Monitoring Product Quantity, Purity and Potency of Biopharmaceuticals in Real-time by Predictive Chemometrics and Soft sensors

Astrid Dürauer[1], Alois Jungbauer[1], and Theresa Scharl[2]

[1]Universitat fur Bodenkultur Department fur Biotechnologie
[2]Universitat fur Bodenkultur Wien Institut fur Statistik

March 16, 2023

Abstract

The biopharmaceutical industry is still running in batch mode, mostly because it is a highly regulated industry sector. In the past, sensors were not readily available and in-process control was mainly executed off-line. The most important product parameters are quantity, purity and potency besides adventitious agents and bioburden. There is increasing economic pressure on time-to-market and also on the environmental sustainability of biopharmaceutical manufacturing. New concepts for manufacturing using disposable single-use technologies and integrated bioprocessing will dominate the future of bioprocessing. In order to ensure the quality of pharmaceuticals initiatives such as Process Analytical Technologies, Quality by Design and Continuous Integrated Manufacturing have been established. The vision must be that these initiatives together with technology development pave the way for process automation and autonomous bioprocessing without any human intervention. Then a real-time release would be realized leading to a highly predictive and robust biomanufacturing system. The steps toward such automated and autonomous bioprocessing are reviewed in context of monitoring and control. Starting from statistical treatment of single and multiple sensors, establishing soft sensors with predictive chemometrics and hybrid models. A scenario is described how to integrate soft sensors and predictive chemometrics into modern process control. This will be exemplified by selective downstream processing steps such as chromatography and membrane filtration, the most common unit operations for separation of biopharmaceuticals.

For submission to Biotechnology and Bioengineering 2023 Special Edition Recovery of Bioproducts

Monitoring Product Quantity, Purity and Potency of Biopharmaceuticals in Real-time by Predictive Chemometrics and Soft sensors

Astrid Dürauer[1*], Alois Jungbauer[1,2], and Theresa Scharl[3]

[1]Institute of Bioprocessing Science and Engineering, University Natural Resources and Life Sciences, Vienna Austria

[2]Austrian Centre of Industrial Biotechnology, Vienna, Austria

[3]Institute of Statistics, University Natural Resources and Life Sciences, Vienna Austria

*Corresponding author:

Institute of Bioprocessing Science and Engineering

Muthgasse 18

1190 Vienna

Austria

e.mail: astrid.duerauer@boku.ac.at

Abstract

The biopharmaceutical industry is still running in batch mode, mostly because it is a highly regulated industry sector. In the past, sensors were not readily available and in-process control was mainly executed off-line. The most important product parameters are quantity, purity and potency besides adventitious agents and bioburden. There is increasing economic pressure on time-to-market and also on the environmental sustainability of biopharmaceutical manufacturing. New concepts for manufacturing using disposable single-use technologies and integrated bioprocessing will dominate the future of bioprocessing. In order to ensure the quality of pharmaceuticals initiatives such as Process Analytical Technologies, Quality by Design and Continuous Integrated Manufacturing have been established. The vision must be that these initiatives together with technology development pave the way for process automation and autonomous bioprocessing without any human intervention. Then a real-time release would be realized leading to a highly predictive and robust biomanufacturing system. The steps toward such automated and autonomous bioprocessing are reviewed in context of monitoring and control. Starting from statistical treatment of single and multiple sensors, establishing soft sensors with predictive chemometrics and hybrid models. A scenario is described how to integrate soft sensors and predictive chemometrics into modern process control. This will be exemplified by selective downstream processing steps such as chromatography and membrane filtration, the most common unit operations for separation of biopharmaceuticals.

Keywords: Real-time release, machine learning; process control, continuous integrated biomanufacturing;

Introduction

The biopharmaceutical industry is still running in batch mode, mostly because it is a highly regulated industry sector (Kumar et al., 2020), but there is increasing economic pressure on time-to-market and also on the environmental sustainability of biopharmaceutical manufacturing. New concepts for manufacturing using disposable single-use technologies are being increasingly implemented and manufacturing is turning away from dedicated manufacturing suites to ballroom facilities suited for manufacturing multiple products (Müller et al., 2022). Integrated continuous biomanufacturing will expedite and leverage these developments because it is a way to reduce the footprint of a manufacturing plant and reduce the time required to produce clinical batches(Cataldo et al., 2020; Ding et al., 2022; Farid, 2019; Konstantinov et al., 2015; Patil et al., 2018). The status-quo is still batch processing with quality by testing. After each step samples are taken, tested and then the subsequent step is executed. The PAT initiative ("Guidance for Industry: PAT-A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance," 2004) has encouraged to implement advanced analytical technologies in form of at-line, on-line and off-line analytics to "design, develop and operate processes consistently to ensure a predefined quality at the end of the manufacturing process" (Figure 1). It was obvious from the start that chemometric approaches will allow us to get better process understanding instead of interpreting the analytical results one by one (Lopes et al., 2004; Wasalathanthri et al., 2020).

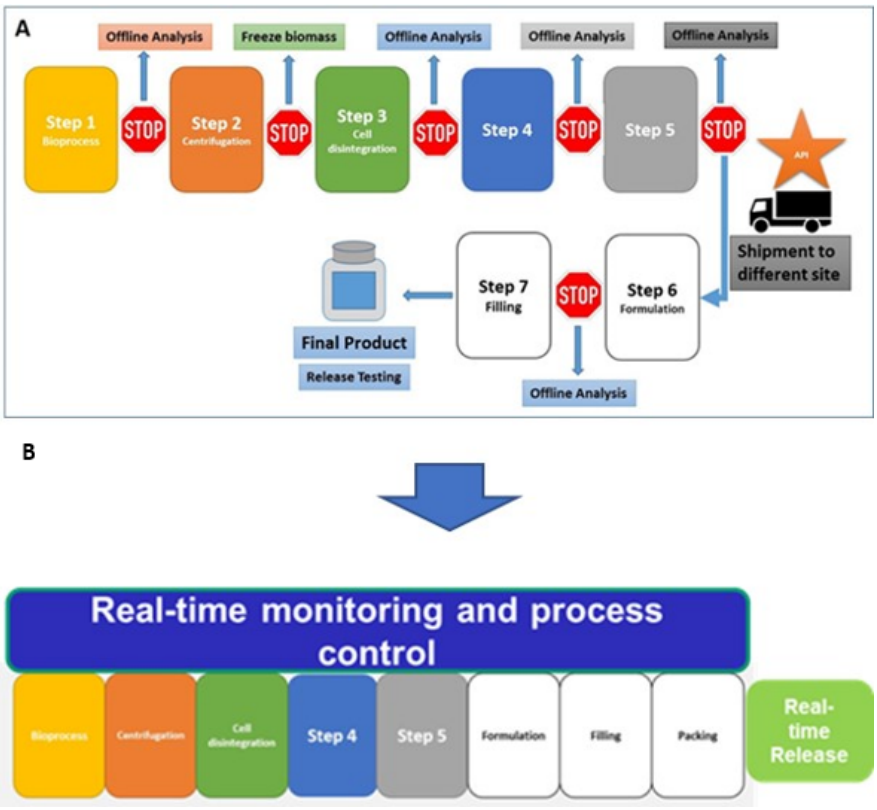


Figure 1: Approaches for Quality Assurance in Biopharmaceutical Production (A) Quality by testing approach – after process steps samples are drawn to determine product quality and quantity in the laboratory. As this is a retrospective approach, it does not allow any process control and leads to failure prone production even though it is costly and time consuming. (B) Quality by Design approach – the process is monitored in real-time and therefore product quality and quantity is known at each stage of the process. If the process is performed within the set specifications, the product will finally have the desired quality. Real-time monitoring of the process allows process control and therefore enhances product quality and reduces time and material consumption for offline analysis.

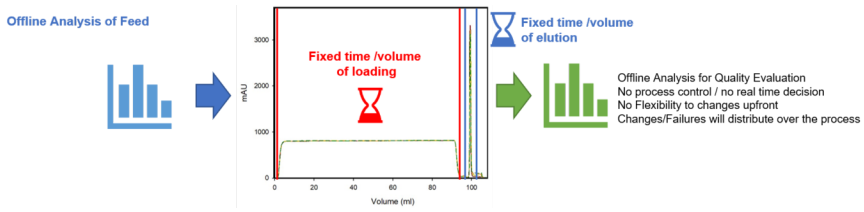
Then the subsequent quality by design initiative's main aims (Q8-Q11 quality guidelines of the International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH) (The International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH),) should "ensure that all sources of variability affecting a process are identified, explained, and managed by appropriate measures. This enables the finished medicine to consistently meet its predefined characteristics from the start". This considers the raw material variations and all variations of the process parameters. Furthermore, recently a guidance has been released to recommend continuous integrated biomanufacturing. This guidance also recommends an in depth process understanding but does not recommend process control. The guideline of real-time release has been developed because "under specific circumstances an appropriate combination of process controls (critical process parameters) together with pre-defined material attributes may provide greater assurance of product quality than end product testing and the context as such be an integral part of the control strategy principle". Furthermore, "RTRT may be applied during the stages of manufacture of chemical and biological products resulting in the elimination of all, or certain, specific tests in the specifications of the finished active substance or finished medicinal product". The ultimate goals must be to monitor and control the process in real-time so that we are able to efficiently control the variability of raw materials, variability of the process parameters and in the inherent variability of the biological

system(Wasalathanthri et al., 2020). If we want to control the process, we need real-time information on quantity, purity and potency, - the critical quality attributes (CQA) - because only then can we control the critical process parameters. This does not differ from current approaches, but everything will be real-time instead of retrospective. It is obvious that we do not have dedicated sensors for quantity, purity and potency. We must use indirect methods, but there is a reluctance in industry and academia to apply soft sensors. A soft sensor (Luttmann et al., 2012; Mandenius et al., 2015; Roch et al., 2016), also known as a virtual sensor, is not a real existing sensor, but a simulation of interdependencies between representative measured variables and a target variable such as product concentration, aggregate content, dsDNA content, or a biological activity of a biopharmaceutical for instance. Then the soft sensors can be used to control a system(Ender et al., 2003). Especially in downstream processing when executed in batch we have limited possibility to really control the process. Therefore, a further incentive for continuous integrated manufacturing beyond the economic gain and improvement of the environmental footprint is the ability to fully control and automate a process and to arrive at system runs autonomously. Such a system would also contribute to a digital twin of a manufacturing plant, but such a twin is not required for an automated process. Process control is already well established in upstream processing but not in downstream processing. Control of oxygen and pH has been already done in the very early days of production of biopharmaceuticals in bioreactors. In an advanced bioreactor control systems air, O_2, CO_2, in the off gas and dissolved oxygen and temperature are measured in-line together with NIR, Raman, turbidity and capacitance to infer nutrient concentration and waste products or on-line sampling is conducted measured it directly by at-line enzymatic assays or HPLC (Zhao et al., 2015).

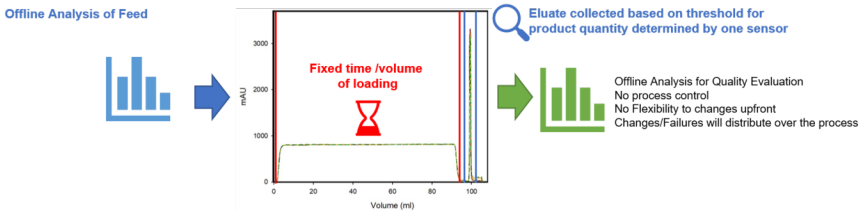
In advanced downstream processing, chromatographic workstations are equipped with UV 280 flow cell, pH, conductivity probes (Carta et al., 2010). Advanced sensors such as NIR, AT -FTIR, Raman multiangle light scattering, and dynamic light scattering are not routinely used as sensors for downstream process monitoring, although they would provide very useful information that goes far beyond the standard information. The signal from such sensors yields a spectrum that must be further interpreted using a so-called deconvolution process, for example, to assign a particular wavelength to a molecular property. These chemometric approaches were developed simultaneously with the development of the sensor hardware. Recently, it was found that the peak profile itself could be used to correlate it with certain properties of the process fluid or biomolecule, such as the glycosylation pattern of antibodies, aggregate content, or truncated variants. In such a case, the entire spectrum is used instead of a specific wavelength, and if the spectral information is trained with off-line data, it can be considered a soft sensor(Feidl et al., 2019; Kornecki et al., 2018).

Decades ago, bioprocessing originally started with fixed time/volume processes and even processes are performed without a single sensor. The next stage is threshold-based processing based on either a single sensor or multiple sensors. More sophisticated methods for using sensor data include chemometric methods, advanced statistical methods, and hybrid models (Figure 2). Hybrid models and model predictive control are the culmination of a monitoring and control system that ultimately enables real-time release.

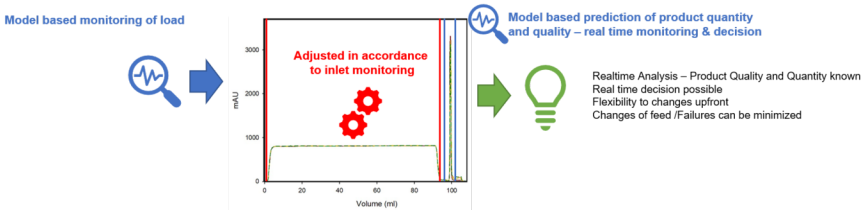
Fixed Time/Volume



Threshold based



Model based online monitoring



Model predictive control

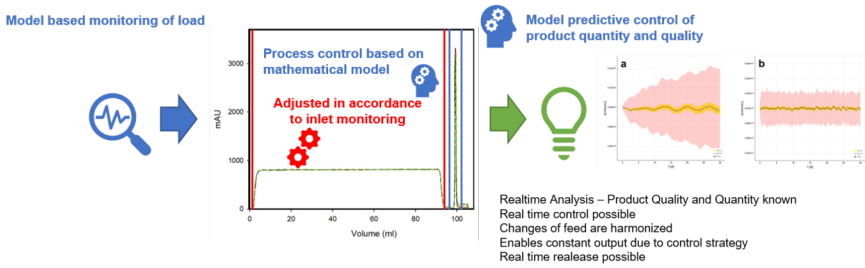


Figure 2: Typical approaches how to conduct a biopharmaceutical process shown for a chromatographic step exemplarily. It can be either (A) Based on Fixed time / volume for loading and elution; (B) Threshold based e.g. stop loading or start collecting the eluate at a certain UV absorbance; (C) Model based online monitoring with sensors implemented before and / or after the column to predict product quality and quantity in feed and / or eluate; (D) Model predictive control. While (A) and (B) rely on time and material offline analysis and do not allow in process control, (C) and (D) do need offline analysis only for the model training and enable real-time decision making and real-time release.

In this review, we will elaborate how we can arrive at a system for monitoring and control to run a bioprocess, autonomously; we will mainly on downstream processing. An ideal bioprocess from the point of view of real-time monitoring, control and real-time release is outlined.

Statistical modelling concepts

In real-time monitoring we rely on establishing correlations between sensor signals, process parameters and off-line measured quantity and quality parameters. It is pivotal how a correlation is established because it will impact highly correlated predictors (Figure 3).

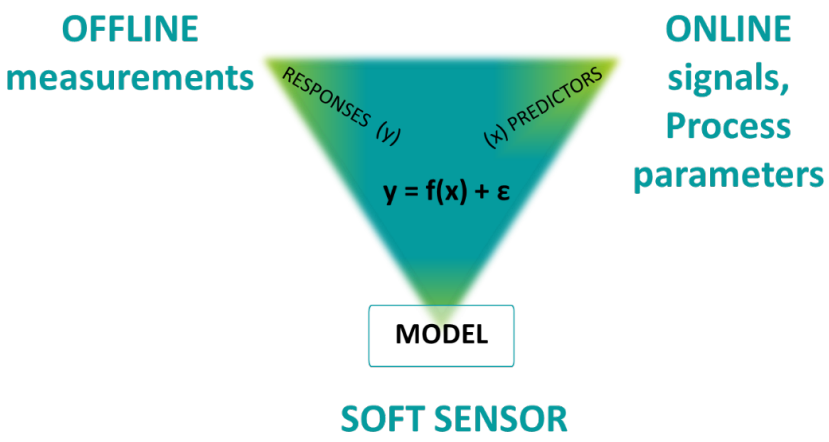


Figure 3: Concept how a soft sensor is generated by correlation of offline measurement with on-line signals from sensors and process parameters for Critical Quality Attributes QCAs.

Correlation analysis is used to measure the extent of the relationship between variables (Varmuza et al., 2009). Typically, only the extent of the linear relationship is considered using Pearson correlation. However, there are also methods available to measure nonlinear relationships, e.g., Spearman's rank correlation, which is a nonparametric measure of rank correlation, reporting the statistical relationship between the rankings of two variables (Lee et al., 2000). The correlation structure in the data has a major impact on the subsequent analysis of the data. It is important to note that correlation does not imply a causal relationship between the variables, i.e., one variable is affected by another.

Often, we want to model a critical quality attribute (CQA) of a downstream process based on one or many input variables. Typically, measuring the CQA is laborious, cost intensive and takes hours or even days. On the other hand, the input variables are easy to measure online using sensors like the standard UV detector, the standard pH and conductivity probe or spectroscopic data. We distinguish between situations where the relationship can be described by a fundamental scientific law (first-principle model), by a relatively simple mathematical equation (based on physical/chemical knowledge) and purely data-driven models where we only assume that relationships exist (Varmuza et al., 2009). Coming up with a fundamental scientific law in downstream processing is a highly complex task which requires an immense number of experiments, therefore data-driven approaches appear to be more suitable (Rathore et al., 2022a). A compromise is sometimes found in so-called grey-box models or hybrid models (Hong et al., 2018; Simon et al., 2015a). Hybrid modelling approaches have already been successfully used in downstream processing to model the flux evolution and duration of ultrafiltration processes (Krippl et al., 2020; Krippl et al., 2021) and capture chromatography (Narayanan et al., 2021) (Figure 4).

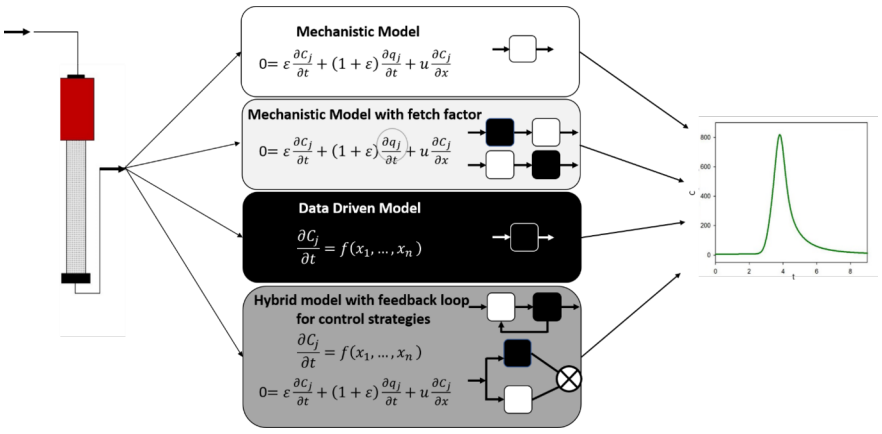


Figure 4: Categories of models based on the underlying data and assumptions exemplified for the mass balance of a chromatographic separation. **Mechanistic models** are based on biophysical relationship and first principle knowledge packed into a mathematical model. These models are sometimes synonymously called "white box models" as the have to be based on full process understanding. All parameters are determined by independent experiments. For**mechanistic models with fetch factor** individual parameters are determined experimentally, they are data driven. **Data driven models** are mathematical models solely based on statistical relationships of data -mostly derived by online sensors and offline analysis. They are not based on biophysical relations. Therefore, they are also called "black box models". However, these models depict the entity of the process space. **Hybrid models** combine mechanistic and data driven models and can compensate the first ones for lack of understanding of processes by data collected throughout the process. They are of special interest for bioprocess modelling as in most cases those processes are not fully understood and describable.

Multiple linear regression (MLR) is one of the most frequently used methods in multivariate data analysis. It can easily be applied and interpretation of the effect of the individual predictors on the response is straightforward. However, this is only true if there are many more observations than predictors in the model and there is no correlation structure among the predictors. In such situations linear regression can be used to model a clearly understood mechanistic relationship between variables. On the other hand, if the predictors are highly correlated, interpretation of the individual effects of those predictors is no longer possible. Individual effects can be masked by several correlated variables. Therefore, it is mandatory to check the correlation structure between the input variables prior to setting up an MLR model. In the case of multi-collinearity, no solution can be calculated at all as the covariance matrix becomes singular (Varmuza et al., 2009). In many practical situations MLR is still used even if the predictors included in the model are moderately correlated. In such situations the model is selected due to fitting the response well rather than for interpretability of the individual effects of the input variables. This is a very pragmatic and purely data-driven approach where model selection is based solely on the performance measure. Spurious correlations are used, there is no need for a causative correlation, because we exploit the data structure, e.g., we control and measure the pumps, pH, conductivity and pressure in a chromatography run and search for a relationship with multiple sensors. Finally, we can predict CQAs.

Model Selection, Training and Overfitting

For model selection the performance of different models on so-called training data must be evaluated accordingly. Typical performance measures are the (MSE), the root-mean-squared-error (RMSE) or the mean relative deviation (MRD).

The MSE measures the squared differences between the observed (measured) values and the ones predicted by the model and takes the mean over all these differences. The MSE is often used during model building

and parameter optimization. However, it is not suitable for evaluation of the prediction error as it is a squared quantity. To get an idea of the size of the error the root of the MSE is used, the RMSE, which is in the same unit as the CQA that was modelled. The MRD is the mean absolute difference between the observed and the predicted value standardized by the observed value. Often it is given in percentages.

It is not valid to calculate a performance measure on the data the model was built on. Instead, an independent test set needs to be used. A split of the available data into a training set (ca. 50% of the objects), a validation set (ca. 25% of the objects) and a test set (ca 25% of the objects) is often recommended (James et al., 2021a; Varmuza et al., 2009).

Here the training data is used for model building and the validation data is used for parameter optimization. This split of the data is very useful to avoid overfitting, i.e., the model performs very good on the training data but does not generalize well to new data the model has not seen during model building. The performance of different models can then be evaluated on the independent test set in order to check how well it generalizes. If only a limited number of observations are available for model building the split must be adapted. Depending on the statistical method used, an individual validation set might not be necessary as parameter optimization is done on the training data via cross-validation, e.g., when using PLS. In this case 75% or even 80% of the data can be used for model training (Westad et al., 2015).

Experimental Design vs Design of Experiments – how much experimental data do we need?

A model will only perform well on the test set if the study was designed properly. Still, as we are working with biological material the choice of splitting the data into training – validation – and test set has a major impact on the performance measure. Therefore, multiple splits and/or cross validation as an efficient method to reuse the data are recommended (Gareth et al., 2021; James et al., 2013, 2021a). When designing a study, it is important to keep the goal of the study in mind. The prediction model could be used in a biopharmaceutical production process where the process conditions are fixed, and only little deviations are expected. Alternatively, real-time monitoring could also be used during process development where many different process conditions need to be explored. In (Felfödi et al., 2020) it was shown how the precision of the analytical method also impacts the data needed to set up a prediction model. For model training, the number of fractions for the off-line analysis together with the number of chromatographic runs performed is crucial. The prediction model can only be of high quality if the off-line analysis is very accurate (expressed as coefficient of variation). Therefore, a compromise must be found between the number of fractions analyzed, the number of chromatographic runs and the performance of the prediction model expressed as RMSE (Figure 5).

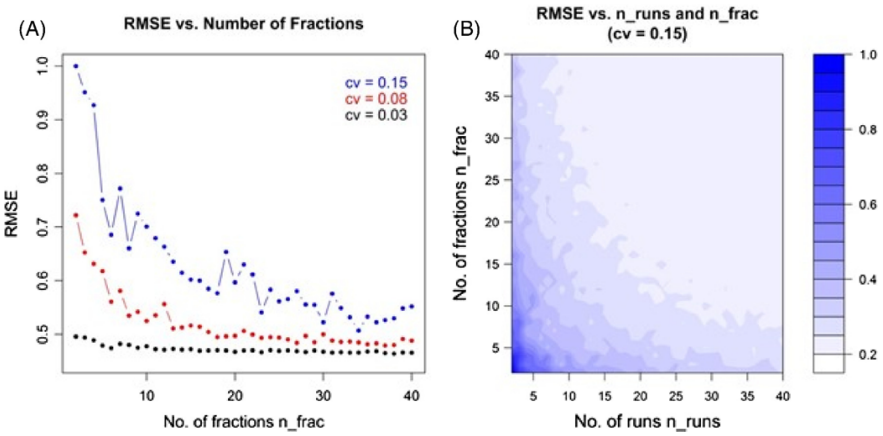


Figure 5: Example to demonstrate how many fractions and how many runs are required to train a statical model in chromatography. Modeling error as a function of number of fractions per run (n_{f}) for values of

coefficient of variation (cv) representing accuracy of analytical method. (B) Pseudo-3D contour plot with prediction error as a function of bothn_{r} and n_{f}. From (Felfödi et al., 2020)

Interpolation versus extrapolation

When setting up a prediction model the design space of critical process parameters has to be considered.

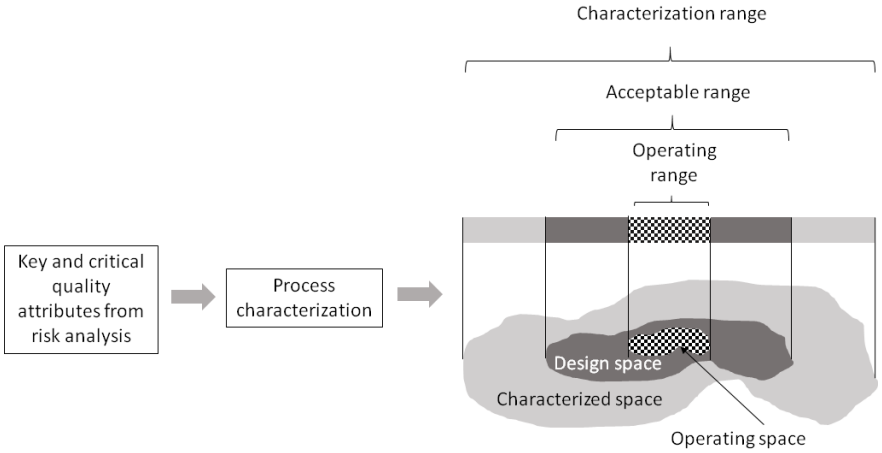


Figure 6: The basic principles of the quality by design approach, where more flexibility is obtained by a characterized space; adapted from (Carta et al., 2010).

Ideally, the full design space is covered by the training, validation and test data. In this ideal situation interpolation within the design space will be very good (Figure 6). Rolinger et al. (Rolinger et al., 2021) five runs in their study. Four runs covered the design space, and the center point was used as an independent test set. It is not surprising that the model performed well. Sauer et al and Walch et al. (Sauer et al., 2019; Walch et al., 2019) on the other hand simulated the situation of a manufacturing run under clearly defined process conditions and used eight runs performed under identical process conditions for model building. The performance was then evaluated on purification runs performed under identical conditions but with biological material from a new fermentation and under different process conditions. It was shown that the model performed well on new biological material (Sauer et al., 2019) and that model extrapolations are possible but are limited to modifications in the column load and small changes in buffer composition (Walch et al., 2019).

Machine Leaning Methods

Multiple statistical methods can be used to set up a predictive model. In simple situations with linear relationships and only a few possible predictors, even linear models can be sufficient. Depending on the aim of the model, fast and simple methods like PLS can be used to get a rough idea of the prediction performance in process development. If the model should be used in real-time in a production process, more advanced non-linear methods such as Random Forest or Neural Networks should be considered. Here we give an overview of the advantages and disadvantages of some selected machine learning methods (Table 1).

Table 1: Most prominent statistical methods used for prediction of quality and quantity in bioprocess monitoring and control

Method	Number of Variable	Variable Selection necessary	Applications
MLR	Small, uncorrelated	yes	Simple, linear
Partial Least Squares Regression PLS	Large, typically correlated	no	First idea of a
Structured Additive Regression STAR	moderate	yes	Production pr

Method	Number of Variable	Variable Selection necessary	Applications
Random Forests RF	large	no	Quick idea of a
Support Vector Machines Regression SVM	large	no	Production pro
Neural Networks NN	moderate	yes	Production pro
Deep Learning DL	huge	no	Production pro
Gaussian Process Regression GPR	moderate	no	Process develo

Partial Least Squares Regression - PLS

When the number of predictors is large or even larger than the number of observations and the predictors are highly correlated, e.g., when using spectroscopic data, Partial Least Squares (PLS) regression (Wold et al., 2001) is frequently used (Brestich et al., 2018; Christler et al., 2021; Felfödi et al., 2020; Rüdt et al., 2017; Walch et al., 2019). PLS can easily be applied and has the big advantage that model training is very fast. PLS transforms the original predictor variables into a set of latent variables, which are also linear functions of the original predictors. Linear combinations are determined such that a maximum covariance between the scores (the values of the latent variables) and the response is achieved. The number of components (latent variables) is an optimization parameter. It is usually determined within the framework of cross-validation (CV). These latent variables are then used as predictors in a multiple linear regression model for the CQA. If you want to optimize a PLS model different subsets of predictors should be considered as inputs as selection of the number of latent variables alone is usually not sufficient (Walch et al., 2019). Ultimately, PLS models are a good starting point for modelling CQA when setting up real-time monitoring or in process development based on spectroscopic data. However, if you want to implement a real-time prediction model more advanced methods should be considered.

Structured Additive Regression - STAR

The relationships between the response and the different input variables are usually non-linear. PLS however, is a linear regression method. Obviously, non-linearities are present in complex natural product mixtures. These can be incorporated using Structured Additive Regression (STAR) models (Fahrmeir et al., 2004), an extension of linear models. Within this framework the machine learning technique boosting is often used for variable selection (Hofner et al., 2015; Hofner et al., 2011). This method works very well when the number of predictors is moderate as optimization is computationally very intensive (Sauer et al., 2019).

Support Vector Machines Regression - SVM

A further method is Support Vector Machines (SVM) regression (e.g., (Li et al., 2007)). SVMs can achieve an optimum network structure by a compromise, balancing the quality of the approximation of the given data and the complexity of the approximation function. SVMs are boundary methods, they do not try to model a group of objects. Because the boundary can be very complex attention should be paid to the problem of overfitting. SVM was originally used for classification purposes, but its principles can be extended easily to the task of nonlinear regression by the introduction of an alternative loss. The basic idea of SVM regression is to map the original data set into the mapped data set in a high dimensional feature space via a nonlinear mapping function (so-called kernel functions), and then perform a linear regression in this feature space. Defining the linear regression function in this feature space, nonlinear function regression in the original space becomes a linear function regression in the feature space. SVMs have been included in recent studies on continuous biomanufacturing (Nikita et al., 2022). The selection of an appropriate kernel function is data dependent and needs expert knowledge. Additionally, hyperparameter tuning must be performed in order to avoid overfitting.

Gaussian Process Regression

Gaussian Process Regression (GPR) is a non-parametric, Bayesian machine learning method that infers probability distributions rather than the exact measured values (Rasmussen et al., 2021). These models

are particularly attractive as they can be used for small data sets, do not only provide flexible models but also give a model-based estimate of the prediction error. Recently, Gaussian Process Regression (GPR) has received increasing attention in the field of biotechnology. di Sciascio (di Sciascio et al., 2008)used GPR for the development of a biomass concentration estimator, whereas Hutter-2021 (Hutter et al., 2020) used GPR to efficiently learn from multiple product spanning process data. However, in contrast to MLR, different software implementations will give different error estimates due to different parameter definitions and different numerical optimization (Erickson et al., 2018). A further limitation is that GPR they do not scale well with increasing data size.

Tree-based methods

Alternative methods are tree-based models like regression trees or Random Forests (Breiman, 2001). A regression tree is a hierarchical model where observations are recursively split into binary partitions based on their predictor values. Random Forests are ensemble learning methods where the predictions are obtained by averaging over hundreds or even thousands of trees built on bootstrap samples, i.e., samples taken from the training data with replacement. Recently it was shown that these methods perform very well in downstream processing (Nikita et al., 2022). Random Forests are very popular due to the build-in permutation-based variable importance measure. This approach was used to find suitable inputs for Artificial Neural Networks (Melcher et al., 2015). Model tuning is very important but setting up a prediction model is straightforward and fast, especially in comparison to ANNs.

Artificial Neural Networks - ANN

Artificial Neural Networks (e.g., (James et al., 2021b; Lecun et al., 1989)) are nonlinear statistical models that have been used in bioprocess modelling since the early days (Glassey et al., 1997). They can simulate highly nonlinear dynamic relationships of the process without prior knowledge of the model structure. In a single-layer-neural-network (often called shallow network, (Lee et al., 2018)), different linear combinations of the input variables are built and then a nonlinear function, e.g., the sigmoid function is applied. Since these new variables are not directly observed, they are called hidden units, and often they are arranged in a graphical representation as a hidden layer. The new variables can be used in a linear or nonlinear regression model resulting in an output variable. In classical ANNs, more hidden layers can be used but with a great tendency to overfitting. Further disadvantages of classical ANNs are related to convergence speed, network topology and bad local minima. After 2010, neural networks gained further attention mainly in the field of image classification with the name Deep Learning where “deep” refers to the number of hidden layers. The architecture of a Convolutional Neural Network (CNN) was particularly successful (Lecun et al., 1989). They are now very successful in comparison to other machine learning techniques due to major computer hardware improvements, the use of graphical processing units and much larger data sources. Nowadays, modern neural networks use the ReLU activation function instead of the sigmoid function and consist of multiple hidden layers. Deep learning was recently used for real-time quality prediction and process control (Nikita et al., 2022). They applied deep neural networks and compared it to SVM, decision trees regression and random forests on time series data of UV, conductivity and pH probes of 84 batches. In this study, random forests and decision trees outperformed the deep neural networks maybe due to the relatively small number of predictors. Rolinger-2021 (Rolinger et al., 2021) compared PLS to CNNs for their ability to quantify the mAb concentration in the column effluent based on UV and Raman spectroscopy. In this study there was also no need to use deep learning as a UV-based PLS model was already sufficiently precise. Deep learning was developed for large or big data in terms of both experiments and input variables.

Variable selection and feature importance

For classical ANNs, variable selection had to be performed prior to setting up the model due to slow convergence. Random forests were successfully used to come up with an optimal set of inputs for the ANN (Melcher et al., 2015) and extensive variable selection was performed prior to establishing the STAR and PLS models (Sauer et al., 2019; Walch et al., 2019). For this purpose, prior knowledge such as amide bands and fingerprint regions of the spectral data was used to select informative predictors. Additionally, highly

correlated wavelengths were removed by reducing the resolution of the spectra. This required a considerable amount of domain knowledge and engineering skills. In deep learning, these preparatory steps are no longer required as the method is now capable of dealing with a very large number of inputs and weights thereof are computed automatically (Lecun et al., 1989). Of course, using a very large number of inputs for a prediction model comes at the expense of interpretability of the model.

Explainable Machine Learning

In the machine learning community, a lot of research is currently going on in trying to explain what is going on inside prediction models (Roscher et al., 2020; Zhong et al., 2022). Statistical, purely data-based models are often called black box models in the field of hybrid modelling (Simon et al., 2015a; Simon et al., 2015b). Of course, they are no black boxes. The impact of individual input parameters on latent variables in PLS or the weights in ANNs can be extracted and interpreted. If the software including the underlying code to generate a certain model is freely available such models can be considered as white-box models as all information is available. Black-box models only occur if the code is not shared as usually the case in commercial software (Winter et al., 2021). However, especially if the machine learning models are based on high-dimensional data, the interpretation of the model parameters can be complex. This process is more complex than understanding first-principal models with a clear mechanistic relationship between the input and output variables. This makes the efforts in the field of explainable machine learning even more important and promising for the future of predictive chemometrics.

Real-time monitoring

Real-time monitoring of critical quality attributes (CQA) or critical process parameters (CPP) using real-time monitoring by means of soft sensors (Mandenius et al., 2015), (Roch et al., 2016) has enormous advantages over the determination of those off-line after the unit operation. It is much faster and can assess attributes, where no dedicated sensor or analytical method is available. It also provides information in real-time if the process is within the operating range. These soft sensors are also called prediction models (e.g., (Sauer et al., 2019)). Often, we also want to monitor other attributes, which are beyond the CQA, because they are relevant for process economics, but not relevant in respect to patient safety. In a conventional process, particularly in DSP the samples are taken from a pool or intermediates and then subjected to further off-line analysis. Interventions in the upstream process do not require a fast reaction because cells grow slowly and a response after up to an hour is sufficient. Even an at-line measurement using automatic sampling devices with a 30-min lag time is considered as a "real-time" measurement because the return of the measurement is fast enough to control the process (Zhao et al., 2015). Whereas in a lot of situations in downstream processes the decision must be made within seconds, e.g., when a stream is diverted from waste to product collection. Then the decision must be made within seconds and time-consuming analytics become useless.

Multiple-sensor predictive chemometrics approach

If only a single quality attribute is monitored such soft sensors are often based on single sensors, e.g., the monitoring of the product quantity by UV spectroscopy (Rolinger et al., 2021) or attenuated total reflection Fourier-transform infrared (ATR-FTIR) spectroscopy (Rudt et al., 2019). However, multiple sensors must be combined even if only the process signals are used that a chromatographic workstation is typically equipped with, e.g., UV, conductivity and pH probe (Nikita et al., 2022). If several CQAs need to be modelled in real-time, it is indispensable to use a multiple sensor approach in order to capture the different properties of these CQAs (Sauer et al., 2019). Such an approach can be regarded as data fusion (Borras et al., 2015; Liggins et al., 2017; Rolinger et al., 2021). A multiple sensor approach has many advantages, e.g., in the case of sensor saturation where the combination of different input variables leads to redundancies that can compensate for the information loss.

Example for predictive chemometrics approach

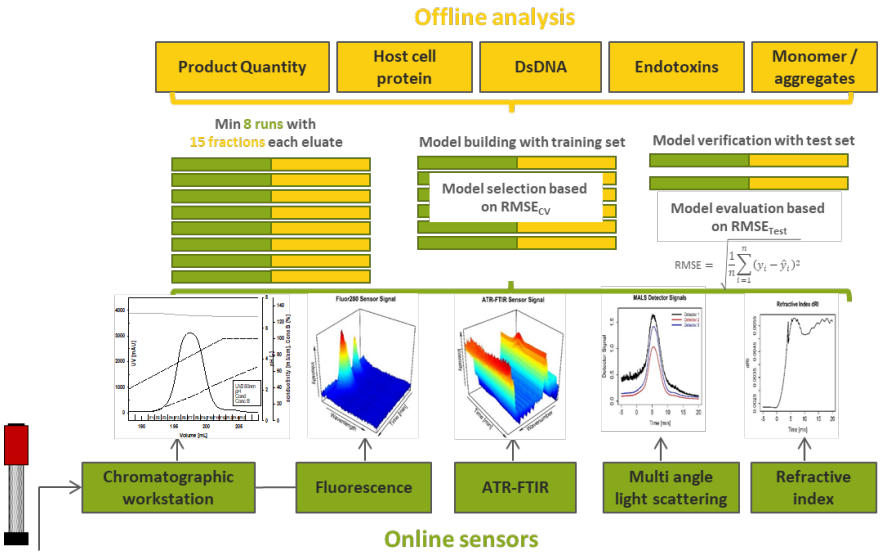


Figure 7: Workflow for model-based real-time monitoring of a chromatographic step adapted from Sauer et a. 2018. In addition to the sensors for pH, conductivity, UV absorbance and pressure which a chromatography workstation is typically equipped with, four online sensors have been implemented in the flow after the column. Online data were obtained for 8 identical runs of a chromatographic purification step. The eluates were aliquoted and collected as 15 fractions and offline analysis was carried out to determine the desired quality attributes, product quantity and impurity content. Part of this data set was then used to establish mathematical models for each quality attribute by relating the offline with the online data. The models were selected via the lowest root mean squared error (RMSE) and then evaluated via their predictability for independent test data sets which have not been used for the model training before. Implemented in the stirring software of the chromatographic workstation, the established models give information on all quality attribute they have been trained on in real-time (<1sec) and enable real-time decision making for e.g. pooling of the eluate for the next step.

Sauer et.al. equipped a chromatographic workstation with multiple sensors (Sauer et al., 2019). Besides standard detectors (UV, pH and conductivity), multi-angle light scattering, refractive index, attenuated total reflection Fourier-transform infrared and fluorescence spectroscopy were included. The real-time monitoring system was used in a cation exchange capture step of fibroblast growth factor 2 expressed in *E. coli* . Eight training runs were performed where 15 fractions of the eluate were analyzed to get information about the product quantity, host cell protein and double-stranded DNA impurities as well as endotoxins and Monomer/aggregates (Figure 7). Prediction models were generated for each individual response variable using cross-validation. The same system was used for an antibody capture process (Walch et al., 2019). The input data of the various devices was time-aligned considering the different void volumes and time resolution. Individual preprocessing methods were applied to the individual sensors together with a variable selection procedure specific for the sensor. Finally, the online signals were averaged over the time intervals of the collected fractions that were analyzed off-line. A multiple sensor approach is only feasible if the chromatographic workstation is equipped with a central database (Oliveira, 2019; Steinwandter et al., 2019). For these multiple sensors the software solution XAMIris (Evon, Austria) was used for the recording of various signals, starting of the chromatographic runs, data export, time-alignment as well as the implementation of soft sensors for real-time monitoring of several CQAs (Christler et al., 2021; Sauer et al., 2019; Walch et al., 2019).

Impact of different sensors

In same multi sensor setup UV/Vis spectroscopy was used as it mainly measures the primary structure, such as the content of aromatic amino acids (UV280nm), polypeptide backbone (UV214nm) or DNA content (UV260nm) (Christler et al., 2021; Sauer et al., 2019; Walch et al., 2019). The refractive index (RI) was included as it was previously used to quantify protein (Zhao et al., 2011). ATR-FTIR can distinguish between HCP and target protein (Capito et al., 2013). Intrinsic fluorescence of the aromatic amino acids can used to measure the tertiary structure of proteins and to detect structural changes induced by polarity (Ghisaidoobe et al., 2014; Rathore et al., 2009). Light scattering methods (Minton, 2016) are used to determine their quaternary structure, for example, protein aggregation. Fluorescence spectroscopy, as well as light scattering techniques, have been used for at-line determination of quality attributes (Patel et al., 2018; Rathore et al., 2009; Yu et al., 2013).

Multiple CQA monitoring- single sensors vs multiple sensors

If multiple CQAs are monitored using multiple sensors decent model selection is required. Extensive investigation of the impact of the individual sensors on the prediction performance was done (Sauer et al., 2019; Walch et al., 2019). Typically, a prediction model is as simple as possible but as complex as necessary. Therefore, models based solely on one single sensor were compared to models with two, three up to all available sensors. The best model was selected based solely on the prediction error, e.g., the root mean squared error (RMSE) of prediction on an independent test set, a purely data-driven approach. However, if the performance of an extensive model including fluorescence and/or ATR-FTIR data only slightly outperformed a basic model, it was still recommended to use the basic model (Sauer et al., 2019). For all investigated CQAs, the finally selected models contained more than one single sensor.

Robustness of the monitoring system - sensor fouling / sensor shift

For the set-up of a real-time monitoring system, it is recommended to implement multiple prediction models as sensor fault, shift or fouling can easily distort the input data and make the prediction models useless. The model used impacts the pooling decision (Walch et al., 2019). Even though the performance of the more complex models was superior on the independent test set, the simpler models were smoother and more robust. An optimal monitoring system should always be based on several prediction models based on different sensor combinations in order to react on sensor fouling or sensor shifts. If one sensor fails, the real-time monitoring can easily be based on an alternative model where the specific sensor is not used. The technology transfer of the monitoring system with multiple sensor (Sauer et al., 2019; Walch et al., 2019) revealed that only a subset of possible prediction models could be used for real-time prediction at the different sites as the fluorescence device was not robust enough (Christler et al., 2021). Simpler models without fluorescence could still be used at the different sites. However, due to the very specific properties of the individual sensors, the performance of the prediction models could be considerably improved by new model training.

Towards real-time monitoring, control and real-time release

Real-time monitoring provides information if the process runs within the specified operating range. This is useful information and can be linked to release criteria, but it could be further exploited for process control. Then raw material variations, fluctuations in process parameters could be controlled and a much more robust process can be obtained. This is then also the step towards automation, because human intervention could be minimized. In bioreactors real-time monitoring is used to control the process to deliver optimal nutrients to reach a high titer or product quality. A typical example is the addition to manganese ions to trigger the glycosylation of antibodies (Tharmalingam et al., 2015) addition of amino acids or control the level of glucose at a minimum level. Bioreactor control is much more advanced compared to downstream processing. Besides pooling there are not a lot of opportunities to control batch chromatography. An interesting approach is to control the gradient shape in batch chromatography to optimize resolution and binding capacity (Sellberg et al., 2017). The authors named it salt trajectory. An optimizer adapts the

shape of the gradient according to the retention behavior of product and impurities and the salt gradient is modulated. The wider application of such a controls system would require a fast-real-time monitoring of the effluent stream of the chromatography column, because they still used off-line measurements of the column effluent. Such a system maximizes productivity while resolution is not affected. Usually increasing productivity will be obtained at the expense of reduced resolution. Implementation of soft sensors such as the (Walch et al., 2019) into a column chromatography with a gradient optimizer would lead to an optimal system where product quality and quantity is monitored, and the information is further utilized to control the system and operate under high resolution.

Another dilemma is the optimization of dynamic binding capacity. When chromatography is operated with constant velocity an optimum of productivity at a certain residence time is achieved. At this residence time the column utilization is rather low, but by increasing the column utilization the productivity is reduced. At high velocity/low residence time the breakthrough curve is shallow and at low velocity steep, therefore the dynamic binding capacity and column utilization is higher at low velocity (Carta et al., 2010; Eslami et al., 2022a; Eslami et al., 2022b). One solution is countercurrent chromatography, which allows maximizing column utilization and productivity, but on expense equipment/instrument complexity(Heeter et al., 1996). Countercurrent loading or periodic counter current chromatography is a way to render batch chromatography into continuous chromatography and in addition column utilization is improved. In its simplest form a column is overloaded, and the breakthrough loaded on the second column. Concurrently the third column is washed, eluted, and regenerated. This concept has been extensively used for purification of recombinant antibodies with protein A affinity chromatography(Badr et al., 2021; Davis et al., 2021; Farid et al., 2015; Gerstweiler et al., 2021; Gomis-Fons et al., 2020; Rathore et al., 2022b; Scheffel et al., 2022; Sun et al., 2022; Vetter et al., 2021; Vogg et al., 2018). In order to control the overloading of the first column a control algorithm using absorbance at 280 nm has been implemented. First a breakthrough of the impurities is observed and when the column is close saturation a second breakthrough is caused by the product. This signal can be used as a control algorithm for switching columns (Chmielowski et al., 2017; Gerstweiler et al., 2022; Godawat et al., 2012). This will only work when rather pure feedstock such as an antibody produced in chemical defined media is purified. Crude feedstocks would fully saturate the UV sensor and the difference between the signals of the impurities and product might be too low in order to get useful information to control the loading. In this case a soft sensor such as the Walch et. al. approach (Walch et al., 2019) will be a general solution to control loading in PCC. Near infrared spectroscopy (NIRS) has been also placed at the column entrance and exit and calibrated with a chemometrics approach (Thakur et al., 2020). It is not yet clear if this sensor alone would work for crude solutions or if only the concentration range is wider compared to a UV-cell. If only a wider concentration range must be covered flow VP is a good solution.

Another way to solve the problem of productivity and column utilization is to change the flow rate during loading. A dual flowrate during loading has been proposed by Lacki and coworkers (Ghose et al., 2004a; Ghose et al., 2004b) and then extend to either multiple steps (Ramakrishna et al., 2022) or a linear flow-rate gradient (Chen et al., 2021)or a model predicted gradient (Eslami et al., 2022a; Eslami et al., 2022b; Gomis-Fons et al., 2021; Sellberg et al., 2018)). The loading is started with a high velocity and then subsequently reduced. In order to establish an optimized system an optimizer and controller are necessary. Such a model predictive control results in high productivity and high column utilization but with less complex equipment. In order to run such a controlled process a soft sensor is required at the column in-let and outlet when crude feedstocks are purified and UV280 is not sufficient to monitor product concentration (Eslami et al., 2022a; Eslami et al., 2022b; Sellberg et al., 2018). Such a controller has been implemented in the XAMIris (Evon, Austria) (Figure 8).

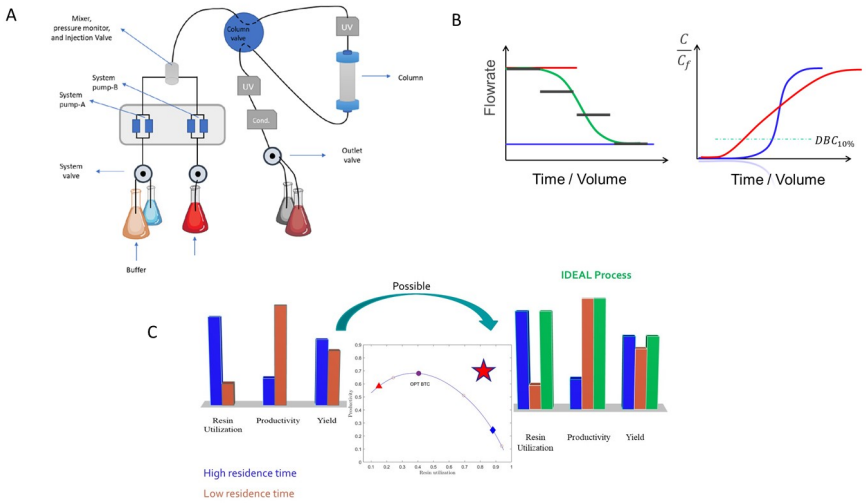


Figure 8: Principle of a flow rate gradient for optimization of productivity and resin utilization. (A) set-up of sensors in the chromatographic workstation (B) change of flow rate vs time or volume loaded onto the column and breakthrough curves at low and high residence time blue line high residence time, redline low residence time, (C) resin utilization productivity and yield of an unoptimized process and a process with model predictive control, data from (Eslami et al., 2022a; Eslami et al., 2022b).

The system can be also used for continuous chromatography. Then two columns are operated in tandem, while one column is loaded and the second column is washed, eluted and regenerated. If the wash, elution, and regeneration cycle is longer than loading then four columns must be installed. The selection of column number follows the same rules as for counter current chromatography. Furthermore, the velocity gradient approach can be extended to wash, elution and regeneration and thereby even higher productivity could be obtained. During loading in counter current chromatography two columns are interconnected and therefore the maximal system pressure and flowrate is constrained by this situation. This is not the case when columns are operated in parallel with velocity gradients during loading.

How fully integrated continuous biomanufacturing may look like was been previously proposed by the Morbidelli group (Feidl et al., 2020). They propose a process wide control of the integrated process. The unit operations are individually controlled and process wide by a supervisory system. The challenge is to integrate all unit operations and to maintain a constant mass flow and to avoid propagation of process errors or deviations. Therefore, always surge tanks between unit operations are placed to dampen fluctuations in flow and pressure and also might be used as intermedia storage when a unit operation stops working. Fluctuations flow rate and concentrations are inherent when a conventional batch manufacturing process is rendered into a continuous. Then preferably perfusion culture is connected to a counter current chromatography. Constants harvest conditions in perfusion culture could be only expected if be fully growth arrested. This is not the case and therefore the excess of cells must be removed from the bioreactor by a so-called bleeding leading to a change in harvest flow rate. The concentration is simultaneously changing. Therefore without proper control these fluctuations would propagate through the entire process and eventually amplify and lead to a derailed process. Model predictive control ensures either constant harvest flow or concentration (Pappenreiter et al., 2022).

Conventional feedback control is not sufficient, because it is designed for rapid changes in process conditions, but in perfusion cell culture conditions change slowly (Zhao et al., 2015) . Then harvest is sent to a surge thank and then captured by a counter current chromatography. The loading flow rate is constant but elution after the columns is saturated leads to a periodically change of mass flow and in order to warrant a constant contact time in the virus inactivation the elution fractions are pooled and then very often subjected to a

batch virus inactivation, although many continuous virus inactivation methods have been established. By following such a process design the benefits of continuous biomanufacturing are reduced in part automation and on-line control and real-time release become very difficult to implement. Surge tanks have a deteriorating effect RTD and wide RTD makes a process slow and batch definition difficult(Lali et al., 2022). In an ideal process the mass flow is never interrupted, and surge tanks are avoided. The perfusion bioreactor produces a constant harvest flow and can be further processed by flocculation and precipitation and polished with flow through chromatography methods. After concentration adjustment by e.g., single pass tangential flow filtration a continuous virus filtration can be added as recently suggested (Figure 9).

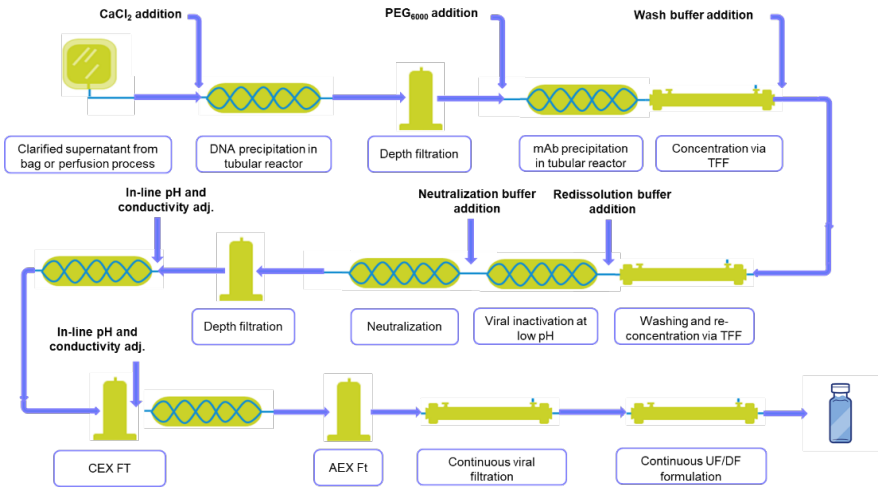


Figure 9: Concept of the ideal process based on precipitation, tow stage filtration and flow through chromatography, data from (Burgstaller et al., 2019; Li et al., 2019) drawing courtesy G. Recanati.

By combing existing elements of continuous biomanufacturing with real-time monitoring concepts it is possible develop a fully automated system (Helgers et al., 2022; Vetter et al., 2021) which runs under optimal conditions without large safety margins it will run autonomously, and a real-time release concept could be realized.

Conclusion

Although the biopharma industry is highly regulated, there are many opportunities to develop better and more robust processes. This is also encouraged by health authorities with the various guidelines that have been adopted in the past, such as PAT, QbD, Continuous Manufacturing and Real-time Release. All the elements for technical implementation are in place and there is acceptance by the authorities. There is a perception that real-time monitoring is difficult to implement and requires a lot of change. It is possible to integrate real-time monitoring gradually and it should be considered from a soft sensor perspective. This concept has been successfully implemented in other industries. It required relatively simple model training and the use of established statistical tools such as multivariate statistics or neural networks. Such approaches have been practiced for decades and are known as chemometrics when multiple analytical methods were required to measure a particular analytical attribute. A soft sensor could also be described with predictive chemometrics. A fully controlled and automated process would also help reduce drug shortages, improve process economics, and reduce the environmental footprint.

Acknowledgments

The COMET center: acib: Next Generation Bioproduction is funded by BMK, BMDW, SFG, Standortagentur Tirol, Government of Lower Austria und Vienna Business Agency in the framework of COMET - Competence Centers for Excellent Technologies. The COMET-Funding Program is managed by the Austrian Research Promotion Agency FFG.

References

Badr, S., Okamura, K., Takahashi, N., Ubbenjans, V., Shirahata, H., & Sugiyama, H. (2021). Integrated design of biopharmaceutical manufacturing processes: Operation modes and process configurations for monoclonal antibody production. *Computers and Chemical Engineering, 153* . doi:10.1016/j.compchemeng.2021.107422

Borras, E., Ferre, J., Boque, R., Mestres, M., Acena, L., & Busto, O. (2015). Data fusion methodologies for food and beverage authentication and quality assessment - A review. *Analytica Chimica Acta, 891* , 1-14. doi:10.1016/j.aca.2015.04.042

Breiman, L. (2001). Random forests. *Machine Learning, 45* (1), 5-32. doi:Doi 10.1023/A:1010933404324

Brestich, N., Rüdt, M., Büchler, D., & Hubbuch, J. (2018). Selective protein quantification for preparative chromatography using variable pathlength UV/Vis spectroscopy and partial least squares regression.*Chemical Engineering Science, 176* , 157-164. doi:10.1016/j.ces.2017.10.030

Burgstaller, D., Jungbauer, A., & Satzer, P. (2019). Continuous integrated antibody precipitation with two-stage tangential flow microfiltration enables constant mass flow. *Biotechnology and Bioengineering, 116* (5), 1053-1065. doi:10.1002/bit.26922

Capito, F., Skudas, R., Kolmar, H., & Stanislawski, B. (2013). Host cell protein quantification by fourier transform mid infrared spectroscopy (FT-MIR). *Biotechnology and Bioengineering, 110* (1), 252-259. doi:10.1002/bit.24611

Carta, G., & Jungbauer, A. (2010). *Protein Chromatography: Process Development and Scale-Up* .

Cataldo, A. L., Burgstaller, D., Hribar, G., Jungbauer, A., & Satzer, P. (2020). Economics and ecology: Modelling of continuous primary recovery and capture scenarios for recombinant antibody production.*Journal of Biotechnology, 308* , 87-95. doi:10.1016/j.jbiotec.2019.12.001

Chen, C. S., Ando, K., Yoshimoto, N., & Yamamoto, S. (2021). Linear flow-velocity gradient chromatography—An efficient method for increasing the process efficiency of batch and continuous capture chromatography of proteins. *Biotechnology and Bioengineering, 118* (3), 1262-1272. doi:10.1002/bit.27649

Chmielowski, R. A., Mathiasson, L., Blom, H., Go, D., Ehring, H., Khan, H., . . . Roush, D. (2017). Definition and dynamic control of a continuous chromatography process independent of cell culture titer and impurities. *Journal of Chromatography A, 1526* , 58-69. doi:10.1016/j.chroma.2017.10.030

Christler, A., Scharl, T., Sauer, D. G., Köppl, J., Tscheließnig, A., Toy, C., . . . Dürauer, A. (2021). Technology transfer of a monitoring system to predict product concentration and purity of biopharmaceuticals in real-time during chromatographic separation. *Biotechnology and Bioengineering, 118* (10), 3941-3952. doi:10.1002/bit.27870

Davis, R. R., Suber, F., Heller, I., Yang, B., & Martinez, J. (2021). Improving mAb capture productivity on batch and continuous downstream processing using nanofiber PrismA adsorbents. *Journal of Biotechnology, 336* , 50-55. doi:10.1016/j.jbiotec.2021.06.004

di Sciascio, F., & Amicarelli, A. N. (2008). Biomass estimation in batch biotechnological processes by Bayesian Gaussian process regression. *Computers & Chemical Engineering, 32* (12), 3264-3273. doi:https://doi.org/10.1016/j.compchemeng.2008.05.015

Ding, C., Ardeshna, H., Gillespie, C., & Ierapetritou, M. (2022). Process design of a fully integrated continuous biopharmaceutical process using economic and ecological impact assessment. *Biotechnology and Bioengineering, 119* (12), 3567-3583. doi:10.1002/bit.28234

Ender, L., & Maciel Filho, R. (2003) Neural networks applied to a multivariable nonlinear control strategies. In*: Vol. 15. Computer Aided Chemical Engineering* (pp. 190-195).

Erickson, C. B., Ankenman, B. E., & Sanchez, S. M. (2018). Comparison of Gaussian process modeling software. *European Journal of Operational Research, 266* (1), 179-192. doi:https://doi.org/10.1016/j.ejor.2017.10.002

Eslami, T., Jakob, L. A., Satzer, P., Ebner, G., Jungbauer, A., & Lingg, N. (2022a). Productivity for free: Residence time gradients during loading increase dynamic binding capacity and productivity. *Separation and Purification Technology, 281* . doi:10.1016/j.seppur.2021.119985

Eslami, T., Steinberger, M., Csizmazia, C., Jungbauer, A., & Lingg, N. (2022b). Online optimization of dynamic binding capacity and productivity by model predictive control. *Journal of Chromatography A, 1680* . doi:10.1016/j.chroma.2022.463420

Fahrmeir, L., Kneib, T., & Lang, S. (2004). Penalized structured additive regression for space-time data: A Bayesian perspective. *Statistica Sinica, 14* (3), 731-761.

Farid, S. S. (2019). Integrated Continuous Biomanufacturing: Industrialization on the Horizon. *Biotechnology Journal, 14* (2). doi:10.1002/biot.201800722

Farid, S. S., Pollock, J., & Ho, S. V. (2015). Evaluating the Economic and Operational Feasibility of Continuous Processes for Monoclonal Antibodies. In *Continuous Processing in Pharmaceutical Manufacturing* (pp. 433-456).

Feidl, F., Garbellini, S., Vogg, S., Sokolov, M., Souquet, J., Broly, H., . . . Morbidelli, M. (2019). A new flow cell and chemometric protocol for implementing in-line Raman spectroscopy in chromatography. *Biotechnology Progress, 35* (5). doi:10.1002/btpr.2847

Feidl, F., Vogg, S., Wolf, M., Podobnik, M., Ruggeri, C., Ulmer, N., . . . Morbidelli, M. (2020). Process-wide control and automation of an integrated continuous manufacturing platform for antibodies. *Biotechnology and Bioengineering, 117* (5), 1367-1380. doi:10.1002/bit.27296

Felfödi, E., Scharl, T., Melcher, M., Dürauer, A., Wright, K., & Jungbauer, A. (2020). Osmolality is a predictor for model-based real time monitoring of concentration in protein chromatography. *Journal of Chemical Technology and Biotechnology, 95* (4), 1146-1152. doi:10.1002/jctb.6299

Gareth, J., Witten, D., Hastie, T., & Tibishirani, R. (2021). *Elements of Statistical Learning* .

Gerstweiler, L., Bi, J., & Middelberg, A. P. J. (2021). Continuous downstream bioprocessing for intensified manufacture of biopharmaceuticals and antibodies. *Chemical Engineering Science, 231* . doi:10.1016/j.ces.2020.116272

Gerstweiler, L., Billakanti, J., Bi, J., & Middelberg, A. P. J. (2022). Control strategy for multi-column continuous periodic counter current chromatography subject to fluctuating inlet stream concentration. *Journal of Chromatography A, 1667* . doi:10.1016/j.chroma.2022.462884

Ghisaidoobe, A. B. T., & Chung, S. J. (2014). Intrinsic Tryptophan Fluorescence in the Detection and Analysis of Proteins: A Focus on Forster Resonance Energy Transfer Techniques. *International Journal of Molecular Sciences, 15* (12), 22518-22538. doi:10.3390/ijms151222518

Ghose, S., Nagarath, D., Hubbard, B., Brooks, C., & Cramer, S. M. (2004a). Erratum: Use and optimization of a dual-flowrate loading strategy to maximize throughput in protein-A affinity chromatography (Biotechnology Progress (2004) 20 (830-840)). *Biotechnology Progress, 20* (5), 1614. doi:10.1021/bp040029x

Ghose, S., Nagrath, D., Hubbard, B., Brooks, C., & Cramer, S. M. (2004b). Use and optimization of a dual-flowrate loading strategy to maximize throughput in protein-A affinity chromatography.*Biotechnology Progress, 20* (3), 830-840. doi:10.1021/bp0342654

Glassey, J., Ignova, M., Ward, A. C., Montague, G. A., & Morris, A. J. (1997). Bioprocess supervision: Neural networks and knowledge based systems. *Journal of Biotechnology, 52* (3), 201-205. doi:Doi 10.1016/S0168-1656(96)01645-8

Godawat, R., Brower, K., Jain, S., Konstantinov, K., Riske, F., & Warikoo, V. (2012). Periodic counter-current chromatography - design and operational considerations for integrated and continuous purification of proteins. *Biotechnology Journal, 7* (12), 1496-1508. doi:10.1002/biot.201200068

Gomis-Fons, J., Andersson, N., & Nilsson, B. (2020). Optimization study on periodic counter-current chromatography integrated in a monoclonal antibody downstream process. *Journal of Chromatography A, 1621* . doi:10.1016/j.chroma.2020.461055

Gomis-Fons, J., Yamanee-Nolin, M., Andersson, N., & Nilsson, B. (2021). Optimal loading flow rate trajectory in monoclonal antibody capture chromatography. *Journal of Chromatography A, 1635* , 461760. doi:https://doi.org/10.1016/j.chroma.2020.461760

Guidance for Industry: PAT-A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance. (2004). *US Department of Health and Human Services Food and Drug Administration* .

Heeter, G. A., & Liapis, A. I. (1996). Multi-component perfusion chromatography in fixed bed and periodic counter current column operation. *Journal of Chromatography A, 734* (1), 105-123. doi:10.1016/0021-9673(95)01147-1

Helgers, H., Schmidt, A., & Strube, J. (2022). Towards Autonomous Process Control—Digital Twin for CHO Cell-Based Antibody Manufacturing Using a Dynamic Metabolic Model. *Processes, 10* (2). doi:10.3390/pr10020316

Hofner, B., Boccuto, L., & Göker, M. (2015). Controlling false discoveries in high-dimensional situations: Boosting with stability selection. *BMC Bioinformatics, 16* (1). doi:10.1186/s12859-015-0575-3

Hofner, B., Hothorn, T., Kneib, T., & Schmid, M. (2011). A framework for unbiased model selection based on boosting. *Journal of Computational and Graphical Statistics, 20* (4), 956-971. doi:10.1198/jcgs.2011.09220

Hong, M. S., Severson, K. A., Jiang, M., Lu, A. E., Love, J. C., & Braatz, R. D. (2018). Challenges and opportunities in biopharmaceutical manufacturing control. *Computers and Chemical Engineering, 110* , 106-114. doi:10.1016/j.compchemeng.2017.12.007

Hutter, C., von Stosch, M., Bournazou, M. N. C., & Butté, A. (2020). Knowledge Transfer Across Cell Lines Using Hybrid Gaussian Process Models With Entity Embedding Vectors. *Knowledge transfer across cell lines using Hybrid Gaussian Process models with entity embedding vectors* .

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. *112* .

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021a). Classification. *An Introduction to Statistical Learning: With Applications in R* , 129-195.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021b). An Introduction to Statistical Learning: with Applications in R.

Konstantinov, K. B., & Cooney, C. L. (2015). White paper on continuous bioprocessing May 20-21, 2014 continuous manufacturing symposium.*Journal of Pharmaceutical Sciences, 104* (3), 813-820. doi:10.1002/jps.24268

Kornecki, M., & Strube, J. (2018). Process analytical technology for advanced process control in biologics manufacturing with the aid of macroscopic kinetic modeling. *Bioengineering, 5* (1).

doi:10.3390/bioengineering5010025

Krippl, M., Dürauer, A., & Duerkop, M. (2020). Hybrid modeling of cross-flow filtration: Predicting the flux evolution and duration of ultrafiltration processes. *Separation and Purification Technology, 248* . doi:10.1016/j.seppur.2020.117064

Krippl, M., Kargl, T., Duerkop, M., & Dürauer, A. (2021). Hybrid modeling reduces experimental effort to predict performance of serial and parallel single-pass tangential flow filtration. *Separation and Purification Technology, 276* . doi:10.1016/j.seppur.2021.119277

Kumar, A., Udugama, I. A., Gargalo, C. L., & Gernaey, K. V. (2020). Why is batch processing still dominating the biologics landscape? Towards an integrated continuous bioprocessing alternative. *Processes, 8* (12), 1-19. doi:10.3390/pr8121641

Lali, N., Jungbauer, A., & Satzer, P. (2022). Traceability of products and guide for batch definition in integrated continuous biomanufacturing. *Journal of Chemical Technology and Biotechnology, 97* (9), 2386-2392. doi:10.1002/jctb.6953

Lecun, Y., Jackel, L. D., Boser, B., Denker, J. S., Graf, H. P., Guyon, I., . . . Hubbard, W. (1989). Handwritten Digit Recognition - Applications of Neural Network Chips and Automatic Learning. *Ieee Communications Magazine, 27* (11), 41-46. doi:Doi 10.1109/35.41400

Lee, C., Lee, J., & Lee, A. (2000). Statistics for Business and Financial Economics, Vol. 1. . *World Scientific 2000* .

Lee, J., Shin, J., & Realff, M. J. (2018). Machine learning: Overview of the recent progresses and implications for the process systems engineering field. *Computers & Chemical Engineering, 114* , 111-121. doi:10.1016/j.compchemeng.2017.10.008

Li, X. Y., Luan, F., Si, H. Z., Hu, Z., & Liu, M. C. (2007). Prediction of retention times for a large set of pesticides or toxicants based on support vector machine and the heuristic method. *Toxicology Letters, 175* (1-3), 136-144. doi:10.1016/j.toxlet.2007.10.005

Li, Z., Gu, Q., Coffman, J. L., Przybycien, T., & Zydney, A. L. (2019). Continuous precipitation for monoclonal antibody capture using countercurrent washing by microfiltration. *Biotechnology Progress, 35* (6). doi:10.1002/btpr.2886

Liggins, M., Hall, D., & Llinas, J. (2017). *Handbook of Multisensor Data Fusion: Theory and Practice* .

Lopes, J. A., Costa, P. F., Alves, T. P., & Menezes, J. C. (2004). Chemometrics in bioprocess engineering: Process analytical technology (PAT) applications. *Chemometrics and Intelligent Laboratory Systems, 74* (2 SPEC.ISS.), 269-275. doi:10.1016/j.chemolab.2004.07.006

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* (8), 1040-1048. doi:10.1002/biot.201100506

Mandenius, C. F., & Gustavsson, R. (2015). Mini-review: Soft sensors as means for PAT in the manufacture of bio-therapeutics. *Journal of Chemical Technology and Biotechnology, 90* (2), 215-227. doi:10.1002/jctb.4477

Melcher, M., Scharl, T., Spangl, B., Luchner, M., Cserjan, M., Bayer, K., . . . Striedner, G. (2015). The potential of random forest and neural networks for biomass and recombinant protein modeling in Escherichia coli fed-batch fermentations. *Biotechnology Journal, 10* (11), 1770-1782. doi:10.1002/biot.201400790

Minton, A. P. (2016). Recent applications of light scattering measurement in the biological and biopharmaceutical sciences.*Analytical Biochemistry, 501* , 4-22. doi:10.1016/j.ab.2016.02.007

Müller, D., Klein, L., Lemke, J., Schulze, M., Kruse, T., Saballus, M., . . . Zijlstra, G. (2022). Process intensification in the biopharma industry: Improving efficiency of protein manufacturing processes from

development to production scale using synergistic approaches. *Chemical Engineering and Processing - Process Intensification, 171* . doi:10.1016/j.cep.2021.108727

Narayanan, H., Seidler, T., Luna, M. F., Sokolov, M., Morbidelli, M., & Butté, A. (2021). Hybrid Models for the simulation and prediction of chromatographic processes for protein capture. *Journal of Chromatography A, 1650* . doi:10.1016/j.chroma.2021.462248

Nikita, S., Thakur, G., Jesubalan, N. G., Kulkarni, A., Yezhuvath, V. B., & Rathore, A. S. (2022). AI-ML applications in bioprocessing: ML as an enabler of real time quality prediction in continuous manufacturing of mAbs. *Computers and Chemical Engineering, 164* . doi:10.1016/j.compchemeng.2022.107896

Oliveira, A. L. (2019). Biotechnology, Big Data and Artificial Intelligence. *Biotechnology Journal, 14* (8). doi:10.1002/biot.201800613

Pappenreiter, M., Döbele, S., Striedner, G., Jungbauer, A., & Sissolak, B. (2022). Model predictive control for steady-state performance in integrated continuous bioprocesses. *Bioprocess and Biosystems Engineering, 45* (9), 1499-1513. doi:10.1007/s00449-022-02759-z

Patel, B. A., Gospodarek, A., Larkin, M., Kenrick, S. A., Haverick, M. A., Tugcu, N., . . . Richardson, D. D. (2018). Multi-angle light scattering as a process analytical technology measuring real-time molecular weight for downstream process control. *Mabs, 10* (7), 945-950. doi:10.1080/19420862.2018.1505178

Patil, R., & Walther, J. (2018) Continuous manufacturing of recombinant therapeutic proteins: Upstream and downstream technologies. In*: Vol. 165. Advances in Biochemical Engineering/Biotechnology* (pp. 277-322).

Ramakrishna, A., Maranholkar, V., Hadpe, S., Iyer, J., & Rathore, A. (2022). Optimization of multi flow rate loading strategy for process intensification of Protein A chromatography. *Journal of Chromatography Open, 2* , 100049. doi:https://doi.org/10.1016/j.jcoa.2022.100049

Rasmussen, C. E., & Gaussian, W. C. (2021). *Processes for Machine Learning* .

Rathore, A., Li, X. H., Bartkowski, W., Sharma, A., & Lu, Y. F. (2009). Case Study and Application of Process Analytical Technology (PAT) towards Bioprocessing: Use of Tryptophan Fluorescence as At-line Tool for Making Pooling Decisions for Process Chromatography. *Biotechnology Progress, 25* (5), 1433-1439. doi:10.1002/btpr.212

Rathore, A., Nikita, S., & Jesubalan, N. G. (2022a). Digitization in bioprocessing: The role of soft sensors in monitoring and control of downstream processing for production of biotherapeutic products. *Biosensors and Bioelectronics: X, 12* . doi:10.1016/j.biosx.2022.100263

Rathore, A., & Shereef, F. (2022b). Innovating manufacturing technology in emerging economies. *Nature Biotechnology, 40* (12), 1714-1716. doi:10.1038/s41587-022-01499-5

Roch, P., & Mandenius, C. F. (2016). On-line monitoring of downstream bioprocesses. *Current Opinion in Chemical Engineering, 14* , 112-120. doi:10.1016/j.coche.2016.09.007

Rolinger, L., Rüdt, M., & Hubbuch, J. (2021). Comparison of UV- and Raman-based monitoring of the Protein A load phase and evaluation of data fusion by PLS models and CNNs. *Biotechnology and Bioengineering, 118* (11), 4255-4268. doi:10.1002/bit.27894

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access, 8* , 42200-42216. doi:10.1109/ACCESS.2020.2976199

Rüdt, M., Briskot, T., & Hubbuch, J. (2017). Advances in downstream processing of biologics – Spectroscopy: An emerging process analytical technology. *Journal of Chromatography A, 1490* , 2-9. doi:10.1016/j.chroma.2016.11.010

Rudt, M., Vormittag, P., Hillebrandt, N., & Hubbuch, J. (2019). Process monitoring of virus-like particle reassembly by diafiltration with UV/Vis spectroscopy and light scattering. *Biotechnology and Bioengineering,*

116 (6), 1366-1379. doi:10.1002/bit.26935

Sauer, D. G., Melcher, M., Mosor, M., Walch, N., Berkemeyer, M., Scharl-Hirsch, T., . . . Dürauer, A. (2019). Real-time monitoring and model-based prediction of purity and quantity during a chromatographic capture of fibroblast growth factor 2. *Biotechnology and Bioengineering, 116* (8), 1999-2009. doi:10.1002/bit.26984

Scheffel, J., Isaksson, M., Gomis-Fons, J., Schwarz, H., Andersson, N., Norén, B., . . . Nilsson, B. (2022). Design of an integrated continuous downstream process for acid-sensitive monoclonal antibodies based on a calcium-dependent Protein A ligand. *Journal of Chromatography A, 1664* . doi:10.1016/j.chroma.2022.462806

Sellberg, A., Holmqvist, A., Magnusson, F., Andersson, C., & Nilsson, B. (2017). Discretized multi-level elution trajectory: A proof-of-concept demonstration. *Journal of Chromatography A, 1481* , 73-81. doi:10.1016/j.chroma.2016.12.038

Sellberg, A., Nolin, M., Löfgren, A., Andersson, N., & Nilsson, B. (2018) Multi-flowrate Optimization of the Loading Phase of a Preparative Chromatographic Separation. In*: Vol. 43. Computer Aided Chemical Engineering* (pp. 1619-1624).

Simon, L. L., Pataki, H., Marosi, G., Meemken, F., Hungerbühler, K., Baiker, A., . . . Chiu, M. S. (2015a). Assessment of recent process analytical technology (PAT) trends: A multiauthor review. *Organic Process Research and Development, 19* (1), 3-62. doi:10.1021/op500261y

Simon, L. L., Pataki, H., Marosi, G., Meemken, F., Hungerbuhler, K., Baiker, A., . . . Chiu, M. S. (2015b). Assessment of Recent Process Analytical Technology (PAT) Trends: A Multiauthor Review. *Organic Process Research & Development, 19* (1), 3-62. doi:10.1021/op500261y

Steinwandter, V., Borchert, D., & Herwig, C. (2019). Data science tools and applications on the way to Pharma 4.0. *Drug Discovery Today, 24* (9), 1795-1805. doi:10.1016/j.drudis.2019.06.005

Sun, Y. N., Shi, C., Zhong, X. Z., Chen, X. J., Chen, R., Zhang, Q. L., . . . Lin, D. Q. (2022). Model-based evaluation and model-free strategy for process development of three-column periodic counter-current chromatography. *Journal of Chromatography A, 1677* . doi:10.1016/j.chroma.2022.463311

Thakur, G., Hebbi, V., & Rathore, A. S. (2020). An NIR-based PAT approach for real-time control of loading in Protein A chromatography in continuous manufacturing of monoclonal antibodies. *Biotechnology and Bioengineering, 117* (3), 673-686. doi:10.1002/bit.27236

Tharmalingam, T., Wu, C. H., Callahan, S., & Goudar, C. T. (2015). A framework for real-time glycosylation monitoring (RT-GM) in mammalian cell culture. *Biotechnology and Bioengineering, 112* (6), 1146-1154. doi:10.1002/bit.25520

Varmuza, K., & Filzmoser, P. (2009). Introduction to multivariate analysis in chemometrics. *CRC Press* .

Vetter, F. L., Zobel-Roos, S., & Strube, J. (2021). Pat for continuous chromatography integrated into continuous manufacturing of biologics towards autonomous operation. *Processes, 9* (3). doi:10.3390/pr9030472

Vogg, S., Wolf, M. K. F., & Morbidelli, M. (2018) Continuous and integrated expression and purification of recombinant antibodies. In*: Vol. 1850. Methods in Molecular Biology* (pp. 147-178).

Walch, N., Scharl, T., Felföldi, E., Sauer, D. G., Melcher, M., Leisch, F., . . . Jungbauer, A. (2019). Prediction of the Quantity and Purity of an Antibody Capture Process in Real Time. *Biotechnology Journal, 14* (7). doi:10.1002/biot.201800521

Wasalathanthri, D. P., Rehmann, M. S., Song, Y., Gu, Y., Mi, L., Shao, C., . . . Li, Z. J. (2020). Technology outlook for real-time quality attribute and process parameter monitoring in biopharmaceutical development—A review. *Biotechnology and Bioengineering, 117* (10), 3182-3198. doi:10.1002/bit.27461

Westad, F., & Marini, F. (2015). Validation of chemometric models - A tutorial. *Analytica Chimica Acta, 893* , 14-24. doi:10.1016/j.aca.2015.06.056

Winter, P., Weissenböck, s., Schwald, C., Doms, T., Vogt, T., Hochreiter, S., & Nessler, B. (2021). Trusted Artificial Intelligence: Towards Certification of Machine Learning Applications. *Vienna, March 17th, 2021, TÜV AUSTRIA Group, Johannes Kepler University Linz – Institute for Machine Learning* .

Wold, S., Trygg, J., Berglund, A., & Antti, H. (2001). Some recent developments in PLS modeling. *Chemometrics and Intelligent Laboratory Systems, 58* (2), 131-150. doi:10.1016/S0169-7439(01)00156-3

Yu, Z., Reid, J. C., & Yang, Y. P. (2013). Utilizing Dynamic Light Scattering as a Process Analytical Technology for Protein Folding and Aggregation Monitoring in Vaccine Manufacturing. *Journal of Pharmaceutical Sciences, 102* (12), 4284-4290. doi:10.1002/jps.23746

Zhao, H. Y., Brown, P. H., & Schuckt, P. (2011). On the Distribution of Protein Refractive Index Increments. *Biophysical Journal, 100* (9), 2309-2317. doi:10.1016/j.bpj.2011.03.004

Zhao, L., Fu, H. Y., Zhou, W., & Hu, W. S. (2015). Advances in process monitoring tools for cell culture bioprocesses. *Engineering in Life Sciences, 15* (5), 459-468. doi:10.1002/elsc.201500006

Zhong, X., Gallagher, B., Liu, S., Kailkhura, B., Hiszpanski, A., & Han, T. Y. J. (2022). Explainable machine learning in materials science.*npj Computational Materials, 8* (1). doi:10.1038/s41524-022-00884-7