PHCbi

70% of surveyed scientists admitted that they could not replicate someone else's research.[1]

50% admitted that they couldn't replicate their own research.[1]

MCO-50M-PA

MCO-50AICL-PA

Compact 1.8 cu.ft., stackable three high, with or without O^2 control.

Grow Cells Stress-Free Every Time

Improve Reproducibility in Clinical and Research Applications

Successful cell cultures require precise CO_2, O_2, temperature, humidity and real-time contamination protection maintained in PHCbi MCO-50 Series laboratory incubators. These compact incubators prevent contamination before it starts with standard inCu-saFe® copper-enriched germicidal surfaces, easy clean integrated shelf channels and condensation control. H_2O_2 vapor and SafeCell™ UV scrubbing combine to increase *in vitro* cell safety.

Learn more at www.phchd.com/us/biomedical/cellculture-incubators

PHC Corporation of North America

PHC Corporation of North America
1300 Michael Drive, Suite A, Wood Dale, IL 60191
Toll Free USA (800) 858-8442, Fax (630) 238-0074
www.phchd.com/us/biomedical

1) Baker, Monya. "1,500 scientists lift the lid on reproducibility." Nature, no. 533 (May 26, 2016): 452-54. doi:10.1038/533452a.

Received: 21 April 2021 | Revised: 16 June 2021 | Accepted: 9 July 2021

DOI: 10.1002/bit.27894

ARTICLE

BIOTECHNOLOGY AND BIOENGINEERING WILEY

Comparison of UV- and Raman-based monitoring of the Protein A load phase and evaluation of data fusion by PLS models and CNNs

Laura Rolinger[1,2] | Matthias Rüdt[1,3] | Jürgen Hubbuch[1]

[1]Institute of Engineering in Life Sciences, Section IV: Biomolecular Separation Engineering, Karlsruhe Institute of Technology, Karlsruhe, Germany

[2]PTDC-P PAT, Hoffmann-La Roche AG, Basel, Switzerland

[3]Haute Ecole d'Ingénierie, HES-SO Valais-Wallis, Sion, Switzerland

Correspondence

Jürgen Hubbuch, Institute of Engineering in Life Sciences, Section IV: Biomolecular Separation Engineering, Karlsruhe Institute of Technology, Fritz-Haber-Weg 2, 76131 Karlsruhe, Germany.
Email: Juergen.Hubbuch@kit.edu

Abstract

A promising application of Process Analytical Technology to the downstream process of monoclonal antibodies (mAbs) is the monitoring of the Protein A load phase as its control promises economic benefits. Different spectroscopic techniques have been evaluated in literature with regard to the ability to quantify the mAb concentration in the column effluent. Raman and Ultraviolet (UV) spectroscopy are among the most promising techniques. In this study, both were investigated in an inline setup and directly compared. The data of each sensor were analyzed independently with Partial-Least-Squares (PLS) models and Convolutional Neural Networks (CNNs) for regression. Furthermore, data fusion strategies were investigated by combining both sensors in hierarchical PLS models or in CNNs. Among the tested options, UV spectroscopy alone allowed for the most precise and accurate prediction of the mAb concentration. A Root Mean Square Error of Prediction (RMSEP) of 0.013 g L^{-1} was reached with the UV-based PLS model. The Raman-based PLS model reached an RMSEP of 0.232 g L^{-1}. The different data fusion techniques did not improve the prediction accuracy above the prediction accuracy of the UV-based PLS model. Data fusion by PLS models seems meritless when combining a very accurate sensor with a less accurate signal. Furthermore, the application of CNNs for UV and Raman spectra did not yield significant improvements in the prediction quality. For the presented application, linear regression techniques seem to be better suited compared with advanced nonlinear regression techniques, like, CNNs. In summary, the results support the application of UV spectroscopy and PLS modeling for future research and development activities aiming to implement spectroscopic real-time monitoring of the Protein A load phase.

KEYWORDS

data fusion, partial-least-squares regression, process analytical technology, Raman spectroscopy, UV spectroscopy

1 | INTRODUCTION

In biopharmaceutical downstream processing of monoclonal antibodies (mAbs), a focus of Process Analytical Technology (PAT) research has been on the monitoring of the Protein A load phase (Feidl, Garbellini, Luna, et al., 2019; Feidl, Garbellini, Vogg, et al., 2019; Thakur et al., 2020; Rüdt et al., 2017) as this application promises the most economic benefits due to the high costs of Protein A resin (Rolinger et al., 2020b). Economic improvements may be achieved due to multiple aspects. In conventional batch production, the Protein A column capacity is typically underused. The acceptance range for the column loading density is set such that it can be kept constant during the resin lifetime. A dynamic termination of the load phase by detecting product breakthrough allows one to use the optimal column capacity throughout resin life time. Furthermore, real-time PAT eliminates the need for completing at- or off-line titer measurements before starting the downstream process resulting in a more streamlined production. As pharmaceutical companies move towards continuous processes, real-time monitoring of the Protein A load phase becomes more interesting to support robust process control. In continuous Protein A chromatography, the effluent of a first column is commonly loaded onto a second column, which allows one to overload the columns without losing product. If a continuous load stream with a variable mAb titer is used, monitoring the product concentration in the breakthrough continuously reduces the dependence of the process on at- or off-line analytics and thus improves the process control.

Different spectroscopic sensors, like, ultraviolet (UV) (Rolinger et al., 2020b; Rüdt et al., 2017), Near-Infrared (NIR; Thakur et al., 2020), and Raman (Feidl, Garbellini, Luna, et al., 2019; Feidl, Garbellini, Vogg, et al., 2019), have been investigated for the purpose of quantifying the mAb concentration in the column effluent with varying success. On the basis of the literature data, UV spectroscopy and Raman spectroscopy seem to be the most promising techniques for the breakthrough monitoring of the Protein A load.

Raman spectroscopy has been successfully implemented to monitor various attributes during the upstream process of mAbs, including the mAb concentration in the complex cell culture fluid (Abu-Absi et al., 2011; Buckley & Ryder, 2017; Li et al., 2013, 2010). A limiting factor for the application of Raman spectroscopy to the downstream process is the long acquisition times to derive a good signal-to-noise ratio. This is important, because process steps in the downstream take hours in comparison to days during the fermentation (Rolinger et al., 2020a). Therefore, Feidl et al. (2019, 2019) applied advanced preprocessing of the spectra and mechanistic modeling the prediction of the mAb concentration to overcome the noise limitation of the Raman spectra due to short measurement times.

For monitoring the downstream process, the application of UV-based PAT methods was proven to be successful for selective mAb concentration measurements (Brestrich et al., 2018, 2015; Rolinger et al., 2020b; Rüdt et al., 2017; Zobel-Roos et al., 2017). Raman spectroscopy has been proven to selectively quantify protein (Wen, 2007) and different buffer components (Saggu et al., 2015), which can be interesting for Ultrafiltration/Diafiltration (UF/DF) steps and formulation. In comparison to Raman-based techniques, UV spectroscopy offers a higher measurement speed and a better signal-to-noise ratio for quantification of proteins in aqueous solutions with the drawback of less selectivity for different protein features (Rolinger et al., 2020a). To compensate the lower selectivity and thereby improve the prediction of the UV-based PAT methods, dynamic background subtraction methods have been investigated to remove the influence of process-related impurities on the UV spectra (Rolinger et al., 2020b; Rüdt et al., 2017). Another drawback of the UV spectroscopy in comparison to Raman spectroscopy is the detector saturation at high protein concentrations. To resolve this, a flow cell with adequate pathlength or with variable pathlength needs to be chosen. Raman spectroscopy has a larger working range due to more possibilities in laser and detector settings to avoid the saturation of the detector.

The comparison of the techniques with results from different studies remains difficult as different sample conditions and different methods for model optimization and model validation can influence the results dramatically. Therefore, a final conclusion can only be drawn, when using the different sensors on the same sample set and by applying the same model methodology. An application to the same sample set can be realized by serial in-line measurements with both sensors. This also enables the application of data fusion algorithms on the multimodal data set. Data fusion from multiple sensors promises advantages over data from a single source, like, the statistical advantage of improving the number of measurements and the improved observability by combining multimodal measurement data (Liggins et al., 2017). The development and use of chemometric data fusion algorithms of multimodal spectroscopic sensors have been driven by food science (Biancolillo et al., 2016; Borràs et al., 2015), but data fusion is starting to be used in biopharmaceutical production as well (Rolinger et al., 2020a). Up to the present, mostly low-level data fusion is used and a thorough investigation into the improved prediction by data fusion methods in comparison to single sensor models is missing.

In this study, Raman spectroscopy and UV spectroscopy are evaluated based on their ability to quantify the mAb concentration in the column effluent of the Protein A column. It is discussed what molecular features the spectroscopic techniques measures to quantify the mAb concentration of complex mixtures. Additionally, data fusion techniques are applied to evaluate the benefit of two orthogonal sensors. First, traditional data fusion techniques, which are based on Partial-Least-Squares (PLS) modeling, are compared with the base PLS models of the individual sensors. Special emphasis is put on the considerations for variable and data block scaling, and on the comparison to the single sensor models. In a second step, the application of Convolutional Neural Networks (CNNs) as nonlinear regression techniques is evaluated for Raman and UV spectroscopy. Lastly, the potential of CNNs as a data fusion technique is explored and compared with the traditional PLS-based data fusion techniques.

2 | MATERIALS AND METHODS

2.1 | Biologic material

All biologic material was stored at 5°C before experimentation after delivery from our industry partner Sanofi-Aventis. To obtain a variable mAb concentration and a variable impurity profile in the load material, the product containing Harvested Cell Culture Fluid (HCCF) with a product concentration of 2 g L^{-1} (Feedstock 1) was mixed with purified product (Feedstock 2) and three different mock HCCF solutions (Feedstocks 3–5). One mock solution was cultivated with a nonproducing cell line. The other two mock solutions were prepared as flow-through by preparative Protein A chromatography. These two mock solutions were derived from HCCFs of two different cell lines which produce two different mAbs, respectively. Before this study, it was ensured that the Protein A flow-through did not contain antibodies in detectable concentrations (based on analytical Protein A chromatography). For product spiking, the used mAb (Feedstock 2) was purified to the second polishing step by our industry partner and was concentrated up to 20 g L^{-1} to reduce dilution effects of the impurities by addition of the concentrated product.

The product containing HCCF, purified mAb, and mock HCCFs was filtered with a cellulose acetate filter with a pore size of 0.22 µm (Pall) before mixing. In Table 1, the used volumina of the different stock materials for each run are shown. The composition of the mixtures between the three mock materials was determined by Latin Hypercube Sampling to provide a random multidimensional distribution.

2.2 | Chromatography runs and sensors

All preparative runs were realized with an Äkta Pure 25 purification system controlled by Unicorn 6.4.1 (Cytiva). The system was equipped with a sample pump S9, a fraction collector F9-C, a column valve kit (V9-C, for up to five columns), a UV-monitor U9-M (2 mm pathlength), a conductivity monitor C9, a pH valve kit (V9-pH) and an I/O-box E9. To monitor the breakthrough by Raman spectroscopy, a MarqMetrix BioReactor Ballprobe (MarqMetrix) was inserted into an in-house made flow cell. The probe was connected to a HyperFlux PRO Plus 785 Raman analyzer with Spectralsoft 2.8.0 (Tornado Spectral Systems). The laser power during acquisition was set to 495 mW with an acquisition time of 800 ms and 10 acquisitions per spectrum. The flow cell was placed after the conductivity monitor of the Äkta system. In Figure 1 the flow cell is displayed. X-, Y- and laser calibration were done before the experiment according to the manual. More information on the Raman measurement setup is given in the Supporting Information Data A.

Additionally, an UltiMate 3000 Diode Array Detector (DAD) equipped with a semipreparative flow cell (0.4 mm optical pathlength) and operated with Chromeleon 6.8 (Thermo Fisher Scientific) was connected to the Äkta Pure. The DAD was positioned between the Raman flow cell and the V9-pH valve.

For the PLS model calibration and validation, breakthrough experiments with variable mAb titers in the feed were performed. The mAb titers in the different load materials were 1, 1.5, 2, 2.5, and 3 g L^{-1}. For each experiment, a prepacked 5 mm × 50 mm, MabSelect SuRe column (0.982 ml; Repligen) was first equilibrated for 5 Column Volumes (CVs) with a 25 mM Tris(hydroxymethyl)aminomethane (TRIS) and 0.1 mM sodium chloride buffer at pH 7.4, and then loaded with 100 mg of mAb. At the beginning of the load phase, the DAD equipped with a semipreparative flow cell (optical pathlength 0.4 mm) was triggered to record absorption spectra between 200 and 800 nm and the column flow-through was collected in 200 µl fractions, as explained in more detail by Rüdt et al. (2017). An additional command was inserted into the MATLAB script (MATLAB version R2019b from the MathWorks Inc.) to trigger the Raman measurements over Transmission Control Protocol/Internet Protocol (TCP/IP).

After the load phase, the column was washed for 4.5 CVs with equilibration buffer, before the mAb was eluted with 20 mM citric acid at pH 3.6. A sanitization was conducted with 50 mM sodium hydroxide and 1 mM sodium chloride for 5 CVs after each run.

2.3 | Analytical chromatography

Reference analysis of the collected fractions was performed using a Vanquish Flex Binary High-Performance Liquid Chromatography (HPLC) system (Thermo Fisher Scientific) by analytical Protein A chromatography. The system consisted of a Binary Pump F,

TABLE 1 Sample composition for the calibration runs 1–4 and the validation run 5 with volumes of the product containing HCCF (Feedstock 1), purified mAb (Feedstock 2), mock HCCF (Feedstock 3), and flow-through 1 and 2 (Feedstocks 4 and 5)

Run number (ml)	Data usage –	HCCF (ml)	mAb (ml)	Flow-through 1 (ml)	Flow-through 2 (ml)	mock HCCF (ml)	Final mAb concentration (g L^{-1})
Run 1	Calibration	52.50	0.00	9.85	21.91	20.74	1.0
Run 2	Calibration	35.00	1.75	14.82	1.36	17.08	1.5
Run 3	Calibration	21.00	3.15	6.48	3.93	7.44	2.5
Run 4	Calibration	17.50	3.50	6.57	6.13	1.30	3.0
Run 5	Validation	26.25	2.63	2.01	12.65	8.96	2.0

Abbreviations: HCCF, Harvested Cell Culture Fluid; mAb, monoclonal antibody.

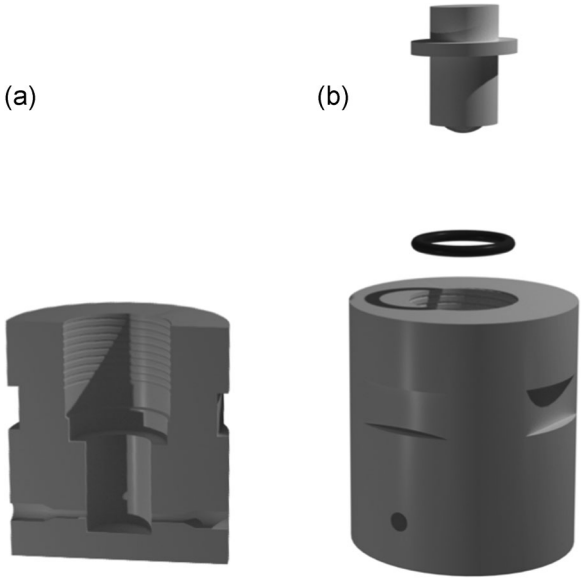

FIGURE 1 Cut of the (a) and exploded view of the in-house made flow cell, O-ring, and MarqMetrix Ballprobe with welded flange (b). The flow cell consists of a block of stainless steel with a PG 13.5-sized threaded borehole to insert the Ballprobe and two boreholes for 1/16 inch Äkta fingertight connectors

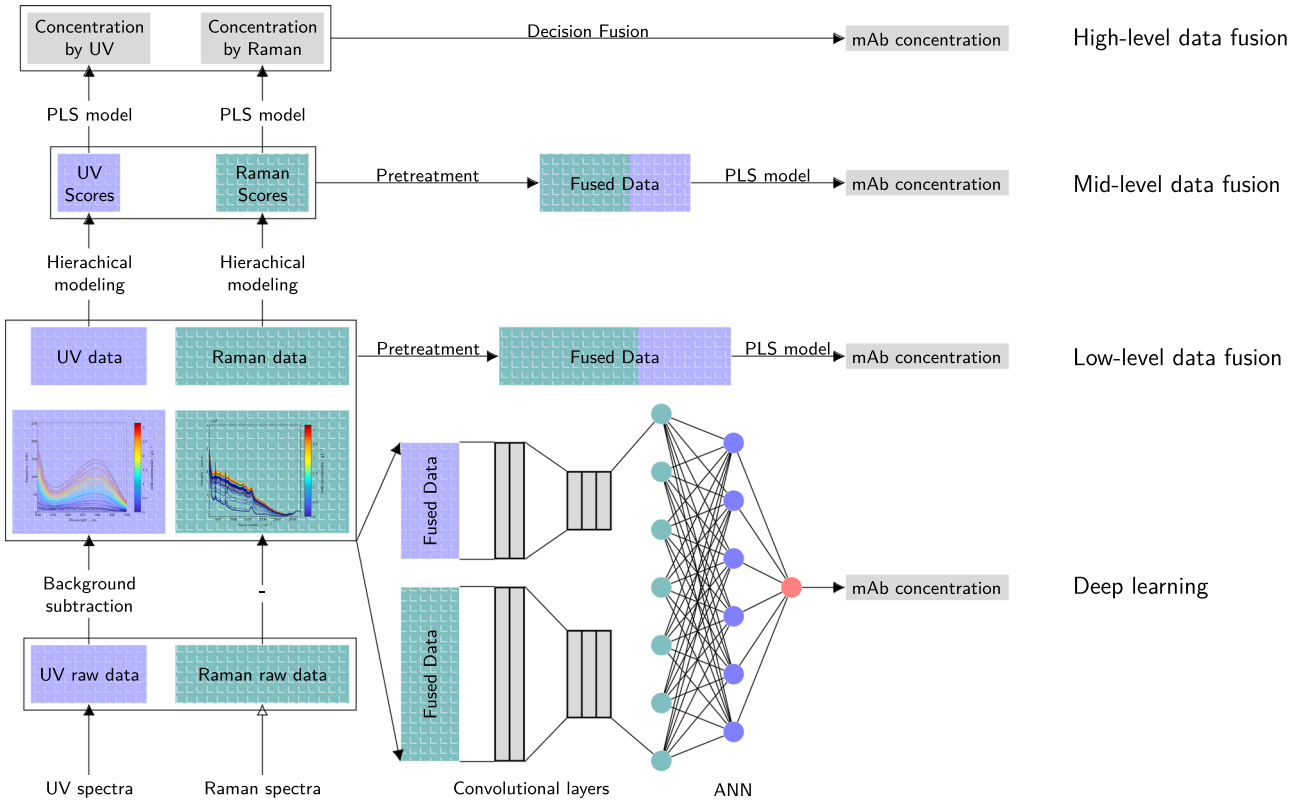

FIGURE 2 Methodology for the applied model building in low-level, midlevel, and high-level data fusion and, additionally, deep learning

Split Sampler FT, Column Compartment H, and a Diode Array Detector HL. Chromeleon Version 7.2 SR4 (Thermo Fisher Scientific) was used to control the HPLC. The collected fractions of all runs were examined by analytical Protein A chromatography to obtain the mAb concentrations. For each sample, a 2.1 mm × 30 mm POROS prepacked Protein A column (Applied Biosystems) was equilibrated with 2 CVs of equilibration buffer, followed by an injection of 20 µl of sample. The column was then equilibrated with 0.8 CVs of equilibration buffer and eluted with 1.4 CVs of elution buffer. The flow rate was 2 ml min^{-1} for all phases and experiments.

Column equilibration was carried out using a buffer with 10 mM phosphate (from sodium phosphate and potassium phosphate) with 0.65 M chloride ions (from sodium chloride and potassium chloride) at pH 7.1. Elution was performed with the same buffer, but titrated to pH 2.6 with hydrochloric acid. All buffer components were purchased from VWR. The buffers were prepared with Ultrapure Water (PURELAB Ultra, ELGA LabWater, Viola Water Technologies), filtrated with a cellulose acetate filter with a pore size of 0.22 µm (Pall), and degassed by sonification.

2.4 | Data analysis

Figure 2 shows an overview of the applied data analysis. First, the sensor signals were gathered and combined with the mAb concentration. For the UV and Raman spectra, various types of preprocessing were evaluated by two-block PLS modeling. Subsequently, the best preprocessing technique was applied to the raw data resulting in the data used for both data fusion by PLS modeling and CNN regression. These data were concatenated and pretreated for low-level data fusion by PLS modeling. Additionally the data were used to build the base PLS model for each spectroscopic technique. From the base models, the scores were concatenated and pretreated for midlevel data fusion by PLS modeling. Additionally, the predictions of the hierarchical models were taken for decision fusion PLS modeling for high-level data fusion. Further details on the raw data analysis, PLS model calibration and evaluation, and CNN training are given below.

2.4.1 | Raw data analysis

The recorded Raman and UV spectra, the measured mAb concentration by analytical chromatography, and run data from the Äkta system were read in and processed with MATLAB R2019b (The MathWorks Inc.). A background subtraction to remove the influence of contaminants on the spectra was evaluated for both spectra sets as described in Rolinger et al. (2020b). After the background subtraction, the spectra were averaged according to the fraction size data from the Äkta. For the calibration/training of the different models, Runs 1, 2, 4, and 5 were used as calibration data set. Run 3 was always used as external validation, because it is the center point of the design space.

2.4.2 | PLS modeling

For the calibration of PLS models, SIMCA 13.0.3 (Sartorius) was used. SIMCA applies a 7-fold cross-validation as internal validation, by splitting the calibration data set into seven parts and leaving each part out of the calibration once. SIMCA applies the Nonlinear Iterative Partial-Least-Squares (NIPALS)-algorithm for PLS model building (Eriksson et al., 2006a). For the UV-based model, no spectral preprocessing was done except the previously explained subtraction of the background. All spectra and the mAb concentration were pretreated by mean-centering. The resulting model was chosen as the base model for all PLS-based data fusion efforts.

For the Raman-based models, first, different spectral preprocessing steps were evaluated to improve the model prediction and linearity during calibration. This involved the use of an Extended Multiplicative Signal Correction (EMSC) filter, first and second derivation, baseline removal, and a background subtraction. Additionally, the different spectral preprocessing options were compared in Solo 8.9 (Eigenvector Research Inc.) with the optimization tool. After the evaluation of different preprocessing options, the best Raman model was chosen as base data along with the UV model for comparing the prediction quality and data fusion purposes.

Often data fusion is grouped into three different levels, namely, low-level, midlevel, and high-level data fusion (Borràs et al., 2015; Cocchi, 2019). In this study, the results of the different fusion levels will be compared with each other. Low-level data fusion is the concatenation of the preprocessed UV and Raman spectra. Midlevel data fusion refers to additional variable selection before the concatenation of the spectra. In this study, hierarchical PLS modeling will be used as the main variable selection technique. With hierarchical PLS modeling, the score vectors of the base model are taken as input variables, also referred to as "super variables," for a new PLS model (Wold et al., 1996). For high-level data fusion, an output fusion of the base PLS models was carried out by hierarchical PLS modeling.

The basis for successful data fusion is proper data alignment (Liggins et al., 2017). Here, both data sets were already aligned timewise and averaged according to the collected fractions before preprocessing or concatenation. Due to the two-dimensional nature of the UV and Raman spectra, no dimension reduction before concatenation was necessary. However, the UV and Raman spectra differ in the number of variables and in the total value of the variables. To prevent the greater influence of one data set onto the model by either the total value of the variables or the number of variables in the data set, proper scaling is important (Eriksson et al., 2006a).

The preprocessing methods used in this study are mean-centering, unit variance scaling, and Pareto scaling. Mean-centering performs a subtraction of the mean value of a signal $\bar{x}_j$ (Equation 1) from the measured values x_{ij} with i being the sample number and j being the signal number. In case of unit variance scaling, the mean-centered value is divided by the standard deviation of the signal s_j (Equation 2) to account for any difference in the signal variance. Pareto scaling is an intermediate between mean-centering and unit

variance scaling, as the mean-centered values are divided by the square root of the standard deviation s_j (van den Berg et al., 2006).

$$\bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}, \tag{1}$$

$$s_j = \sqrt{\frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}{n - 1}}, \tag{2}$$

$$\text{Center } \hat{x}_{ij} = x_{ij} - \bar{x}_j, \tag{3}$$

$$\text{Unit variance } \hat{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \tag{4}$$

$$\text{Pareto } \hat{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}. \tag{5}$$

2.4.3 | CNN

The neural networks were built in Python version 3.6 (Python Software Foundation) using NumPy version 1.18.5 (Harris et al., 2020), pandas version 1.0.5 (McKinney, 2010), and TensorFlow version 2.2.0 (Abadi et al., 2015) as libraries. For all models, a hyperparameter optimization was done via Bayesian optimization (Keras Tuner, version 1.0.1; O'Malley et al., 2019).

The structure of the used CNNs may be broadly split into convolutional blocks and a fully connected block. Every convolutional block consisted of a convolutional layer, a pooling layer, and a dropout layer. The number of such convolutional blocks was optimized in the range from 1 to 3 and from 1 to 2 for the Raman- and UV-based model, respectively. The window width of the first convolutional layer was allowed to change from 60 to 130 for the Raman-based model and from 4 to 30 for the UV-based model. To initialize the kernel of the first convolutional layer of the Raman model, a first and second derivative Gaussian wavelet was used. Thereafter, a dense layer with 1–52 neurons was optimized. Swish was used as an activation function (Ramachandran et al., 2017). As beta was not specified, Swish is equivalent to a Sigmoid-weighted Linear Unit. The output layer was fixed with one densely connected neuron with a leaky rectified linear unit (ReLu) activation function (alpha of 0.1) and a bias. This was chosen due to the linearity of the ReLu function in the positive domain and the attenuation of negative values. The weights of the neurons were optimized with Adaptive Moment Estimation (Adam; Kingma & Ba, 2017). The learning rate of Adam optimizer was a further hyperparameter varied by Bayesian optimization. As loss function Mean Square Error (MSE) was used.

For the combined Raman and UV-based CNN model, only a hyperparameter optimization of an additional dense layer on top of the individual dense layers was done to combine both models. Bayesian optimization was used again with a range between 12 and 64 neurons in the dense layer and the same conditions on the learning rate as for single sensor models

3 | RESULTS AND DISCUSSION

This paper focuses on a comparison of UV- and Raman-based monitoring of the Protein A breakthrough as well as the evaluation of data fusion techniques for both sensor signals. UV data were preprocessed as described by Rolinger et al. (2020b), which leads to a significantly improved prediction as it suppresses absorption from interfering co-eluting species. For an analysis of the UV spectra during the load phase, a comparison to elution spectra, and a detailed discussion on the effects of the preprocessing, we refer to Rolinger et al. In the following, the focus is set towards an analysis of the Raman spectra and the comparison of the prediction quality based on UV- and Raman-based models. First, the observable features of Raman spectra will be analyzed followed by a discussion on the performance of the different PLS models for the Raman spectra and data fusion. Finally, the results from the CNN models are introduced and discussed for the individual sensors and the fused data.

3.1 | Raman spectra

Figure 3 shows the Raman raw spectra, the first and second derivatives colored according to mAb concentration. For further data analysis, only the raw spectra were used. The first and second derivatives are plotted to show the influence of the background removal on the spectra. It is interesting to note that the raw spectra show an underlying baseline effect that increases with increasing run time. The intensity of this effect varies for every feedstock. The background spectra for each run are shown in the Supporting Information Data B. Therefore, when looking at multiple runs, the raw spectra are not primarily sorted by mAb concentration but rather by run-specific baseline effects. For every individual run, a trend of increasing baseline with increasing run time after the impurity breakthrough is apparent. Within each run, the baseline increase is visually the strongest effect over the run time in the spectra. The first derivative mostly removes the baseline effects except for the steep increase below 400 cm^{-1}. The second derivative removes the baseline effect completely. However, it also becomes obvious that very little change remains in the spectra after removal of the baseline by derivation. Additionally, the signal-to-noise ratio is decreased by the derivation.

In Figure 4 the Raman spectra over the course of Run 2 are plotted to show the formation of the Raman bands over the process time. The most prominent effect, which also partly correlates with the mAb concentration, is the increase in background scattering. The spectrum with the lowest overall intensities is the first spectrum of the run, where only buffer is measured. The sapphire band at 418 cm^{-1} is the strongest band in the spectrum. No wavenumber-dependent intensity correction was performed. Otherwise the water bands around 3000 cm^{-1} would be more prominent as well. Proteins have low Raman scatter efficiencies (Rolinger et al., 2020a), which makes the contribution of water in the spectrum more prominent. The strongest protein bands seem to be caused by phenylalanine (1006 cm^{-1}), tryptophan (1360 cm^{-1}), CH deformations

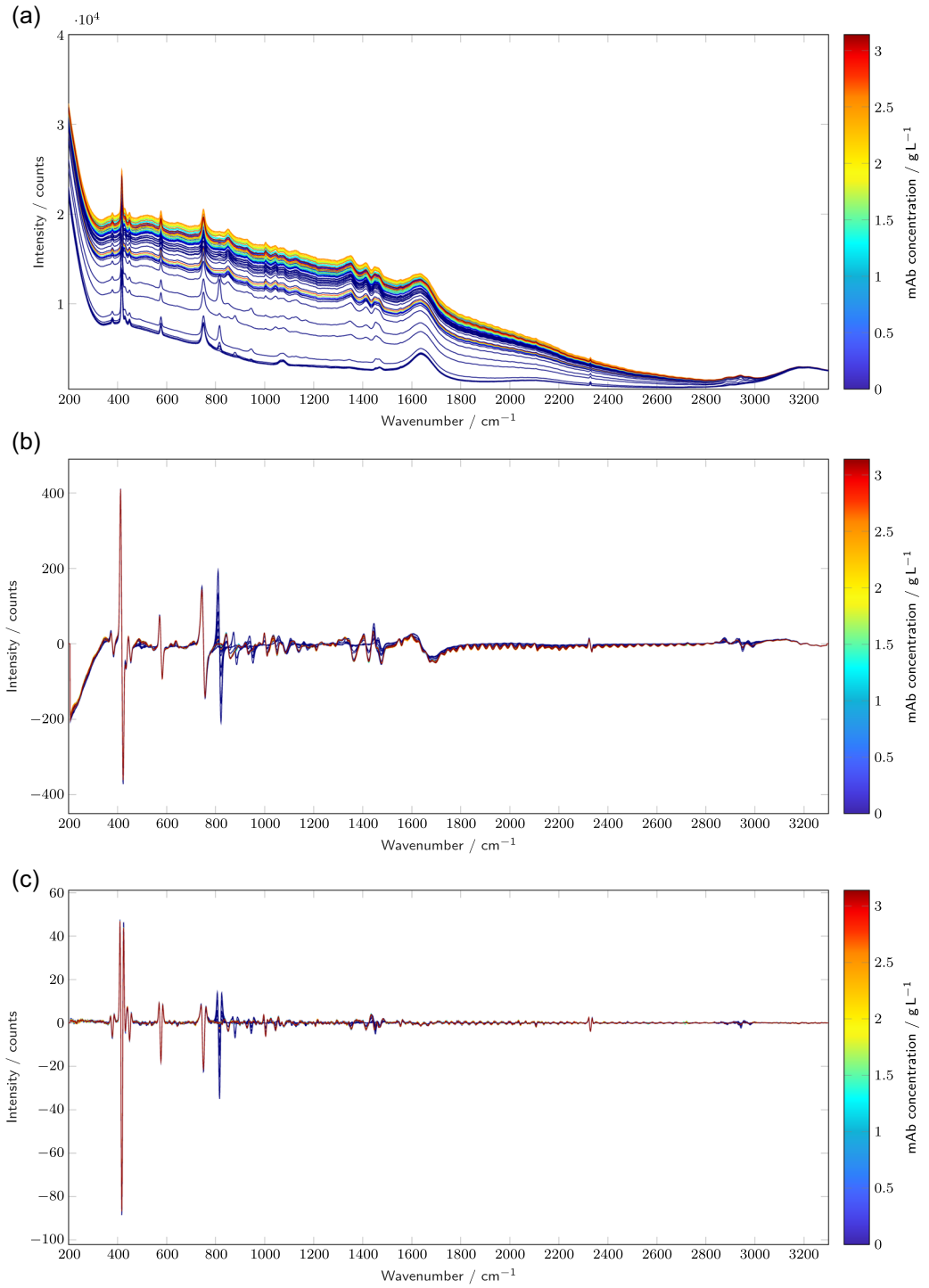

FIGURE 3 The raw (a), first derivative (b), and second derivative (c) spectra of the calibration runs. The spectra are colored by mAb concentration

(1421 and1468 cm^{-1}; Rygula et al., 2013; Silveira et al., 2019) and C–H stretching at 2952 cm^{-1} (Jiskoot & Crommelin, 2005). Overall, with increasing run time there are more weak protein-based peaks present in the spectral range 500–1700 cm^{-1}, which are corrupted by noise.

Jiskoot et al. estimate the limit of quantification for proteins in aqueous solutions to range between 1% and 5% (Jiskoot & Crommelin, 2005) which corresponds to a concentration 10–50 g L^{-1}. Wen et al. claim that therapeutical proteins can be quantified from 1 g L^{-1} due to significant instrument improvements (Wen, 2007). From the shown spectra, it seems that a quantification to lower concentrations is possible with our setup. In general, the quantification does not seem to rest on features generated by the protein backbone, that is, the amid bands, but rather on bands related to

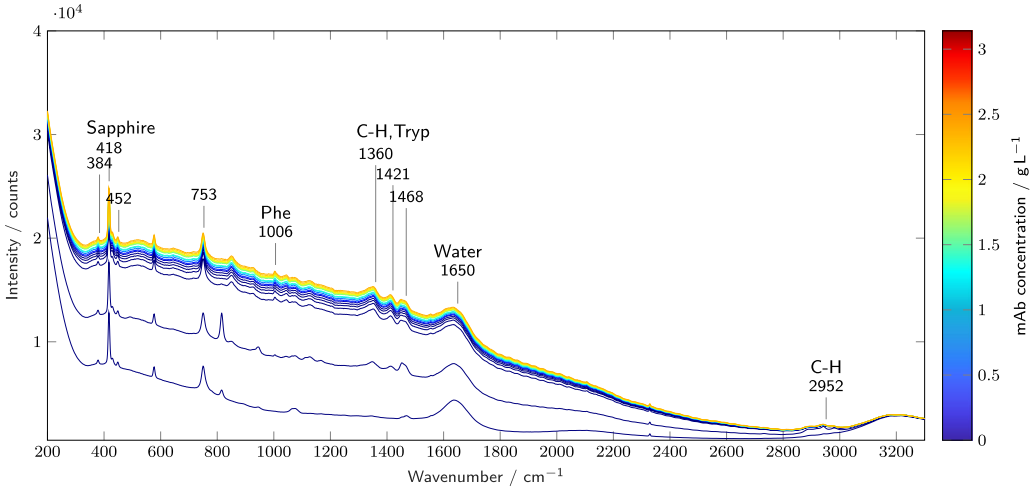

FIGURE 4 Every 10th Raman spectrum of Run 2 is plotted and colored by the mAb concentration. The prominent bands in the spectra are assigned to the generating species sapphire glass, water, buffer, and protein

aromatic groups and C–H vibrations. A selective quantification by Raman spectroscopy between different protein species, based on other protein structure elements than aromatic groups and C–H vibrations, in the investigated concentration range seems difficult due to the low signal-to-noise ratio of the amide bands.

Figure 5 compares the raw signals of UV absorption at 280 nm with the Raman intensity at 400 cm^{-1} over the run time. At a wavenumber of 400 cm^{-1}, no relevant Raman scattering of proteins exists (Rygula et al., 2013; Wen, 2007), that is, any change may be considered a background effect. A distinct increase over the process run time is visible for the Raman intensity similar to the trend of the UV absorption. This background effect is sometimes attributed to the fluorescence of cell culture components (Goldrick et al., 2020; Whelan et al., 2012). However, the same background effect is seen in aqueous protein solutions with increasing protein concentration (Parachalil et al., 2018). As the intrinsic protein fluorescence does not reach above 500 nm, the observed background effect is probably not caused by fluorescence (Lakowicz, 2013). It seems more likely that Rayleigh's scattered light is the incomplete blocking of the Rayleigh scattered light by the notch filter and optical grating (Parachalil et al., 2018). The increase in scattered light could also be attributed to the change in refractive index, which is correlated to protein concentration. During the load phase, impurities with large molecular weight (e.g., deoxyribonucleic acid [DNA] and Host Cell Proteins [HCPs]) elute from the column and lead to an increased amount of Rayleigh scattering, before the mAb breaks through.

3.2 | Comparison of UV- and Raman-based PLS models

For the UV-based PLS model, it was previously established that a background subtraction significantly improves the precision of the UV-based PLS model (Rolinger et al., 2020b; Rüdt et al., 2017). On the basis of the high quality of the prediction, the conductivity-based background subtraction was chosen as preprocessing. No further preprocessing was performed for the UV spectra.

For the calibration of the Raman-based PLS model, different preprocessing methods were evaluated. The model with the best calibration results by cross-validation was chosen as base model. The tested preprocessing methods were conductivity-based background subtraction, derivatives, and baseline removal by extended multiplicative scatter correction and asymmetric Whittaker smoothing. However, the raw data provided the best results during cross-validation. This could be caused by the noise increase in the data due to a subtraction of a noisy background spectrum or due to the amplification of noise by derivation, respectively. It is also interesting, that a baseline removal did not yield a better model compared with the raw data. Apparently, the PLS model uses the background scattering effect to improve the prediction quality.

In Figure 5, the calibration results of the UV-based and the Raman-based PLS models are plotted and compared with the reference analytics. Additionally, as discussed in Section 3.1, the UV absorption at 280 nm and the Raman intensity at 400 cm^{-1} are compared. The results of the UV-based and Raman-based PLS models are listed in Table 2.

The UV-based PLS model has a better prediction accuracy with a higher coefficient of determination R^2, a higher coefficient of determination during cross-validation Q^2, and a lower Root Mean Square Error of Cross-Validation (RMSECV). Regarding the Root Mean Square Error of Prediction (RMSEP), the difference between the models is even more pronounced. The RMSEP of the UV-based PLS model is 0.013 g L^{-1} while it is 0.232 g L^{-1} for the Raman-based PLS model. In Figure 6, the model predictions are depicted. The UV-based model prediction and the reference mAb concentration show only minimal differences. The Raman-based prediction shows an offset to the reference mAb concentration. Additionally, the difference between prediction and measured concentration increases starting at a mAb concentration higher than

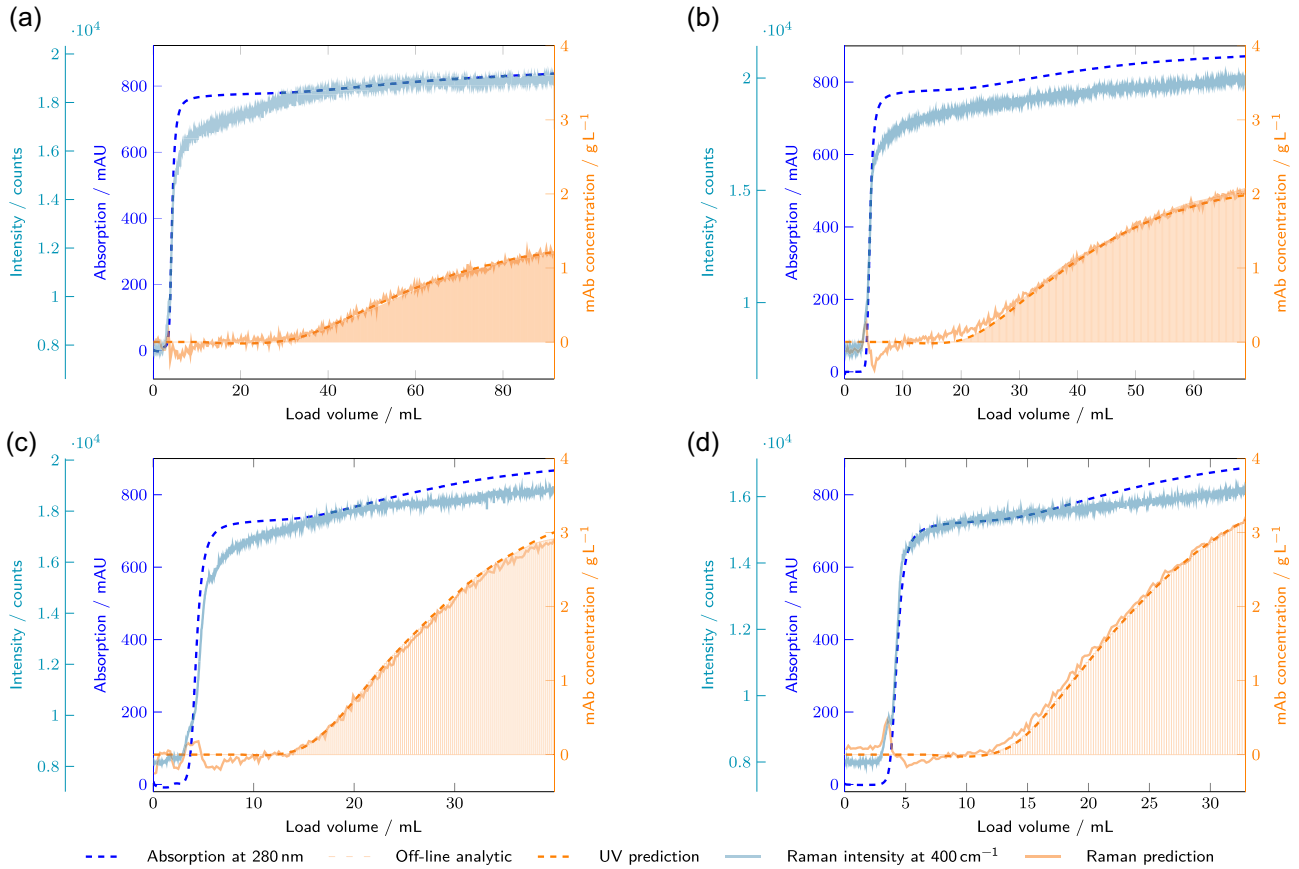

FIGURE 5 Results of the PLS model calibration for Raman and UV-based PLS models. The UV absorption at 280 nm A_{280} (displayed as dashed blue line) and Raman intensity at 400 cm^{-1} (displayed as a solid cerulean line) are compared with the results of the off-line analytics for mAb quantification (orange bars). The UV-based PLS model prediction is illustrated as dashed orange line. The Raman-based PLS model prediction is illustrated as orange line. The four runs exhibited variable mAb titers in the feed (a) 1 g L^{-1}, (b) 1.5 g L^{-1}, (c) 2.5 g L^{-1}, and (d) 3 g L^{-1}. mAb, monoclonal antibody; PLS, Partial-Least-Square; UV, ultraviolet

1.9 g L^{-1}. This seems to be a nonlinear behavior. When looking at the loadings of the Raman-based PLS model, the first loading has a high similarity to the background effect and the following loadings show protein bands. It seems, that the PLS model uses both the background effect and the protein bands to estimate the mAb concentration. Even though the background effect increases with increasing mAb concentration, the background effect alone cannot be used as a sole predictor for the mAb concentration in this data set, because the initial intensity of the background spectrum depends on the feedstock composition. The use of the background effect, which has an offset between the different runs, could impede the linearity between spectra and protein concentration. The deviation from the linearity between concentration and certain Raman peaks could also be caused by the measurement with the ball probe, the influence of the refractive index when protein concentration is increasing or inhomogeneities in the sample flow in the flow cell.

In the performed experiments, the RMSEPs of both PLS models are expected to be comparable with the RMSECV or lower, because the validation run lays in the middle of the calibration design space. For the Raman-based model, the RMSEP is, however, higher compared with RMSECV, which indicates an overfitting as the validation run should be in the center of the design space. The increased RMSEP of the Raman-based model could be caused by the relatively high number of seven Latent Variables (LVs) in comparison to two LVs used by the UV-based PLS model.

It is also worth noting that the prediction of the Raman-based model appears to be more corrupted by white noise (less precise) than the prediction of the UV-based model. This indicates that the Raman-based prediction is more strongly affected by measurement noise than the UV-based predictions. Improvements in measurement quality of the Raman spectra could thus potentially improve the prediction quality.

Additionally, the correlation of prediction of the Raman-based model and mAb reference concentration starts to deviate from the linear relation, especially for Run 3 and mAb concentration above 1.9 g L^{-1} (see also Supporting Information Data D for an observed vs. predicted plot). The UV-based model shows only very little deviation from the linear relation, probably caused by errors in the reference analytic. The stronger deviation from the linear correlation of the Raman-based model could explain why a higher number of LVs is necessary for the Raman-based model in comparison to the

TABLE 2 Input data, data fusion level, scaling, block scaling, R^2, Q^2, RMSEC, RMSECV, RMSEP, and number of LVs for the PLS models

Input data	Data fusion level	Hierarchical level	Scaling	Block scaling	R^2	Q^2	RMSEC (g L^{-1})	RMSECV (g L^{-1})	RMSEP (g L^{-1})	Number of LVs
UV	–	Base	Center	–	0.999	0.999	0.025	0.025	0.013	2
Raman	–	Base	Center	–	0.992	0.992	0.073	0.076	0.232	7
Both	Low	–	Center	–	0.986	0.986	0.100	0.101	0.290	6
Both	Low	–	Pareto	–	0.976	0.976	0.129	0.129	0.092	4
Both	Low	–	Unit var.	–	0.999	0.999	0.025	0.025	0.044	5
Both	Low	–	Center	1/sqrt	0.987	0.987	0.096	0.096	0.155	4
Scores	Mid	Top	Center	–	0.976	0.975	0.013	0.131	0.433	4
Scores	Mid	Top	Pareto	–	0.986	0.986	0.100	0.100	0.313	3
Scores	Mid	Top	Pareto	1/sqrt	0.990	0.990	0.082	0.082	0.231	3
Scores	Mid	Top	Unit var.	–	0.998	0.998	0.040	0.040	0.118	1
Scores	Mid	Top	Unit var.	1/sqrt	0.998	0.998	0.040	0.040	0.129	2
Output	High	Top	Center	–	0.998	0.998	0.040	0.040	0.129	1
Output	High	Top	Unit var.	–	0.998	0.998	0.040	0.040	0.129	1

Abbreviations: LV, Latent Variable; PLS, Partial-Least-Square; RMSEC, Root Mean Square Error of Calibration; RMSECV, Root Mean Square Error of Cross-Validation; RMSEP, Root Mean Square Error of Prediction; UV, ultraviolet.

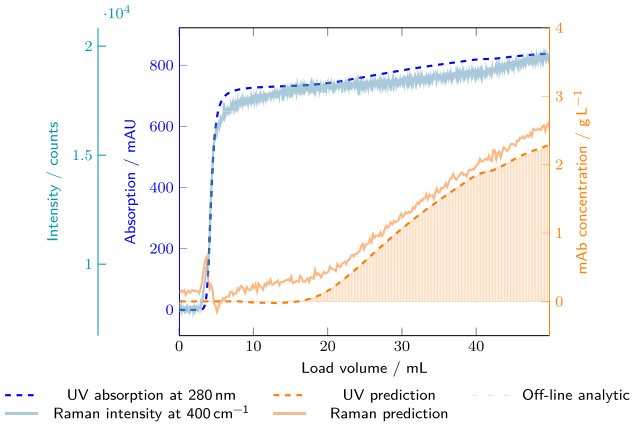

FIGURE 6 Results of the PLS model validation of Run 4 for Raman and UV-based PLS models. The UV absorption at 280 nm A_{280} (displayed as dashed blue line) and Raman intensity at 400 cm^{-1} (displayed as solid cerulean line) are compared with the results of the off-line analytics for mAb quantification (orange bars). The UV-based PLS model prediction is illustrated as dashed orange line. The Raman-based PLS model prediction is illustrated as an orange line. mAb, monoclonal antibody; PLS, Partial-Least-Square; UV, ultraviolet

UV-based model. PLS models can approximate nonlinearities by including additional LVs (Martens & Naes, 1992).

In summary, for the investigated experimental conditions, UV spectroscopy is better suited for monitoring the mAb breakthrough during Protein A chromatography than used Raman spectroscopy setup. The UV-based PLS model reaches a more than 10-fold lower RMSEP compared with the Raman-based PLS model. While there might still be chromatographic capture steps, where a Raman-based PLS model performs better (e.g., high mAb concentration and high variation in UV absorbing background species), the distinctively lower RMSEP of the UV-based model indicates a competitive advantage for most applications involving mAbs. The competitive advantage is further supported by the simpler equipment requirements for UV spectroscopy which may simplify implementation in production environments. Additionally, the used Raman setup might not work for all feedstocks due to autofluorescence (Matthews et al., 2018). The only solution in the case of large autofluorescence is to switch to a longer laser wavelength by using a different equipment. As longer laser wavelengths will cause a weaker Raman signal, the exposure times need to be longer to achieve the same signal-to-noise ratio, which might not be feasible for the typical measurement times in chromatography.

3.3 | Data fusion for UV- and Raman-based PLS models

The results of the different data fusion levels and data pretreatments are compared in Table 2. For low-level data fusion, both spectra were scaled individually and block scaling was eventually applied. With only mean-centering, an RMSEP of 0.290 g L^{-1} is achieved in comparison to an RMSEP of 0.092 g L^{-1} with Pareto scaling and an RMSEP of 0.044 g L^{-1} with unit variance scaling. When comparing the results of the low-level data fusion models without block-scaling, it is noticeable, that the less influence the Raman data have on the model prediction, the better the fused model gets. This is expected as the solely UV-based model has better performance than the corresponding Raman model. Without scaling, the Raman spectra reach intensities of more than 30,000 counts in comparison to the around 200 mAU reached by the UV spectra. The absolute change in variables of the Raman spectra is larger as well due to the scale of the spectra. When only applying mean-centering, this larger variance in the Raman spectra biases the PLS model to mostly include Raman-based signals into the first LVs (i.e., the high variance variables).

In contrast to mean-centering, unit variance scaling additionally divides each variable by their standard deviation. Therefore, the scale of the variables gets removed. The advantage of unit variance scaling is, that not a few variables dominate the total variance of all variables. Thus, also variables with smaller variance and a good correlation to the response may become relevant for model building. The disadvantage of the unit variance scaling is the noise inflation, which usually reduces the performance of PLS models (van den Berg et al., 2006). Pareto scaling is an intermediate between mean-centering and unit variance scaling as variables are scaled by the square root of the standard deviation. When little is known about the importance of the different blocks for the response prediction, unit variance scaling seems a good option even though a less accurate model is achieved than by only using the UV block for prediction.

As the Raman spectra have 3101 variables in comparison to the UV spectra with 171 variables, the contributed variance of the Raman spectra to the complete X block is larger even after unit variance scaling. To avoid this bias after preprocessing, the different blocks can be multiplied by different weights. These weights typically consist of a term to make the scale of the different blocks more even. Here, the mean-centered blocks were scaled by the reciprocal square root of the number of variables in each block (Eriksson et al., 2006b). By block scaling, the RMSEP of 0.290 g L^{-1} of the mean-centered model was lowered to 0.155 g L^{-1} as a large number of variables from the Raman spectrum had less influence on the prediction.

As an approach for midlevel data fusion, hierarchical PLS modeling was chosen. In hierarchical modeling, the individual spectra are multiplied by the loadings of each LV to calculate the scores of each spectrum. The different loadings of the UV- and Raman-based PLS model are displayed in the Supporting Information Data C. When using hierarchical modeling, the same consideration for the scaling are necessary as in low-level data fusion. Again, as with low-level data fusion, the closer the scores are scaled to unit variance, the lower the RMSEP becomes. With only mean-centering and midlevel data fusion, an RMSEP of 0.433 g L^{-1} is achieved in comparison to an RMSEP of 0.313 g L^{-1} with Pareto scaling and an RMSEP of 0.118 g L^{-1} with unit variance scaling. Interestingly, the RMSEPs of the unit variance scaled and Pareto scaled midlevel data fusion models are higher than the original RMSEP of the Raman-based PLS model. An explanation for this could the low linearity of the Raman spectrum with regard to the mAb concentration. The Raman-based model uses the background effect to a certain degree to allow for a better prediction. With midlevel data fusion, the number of LVs are generally lower and an approximation of the nonlinearities is more difficult, because fewer colinear parameters are available for the fit.

High-level data fusion was realized as output fusion in this study, where the predictions of the base models were fused by a PLS model. In the case of output fusion, the scaling of the variables is not important as they are already on the same scale. Therefore, different scaling methods, have the same result in our case. An RMSEP of 0.118 g L^{-1} is achieved. This RMSEP is almost the average of the two base models with leveraging the UV-based model more due to a regression coefficient of 0.503 in comparison to 0.497. As an alternative to PLS, other techniques, like, Bayesian belief networks could be used as well.

We conclude, that the best way of optimizing a prediction is to choose the right sensor from the start (Andersen & Bro, 2010; Hall & Steinberg, 2001). For the purpose of monitoring the mAb concentration in the effluent of a Protein A column, UV spectroscopy is better suited than Raman spectroscopy due to a higher sensitivity and better linearity. Often the limited selectivity of UV spectroscopy is mentioned as a drawback, but for this application case the sensitivity seems to be no issue possibly due to the applied background subtraction. Even though data fusion has been reported as a useful tool, when combining a good sensor with a sensor with limited observation ability of the effect in focus, data fusion can do very little beyond the capacity of the best sensor. We therefore would like to issue a word of caution on the application of data fusion for data sets with poor sensors or without understanding the possible benefit of data fusion. Even though we have seen an increasing body of literature where data fusion is applied (Felfödi et al., 2020; Sauer et al., 2019; Walch et al., 2019), data fusion methods should be considered skeptically. If a sensor cannot quantify a concentration on its own, a fusion with a different sensor will likely not lead to meaningful results in regression. The risk of coincidental correlations and overfitting is increased. In our case, the prediction was always worse when combining UV and Raman spectra than the UV-based prediction alone. A solution could be the application of nonlinear models, like, Artificial Neural Networks (ANNs) to improve the prediction ability of the Raman models and thereby the accuracy of the fusion models.

3.4 | CNNs for UV and Raman data

Table 3 shows the hyperparameters after the Bayesian optimization.

Even though the UV-based CNN and Raman-based CNN were given similar boundaries for the optimization, the optimum of the UV-based CNN has less convolutional layers, less filters, and smaller window widths, which implies that less data 'preprocessing' is required for the UV-based CNN. The first convolutional layer in the Raman-based CNN was initialized by wavelets which imitate a first and second derivation. Otherwise the optimization did not converge on an optimum of comparable quality as a PLS model. The output of the convolutional layers for the UV- and Raman-based model are displayed in Section E. Figure 7 shows the predictions of the UV-based, Raman-based, and combined CNN model for the external validation run.

Table 4 lists the Root Mean Square Error of Calibration (RMSEC) and RMSEP of the CNN models. The UV-based CNN predicts the mAb concentration accurately with an RMSEP of 0.013 g L^{-1}. The Raman-based CNN has a prediction, which is more corrupted by noise in comparison to the UV-based CNN. The higher RMSEP of 0.220 g L^{-1} is not only caused by the increased noise, but also by an offset. Both CNNs deliver comparable results to the base PLS models. The CNN with the combined data had 21 neurons in the additional fully connected layer after optimization. With this, an RMSEP of 0.050 g L^{-1} was reached. The CNN with the combined data lays between the results of the individual models with regard to noise in the prediction and RMSEP.

For the presented study, the use of CNNs in comparison to PLS models only offers a limited benefit. The training of CNNs needs more resources and wrong setting of the initial start conditions can lead to a divergence of the training. In our case, the training set with 1169 training spectra was bigger compared with usual spectroscopic training sets. A lower amount of training spectra will probably cause problems for CNNs due to the high number of parameters.

TABLE 3 Hyperparameter found by Bayesian optimization for the Raman and the UV-based CNNs

Hyperparameter	Raman	UV
Number of convolutional layers	3	2
Window width convolutional layer zero	90	4
Pooling width convolutional layer zero	11	1
Number of filters in convolutional layer zero	8	2
Window width convolutional layer one	16	6
Pooling width convolutional layer one	1	1
Number of filters in convolutional layer one	8	8
Window width convolutional layer two	28	–
Pooling width convolutional layer two	1	–
Number of filters in convolutional layer two	6	–
Number of neurons in fully connected layer	46	31
Learning rate	0.001	0.001

Abbreviations: CNN, Convolutional Neural Network; UV, ultraviolet.

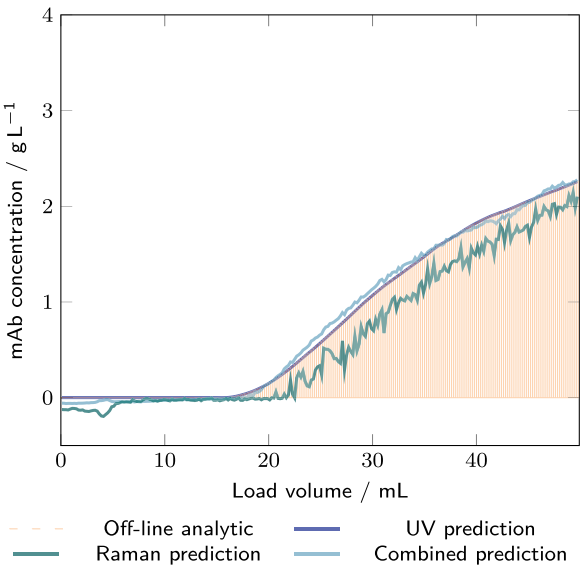

FIGURE 7 Results of the CNN model validation of Run 4 for the Raman, UV-based and combined CNN models. The UV-based model prediction (displayed as solid blue line), the Raman-based model prediction (displayed as solid teal line), and the combined model prediction (displayed as solid cerulean line) are compared with the results of the off-line analytics for mAb quantification (orange bars). CNN, Convolutional Neural Network; mAb, monoclonal antibody; UV, ultraviolet

TABLE 4 RMSEC, RMSEP of the Raman, UV-based and combined CNNs

Input data	RMSEC (g L^{-1})	RMSEP (g L^{-1})
UV	0.019	0.013
Raman	0.078	0.220
both	0.047	0.050

Abbreviations: CNN, Convolutional Neural Network; RMSEC, Root Mean Square Error of Calibration; RMSEP, Root Mean Square Error of Prediction; UV, ultraviolet.

4 | CONCLUSION AND OUTLOOK

In this study, Raman and UV spectroscopy have been compared in their ability to predict the mAb concentration in the column effluent during the load phase of the Protein A capture step. Additionally, data fusion strategies based on PLS models and CNNs were presented and compared with the single sensor models.

We conclude that UV spectroscopy achieves a better prediction accuracy in comparison to Raman spectroscopy. UV- and Raman-

based PLS models required two, respectively, seven LVs. The high number of LVs of the Raman-based PLS model may be related to nonlinearities, which are more difficult to fit by the linear PLS model. Of all fusion approaches, no model was better than the simple UV PLS model or the corresponding CNN model, which both achieved an RMSEP of 0.013 g L^{-1}. Data fusion for regression purposes seems not to be beneficial, if one sensor already provides a very good accuracy and an additional sensor could only contribute noise. For Raman spectroscopy, the application of CNNs in comparison to traditional PLS models improved the prediction of the mAb concentration from 0.232 g L^{-1} (PLS model) to 0.220 g L^{-1}. The training and optimization of CNNs for both UV and Raman data was time-consuming. The success was dependent on establishing proper boundaries and starting conditions for model optimization. In our opinion, it seems generally not worth the effort to apply nonlinear models to the monitoring of the mAb breakthrough, because a similar prediction accuracy can be reached with traditional PLS models (Kjeldahl & Bro, 2010).

For future technology evaluations for the implementation of real-time monitoring of the Protein A capture step, we consider UV spectroscopy to have a competitive advantage compared with Raman spectroscopy due to the better prediction quality and the simpler equipment. Raman spectroscopy may be of interest, if alternative chemicals should be monitored in the column effluent which does not have a UV absorption.

ACKNOWLEDGMENTS

Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

AUTHOR CONTRIBUTIONS

L. Rolinger designed the study, carried out the experiments, analyzed the data, and wrote the manuscript. M. Rüdt contributed to the study design and supported the data analysis and writing of the manuscript. J. Hubbuch supervised the project and reviewed the manuscript.

ORCID

Laura Rolinger http://orcid.org/0000-0002-6061-654X
Matthias Rüdt http://orcid.org/0000-0002-8583-6645

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., …Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems*. Software available from tensorflow.org. https://www.tensorflow.org/

Abu-Absi, N. R., Kenty, B. M., Cuellar, M. E., Borys, M. C., Sakhamuri, S., Strachan, D. J., Hausladen, M. C., & Li, Z. J. (2011). Real time monitoring of multiple parameters in mammalian cell culture bioreactors using an in-line Raman spectroscopy probe. *Biotechnology and Bioengineering*, *108*(5), 1215–1221.

Andersen, C. M., & Bro, R. (2010). Variable selection in regression—A tutorial. *Journal of Chemometrics*, *24*(11–12), 728–737.

Biancolillo, A., Liland, K. H., MÅge, I., Næs, T., & Bro, R. (2016). Variable selection in multi-block regression. *Chemometrics and Intelligent Laboratory Systems*, *156*, 89–101.

Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L., & Busto, O. (2015). Data fusion methodologies for food and beverage authentication and quality assessment—A review. *Analytica Chimica Acta*, *891*, 1–14.

Brestrich, 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.

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.

Buckley, K., & Ryder, A. G. (2017). Applications of Raman spectroscopy in biopharmaceutical manufacturing: A short review. *Applied Spectroscopy*, *71*(6), 1085–1116.

Cocchi, M. (2019). *Data fusion methodology and applications*. Elsevier.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikström, C., & Wold, S. (2006a). *Multi- and megavariate data analysis* (Vol. 1). Umetrics Ab Umea.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikström, C., & Wold, S. (2006b). *Multi- and megavariate data analysis: Part II: Advanced applications and method extensions*. Umetrics Inc.

Feidl, F., Garbellini, S., Luna, M. F., Vogg, S., Souquet, J., Broly, H., Morbidelli, M., & Butté, A. (2019). Combining mechanistic modeling and Raman spectroscopy for monitoring antibody chromatographic purification. *Processes*, *7*(10), 683.

Feidl, F., Garbellini, S., Vogg, S., Sokolov, M., Souquet, J., Broly, H., Butté, A., & Morbidelli, M. (2019). A new flow cell and chemometric protocol for implementing in-line Raman spectroscopy in chromatography. *Biotechnology Progress*, *35*(5), e2847.

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 & Biotechnology*, *95*(4), 1146–1152.

Goldrick, S., Umprecht, A., Tang, A., Zakrzewski, R., Cheeks, M., Turner, R., Charles, A., Les, K., Hulley, M., Spencer, C., & Farid, S. S. (2020). High-throughput Raman spectroscopy combined with innovative data analysis workflow to enhance biopharmaceutical process development. *Processes*, *8*(9), 1179.

Hall, D. L., & Steinberg, A. (2001). *Dirty secrets in multisensor data fusion* (Technical report). Pennsylvania State University Park Applied Research Lab.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Rio, J. F., Wiebe, M., Peterson, P., …Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Jiskoot, W., & Crommelin, D. (Eds.). (2005). *Methods for structural analysis of protein pharmaceuticals*. American Association of Pharmaceutical Scientists.

Kingma, D. P. & Ba, J. (2017). Adam: A method for stochastic optimization. arXiv: 1412.6980

Kjeldahl, K., & Bro R. (2010). Some common misunderstandings in chemometrics. *Journal of Chemometrics*, *24*, 558–564.

Lakowicz, J. R. (2013). *Principles of fluorescence spectroscopy*. Springer Science & Business Media.

Li, B., Ray, B. H., Leister, K. J., & Ryder, A. G. (2013). Performance monitoring of a mammalian cell based bioprocess using Raman spectroscopy. *Analytica Chimica Acta*, *796*, 84–91.

Li, B., Ryan, P. W., Ray, B. H., Leister, K. J., Sirimuthu, N. M., & Ryder, A. G. (2010). Rapid characterization and quality control of complex cell culture media solutions using Raman spectroscopy and chemometrics. *Biotechnology and Bioengineering*, *107*(2), 290–301.

Liggins II, M., Hall, D., & Llinas, J. (2017). *Handbook of multisensor data fusion: Theory and practice*. CRC Press.

Martens, H., & Naes, T. (1992). *Multivariate calibration*. John Wiley & Sons.

Matthews, T. E., Smelko, J. P., Berry, B., Romero-Torres, S., Hill, D., Kshirsagar, R., & Wiltberger, K. (2018). Glucose monitoring and adaptive feeding of mammalian cell culture in the presence of strong autofluorescence by near infrared Raman spectroscopy. *Biotechnology Progress*, *34*(6), 1574–1580.

McKinney, W. (2010). Data structures for statistical computing in Python. In I. Stéfan van der Walt, & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L., et al. (2019). *Keras tuner*. https://github.com/keras-team/keras-tuner

Parachalil, D. R., Brankin, B., McIntyre, J., & Byrne, H. J. (2018). Raman spectroscopic analysis of high molecular weight proteins in solution—Considerations for sample analysis and data pre-processing. *Analyst*, *143*(24), 5987–5998.

Ramachandran, P., Zoph, B. & Le, Q. V. (2017). Searching for activation functions. arXiv preprint arXiv:1710.05941.

Rolinger, L., Rüdt, M., & Hubbuch, J. (2020a). A critical review of recent trends, and a future perspective of optical spectroscopy as PAT in biopharmaceutical downstream processing. *Analytical and Bioanalytical Chemistry*, *412*, 2123–2136.

Rolinger, L., Rüdt, M., & Hubbuch, J. (2020b). A multisensor approach for improved Protein A load phase monitoring by conductivity-based background subtraction of uv spectra. *Biotechnology and Bioengineering*.

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.

Rygula, A., Majzner, K., Marzec, K. M., Kaczor, A., Pilarczyk, M., & Baranska, M. (2013). Raman spectroscopy of proteins: A review. *Journal of Raman Spectroscopy*, *44*(8), 1061–1076.

Saggu, M., Liu, J., & Patel, A. (2015). Identification of subvisible particles in biopharmaceutical formulations using Raman spectroscopy provides insight into polysorbate 20 degradation pathway. *Pharmaceutical Research*, *32*(9), 2877–2888.

Sauer, D. G., Melcher, M., Mosor, M., Walch, N., Berkemeyer, M., Scharl-Hirsch, T., Leisch, F., Jungbauer, A., & 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.

Silveira, L., Pasqualucci, C. A., Bodanese, B., Pacheco, M. T. T., & Zângaro, R. A. (2019). Normal-subtracted preprocessing of Raman spectra aiming to discriminate skin actinic keratosis and neoplasias from benign lesions and normal skin tissues. *Lasers in Medical Science*, *35*, 1141–1151.

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.

van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics*, *7*(1), 142.

Walch, N., Scharl, T., Felföldi, E., Sauer, D. G., Melcher, M., Leisch, F., Dürauer, A., & Jungbauer, A. (2019). Prediction of the quantity and purity of an antibody capture process in real time. *Biotechnology Journal*, *14*(7), 1800521.

Wen, Z.-Q. (2007). Raman spectroscopy of protein pharmaceuticals. *Journal of Pharmaceutical Sciences*, *96*(11), 2861–2878.

Whelan, J., Craven, S., & Glennon, B. (2012). In situ Raman spectroscopy for simultaneous monitoring of multiple process parameters in mammalian cell culture bioreactors. *Biotechnology Progress*, *28*(5), 1355–1362.

Wold, S., Kettaneh, N., & Tjessem, K. (1996). Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *Journal of Chemometrics*, *10*(5–6), 463–482.

Zobel-Roos, S., Mouellef, M., Siemers, C., & Strube, J. (2017). Process analytical approach towards quality controlled process automation for the downstream of protein mixtures by inline concentration measurements based on ultraviolet/visible light (UV/VIS) spectral analysis. *Antibodies*, *6*(4), 24.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: 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*, 1–14. https://doi.org/10.1002/bit.27894