Journal of Chromatography A 1756 (2025) 466069

Contents lists available at ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

Digital Butterworth filter as preprocessing method for implementing Raman spectroscopy as an analytical method in downstream processing of biopharmaceuticals

Jingyi Chen [a,b], José Munoz Reyes [a,b], Robin Schiemer [a], Gang Wang [a], Joey Studts [a], Matthias Franzreb [b,*]

[a] Boehringer Ingelheim Pharma GmbH / Co. KG, Biberach an der Riss, Germany
[b] Institute of Functional Interfaces, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, 76344, Germany

ARTICLE INFO

Keywords:
Therapeutic antibody
Butterworth filter
Raman spectroscopy
Process analytical technology
Data preprocessing
Machine learning

ABSTRACT

For implementing Raman spectroscopy as an analytical method in downstream processing, extracting molecular information related to biopharmaceuticals is still challenging due to spectral variations caused by spectrometer, setup and fluorescence. This study explores the potential of the Butterworth filter as a preprocessing method for baseline correction and noise reduction in Raman spectra. We first investigate the Butterworth highpass filter's working principle and its optimization by introducing disturbances to spectral baselines and assessing the cutoff frequency ω_c's effect on minimizing baseline variations and enhancing the linear correlation (r^2) between Raman signals and protein concentrations. The optimal ω_c range (0.004 to 0.008 cm) yields an $r^2 \geq 0.85$, outperforming the Savitzky-Golay derivative filter's 0.68. Further, we explore a Butterworth bandpass filter, adjusting low and high cutoff frequencies, showing an 11.6–15 % improvement in r^2 over the highpass design. Our results suggest the necessity of specific cutoff frequency selection when applying the bandpass design to the Raman spectra of individual protein molecules and the method for this selection is discussed. By applying the optimization outputs, we developed chemometric models linking Critical Quality Attributes to the Raman data preprocessed by the Butterworth bandpass filter, covering concentrations up to 25.6 mg/mL for a biopharmaceutical immunoglobulin G (IgG) antibody and 4.2 mg/mL for Transferrin. When validated in Cation Exchange Chromatography runs with gradient lengths of 5 and 10 column volume for in-line predictions, the models show high predictability, achieving a coefficient of determination R^2 of 0.99 for IgG and 0.95 for Transferrin.

1. Introduction

The Process Analytical Technology (PAT) framework is increasingly being advocated in biopharmaceutical manufacturing landscape [1], particularly in the downstream processing (DSP) of monoclonal antibodies (mAbs) [2]. This shift towards PAT is driven by its potential to streamline process development, enhance detection of critical quality attributes (CQAs), and reduce time and costs [3]. Within the PAT framework, Raman spectroscopy has emerged as a promising analytical tool for monitoring quality attributes across a range of processes [4–7]. Unlike the ultraviolet/visible (UV/Vis) spectroscopy, which allows straightforward protein concentration determination at a specific wavelength, Raman spectroscopy has the potential to correlate to a broader range of CQAs but faces a challenge due to the complex nature of bands. This complexity necessitates the use of chemometrics models [8] for calibrating analytes in Raman spectroscopy. A variety of chemometric model procedures [8,9] have been reported, highlighting the importance of data preprocessing. The preprocessing methods handle several tasks such as spectral variation removal, noise reduction, outlier detection and normalization. Given the high sensitivity of Raman acquisition to their measurement conditions, identical samples can yield spectra with baseline deviations due to minor variations in the configuration of spectrometer and setup [10].

Raman signals are comprised of information about the measured sample as well as several side effects, such as spectral baseline, artifacts, and noises. These can originate from both intrinsic system and extrinsic sources related to the detector or environmental conditions. Charge-Coupled Devices (CCD) detectors, widely used in Raman spectrometer

* Corresponding author.
E-mail address: matthias.franzreb@kit.edu (M. Franzreb).

https://doi.org/10.1016/j.chroma.2025.466069
Received 25 February 2025; Received in revised form 7 May 2025; Accepted 18 May 2025
Available online 19 May 2025

for their efficiency in detecting weak Raman signals, are prone to capturing cosmic high-energy particles, leading to spectral artifacts in shape of narrow-bandwidth spikes [11]. Spectral smoothing is employed to effectively eliminate high-frequency noises [12] and amplify the signal-to-noise ratio (SNR), thereby facilitating the extraction of molecular information related to a specific biopharmaceutical. A significant challenge in Raman signal recovery is the predominant spectral baseline, which is typically several orders of magnitude [13] stronger than the fingerprint peaks of biopharmaceutical samples. Various factors contribute to the Raman spectral baseline, including fluorescence background [14], thermal fluctuations in CCDs [11], variations in flow rates through the flow-cell [15], and increasing run time [16].

For practical application in DSP scenarios, Raman spectroscopy requires extensive data-driven chemometric models that are trained and validated across a range of DSP scales and operational setups. Considering the sensitivity of Raman acquisition, which can result in fluctuations in the spectral baseline between datasets, it is crucial to maintain consistency in Raman datasets for identical samples, regardless of the scale or system. A comprehensive collaborative study on Raman data comparability [17] has highlighted that variations in setups can lead to discrepancies in Raman data, further emphasizing the need for consistent data handling and preprocessing methods.

Digital filters have shown promise in their ability to filter out background interferences and to improve the characteristics of Raman spectroscopy. For instance, Savitzky-Golay (SG) filtering is widely employed in chemometric models, usually coupled with Principal Component Regression (PCR) or Partial Least Square Regression (PLSR) [18,19]. Wei and coworkers [20] showcased the capability of the SG derivative to eliminate spectral baseline variations caused by Raman spectrometers. Wang et al. [15] explored the potential of Butterworth highpass filter and demonstrated its efficacy in removing the effect of flow rate on the spectral baseline. There is a demand for a simplified preprocessing method to ensure Raman data consistency and also improve signal recovery of biopharmaceuticals from the side effects within a single method. This would further enhance the practical application of Raman spectroscopy in biopharmaceutical manufacturing.

In our selection of filtering techniques, the Butterworth filter was chosen for its distinct advantages in meeting the requirements of our study. Our analysis focuses on the Raman signals, which are complex mixtures of fluorescence interference and scattering effects from various components within the solution. The Butterworth highpass filter has proven effective in filtering low-frequency broad signals, such as fluorescence interference [4], which significantly improve signal quality of components of interests. Additionally, the bandpass design [21] of Butterworth filter can enhance model's predictability based on Near Infrared (NIR) spectroscopic data, by suppressing certain high-frequency components. Furthermore, the Butterworth filter is designed to achieve the maximal flatness in the passband for the given filter order. This is particularly advantageous compared to traditional Fourier filtering methods, which requires extensive control to prevent artifacts and ripples in passband response. Despite the adaptive denoising capability of wavelet transform [22], it necessitates computation complexity and cost on the determination of wavelet function type and multi-levels of decomposition. Empirical Mode Decomposition (EMD) [23] is a data-driven filtering technique that decomposes a signal into various Intrinsic Mode Functions (IMFs). However, the number of IMFs can increase with the spectrum complexity and complicate the correlation of the decomposed data with various CQAs in downstream processes. This article presents the first representative interpretation of the Butterworth filter's capability to decompose Raman spectra of biopharmaceuticals into multiple Butterworth frequency regions that effectively filter out irrelevant frequencies, or spectral components. We further explored the impact of the cutoff frequency on the Butterworth highpass filter's efficacy in eliminating synthetic spectral baselines. Through these investigations, we identified the most effective Butterworth frequency region for baseline removal. We also employed the SG derivative filter, a common preprocessing method in Raman spectroscopy, and compare its performance with the Butterworth filter. In our interest to maximize the recovery of protein-related spectral features using a single preprocessing method, we evaluated the Butterworth bandpass filter's effectiveness in eliminating both baseline fluctuations and high-frequency noises. Our comprehensive screening of low and high cutoff frequencies revealed that a bandpass design enhances the linear correlation between Raman signals and protein concentrations by 11.6–15 %, surpassing the highpass design. Additionally, our research emphasized the necessity of specific cutoff frequency selection when applying the bandpass design to the Raman spectra of individual protein molecules. This finding enables the precise tuning of bandpass parameters, crucial for CQA specific decomposition of Raman signals, thereby enhancing the implementation of Raman spectroscopy in the downstream processing of biopharmaceuticals. A biopharmaceutical immunoglobulin G (IgG) antibody and Transferrin molecule representing a model impurity were studied to develop calibration models for determination of the two molecules in Cation Exchange Chromatography (CEX). When the specific Butterworth frequencies identified from the screening were applied, the IgG model achieved a coefficient of determination R^2 of 0.99, while the Transferrin model achieved a R^2 of 0.95.

2. Materials and methods

2.1. Raman spectrometer setup and two Raman detection systems

In the study, we employed a HyperFlux Pro Plus Raman spectrometer (Tornado Spectral Systems, Mississauga, Ontario, Canada), controlled by SpectralSoft 3.4. software. This spectrometer was excited by a 785 nm emission laser, covering a wavenumber range from 200 to 3300 cm^{-1} with a resolution of 1 cm^{-1}. A Hudson Probe with a 45 μL Micro Flow Cell (MFC) (Tornado Spectral Systems, Mississauga, Ontario, Canada) was connected to the Raman spectrometer for fluidic measurement. We followed a consistent acquisition setup for all Raman measurements in this study [15]. To maximize the Raman signal intensity and signal-to-noise ratio, the laser was set to its maximum power of 495 mW, which did not cause detector saturation or sample damage. In our previous studies, we found that using the maximum laser power led to detector saturation and distorted Raman signals during capture chromatography process step [4]. Increased laser power can potentially result in heating of the sample due to high energy, especially after long exposure it can burn biological samples. The exposure time to 500 ms with 15 averages, resulting in a scan time 7.5 s/per scan. For off-line measurement of well-mixed samples, the Raman spectrometer was mounted on a Tecan Fluent 780 Liquid Handler (Tecan, Männedorf, Switzerland), as described in references [4,15,24]. Each 300 μL sample was automatically injected into the MFC, with each off-line Raman detection lasting 90 s and yielding 11 spectra per sample. We utilized a second Raman detection system for in-line measurement of elution samples from chromatography runs using the same spectrometer. This was performed on an ÄKTA Avant 25 system (Cytiva, Uppsala, Sweden) controlled by UNICORN™ 7.5 software. The MFC was positioned between the conductivity and pH sensors for Raman detection.

2.2. Molecules and experimental designs

2.2.1. One-component dilution series of two proteins

We utilized a pharmaceutical IgG antibody, referred to in subsequent text as mAb1, provided by Boehringer Ingelheim Pharma GmbH & Co. KG (Biberach, DE), Additionally, human Transferrin (Sigma Aldrich, Burlington, Massachusetts, US) was employed as a model impurity. A mAb1 drug substance solution with 50 mg/mL was ultra-filtrated into purified water (Unagi, Unchained Labs, Pleasanton, California, USA), and diluted to a stock solution with a concentration of 19.5 mg/mL.

Lyophilized transferrin was dissolved in purified water and adjusted to 20 mg/mL stock solution. Two dilution series were separately prepared with 11 levels, by mixing the two stock solutions with purified water. Each mixed sample was measured off-line, with 40 Raman spectra collected for each sample.

2.2.2. Three-component calibration experiments mixing two proteins with salt buffer

The further study of preprocessing method focused on a three-component system containing Transferrin, mAb1, and salt concentration. For the calibration experiment, we set the mAb1 concentration calibration range from zero to the maximum 25.6 mg/mL, and Transferrin up to a maximum 4.2 mg/mL, both tested at 11 equally spaced levels. The third factor, salt concentration, was regulated by adjusting the ratio of two solutions: Buffer A (50 mM acetate, pH 5) and Buffer B (50 mM acetate, pH 5, 1 M NaCl). We designed an experiment ($N = 132$) involving these three factors, ensuring minimal correlation among them while minimizing the protein materials. This customized design is graphically visualized in Figure S1a). The correlation matrix for the three factors is presented in Figure S1b, with all correlation coefficients below 0.13. The experimental window was divided into two triangular sections (I and II) by a diagonal line, resulting in two sub-experiments with each mixing three feed solutions (N_1=N_2=66). In the first sub-experiment (section I), Transferrin stock solution was added to mAb1 drug substance solution, then ultra-filtrated into Buffer A, reaching the concentration of 25.2 mg/mL for mAb1 and of 4.19 mg/mL for Transferrin (feed solution F1-A). Another mAb1 drug substance solution was directly buffer-exchanged and diluted in Buffer B to 24.02 mg/mL (feed solution F2-B). A 1:1 mixture of Buffer A and B was used as feed solution F3. Similarly, the sub-experiment (section II) mixed another feed solutions (F1-B, F2-A and F3). In the second sub-experiment, feed solutions F1-B, F2-A and F3 were mixed. F1-B was a mixture of 4.2 mg/mL Transferrin and 25.6 mg/mL buffer-exchanged in Buffer B. F2-A was a Transferrin solution diafiltrated in Buffer A at concentration of 4.15 mg/mL.

Both sub-experiments utilized the Tecan Fluent 780 system for automatic mixing of feed solutions. The samples were then injected into the MFC for off-line Raman detection (Section 2.1). All samples from each sub-experiment were pooled post-measurement, stored at 4 °C until being reused as loading materials for the subsequent chromatography runs. Due to a wash step in between two measurements, the two sample pools were diluted with purified water. All the Raman measurements collected from the 130 samples are provided as the training dataset for calibration models of quantifying mAb1 and Transferrin concentrations.

2.2.3. CEX runs with fractionations

Three CEX runs were performed on an ÄKTA Avant 25 system for validating the calibration models (Section 2.3). The CEX column used was packed with Poros XS resin (Thermo Fisher Scientific, Waltham, USA) and had a diameter of 1 cm and a column volume (CV) of 17.14 mL. The CEX runs initiated with an equilibration phase using Buffer A, followed by load and wash phases. During the elution phase, a salt gradient setup was applied, where Buffer A and B were pumped with an isocratic volumetric percentage from 0 % to 100 % Buffer B, over a specific gradient length in CV. As detailed in Table 1, a gradient length of 10 CV was used in the Validation 1 run, while 5 CV in the two Validation 2 and Test runs. For Validation 1 and Validation 2, two pooled solutions from the sub-experiments were titrated to pH 5.0 using 99 % acetic acid (Aug. Hedinger GmbH & Co. KG), just before loading. The loading masses of Transferrin and mAb1 were listed in Table 1. All runs were fractionated using a built-in fraction collector into fractions of 1 mL. The in-line Raman spectra recorded were averaged for each of the collected fractions. The concentrations of Transferrin and mAb1 in each fraction were determined by performing Ultra-performance Size Exclusion Chromatography (UP-SEC). This was done using an Acquity UPLC BEH200 SEC Column on an Acquity Premier system controlled by Empower 4 (all from Water Corporation, Milford, MA, USA).

2.3. Preprocessing and modelling algorithms

2.3.1. Role of cutoff frequencies in Butterworth filtering of Raman spectra

The Butterworth filter is a signal processing filter designed to ensure that the amplitude of the frequency response within the passband is as flat as possible, thereby transmitting the desired signals with minimal signal distortion. Depending on specific requirements, the filter can be designed as lowpass, highpass, bandpass, or band-stop. In the case of a low- or highpass design, there are two hyperparameters: cutoff frequency f_c, which refers to a Butterworth frequency marking the half-power point between passband and stopband, and filter order n, which describes the steepness of transmission from passband to stopband. A Raman spectrum is a composition of spectral baseline, noise, narrow and broad peaks. Each of these components can be approximated using an arbitrarily number of periodic functions. In the context of Raman spectroscopy, the application of the Butterworth filter requires an understanding that these waveform signals can be viewed as periodic functions. Regarding a waveform in Raman signal, the period of the function corresponds to a specific wavenumber region, denoted in units of cm^{-1}. This allows us to perceive Raman signals as existing within a time-domain represented by the unit cm^{-1}. In signal processing, the term Fourier-transformation is utilized to convert signals in time-domain to a discrete frequency domain. Specifically, when preprocessing Raman spectra, the Butterworth filter Fourier-transforms (FT) Raman signals from the time-domain to a frequency-domain, represented in a reverse unit of $\frac{1}{cm^{-1}}$. Considering the entire wavenumber region totaling 3101 variables as a single sine function, the period of the waveform is 1550.5 cm^{-1} and the corresponding Butterworth frequency is the inverse equaling 0.00064 $\frac{1}{cm^{-1}}$ To simplify the term "frequency" and avoid confusion with the unit [cm], we use a normalized frequency f/f_s throughout the text. Here, f_s represents the sampling frequency in unit of $\frac{1}{cm^{-1}}$, as we sample a Raman spectrum at an interval of 1 cm^{-1}. To understand the behavior of the cutoff frequencies on Raman spectra, the 40 raw spectra of 20 mg/mL Transferrin stock solution (Section 2.2.1) were Butterworth-filtered, applying low- and highpass with $n = 5$ (ten-pole) [15] and varied cutoff frequencies $[0.002,\ 0.006,\ 0.02,\ 0.2]f_s$. The whole study was programmed using Python 3.9, and all the

Table 1
Process parameters of the calibration, validation, and test experiments.

Runs	Setup	Transferrin [mg]	mAb1 [mg]	Gradient length [CV]	Usage	Sample numbers
Calibration	Tecan (off-line)	–	–	–	Training	132
Validation 1	ÄKTA (in-line)	41.9	492.2	10	Validation, Hyperparameter screening	60
Validation 2	ÄKTA (in-line)	83.5	256	5		51
Test	ÄKTA (in-line)	40	225	5	Final model test	54

preprocessing methods were applied using packages *NumPy* and *SciPy*.

2.3.2. Assessing the Butterworth filter's robustness against disturbances in spectral baseline

Numerical experiments were conducted on the two protein dilution series (Section 2.2.1, mAb1 and Transferrin) to assess the impact of cutoff frequencies on baseline removal, and to evaluate the Butterworth filter's robustness against synthetic disturbances added in spectral baseline. Synthetic disturbances were generated by replacing random data points located in baseline with varied values. In Raman data, real variations in baseline can manifest as fluctuations and intensity shift in the low wavenumber range of 400 to 1800 cm^{-1}. To simulate these real variations, 20 data points (1.4 %) were randomly selected within the [400, 1800] cm^{-1} range from the baseline. For each Raman spectrum, its baseline was estimated using a fourth-degree Improved Modified Polynomial fit [25]. The magnitude of the added variations was sampled from a uniform distribution of [0.8,1.2] of their original value [26]. The resulting 20 new points were subsequently fitted with a new four-degree polynomial, and a disturbance was defined as the difference between the original and disturbed baselines. This disturbance was then added to the original spectrum, obtaining a disturbed spectrum. For baseline removal, the highpass Butterworth filter was applied to both original and disturbed spectra, with a filter order of 5 (ten-pole) and varying cutoff frequencies in the range from 0.001 to 0.020 f_s with a step size of 0.001 f_s To account for randomness, at each concentration, the numerical experiment was independently repeated 40 times for each spectrum. A second-derivative SG filter was applied as a reference using a window length of 11 and a 2nd order polynomial [20].

The robustness of the highpass Butterworth filter against baseline variations was assessed by comparing the disturbed and undisturbed spectra after treatment with the filter, using two different metrics: the cosine similarity θ and the averaged squared Pearson correlation coefficient r^2. The cosine similarity [27,28] is defined as the dot product between spectra divided by the product of their Euclidean norms as follows:

$$\theta = \frac{\left(\boldsymbol{x}_{c_i}(\nu),\ \widetilde{\boldsymbol{x}}_{c_i}(\nu)\right)}{\| \boldsymbol{x}_{c_i}(\nu)\|_2 \|\widetilde{\boldsymbol{x}}_{c_i}(\nu)\|_2}, \tag{1}$$

where $\boldsymbol{x}_{c_i}$ and $\widetilde{\boldsymbol{x}}_{c_i}$ are the preprocessed spectra of raw spectrum and disturbed spectrum at a single concentration c_i, respectively. ν is one single wavenumber in the Raman spectrum and m the total number of wavenumber variables. The use of cosine similarity aims to quantify the overlap between the two processed spectra $\boldsymbol{x}_{c_i}$ and $\widetilde{\boldsymbol{x}}_{c_i}$ with and without disturbances. To verify the degree of how the protein concentration signals were attenuated, we took the averaged squared Pearson correlation coefficient correlation r^2 across all wavenumbers as a measure of averaged concentration-dependence, regarding raw or preprocessed spectra. The r^2 was defined as follows:

$$r^2 = \frac{1}{m}\sum_{j=1}^{m} {r_j}^2 = \frac{1}{m}\sum_{j=1}^{m}\left[\frac{\sum_{c_i}\left(x_{c_i,j}-\overline{x_j}\right)(c_i-\overline{c})}{\sqrt{\sum_{c_i}\left(x_{c_i,j}-\overline{x_J}\right)^2\sum_{c_i}(c_i-\overline{c})^2}}\right]^2, \tag{2}$$

where x_{c_i} indicates one investigated spectrum (raw or preprocessed) at a concentration c_i.We measured the linear correlation r between the concentration vector and their corresponding spectra. This was conducted by comparing the concentrations and signal intensities at a same wavenumber j from total variable numbers m. Then, squared correlation ${r_j}^2$ was summed and averaged across all wavenumbers.

2.3.3. Screening low and high cutoff frequencies of a Butterworth bandpass filter

A further investigation on Butterworth bandpass design was carried out using the two protein dilution series. The purpose of this design was to remove not just spectral baseline, but also high-frequency noise, thereby maximizing the extraction of protein-related spectral features. A bandpass Butterworth filter can remove Butterworth frequencies that are either lower than a specific low cutoff frequency $f_{c,low}/f_s$ or higher than a certain high cutoff frequency$f_{c,high}/f_s$. For both proteins, the spectra consisting of 11 concentration levels underwent treatment with various bandpass filters with $n = 10$ (ten-pole). But the low frequency varied between 0.001 and 0.014 f_s in a step of 0.0005 f_s, while the high frequency ranged from 0.015 to 0.5 f_s in a step of 0.01 f_s The effectiveness of recovering protein-related features was assessed using the same averaged r^2 mentioned in Section 2.3.2.

2.3.4. Performance evaluation of the Butterworth bandpass filter in real downstream process

To assess the performance of the Butterworth bandpass filter in a downstream process, a calibration experiment along with two CEX runs were carried out to collect training data (Section 2.2.2) and two validation datasets (Section 2.2.3). The objective was to evaluate the ability of preprocessing method to manage baseline variations caused by instruments and systems. Ham et al. [21] recommend evaluating the effectiveness of preprocessing method by executing the complete workflow instead of a single preprocessing step. Alterations in the workflow can lead to different model outputs. Therefore, we first conducted a comprehensive workflow screening that included spectral variable truncation, preprocessing method, and regression models. As listed in Table 2, various options of each step were tested along with their hyperparameters. This included 780 Butterworth bandpass filters and 800 SG filters, and 20 negative controls (no preprocessing), resulting in 1610 workflow candidates for predicting concentrations of mAb1 and Transferrin. A 20-fold cross validation procedure was employed to internally evaluate the performance. The score was computed using Root-mean-square deviation (RMSE) [15] and Coefficient of determination (R^2) [24]. Each model was subsequently validated using the two external validation datasets. Regarding protein concentration prediction, the model performance was evaluated by the R^2 for training dataset and the coefficient of prediction Q^2 for two validation datasets. Therefore, to rank the candidates in a simplified manner, we choose the lowest value among the three coefficients as a single new figure of merit f, which describes the most tolerable model performance within the three datasets. All the regression models were built using package *scikit-learn*.

Normalization is necessary to handle multiplicative effects that arise from variations in laser power, spectrometer drift, inherent intensity variability of the sample, or alterations in the medium's refractive index [29]. During the final model tuning, we applied an additional normalization procedure using a weighted multiplicative scatter correction (MSC) algorithm [30–32]. In this normalization algorithm, we assigned

Table 2
Category and the approach choices in the workflow screening.

Category	Choice	Hyperparameter
Wavenumber Truncation	Full range: 200 to 3300 cm^{-1} Range 2: 800 to 1800 cm^{-1}	
Preprocessing	Butterworth Bandpass filter (N_{total}= 780)	$f_{c,low}/f_s$ = [0.002, 0.009] in 0.001 $f_{c,high}/f_s$ = [0.02, 0.11] in 0.01 5-order (10-pole)
	Savitzky-Golay derivative (N_{total}= 800)	W_l = [11, 51] in a step of 2 n_{poly} = 2 or 3 $n_{derivative}$ = 1 or 2
	Control: no preprocessing (N_{total}= 20)	
Model regression	Partial Least Square Regression (N_{total}= 960)	$n_{component}$ = 3, 4, 5
	Support Vector Regression (polynomial) (N_{total}= 640)	C = 1000 or 100 ε = 0.01, degree = 3

a weight of 1 to the region between 2100 and 2400 cm^{-1}. These weights were determined using a variable sorting for normalization (VSN) algorithm (refer to Supplementary methods). The Support Vector Regression (SVR) with a polynomial kernel was chosen as the model regressor. The polynomial kernel function has a degree of three, with an epsilon (ϵ) value of 0.01, and a regularization parameter (C) of 1000.

3. Results and discussion

3.1. The capability of the Butterworth filter on Raman spectra decomposition

Different low- and highpass Butterworth filters were used to preprocess the raw spectrum of a 20 mg/mL human Transferrin solution in purified water, using four different cutoff frequencies$[0.002,\ 0.006,\ 0.02,\ 0.2]\ f_s$ and ten-pole. This aims to investigate the working principle of Butterworth filter with low- and high-pass designs and the effect of the cutoff frequency. Fig. 1 represents the behavior of a Butterworth filter with low- and highpass designs, displaying the results using different cutoff frequencies in rows from A to D. The second column displays lowpass components with varied cutoff frequencies, using a raw spectrum as a reference. The third column presents highpass components. The last column in Fig. 1 represents the power plot of FT data, displaying the signals in the Butterworth frequency domain. The signal power is the squared amplitude of the sine function at each Butterworth frequency. For the limit case of a filter with an ideal cutoff behavior ($n\to\infty$), the decomposition of the spectra can be simplified as: raw spectrum = lowpass component signals + highpass component signals.

Taking the example of a filter with frequency of 0.002 f_s (row A), by passing through the Butterworth frequencies below the frequency value, a lowpass component is shown and it can be considered a baseline in a broad waveform with low Butterworth frequency region (A2). A high-pass design improves the resolution of significant Raman peaks by filtering out the lowpass component i.e. baseline and reducing the order of baseline's magnitudes (A3). Similarly, in the power plot (A4), we observe a significant amplitude drop in the low Butterworth frequency range below 0.002 f_s, relative to the raw spectrum. The orange profile behaviors sharper and more intensive than the blue, suggesting that the remaining Raman data is amplified post-filtering. The data within the mid-range between 0.002 and 0.1 f_s could potentially contain abundant protein information, and this amplification could contribute to model enhancement. Although the region above 0.1 f_s, which primarily contains spectral noise data, is also amplified, it is at least four orders of magnitudes weaker than the true peaks and can therefore be ignored.

By increasing the cutoff frequency from 0.002 f_s to 0.2 f_s (row from A to D), the lowpass component profile tends to overlap the original spectrum, while the highpass component profile has a transition from broad true peaks to artificial spikes or noises. Those artifacts or noises, which are extremely narrow and sharp peaks, are Fourier-transformed into the high Butterworth frequency region. The subfigure D3, a high-pass filter with a frequency of 0.2 f_s, represents its highpass component composed of almost only noise and/or residuals of those sharp and narrow Raman peaks, such as the sapphire peak at 418 cm^{-1}. As for its power plot (D4), true spectral peaks here disappear, and their values are forced to be a certain constant, resulting in predominant spectral noises. For a frequency of 0.006 f_s (row B), the lowpass component (B2) tends to penetrate wider peaks like the water peak at 1640 cm^{-1} while leaving sharper peaks like the sapphire peak at 418 cm^{-1} unchanged. For a frequency of 0.02 f_s (row C), the lowpass components (C2) penetrates even sharp and intense peaks. Broad spectral peaks that have relative low Butterworth frequencies are attenuated, such as power signals in a range between 0.002 f_s and 0.01 f_s

The Raman spectrum is represented as the sum of the pure Raman signal of measured sample, the spectral baseline, and noise or artifacts along all the wavenumbers or Raman shifts. In the context of Raman spectroscopy, the use of a Butterworth filter facilitates the Fourier transformation of Raman signals from the wavenumber domain to the inverse Butterworth frequency domain. This transformation to the frequency domain decomposes the original signals into multiple frequency

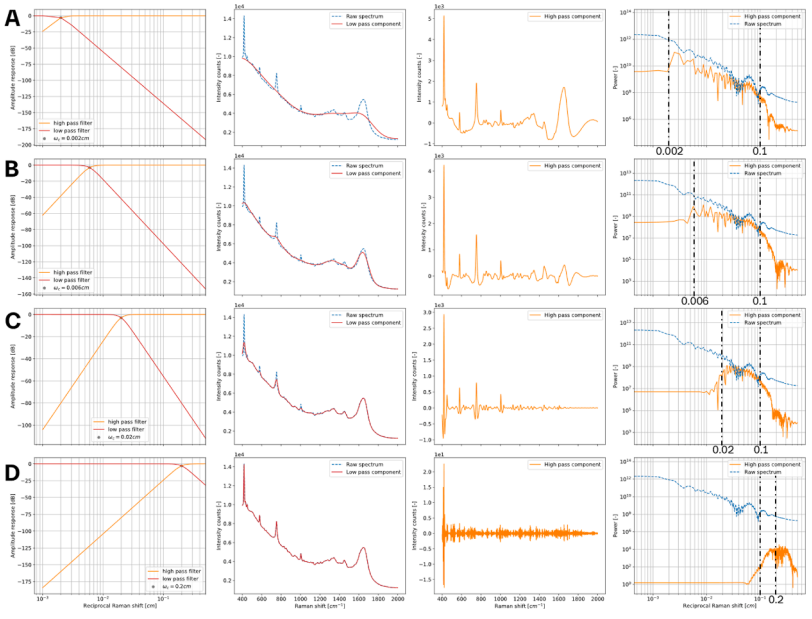

Fig. 1. Working principle of the Butterworth filter for spectral preprocessing. A raw spectrum of a Transferrin sample at 20 mg/mL was taken as an example and truncated to the region between 400 and 2000 cm^{-1}. This spectrum was preprocessed by Butterworth filters with varied cutoff frequencies in range of $[0.002,\ 0.006,\ 0.02,\ 0.2]\ f_s$, from row A to D, respectively. Each row corresponds to a Butterworth filter applying an order of 10 and one cutoff frequencies. Different response and filter components of a Butterworth filter are represented along the columns. The first column shows the filter amplitude responses of the respective Butterworth filter using high- and low-pass. The second column shows the lowpass components applying lowpass design with varied cutoff frequencies. An original raw spectrum is provided as a reference. The third column shows the highpass components applying high pass design with varied cutoff frequencies. The fourth column shows the spectra of the highpass component and raw data in form of power, cataloged by reciprocal Raman shift or Butterworth frequency in cm.

categories and facilitates the filtering of irrelevant Butterworth frequencies. Consequently, a frequency sequence can be hypothesized, ascending from low to high values, that can be classified as follows: spectral baseline < broad peaks < narrow and sharp peaks < noises. By changing the cutoff frequency on the identical raw Transferrin spectrum, the undesirable signals can be effectively removed by preventing their corresponding frequencies from passing through. Thus, the selection of an optimal threshold or cutoff frequency is critical for baseline removal and smoothing in the application of the Butterworth filter.

3.2. Cutoff frequency selection for robust and efficient baseline removal

On two dilution series (mAb1 and Transferrin), we aim to choose the optimal cutoff frequency of Butterworth highpass filter for filtering out baseline. The experiment is based on synthetic disturbances randomly added to the baseline of raw spectra, which results in baseline-disturbed spectra. The cosine similarity θ is used to evaluate the preprocessing method's ability and robustness in filtering out the disturbed baseline. At each protein concentration, the synthetic baseline disturbances were randomly replicated for 40 times and the average cosine similarity θ are computed over all the concentrations. The averaged squared Pearson correlation coefficient r^2 is another metric used to assess the correlation between protein concentration and a given dataset. A higher r^2 value indicates a stronger dependence of the dataset on protein concentration.

Fig. 2 presents the average cosine similarity and averaged Pearson correlation coefficients for the cutoff frequency in a range of 0.001 to 0.02 f_s After an initial increase from 0.4 to 0.8 f_s for cutoff frequencies smaller than 0.005 f_s, the mean r^2 shows a plateau at approximately 0.9 for mAb1 for frequency in the range of 0.007 and 0.015 f_s, while for transferrin a maximum of 0.85 is reached in the range of 0.01 and 0.015. In Fig. 2**a**, a significant r^2 drop occurs from a cutoff frequency of 0.015 f_s for both proteins. In the given cut off frequency range of 0.001 to 0.02 f_s, mAb1 shows a higher r^2 value than Transferrin. In contrast, the SG derivative filter was applied on the identical datasets but obtains a lower r^2 value, approximately 0.68. Fig. 2**b** illustrates a sharp increase in the mean cosine similarity as the frequency rises to 0.005 f_s After this, the mean cosine similarity reaches a maximum exceeding 0.999. The SG derivative filter also shows a high cosine similarity above 0.999. Fig. 3 visualizes the effect of baseline removal using Transferrin as an example. It displays both the undisturbed and disturbed spectra, along with their spectra preprocessed using a Butterworth highpass with a cutoff frequency of 0.005 f_s and a second-derivative SG filter. In Fig. 3**a** and **d**, we observe large deviations in the undisturbed and disturbed spectra at varying concentrations (indicated by colors). These deviations are vastly eliminated or removed after preprocessing, as shown in Fig. 3**e** and **f**. The preprocessed data displayed in Fig. 3**b** and **e** (or Fig. 3**c** and **f**) are nearly identical, making them comparable at each concentration. When applied to the undisturbed raw data, the Butterworth filter improves the mean r^2 value from 0.771 to 0.851 (+10.4 %), whereas the SG filter decreases the value to 0.762 (−1.2 %). Despite the addition of synthetic disturbances to the baseline reducing the value to 0.396, the Butterworth filter achieves a comparable value of 0.797 (+101.3 %) close to 0.851. The SG filter also reaches a comparable value of 0.734 (+85.4 %).

In our analysis of two proteins, mAb1 and Transferrin, ranging from zero to ca. 20 mg/mL, the Butterworth highpass filter effectively removed the random numerical perturbations added to the baseline of the original spectra (Fig. 2**,** Fig. 3). Within the cutoff frequency range of 0.001 to 0.015 f_s, choosing an increasing cutoff frequency enhances the linear dependency between signals and protein concentration. It could recover the signals that were distorted by artificial perturbations, as shown by the rising r^2 value. However, the r^2 profile is not monotonic. The linear correlation starts to decrease significantly when the cutoff frequency reaches 0.015 f_s Interestingly, at a frequency of 0.005 f_s, the r^2 profiles of two proteins diverge, suggesting a difference in their protein molecular information at this point. The Butterworth frequency range from 0.005 to 0.015 f_s could be a transitional zone. In this zone, the primary information source might shift from the spectral baseline to the protein molecular structure.

This disturbance numerical experiment clearly demonstrates the capability of Butterworth filter in removing baseline and its disturbances. It also reveals that a Butterworth frequency larger than 0.015 f_s does not bring about significant improvements in the two metrics and may even lower the protein concentration dependence. Therefore, to maximize the spectral discrepancies in pharmaceutical structures, the right cutoff frequencies are critical. Our findings suggest that the low to middle range, specifically 0.004 to 0.008 f_s, is optimal for removing the Raman spectral baseline and baseline disturbances. Based on these findings, we decided to conduct a further study using a bandpass Butterworth filter. In the previous study, a cutoff frequency of 0.004 f_s (a coefficient of 2) was identified as optimal for eliminating the baseline effect caused by flow rate [15]. Our current research expanded on this by evaluating the full frequency range, rather than limiting to the three frequencies of 0.001, 0.004 and 0.008 f_s This broader evaluation confirmed that the 0.004 f_s significantly outperformed the 0.001 f_s and demonstrated comparable performance to the 0.008 f_s

The SG filter also demonstrated robust performance in baseline removal (with an 85.4 % improvement), even when handling synthetic baseline deviations that rarely manifest in practical scenarios. However,

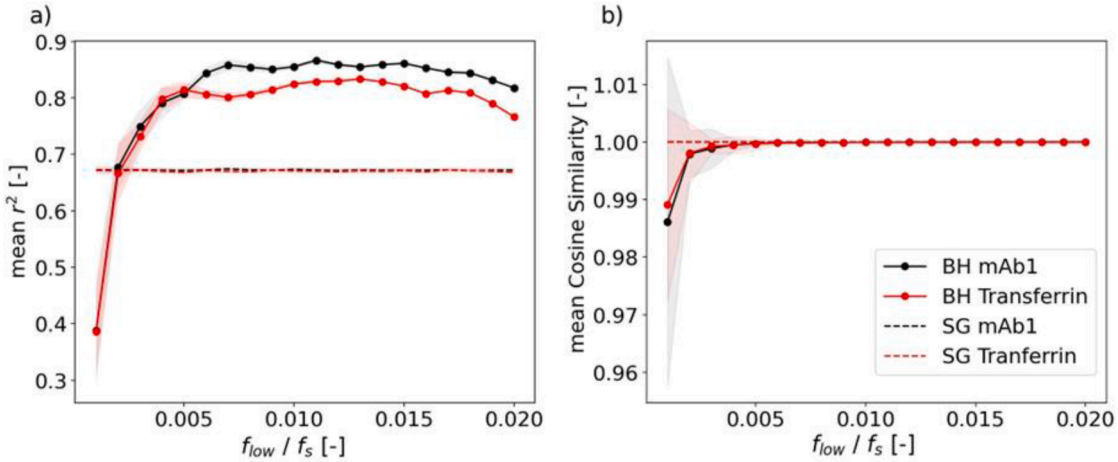

Fig. 2. The effect of Butterworth highpass cutoff frequencies on baseline-disturbance removal. The computation was carried out on two dilution series of mAb1 and Transferrin respectively. Average r^2 a) and cosine similarity b) were computed between the Butterworth filtered data of the baseline-disturbed and -undisturbed spectra, at different concentrations with standard deviation.

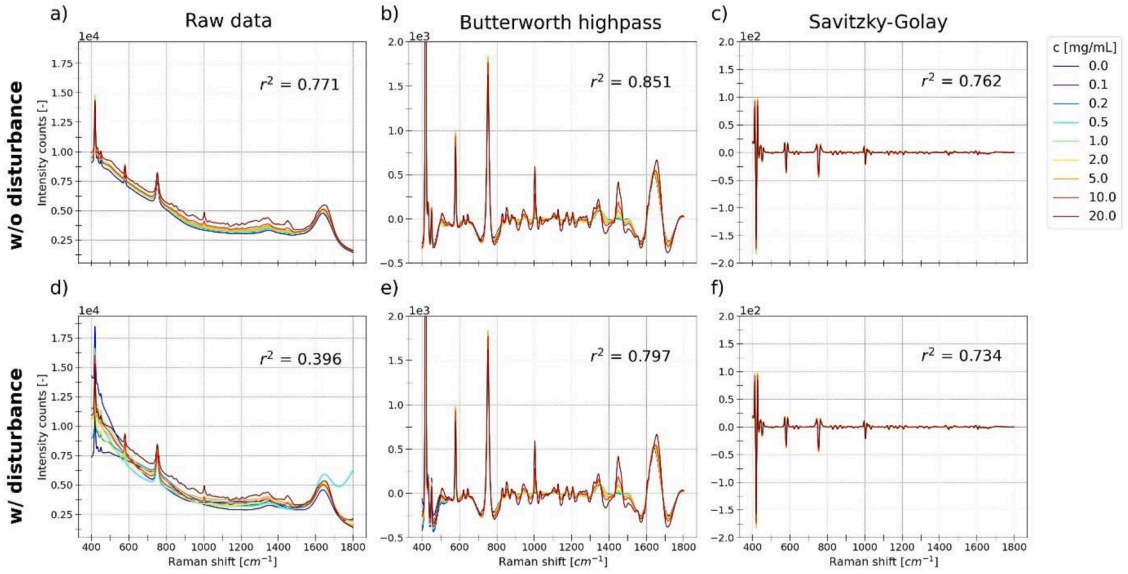

Fig. 3. Schematic representation of baseline distribution and correction by a high-pass Butterworth filter. Raw spectra of Transferrin samples in range from 0 to 20 g/L were used for investigation. a) Raw spectra. b) Raw spectra baseline corrected by a highpass Butterworth filter with $\boldsymbol{n} = 10$ and a cutoff frequency of 0.005 $\boldsymbol{f_s}$ c) Spectra disturbed by random polynomial baseline disturbance. d) Disturbed spectra baseline corrected by the same Butterworth filter as for b). With refer to concentrations, all spectra were plotted in different colors. Average $\boldsymbol{r^2}$ is given in every plot.

the Butterworth filter outperformed the SG filter in achieving a stronger linear dependency between signals and protein concentration in this study. The specific reasons for this difference remain unclear, but potential contributing factors could include the selection of window size and polynomial order for the SG filter, which may not have been optimal for the dataset used in this study. Future studies could involve a systematic evaluation of filtering techniques, focusing on their impact on specific spectral peak and band such as relative peak height and location after filtering [33].

3.3. Butterworth bandpass parameter screening

Unlike the highpass design that employs a single cutoff frequency, the bandpass design utilizes both low and high cutoff frequencies. This is done to extract only the middle frequency region. In our use-case, the goal is not just to remove baseline interferences, but also to avoid amplifying noises at high frequencies. Fig. 4 presents the averaged r^2 results obtained from screening the low and high cutoff frequencies across the two protein dilution series. The low cutoff frequency was tested within a range of 0.001 to 0.014 f_s, where the high cutoff frequency was examined in a broad range of [0.015, 0.5] f_s. Only results from the [0.015, 0.4] f_s and [0.015, 0.1] f_s regions are displayed for mAb1 and Transferrin, respectively. Analogous to Fig. 2, the mean r^2, represented in various colors, measures the protein concentration dependency post the preprocessing step. Despite the identical buffer composition and comparable concentration ranges up to 20 mg/mL (mAb1) and 19.5 mg/mL (Transferrin), the two screenings yielded different patterns for the two proteins. Moreover, mAb1 maximizes the averaged r^2 at $\left(f_{c,low},\ f_{c,high}\right) = (0.008,\ 0.016) f_s$ (in Fig. 4**a**), while Transferrin achieves r^2 maximum near the point $(0.004,\ 0.07)\ f_s$ (in Fig. 4**b**). A decreasing trend of averaged r^2 along the low cutoff frequency is observed, starting from $0.008 f_s$ for mAb1 and $0.005\ f_s$ for Transferrin. In general, mAb1 has a higher value than Transferrin. These findings closely align with the curves in Fig. 2**a**, emphasizing the middle range of low frequency as crucial for effective preprocessing. When compared to the r^2 maximums in Fig. 2**a**, the optimized bandpass design results in an improvement of 11.6 % (from 0.86 to 0.96) for mAb1, and

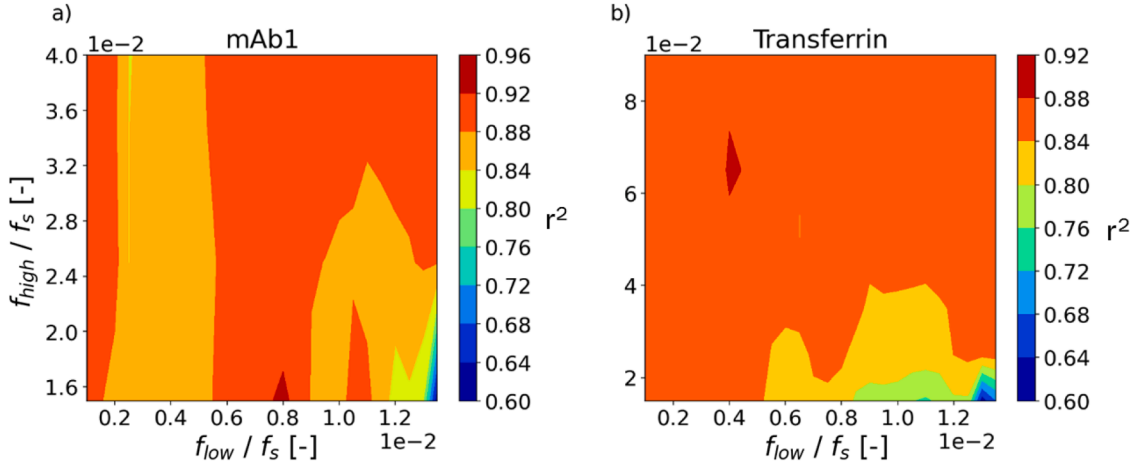

Fig. 4. Preprocessing performances of low and high cutoff frequencies of a band-pass Butterworth filter design. The computed average $\boldsymbol{r^2}$ for a) mAb1 and b) Transferrin under varied high and low cutoff frequencies, applying bandpass Butterworth filters with $n = 10$.

15 % (from 0.80 to 0.92) for Transferrin. The different patterns and maximums underscore the importance of selecting specific cutoff frequencies when implementing the Butterworth bandpass filter for Raman spectra of individual protein molecules.

This line of experiments demonstrated that the bandpass design is more effective than the highpass design in extracting protein concentration dependence from raw Raman spectra. This could be due to the bandpass design's ability to separate high-frequency noise [34] from the Raman data by filtering out specific frequency regions. The Butterworth filter treats one spectral signal as a diversity of periodic waves, each with distinct frequencies or periods. In the bandpass design, only those waves with frequencies or periods below the high cutoff frequency remained post-filtering. Specifically, the optimal cutoff frequency used was from 0.008 to 0.016 f_s for mAb1, corresponding to a wave period of from 62.5 to 125 cm^{-1}, compared to a broader region from 0.004 to 0.07 f_s (or from 14.3 to 250 cm^{-1}) for Transferrin. Given the complex nature of biological molecules, the precise assignment and interpretation of Raman bands linked to proteins poses a challenge. For example, a common feature of biomolecules is the strong sharp band resulting from the vibration of the symmetric ring in phenylalanine, typically observed at 1000–1006 cm^{-1} [35]. The Amide III region, ranging from 1225 to 1280 cm^{-1}, includes common protein secondary structures: random coil (1225–1240 cm^{-1}), beta sheet (1240–1260 cm^{-1}) and alpha-helix (1260–1280 cm^{-1}) [35]. In our study, mAb1 showed more intense and broader Amide III response compared to Transferrin. To isolate and amplify the broad band in the Amide III region, varying the high cutoff threshold could be a strategy. By applying a 4-times higher frequency from 0.016 to 0.07 f_s, it's possible to divide this overlapping band of 55 cm^{-1} (1225–1280 cm^{-1}) into smaller segments, potentially achieving a resolution of 14.3 cm^{-1}. This approach could help in distinguishing between the closely overlapping features, thereby enhancing the analysis of protein secondary structures. Nevertheless, not all values in the screening range for the bandpass design prove effective. Different patterns, each with their unique optima, were identified in the r^2 plots of the two proteins. In the previous study, Wang et al. [4] suggested the augmentation of multiple Butterworth highpass filters across a wide range of cutoff frequencies to generate a 2D Raman image dataset. Their success of predicting multiple CQAs might result from the 2D Raman image dataset, which could potentially include the patterns of different CQAs. While the earlier studies determined the cutoff threshold through empirical experience [4] or by screening within a training dataset [24], our approach employs a screening strategy grounded in experiments with pure protein. This approach not only reduces computational costs but also minimizes the need for extensive wet-lab work, making it a more practical and efficient alternative.

This study also suggests that a specific parameter configuration might be necessary to extract the most relevant data for training a predictive model of an individual protein molecule. To develop a predictive model of a single molecule, the optimal bandpass configuration can be determined by performing a dilution series of the desired molecule. This approach aims to identify the configuration that best isolates the spectral features of interest, enhancing the model's predictability. However, these optimized parameters for the model impurity cannot be simply transferred to real downstream related CQAs due to their different Raman bands. Successfully applying this approach on more complex matrices with interfering species requires a high purity of the molecule. For example, dilution series can be easily made on buffer excipients. In the real downstream processing, achieving high purity and quality of process-related impurities, such as high and low molecular weight species (H/LMWs), can be challenging but can be done by conducting preparative size exclusion chromatography. Specific recombinant host cell proteins can also be purified through capture and size exclusion chromatography steps [36]. Additionally, H/LMWs are size variants of the target protein with structure changes. This change might be less pronounced than those observed in Transferrin, leading to unsatisfactory optimization. Furthermore, the complexity of peak assignment and interpretation can significantly increase after applying filtering techniques. Proper interpretation of these peaks could lead to a deep understanding of the true features versus redundant signals, thereby simplifying the preprocessing method.

3.4. Workflow screening and final model tuning

In the study, the same Raman spectrometer was used to measure all Raman spectra. This spectrometer was mounted on a Tecan system to perform off-line measurements of the mixed samples (Section 2.2.2). The spectrometer was then transferred to an ÄKTA system to gather in-line Raman data for all CEX runs (Section 2.2.3). The raw Raman spectra from these measurements are overlapped and plotted in **Figure S2**. However, due to the use of different Tecan and ÄKTA systems, the spectra measured off-line and in-line show significant differences in their baselines. The training data displays strongly shifted spectral baselines, where the intensities are higher than the rest of the data. In contrast, no significant discrepancy is observed between the two sets of validation data (**Figure S2a, S2b**). This demonstrated the variation in spectral baseline when the same spectrometer is used on different systems. In **Figure S2c**, the baseline shape of test data (shown in yellow) is slightly pulled towards the upper right compared to the training data (shown in black). A significant drop, approximately 1000 counts, is detected in the Raman shift region between 400 and 1720 cm^{-1}. Additionally, in the region from 2000 to 3000 cm^{-1}, the intensities greatly exceed those of the training data. This could be due to the Raman spectrometer's declining performance or instrumental interferences, such as an unstable detector and laser power. This requires a normalization procedure to correct datasets. To successfully implement Raman-based models in downstream processing, it is crucial to ensure data comparability across spectra with deviated baselines. Any variations or inconsistencies in the instrument and system can affect the data comparability, posing challenges to the development of robust regression models. Therefore, a robust data preprocessing workflow should be used to harmonize the data.

Different parameter configurations can be applied to a preprocessing method. However, if the parameters are not configured correctly, the model's predictability may be unsuccessful. To test various preprocessing parameter configurations, we performed a workflow screening, coupled with different model regression algorithms. The model performance of this screening is presented in Fig. 5, using two key metrics, f_1 for Transferrin and f_2 for mAb1. Out of 1620 workflow screenings, only 162 (10 %) workflow candidates yielded both positive results for both f_1 and f_2. The remaining 90 % of workflows failed to predict protein concentrations with negative f_1 or f_2. This low success rate of 10 % highlights the importance of an appropriate workflow that include the preprocessing method and model regression, to ensure successful modelling and data comparability. In Fig. 5**a**, the utopia point, where $f_1 = f_2 = 1$, serves as a reference for the ideal scenario. The six Pareto front candidates, indicated by red circles, are considered the best candidates. Fig. 5**b** shows that among the workflows with positive results, those using the Butterworth filter (green dots) outperformed those using the SG filter. The Butterworth filter was used in 89 workflows (55 %), while the SG filter was used in 73 workflows (45 %). Although the workflows using the SG filters showed promising results in estimating mAb1 concentration with $f_2 > 0.95$, their performance in predicting Transferrin concentration was unsatisfactory. The majority of workflows using SG filter were characterized by f_1 value below 0.7. This underperformance can be attributed to the suboptimal preprocessing by the SG filter for Transferrin, which serves as a model impurity. It is potentially because of the low concentration in the experimental design (max. 4.2 mg/mL, Section 2.2.2). While these workflows are prone to accurately predict the concentration of the primary molecule, mAb1, due to its intense signals, they have less capability in detecting impurities hat are present at low concentration in the bioproduct. In contrast,

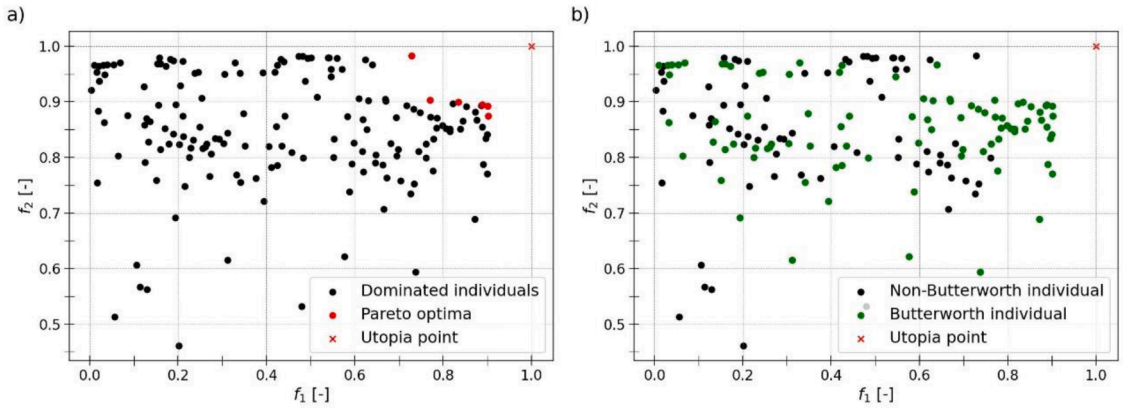

Fig. 5. Workflow screening and optimization results. f_1, f_2 were calculated and used as model performance evaluation. All the workflows of positive f_1, f_2 are shown on the (f_1, f_2) plane in both plots. a) The pareto front points are shown in red diamond. b) the workflows using the Butterworth filter are highlighted in green. The utopia point, where f_1, f_2 equal 1, is highlighted in red cross as reference.

only the workflows using Butterworth filter showed high performance for both mAb1 and Transferrin. This was evident in the performance plot, where both evaluation metrics, f_1 and f_2, exceed 0.8.

Of the six pareto front candidates, the highest value achieved for f_1 was 0.9022, while for f_2 it was 0.9825. All these six candidates have both their metrics exceeding 0.7. Five of these candidates employed the Butterworth bandpass filter but with different hyperparameters. This observation underlines the Butterworth bandpass filter's ability to harmonize the spectral datasets with baseline deviations, therefore enhancing data comparability and model performance. In these five candidates, full-range spectral variables and the SVR were paired with the Butterworth bandpass filter. It was observed that the Butterworth-filtered data showed a better fit with the SVR with a polynomial kernel than with the PLS regressor. Furthermore, the five top candidates shared a common low cutoff frequency of 0.005 f_s, a finding that aligns with the previous results shown in Figs. 2 and 3. Detailed performance metrics and corresponding workflows for these candidates can be found in Table 3.

As mentioned earlier, the spectral baseline in Test data has a significant shift (**Figure S2c**) and necessitates a normalization procedure for correction [29]. Consequently, during the final model tuning, we applied an additional normalization procedure. Based on the screening results, full spectral variables and the SVR with a polynomial kernel were used to develop single-output models. These models were designed to predict concentrations of Transferrin and mAb1 in the CEX runs. The final tuning tested the best bandpass configurations derived from previous study results (Section 3.3). In the final models, the optimal configurations identified in the bandpass screening (Fig. 4) were applied. The mAb1 model utilized a low cutoff frequency of 0.008 f_s and a high cutoff frequency of 0.016 f_s For the Transferrin model, a low cutoff frequency of 0.004 f_s and a high cutoff frequency of 0.07f_s were used. This confirms that the optimal configurations are both promising and effective, as evidenced not only in the r^2 plots but in model development. It suggests that the averaged r^2 may be a meaningful representation for the spectral extraction of relevant data. Therefore, using the r^2 value could be a promising approach to identify the optimal parameter.

Table 3
The applied approaches and model outputs of nine workflow candidates.

Candidate	Truncation	Preprocessing workflow	Regressor	f_1	f_2
1	Full-range	Butterworth bandpass (10-pole, [0.005, 0.04] f_s)	SVR*	0.9022	0.8745
2	Full-range	Butterworth bandpass (10-pole, [0.005, 0.06] f_s)	SVR*	0.9007	0.8921
3	Full-range	Butterworth bandpass (10-pole, [0.005, 0.07] f_s)	SVR*	0.8881	0.8953
4	Full-range	Butterworth bandpass (10-pole, [0.005, 0.08] f_s)	SVR*	0.8344	0.8996
5	Full-range	Butterworth bandpass (10-pole, [0.005, 0.09] f_s)	SVR*	0.7700	0.9027
9	Range 2	SavGol derivative ($n_{diff} = 2$, window = 89, $n_{poly} = 3$)	PLSR**	0.7288	0.9825

SVR*: Support Vector Regression, $C = 1000$, epsilon = 0.01, degree = 3.
PLSR**: Partial Least Square Regression, number of components = 5.

Fig. 6 presents the in-line predictions given by the final model for two validation and test datasets over the elution time. In Fig. 6**a, c** and **e**, the concentration predictions for mAb1 (in blue) and Transferrin (in red) are shown. These predictions are displayed in form of chromatograms alongside off-line concentrations, in-line pH, and conductivity profiles. The in-line prediction curves for both proteins align well with the off-line measurements. Fig. 6**b, d** and **f** show a parity plot, which compares the off-line measured concentrations with the model predictions for each dataset. As annotated in these figures, the R^2 or Q^2 values reach the highest at 0.99 for mAb1 (Fig. 6**b**) and 0.95 for Transferrin (Fig. 6**d**). In all three datasets, the mAb1 model demonstrated robust predictability, with R^2 and Q^2 values exceeding 0.95. Despite precisely predicting the concentration profiles that closely match off-line measurements, the Transferrin model exhibits weaker predictability as shown in Fig. 6**b** and 6**f**. In *Validation 2*, the reduction in salt gradient length led to earlier elution from the column and narrowed protein profiles. This caused a significant overlap of proteins between the 6 to 9 min, a region specifically targeted to test the model's ability to differentiate the protein mixture. In this scenario, the Transferrin model made accurate predictions. However, the mAb1 model failed to accurately predict this high overlap, unexpectedly revealing a mAb1 peak from 4 to 7 min. As the Transferrin concentration here exceeded the calibration range of 4.2 mg/mL, we replicated *Validation 2* using same experimental setup, but with a reduced amount of Transferrin. In this replication, referred to as the *Test* experiment, the mAb1 model predicted the overlap with greater accuracy, without showing the unexpected peak. Furthermore, the *Test* experiment greatly improved due to the normalization procedure, which effectively corrected the baseline shift.

For evaluating the feasibility of the three preprocessing methods for real-time monitoring, we also assessed their computational based on the

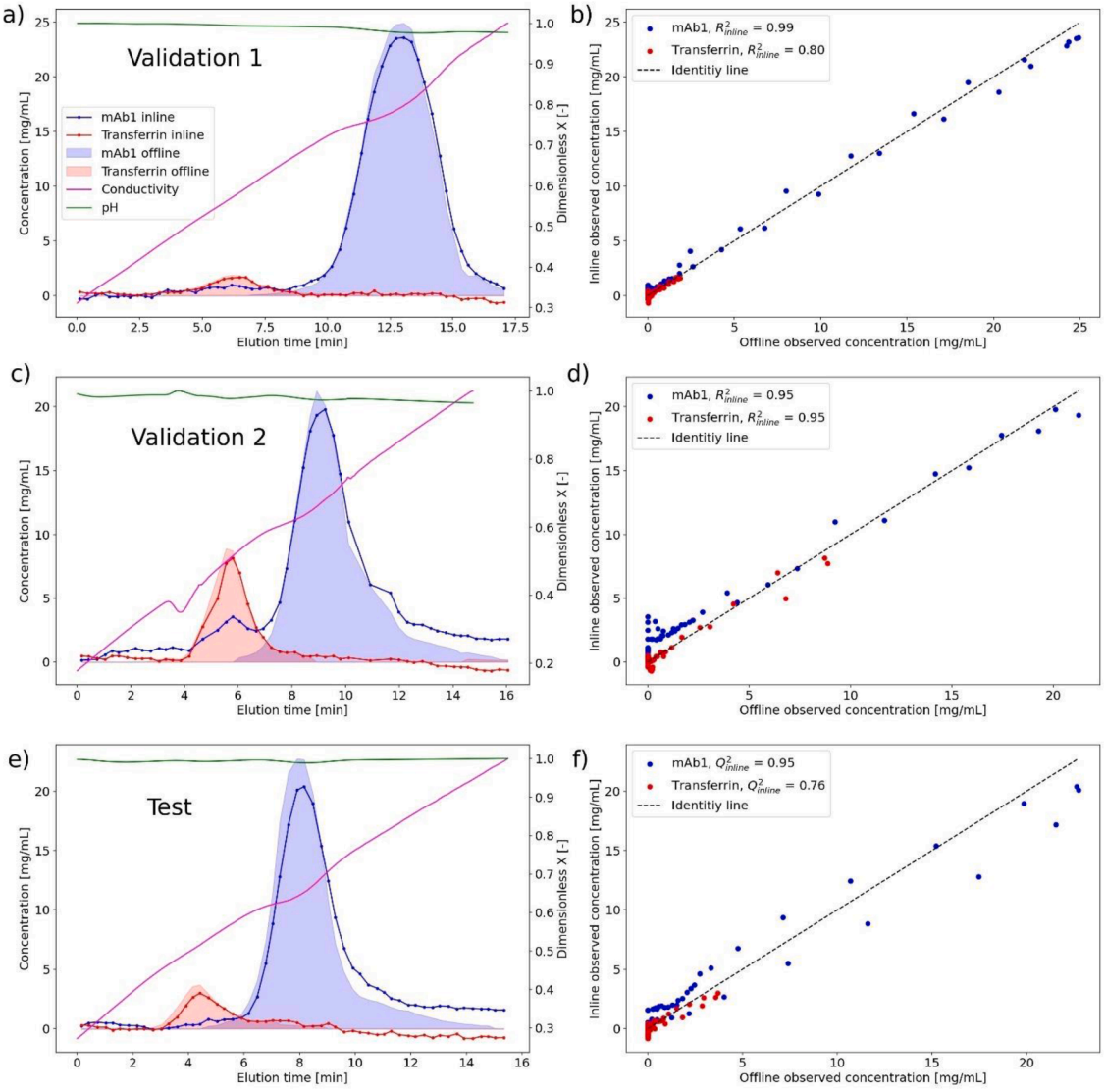

Fig. 6. Chromatograms of three chromatography experiments and comparison of their in-line Raman predictions and off-line measurements. The three rows show the results of validation 1, validation 2, test experiments, respectively, from top to bottom. On the left side, their chromatograms are represented with off-line measurements (filled area), in-line Raman predictions (line with dots), in-line conductivity (in pink) and in-line pH (in green). Right next to each chromatogram, the off-line observation concentrations were compared with the in-line Raman predictions in a single plot with refer to Transferrin (in red) and mAb11 (in blue). A diagonal line is shown as reference.

in-line Raman data collected from the three chromatography runs. The SG filter demonstrated the lowest computational cost, with an average processing time of 27.8 ms per spectrum. The Butterworth bandpass filter required slightly more time, averaging 30.8 ms per spectrum. In contrast, the Butterworth highpass filter had a higher computational cost, with an average processing time of 37.9 ms per spectrum. Including the model prediction time of approx. 20 – 70 ms, the total processing time ranged from 50 to 100 ms, significantly faster than the spectrum acquisition of 7.5 s. This suggests that the integration of these methods into real-time Raman analysis would not pose a computational bottleneck.

4. Conclusion

The study successfully illustrated the potential of the Butterworth filter for decomposing Raman spectra in the downstream processing of biopharmaceuticals. A detailed investigation was conducted on the highpass design to understand the basics of cutoff frequency and the frequency domain. This investigation focused on the impact of the cutoff frequency on filtering performance and the extraction of protein-related Raman data. This digital filter is able to decompose original Raman signals into multiple frequency categories, enabling the filtering out of irrelevant frequencies by preventing them from passing through. It was found that the spectral baseline could be transformed into a middle range of cutoff frequency, specially from 0.004 to 0.008 f_s This range is crucial for baseline removal when decomposing Raman spectra of biopharmaceuticals. This finding is also in line with the optimal cutoff frequency of 0.004 f_s in eliminating flowrate effect in previous study [15]. The study also revealed that a cutoff frequency of 0.005 f_s is an optimal threshold. This enhances protein signal recovery and model predictability in chromatography runs.

Following the understanding of highpass, the study delved into the bandpass design. This design filters out not only the baseline but also high-frequency noises, though the screening of low and high cutoff frequencies. The findings indicate that the Butterworth bandpass filter outperforms the highpass design in improving the linear correlation between Raman signals and protein concentrations. The study also underscores the importance of specific cutoff frequency selection when

applying the bandpass design to the Raman spectra of individual protein molecules. Our approach determines the parameter configuration using experiments with pure protein, offering a practical and efficient alternative to empirical methods or computationally intensive screening, with reduced costs and minimal lab work. The optimal configurations identified through screening was successfully validated in real downstream chromatography scenarios. Directly applying these optimal configurations in developing predictive models, high coefficients of determination were achieved for IgG ($R^2 = 0.99$) and Transferrin ($R^2 = 0.95$). Efficient bandpass parameter tuning can provide precise frequency configuration for the post modeling step, thereby enhancing the implementation of Raman spectroscopy in the downstream processing of biopharmaceuticals. While the proposed Butterworth bandpass with a screening strategy delivers an efficient refinement on developing chemometric models for proteins in downstream processes, its transferability to real CQAs has to be further studied and verified, gaining more comprehensive understanding on the peak assignment and interpretation.

CRediT authorship contribution statement

Jingyi Chen: Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology. **José Munoz Reyes:** Writing – original draft, Visualization, Methodology, Investigation. **Robin Schiemer:** Writing – review & editing. **Gang Wang:** Supervision, Investigation. **Joey Studts:** Writing – review & editing, Resources, Project administration. **Matthias Franzreb:** Supervision, Investigation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jingyi Chen reports financial support was provided by Boehringer Ingelheim Pharma GmbH & Co KG If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.chroma.2025.466069.

Data availability

The authors do not have permission to share data.

References

[1] K.A. Esmonde-White, M. Cuellar, I.R. Lewis, The role of Raman spectroscopy in biopharmaceuticals from development to manufacturing, Anal. Bioanal. Chem. 414 (2022) 969–991, https://doi.org/10.1007/s00216-021-03727-4.

[2] J. Glassey, K.V. Gernaey, C. Clemens, T.W. Schulz, R. Oliveira, G. Striedner, C.-F. Mandenius, Process analytical technology (PAT) for biopharmaceuticals, Biotechnol. J. 6 (2010) 369–377, https://doi.org/10.1002/biot.201000356.

[3] Y.K. Lin, H.Y. Leong, T.C. Ling, D.-Q. Lin, S.-J. Yao, Raman spectroscopy as process analytical tool in downstream processing of biotechnology, Chinese J Chem Eng 30 (2021) 204–211, https://doi.org/10.1016/j.cjche.2020.12.008.

[4] J. Wang, J. Chen, J. Studts, G. Wang, Simultaneous prediction of 16 quality attributes during protein A chromatography using machine learning based Raman spectroscopy models, Biotechnol. Bioeng. (2023), https://doi.org/10.1002/bit.28679.

[5] C.H. Wegner, S.M. Eming, B. Walla, D. Bischoff, D. Weuster-Botz, J. Hubbuch, Spectroscopic insights into multi-phase protein crystallization in complex lysate using Raman spectroscopy and a particle-free bypass, Front. Bioeng. Biotechnol. 12 (2024) 1397465, https://doi.org/10.3389/fbioe.2024.1397465.

[6] A. Dietrich, R. Schiemer, J. Kurmann, S. Zhang, J. Hubbuch, Raman-based PAT for VLP precipitation: systematic data diversification and preprocessing workflow identification, Front. Bioeng. Biotechnol. 12 (2024) 1399938, https://doi.org/10.3389/fbioe.2024.1399938.

[7] B.S. McAvan, L.A. Bowsher, T. Powell, J.F. O'Hara, M. Spitali, R. Goodacre, A. J. Doig, Raman spectroscopy to monitor post-translational modifications and degradation in monoclonal antibody therapeutics, Anal Chem 92 (2020) 10381–10389, https://doi.org/10.1021/acs.analchem.0c00627.

[8] S. Guo, T.Bocklitz J.Popp, Chemometric analysis in Raman spectroscopy from experimental design to machine learning–based modeling, Nat. Protoc. 16 (2021) 5426–5459, https://doi.org/10.1038/s41596-021-00620-3.

[9] F. Feidl, S. Garbellini, S. Vogg, M. Sokolov, J. Souquet, H. Broly, A. Butté, M. Morbidelli, A new flow cell and chemometric protocol for implementing in-line Raman spectroscopy in chromatography, Biotechnol. Prog. 35 (2019) e2847, https://doi.org/10.1002/btpr.2847.

[10] S. Guo, C. Beleites, U. Neugebauer, S. Abalde-Cela, N.K. Afseth, F. Alsamad, S. Anand, C. Araujo-Andrade, S. Aškrabić, E. Avci, et al., Comparability of Raman spectroscopic configurations: a large scale cross-laboratory study, Anal. Chem. 92 (2020) 15745–15756, https://doi.org/10.1021/acs.analchem.0c02696.

[11] H.J. Butler, L. Ashton, B. Bird, G. Cinque, K. Curtis, J. Dorney, et al., Using Raman spectroscopy to characterize biological materials, Nat. Protoc. 11 (2016) 664–687, https://doi.org/10.1038/nprot.2016.036.

[12] C. Battistoni, G. Mattogno, G. Righini, Spectral noise removal by new digital smoothing routine, J. Electron Spectrosc. Relat. Phenom. 74 (1995) 159–166, https://doi.org/10.1016/0368-2048(95)02363-1.

[13] D. Wei, S. Chen, Q. Liu, Review of Fluorescence Suppression Techniques in Raman Spectroscopy, Appl. Spectrosc. Rev. 50 (2015) 387–406, https://doi.org/10.1080/05704928.2014.999936.

[14] S. Goldrick, D. Lovett, G. Montague, B. Lennox, Influence of incident wavelength and detector material selection on fluorescence in the application of raman spectroscopy to a fungal fermentation process, Bioeng. (Basel, Switz.) 5 (2018) 79, https://doi.org/10.3390/bioengineering5040079.

[15] J. Wang, J. Chen, J. Studts, G. Wang, In-line product quality monitoring during biopharmaceutical manufacturing using computational Raman spectroscopy, mAbs 15 (2023) 2220149, https://doi.org/10.1080/19420862.2023.2220149.

[16] L. Rolinger, M. Rüdt, J. Hubbuch, Comparison of UV- and Raman-based monitoring of the Protein A load phase and evaluation of data fusion by PLS models and CNNs, Biotechnol. Bioeng. 118 (2021) 4255–4268, https://doi.org/10.1002/bit.27894.

[17] S. Guo, R. Heinke, S. Stöckel, P. Rösch, J. Popp, t. Bocklitz, Model transfer for Raman-spectroscopy-based bacterial classification, J. Raman Spectrosc. 49 (2018) 627–637, https://doi.org/10.1002/jrs.5343.

[18] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639, https://doi.org/10.1021/ac60214a047.

[19] C.D. Brown, P.D. Wentzell, Hazards of digital smoothing filters as a preprocessing tool in multivariate calibration, J. Chemom. 13 (1999) 133–152, https://doi.org/10.1002/(SICI)1099-128X(199903/04)13:2<133::AID-CEM533>3.0.CO;2-C.

[20] B. Wei, N. Woon, L. Dai, R. Fish, M. Tai, W. Handagama, A. Yin, J. Sun, A. Maier, D. McDaniel, E. Kadaub, J. Yang, M. Saggu, A. Woys, O. Pester, D. Lambert, A. Pell, Z. Hao, G. Magill, J. Yim, J. Chan, L. Yang, F. Macchi, C. Bell, G. Deperalta, Y. Chen, Multi-attribute Raman spectroscopy (MARS) for monitoring product quality attributes in formulated monoclonal antibody therapeutics, mAbs 14 (2022) 2007564, https://doi.org/10.1080/19420862.2021.2007564.

[21] F.M. Ham, I.N. Kostanic, G.M. Cohen, B.R. Gooch, Determination of glucose concentrations in an aqueous matrix from NIR Spectra using optimal time-domain filtering and partial least-squares regression, IEEE Trans. Biomed. Eng. 44 (1997) 475–485, https://doi.org/10.1109/10.581938.

[22] H. Chen, W. Xu, N. Broderick, J. Han, An adaptive denoising method for Raman spectroscopy based on lifting wavelet, J. Raman Spectrosc. 49 (2018) 1529–1539, https://doi.org/10.1002/jrs.5399.

[23] L. Fabiola, R.E. Miguel, M. Martin O, D.M. Guadalupe, R.A. Ma del Carmen, A. Alfonso, Denoising of Raman spectroscopy for biological samples based on empirical mode decomposition, Int. J. Mod. Phys. C. 28 (2007) 1750116, https://doi.org/10.1142/S0129183117501169.

[24] J. Chen, J. Wang, R. Hess, G. Wang, J. Studts, M. Franzreb, Application of Raman spectroscopy during pharmaceutical process development for determination of critical quality attributes in Protein A chromatography, J. Chromatogr. A (2024) 464721, https://doi.org/10.1016/j.chroma.2024.464721.

[25] J. Zhao, H. Liu, D.I. McLean, H. Zeng, Automated Autofluorescence background subtraction algorithm for biomedical Raman spectroscopy, Appl. Spectrosc. 61 (2007) 1225–1232. https://opg.optica.org/as/abstract.cfm?URI=as-61-11-1225.

[26] A.K. Das, S. Halder, Baseline wander correction and impulse noise suppression using cascaded empirical mode decomposition and improved morphological algorithm, IJAER 12 (2017) 2329–2337, https://doi.org/10.37622/IJAER/12.10.2017.2329-2337.

[27] A.S. Moorthy, A.J. Kearsley, Pattern Similarity Measures Applied to Mass Spectra, Progress in Industrial Mathematics: Success Stories, 5, SEMA SIMAI Springer SeriesSpringer, Cham, 2020, https://doi.org/10.1007/978-3-030-61844-5_4.

[28] M. Agaria, S. Bianco, L. Celona, R. Schettini, M. Tchobanou, An analysis of spectral similarity measures, in: Proc. IS&T 29th Color and Imaging Conf. 29, 2021, pp. 300–305, https://doi.org/10.2352/issn.2169-2629.2021.29.300.

[29] N.K. Afseth, A. Kohler, Extended multiplicative signal correction in vibrational spectroscopy, a tutorial, Chemom. Intell. Lab. Syst. 117 (2012) 92–99, https://doi.org/10.1016/j.chemolab.2012.03.004.

[30] K.H. Liland, A. Kohler, N.K. Afseth, Model-based preprocessing in Raman spectroscopy of biological samples, J. Raman Spectrosc. 47 (2016) 643–650, https://doi.org/10.1002/jrs.4886.

[31] G. Rabatel, F. Marini, B. Walczak, J. Roger, VSN: variable sorting for normalization, J. Chemom. 34 (2020), https://doi.org/10.1002/cem.3164.

[32] P. Beumers, Physically-Based Models for the Analysis of Raman Spectra, Wissenschaftsverlag Mainz GmbH, 2019.
[33] M. Schmid, D. Rath, U. Diebold, Why and how Savitzky–Golay filters should be replaced, ACS Meas. Sci. Au 2 (2022) 185–196, https://doi.org/10.1021/acsmeasuresciau.1c00054.
[34] D. Plazas, F. Ferranti, Q. Liu, M.L. Choobbari, H. Ottevaere, A study of high-frequency noise for microplastics classification using raman spectroscopy and machine learning, Appl. Spectrosc. 78 (2024) 567–578, https://doi.org/10.1177/00037028241233304.
[35] D.W. Shipp, F. Sinjab, L. Notingher, Raman spectroscopy: techniques and applications in the life sciences, Adv. Opt. Photon. 9 (2017) 315–428, https://doi.org/10.1364/AOP.9.000315.
[36] M. Maier, S. Schneider, L. Weiss, S. Fischer, D. Lakatos, J. Studts, M. Franzreb, Tailoring polishing steps for effective removal of polysorbate-degrading host cell proteins in antibody purification, Biotechnol. Bioeng. (2024), https://doi.org/10.1002/bit.28767.