



ARTICLE

Real-time monitoring and model-based prediction of purity and quantity during a chromatographic capture of fibroblast growth factor 2

Dominik Georg Sauer¹ | Michael Melcher^{1,2} | Magdalena Mosor¹ | Nicole Walch³ |
Matthias Berkemeyer⁴ | Theresa Scharl-Hirsch^{1,2} | Friedrich Leisch^{1,2} |
Alois Jungbauer^{1,5}  | Astrid Dürauer^{1,5} 

¹Austrian Centre of Industrial Biotechnology, Vienna, Austria

²Institute of Applied Statistics and Computing, University of Natural Resources and Life Sciences Vienna, Vienna, Austria

³Biopharmaceuticals Operations Austria, Manufacturing Science, Boehringer Ingelheim Regional Center Vienna GmbH & Co KG, Vienna, Austria

⁴Biopharma Process Science Austria, Boehringer Ingelheim Regional Center Vienna GmbH & Co KG, Vienna, Austria

⁵Department of Biotechnology, University of Natural Resources and Life Sciences Vienna, Vienna, Austria

Correspondence

Astrid Dürauer, Department of Biotechnology, University of Natural Resources and Life Sciences Vienna, Muthgasse 18, 1190 Vienna, Austria.

Email: astrid.duerauer@boku.ac.at

Funding information

Österreichische

Forschungsförderungsgesellschaft, Grant/Award Number: 824186; the Austrian Federal Ministry for Digital and Economic Affairs (bmwd); the Federal Ministry for Transport, Innovation and Technology (bmvit); the Styrian Business Promotion Agency (SFG); the Standortagentur Tirol; the Government of Lower Austria and ZIT - Technology Agency of the City of Vienna through the COMET-Funding Program

Abstract

Process analytical technology combines understanding and control of the process with real-time monitoring of critical quality and performance attributes. The goal is to ensure the quality of the final product. Currently, chromatographic processes in biopharmaceutical production are predominantly monitored with UV/Vis absorbance and a direct correlation with purity and quantity is limited. In this study, a chromatographic workstation was equipped with additional online sensors, such as multi-angle light scattering, refractive index, attenuated total reflection Fourier-transform infrared, and fluorescence spectroscopy. Models to predict quantity, host cell proteins (HCP), and double-stranded DNA (dsDNA) content simultaneously were developed and exemplified by a cation exchange capture step for fibroblast growth factor 2 expressed in *Escherichia coli*. Online data and corresponding offline data for product quantity and co-eluting impurities, such as dsDNA and HCP, were analyzed using boosted structured additive regression. Different sensor combinations were used to achieve the best prediction performance for each quality attribute. Quantity can be adequately predicted by applying a small predictor set of the typical chromatographic workstation sensor signals with a test error of 0.85 mg/ml (range in training data: 0.1–28 mg/ml). For HCP and dsDNA additional fluorescence and/or attenuated total reflection Fourier-transform infrared spectral information was important to achieve prediction errors of 200 (2–6579 ppm) and 340 ppm (8–3773 ppm), respectively.

KEYWORDS

ATR-FTIR, dsDNA, fluorescence, HCP, MALS, online sensors

Abbreviations: ATR-FTIR, attenuated total reflection Fourier-transform infrared; dsDNA, double-stranded DNA; FGF-2, fibroblast growth factor 2; FDA, food and drug administration; HCP, host cell protein; MRD, mean relative deviation; MALS, multi-angle light scattering; RI, refractive index; RP-HPLC, reverse phase HPLC; RMSE, root-mean-square error; STAR, structured additive regression.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2019 The Authors. *Biotechnology and Bioengineering* Published by Wiley Periodicals, Inc.

1 | INTRODUCTION

Real-time monitoring and model-based prediction of purity and quantity are key steps towards real-time release in the manufacturing of biopharmaceuticals (Jiang et al., 2017). According to the ICH guidelines (Q8 R2), real-time release testing is defined as “the ability to evaluate and ensure the quality of an in-process and/or final drug product based on process data, which typically includes a valid combination of measured material attributes and process controls” (Holm, Allesø, Bryder, & Holm, 2017). There is not a single online sensor available that allows a direct measurement of the quality of a biopharmaceutical. Therefore, the information must be extracted from multiple sensor signals. In addition, model-based prediction of purity and quantity can be used for process control, in particular for real-time decisions on pooling. Currently, the determination of the critical quality attributes (CQAs) of biopharmaceutical products during processing is performed offline after the unit operation has been completed. This requires extensive sample preparation, is time-consuming, and adds to the manufacturing costs. The acquired information is available with a substantial time delay. Such a Quality-by-Testing approach using offline analysis is a retrospective quality control and is not suitable for continuous manufacturing (Food & Drug Administration, 2004; Löfgren et al., 2017). In contrast, the fundamental concept of the Quality-by-Design (QbD) approach is that a process is controlled to perform in defined design space and thus guarantees a continuous quality output (Rathore, 2016; Read et al., 2010; Scott & Wilcock, 2006; Yu et al., 2014). Process analytical technology (PAT), as one strategy of the QbD approach, includes the tasks of designing, analyzing and controlling production processes based on real-time monitoring of quality attributes to allow continuous manufacturing and enhanced flexibility. Online monitoring by employing fast and noninvasive mostly spectroscopic technologies collect real-time data of the process which have to be converted into relevant process information by appropriate statistical models. Therefore, online monitoring and model predictive control are mandatory for the realization of a PAT approach and to significantly reduce the need for offline analyses. Besides process control, PAT can be applied efficiently to increase the process understanding during development (Rathore, 2016) and for prospective real-time release (Jiang et al., 2017). So far, QbD and PAT have limited applications in biomanufacturing. Several studies for real-time monitoring of upstream processes were performed where the most important criteria to be controlled are the product formation and feed strategies and corresponding offline methods (e.g., cell-based assays) can last up to days (Dabros, Amrhein, Bonvin, Marison, & von Stockar, 2009; Luchner et al., 2012; Melcher et al., 2015; Melcher, Scharl, Luchner, Striedner, & Leisch, 2017; Pais, Carrondo, Alves, & Teixeira, 2014; von Stosch, Hamelink, & Oliveira, 2016). Chromatography is the main purification method for biopharmaceutical proteins. Process-related impurities such as host cell proteins (HCP), DNA, endotoxins, and product-related impurities (e.g., product variants or aggregates) have to be depleted to deliver a product that meets defined quality standards (Food & Drug Administration, 2003). Controlled loading and pooling of the eluates

are critical for the overall chromatography performance to ensure consistent product quality (Borg et al., 2014). The knowledge of the eluate composition is crucial for subsequent downstream steps.

Currently, process decisions are based mainly on online monitoring of UV absorbance, which is beneficial for the estimation of overall protein content or the protein/DNA ratio, but is rather unspecific for typical co-eluting critical impurities. It is challenging when HCP and product variants possess physicochemical properties similar to the target protein. For reliable control of preparative chromatographic processes, monitoring methods have to discriminate the product from relevant impurities within short response times (i.e. seconds), as quality attributes of the effluent are changing quickly. Another challenge is the complex matrix of buffers and product samples that change throughout the purification steps. Single sensor methods have been examined for their feasibility in PAT for pooling of chromatographic methods (Großhans et al., 2018; Rathore, Li, Bartkowski, Sharma, & Lu, 2009; Rathore, Wood, Sharma, & Dermawan, 2008; Rathore, Yu, Yeboah, & Sharma, 2008; Read et al., 2010; Rüdts, Briskot, & Hubbuch, 2017). Most commonly, UV/Vis absorption and attenuated total reflection Fourier-transform infrared (ATR-FTIR) spectroscopy are used for PAT applications (Brestrich, Briskot, Osberghaus, & Hubbuch, 2014; Brestrich et al., 2015; Großhans et al., 2018; Rathore et al., 2008). In all these cases, only a single quality attribute was monitored and used as a pooling criterion. Spectroscopic methods have been widely used as tools, as they deliver information on the primary, secondary, tertiary, and quaternary structures of proteins. These techniques are useful as noninvasive, nondestructive, rapid, sensitive, and automatable methods with the ability to provide information simultaneously on different proteins, conformational variations, or DNA (Brestrich et al., 2014; Flatman, Alam, Gerard, & Mussa, 2007; Rüdts et al., 2017; Workman, Koch, & Veltkamp, 2007). UV/Vis spectroscopy mainly measures the primary structure, such as the content of aromatic amino acids ($UV_{280\text{ nm}}$) or polypeptide backbone ($UV_{214\text{ nm}}$). The $UV_{260\text{ nm}}$ absorbance provides an estimation of DNA content (Antosiewicz & Shugar, 2016). The refractive index (RI) has also been applied for protein quantification (Zhao, Brown, & Schuck, 2011). The secondary structure can be measured by vibrational spectroscopy such as FTIR, circular dichroism, and Raman spectroscopy (Flatman et al., 2007; Rüdts, Briskot et al., 2017; Workman et al., 2007). At-line ATR-FTIR can distinguish between HCP and target protein (Capito, Skudas, Kolmar, & Stanislawski, 2013). The benefit of ATR is the lack of complex sample preparation, as part of the totally reflected infrared beam in the ATR crystal enters the sample interface. In the spectral regions where the sample absorbs energy, the beam is attenuated (Barth, 2007). The tertiary structure of proteins can be measured via intrinsic fluorescence of the aromatic amino acids and structural changes induced by polarity changes can be detected (Ghisaidoobe & Chung, 2014; Rathore et al., 2009). Their quaternary structure, for example, protein aggregation, can be determined by light scattering methods (Lorber, Fischer, Bailly, Roy, & Kern, 2012; Minton, 2016). Fluorescence spectroscopy, as well as light scattering techniques, have been used for at-line determination of quality attributes (Rathore et al., 2009; Yu, Reid, & Yang, 2013). All those spectroscopic data are complex with limited first principle

knowledge. Online ion-exchange liquid chromatography has been applied to monitor antibody variants in a continuous process, however one sample measurement lasts 15 min (Patel et al., 2017).

State of the art statistical methods to relate many sensor signals to offline analyses include Partial Least Squares (PLS), tree-based methods (e.g., Random Forests) or boosted structured additive regression (STAR), which extends the well-known multiple linear regression models with interaction effects or nonlinear spline functions (Bühlmann & Hothorn, 2007; Hothorn, Bühlmann, Kneib, Schmid & Hofner, 2015).

No studies are available using a combination of sensors that would allow the simultaneous monitoring and prediction of many product quality attributes in parallel (Borg et al., 2014; Großhans et al., 2018; Rathore et al., 2009; Rüdts, Brestrich, Rolinger, & Hubbuch, 2017; Rüdts, Briskot et al., 2017; von Stosch et al., 2016; Yu et al., 2013).

The aim of the present study was a comprehensive and efficient real-time monitoring of a protein purification step using a panel of online sensors. UV/Vis, pH and conductivity probes are standard sensors of a chromatographic workstation. Multi-angle light scattering (MALS) and RI detectors, ATR-FTIR, and fluorescence spectroscopic sensors were additionally integrated into a commercially available chromatographic workstation for model-based prediction of product quantity and impurity content. As HCP and dsDNA are critical host cell impurities they were addressed in this study. Data analysis was performed with boosted STAR, which gives promising results in settings, where the number of variables (greatly) exceeds the number of observations, as it is common, when spectroscopic sensor systems are involved (Melcher et al., 2017).

2 | MATERIALS AND METHODS

2.1 | Fibroblast growth factor 2

Fibroblast growth factor 2 (FGF-2) was overexpressed in *Escherichia coli* (*E. coli*) BL21 cells (Sauer et al., 2018). FGF-2 has a molecular weight of 17 kDa and an isoelectric point pI of 9.6 (Gasparian et al., 2009).

2.2 | Chromatographic capture step by ion exchange

A cation exchange resin was used to purify FGF-2 from the clarified *E. coli* homogenate. A Tricorn column ($d = 10$ mm; $h = 150$ mm; GE Healthcare; Uppsala, Sweden) was packed with Carboxymethyl Sepharose Fast Flow resin (GE Healthcare; Uppsala, Sweden) resulting in a column volume (CV) of 11.8 ml. The FGF-2 capture method was conducted at a flow rate of 1 ml/min. 10 CV of clarified *E. coli* homogenate (1.7 ± 0.2 mg/ml) were loaded and elution was performed with a linear gradient from 0 to 1 M NaCl in 100 mM Na-phosphate (pH 7.0). For each performed chromatographic run, 15 fractions (UV_{280 nm} signal >50 mAU) of 1 ml were collected and used for all offline assays (Sauer et al., 2018). Clarified homogenate compositions of fermentation batches are provided in Table S1.

2.3 | Offline monitoring

Product quantity (mg/ml) was determined by reverse-phase high-performance liquid chromatography (RP-HPLC). Based on the FGF-2 concentration, the relative dsDNA content (ppm) was quantified by PicoGreen[®] assay and the HCP content (ppm) by enzyme-linked immunosorbent assay (ELISA). The intra-assay variabilities were 20% for HCP ELISA, 15% for the PicoGreen[®] assay, and 5% for the RP-HPLC quantification (Sauer et al., 2018).

2.4 | Online monitoring

The chromatographic workstation Äkta Pure 25 (GE Healthcare, Uppsala, Sweden) comprises a multi-wavelength UV/Vis detector (UV_{214 nm}, UV_{260 nm}, and UV_{280 nm}), a conductivity and a pH probe and therefore contributes five variables/predictors in the subsequent prediction models. The mid infrared spectrometer MATRIX-FM (Bruker; Billerica) based on ATR was chosen to measure FTIR spectra from 3500 to 750 cm⁻¹ with a resolution of 2 cm⁻¹ (resulting in 1427 channels/predictors). 16 ATR-FTIR scans were performed per spectrum and averaged within 4 s. For fluorescence detection, the setup consisted of a laser-induced xenon lamp (EQ-99XFC LDLS; Energetiq; Woburn), a fiber optic multiplexer (Avantes, Apeldoorn, The Netherlands), a flow cell (FIALab Instruments; Seattle, USA), and the spectrometer (AvaSpec-ULS-TEC with 600 L/mm grating; Avantes; Apeldoorn, The Netherlands). This fluorescence sensor enabled excitation with seven different wavelengths (265 ± 10 nm, 280 ± 10 nm, 289 ± 10 nm, 300 ± 10 nm, 300 ± 40 nm, 340 ± 10 nm, 400 ± 10 nm). For each excitation wavelength, emission spectra were detected over a range of 236–795 nm at a resolution of 0.3 nm and an integration time of 1 s per excitation wavelength (giving 3215 channels/predictors after data reduction to the spectral region of interest containing emission bands). The measurement of all seven emission spectra took 16 s, including multiplexer switching time. A differential RI detector was used (Optilab T-rEX; Wyatt; Santa Barbara), with a differential RI in the range of -0.0047 to $+0.0047$ RIU. The RI also traced an additional forward monitor and LED monitor intensities (three predictors). The MALS detector (miniDAWN TREOS; Wyatt; Santa Barbara) recorded light scattering signals from three integrated detectors at angles of 43.6° (LS1), 90° (LS2), and 136.4° (LS3) plus forward monitor intensity (four predictors). All buffers applied were aqueous based, therefore water was used as the common blank. All sensors were integrated in the liquid stream after the chromatography column in the order of increasing flow cell void volume.

2.5 | Data preprocessing and statistical modeling

On- and offline data were collected for 19 chromatographic runs, 13 of which were used for model building (training runs) and the remaining six served as test runs. All runs were performed under identical experimental conditions and are expected to differ only by random, biological variation. The test runs originate from two

different fermentation batches and therefore represent a typical field of application of the derived models. While the test runs are complete data sets, there are some missing sensor data in the training set.

The major preprocessing operations consisted of two time-alignment steps:

- (1) The correction of the time shift in the data between the online sensors resulting from void volumes in the setup. The delay volume between the first (UV) and last sensor (RI) was 1.39 ml and between the last sensor and the outlet valve 0.37 ml.
- (2) Offline data were available for 1 ml fractions (collection time: 60 s, 15 fractions per run), whereas the online signals were measured on a time grid of typically one (UV, pH, conductivity, MALS, RI) or a few seconds (fluorescence, ATR-FTIR). Consequently, the online signals were averaged over the time interval of 60 s to achieve corresponding online/offline data pairs required for modeling. The so composed training matrix \mathbf{X} consists of $n = 225$ observations (one row per offline fraction for 13 runs, 15 fractions per run and triplicate offline measurements for a single run) and up to $p \approx 4650$ variables/predictors x_j .

Preprocessing methods for fluorescence and ATR-FTIR sensor signals are described in the supplementary material. Structured additive regression (STAR) was used as a statistical learning technique (Fahrmeir, Kneib, & Lang, 2004). Based on a $n \times p$ predictor matrix \mathbf{X} with entries x_{ij} and a n -vector of responses \mathbf{y} , a response value y_i is modeled with univariate linear (f_j^{lin}) or smooth terms (e.g., spline functions f_j^{s}) and (eventually) bivariate interaction terms $f_{j,k}^{\text{ia}}$ resulting in the following model:

$$y_i = \sum_j f_j^{\text{lin}}(x_{ij}) + \sum_k f_k^{\text{s}}(x_{ik}) + \sum_{j,k} f_{j,k}^{\text{ia}}(x_{ij}, x_{ik}) + \varepsilon_i \quad (1)$$

The sums extend over predictors (or pairs of predictors in the interaction case) and ε_i is a random error term capturing the unexplained variation in the data. Due to computational reasons, besides linear and penalized regression (P-) splines (Eilers & Marx, 1996) only product interactions between non-spectroscopic predictors were used.

Due to the high number of predictors, an efficient variable selection technique is required. The boosting algorithm builds the model stepwise, where in each step a single linear, smooth, or interaction term (in this context called a baselearner) is added to the current model starting with a simple intercept model and stopping after the addition of m (not necessarily different) baselearners with m (the number of iterations) usually being determined by a form of cross-validation (Bühlmann & Hothorn, 2007; Bühlmann & Yu, 2003). Typically, between a few hundred up to several thousand iterations are required. This model optimization is performed on the training set using leave-one-run-out cross-validation, that is a model is built on all observations except those from a single run and applied to the left-out observations.

This process is repeated until each run was left out once. As an error measure, the root-mean-square error is chosen:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

where n is the number of observations, y_i the measured and \hat{y}_i the predicted response. If the predictions \hat{y}_i in equation (2) are obtained in a CV framework on the training set, these errors are termed RMSE_{CV} (or cross-validation errors) and are used to select a single or a few best model(s). Applying this/these model(s) to the test set gives an independent estimate of the model performance by the so-called test error $\text{RMSE}_{\text{Test}}$. For descriptive purposes, in a few cases also the mean relative deviation (MRD [%]) will be given

$$\text{MRD} = \frac{100}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} \quad (3)$$

All computations were performed using the software platform R version 3.4.2 (R Core Team,) using the packages mboost (Hofner, Boccuto, & Göker, 2015; Hofner, Hothorn, Kneib, & Schmid, 2011; Hothorn et al., 2015), signal and baseline.

3 | RESULTS

The model building workflow is depicted in Figure 1. Standard signals from sensors of the chromatographic workstation (UV_{280 nm}, UV_{260 nm}, UV_{214 nm}, pH, conductivity) were complemented with MALS, RI, fluorescence spectroscopy, and ATR-FTIR (Figure 1a-e). The clarified *E.coli* homogenate was loaded on a CM-Sepharose Fast Flow column, washed and eluted with a linear salt gradient. Overlays of chromatograms are shown in Figure S1a,b. The eluate was fractionated and analyzed (Figure 1f-h). The FGF-2 concentration in the eluate fractions ranged from 0.1 to 28 mg/ml (Figures 2S a,b), HCP from 3 to 6579 ppm (Figure 3S a,b), and the dsDNA content from 8 to 3773 ppm (Figures 4S a,b).

We define a basic model as one containing only standard signals from the chromatographic workstation (UV_{280 nm}, UV_{260 nm}, UV_{214 nm}, pH, conductivity) as predictors, whereas in a medium model additionally MALS and/or RI predictors are considered. Both model types contain a small number of predictors ($p \leq 12$) and are simple to handle. On the other hand, an extensive model also contains spectroscopic predictors (ATR-FTIR and/or fluorescence). By using a stepwise approach (basic \rightarrow medium \rightarrow extensive model) the benefit of a sensor system for predicting a response can be assessed by comparing the RMSE_{CV} values for models with or without this sensor type. A more complex model (with respect to the number of sensors and/or predictors) is considered as superior only if a considerable reduction in RMSE_{CV} is achieved which compensates the loss in robustness (due to missing data extensive models can only be based on 7 instead of 13 runs as for the basic and medium models) and the higher computational costs.

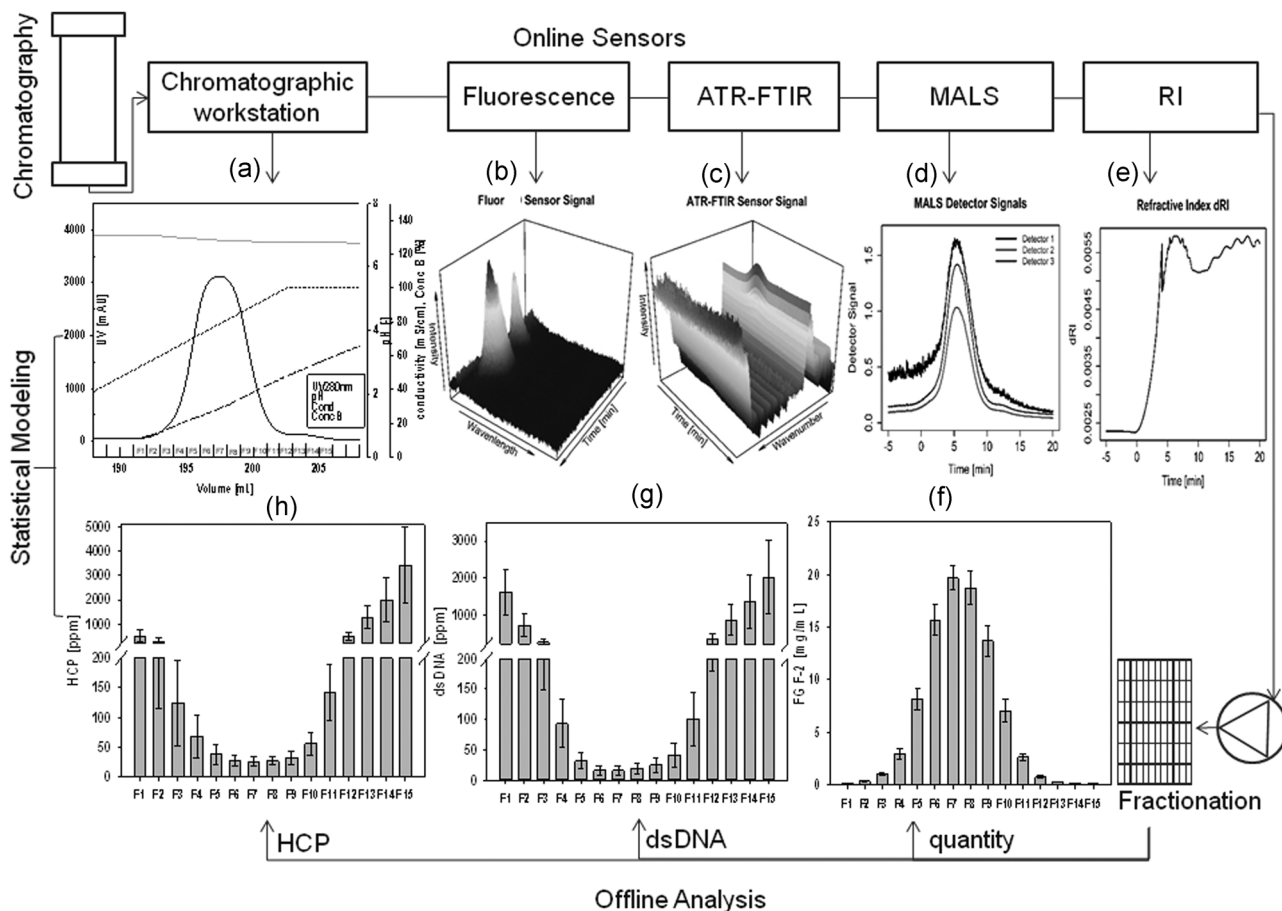


FIGURE 1 Real-time monitoring of a chromatographic step using a commercial chromatographic workstation equipped with additional online sensors. For each run, 15 fractions of the elution peak were collected and analyzed by offline assays for their FGF-2 concentration and impurity profile of HCP and dsDNA. Error bars in all figures represent \pm one standard deviation of the mean of each fraction calculated from 13 training runs. The online signals (a) $UV_{280\text{ nm}}/UV_{260\text{ nm}}/UV_{214\text{ nm}}/\text{conductivity}/\text{pH}$ were provided by the chromatographic workstation, (b) fluorescence sensor, (c) ATR-FTIR, (d) MALS, and (e) RI; offline data included: (f) FGF-2 quantity, (g) HCP content, and (h) dsDNA content. ATR-FTIR: attenuated total reflection Fourier-transform infrared; dsDNA: double-stranded DNA; HCP: host cell protein; FGF-2: fibroblast growth factor-2; MALS: multi-angle light scattering; RI: refractive index

3.1 | Prediction of FGF-2 quantity

Already the basic model enables an accurate prediction of the product quantity with a $RMSE_{CV}$ of 0.51 mg/ml (range in training data: 0.1–28 mg/ml), which corresponds to a relative error of about 6.4% if only fractions with a protein concentration above 1 mg/ml are considered. Further extension of the predictor set does not improve the model performance: The medium model based on the same training set gives the same $RMSE_{CV}$ of 0.51 mg/ml (Figure 2a). The extensive model (based on seven training runs) results in an error of 0.32 mg/ml which is only a negligible improvement to an $RMSE_{CV}$ of 0.33 mg/ml obtained by the basic model based on this smaller data set (Figure 2b), but at the cost of a much more complex model. The performance of the basic model on the test runs with a $RMSE_{Test}$ of 0.85 mg/ml is satisfactory (Figure 2c). This example demonstrates that a more complex model (with respect to the number of predictors) does not necessarily imply a more powerful model. This finding is in accordance with the state-of-the-art monitoring where estimation of product

concentration based on UV/Vis signals is well established. In fact, it turns out that omitting pH and using just the four signals $UV_{280\text{ nm}}$, $UV_{260\text{ nm}}$, $UV_{214\text{ nm}}$ and conductivity is sufficient to obtain the same model quality. There are a number of alternative models capable of predicting the protein quantity with slightly higher errors, among them models based solely on the fluorescence or ATR-FTIR sensor signals, which are neither contained in the previously suggested model. The fact that a sensor is not selected does not necessarily imply its uselessness for predicting response. However, it is natural that a model with only 4 predictors (UV and conductivity) is preferred.

In our purification process, a low HCP content of <20 ppm was reached in the fractions of the highest FGF-2 concentration and up to 6500 ppm in later eluting fractions. Therefore, the unspecific influence and contribution of HCP to the UV absorbance was not significant. For purification protocols, where a higher content of co-eluting impurities will be present, models including MALS, RI, ATR-FTIR, or fluorescence signals are expected to be superior over the basic model.

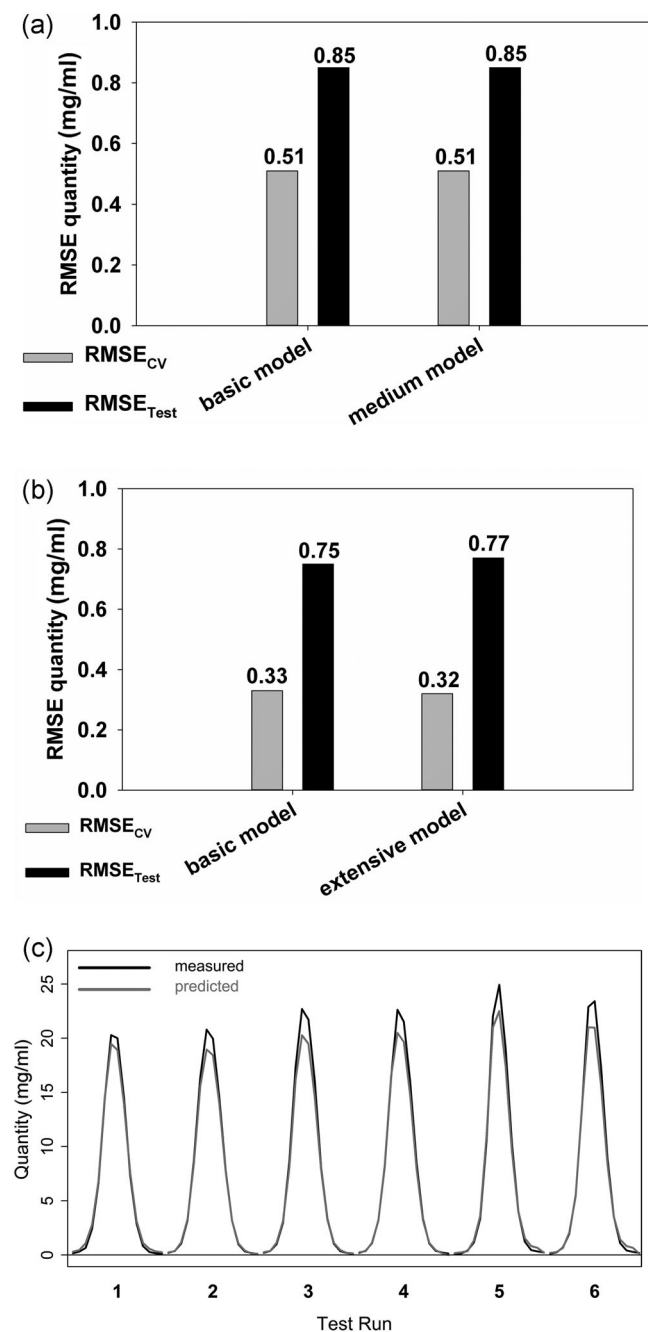


FIGURE 2 Performance of the prediction models for FGF-2 quantity (mg/ml) based on different sensor combinations. Comparison of RMSE_{CV} and RMSE_{Test} for (a) the basic and medium models based on 13 training runs and (b) basic and extensive models based on seven training runs. The basic model is the preferred one. (c) Comparison of the measured (black) and predicted (gray) values for 15 fractions in each of the six test runs (RMSE_{Test} = 0.85 mg/ml for the basic model). FGF-2: fibroblast growth factor 2; RMSE: root mean square error

3.2 | Prediction of HCP content

Basic and medium models for HCP content show similar prediction errors of 563 and 582 ppm, respectively, and hence no benefit of the MALS and RI sensors is determined (Figure 3a). Accurate prediction

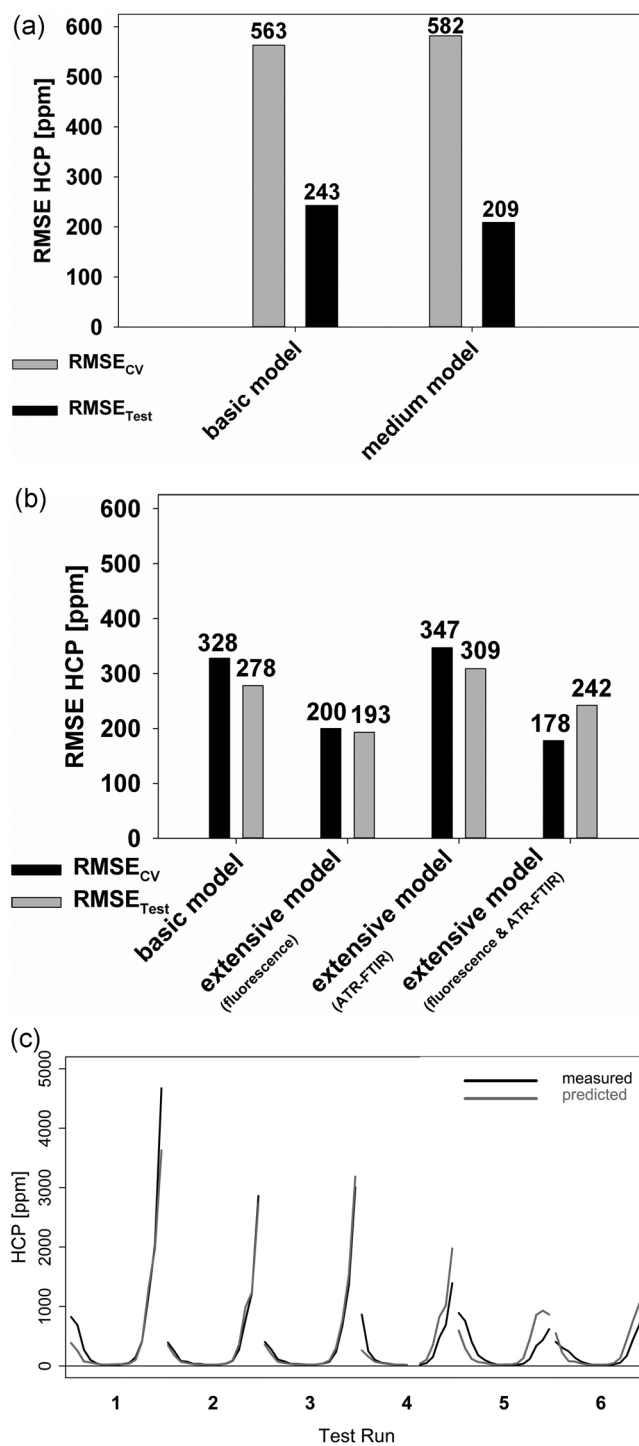


FIGURE 3 Performance of the prediction models for the HCP (ppm) based on different sensor combinations. Comparison of RMSE_{CV} and RMSE_{Test} for (a) basic and medium models based on 13 training runs and (b) basic and extensive models of several sensor combinations on seven training runs. The final contains predictors of the UV, conductivity, and fluorescence sensors. (c) Comparison of measured (black) and predicted (gray) values for the six test runs (overall test error of 193 ppm). HCP: host cell protein; RMSE: root mean square error

of HCP requires spectroscopic sensors, which becomes evident in Figure 3b depicting a comparison of the performance of the basic and extensive models on the smaller 7-run training set. A RMSE_{CV} of

200 ppm is obtained by adding fluorescence predictors to the UV and conductivity signals in the basic model, which is a 40% reduction in the prediction error. On the other hand, ATR-FTIR signals slightly decrease the model performance to 347 ppm and result in a 10% improvement to 178 ppm when added to the previous model already containing fluorescence predictors. The marked increase in the corresponding test error from 193 to 242 ppm might be an indication of overfitting, hence we consider a model based on UV, the conductivity and fluorescence signals as the final model. Omitting any of these sensors results in models with at least 50% increased error. The requirement of the extensive predictor set is obvious because the target product is present in excess and UV properties of the product and HCP are very similar and do not allow differentiation. The HCP prediction profiles are in good agreement with the offline measurements for the test runs (Figure 3c), even though the HCP levels in test runs 4–6 are significantly lower than those in test runs 1–3. The mean relative deviation (MRD) for HCP is 46% (with a median of 33%).

3.3 | Prediction of dsDNA content

Figure 4b compares the basic, medium and extensive models for the dsDNA content on the small 7-run data set indicating decreasing errors with increasing complexity of the models - basic (510 ppm) → medium (396 ppm) → extensive (339 ppm) (Figure 4b). When comparing the two former models on the larger (13-run) data set (Figure 4a), similar $RMSE_{CV}$ values of 321 and 341 ppm are obtained (due to missing fluorescence sensor data the extensive model cannot be evaluated on this data set). As these two models are based on approximately twice as many observations as the extensive model (13 vs. 7 runs), this might explain the lower test error of approximately 260 ppm (vs. \approx 360 ppm for the extensive model). The choice between the more robust and parsimonious basic model and a more complex extensive model, which outperforms the former on the 7-run data set, is somewhat subjective. However, Figure 4c shows exemplarily the performance of the extensive model containing UV, fluorescence, and ATR-FTIR signals on the test runs ($RMSE_{Test} = 359$ ppm). There are some obvious deviations between predicted and measured dsDNA values. Low dsDNA fractions are sometimes predicted negatively, but the model captures the general U-shape sufficiently well for all test runs. Predictions of negative concentrations could be avoided by modeling the log-transformed response (as was done for HCP), which results in strictly positive values upon back transformation using the exponential function. However, in the case of dsDNA, the overall errors (both cross-validated and test errors) were significantly lower for models based on the original response data. The benefit of the sensors for dsDNA quantification is lower than for protein-based information.

3.4 | Process control – model-based pooling

After integration into the chromatographic system, these models can be used in process control as a PAT application and enable

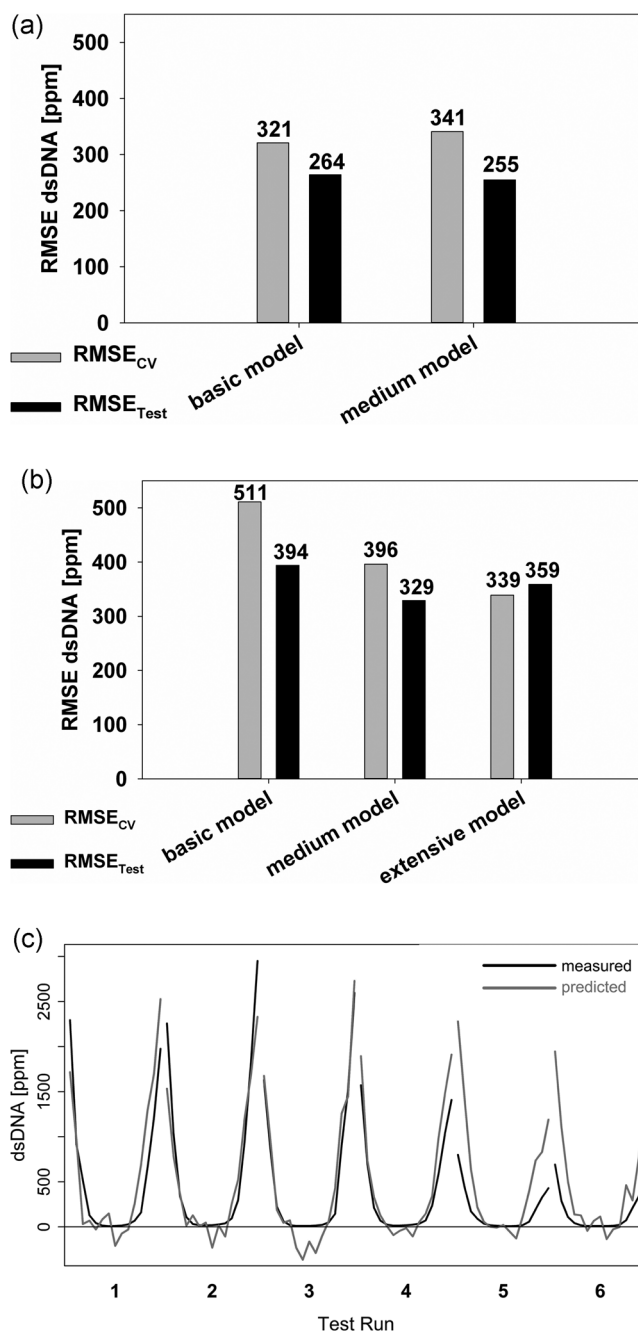


FIGURE 4 Performance of the prediction models for the dsDNA (ppm) based on different sensor combinations. Comparison of $RMSE_{CV}$ and $RMSE_{Test}$ for the (a) basic and medium models based on 13 training runs and (b) basic, medium, and extensive models based on seven training runs. The final model contains predictors of the UV, fluorescence and ATR-FTIR sensors. (c) Comparison of measured (black) and predicted (gray) values for the six test runs (overall test error of 359 ppm). ATR-FTIR: attenuated total reflection Fourier-transform infrared; ds DNA: double stranded DNA; RMSE: root mean square error

predictions of the quality attributes in real-time, which is the overall goal of the online monitoring system. The methodology presented can be used for online pooling by switching the collection valve, for decisions if the process stream is out of specification or for a

TABLE 1 Average pool composition and standard deviation of the six test runs based on offline and model-based pooling decisions.

	FGF-2 quantity (mg/ml)	Pool volume (ml)	HCP (ppm)	dsDNA (ppm)	Yield (%)
Offline pooling	8.9 ± 0.5	11.0 ± 0.8	32 ± 4	25 ± 8	98.1 ± 2.5
Model- based pooling	9.6 ± 0.4	9.3 ± 0.5	33 ± 1	29 ± 17	95.0 ± 5.2

Note. dsDNA: double-stranded DNA; FGF-2: fibroblast growth factor 2; HCP: host cell protein.

fraction-wise pooling after the process is completed resulting in a reduction of holding times and offline analytics. As process control application a fraction-wise pooling based on the online signals is demonstrated for the test runs. The results of the model-based pooling were compared to the conventional offline pooling based on measured values regarding yield, impurity content and product quantity. We assumed the following minimum pooling criteria: HCP content < 35 ppm, dsDNA content < 60 ppm and FGF-2 concentration > 1 mg/ml. The pools should meet these criteria at least on average and highest possible yield. Our model-based pooling leads to the collection of less fractions, but only with about 3% reduced product yield. The pools calculated on the model-based prediction compared to offline analysis showed 33 ± 1 ppm and 32 ± 4 ppm for the HCP content and 29 ± 17 ppm and 25 ± 8 ppm for the dsDNA, respectively. The standard deviation for the HCP content was even lower in the model-based pooling, whereas dsDNA content showed higher variation compared to the offline pooling (Table 1). Our attempt has already proved a model-based pooling of high accuracy, as HCP, dsDNA content and yield were in the same range of the offline values. Holding times and offline analytics can therefore be tremendously reduced.

4 | DISCUSSION

Our study demonstrates that the selected analytical spectroscopic methods provide a promising sensor combination for simultaneous determination of several biopharmaceutical quality attributes. By applying the final models (Table 2) on independent test runs, their validity could be shown and estimates of the future performance could be derived. The selection and combination of sensors in a specific situation will depend on the product, the unit operation and the stage of the downstream processing (capture, intermediate or polishing) monitored. In the Supporting Information Material prediction results for high molecular weight impurities are presented (Figure S5). The rapid, non-destructive spectroscopic methods enable real-time monitoring and control options for bioprocesses. Our model-based predictions can be computed within a few seconds and are suitable for process intervention. Structured additive regression in combination with boosting for variable selection proves to be a

TABLE 2 Summary (in terms of RMSE_{CV}, RMSE_{Test}) of the final prediction models for all responses.

Response	Final predictor set	Final model	
		RMSE _{CV}	RMSE _{Test}
FGF-2 (mg/ml)	Basic model (UV _{280 nm} , UV _{260 nm} , UV _{214 nm} , conductivity)	0.51	0.85
HCP (ppm)	Extensive model (UV _{280 nm} , UV _{260 nm} , UV _{214 nm} , conductivity, Fluorescence)	200	193
DsDNA (ppm)	Extensive model (UV _{280 nm} , UV _{260 nm} , UV _{214 nm} , conductivity, Fluorescence, ATR-FTIR)	339	359

Note. Ds DNA: double stranded DNA; FGF-2: fibroblast growth factor 2; HCP: host cell protein; RMSE: root mean square error.

high-performance modeling technique, particularly if predictor/response relations are nonlinear and the pool of potential predictors is large. Nevertheless, there are still several challenges related to real-time monitoring and model-based prediction. The main issue is that the prediction accuracy is restricted by the measurement error and variation of the offline assays. It is obvious that a model cannot be more accurate than the offline assay applied for data generation and can, therefore, be considered as a starting point for improvements to become more precise. The models used for process development need to be applied to highly varying training sets. The training set(s) need to include larger variations (in the measured data), enabling a large prediction range. In contrast, models used for process control in manufacturing have to perform accurately within the defined operation space to detect deviations/drifts and prevent the process to be out of the specifications by control actions (e.g., defined pooling). The presented models were established for process control as the training set used was generated based on typical inherent process variations caused by preprocessing of the *E. coli* homogenate of one fermentation batch and verified by test runs processing additional fermentation batches. Real-time monitoring data of chromatographic runs from deliberately varied process conditions were not included. As bioprocesses are comprised of integrated unit operations wider ranges of validity have to be tested and enlarged, including chromatographic training runs where process parameters will be varied in a controlled way. Such input process parameters influencing the output of quality attributes have to be selected based on a risk assessment. Robustness could then be improved within certain experimental variations to enhance the operation space, for example, the different starting material, spiking of impurities, buffer composition as predictions models are then trained on these deviations. A challenge in our application for prediction of product and impurity content during elution was the gradient elution. The buffer varies with progressing elution. Certain sensors are very sensitive to the electrolyte composition, for example, ATR-FTIR and RI. Except for the pH probe, all online sensors have a fast response time and can be used with process relevant flow rates. It is known that several salts and buffer

components give signals in certain regions of the ATR-FTIR spectra (Capito, Skudas, Stanislawski, & Kolmar, 2013; Rathore et al., 2009). Usually, the buffer background is subtracted before the spectrum is deconvoluted. This is not possible with a linear gradient or suitable for real-time monitoring. We circumvented this problem by using preprocessing operations such as spectra differences and normalization. For spectroscopic data evaluation, statistical models are especially useful to extract information as spectral overlaps of matrix background and analytes are common (Esmonde-White, Cuellar, Uerpmann, Lenain, & Lewis, 2017). The models established in our setup will be valid for minor process deviations as they have been verified with the independent test runs. The process variability represented in the training set defines the model applicability (Esmonde-White et al., 2017). Therefore transferability of the models to other processes relies on measured data in the training set (Craven, Shirsat, Whelan, & Glennon, 2013; Kroll, Hofer, Ulonska, Kager, & Herwig, 2017; Pernot, 2017). The established methodology allows simultaneous real-time prediction of quantity, HCP, and dsDNA. This attempt will be a basis for process control and real-time release. This real-time monitoring approach requires the cooperation of several disciplines: data science, biotechnology, biophysics, and software engineering. More work like the prediction of product potency is required to reach real-time release requirements as end-product testing cannot be replaced so far, but already in-process offline testing can be reduced and real-time control can be conducted for process consistency.

5 | CONCLUSION

Real-time monitoring of a chromatographic capture step was successfully implemented and provided a model-based prediction of HCP and dsDNA content and quantity of a biopharmaceutical. STAR modeling can be applied for the prediction of (critical) quality attributes in the eluate within seconds, despite the co-elution of many protein and non-protein impurities. A small set of online signals from the chromatographic workstation enabled the adequate prediction of protein quantity. However, the prediction of HCP and dsDNA content demanded more complex models including spectroscopic sensors. The online sensors setup and the predictive models are the basis for real-time interventions, either for process control (e.g., pooling) or real-time release. In chromatography, the production efficiency, yield, and product quality can be improved and time-consuming offline analyses are reduced. Our findings pave the way towards PAT implementation in biopharmaceutical manufacturing.

ACKNOWLEDGMENTS

This work was supported by the Austrian Federal Ministry for Digital and Economic Affairs (bmwd), the Federal Ministry for Transport, Innovation and Technology (bmvit), the Styrian Business Promotion Agency (SFG), the Standortagentur Tirol,

the Government of Lower Austria and ZIT - Technology Agency of the City of Vienna through the COMET-Funding Program (grant number 824186) managed by the Austrian Research Promotion Agency (FFG). The funding agencies had no influence on the conduct of this research. The computational results presented were achieved using the Vienna Scientific Cluster (VSC).

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

ORCID

Alois Jungbauer  <http://orcid.org/0000-0001-8182-7728>

Astrid Dürauer  <http://orcid.org/0000-0002-6007-7697>

REFERENCES

- Antosiewicz, J. M., & Shugar, D. (2016). UV-Vis spectroscopy of tyrosine side-groups in studies of protein structure. Part 1: Basic principles and properties of tyrosine chromophore. *Biophysics Review*, 8(2), 151–161. <https://doi.org/10.1007/s12551-016-0198-6>
- Barth, A. (2007). Infrared spectroscopy of proteins. *Biochimica et Biophysica Acta*, 1767(9), 1073–1101. <https://doi.org/10.1016/j.bbapbio.2007.06.004>
- Borg, N., Brodsky, Y., Moscariello, J., Vunnum, S., Vedantham, G., Westerberg, K., & Nilsson, B. (2014). Modeling and robust pooling design of a preparative cation-exchange chromatography step for purification of monoclonal antibody monomer from aggregates. *Journal of Chromatography A*, 1359, 170–181. <https://doi.org/10.1016/j.chroma.2014.07.041>
- Brestrich, N., Briskot, T., Osberghaus, A., & Hubbuch, J. (2014). A tool for selective inline quantification of co-eluting proteins in chromatography using spectral analysis and partial least squares regression. *Biotechnology and Bioengineering*, 111(7), 1365–1373. <https://doi.org/10.1002/bit.25194>
- Brestrich, N., Sanden, A., Kraft, A., McCann, K., Bertolini, J., & Hubbuch, J. (2015). Advances in inline quantification of co-eluting proteins in chromatography: Process-data-based model calibration and application towards real-life separation issues. *Biotechnology and Bioengineering*, 112(7), 1406–1416. <https://doi.org/10.1002/bit.25546>
- Bühlmann, P., & Yu, B. (2003). Boosting with the L_2 loss: Regression and classification. *Journal of the American Statistical Association*, 98, 324–339. <https://doi.org/10.1198/016214503000125>
- Bühlmann, P., & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22, 477–505. <https://doi.org/10.1214/07-STS242>
- 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. <https://doi.org/10.1002/bit.24611>
- Capito, F., Skudas, R., Stanislawski, B., & Kolmar, H. (2013). Matrix effects during monitoring of antibody and host cell proteins using attenuated total reflection spectroscopy. *Biotechnology Progress*, 29(1), 265–274. <https://doi.org/10.1002/btpr.1643>
- Craven, S., Shirsat, N., Whelan, J., & Glennon, B. (2013). Process model comparison and transferability across bioreactor scales and modes of operation for a mammalian cell bioprocess. *Biotechnology Progress*, 29(1), 186–196. <https://doi.org/10.1002/btpr.1664>

- Dabros, M., Amrhein, M., Bonvin, D., Marison, I. W., & von Stockar, U. (2009). Data reconciliation of concentration estimates from mid-infrared and dielectric spectral measurements for improved on-line monitoring of bioprocesses. *Biotechnology Progress*, 25(2), 578–588. <https://doi.org/10.1002/btpr.143>
- Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11, 89–121.
- Esmonde-White, K. A., Cuellar, M., Uerpmann, C., Lenain, B., & Lewis, I. R. (2017). Raman spectroscopy as a process analytical technology for pharmaceutical manufacturing and bioprocessing. *Analytical and Bioanalytical Chemistry*, 409(3), 637–649. <https://doi.org/10.1007/s00216-016-9824-1>
- Fahrmeir, L., Kneib, T., & Lang, S. (2004). Penalized structured additive regression for space-time data: A bayesian perspective. *Statistica Sinica*, 14, 109–118.
- Flatman, S., Alam, I., Gerard, J., & Mussa, N. (2007). Process analytics for purification of monoclonal antibodies. *Journal of Chromatography B*, 848(1), 79–87. <https://doi.org/10.1016/j.jchromb.2006.11.018>
- Food and Drug Administration, H. S. S. (2003). International Conference on Harmonisation; revised guidance on Q3A impurities in new drug substances; availability.. *Notice. Fed Regist*, 68(28), 6924–6925. <https://www.federalregister.gov/documents/2003/02/11/03-3352/international-conference-on-harmonisation-revised-guidance-on-q3a-impurities-in-new-drug-substances>
- Food and Drug Administration (2004). Guidance for Industry. PAT – A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance. <http://www.fda.gov/cder/OPS/PAT.htm>
- Gasparian, M. E., Elistratov, P. A., Drize, N. I., Nifontova, I. N., Dolgikh, D. A., & Kirpichnikov, M. P. (2009). Overexpression in *Escherichia coli* and purification of human fibroblast growth factor (FGF-2). *Biochemistry (Mosc)*, 74(2), 221–225.
- Ghisaidoobe, A. B., & Chung, S. J. (2014). Intrinsic tryptophan fluorescence in the detection and analysis of proteins: A focus on Förster resonance energy transfer techniques. *International Journal of Molecular Sciences*, 15(12), 22518–22538. <https://doi.org/10.3390/ijms151222518>
- Großhans, S., Rüdter, M., Sanden, A., Brestrich, N., Morgenstern, J., Heissler, S., & Hubbuch, J. (2018). In-line Fourier-transform infrared spectroscopy as a versatile process analytical technology for preparative protein chromatography. *Journal of Chromatography A*, 1547, 37–44. <https://doi.org/10.1016/j.chroma.2018.03.005>
- Hofner, B., Boccuto, L., & Göker, M. (2015). Controlling false discoveries in high-dimensional situations: Boosting with stability selection. *BMC Bioinformatics*, 16, 144. <https://doi.org/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, 956–971. <https://doi.org/10.1198/jcgs.2011.09220>
- Holm, P., Alleso, M., Bryder, M. C., & Holm, R. (2017). ICH quality guideline: An implementation guide. In E. Teasdale, D. Elder, & R. W. Niims (Eds.), Q8(R2). Hoboken, NJ: John Wiley & Sons, Inc. <https://doi.org/10.1002/9781118971147.ch20>
- Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., Hofner, B. (2015). mboost: Model-based boosting. In (Vol. R package version 2.4-2). <http://CRAN.R-project.org/package=mboost>.
- Jiang, M., Severson, K. A., Love, J. C., Madden, H., Swann, P., Zang, L., & Braatz, R. D. (2017). Opportunities and challenges of real-time release testing in biopharmaceutical manufacturing. *Biotechnology and Bioengineering*, 114(11), 2445–2456. <https://doi.org/10.1002/bit.26383>
- Kroll, P., Hofer, A., Ulonska, S., Kager, J., & Herwig, C. (2017). Model-based methods in the biopharmaceutical process lifecycle. *Pharmaceutical Research*, 34(12), 2596–2613.
- Löfgren, A., Andersson, N., Sellberg, A., Nilsson, B., Löfgren, M., & Wood, S. (2017). Designing an autonomous integrated downstream sequence from a batch separation process - An industrial case study. *Biotechnology Journal*, 13, 1700691. <https://doi.org/10.1002/biot.201700691>
- Lorber, B., Fischer, F., Bailly, M., Roy, H., & Kern, D. (2012). Protein analysis by dynamic light scattering: Methods and techniques for students. *Biochemistry and Molecular Biology Education*, 40(6), 372–382. <https://doi.org/10.1002/bmb.20644>
- Luchner, M., Gutmann, R., Bayer, K., Dunkl, J., Hansel, A., Herbig, J., & Striedner, G. (2012). Implementation of proton transfer reaction-mass spectrometry (PTR-MS) for advanced bioprocess monitoring. *Biotechnology and Bioengineering*, 109(12), 3059–3069. <https://doi.org/10.1002/bit.24579>
- Melcher, M., Scharl, T., Luchner, M., Striedner, G., & Leisch, F. (2017). Boosted structured additive regression for *Escherichia coli* fed-batch fermentation modeling. *Biotechnology and Bioengineering*, 114(2), 321–334. <https://doi.org/10.1002/bit.26073>
- 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. <https://doi.org/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. <https://doi.org/10.1016/j.ab.2016.02.007>
- Pais, D. A., Carrondo, M. J., Alves, P. M., & Teixeira, A. P. (2014). Towards real-time monitoring of therapeutic protein quality in mammalian cell processes. *Current Opinion in Biotechnology*, 30, 161–167. <https://doi.org/10.1016/j.copbio.2014.06.019>
- Patel, B. A., Pinto, N. D. S., Gospodarek, A., Kilgore, B., Goswami, K., Napoli, W. N., & Richardson, D. D. (2017). On-line ion exchange liquid chromatography as a process analytical technology for monoclonal antibody characterization in continuous bioprocessing. *Analytical Chemistry*, 89(21), 11357–11365. <https://doi.org/10.1021/acs.analchem.7b02228>
- Pernot, P. (2017). The parameter uncertainty inflation fallacy. *The Journal of Chemical Physics*, 147(10), 104102. <https://doi.org/10.1063/1.4994654>
- R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from <https://www.R-project.org/>.
- Rathore, A. S. (2016). Quality by design (QbD)-based process development for purification of a biotherapeutic. *Trends in Biotechnology*, 34(5), 358–370. <https://doi.org/10.1016/j.tibtech.2016.01.003>
- Rathore, A. S., Yu, M., Yeboah, S., & Sharma, A. (2008). Case study and application of process analytical technology (PAT) towards bioprocessing: Use of on-line high-performance liquid chromatography (HPLC) for making real-time pooling decisions for process chromatography. *Biotechnology and Bioengineering*, 100(2), 306–316. <https://doi.org/10.1002/bit.21759>
- Rathore, A. S., Wood, R., Sharma, A., & Dermawan, S. (2008). Case study and application of process analytical technology (PAT) towards bioprocessing: II. Use of ultra-performance liquid chromatography (UPLC) for making real-time pooling decisions for process chromatography. *Biotechnology and Bioengineering*, 101(6), 1366–1374. <https://doi.org/10.1002/bit.21982>
- Rathore, A. S., Li, X., Bartkowski, W., Sharma, A., & Lu, Y. (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. <https://doi.org/10.1002/btpr.212>
- Read, E. K., Park, J. T., Shah, R. B., Riley, B. S., Brorson, K. A., & Rathore, A. S. (2010). Process analytical technology (PAT) for biopharmaceutical products: Part I. concepts and applications. *Biotechnology and Bioengineering*, 105(2), 276–284. <https://doi.org/10.1002/bit.22528>
- Rüdter, 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. <https://doi.org/10.1016/j.chroma.2016.11.010>

- Rüdt, M., Brestrich, N., Rolinger, L., & Hubbuch, J. (2017). Real-time monitoring and control of the load phase of a protein A capture step. *Biotechnology and Bioengineering*, 114(2), 368–373. <https://doi.org/10.1002/bit.26078>
- Sauer, D. G., Mosor, M., Frank, A. C., Weiß, F., Christler, A., Walch, N., & Dürauer, A. (2018). A two-step process for capture and purification of human basic fibroblast growth factor from *E. coli* homogenate: Yield versus endotoxin clearance. *Protein Expression and Purification*, 153, 70–82. <https://doi.org/10.1016/j.pep.2018.08.009>
- Scott, B., & Wilcock, A. (2006). Process analytical technology in the pharmaceutical industry: A toolkit for continuous improvement. *PDA Journal of Pharmaceutical Science and Technology*, 60(1), 17–53.
- von Stosch, M., Hamelink, J. M., & Oliveira, R. (2016). Hybrid modeling as a QbD/PAT tool in process development: An industrial *E. coli* case study. *Bioprocess and Biosystem Engineering*, 39(5), 773–784. <https://doi.org/10.1007/s00449-016-1557-1>
- Workman, J., Koch, M., & Veltkamp, D. (2007). Process analytical chemistry. *Analytical Chemistry*, 79(12), 4345–4363. <https://doi.org/10.1021/ac070765q>
- Yu, L. X., Amidon, G., Khan, M. A., Hoag, S. W., Polli, J., Raju, G. K., & Woodcock, J. (2014). Understanding pharmaceutical quality by design. *The AAPS Journal*, 16(4), 771–783. <https://doi.org/10.1208/s12248-014-9598-3>
- 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. <https://doi.org/10.1002/jps.23746>
- Zhao, H., Brown, P. H., & Schuck, P. (2011). On the distribution of protein refractive index increments. *Biophysical Journal*, 100(9), 2309–2317. <https://doi.org/10.1016/j.bpj.2011.03.004>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Sauer DG, Melcher M, Mosor M, et al. Real-time monitoring and model-based prediction of purity and quantity during a chromatographic capture of fibroblast growth factor 2. *Biotechnology and Bioengineering*. 2019;116: 1999–2009. <https://doi.org/10.1002/bit.26984>