Received: 14 November 2023 | Revised: 29 January 2024 | Accepted: 10 February 2024

DOI: 10.1002/bit.28679

ARTICLE

BIOTECHNOLOGY AND BIOENGINEERING WILEY

Simultaneous prediction of 16 quality attributes during protein A chromatography using machine learning based Raman spectroscopy models

Jiarui Wang[1] | Jingyi Chen[1,2] | Joey Studts[1] | Gang Wang[1]

[1]Late Stage Downstream Process Development, Boehringer Ingelheim Pharma GmbH/Co. KG, Biberach an der Riss, Germany

[2]Bioprocess development and modelling, Karlsruhe Institute of Technology, Karlsruhe, Germany

Correspondence
Gang Wang
Email: gang_3.wang@boehringer-ingelheim.com

Abstract

Several key technologies for advancing biopharmaceutical manufacturing depend on the successful implementation of process analytical technologies that can monitor multiple product quality attributes in a continuous in-line setting. Raman spectroscopy is an emerging technology in the biopharma industry that promises to fit this strategic need, yet its application is not widespread due to limited success for predicting a meaningful number of quality attributes. In this study, we addressed this very problem by demonstrating new capabilities for preprocessing Raman spectra using a series of Butterworth filters. The resulting increase in the number of spectral features is paired with a machine learning algorithm and laboratory automation hardware to drive the automated collection and training of a calibration model that allows for the prediction of 16 different product quality attributes in an in-line mode. The demonstrated ability to generate these Raman-based models for in-process product quality monitoring is the breakthrough to increase process understanding by delivering product quality data in a continuous manner. The implementation of this multiattribute in-line technology will create new workflows within process development, characterization, validation, and control.

KEYWORDS

in-line quality attributes measurement, machine learning, process analytical technology, Raman spectroscopy

1 | INTRODUCTION

Many therapeutic antibodies have demonstrated the ongoing value biopharmaceutical products bring to patients as they continue to address large unmet clinical needs such as the treatment of Alzheimer's disease (Rashad et al., 2022; Reardon, 2023). The emergence of a wide variety of efficacy and safety profiles observed across multiple antibody products targeting amyloid beta underscores the criticality of implementing advanced chemistry manufacturing and controls (CMC) strategies. Regulatory agencies have been pushing for advanced continuous manufacturing capabilities in recent years (ICH, 2021), especially when paired with continuous process analytical technology (PAT) tools to monitor product quality for rapid feedback control (ICH, 2022). Furthermore, advanced development and manufacturing capabilities critical for responding to future pandemics (Kelley, 2020) highlight the urgency

of adopting new technologies for biopharmaceutical manufacturing (Khanal & Lenhoff, 2021).

One key missing technology in the development of continuous and advanced biopharmaceutical processing methodologies is the ability to measure product quality attributes in a rapid and noninvasive manner (Silva et al., 2022), similar to conventional optical methods for measuring UV absorbance. The search for more advanced PAT tools is an ongoing process with a history of several years of research (Gillespie et al., 2022), with Raman spectroscopy being a promising technique (Rolinger et al., 2020). The successful implementation of in-line PAT tools employing a rapid, noninvasive optical sensor capable of measuring multiple product quality attributes would enable the continuous collection of product quality data impacting the development of several key technologies, including continuous manufacturing (Khanal & Lenhoff, 2021), high-throughput process development (HTPD) (Hubbuch, 2012; Silva et al., 2022), and mechanistic bioprocess modeling (Saleh et al., 2020). Conventional offline analytical measurements often require lengthy manual methods that lead to the "low-N" problem (Tulsyan et al., 2019). The growing burden of analyzing increasing numbers of samples created by continuous bioprocessing or automation efforts have contributed to new algorithms being developed (Tulsyan et al., 2020) to accommodate the undesirable yet unavoidable need to calibrate in-process measurement models trained against small numbers of off-line samples (Müller et al., 2023).

Raman spectroscopy has the advantage of being able to analyze several product quality attributes optically (Wei et al., 2021), although this capability has not yet been extended to real-time monitoring of a bioprocessing unit operation. Studies evaluating Raman for in-line process monitoring achieve great results for product concentration (Rolinger et al., 2021), but for other quality attributes such as product aggregation, nuclear magnetic resonance (Taraban et al., 2019) or multiangle light scattering (Patel et al., 2018) techniques have been implemented. Raman spectra are often complex combinations of many fundamental signals (Zhu et al., 2011) and require complex preprocessing steps (Guo et al., 2021) designed for particular use cases, for example, to bridge samples measured under different temperatures and flow rates. Although several studies use some combination of a first-order derivative filter, normalization, and data smoothing (Goldrick et al., 2020; Wei et al., 2021; Yilmaz et al., 2020) for preprocessing, older methods such as those based on signal processing principles (Zhang, Chen, Liang, Liu, et al., 2010) add further factorial possibilities for Raman signal decomposition. Experimentation with different combinations of preprocessing steps is required to maximize the amount of information that can be extracted from raw Raman signals before chemometric modeling (Smith & Dent, 2019).

Designing data processing steps to transform raw Raman signals enables the building of models to correlate in-line measured signals to multiple product quality attributes in real time. This kind of model building is typically done using algorithms leveraging dimensionality reduction, such as partial least squares regression (Feidl et al., 2019; Wei et al., 2021) or principal component regression (Ramakrishna et al., 2022; McAvan et al., 2020). Improvements in laboratory automation hardware allowing the miniaturization of in-line flow cells (Pedro et al., 2021) increased the amount of calibration data available for modeling, enabling the use of more advanced models such as convolutional neural networks (Rolinger et al., 2021) that take advantage of larger data sets without the need for dimensionality reduction. Affinity product capture is usually the first preparative chromatography step in bioprocessing, containing a wide variety of product quality variants as well as other impurities (Pabst et al., 2018). These great initiatives, driven by a mutual vision of the PAT community in the biopharma industry and academia, have resulted in a solid foundation for applying this technology in process development and process control. However, to make a significant impact and difference when implementing Raman-based PAT, it is absolutely necessary to predict all relevant quality attributes simultaneously.

In this paper, we demonstrate, for the very first time, the prediction of 16 quality attributes, covering all relevant quality attributes during the capture step in downstream processing of monoclonal antibodies. In this case, the challenging Protein A chromatography elution pool, which is prone to potential interferences between signals due to high impurity levels, serves as a suitable system for the application of Raman-based models in measuring multiple product quality attributes in real-time during biopharmaceutical manufacturing.

2 | MATERIALS AND METHODS

2.1 | Preparative chromatography

We demonstrate a new PAT during the first preparatory chromatography unit operation in a biopharmaceutical process. Harvested cell culture fluid was used as the starting material and was obtained internally (Boehringer Ingelheim Pharma). Affinity capture—preparative-chromatography was then carried out using the following parameters. Mab Select PrismA resin (Cytiva) was packed using a 2.6 cm diameter column to a height of 19.4 cm, resulting in a column volume (CV) of 103 mL. Column packing and chromatography runs were performed on an AKTA Avant 150 system (Cytiva). A constant flow rate of 20.6 mL/min was used throughout the run. The column was equilibrated with five CVs of 25 mM Tris with pH 7.5 and conductivity 1.98 mS/cm. Next, 6.2 g of harvested fluid was loaded to a density of 60 g/L. Following the loading phase, an additional 3 CV of equilibration buffer was used for an initial wash, followed by 3 CV of a secondary wash at pH 7.5, followed by 3 CV of equilibration buffer for a final wash. The elution of 3 CV was conducted at pH 3.5 and 1.25 mS/cm, including fractionation every 12 mL, resulting in 26 fractions in total or 25 fully filled fractions. Column regeneration was performed using acidic and basic wash buffers following conventional protocols.

In addition to conventional in-line monitoring probes (UV, pH, conductivity), we introduced an in-line Raman spectroscopy probe to test whether we could measure multiple product quality attributes simultaneously in real time. We used a HyperFlux Pro Plus Raman

spectrometer (Tornado Spectral Systems) with an emission wavelength of 785 nm. This spectrometer was equipped with a 200 µL dead volume flow cell (Marqmetrix) that was connected in-line to an Akta Avant 150 preparative chromatography system after the UV flow cell. The Raman exposure time was set to 500 ms with an averaging of 15 spectra (resulting in an effective acquisition time of 7.5 s per spectrum) and the laser power was gradually decreased from the maximum of 495–350 mW to minimize detector saturation while maximizing the signal-to-noise ratio. Detector saturation was observed during peak product concentrations during the load and elution phases of affinity capture. Despite saturation, the Raman probe achieved a wider dynamic range in comparison to conventional UV probes (Figure 3a). The Raman spectra were collected continuously during the unit operation and was aligned to the AKTA data set using timestamps from both the Raman SpectralSoft (Tornado Spectral Systems) and AKTA Unicorn software (Cytiva). One Raman collection epoch resulted in one file, representing an average over 7.5 s of recording. An average and standard deviation was then calculated every four nonoverlapping epochs, representing 37.5 s of recording. These averages were then assigned to preparative chromatography fractions based on timestamp overlaps.

We performed the calibration of the Raman signals on a separate hardware system that integrated the Raman flow cell with liquid handling robotics. We set up a method for sample mixing and injection on the Fluent liquid handling platform (Tecan). The Fluent was set up with a carrier that was compatible with 600 µL RoboColumn hardware (Repligen) that was modified to provide a direct interface between the Raman flow cell and the liquid-handling tips, a module on Fluent system. The tips moved samples through the RoboColumn body as the interface to the Raman flow cell. The RoboColumn body used in the study was prepared by disassembling a commercial RoboColumn to two pieces: an upper cup and the main column body. Prepacked resin was removed from the column body, and a hole was drilled from the bottom of the upper cup. A connection to Raman flow cell was achieved by fixing Peak tubing by friction into the hole drilled in the cup. A light sensor was implemented at the top of the column body to record timestamps of Fluent tips contact events. The interface was cleaned with 1 M NaOH, 1 M acetate, and water after each sample injection. A Fluent worklist was generated using a custom Python script that automatically determined the sample handling sequence given the number of preparative chromatography fractionation samples and the number of intermediate calibration samples to create per pair of samples. For example, a calibration run involving 25 fractionation samples and six mixing samples produced 169 new mixed samples for Raman flow-cell injection, requiring 2 mL of volume per sample. Each of 169 injections was measured for an average of 2.8 min, generating means and standard deviations for each of 3101 wavenumbers from 200 to 3300 cm^{-1}.

2.2 | Analytical testing

Each of 25 preparative chromatography fractions were analyzed by five analytical assays to generate 16 product quality attributes, assessing protein concentration, host cell protein (HCP), product size variance, product charge variance, and N-glycan analysis. The product protein concentration was measured using a Lunatic (Unchained Labs) analyzer. HCP quantification was performed on the Octet analyzer using an Anti-CHO HCP Kit (Sartorius AG). Product molecular weight variants were analyzed using ultra-performance liquid size exclusion chromatography on an Acquity BEH SEC. 200 column (Waters Corporation). Product charge variants were analyzed by high-performance liquid chromatography on a MAbPac SCX-10 column (Thermo Fisher Scientific). Due to changes in quality attributes over time, the panel of assays were performed before and after calibration experiments, yielding standard deviation estimates that were important to the training and validation steps.

2.3 | Raman image preprocessing and modeling

Raman spectra were collected in the form of raw CSV files and required preprocessing to remove potentially confounding baseline effects (Ryabchykov et al., 2018). Because the amount of filtering is dependent on the amount of noise and smoothness of background signals, an approach using a range of cutoff frequency coefficients inspired by the prior methodology with continuous wavelet transformation within a range of scales (Zhang, Chen, Liang, Liu, et al., 2010) was introduced. We applied a series of Butterworth high-pass filters tuned to different cutoff frequency coefficients. This coefficient was set such that a value of 1 indicated oscillations across the Raman spectra with a period of 500 wavenumbers, which is equivalent to approximately six cycles over the whole range of wavenumbers 200–3300 cm^{-1}. Next, a range of coefficients was enumerated following a geometric scale over the range 2^{-2}–2^{2} (i.e., 0.25–4) resulting in 155 coefficients. A series of 155 Butterworth filters each using one of the enumerated coefficients was then applied to each raw Raman spectrum followed by minimum–maximum normalization, resulting in a final 2D Raman image of dimension 155 by 3101 (Figure 1a). This way, no new information was generated; instead, signals attributed to fluorescence interference and other analytes were removed, resulting in a Raman spectrum of improved quality.

A calibration model was built using a k-nearest neighbor (KNN) regressor to predict each of 16 quality attributes simultaneously given a 2D Raman image. The KNN regressor was chosen over other more conventional regression models, as it reduced the mean absolute error (MAE) of the high molecular weight (HMW) prediction three-fold compared to partial least squares regressoin (PLS) and principal component regression (PCR) (Wang et al., 2023). The MAE is defined as the sum of the absolute differences between the model output and the offline analytical measurement, divided by the total number of samples. The training data set for the model consisted of 169 samples mixed by the integrated Tecan-Raman system for which both the 2D Raman images and product quality attribute measurements by conventional offline analytical methods were available. The resulting quality attributes were mathematically derived from the analytical measurements of the original fraction samples and

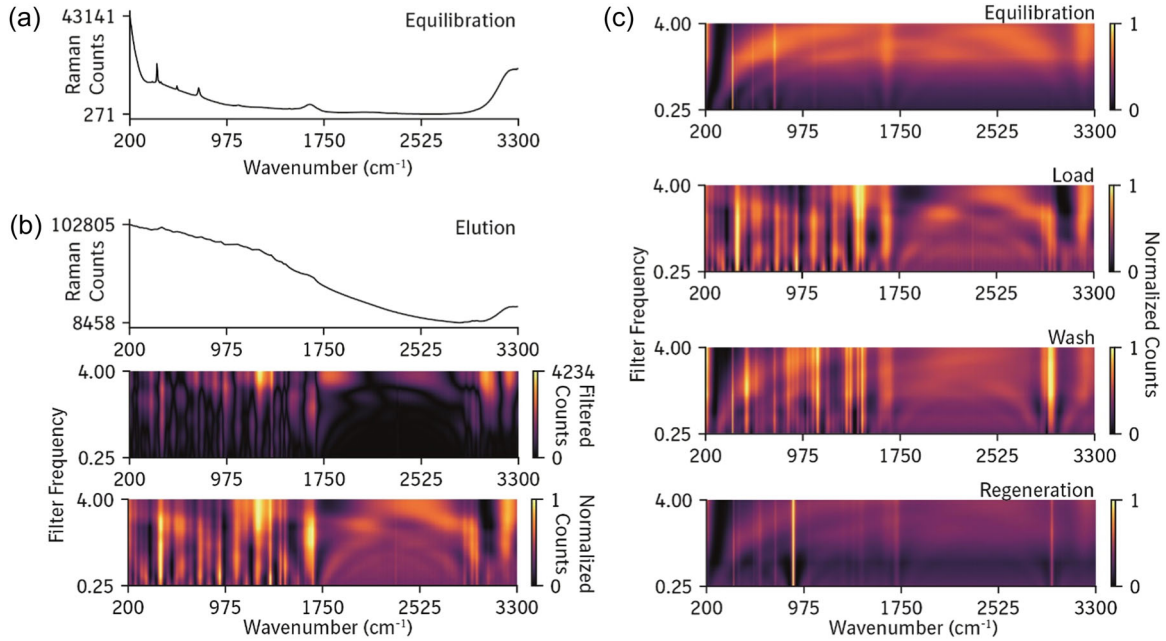

FIGURE 1 Raman 2D image preprocessing reveals features for product quality attribute classification and quantification. A raw Raman spectrum collected during the equilibration phase of an affinity capture preparative chromatography operation (a) contains several features not related to the product which need to be removed through preprocessing. The filtering process is demonstrated using an exemplar spectrum collected during the elution phase (b) where the raw spectrum (top) is subjected to a series of high-pass Butterworth filters tuned to different cutoff frequencies (middle) and is subsequently normalized (bottom). (c) The normalized 2D spectra observed during other phases of preparative chromatography reveal noticeably different features useful for the identification and quantification of product variant species.

the respective mixing ratios. The testing data set for the model consisted of 25 preparative chromatography fractions for which both the online 2D Raman images and offline analytical measurements were available. Data augmentation was performed on the training data set by randomly sampling the Raman images from a Gaussian distribution using a standard deviation that was 20% of the observed deviation and by a factor of 10 resulting in 1690 training samples. Each of the 16 analytical measurement labels were similarly augmented by sampling from a Gaussian distribution using a standard deviation that was 10% of the observed deviation by the same factor of 10. Next, a KNN regressor was trained on the augmented training data set using a k-value of 84. The inverse distance was used to weigh predictions and produced better results compared to the commonly used uniform weighing. Prediction of the online results also followed data augmentation where the online 2D Raman image was sampled from a Gaussian distribution using the same coefficient of 20% of the observed standard deviation. This sampled Raman image was then converted into a set of 16 product quality attribute predictions using the trained KNN model.

3 | RESULTS

We investigated whether Raman spectroscopy could be used as a noninvasive, optical analyzer suitable for real-time multiple product quality monitoring during bioprocessing. The relationship between distinct Raman peaks and the identity and quantity of free amino acids have been successfully modeled and understood (Zhu et al., 2011), however, the modeling of complex therapeutic proteins with several product-related variants remains a challenge (Rolinger et al., 2021). While the Raman background is relatively free of confounding effects such as baseline shifts due to water (Figure 1a), the product spectrum is often too complex for either conventional or computational analysis. We adapted a preprocessing method using a range of cutoff frequencies which leverages spectral decomposition to differentiate underlying signals not immediately eminent in the captured signal. Taking the in-line Raman spectrum from the elution phase of an affinity capture preparatory chromatography unit operation as an example, a complex 1D Raman signal without any distinguishing features is decomposed using a series of high-pass frequency filters tuned to different coefficients (see Section 2) to reveal several noticeable features on the resulting 2D Raman image (Figure 1b). Small peaks that often overlap and become negated in the complex 1D signal are resolved on the orthogonal second dimension as signals are separated by their relative sharpness. Critically, many of these product elution features are not present in the equilibration and regeneration phases where no product is present (Figure 1c), especially new features seen in the regions below 1750 cm^{-1}. Normalization is applied to prevent outlier effects from particularly strong features, for example, peaks attributed to sapphire glass present in the Raman flow cell. A subset of these features overlapped with those present during load and wash phases, which are known to contain quantities of product, product-related variants, and HCPs. While it is not known how these decomposed features

relate mechanistically to a given biochemical transformation, we hypothesized that it would be feasible to apply computational modeling to elucidate potential interactions to simultaneously predict multiple product variants from the processed Raman signal.

Modeling relationships between 2D Raman image features and product variant concentrations required a training data set that captured sufficient variability in relevant parameters. For this purpose, we used 25 fractionation samples from the elution phase of an affinity preparative chromatography unit operation. Protein A chromatography results in measurable proportions of variants in addition to the desired product (Pabst et al., 2018). We performed five analytical assays to determine 16 product quality attributes for each of the 25 fractionation samples (see Section 2) and designated these observations as the test data set, or the observed ground truth. We then used an automated Tecan-Raman system to generate a training data set with 169 observations of the same 16 product quality attributes. Each of 169 off-line samples were analyzed by Raman with an acquisition time of 2.8 min. The system automated sample mixing and injection into the Raman flow cell, enabling lower manual labor effort and a turnaround time of less than 1 week. The mixing system was able to systematically enumerate novel mixtures of samples of known product variant concentrations, resulting in 169 sets of 16 product quality attribute values usable for calibration model training from the analysis of only 25 offline samples following laborious conventional protocols. The resulting training and test data sets were then preprocessed following 2D Raman imaging (Figure 1), and a KNN model was fitted using the training data set. Hyperparameter optimization was performed on KNN parameters to select the optimal number of neighbors and data augmentation parameters critical for minimizing overfitting (see Section 2). The predicted values of 16 in-line product quality attributes plotted against their measured values are shown in Figure 2. The full names of each quality attribute are shown in Table 1. Detector saturation occurred around the elution peak where measured concentrations were maximal, resulting in elevated errors in attribute prediction. Otherwise, there was generally a good agreement between predicted and observed values, demonstrating the feasibility to measure 16 quality attributes during bioprocessing in real time. The key advantages of using this in-line Raman spectroscopy-based method are the short acquisition time of 30 s and the noninvasive nature of the optical measurement.

The successful application of 2D Raman imaging for multiple product quality attribute monitoring was demonstrated during a key preparative chromatography process. Conventionally, pH, conductivity, and UV absorbance at 280 nm are captured and monitored in real time. Dynamic in-line measurements are often used to set process-critical triggers to control the start or stop of product pooling. Here, we demonstrate the ability to monitor total protein concentration in-process with a dynamic range much wider than that of traditional UV sensors (Figure 3a). We achieved concentration monitoring with a linearity of $Q = 0.95$ and an MAE of 3.4 g/L (Table 1), where most of the error was attributable to detector saturation at peak concentrations. We then calibrated our Raman model to monitor the amount of undesirable HCP, which was the total concentration minus product and product variant concentrations. We achieved a linearity of $Q = 0.84$ and an MAE of 1.6 mg/L. Next, we trained our model to monitor results of the N-glycan assay (Figure 3b), revealing product variants with potentially undesirable Fc-mediated effector functions. We found linearity of at least $Q = 0.9$ (Table 1) across detected N-glycan species, again with key deviations attributable to detector saturation. Furthermore, we trained our model to monitor product molecular weight variants including product aggregates and fragments (Figure 3c) with linearity of at least $Q = 0.95$. Finally, we implemented our model to measure product charge variants including acidic and basic peak groups (Figure 3d), resulting in linearity of at least $Q = 0.93$. A summary of accuracy metrics, including additional metrics such as mean absolute percent error, for all 16 product quality attributes are shown in Table 1.

4 | DISCUSSION

Monitoring the quality attributes of pharmaceutical antibody products during manufacturing in real-time promises to enable advanced manufacturing capabilities in future processes and improve quality control in conventional processes. In this study, we demonstrate the feasibility of simultaneously monitoring 16 diverse product quality attributes during a key preparative chromatography process step. The resulting images show visibly identifiable features that characterized each distinct phase of affinity capture chromatography. Only through the combination of this preprocessing step with the supervised machine learning algorithm was it possible to build models to correlate such a broad range of quality attributes. An automation system consisting of a Raman spectrometer integrated into a liquid handling robot was used to mix chromatography fraction samples, automating the Raman spectra collection process. A KNN regression model was then trained on the preprocessed Raman spectra and applied to predict 16 product quality attributes during affinity chromatography in real time (Figure 2). The calibration model made accurate predictions (Table 1), with errors mostly attributable to detector saturation under high-concentration conditions (Figure 3). The reason for this concentration limitation lies in the saturation of Raman signals caused by the hardware settings, especially exposure time, utilized. It was chosen with the idea to cover lower concentrated quality attributes, such as HCP or N-glycan variants, as well. Therefore, future works need to be aware of the trade-off when selecting the concentration range. In summary, these findings demonstrate the simultaneous in-line measurement of 16 product quality attributes potentially critical to process control strategies.

Without holding back, we would like to share our honest opinion on the potential of Raman-based PAT to accurately forecast the CQAs in this investigation. For HMW, there are multiple mechanisms known such as nucleation-dominated, chain polymerization, and associated polymerization. The formation of beta sheets can be found to be reflected in the Amide III region near 1200 cm^{-1}, while disulfide structural changes are typically observed in the 500–700 cm^{-1} regions (Barnett et al., 2015; Li & Li, 2009). Critical regions related

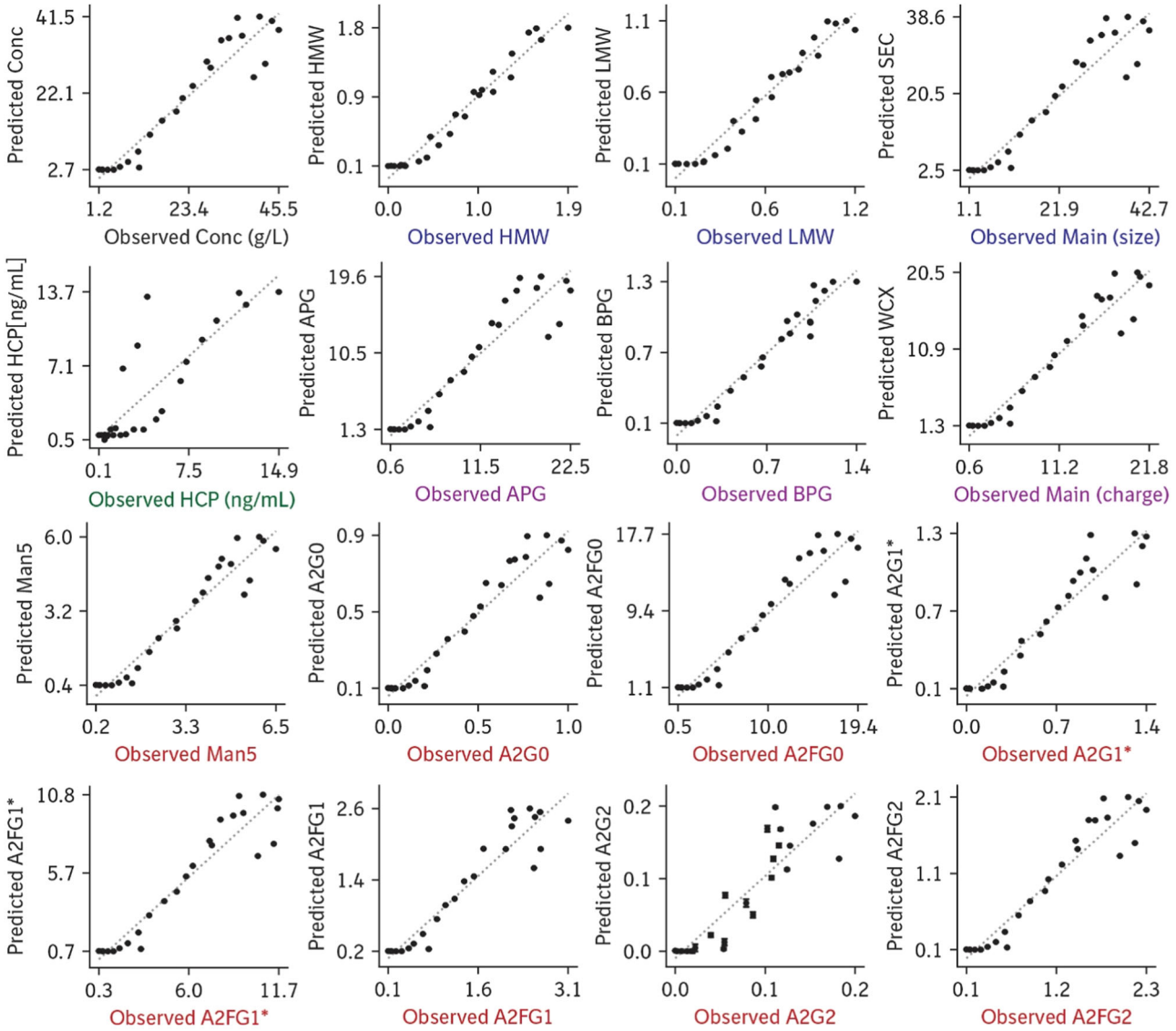

FIGURE 2 Comparison between conventional off-line analytical measurements and in-line predictions by the Raman model. After calibrating a k-nearest neighbor model on an offline data set ($n = 169$) linking preprocessed Raman spectra to 16 product quality attributes derived from five analytical assays, each of 25 samples from the elution phase of an affinity capture preparative chromatography operation was analyzed by conventional assays, plotted as the "Observed" values, as well as by the Raman model, plotted as the "Predicted" values. A list of each attribute and accuracy performance values is shown in Table 1.

to protein folding and unfolding can also be identified at 770 and 900 cm^{-1} (Gómez de la Cuesta et al., 2014). Regarding LMW, the formation is similarly diverse. Fragments like FC domains and antigen-binding domains can be detected through the C-S and S-S stretch band between 670 and 780 cm^{-1} and 425 and 550 cm^{-1}, respectively, which results from disulfide reduction and higher-order structural changes in the protein. These changes also lead to variations in the Amide I region and other regions reflecting aromatic amino acids. Charge variants can also be attributed to various factors, including deamidation and glycosylation. Deamidation, a common posttranslational modification, converted the amide functional group in asparagine or glutamine residues to aspartic acid or glutamic acid, respectively, creating a new acidic site. This conversion resulted in changes in the Raman spectrum, particularly in the Amide I and Amide III bands, although the specific changes depend on the location of deamidation and the presence of other modifications. Glycosylation patterns also influence charge variants by altering the charge of the mAb molecule. The specific wavenumbers at which glycan groups can be detected in the Raman spectrum vary based on the structure and environment of the molecule, making it challenging to assign specific bands to individual sugar residues. Therefore, the unexpected proximity of the model's prediction in this context is noteworthy. The application of this in-line multiple attribute monitoring system is expected to meet the currently unmet demand for high volumes of analytical feedback in high-throughput experimentation and continuous manufacturing strategies. High-throughput

TABLE 1 Performance of the in-line measurement of 16 product quality attributes.

Quality attribute	Q^2	Q	MAE	MAPE [%]
Conc	0.909	0.953	3.40 [g/L]	35.78
HCP	0.712	0.844	1.62 [ng/L]	50.17
HMW	0.962	0.981	0.10 [g/L]	35.44
LMW	0.957	0.978	0.07 [g/L]	30.34
Main peak of size variants (SEC)	0.902	0.950	3.26 [g/L]	36.07
APG	0.880	0.938	1.86 [g/L]	37.01
BPG	0.955	0.977	0.09 [g/L]	34.17
Main peak of charge variants (WCX)	0.927	0.963	1.51 [g/L]	35.01
N-glycan (Man5)	0.932	0.965	0.44 [g/L]	34.75
N-glycan (A2G0)	0.910	0.954	0.07 [g/L]	35.49
N-glycan (A2FG0)	0.905	0.952	1.47 [g/L]	35.94
N-glycan (A2G1*)	0.900	0.949	0.10 [g/L]	38.72
N-glycan (A2FG1*)	0.902	0.949	0.90 [g/L]	35.94
N-glycan (A2FG1)	0.894	0.946	0.23 [g/L]	36.53
N-glycan (A2G2)	0.810	0.900	0.02 [g/L]	47.65
N-glycan (A2FG2)	0.910	0.954	0.17 [g/L]	35.73
Absorbance 280 nm (UV)	0.811	0.901	0.29 [mAU]	27.85
pH	0.348	0.590	1.08 [-]	23.43
Cond	0.630	0.793	0.09 [mS/cm]	9.89

Note: The accuracy of the in-line measurement is computed using the predictive correlation coefficient Q and Q^2, the MAE, and the MAPE. Errors in in-line prediction are mainly attributable to Raman detector saturation at higher concentrations. The performance of using the Raman model to predict three conventionally monitored in-line attributes (UV, pH, and conductivity) are shown in the bottom three rows.

Abbreviations: APG, acidic peak group; BPG, basic peak group; Conc, concentration; Cond, conductivity; HCP, host cell proteins; HMW, high molecular weight; LMW, low molecular weight; MAE, mean absolute error; MAPE, mean absolute percent error.

chromatography experimentation generates large numbers of chromatograms, especially using 96-well format experiments (Keller et al., 2022). These large numbers of chromatograms (e.g., $n = 48$) can each be significantly augmented with 16 additional attribute traces using our in-line monitoring approach, providing new scientific opportunities not possible under current off-line analytical workflows. Automation robotics generate large numbers of samples, however, an optical analytical method for product quality attribute measurement is currently missing and highly desired to realize the full value of automation (Silva et al., 2022). There is an industry-wide consensus that continuous manufacturing under a good manufacturing practice setting may only be possible once sufficient in-line product quality monitoring tools become widely applied (Esmonde-White et al., 2022). Here we demonstrate one solution that will fit this unmet need in advanced manufacturing. We highlight our ability to monitor several attributes of potentially heightened interest such as acidic charge and N-glycan variants, or the presence of HCPs (Table 1). Overall, the presented method consisting of KNN regression and Butterworth filter series shows promising results, demonstrating its effectiveness in realizing PAT for biopharmaceutical downstream processing. Before the implementation of the Butterworth filter series, the prediction of size variants was only achieved at a satisfactory level (Wang et al., 2023). It is important to highlight that the purpose of applying the Butterworth filter is to selectively remove noise and interference from the data, particularly fluorescence signals, to improve the quality of Raman spectra for the relevant quality attributes of interest. The presence of intense fluorescence emissions in Raman spectroscopy can obscure weaker vibrational fingerprints and degrade the overall quality of the spectra. By utilizing data preprocessing methods like the Butterworth filter, fluorescence interference can be effectively eliminated. Furthermore, multicomponent real-life systems as used in this study, pose a major challenge when aiming to extract a specific quality attribute of interest from the overall Raman signals, since the scattering contribution from all other components can be considered as background noise, which comes on top to the fluorescence interference. This complexity necessitated the application of data preprocessing techniques, here a series of Butterworth filters has been shown to be effective separating the desired Raman signal from the background noise.

The application of this method will need to be harmonized with ongoing parallel work in upstream systems (Graf et al., 2022) to realize the long-term goal of shortened project timelines and continuous in-process controls across all unit operations. This point is critical since time savings in a limited number of unit operations does not reduce either the throughput or duration of the overall development pipeline due to the parallelized nature of process/analytical development and characterization strategies across upstream, downstream, formulation, and analytical functions (Kelley, 2020). It is, therefore, critical for industry leaders seeking accelerated product-to-market timelines to find strategic technology-implementation fit for the key results presented here (Gillespie et al., 2022).

UV absorbance is a critical part of bioprocessing despite its limited dynamic range at 280 nm (Ramakrishna et al., 2022) and sensitivity to interfering nonprotein chromophores, requiring process limitations to avoid false positive detection of concentration. Raman spectroscopy operates on an equivalent mechanism as UV absorbance in that it is deployed in-line and relies on a light source to noninvasively monitor product and impurities dynamically under flow. Common limitations affecting UV absorbance measurement also impact Raman imaging, including constraints on dynamic range due to detector saturation and sources of confounding variables from process parameter variations, often requiring supplementary calibration data to fine-tune the original model (Smith & Dent, 2019). We demonstrate in this study the feasibility of using Raman to quantify up to a maximum concentration of around 40 g/L while producing meaningful results, a range significantly broader than previously investigated (Wei et al., 2021). We accounted for variations of

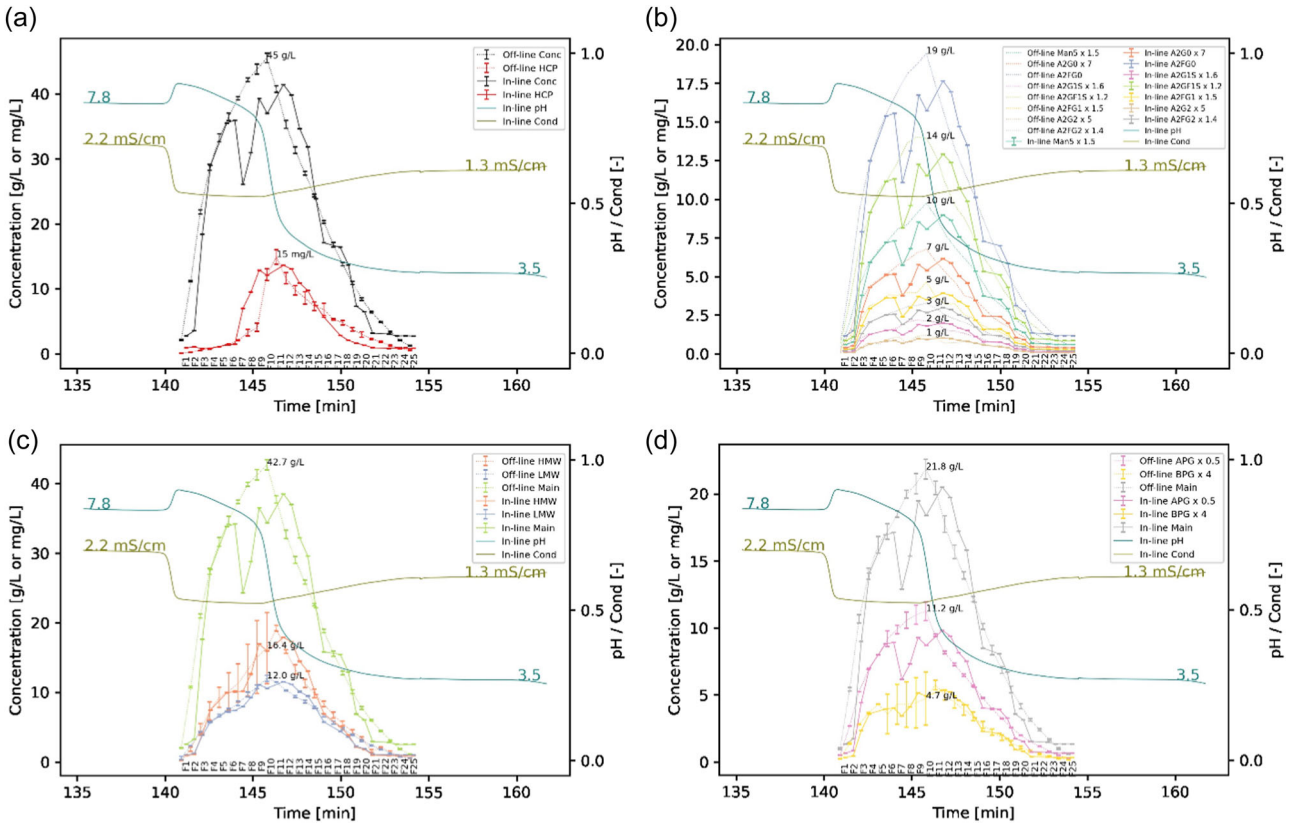

FIGURE 3 Chromatograms demonstrating the in-line prediction of 16 product quality attributes simultaneously by applying the Raman model. After applying the Raman model, the predicted attribute values are shown over time during affinity capture preparative chromatography. Conventional assays used for comparison include product concentration and ELISA host cell protein (HCP) quantification (a), N-glycan analysis by capillary gel electrophoresis (CGE) (b), size variant analysis by ultra performance liquid chromatography (UPLC) (c), and charge variant analysis by high-performance liquid chromatography (d). Dotted lines indicate results from off-line analyses while solid lines indicate in-line predictions by the Raman model. Error bars indicate one standard deviation. Saturation of the Raman detector occurred near the elution peak and was responsible for abnormal shifts in the predicted concentrations. It is important to note that in subfigures (b) and (d), the quality attributes with extremely low concentrations were magnified. This is indicated by the numbers after the multiplication operator in the legend of the respective subfigures, which were used to enhance visualization and ease of reading.

potential sources of background contamination by augmenting existing single parameter preprocessing methods (Zhang, Chen, Liang, et al., 2010) with multiple coefficient filtering (Zhang, Chen, Liang, Liu, et al., 2010), resulting in 2D Raman images with a broad set of identifying features (Figure 1). Since we were able to capture spectra sufficient for in-line quality monitoring within 30 s, improving on current methods requiring several minutes (Wei et al., 2021), we anticipate that collecting and synthesizing spectra using other exposure time values in addition to 500 ms will increase the dynamic range further. Increasing the frequency cutoff coefficient range during preprocessing (see Section 2) is expected to decrease the potential for interfering signals to avoid detection, minimizing risks of false detection and quantification. Other sources of process variations such as pH or conductivity should be minimized during process development to further diminish the risk of calibration model mismatch.

The relatively high number of feature variables ($n = 3101 * 155$) compared to the small number of observations ($n = 169$) in this study is similarly apparent in other studies dealing with biological and clinical data (Berisha et al., 2021; Tulsyan et al., 2019). This high-dimensionality problem arises due to the scarce nature of biological samples that frequently employ laborious analytical testing and expert interpretation. We contributed toward diminishing this problem from implementing both hardware and software solutions. Our integrated Tecan-Raman system increased the number of training observations by automatically mixing calibration samples. From the software side, we implemented data augmentation to virtually increase our number of observations by a factor of 10 (see Section 2). Although we increased the dimensionality of our feature set by a factor of 155 by constructing a 2D Raman image, the number of meaningful features related to peak characteristics is likely much smaller than the total number of variables used for our KNN regression model. This suggests that regression algorithms that automate feature selection such as convolutional neural networks (Rolinger et al., 2021) may be able to meaningfully represent these 2D features using a smaller or more reasonable number of parameters. One advantage of using the KNN regressor is the absence of training and thus an absence of a need to apply regularization or optimization during training. The number of tunable

parameters is limited to the single *k* parameter and the relatively straightforward data augmentation parameters that determined robustness to measurement noise (see Section 2).

We carried out calibration experiments at the benchtop scale and anticipate the need for both smaller and larger scales to further realize the potential of in-line product quality monitoring. Toward the miniaturization of experiments following the HTPD strategy (Hubbuch, 2012), we anticipate the potential for chromatography experiments to be carried out using CV of under 1 mL. A smaller Raman flow cell with a dead volume of 45 uL is commercially available, which when compared with the flow cell used in this study with a dead volume of 200 uL, presents exciting future opportunities for investigating possibilities at the automation scale. We required 1 mL of sample volume for calibrating the flow cell, which was similar to the material demand of comparable studies (Wei et al., 2021), thus if we were to proportionally scale our experiment using the smaller flow cell, we would require 4.4 times less sample for calibration at 225 uL per injection. In addition to reducing sample requirement, the miniaturization of this technique can automate larger numbers of on-column experiments using chromatography columns smaller than 1000 uL directly in line with multiple product quality analytical capabilities. This would relieve high burdens for analytical testing, a significant limitation currently constraining HTPD applications (Silva et al., 2022). Miniaturization promises to increase process understanding by significantly increasing available product quality analytical measurements that facilitate process development, characterization, transfer, and monitoring in both upstream and downstream bioprocessing unit operations (Rathore, 2014; Silva et al., 2022). Toward the scaling-up of experimental scale, we anticipate significant improvements to process monitoring and control for continuous manufacturing strategies (Khanal & Lenhoff, 2021).

Although the findings in the presented study are highly encouraging and exciting to be expanded and implemented into the industrial downstream process development and other stage-one validation activities, the authors herewith highlight the importance of considering the real-world complexities and interfaces in implementing the technology within a project environment. Based on the available data, it is not evident that a model trained on a specific molecule can be directly applied to a new molecule of similar format. Similarly, there is no evidence to support the direct transferability of a model from one unit operation to another within the process stream. However, process condition changes could be accommodated within the calibration space. Ultimately, regardless of the regression model used, such as KNN, PLS, PCR, or CNN, there is no mechanistic foundation to ensure accurate and reliable extrapolation beyond the calibration ranges of the quality attributes.

5 | CONCLUSION

In conclusion, this study demonstrates the feasibility of using Raman spectroscopy combined with a 2D preprocessing method and machine learning algorithms to simultaneously predict 16 different quality attributes during Protein A chromatography in biopharmaceutical manufacturing. The novel preprocessing method allows for the extraction of a larger number of spectral features, which, when paired with a machine learning algorithm, enables the training of a model for in-line quality monitoring. At process development stage, the integration of this technology allows for high volumes of analytical feedback, such as charge variants, size variants, glycan pattern, and HCPs, supporting scientifically sound decision making. Further, the ability to predict multiple quality attributes in real time enhances process control and understanding in biopharmaceutical manufacturing. The novel technology addresses the need for rapid and noninvasive measurement of process quality attributes, in alignment with the industry-wide push for continuous manufacturing. The real-time collection of quality attribute data enables adaptive control and improves quality control in conventional processes. While this study showed the potential of the proposed technology based on Protein A chromatography, further development is needed to apply this technology to other unit operations and scales, especially for the ultrafiltration/diafiltration step, with the need to monitor both high protein concentrations well beyond 40 g/L and the lowest levels of impurities and buffer species could be a stretch.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Markus Wendeler and Dr. Jan Visser for their constant support.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Research data are not shared.

ORCID

Gang Wang http://orcid.org/0009-0000-8262-2951

REFERENCES

Barnett, G. V., Qi, W., Amin, S., Lewis, E. N., Razinkov, V. I., Kerwin, B. A., Liu, Y., & Roberts, C. J. (2015). Structural changes and aggregation mechanisms for anti-streptavidin IgG1 at elevated concentration. *The Journal of Physical Chemistry B*, *119*(49), 15150–15163.

Berisha, V., Krantsevich, C., Hahn, P. R., Hahn, S., Dasarathy, G., Turaga, P., & Liss, J. (2021). Digital medicine and the curse of dimensionality. *NPJ Digital Medicine* *4*(1), 153.

Esmonde-White, K. A., Cuellar, M., & Lewis, I. R. (2022). The role of Raman spectroscopy in biopharmaceuticals from development to manufacturing. *Analytical and Bioanalytical Chemistry*, *414*, 969–991.

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*:e2847.

Gillespie, C., Wasalathanthri, D. P., Ritz, D. B., Zhou, G., Davis, K. A., Wucherpfennig, T., & Hazelwood, N. (2022). Systematic assessment of process analytical technologies for biologics. *Biotechnology and Bioengineering*, *119*, 423–434.

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 innovate data analysis workflow to enhance biopharmaceutical. *Processes*, *8*, 1179.

Gómez de la Cuesta, R., Goodacre, R., & Ashton, L. (2014). Monitoring antibody aggregation in early drug development using raman spectroscopy and perturbation-correlation moving windows. *Analytical Chemistry*, *86*, 11133–11140.

Graf, A., Lemke, J., Schulze, M., Soeldner, R., Rebner, K., Hoehse, M., & Matuszczyk, J. (2022). A novel approach for non-invasive continuous in-line control of perfusion cell cultivations by Raman spectroscopy. *Frontiers in Bioengineering and Biotechnology*, *10*, 719614.

Guo, S., Popp, J., & Bocklitz, T. (2021). Chemometric analysis in Raman spectroscopy from experimental design to machine learning–based modeling. *Nature Protocols*, *16*, 5426–5459.

Hubbuch, J. (2012). Editorial: High-throughput process development. *Biotechnology Journal*, *7*, 1185.

ICH. (2021). Continuous manufacturing of drug substances and drug products Q13. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use.

ICH. (2022). Analytical procedure development Q14. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use.

Keller, W. R., Evans, S. T., Ferreira, G., Robbins, D., & Cramer, S. M. (2022). Understanding the effects of system differences for parameter estimation and scale-up of high throughput chromatographic data. *Journal of Chromatography A*, *1661*, 462696.

Kelley, B. (2020). Developing therapeutic monoclonal antibodies at pandemic pace. *Nature Biotechnology*, *38*, 540–545.

Khanal, O., & Lenhoff, A. M. (2021). Developments and opportunities in continuous biopharmaceutical manufacturing. *mAbs*, *13*, 1903664.

Li, C., & Li, T. (2009). Application of vibrational spectroscopy to the structural characterization of monoclonal antibody and its aggregate. *Current Pharmaceutical Biotechnology*, *10*(4), 391–399.

McAvan, B. S., Bowsher, L. A., Powell, T., O'Hara, J. F., Spitali, M., Goodacre, R., & Doig, A. J. (2020). Raman spectroscopy to monitor post-translational modifications and degradation in monoclonal antibody therapeutics. *Analytical Chemistry*, *92*, 10381–10389.

Müller, D. H., Flake, C., Brands, T., & Koß, H. (2023). Bioprocess in-line monitoring using Raman spectroscopy and Indirect Hard Modeling (IHM): A simple calibration yields a robust model. *Biotechnology and Bioengineering*. *120*(7), 1857–1868.

Pabst, T. M., Thai, J., & Hunter, A. K. (2018). Evaluation of recent Protein A stationary phase innovations for capture of biotherapeutics. *Journal of Chromatography A*, *1554*, 45–60.

Patel, B. A., Gospodarek, A., Larkin, M., Kenrick, S. A., Haverick, M. A., Tugcu, N., Brower, M. A., & Richardson, D. D. (2018). Multi-angle light scattering as a process analytical technology measuring real-time molecular weight for downstream process control. *mAbs*, *10*, 945–950.

Pedro, M. N. S., Klijn, M. E., Eppink, M. H., & Ottens, M. (2021). Process analytical technique (PAT) miniaturization for monoclonal antibody aggregate detection in continuous downstream processing. *Journal of Chemical Technology & Biotechnology*. *97*(9), 2347–2364.

Ramakrishna, A., Prathap, V., Maranholkar, V., & Rathore, A. S. (2022). Multi-wavelength UV-based PAT tool for measuring protein concentration. *Journal of Pharmaceutical and Biomedical Analysis*, *207*, 114394.

Rashad, A., Rasool, A., Shaheryar, M., Sarfraz, A., Sarfraz, Z., Robles-Velasco, K., & Cherrez-Ojeda, I. (2022). Donanemab for Alzheimer's disease: A systematic review of clinical trials. *Healthc*. *11*, 32.

Rathore, A. S. (2014). QbD/PAT for bioprocessing: Moving from theory to implementation. *Current Opinion in Chemical Engineering*, *6*, 1–8.

Reardon, S. (2023). Alzheimer's drug donanemab: What promising trial means for treatments. *Nature*, *617*, 232–233.

Rolinger, L., Rüdt, M., & Hubbuch, J. (2020). A critical review of recent trends, and a future perspective of optical spectroscopy as PAT in biopharmaceutical downstream processing. *Analytical and Bioanalytical Chemistry*, *412*, 2047–2064.

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

Ryabchykov, O., Guo, S., & Bocklitz, T. (2018). Analyzing Raman spectroscopic data. *Physical Sciences Reviews*, *4*, 20170043.

Saleh, D., Wang, G., Müller, B., Rischawy, F., Kluters, S., Studts, J., & Hubbuch, J. (2020). Straightforward method for calibration of mechanistic cation exchange chromatography models for industrial applications. *Biotechnology Progress*, *36*, e2984.

Silva, T. C., Eppink, M., & Ottens, M. (2022). Automation and miniaturization: Enabling tools for fast, high-throughput process development in integrated continuous biomanufacturing. *Journal of Chemical Technology & Biotechnology*, *97*, 2365–2375.

Smith, E., & Dent, G. (2019). Modern Raman Spectroscopy.

Taraban, M. B., Briggs, K. T., Merkel, P., & Yu, Y. B. (2019). Flow water proton NMR: In-line process analytical technology for continuous biomanufacturing. *Analytical Chemistry*, *91*, 13538–13546.

Tulsyan, A., Garvin, C., & Undey, C. (2019). Industrial batch process monitoring with limited data. *Journal of Process Control*, *77*, 114–133.

Tulsyan, A., Wang, T., Schorner, G., Khodabandehlou, H., Coufal, M., & Undey, C. (2020). Automatic real-time calibration, assessment, and maintenance of generic Raman models for online monitoring of cell culture processes. *Biotechnology and Bioengineering*, *117*, 406–416.

Wang, J., Chen, J., Studts, J., & Wang, G. (2023). In-line product quality monitoring during biopharmaceutical manufacturing using computational Raman spectroscopy. *mAbs*, *15*, 2220149.

Wei, B., Woon, N., Dai, L., Fish, R., Tai, M., Handagama, W., Yin, A., Sun, J., Maier, A., McDaniel, D., Kadaub, E., Yang, J., Saggu, M., Woys, A., Pester, O., Lambert, D., Pell, A., Hao, Z., Magill, G., … Chen, Y. (2021). Multi-attribute Raman spectroscopy (MARS) for monitoring product quality attributes in formulated monoclonal antibody therapeutics. *mAbs*, *14*, 2007564.

Yilmaz, D., Mehdizadeh, H., Navarro, D., Shehzad, A., O'Connor, M., & McCormick, P. (2020). Application of Raman spectroscopy in monoclonal antibody producing continuous systems for downstream process intensification. *Biotechnology Progress*, *36*, e2947.

Zhang, Z., Chen, S., Liang, Y., Liu, Z., Zhang, Q., Ding, L., Ye, F., & Zhou, H. (2010). An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy. *Journal of Raman Spectroscopy*, *41*, 659–669.

Zhang, Z.-M., Chen, S., & Liang, Y.-Z. (2010). Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst*, *135*, 1138–1146.

Zhu G., Zhu X., Fan Q., Wan X. 2011. Raman spectra of amino acids and their aqueous solutions. *Spectrochimica Acta*, *Part A: Molecular and Biomolecular Spectroscopy* 78:1187–1195.

How to cite this article: Wang, J., Chen, J., Studts, J., & Wang, G. (2024). Simultaneous prediction of 16 quality attributes during protein A chromatography using machine learning based Raman spectroscopy models. *Biotechnology and Bioengineering*, *121*, 1729–1738. https://doi.org/10.1002/bit.28679