Biotechnology and Bioengineering

WILEY

BIOTECHNOLOGY AND BIOENGINEERING

ARTICLE OPEN ACCESS

Development of Raman Spectroscopy and Machine Learning Methods for Protein Aggregate Quantification: Application to BSA in Chromatographic Processes

Jakob Heyer-Müller[1] | Robin Schiemer[1] | Lars Robbel[2] | Michael Schmitt[2] | Jürgen Hubbuch[1]

[1]Institute of Process Engineering in Life Sciences—Section IV: Biomolecular Separation Engineering, Karlsruhe Institute of Technology (KIT), Fritz-Haber-Weg 2, Karlsruhe, Baden-Württemberg, Germany | [2]CSL Innovation GmbH Germany, Michelbacher Straße, Marburg, Hessen, Germany

Correspondence: Jürgen Hubbuch (juergen.hubbuch@kit.edu)

Received: 3 September 2025 | **Revised:** 5 December 2025 | **Accepted:** 13 January 2026

Keywords: aggregation | chemometrics | chromatography | convolutional neural network | process analytical technology | Raman spectroscopy | structural markers

ABSTRACT

Protein aggregation poses a significant risk to biopharmaceutical product quality, as even minor amounts of oligomeric species can compromise efficacy and safety. Rapid and reliable detection of protein aggregates thus remains a major challenge in biopharmaceutical manufacturing. Although traditional offline methods such as size-exclusion chromatography provide accurate results, their inherent time delays limit real-time process control capabilities. Consequently, there is an urgent scientific need for inline analytical techniques capable of selectively quantifying protein monomers and aggregates in real time to facilitate immediate corrective actions and enhance overall process robustness. Raman spectroscopy, as a tool for a process analytical technology application, is especially suitable due to its molecular specificity, rapid data acquisition, and compatibility with aqueous solutions commonly used in biopharmaceutical manufacturing. Addressing this need, this study establishes a Raman spectroscopy-based strategy for the selective detection and quantification of monomeric and aggregated forms of a model protein (bovine serum albumin). Controlled stress conditions were applied to generate aggregated species reproducibly, and a Latin Hypercube sampling design was used to independently vary protein concentration and aggregate fraction, ensuring that observed spectral effects were attributable to aggregation rather than concentration differences. Furthermore, spectral markers identified in spectra acquired from multiple chromatographic runs were qualitatively compared with offline reference measurements from size-exclusion chromatography. This limitation in real-time applicability was circumvented by chemometric machine learning approaches. The use of convolutional neural networks enabled the selective quantification of the protein monomers and aggregates and delivered superior predictive performance and robustness across cross-validation, independent testing, and synthetic perturbation scenarios compared to traditional chemometric approaches. Collectively, these results demonstrate that the selected Raman spectral markers, combined with advanced chemometric modeling, enable reliable, real-time monitoring of protein size variants in biopharmaceutical downstream processes.

1 | Introduction

Protein aggregation is a critical quality attribute (CQA) and a major challenge in the manufacturing of biopharmaceuticals. Aggregates can form throughout the product lifecycle—including upstream production, downstream purification, formulation, and storage (Vázquez-Rey and Lang 2011; Roberts 2014; Pham and Meng 2020)—potentially compromising both the safety and efficacy of therapeutic proteins

Jakob Heyer-Muller and Robin Schiemer contributed equally.

(Rosenberg 2006; Moussa et al. 2016). To address this issue, there is an increasing need for real-time monitoring of protein structural integrity under application-relevant conditions. Implementing such monitoring, particularly in downstream processing, is essential to ensure consistent product quality and meet regulatory requirements.

The introduction of the process analytical technology (PAT) initiative by the US Food and Drug Administration (FDA) aimed to enable real-time, inline measurement and control of critical process parameters and quality attributes (FDA 2004). Univariate sensors are commonly applied to monitor critical process parameters in upstream and downstream unit operations. Modern PAT frameworks integrate spectroscopic tools and chemometric models (Rolinger et al. 2020a) to enable real-time impurity quantification (Brestrich et al. 2015; Capito et al. 2015; Wei et al. 2022; Chen et al. 2024), protein structure assessment (Rüdt et al. 2019), reaction monitoring (Schiemer et al. 2023), or precipitation behavior (Dietrich et al. 2024). These innovations support a shift toward Quality by Design (QbD) approaches in biopharmaceutical production (Glassey et al. 2011).

Early studies using vibrational spectroscopy techniques, such as Raman and FTIR, reported changes in the amide I and III bands and the appearance of intermolecular β-sheet structures for proteins subjected to heat stress (Alizadeh-Pasdar et al. 2002; Militello et al. 2004). It has been shown that environmental stresses (e.g., pH, temperature, and ionic strength) induce specific conformational changes that can be spectroscopically monitored and are commonly associated with protein aggregation (Ettah and Ashton 2018). Later studies reported similar spectral changes in the amide bands for different naturally occurring proteins as well as for common therapeutic proteins such as immunoglobulin G (IgG) (Gómez de la Cuesta et al. 2014; Lewis et al. 2014; Barnett et al. 2015; Zhou et al. 2015). Aside from the changes in the amide bands, multiple studies also report the shifts in the vibrational bands of aromatic amino acid side chains like tryptophan, tyrosine, or phenylalanine, which undergo changes in their solvatization during protein aggregation (Barnett et al. 2015; Zhou et al. 2015). However, the vibrational spectra are not only affected by protein aggregation but also by other environmental factors such as pH, protein concentration, or excipients (Ota et al. 2016; Sato et al. 2023; Makki et al. 2021) as well as intrinsic fluorescence and Rayleigh scattering of the protein analytes requiring targeted preprocessing of the collected Raman spectra (Rolinger et al. 2020a; Dietrich et al. 2024).

Due to fine differences between the native and the aggregated proteins and the influence of environmental factors on the Raman spectra, multivariate data analysis is commonly used to monitor purification processes and, in particular, protein aggregation during process development and early formulation stages (Gómez de la Cuesta et al. 2014; Zhang et al. 2019; Wei et al. 2022; Wang et al. 2023). While these methods often indicate a correlation between spectral features and aggregation state, the specific spectral changes directly linked to aggregation and the detection limits for low aggregate levels remain uncertain. Hence, despite the reported evidence of Raman effects corresponding to protein aggregation, a reliable method for real-time quantification of protein aggregates in downstream process operations using structural markers is not yet fully established.

To be used for real-time process monitoring via spectroscopic techniques, multivariate regression models have been established for various separation processes such as precipitation (Dietrich et al. 2024), chromatography (Rüdt et al. 2017), or filtration (Rolinger et al. 2020b). Raman spectroscopy has previously been used in bioseparations, but has been shown to have limited accuracy due to the inherent variation in spectral signals (Feidl et al. 2019; Rolinger et al. 2020b, 2023). Previous studies have shown that these limitations can be circumvented when combining chemometric models with mechanistic process models (Rolinger et al. 2023; Schiemer et al. 2023) or using machine learning models such as convolutional neural networks (CNNs) (Schiemer et al. 2024).

To address these limitations, in this study, protein aggregates are generated under controlled conditions, using bovine serum albumin (BSA) as a model system, to isolate and characterize their spectral signatures. The objective is to identify reliable spectral markers capable of distinguishing between native and aggregated BSA species. These markers should enable a quantitative assessment of monomer and aggregate content. Subsequently, these markers are evaluated for their suitability in the inline monitoring of the separation of BSA monomers and dimers via anion-exchange chromatography (AEX). Finally, the recorded spectral data are used to train chemometric partial least squares (PLS) and CNN models that enable real-time discrimination and quantification of BSA size variants.

2 | Materials and Methods

2.1 | Experimental

2.1.1 | Forced Degradation by Heat Incubation

Forced degradation was performed to generate BSA aggregates and to identify Raman spectral features associated with aggregation. A BSA stock solution at a concentration of approximately 60 g/L was prepared by dissolving lyophilized BSA (Sigma-Aldrich, St. Louis, Missouri, the United States) in ultrapure water (PURELAB Ultra, ELGA LabWater, Celle, Germany). The stock solution was subsequently diluted 1:2 using a mixture of 500 mM NaCl solution and ultrapure water to prepare BSA samples with final NaCl concentrations of 0, 50, 100, and 250 mM, maintaining a total protein concentration of approximately 30 mg/mL. Based on these experiments, a NaCl concentration of 150 mM was identified as providing reproducible aggregation kinetics and was therefore selected for all subsequent studies. Aliquots of these BSA solutions were transferred into 1 mL Eppendorf tubes (Eppendorf SE, Hamburg, Germany), which were incubated at 65°C for durations of 0, 30, 60, 90, and 120 min in a stationary state (without agitation) using a heat block (Eppendorf SE, Hamburg, Germany). Each condition was prepared and incubated in triplicate. Following incubation, samples were filtered using 0.2 μm cellulose acetate syringe filters (Sigma-Aldrich, St. Louis, Missouri, the United States). Raman spectroscopy measurements were conducted offline and automated using a custom-built automated Raman measurement platform (ARMP) (described in Section 2.1.4). Residual samples were diluted 1:100

and analyzed by ultrahigh-performance–size-exclusion chromatography (UHP-SEC) for reference analytics.

2.1.2 | *Mixing Studies*

Mixing studies were conducted to evaluate whether previously identified spectral features are sustained while varying the total protein concentration and aggregate content in an independent manner. Therefore, an aggregate-enriched BSA solution was prepared by heat incubation. A 300 mL BSA stock solution was prepared (225 mL centrifuge tubes, Thermo Scientific, Waltham, the United States) at a protein concentration of 30 g/L and a NaCl concentration of 50 mM. All subsequent experiments were conducted at a NaCl concentration of 150 mM, chosen as it provided reproducible aggregation kinetics, while heating time was systematically varied to control aggregate formation. A 200 mL aliquot was incubated at 65°C for 24 hours (Incubated orbital shaker, Thermo Scientific, Waltham, the United States) to induce aggregation, while the rest was left unstressed and stored at 4°C until further use. After incubation, the stressed solution was filtered using a 0.2 μm cellulose acetate vacuum filtration membrane (Sigma-Aldrich, St. Louis, Missouri, the United States).

The methodology resulted in two BSA stock solutions with concentrations of 30 g/L and 19.85 g/L, containing 25% and 50% aggregates, respectively, as determined by the ratio of the aggregate peak area to the total peak area in the UHP-SEC analysis. This aggregate yield in the stressed stock is considerably lower than for the forced degradation experiments. This is believed to be due to the scaling of the volume from 1 mL to 200 mL. The lower aggregate yield restricted the covered aggregate range in the mixing study.

For the remainder of this manuscript, we will refer to these as the unstressed stock and stressed stock. To prepare samples where the protein concentration and aggregate fraction are varied independently, a Latin hypercube sampling (LHS) design with two dimensions, with 48 samples, was used. The LHS was scaled to cover the ranges from 0 to 15 g/L total protein concentration and 25%–50% aggregates. Each condition was prepared in duplicate at a total sample volume of 800 μL, resulting in 96 samples in total. For each sample, a Raman spectrum was collected using the ARMP, with 20 acquisitions averaged in the mixing experiments to improve the signal-to-noise ratio (SNR). The aggregate content was determined via UHP-SEC. Comprehensive information on sample compositions, mixing ratios, concentrations, and the replicate structure is provided in Table S1. The corresponding experimental workflow is depicted in Figure 1.

2.1.3 | *Bind-Elute AEX Experiments*

To generate process data for the separation of BSA monomers and aggregates, bind-elute AEX experiments with inline Raman measurements were conducted using a 5 mL Eshmuno Q column (Merck KGaA, Darmstadt, Germany) on an Äkta Pure system (Cytiva, Uppsala, Sweden). Unstressed BSA stock solutions at a concentration of approx. 15 g/L and 25% aggregate content, prepared in 20 mM Tris, pH 8, were used as feedstock. The column was equilibrated for 5 column volume (CV) using 20 mM Tris, pH 8, prior to sample loading. Following sample loading, the column was washed for 5 CV using equilibration buffer. Gradient elution was performed at different lengths ranging from 0 to 500 mM NaCl, followed by a strip phase at 1 M NaCl for 5 CVs. The column was subsequently regenerated using 1 M NaOH for 5 CVs and reequilibrated using equilibration buffer. During the elution phase using a flow rate of 1 mL/min, fractions were collected at intervals of 200 μL, and Raman spectra were recorded at an exposure time of 500 ms at a laser power of 495 mW, resulting in 24 acquisitions per fraction.

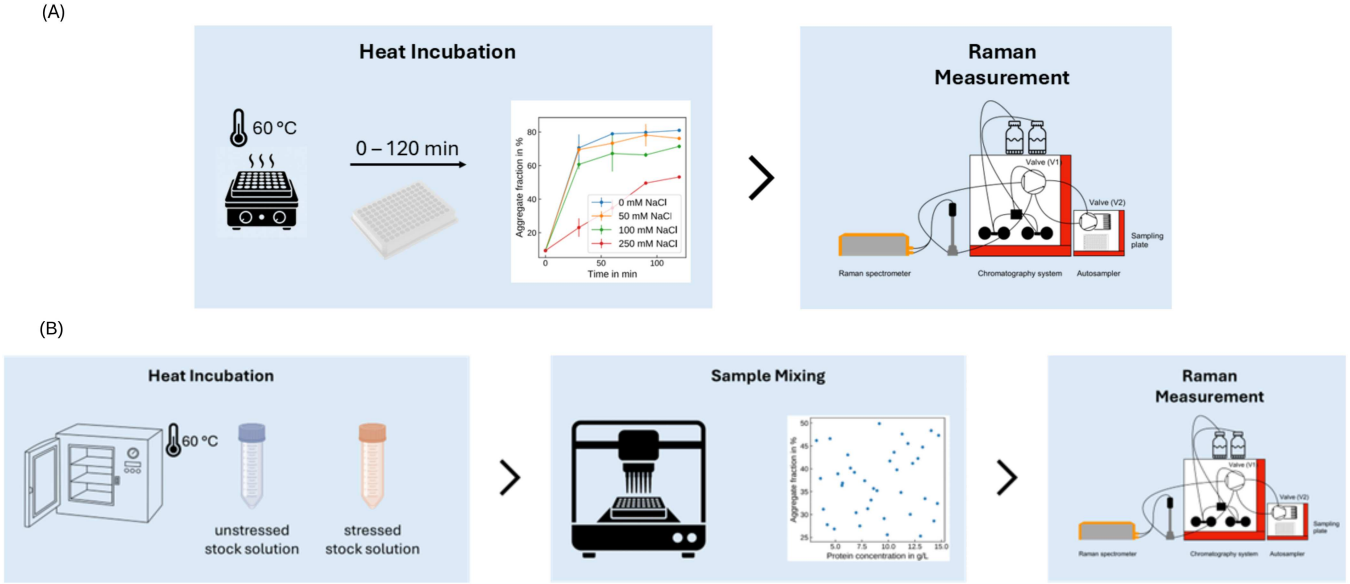

FIGURE 1 | Schematic representation of the forced degradation (A) and mixing study workflows (B). (A) Heat-induced BSA aggregation was generated by incubating samples at 60°C for 0–120 min at defined NaCl concentrations, followed by offline Raman spectroscopy and UHP-SEC reference analytics. (B) To assess spectral features under independently varied protein concentration and aggregate levels, unstressed and heat-stressed BSA stock solutions were combined according to a Latin hypercube design spanning 0–15 g/L protein and 25%–50% aggregates, with Raman spectra and UPSEC data collected for all conditions.

In contrast to the spectra generated during the mixing studies, where 20 acquisitions were averaged to improve the SNR, single 500 ms acquisitions were used during inline AEX to preserve temporal resolution. These single acquisitions were subsequently averaged over the time intervals corresponding to each UPLC-analyzed fraction. The Raman spectrometer was incorporated into the flow path between the outlet valve and the fractionator. In total, 7 AEX experiments were performed using different loading densities, gradient lengths, and feedstocks, which were prepared from two different batches. An overview of the experimental conditions is presented in Table 1.

2.1.4 | *Automated Raman Measurement Platform*

Offline Raman spectroscopy measurements were conducted using an in-house developed system referred to as ARMP. The platform consists of an Äkta purifier system equipped with an autosampler A-905 combined with a Raman spectrometer. The Raman BioReactor BallProbe, inserted a Raman flow cell with a dead volume of 240 μL (both MarqMetrix, Seattle, Washington, the United States), was connected to a HyperFlux^{TM} PRO Plus 785 spectrometer, operated by SpectralSoft 3.2.600.1 software (Tornado Spectral Systems, Mississauga, Ontario, Canada). The ARMP automatically injects samples into the Raman flow cell and collects Raman spectra using predefined configurations. For technical details about the ARMP, we refer the reader to Heyer-Müller et al. (2025).

2.1.5 | *Raman Spectroscopy—Acquisition Settings*

All measurements were conducted using the maximum laser power of 495 mW. For the forced degradation experiments and the mixing studies, the exposure time was set to 1200 ms per acquisition. Each collected spectrum was averaged from 20 acquisitions. For inline measurements during AEX experiments, the exposure time was set to 500 ms per acquisition. The raw Raman signal was treated by a Y-correction and a cosmic ray filter as directly implemented in the operating software SpectralSoft. The Raman spectrometer was calibrated using a white light source at regular intervals.

2.1.6 | *UHP-SEC*

To enable precise determination of size-variant distributions, samples were analyzed via UHP-SEC. Therefore, a TSKgel SuperSW mAb HTP column (4 μm particle size, 4.6 × 150 mm) operated at a flow rate of 0.3 mL/min was used. The mobile phase consisted of 15 mM sodium phosphate buffer at pH 6.2. Before injection, all samples were diluted to achieve a protein concentration within the range of 0–1 mg/mL. Analysis was conducted using a Vanquish UHPLC system controlled by Chromeleon software (version 7.2) (both Thermo Fisher Scientific, Waltham, Massachusetts, the United States).

2.2 | Data Analysis

All data analysis and computations were performed in Python 3.12.7.

2.2.1 | *Spectral Preprocessing*

Raman spectral processing consisted of multiple steps, including cropping, normalization, baseline and background correction, smoothing, and derivative computation. Each operation and its rationale are detailed below, with processing chains optimized based on empirical testing across different data types.

All spectra were first cropped to 500–3250 cm^{-1}, followed by normalization to the water band at 3250 cm^{-1}. Truncation was performed to remove irrelevant spectral regions and improve the quality of correction steps downstream. The lower bound of 500 cm^{-1} was chosen as smaller wavenumbers solely contain baseline drift and bands stemming from the sapphire glass built into the probe head. The upper bound of 3250 cm^{-1} corresponds to the maximum of the water band in the covered bandwidth (Furić et al. 2000) and was hence used for normalization. Normalization was done to compensate for small variations in total intensity potentially induced by turbidity (Dietrich et al. 2024). Subsequently, background subtraction was performed. For forced degradation and mixing experiments, a water spectrum was subtracted. For AEX experiments, a buffer spectrum originating from the start of the respective elution phase was subtracted. Background subtraction was performed to remove the overlapping water and buffer bands and enable better visualization of the spectral effects. Furthermore, baseline correction was performed using a Whittaker filter through the derivative peak-screening asymmetric least squares algorithm (DERPSALSA) algorithm, as implemented in the *pybaselines* library (v. 1.1.0), with parameters $\lambda = 10^5$, $k = 0.02$, and $d = 2$ (Erb 2025). Finally, the Savitzky–Golay (SG) filter as implemented in *scipy* (v. 1.14.1) was employed both for spectral

TABLE 1 | Experimental conditions for the bind-elute AEX experiment used in the inline Raman monitoring of BSA monomer and aggregate separation.

Run ID	Gradient length (CV)	Load density (mg/mL resin)	Total feed concentration (mg/mL)	HMWC content (%)	Dataset (–)
Batch 1 5 CV	5	15	15	25.33	Train
Batch 1 15 CV	15	22.5	15	25.33	Train
Batch 2 10 CV	10	30	15	15.67	Train
Batch 1 10 CV	10	17.5	15	25.33	Test
Batch 1 Step	Step	22.5	15	25.33	Test
Batch 2 5 CV	5	22.5	15	15.67	Test
Batch 2 15 CV	15	45	15	15.67	Test

smoothing and derivative computation to enhance subtle spectral variations. Here, a second-order polynomial and a window size of 21 were used in all cases. In the following, spectral data are displayed either after smoothing only or after second derivative computation.

2.2.2 | *Evaluation of Spectral Bands and Correlation*

To evaluate the changes of spectral bands and their correlation with the aggregate content, multiple spectral markers or metrics were used throughout this study.

It is known that the amide I and III vibrations carry the most structural information of proteins, which are located at 1500–1800 cm^{-1} and 1200–1400 cm^{-1}, respectively (Lippert et al. 1976; Williams 1986; Rygula et al. 2013). While the amide I region is largely overlapped with the water bending vibration, the amide III region is largely unaffected by water and, depending on the used buffer system, free of other interfering species. In addition, the Fermi doublet of tryptophan side chains is known to be located at the frequencies of approx. 1340 and 1360 cm^{-1} (Rygula et al. 2013) and to provide an informative marker of local solvent environment (Harada et al. 1986; Eisenberg and Juszczak 2012). Given the local superposition of the amide III and tryptophan vibrations, we chose to observe the ratio of the intensities at 1341 and 1320 cm^{-1} as the most prominent bands in this region. We will refer to this intensity ratio as the amide III ratio for the remainder of this manuscript.

To compress the information contained in the amide I band into a single spectral marker, the center of mass was calculated as the intensity-weighted average wavenumber within the chosen spectral range (1500–1800 cm^{-1}),

$$\bar{\nu} = \frac{\sum_i I_i \nu_i}{\sum_i I_i} \tag{1}$$

where I_i and ν_i are the intensity and wavenumber at point i, respectively. The center of mass is a common marker summarizing observed spectral shifts in an easy-to-comprehend manner. To improve the precision of the calculation, the acquired spectra were interpolated to a resolution of 1000 steps in the given spectral range.

Finally, the SNR was used as a correlation metric between the spectral intensity and the aggregate content. The SNR was calculated as

$$SNR = \frac{\mathrm{Var}(\mathbf{x}_\lambda \hat{\boldsymbol{\beta}})}{\hat{\sigma}^2} \tag{2}$$

with $\hat{\boldsymbol{\beta}}$ being the regression coefficients of a univariate linear model of type $\mathbf{y} = \mathbf{x}_\lambda \boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with normally distributed errors ε and $\hat{\sigma}^2$ being the residual variance of the linear model for wavenumber λ according to Soch and Allefeld (2018). To calculate SNR for the Raman spectroscopy data using Equation (2), a linear model was built for each wavenumber individually. A spectral feature is considered to be correlated with aggregate content if it shows high SNR. Low values indicate a high degree of noise and no linear correlation. For all spectral markers in the mixing study (Figure 3), error bars represent the standard deviation of two independent replicates, each derived from an averaged spectrum comprising 20 acquisitions.

2.2.3 | *Regression Modeling*

Data Split Regression models were calibrated using inline collected data from AEX experiments with their respective assignments to training and test sets presented in Table 1. Raman spectra were preprocessed as laid out in Section 2.2.1 and used as inputs to the regression models. As outputs, the absolute concentrations of monomers and aggregates were chosen as determined by UHP-SEC for all collected fractions. A leave-one-batch-out (LOBO) scheme was chosen for cross-validation for all regression models. The LOBO cross-validation iterates through all experiments included in the training set, leaving each of them out once. The LOBO scheme has been found to provide more realistic estimates of prediction errors than more common *K*-fold cross-validation schemes and was hence selected for model evaluation (Rolinger et al. 2020b; Dietrich et al. 2024).

PLS Regression PLS regression models were calibrated as single-output (PLS1) and multi-output (PLS2) models. To improve model accuracy, additional preprocessing was applied to the input data. Additional preprocessing involved truncating the Raman spectra to 600–1800 cm^{-1}, followed by SG filtering with a window size of 15 and varying derivative orders (0, 1, or 2). Finally, the input data was mean-centered. The number of latent variables and the SG derivative order were optimized during cross-validation by grid search in the ranges 2–14 and 0–2, respectively. In total, eight different PLS model variants were calibrated. For each combination of preprocessing operations, a PLS1 and PLS2 model was calibrated. Aside from the previously mentioned preprocessing operations, there are two PLS variants that are calibrated using synthetic data. For more details on the generation of these synthetic data, we refer to Section 2.2.3.

CNN CNN models were calibrated as multi-output models, predicting the monomer and aggregate concentrations simultaneously. The CNNs were implemented in *tensorflow* (v. 2.11). The model architecture and training methodology were similar to Schiemer et al. (2024). The model architecture consisted of three convolutional layers with defined kernel sizes of 15, 7, and 3 and a rectified linear unit (ReLU) activation function. Average pooling with a pool size of 2 was employed after each convolutional layer. Each convolutional layer was configured to use 5 kernels. After the convolutional block, a flattening operation was implemented to project all kernel outputs onto a single vector. A fully connected (FC) layer with 50 units and a ReLU activation function was used as the subsequent layer. Finally, the output layer was configured to use 2 FC units and a ReLU activation function.

CNN models were calibrated with and without synthetic data augmentation, which will be denoted CNN and CNN-Aug for the remainder of this manuscript. For non-augmented CNN and the augmented CNN, training was conducted over 500 and 50 epochs, respectively, with a batch size of 32 using the mean squared error (MSE) as the loss function. The target variables were standardized using a min–max scaler to ensure equal contribution of both targets despite absolute concentration differences. The learning rate was initially set to 0.001, and a

learning rate scheduler was configured to halve the learning rate if the validation loss was not reduced for 5 consecutive epochs. The non-augmented CNN used the original training data as input data. To be able to compute the validation loss at every epoch of the training cycle, the training data was randomly divided into calibration and validation sets, reserving 25% of the original training data points for calculating the validation loss. In the case of the augmented CNN, data augmentation was employed to generate a set of 10,000 synthetic data points, which served for calibration. The validation loss was calculated using 100% of the original training data. For the details of the data augmentation method, we refer to Section 2.2.3.

Data Augmentation Data augmentation was used to enhance model robustness, utilizing the local subset augmentation (LSA) technique introduced in Schiemer et al. (2024). In short, the LSA technique extracts the theoretical pure component spectra from a subset of the training data and generates synthetic spectra by recombining the pure component spectra. A local subset size of 39 was used in all instances. The standardized Euclidean distance as given by $\left(d(x,y) = \sqrt{\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{S_i^2}}\right)$ was used as a metric to select the local subset. Random white noise with a standard deviation of 0.0001 intensity counts and normally distributed shifts along the wavenumber axis with a standard deviation of $0.25 cm^{-1}$ were further introduced to the synthetic spectra.

Error Metrics Model performance was assessed using the R^2 and root mean squared error (RMSE) metrics. In all cases, these metrics are reported separately for the monomer and aggregate species. Additionally, R^2 and RMSE are provided for the aggregate fraction as determined by $v_{\text{agg}} = c_{\text{agg}}/(c_{\text{agg}} + c_{\text{mono}})$, where c_{mono} and c_{agg} denote the absolute concentration of monomer and aggregate species given in g/L, respectively. To avoid division by small values, v_{agg} was calculated only for samples where the denominator summed to values greater than $0.1\ g/L$. The metrics for cross-validation were calculated based on the predicted values for the iterations when the respective experiments were assigned to the validation set.

Augmentation Test To assess the selectivity of the evaluated models, an augmentation test was conducted using the LSA technique described in Section 2.2.3 to evaluate whether the models correctly predict monomer and aggregate combinations that were not covered in the experimental data. To provide a challenging dataset, the synthetic spectra were generated based on the experimental data originating from the test set. The LSA algorithm was used in the same configuration as laid out in Section 2.2.3. The monomer and aggregate combinations covered in the augmentation test were sampled from a uniform distribution in the ranges [0, 20] and [0, 5] for the monomer and aggregate species, respectively. In total, 1000 random samples were drawn for which synthetic spectra were generated. The calibrated models were finally used to predict the monomer and aggregate concentrations from the synthetic spectra, and the model accuracy was evaluated based on the sampled values as the ground truth. The distribution of the sampled values is visualized in Figure S1. This methodology makes use of the test set data by diversifying the obtained data through synthetic amplification. Further, it enables a more systematic evaluation of the model selectivity for the individual species.

Test Set Rotation Test To further compare the robustness and accuracy of the different model variants, we conducted a test set rotation based on a Leave-*P*-Batches-Out scheme with $P = 4$. With 7 experiments in total, this yielded 35 different train-test splits. We conducted the rotation test with the best-performing PLS model and the two different CNN models, one without and one with data augmentation. In each rotation, the respective model was calibrated using the selected training batches. For both model types, the methodology introduced in Section 2.2.3 was used. For the PLS model, the number of components and the derivative order were fixed to the optimal setting found for the training–test split used in the rest of this study; that is, no individual optimization was performed for each rotation. The optimal number of components was 2 and 14, and the order of derivative was 0 and 1 for the monomer and aggregate species, respectively, for the Crop-SG-PLS1 model.

3 | Results and Discussion

3.1 | Forced Degradation Studies

The primary objective of forced degradation studies was to identify Raman spectral features associated with aggregation. Therefore, samples with a concentration of 30 g/L BSA were exposed to thermal stress at 65°C and analyzed for their aggregate content via UHP-SEC and Raman spectroscopy. Figure 1A schematically presents the methodology and the aggregate content over time. The UHP-SEC results illustrated in Figure 1 demonstrate a time-dependent formation of aggregated species, which saturates after roughly 90 min. Comparative retention time analysis showed that heat-induced aggregates elute earlier than native aggregates, suggesting the formation of higher-order multimers (data not shown). Notably, the data indicate that with increasing ionic strength, heat-induced aggregate formation decreased. Forced degradation has been commonly used in Raman spectroscopy studies for aggregate detection (Boulet-Audet et al. 2014; Barnett et al. 2015; Zhou et al. 2015), as batch-to-batch variations do not cover the experimentally interesting range to study the spectral effects of protein aggregation. However, the aggregates produced by, for example, heat stress are not necessarily structurally coherent with naturally occurring aggregates. Nevertheless, forced degradation is a helpful tool to efficiently produce aggregated proteins for further investigation. The authors considered the use of non-native aggregates sufficient for initially investigating spectral features affected by the structural changes induced by aggregation, but acknowledge the need for validation in native aggregate species.

To further investigate the structural changes affected by aggregation, offline-recorded Raman spectra of the thermally stressed samples were evaluated. Figure 2A,F shows background-corrected and second derivative Raman spectra for the different sampling time points, respectively. The upper panels in (B) and (G) depict the SNR between the single spectral features and the aggregate content. The top correlative features together with manually selected (manual selection was based on visual inspection of all wavenumber ranges, in addition to those previously identified in literature) ones are shown in the lower panels (B–E) and (G–K) over time with their respective standard deviations. In the background-corrected spectra (cf. Figure 2A), the top correlative features appear in the region of the phenylalanine band between 990 and $1010\,cm^{-1}$. Spectral

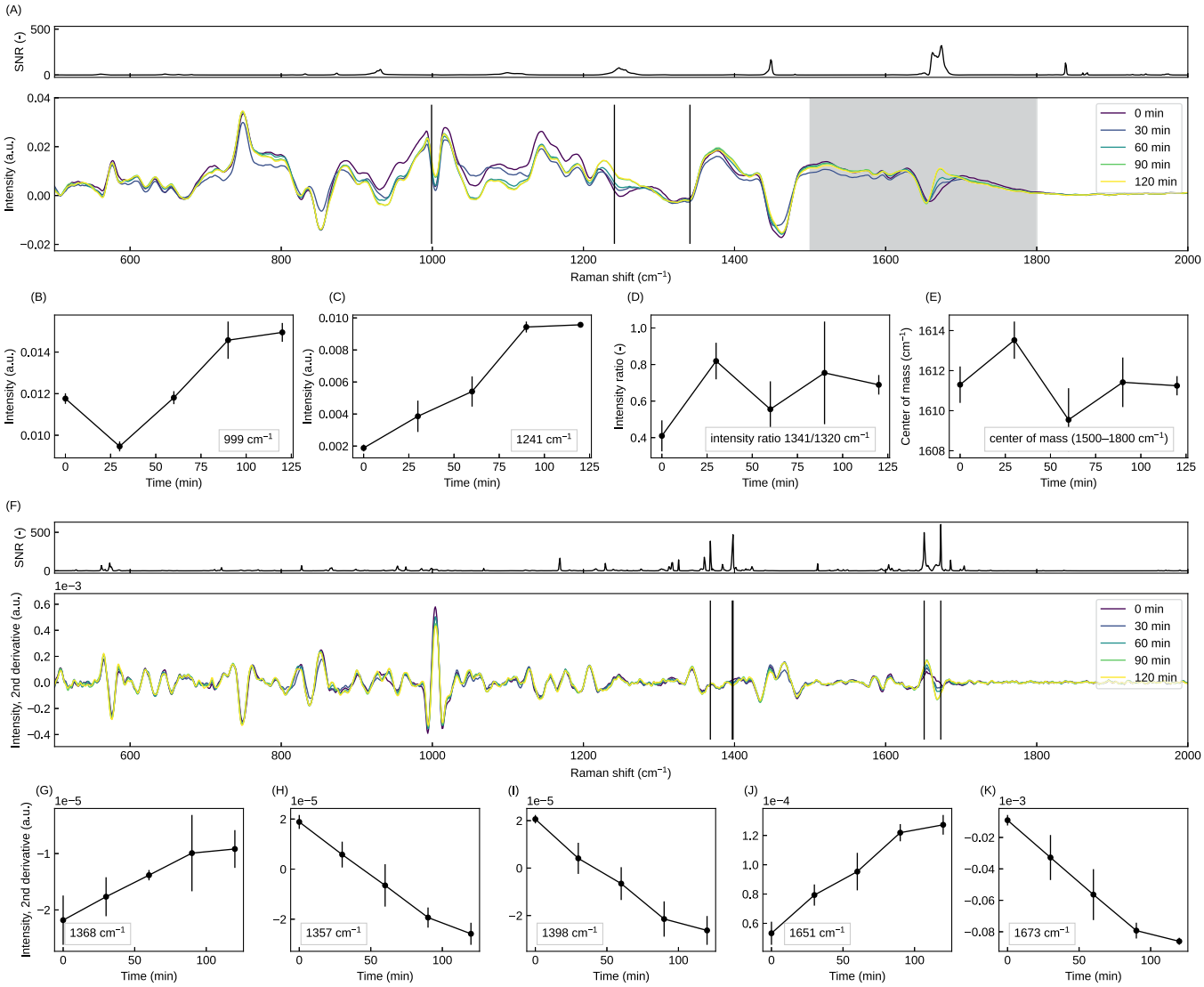

FIGURE 2 | Results of the forced degradation experiment. Offline recorded Raman spectra and SNR for each wavenumber (A). Spectra are shown after preprocessing as described in Section 2.2.1. The color gradient reflects the increasing duration of heating. Intensities at 999 and 1241, the ratio 1341/1320 cm^{-1}, and the center of mass of the amide I region (1500–1800 cm^{-1}) were extracted from the preprocessed Raman spectra and highlight the evolution over time (B–E). Second-derivative Raman spectra and corresponding SNR (F). Temporal evolution of selected wavenumbers (1368, 1397, 1398, 1651, and 1673 cm^{-1}) corresponding to the top five SNR values (G–K).

features in the amide regions (amide III at 1241 cm^{-1} and amide I between 1550 and 1700 cm^{-1}, as expressed by the center of mass show minor correlations with higher standard deviations. The center of mass of the amide I band was calculated according to Equation (1). The amide III ratio (I_{1340}/I_{1321}) was not affected. When additionally treated with the second derivative SG filter, the SNR displays a high degree of correlation in the fingerprint region, mainly between 1200 and 1400 cm^{-1}, as well as in the amide I between 1550 and 1700 cm^{-1}, with values of up to 300 and 500, respectively.

The amide I region has previously been identified as a reliable marker for aggregation in Raman spectroscopy studies (Zhou et al. 2015; Shivu et al. 2013). Especially for proteins with predominantly α-helical structures such as BSA, the amide I band is considered a reliable structural marker. In particular, the formation of intermolecular β-sheets is known to cause a shift in the amide I band with decreasing intensity around 1650 cm^{-1} and increasing intensity around 1670 cm^{-1}. This shift may either be observed in second-derivative spectra thanks to the enhanced selectivity by this treatment or on a macroscopic scale by a shift in center of mass (Militello et al. 2004). Moreover, the amide I is the most promising structural marker in Raman spectroscopy of proteins due to the high intensity compared to amide II and III vibration modes or those of tyrosine and tryptophan side chains (Zhou et al. 2015). However, the overlapping water band, which was removed in this analysis, may pose a challenge in determining accurate amide I patterns as the water band may be affected by the solution environment, such as salt concentrations (Fontana et al. 2013).

3.2 | Mixing Studies

Following the identification of Raman spectral features associated with aggregation, samples were prepared from an unstressed and a stressed BSA stock solution, where protein concentration and aggregate content were systematically varied

(cf. Section 2.1.2) to evaluate the robustness of the spectral effects found in Section 3.1. LHS was employed as a statistical approach to efficiently cover the possible design space with a given number of samples. For each sample, Raman spectroscopy measurements were subsequently performed offline using the ARMP. Figure 1B schematically depicts the described workflow, and Figure 3A–F presents the strongest correlations of individual spectral features with overall protein concentration and aggregate content.

The spectral intensities at 999 cm^{-1}, representing the phenylalanine band, closely follow protein concentration in the samples with standard deviations below 5 %. A slight nonlinear trend toward higher concentrations, along with a few outliers at high concentrations, can be observed. The nonlinearity highlights the complexity of Raman spectroscopy data, as the raw signals undergo multiple preprocessing steps, which may influence the extent of nonlinearity observed. Amide I-related spectral features, such as the 1673 cm^{-1} bands in the second derivative spectra, as well as the center of mass of the amide I, linearly follow the aggregate content. In all instances, outliers can be observed. These outliers mainly originate from samples with low protein concentrations (< 3 g/L), where analysis becomes more challenging due to the overall low signal intensity. These results corroborate findings from forced degradation experiments and demonstrate the potential of the identified spectral features as reliable markers for quantifying protein aggregation.

The selection of the 1673 and 1651 cm^{-1} features in Figure 3 reflects the outcomes of the forced degradation study (Figure 2), where the amide I region exhibited the strongest and most consistent correlation with aggregation. These wavenumber regions correspond to opposing components of the amide I band associated with shifts from native α-helical structure (approximately 1651 cm^{-1}) toward intermolecular β-sheet formation (approximately 1673 cm^{-1}). Their inclusion in the mixing-study analysis therefore provides a mechanistic link to the structural transitions identified earlier and enables an evaluation of whether these aggregation-related spectral changes remain robust when protein concentration and aggregate content are varied independently.

Beyond amide I-related markers, various Raman intensity ratios involving protein-specific bands were investigated as potential indicators of protein aggregation. Specifically, the amide III ratio (I_{1340}/I_{1321}) and the peak maximum within the 920–950 cm^{-1} range were assessed (Ashton and Blanch 2010; Militello et al. 2004; Li and Li 2009; Zhou et al. 2015; Alizadeh-Pasdar et al. 2002). However, both showed only weak correlations with aggregate levels, accompanied by large standard deviations (> 20%) and the presence of significant outliers, limiting their reliability. Although the 1341/1320 cm^{-1} ratio is included for completeness because it has been reported as a potential aggregation-sensitive marker in prior Raman studies, our results demonstrate that it is not consistently responsive under the conditions tested here. We therefore treat it as an auxiliary rather than a primary structural marker and do not rely on it for quantitative interpretation.

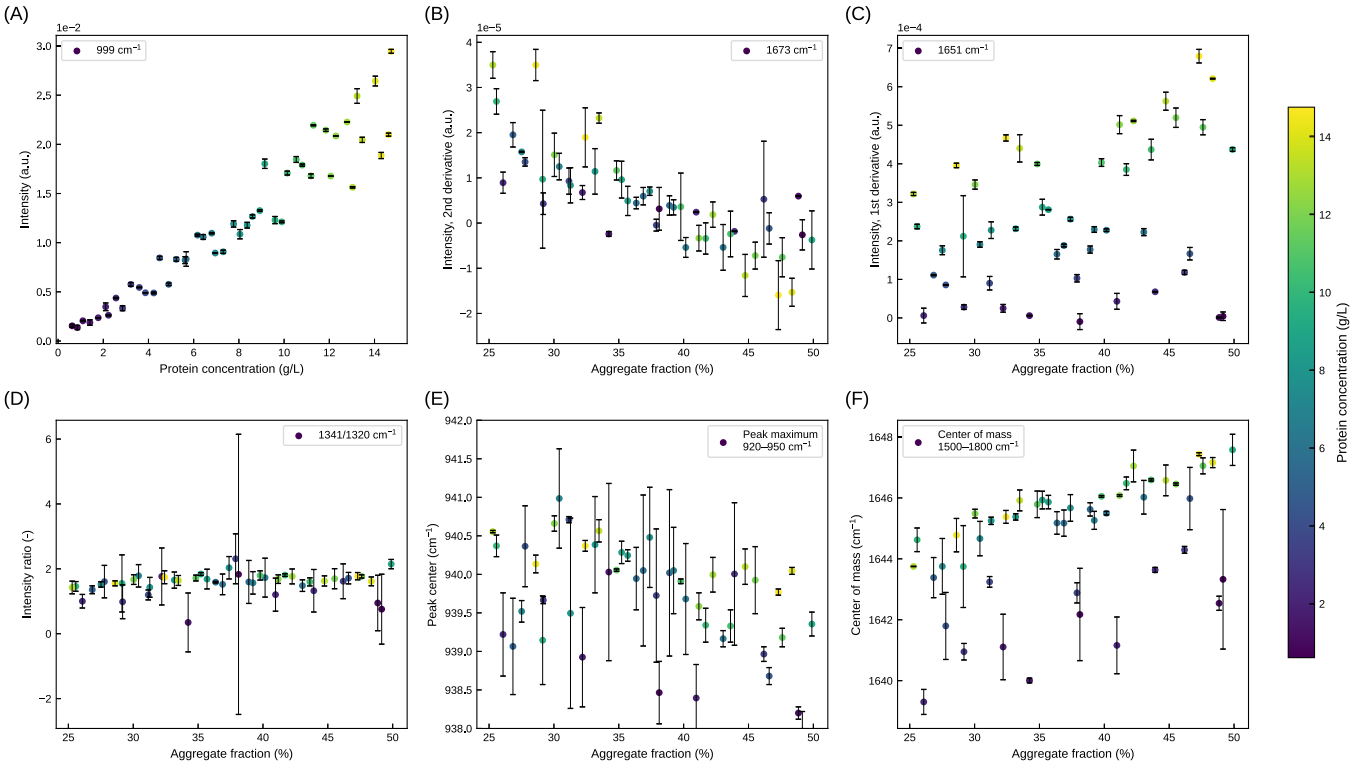

FIGURE 3 | Spectral features observed in offline-recorded Raman spectra of BSA obtained during mixing studies, preprocessed as described in Section 2.2.1. Raman intensities, peak shifts, and band ratios are plotted as functions of total protein concentration (g/L), aggregate fraction (%), and high molecular weight component (HMWC) content (%). Error bars represent standard deviations of replicate measurements. Color scale from dark to light indicates increasing protein concentration (g/L). The Raman bands or markers 1004, 1673, 1651, 1341/1320, Peak Maximum of 920–950 and the center of mass of 1500–1800 are shown in (A–F), respectively.

These findings highlight that while spectral features beyond the Amide I band can provide supplementary structural insights into aggregation processes, their robustness and reproducibility for aggregate quantification remain limited. The phenylalanine band, as well as the amide I region, remains the most reliable spectral marker for accurately monitoring aggregation dynamics. In contrast, other markers showed high variability and limited statistical significance, reducing their suitability for quantitative analysis.

3.3 | Bind-Elute AEX Experiments

3.3.1 | Qualitative Spectral Analysis

Building on insights from the mixing studies, bind-elute AEX experiments were conducted to assess inline Raman detection capabilities for naturally occurring BSA aggregates under operational conditions. Chromatographic conditions comprised gradient elutions with varying loading densities from two BSA batches and one step elution (cf. Table 1). Continuous Raman spectra acquisition during the elution phases facilitated direct comparison with UHP-SEC analyses of collected fractions. Compared to the heat-induced aggregation system, where aggregates with higher-order structures formed rapidly at elevated temperatures and low ionic strength, the chromatography load material was solely comprised of monomeric and dimeric BSA species.

In Figure 4, fraction analytics from all AEX experiments are overlaid with selected spectral markers, namely the amide III ratio I_{1341}/I_{1320} and the center of mass of the amide I band, obtained from inline acquired Raman spectra. The raw spectra were averaged over the collection interval of each fraction and preprocessed the same way as in the forced degradation and mixing experiments. Due to the buffer system used in the chromatographic separations, the spectral background correction was performed using the first acquired spectrum in the elution phase. In case the previously defined spectral markers prove selective for aggregate detection, we expect the following behavior during AEX elutions: Due to the timely separation of monomeric and dimeric species, the spectral markers would initially remain constant until the co-elution of the dimeric species, which is more strongly retained by the chromatography resin. When the co-elution begins, the spectral marker is expected to increase as the relative content of the dimeric species is successively increasing until the elution peak ends. For non-separating conditions, the spectral markers are expected to remain constant over the entire elution phase. The absolute value depends on the aggregate content in the elution pool.

In Figure 4, both the amide III ratio and the center of mass follow the described behavior for all tested chromatographic conditions, despite some non-idealities. For runs from Batch 1 (A–C), the loading densities ranged from 15 to 22.5 g/L resin. For the longer gradients of 10 and 15 CVs (B, C), this led to an eluate pool dilution with maximum monomer concentrations of < 15 g/L. In those cases, the spectral markers appear unstable with a high degree of scattering. As a result, it is vaguely possible to deduce the onset of the co-elution of the dimeric species. For experiments from Batch 2 (D–F), for which loading densities ranged from 22.5 to 40 g/L, the spectral markers show a lower degree of scattering, and fewer outliers were observed. While for the 10 and 15 CV runs (E, F), the actual onset of co-elution after approximately 52 and 62 min, respectively, is

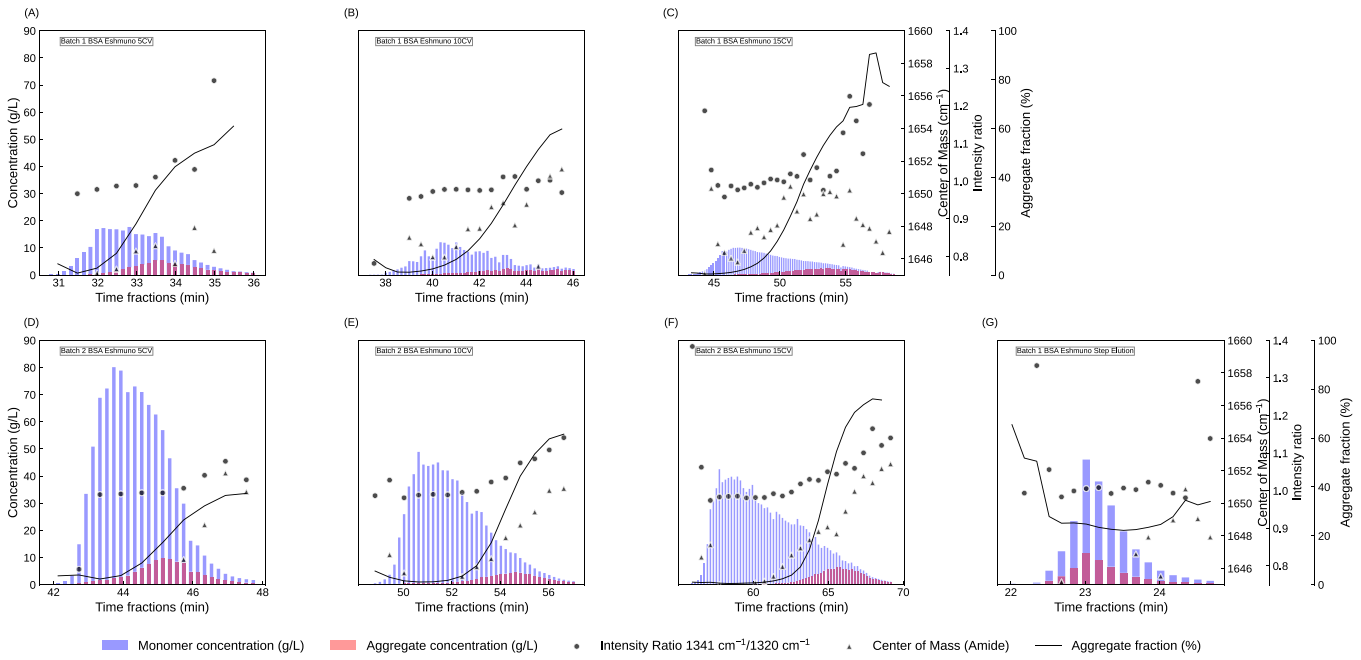

FIGURE 4 | Elution profiles and structural markers of BSA under varying process conditions. Protein separation from two BSA batches was performed using different gradient lengths: (A, D) 5 CV; (B, E) 10 CV; (C, F) 15 CV; and (G) step elution. Blue and red bars represent monomer and aggregate concentrations (g/L), respectively. Black lines indicate aggregate fractions across elution time. Black circles and black triangles denote Raman spectral markers: intensity ratio (1341 cm^{-1}/1320 cm^{-1}) and amide I band center of mass (cm^{-1}), respectively. To improve legibility for panels (A–F), Raman marker traces were smoothed by plotting the mean of three consecutive spectra at the corresponding mean elution time; Panel (G) shows all measured points.

accurately represented by the spectral markers, the spectral markers in the 5 CV run (C) do not match the fraction analytics. Notably, the step elution experiment (G) fulfills the expectations, and the spectral markers remain constant in the width of the elution profile. As the elution occurred within approximately 2 min, the number of collected fractions and corresponding Raman spectra is comparably low. In all experiments, outliers can be observed, which mostly correspond to fractions at the front and end of the elution peak with low concentrations (< 3 g/L).

The depicted selection of spectral markers for AEX experiments is not exhaustive. However, they showed the best qualitative behavior among the studied markers. Notably, previously identified amide I-related markers such as the 1651 and 1673 cm^{-1} intensities in the second derivative spectra were not found to be selective for aggregate content as they were affected by the total protein concentration and therefore followed the overall elution profile. Hypothetically, the higher-order structure aggregates generated during the heat treatment may form more intermolecular β-sheets than natively occurring dimers and hence show a pronounced effect in the amide I region, which upholds variations in protein concentration (Barnett et al. 2015; Zhou et al. 2015). Inversely, the amide III ratio, which showed no significant effect in forced degradation or mixing experiments, served as a valid indicator in AEX experiments. This is likely because pH and salt gradients in AEX can alter hydrogen bonding and polarity around tyrosine residues, enhancing spectral differences between monomeric and aggregated forms (Antosiewicz and Shugar 2016). Despite the ineffectiveness of the amide I-related single-variable intensities, the cumulative measure of the center of mass proved suitable across all experimental modes. Across the studied literature, the center of mass appears to be the most reliable qualitative spectral marker for protein aggregation. However, it is difficult to use this marker for direct quantification of aggregate content, due to possible nonlinearities or background effects. Especially when used in biopharmaceutical processes, such as chromatography, changing buffer environments or excipients is common. Even in AEX gradient elutions, the salt concentration is constantly changing, which has an effect on the water band (Furić et al. 2000) and therefore affects background correction steps in Raman spectroscopy. In summary, these results present the first direct illustration of how protein aggregation affects Raman spectra in chromatographic separation processes and shows applicability of Raman spectroscopy to monitoring biopharmaceutical downstream operations. Due to concentration variations and spectral noise, the use of Raman spectroscopy for direct quantification necessitates the calibration of statistical models.

3.3.2 | *Statistical Model Building*

To enable quantitative prediction of protein size variants from Raman spectra, multiple regression models were calibrated targeting the concentrations of monomeric and dimeric BSA species. Model training incorporated spectra from three representative bind-elute AEX experiments, while testing was performed on four independent datasets to assess generalizability across varying chromatographic conditions (cf. Table 1). Multiple different statistical model variants were compared using different performance criteria, namely their prediction error on training, cross-validation, augmentation test, and external test sets. For PLS models, every preprocessing variant was tested using a single-output (PLS1) and a multi-output version (PLS2). The studied CNNs were exclusively multi-output. The performance metrics of the different model variants are reported in terms of R^2 and RMSE in Tables 2A and 2B.

Figure S2 presents RMSE values for monomer and aggregate components in absolute terms (A and B), as well as RMSE for the aggregate fraction calculated from the predicted monomer and aggregate concentrations (C). This systematic comparison reveals distinct performance patterns both within and between PLS and CNN model families across different evaluation scenarios. For monomer prediction (A), CNN models demonstrate lower RMSE during training, cross-validation, and augmentation test phases. However, this advantage largely disappears in the external test, where PLS and CNN models achieve comparable performance. This convergence suggests that the apparent superiority of CNN models in earlier phases may reflect overfitting rather than genuine generalization capability for monomer prediction. The aggregate prediction results (B) show a more consistent pattern, with CNN models maintaining lower RMSE across all evaluation splits. While individual PLS variants occasionally match CNN performance for specific datasets, no single PLS model achieves comparable accuracy across all scenarios, indicating more robust performance by the CNN approach for this target. Figure S2C confirms this trend for aggregate fraction prediction, with selected PLS models approaching but not consistently matching CNN-type model performance. Notably, data augmentation consistently increases RMSE relative to non-augmented variants across all model types and evaluation scenarios, suggesting that the augmentation strategy employed does not improve generalization and may introduce artifacts that degrade prediction quality.

To better visualize the accuracy of the different model variants, parity plots of observed versus predicted HMWC (%) for all calibrated PLS models are shown in Figures S4–S13, where Panels (A)–(D) represent the model performance on the training, cross-validation, augmentation test, and test datasets, respectively.

To resolve the ambiguity regarding data augmentation effectiveness, we conducted a comprehensive leave-4-groups-out rotation test (Figure S3), evaluating all possible training/test splits for each model type. The results demonstrate that the base CNN performs slightly better than the augmented CNN variant for both monomer and aggregate predictions. Both CNN approaches achieve substantially lower RMSE on average compared to the best-performing PLS model (Crop-SG-PLS1), particularly for aggregate prediction, where the median RMSE is reduced by approximately 50%. The more compact distributions and lower median values for CNN models confirm their superior generalization capability when evaluated across diverse data splits.

Figure 5 presents the model predictions for the Crop-SG-PLS1 and the CNN models in dashed and solid lines, respectively, for training (A), cross-validation (B), and test (C) sets. For the AEX runs in the training set, both models fit the UHP-SEC data accurately with minor deviations. For the Batch-1-5CV run, the PLS model closely follows the sharp dimer elution profile, while

TABLE 2(A) | Coefficient of determination (R^2) for the prediction of monomer and aggregate concentrations across all data splits and models.

Model	Analyte	Train	CV	Augmentation test	External test
PLS1	Monomer (g/L)	0.9761	0.9417	−0.1489	0.8474
	Aggregate (g/L)	0.9765	−0.5345	0.6823	0.6210
PLS2	Monomer (g/L)	0.9911	0.8710	−0.1621	0.8389
	Aggregate (g/L)	0.9878	−0.7969	0.5123	0.5756
SG-PLS1	Monomer (g/L)	0.9709	0.9422	0.1993	0.9030
	Aggregate (g/L)	0.6014	−0.3389	−0.2459	0.0365
SG-PLS2	Monomer (g/L)	0.9670	0.9313	0.5027	0.9323
	Aggregate (g/L)	0.5195	−0.2045	−0.4069	0.0371
Crop-SG-PLS1	Monomer (g/L)	0.9706	0.9529	−0.2536	0.7954
	Aggregate (g/L)	0.9642	0.2039	−2.9519	−0.2070
Crop-SG-PLS2	Monomer (g/L)	0.9772	0.9555	0.3686	0.9108
	Aggregate (g/L)	0.7694	−1.4017	0.8421	0.6350
PLS1-augmented	Monomer (g/L)	0.9632	0.9464	0.9152	0.8779
	Aggregate (g/L)	−0.0479	−0.1968	0.7373	−1.2896
PLS2-augmented	Monomer (g/L)	0.9625	0.9438	0.8955	0.8841
	Aggregate (g/L)	−0.0915	−0.2479	0.7803	−1.2446
CNN	Monomer (g/L)	0.9850	0.9752	0.9126	0.8274
	Aggregate (g/L)	0.9525	0.7804	0.6874	0.6894
CNN-augmented	Monomer (g/L)	0.6775	0.8584	0.7112	0.5756
	Aggregate (g/L)	0.4114	0.3097	−0.0137	0.3051

TABLE 2(B) | Root-mean-square error (RMSE, g/L) for prediction of monomer and aggregate concentrations across all data splits and models.

Model	Analyte	Train	CV	Augmentation test	External test
PLS1	Monomer	1.6568	2.5848	6.0818	6.8610
	Aggregate	0.1945	1.5711	0.8141	1.4377
PLS2	Monomer	1.0121	3.8464	6.1169	7.0501
	Aggregate	0.1399	1.7001	1.0086	1.5213
SG-PLS1	Monomer	1.8253	2.5742	5.0772	5.4720
	Aggregate	0.8007	1.4675	1.6122	2.2922
SG-PLS2	Monomer	1.9438	2.8073	4.0014	4.5705
	Aggregate	0.8792	1.3919	1.7132	2.2914
Crop-SG-PLS1	Monomer	1.8360	2.3248	6.3529	7.9450
	Aggregate	0.2398	1.1316	2.8712	2.5655
Crop-SG-PLS2	Monomer	1.6181	2.2585	4.5088	5.2476
	Aggregate	0.6091	1.9655	0.5739	1.4108
PLS1-augmented	Monomer	2.0550	2.4796	1.6523	6.1382
	Aggregate	1.2983	1.3875	0.7403	3.5335
PLS2-augmented	Monomer	2.0742	2.5386	1.8343	5.9808
	Aggregate	1.3250	1.4168	0.6769	3.4986
CNN	Monomer	1.3103	1.6861	1.6777	7.2968
	Aggregate	0.2765	0.5944	0.8075	1.3015
CNN-augmented	Monomer	6.0814	4.0293	3.0495	11.4429
	Aggregate	0.9731	1.0537	1.4542	1.9467

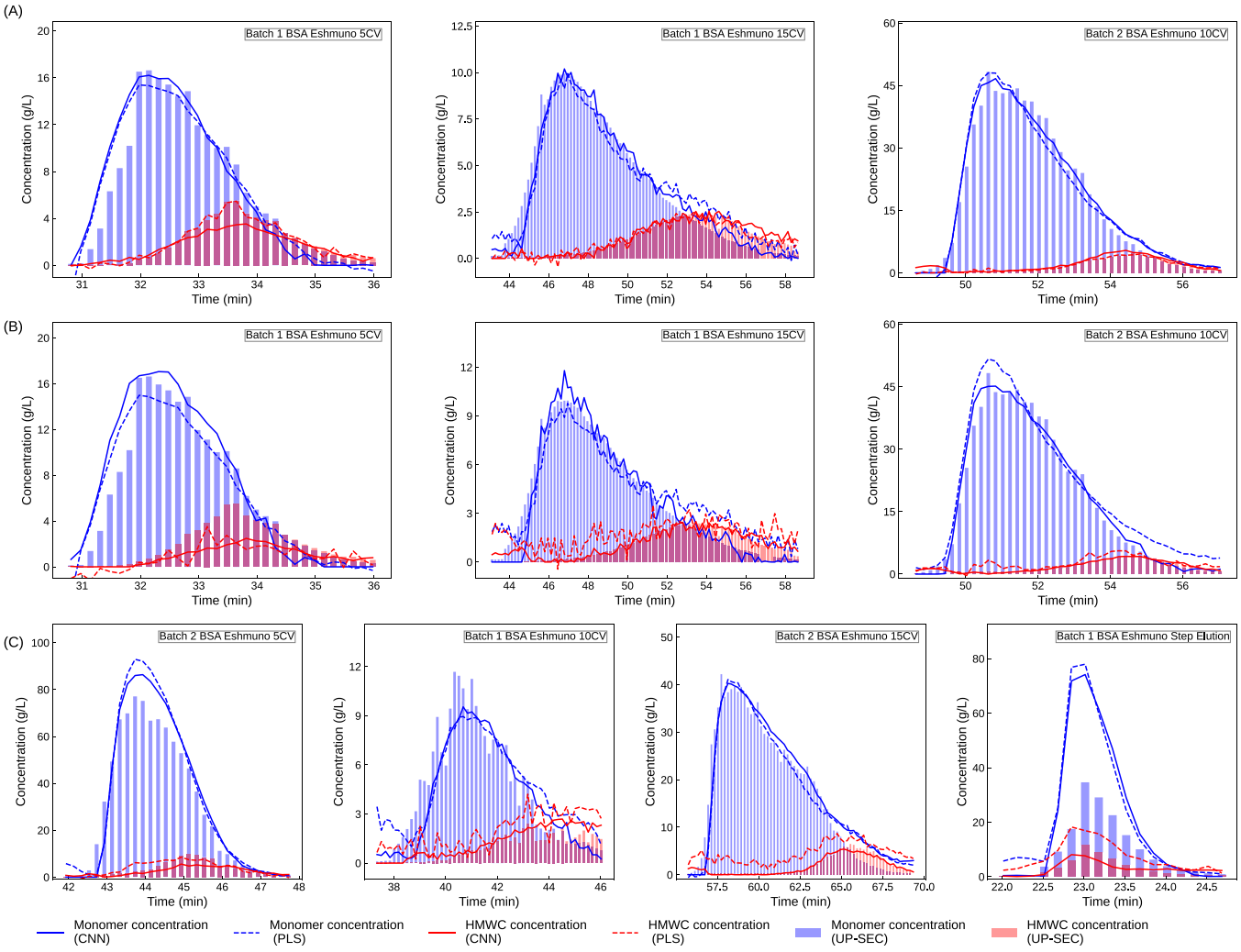

FIGURE 5 | Predictions of Crop-SG-PLS1 (dashed lines) and CNN (solid lines) regression models for training (A), cross-validation (B), and test (C). Blue and red bars represent monomer and dimer concentrations as determined by UHP-SEC.

the CNN model predicts a more flattened, round profile. For the Batch-1-15CV, both models exhibit scattered predictions, with the CNN being slightly less affected for both species. In the Batch 2-10CV run, both models match the UHP-SEC data accurately, with the concentrations being the highest among the training runs. In the cross-validation, the accuracy and robustness of the CNN predictions are superior to the PLS predictions. In addition to the higher degree of scattering of the PLS model, the initial phases where only the monomer is eluting are wrongly predicted by the PLS model, and the selectivity for the monomeric species during co-elution is compromised (cf. Batch-2-10CV). The predictions for the test runs corroborate these findings and suggest a superiority of the CNN model. For all gradient elutions (Batch-2-5CV, Batch-1-10CV, and Batch-2-15CV), the CNN correctly predicts the onset of co-elution and the absolute dimer content, while the PLS predictions are highly scattered and predict early co-elution onsets. Finally, the step elution confirms that both models are able to differentiate between monomeric and dimeric species as a negative control under non-separating conditions. Both models accurately predict a simultaneous elution of monomers and dimers despite an absolute offset in the monomer concentration compared to UHP-SEC data.

Considering all emphasized points, the CNN model is better than the studied PLS-based models, as it combines high accuracy with robustness across cross-validation, test, and augmentation test for both target species. These findings confirm the work presented in (Schiemer et al. 2024), where the CNN models have been established for quantification of different proteins and monoclonal antibody size variants from ultraviolet-visible absorption spectra. In contrast to Schiemer et al. (2024), the augmented CNN model used within this study achieved lower performance than the non-augmented variants across all datasets and test scenarios. In direct comparison with the PLS model, the augmented CNN was still found to be superior. Although data augmentation was not found to be beneficial for this particular case, it is notable that the transfer of the LSA algorithm to Raman spectra was successful. Another recent example of data augmentation for Raman spectra showed a beneficial effect for chemometric modeling (Lange et al. 2024).

Compared to established spectroscopy-based PAT applications in biopharma, such as glucose monitoring in fermentation processes (Cavinato et al. 1990; Müller et al. 2024; Pontius et al. 2020) or UFDF monitoring (Rolinger et al. 2020b, 2023; Milewska et al. 2022), spectroscopic methods for aggregate

quantification remain limited (Wei et al. 2022; Brestrich et al. 2018). Brestrich et al. Brestrich et al. (2018) relied on only four experimental runs, with just one reserved for validation, whereas Wei et al. Wei et al. (2022) utilized aggregates derived from forced degradation experiments, while the cross-correlation structure in the provided data is unclear. Furthermore, reported results were predominantly acquired offline, allowing longer spectral acquisition times and eliminating the challenges associated with real-time process variations.

Although BSA serves as a practical and well-characterized surrogate for method development, we acknowledge that its structural simplicity and spectral characteristics do not fully capture the complexity of industrially relevant biotherapeutics, like monoclonal antibodies. Extending the approach to mAbs requires consideration of their distinct secondary-structure compositions, particularly their substantially higher β-sheet content relative to the predominantly α-helical structure of BSA. This is expected to dampen the spectral difference and hence make the distinction of monomeric from dimeric species more difficult. Additional factors include the higher concentration ranges typically encountered for therapeutic antibodies in downstream processing (20–60 g/L), the aggregate levels relevant to industrial polishing steps (reduction from roughly 3% to $\leq$ 0.5%), and detection-limit requirements of drug product formulations (0.1%–1% aggregates). Ongoing follow-up studies with therapeutic antibodies aim to evaluate these aspects and assess the transferability and robustness of the method across biopharmaceutical modalities.

In summary, the use of CNN models demonstrated suitability by generating selective regression models capable of accurately quantifying monomer and dimer contents in real-time chromatographic processes. However, the development of robust chemometric models requires extensive evaluation of their selectivity and predictive performance through larger experimental datasets and systematic algorithmic validation.

4 | Conclusions

This study demonstrates the feasibility of real-time aggregate quantification in biopharmaceutical downstream processing using Raman spectroscopy and structural spectral markers. By applying controlled aggregation stress and an LHS design, distinct Raman features—particularly in the amide I region—were identified as reliable indicators of protein aggregation across varying concentrations and conditions.

These spectral markers were validated in inline measurements during bind-elute AEX, where they successfully tracked monomer-aggregate separation dynamics. The combination of Raman spectroscopy with multivariate modeling approaches, including PLS and a CNN enhanced through synthetic data augmentation, enabled robust and accurate quantification of size variants under process-relevant conditions. While the CNN model outperformed traditional approaches in terms of predictive performance and selectivity, limitations arose from low signal intensities at low protein concentrations and environmental influences on background correction.

Despite these limitations, this work highlights the potential of Raman-based structural markers and machine learning models for real-time monitoring of protein aggregation in downstream unit operations.

Future work should also aim to expand the experimental space and enhance spectral preprocessing to improve model robustness and transferability across different process steps and environmental conditions.

Author Contributions

Jakob Heyer-Müller and Robin Schiemer both evolved the concepts and methods presented in this manuscript, supervised experimental work, analyzed and interpreted the data, drafted the figures, and wrote the final manuscript. Lars Robbel, Michael Schmitt, and Jürgen Hubbuch supervised the work. Jürgen Hubbuch acquired funding and initiated the work. All authors read and approved the final manuscript.

Acknowledgments

Open Access funding enabled and organized by Projekt DEAL.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request. The data are not publicly available due to privacy or ethical restrictions.

References

Alizadeh-Pasdar, N., S. Nakai, and E. C. Y. Li-Chan. 2002. "Principal Component Similarity Analysis of Raman Spectra to Study the Effects of pH, Heating, and κ -Carrageenan on Whey Protein Structure." *Journal of Agricultural and Food Chemistry* 50, no. 21 (October): 6042–6052.

Antosiewicz, J. M., and D. Shugar. 2016. "UV-Vis Spectroscopy of Tyrosine Side-Groups in Studies of Protein Structure. Part 2: Selected Applications." *Biophysical Reviews* 8 (May): 163–177.

Ashton, L., and E. W. Blanch. 2010. "pH-Induced Conformational Transitions in α -Lactalbumin Investigated With Two-Dimensional Raman Correlation Variance Plots and Moving Windows." *Journal of Molecular Structure* 974, no. 1 (June): 132–138.

Barnett, G. V., W. Qi, S. Amin, et al. 2015. "Structural Changes and Aggregation Mechanisms for Anti-Streptavidin IgG1 at Elevated Concentration." *Journal of Physical Chemistry. B* 119, no. 49 (December): 15150–15163.

Boulet-Audet, M., B. Byrne, and S. G. Kazarian. 2014. "High-Throughput Thermal Stability Analysis of a Monoclonal Antibody by Attenuated Total Reflection FT-IR Spectroscopic Imaging." *Analytical Chemistry* 86, no. 19 (October): 9786–9793.

Brestrich, N., M. Rüdt, D. Büchler, and J. Hubbuch. 2018. "Selective Protein Quantification for Preparative Chromatography Using Variable Pathlength UV/Vis Spectroscopy and Partial Least Squares Regression." *Chemical Engineering Science* 176, no. C (February): 157–164.

Brestrich, N., A. Sanden, A. Kraft, K. McCann, J. Bertolini, and J. Hubbuch. 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, no. 7 (July): 1406–1416.

Capito, F., R. Skudas, H. Kolmar, and C. Hunzinger. 2015. "At-Line Mid Infrared Spectroscopy for Monitoring Downstream Processing Unit Operations." *Process Biochemistry* 50, no. 6: 997–1005.

Cavinato, A. G., D. M. Mayes, Z. Ge, et al. 1990. "Noninvasive Method for Monitoring Ethanol in Fermentation Processes Using Fiber-Optic Near-Infrared Spectroscopy." *Analytical Chemistry* 62, no. 18 (September): 2058–2063.

Chen, J., J. Wang, R. Hess, G. Wang, J. Studts, and M. Franzreb. 2024. "Application of Raman Spectroscopy During Pharmaceutical Process Development for Determination of Critical Quality Attributes in Protein A Chromatography." *Journal of Chromatography A* 1718 (March): 464721.

Dietrich, A., R. Schiemer, J. Kurmann, S. Zhang, and J. Hubbuch. 2024. "Raman-Based PAT for VLP Precipitation: Systematic Data Diversification and Preprocessing Pipeline Identification." *Frontiers in Bioengineering and Biotechnology* 12.

Eisenberg, A. S., and L. J. Juszczak. 2012. "Correlation of TrpGly and GlyTrp Rotamer Structure With W7 and W10 UV Resonance Raman Modes and Fluorescence Emission Shifts." *Journal of Amino Acids* 2012 (July): 1–10.

Erb, D. 2025. pybaselines: A Python Library of Algorithms for the Baseline Correction of Experimental Data.

Ettah, I., and L. Ashton. 2018. "Engaging With Raman Spectroscopy to Investigate Antibody Aggregation." *Antibodies* 7, no. 3 (July): 24.

FDA. September, 2004. Guidance for Industry PAT—A Framework for Innovative Pharmaceutical Development, Manufacuring, and Quality Assurance. Technical report, United States Food and Drug Administration.

Feidl, F., S. Garbellini, M. F. Luna, et al. 2019. "Combining Mechanistic Modeling and Raman Spectroscopy for Monitoring Antibody Chromatographic Purification." *Processes* 7, no. 10 (October): 683.

Fontana, M. D., K. Ben Mabrouk, and T. H. Kauffmann. 2013. "Raman Spectroscopic Sensors for Inorganic Salts." In Spectroscopic Properties of Inorganic and Organometallic Compounds, edited by J. Yarwood, R. Douthwaite, and S. Duckett, volume 44, 40–67. RSC Publishing.

Furić, K., I. Ciglenečki, and B. Ćosović. 2000. "Raman Spectroscopic Study of Sodium Chloride Water Solutions." *Journal of Molecular Structure* 550–551 (September): 225–234.

Glassey, J., K. V. Gernaey, C. Clemens, et al. 2011. "Process Analytical Technology (PAT) for Biopharmaceuticals." *Biotechnology Journal* 6, no. 4 (April): 369–377.

Gómez de la Cuesta, R., R. Goodacre, and L. Ashton. 2014. "Monitoring Antibody Aggregation in Early Drug Development Using Raman Spectroscopy and Perturbation-Correlation Moving Windows." *Analytical Chemistry* 86, no. 22 (November): 11133–11140.

Harada, I., T. Miura, and H. Takeuchi. 1986. "Origin of the Doublet at 1360 and 1340 cm1 in the Raman Spectra of Tryptophan and Related Compounds." *Spectrochimica Acta Part A: Molecular Spectroscopy* 42, no. 2–3 (January): 307–312.

Heyer-Müller, J., R. Schiemer, M. Lopinski, et al. 2025. "A novel Raman-Chromatography Assembly for Automated Calibration and In-Line Monitoring in Bioprocessing." Engineering in Life Sciences. Manuscript accepted and in production.

Lange, C., I. Thiele, L. Santolin, et al. 2024. "Data augmentation Scheme for Raman Spectra With Highly Correlated Annotations." In 34th European Symposium on Computer Aided Process Engineering/15th International Symposium on Process Systems Engineering, volume 53 of *Computer Aided Chemical Engineering*, edited by F. Manenti, G. V. Reklaitis, 3055–3060. Elsevier.

Lewis, E. N., W. Qi, L. H. Kidder, S. Amin, S. M. Kenyon, and S. Blake. 2014. "Combined Dynamic Light Scattering and Raman Spectroscopy Approach for Characterizing the Aggregation of Therapeutic Proteins." *Molecules* 19, no. 12 (December): 20888–20905.

Li, C. H., and T. Li. 2009. "Application of Vibrational Spectroscopy to the Structural Characterization of Monoclonal Antibody and Its Aggregate." *Current Pharmaceutical Biotechnology* 10, no. 4: 391–399.

Lippert, J. L., D. Tyminski, and P. J. Desmeules. 1976. "Determination of the Secondary Structure of Proteins by Laser Raman Spectroscopy." *Journal of the American Chemical Society* 98, no. 22 (October): 7075–7080.

Makki, A. A., V. Massot, H. J. Byrne, et al. 2021. "Understanding the Discrimination and Quantification of Monoclonal Antibodies Preparations Using Raman Spectroscopy." *Journal of Pharmaceutical and Biomedical Analysis* 194 (February): 113734.

Milewska, A., G. Baekelandt, S. Boutaieb, V. Mozin, and A. Falconbridge. 2022. "In-Line Monitoring of Protein Concentration With Mid-Infrared Spectroscopy During Ultrafiltration/Diafiltration." *Engineering in Life Sciences* 23, no. 2 (December): e2200050.

Militello, V., C. Casarino, A. Emanuele, A. Giostra, F. Pullara, and M. Leone. 2004. "Aggregation Kinetics of Bovine Serum Albumin Studied by Ftir Spectroscopy and Light Scattering." *Biophysical Chemistry* 107, no. 2 (February): 175–187.

Müller, D. H., M. Börger, J. Thien, and H.-J. Koß. 2024. "Bioprocess In-Line Monitoring and Control Using Raman Spectroscopy and Indirect Hard Modeling (IHM)." *Biotechnology and Bioengineering* 121: 2225–2233. https://doi.org/10.1002/bit.28724.

Moussa, E. M., J. P. Panchal, B. S. Moorthy, et al. 2016. "Immunogenicity of Therapeutic Protein Aggregates." *Journal of Pharmaceutical Sciences* 105, no. 2 (February): 417–430.

Ota, C., S. Noguchi, S. Nagatoishi, and K. Tsumoto. 2016. "Assessment of the Protein–Protein Interactions in a Highly Concentrated Antibody Solution by Using Raman Spectroscopy." *Pharmaceutical Research* 33, no. 4 (April): 956–969.

Pham, N. B., and W. S. Meng. 2020. "Protein Aggregation and Immunogenicity of Biotherapeutics." *International Journal of Pharmaceutics* 585 (July): 119523.

Pontius, K., H. Junicke, K. V. Gernaey, and M. Bevilacqua. 2020. "Monitoring Yeast Fermentations by Nonlinear Infrared Technology and Chemometrics – Understanding Process Correlations and Indirect Predictions." *Applied Microbiology and Biotechnology* 104, no. 12 (June): 5315–5335.

Rüdt, M., N. Brestrich, L. Rolinger, and J. Hubbuch. 2017. "Real-Time Monitoring and Control of the Load Phase of a Protein a Capture Step." *Biotechnology and Bioengineering* 114, no. 2 (February): 368–373.

Rüdt, M., P. Vormittag, N. Hillebrandt, and J. Hubbuch. 2019. "Process Monitoring of Virus-Like Particle Reassembly by Diafiltration With Uvvis Spectroscopy and Light Scattering." *Biotechnology and Bioengineering* 116: 1366–1379.

Roberts, C. J. 2014. "Protein Aggregation and Its Impact on Product Quality." *Current Opinion in Biotechnology* 30 (December): 211–217.

Rolinger, L., J. Hubbuch, and M. Rüdt. 2023. "Monitoring of Ultra- and Diafiltration Processes by Kalman-Filtered Raman Measurements." *Analytical and Bioanalytical Chemistry* 415, no. 5: 841–854.

Rolinger, L., M. Rüdt, J. Diehm, et al. 2020a. "Multi-Attribute Pat for Uf/Df of Proteins–Monitoring Concentration, Particle Sizes, and Buffer Exchange." *Analytical and Bioanalytical Chemistry* 412, no. 9: 2123–2136.

Rolinger, L., M. Rüdt, and J. Hubbuch. 2020b. "A Critical Review of Recent Trends, and a Future Perspective of Optical Spectroscopy as Pat in Biopharmaceutical Downstream Processing." *Analytical and Bioanalytical Chemistry* 412, no. 9 (April): 2047–2064.

Rosenberg, A. S. 2006. "Effects of Protein Aggregates: An Immunologic Perspective." *AAPS Journal* 8, no. 3 (September): 59.

Rygula, A., K. Majzner, K. M. Marzec, A. Kaczor, M. Pilarczyk, and M. Baranska. 2013. "Raman Spectroscopy of Proteins: A Review." *Journal of Raman Spectroscopy* 44, no. 8: 1061–1076.

Sato, Y., S. Nagatoishi, S. Noguchi, and K. Tsumoto. 2023. "Raman Spectroscopic Analysis of Highly-Concentrated Antibodies under the Acid-Treated Conditions." *Pharmaceutical Research* 40, no. 7: 1853–1864.

Schiemer, R., M. Rüdt, and J. Hubbuch. 2024. "Generative Data Augmentation and Automated Optimization of Convolutional Neural Networks for Process Monitoring." *Frontiers in Bioengineering and Biotechnology* 12. https://www.frontiersin.org/journals/bioengineering-and-biotechnology/articles/10.3389/fbioe.2024.1228846.

Schiemer, R., J. T. Weggen, K. M. Schmitt, and J. Hubbuch. 2023. "An Adaptive Soft-Sensor for Advanced Real-Time Monitoring of an Antibody-Drug Conjugation Reaction." *Biotechnology and Bioengineering* 120, no. 7: 1914–1928.

Shivu, B., S. Seshadri, J. Li, K. A. Oberg, V. N. Uversky, and A. L. Fink. 2013. "Distinct -Sheet Structure in Protein Aggregates Determined by Atr–Ftir Spectroscopy." *Biochemistry* 52, no. 31 (August): 5176–5183.

Soch, J., and C. Allefeld. 2018. "Macs – a New Spm Toolbox for Model Assessment, Comparison and Selection." *Journal of Neuroscience Methods* 306 (August): 19–31.

Vázquez-Rey, M., and D. A. Lang. 2011. "Aggregates in Monoclonal Antibody Manufacturing Processes." *Biotechnology and Bioengineering* 108, no. 7 (July): 1494–1508.

Wang, J., J. Chen, J. Studts, and G. Wang. 2023. "In-Line Product Quality Monitoring During Biopharmaceutical Manufacturing Using Computational Raman Spectroscopy." *mAbs* 15, no. 1 (December): 2220149.

Wei, B., N. Woon, L. Dai, et al. 2022. "Multi-Attribute Raman Spectroscopy (Mars) for Monitoring Product Quality Attributes in Formulated Monoclonal Antibody Therapeutics." *mAbs* 14, no. 1. https://doi.org/10.1080/19420862.2021.2007564.

Williams, R. W. 1986. "[14] Protein Secondary Structure Analysis Using Raman Amide I and Amide III Spectra." In *Methods in Enzymology*, vol. 130, 311–331. Elsevier.

Zhang, C., J. S. Springall, X. Wang, and I. Barman. 2019. "Rapid, Quantitative Determination of Aggregation and Particle Formation for Antibody Drug Conjugate Therapeutics With Label-Free Raman Spectroscopy." *Analytica Chimica Acta* 1081: 138–145.

Zhou, C., W. Qi, E. Neil Lewis, and J. F. Carpenter. 2015. "Concomitant Raman Spectroscopy and Dynamic Light Scattering for Characterization of Therapeutic Proteins at High Concentrations." *Analytical Biochemistry* 472 (March): 7–20.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.
Raman_Protein_Aggregation_supplementary_revised.