz. 1998;9(1):112–47.

88. Das S, Suganthan PN. Differential evolution: a survey of the state-of-the-art. IEEE T Evolut Comput. 2011;15(1):4–31.

89. Wächter A, Biegler LT. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. Math Program. 2006;106(1):25–57.

90. Miettinen K. Nonlinear multiobjective optimization. New York: Springer Science & Business. Media. 2012;

91. Logist F, Houska B, Diehl M, Van Impe JF. Robust multi-objective optimal control of uncertain (bio)chemical processes. Chem Eng Sci. 2011;66(20):4670–82.

92. Franco-Lara E, Link H, Weuster-Botz D. Evaluation of artificial neural networks for modelling and optimization of medium composition with a genetic algorithm. Process Biochem. 2006;41(10): 2200–6.

93. Jacobson D, Gershwin S, Lele M. Computation of optimal singular controls. IEEE Trans Autom Control. 1970;15(1):67–73.

94. Menawat A, Mutharasan R, Coughanowr DR. Singular optimal control strategy for a fed-batch bioreactor: Numerical approach. AICHE J. 1987;33(5):776–83.

95. Lee J, Ramirez WF. Optimal fed-batch control of induced foreign protein production by recombinant bacteria. AICHE J. 1994;40(5):899–907.

96. Charaniya S, W-S H, Karypis G. Mining bioprocess data: opportunities and challenges. Trend Biotechnol. 2008;26(12):690–9.

97. Kamimura RT, Bicciato S, Shimizu H, Alford J, Stephanopoulos G. Mining of biological data I: identifying discriminating features via mean hypothesis testing. Metab Eng. 2000;2(3):218–27.

98. Kamimura RT, Bicciato S, Shimizu H, Alford J, Stephanopoulos G. Mining of biological data II: assessing data structure and class homogeneity by cluster analysis. Metab Eng. 2000;2(3):228–38.

99. Coleman M, Block D, editors. Retrospective time-dependent optimization of recombinant E. coli fermentations using historical data and hybrid neural network models. Abstr Pap Am Chem S; 2003: Amer Chemical Soc 1155 16th St, Nw, Washington, DC 20036 USA.

100. Subramanian V, Buck KKS, Block DE. Use of decision tree analysis for determination of critical enological and viticultural processing parameters in historical databases. Am J Enol Viticult. 2001;52(3):175–84.

101. Vlassides S, Ferrier JG, Block DE. Using historical data for bioprocess optimization: Modeling wine characteristics using artificial neural networks and archived process information. Biotechnol Bioeng. 2001;73(1):55–68.

102. Xiao X, Hou YY, Liu Y, Liu YJ, Zhao HZ, Dong LY, et al. Classification and analysis of corn steep liquor by UPLC/Q-TOF MS and HPLC. Talanta. 2013;107:344–8.

103. Hofer A, Herwig C. Quantitative determination of nine water-soluble vitamins in the complex matrix of corn steep liquor for raw material quality assessment. J Chem Technol Biotechnol. 2017;92(8):2106–13.

104. Gao Y, Yuan YJ. Comprehensive quality evaluation of corn steep liquor in 2-keto-l-gulonic acid fermentation. J Agr Food Chem. 2011;59(18):9845–53.

105. Jose GE, Folque F, Menezes JC, Werz S, Strauss U, Hakemeyer C. Predicting mab product yields from cultivation media components, using near-infrared and 2D-fluorescence spectroscopies. Biotechnol Prog. 2011;27(5):1339–46.

106. Kirdar AO, Chen GX, Weidner J, Rathore AS. Application of near-infrared (NIR) spectroscopy for screening of raw materials used in the cell culture medium for the production of a recombinant therapeutic protein. Biotechnol Prog. 2010;26(2):527–31.

107. Li B, Ryan PW, Ray BH, Leister KJ, Sirimuthu NMS, Ryder AG. Rapid characterization and quality control of complex cell culture media solutions using raman spectroscopy and chemometrics. Biotechnol Bioeng. 2010;107(2):290–301.

108. Xiao X, Hou YY, Du J, Liu Y, Liu YJ, Dong LY, et al. Determination of main categories of components in corn steep liquor by near-infrared spectroscopy and partial least-squares regression. J Agr Food Chem. 2012;60(32):7830–5.

109. Afseth NK, Segtnan VH, Wold JP. Raman spectra of biological samples: a study of preprocessing methods. Appl Spectrosc. 2006;60(12):1358–67.

110. Rinnan A, van den Berg F, Engelsen SB. Review of the most common pre-processing techniques for near-infrared spectra. Trac-Trend Anal Chem. 2009;28(10):1201–22.

111. Roggo Y, Chalus P, Maurer L, Lema-Martinez C, Edmond A, Jent N. A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. J Pharmaceut Biomed. 2007;44(3):683–700.

112. Xu L, Zhou YP, Tang LJ, HL W, Jiang JH, Shen GL, et al. Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration. Anal Chim Acta. 2008;616(2):138–43.

113. Kroll P, Sagmeister P, Reichelt W, Neutsch L, Klein T, Herwig C. Ex situ online monitoring: application, challenges and

opportunities for biopharmaceuticals processes. Pharm Bioprocessing. 2014;2(3):285–300.

114. Norgaard L, Saudland A, Wagner J, Nielsen JP, Munck L, Engelsen SB. Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. Appl Spectrosc. 2000;54(3):413–9.

115. Barker M, Rayens W. Partial least squares for discrimination. J Chemom. 2003;17(3):166–73.

116. Brereton RG, Lloyd GR. Partial least squares discriminant analysis: taking the magic away. J Chemom. 2014;28(4):213–25.

117. Stenlund H, Gorzsas A, Persson P, Sundberg B, Trygg J. Orthogonal projections to latent structures discriminant analysis modeling on in situ FT-IR spectral imaging of liver tissue for identifying sources of variability. Anal Chem. 2008;80(18):6898–906.

118. Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). J Chemom. 2002;16(3):119–28.

119. Balabin RM, Lomakina EI. Support vector machine regression (SVR/LS-SVM)-an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data. Analyst. 2011;136(8):1703–12.

120. Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometr Intell Lab. 1987;2(1–3):37–52.

121. Guebel DV, Canovas M, Torres NV. Analysis of the escherichia coli response to glycerol pulse in continuous, high-cell density culture using a multivariate approach. Biotechnol Bioeng. 2009;102(3):910–22.

122. Lugli E, Roederer M, Cossarizza A. Data analysis in flow cytometry: the future just started. Cytometry Part A. 2010;77a(7):705–13.

123. Huang J, Kaul G, Cai C, Chatlapalli R, Hernandez-Abad P, Ghosh K, et al. Quality by design case study: an integrated multivariate approach to drug product and process development. Int J Pharm. 2009;382(1):23–32.

124. Eros D, Keri G, Kovesdi I, Szantai-Kis C, Meszaros G, Orfi L. Comparison of predictive ability of water solubility QSPR models generated by MLR, PLS and ANN methods. Mini Rev Med Chem. 2004;4(2):167–77.

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

126. Wold, Herman. "Partial least squares." Encyclopedia of statistical sciences (1985).

127. Næs T, Isaksson T, Fearn T, Davies T. A user friendly guide to multivariate calibration and classification. Chichester: NIR publications; 2002.

128. Landgrebe D, Haake C, Hopfner T, Beutel S, Hitzmann B, Scheper T, et al. On-line infrared spectroscopy for bioprocess monitoring. Appl Microbiol Biotechnol. 2010;88(1):11–22.

129. Sivakesava S, Irudayaraj J, Demirci A. Monitoring a bioprocess for ethanol production using FT-MIR and FT-Raman spectroscopy. J Ind Microbiol Biot. 2001;26(4):185–90.

130. Hamburg JH, Booth DE, Weinroth GJ. A neural network approach to the detection of nuclear material losses. J Chem Inf Comp Sci. 1996;36(3):544–53.