Vibrational Spectroscopy 147 (2026) 103960

Contents lists available at ScienceDirect

Vibrational Spectroscopy

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

From raw to refined: Quality-metric-guided preprocessing optimization for Raman spectroscopy in biopharmaceutical downstream processing

Yue Guo [a,b], Robin Schiemer [a], Gang Wang [a], Joey Studts [a], Jürgen Hubbuch [b,*]

[a] Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, 88397, Germany
[b] Institute of Process Engineering in Life Sciences Section IV: Biomolecular Separation Engineering, Karlsruhe Institute of Technology (KIT), Fritz-Haber-Weg 2, Karlsruhe, 76131, Germany

ARTICLE INFO

Keywords:
Raman spectroscopy
Data preprocessing
Baseline correction
Chemometrics
Process analytical technology
Therapeutic antibody

ABSTRACT

The application of Raman spectroscopy is increasingly applied in biopharmaceutical manufacturing for real-time process monitoring and control. In downstream bioprocessing, it offers high potential for in-line quantification of CQAs directly from spectral data using chemometric and machine learning models. Successful implementation, however, strongly depends on robust spectral preprocessing to remove noise, baseline drifts, and interfering signals. Baseline correction remains particularly challenging due to signal variability across various process conditions. Current preprocessing optimization practices typically emphasize minimizing model prediction errors but often disregard the chemical validity of the resulting spectra. To address this gap, we introduce a regression-oriented metric that assesses preprocessing quality based on chemically meaningful criteria by quantifying baseline correction artifacts and spectral intensity-concentration linearity. Paired with an empirically identified preprocessing sequence, this metric is integrated into an automated workflow for systematic baseline correction optimization. Application across datasets comprising mAb and downstream buffer samples demonstrates that the metric reliably ranks preprocessing performance and consistently selects configurations that suppress artifacts and preserve analyte-specific spectral features. Within the evaluated datasets and baseline correction algorithms, benchmarking against model-centric optimization highlights the trade-off between prediction accuracy and spectral interpretability when relying solely on model RMSE. In contrast, the quality metric-driven approach not only aligns with modeling objectives but also safeguards chemically relevant spectral information, yielding sufficiently high predictive power and improved robustness under changing sample conditions. The quality-metric-guided workflow provides a practical methodology for developing preprocessing pipelines for quantitative analysis that supports robust Raman-based PAT in downstream bioprocessing.

1. Introduction

Protein-based therapeutics such as monoclonal antibody (mAb) has profoundly advanced modern medicine by delivering unparalleled precision and efficacy in the treatment of complex diseases [1,2]. The growing demand for biologics has spurred continuous innovation in biomanufacturing to ensure product consistency while improving process efficiency. Guided by Quality by Design (QbD) principles and Process Analytical Technology (PAT) frameworks introduced by the U.S. Food and Drug Administration (FDA) about two decades ago [3], the biopharmaceutical industry is transitioning from conventional end-point batch testing towards real-time monitoring of critical quality attributes (CQAs) [4], with the aim to streamline manufacturing with reduced time and cost. Among the available spectroscopic tools, Raman spectroscopy distinguishes itself through its non-invasive, label-free nature and insensitivity to water interference, making it particularly well-suited for tracking aqueous biological systems [5]. Its high molecular specificity provides a distinct spectral fingerprint, enabling simultaneous resolution of multiple CQAs [6,7]. In downstream processing (DSP), Raman-based PAT solutions have been explored across various stages from chromatography [6,8,9] to filtration [10] for enhanced process control and quality assurance.

Despite its broad adoption, the practical implementation of Raman spectroscopy in bioprocessing remains challenging due to the inherent complexity of biological spectra. These spectra typically exhibit overlapping vibrational bands from proteins and matrix components,

* Corresponding author.
E-mail address: juergen.hubbuch@kit.edu (J. Hubbuch).

https://doi.org/10.1016/j.vibspec.2026.103960
Received 16 May 2026; Received in revised form 28 August 2026; Accepted 9 September 2026
Available online 11 September 2026

resulting in highly convoluted data that hinders straightforward interpretation [11]. Moreover, raw Raman signals are frequently distorted by various artifacts. A major contributor is fluorescence-induced baseline drift [12,13]. Baseline signals often exceed Raman bands by several orders of magnitude [14], hence obscuring subtle features critical for discriminating and quantifying CQAs. Raman measurements are also affected by multiple noise sources originating from samples, instruments, and environmental factors [15,16]. Variations in process conditions such as temperature, concentrations, flow rate introduce additional inconsistencies like intensity offset, further complicating spectral analysis [17,18]. To extract meaningful information, chemometric approaches [19]—combining mathematical preprocessing algorithms with multivariate calibration techniques such as partial least squares (PLS) regression—are indispensable for transforming raw spectra into reliable predictive models that provides actionable process insights.

Preprocessing is the critical step aimed at eliminating the undesirable artifacts and standardizing the Raman measurement before chemometric modeling. Commonly applied operations include baseline correction, smoothing, cropping, and scaling, each contributing to improved spectral quality in different aspects. The consecutive application of different preprocessing operations is usually referred to as a preprocessing pipeline. Among these steps, baseline correction is generally considered the most influential because baseline removal significantly impacts the peak visibility. Therefore, our study is grounded in the fundamental assumption that, similar to other spectroscopic methods, baseline correction also benefits Raman spectra analysis by enabling clearer peak isolation and facilitating spectral standardization.

To date, numerous baseline subtraction methods have been developed [20,21]. Among these, asymmetric least squares (AsLS) represents a simple yet effective category. Although many algorithmic variations exist under this umbrella [22–27], they all share the core principle of balancing the fidelity and smoothness of the fitted baseline by asymmetrically weighting deviations between the observed signal and the estimated baseline. This approach's strength lies in handling complex, nonlinear baselines, where polynomial fitting often fails, while avoiding noise amplification, a common drawback of derivative-based techniques [20,28]. Despite the availability of advanced baseline correction methods based on deep learning [29,30], wavelet transform [31] and convex optimization [32], AsLS-family algorithms remain an attractive and practical choice for industrial Raman PAT applications, because they require no labelled baselines, are computationally efficient and straightforward to implement [26,33]. However, no single baseline correction method is universally optimal. Thus, optimizing preprocessing through careful algorithm selection, parameter tuning, and pipeline order determination is essential, as these decisions ultimately govern the accuracy, robustness, and interpretability of the final model.

Currently, no standardized procedure exists to specify when certain preprocessing methods should be applied or avoided. Identifying an optimal preprocessing pipeline for Raman spectra is typically case-specific. Most published quantitative Raman studies select preprocessing by minimizing model errors, such as root mean square error (RMSE) of cross validation (CV) or prediction [10,21,34–38]. Although this model-oriented approach is fit for predictive performance, it introduces two major risks relevant for analyte concentration regression. First, a low error alone does not guarantee that the processed spectra preserve chemically meaningful, analyte-specific spectral responses. There is a likelihood that regression models may exploit residual artifacts such as fluorescence background, noise patterns or peak distortions that are inadvertently correlated with the target variables, rather than chemically interpretable signals [10,39]. Consequently, predictive performance alone cannot establish that the processed spectra adequately represent the underlying analyte information [19]. Second, this model-centric approach often makes preprocessing model- or dataset-specific [40]. A preprocessing pipeline optimized for a specific model may become ineffective when the model architecture or parameterization changes. Likewise, pipelines highly tailored to the variance structure of one dataset may fail under different experimental conditions [41]. This limited transferability can reduce model robustness and necessitate repeated preprocessing optimization, increasing both development effort and maintenance burden. Therefore, independent evaluation of spectral quality and preprocessing effects alongside model performance is necessary to ensure both chemical interpretability and long-term model reliability [19,39,41].

Alternatively, a quality-parameter-based workflow has emerged [20], where a predefined metric objectively evaluates the performance of preprocessing steps based on the quality of the processed spectra. This approach focuses on artifact removal independently of the predictive model. Without the need to construct statistical models, this method can reduce computational demands and improve optimization efficiency. It may also promote better model transferability. Initial efforts by Guo *et al.* developed a numeric metric, R^{12}, for evaluating baseline correction quality in a classification setting [39]. The R^{12} metric compares signal retained in expected peak regions with residual signal in non-peak regions, thereby rewarding peak preservation and baseline removal. Since this study mainly focused on tackling classification objectives, the R^{12} metric misses a dedicated term that assesses spectral response to analyte concentration, which is vital for quantitative concentration regression questions routinely required in downstream process analyses.

To circumvent the limitations of model-oriented approaches, our study aims to develop a quality-guided preprocessing workflow specifically for quantitative Raman regression applications in biopharmaceutical downstream PAT. To achieve this goal, diverse Raman spectral datasets were generated from single-component concentration series comprising a mAb and a range of downstream-relevant buffers. These datasets were collected under varying process parameters (pH and concentration) and different Raman measurement settings (exposure time, laser power, flow rate), thereby capturing realistic process variability in spectral preprocessing. Building on these datasets, the sequence of preprocessing steps was first examined through manual trials, where inconsistencies in baseline correction are revealed and synergistic effects among the preprocessing steps are briefly discussed. Subsequently, a novel metric that combines both evaluation of artifacts removal and preservation of spectral intensity-concentration linearity is introduced. This metric was integrated into an automated workflow for screening baseline correction algorithms and tuning associated parameters across three classical AsLR-based baseline estimation algorithms. In the present work, a particular emphasis is placed on improving the baseline correction step due to its dominant impact on the overall chemometric analysis. In order to clearly isolate the effects of process variability on baseline preprocessing, the study is restricted to single-component samples. Compared to extensively utilized model-based optimization, quality metric-based approaches for selecting Raman preprocessing remain scarcely discussed in literature. To the best of the authors' knowledge, this is the first study that systematically accounts for process variability and specifically addresses Raman preprocessing in-depth within the context of PAT implementation for biopharmaceutical downstream processing.

2. Materials and methods

2.1. Experiments

2.1.1. Single-component spectra library

A pharmaceutical IgG monoclonal antibody (hereafter referred to as mAb) was provided by Boehringer Ingelheim Pharma GmbH & Co. KG (Biberach, Germany). The mAb drug substance was concentrated to 50 g/L by ultrafiltration and diafiltrated (Unagi, Unchained Labs, Pleasanton, CA, USA) into purified water. Subsequently, it was diluted to 18 concentration levels ranging from 0.1 to 50 mg/mL (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 2.5, 5, 7.5, 10, 15, 20, 25, 50 mg/mL) using purified water.

Buffer systems investigated included acetate, citrate, and glycine. Acetate stock solution at 100 mM was prepared and titrated to six pH values (4.5, 5.0, 5.5, 6.0, 6.5, 7.0), then diluted to eleven concentrations (2, 4, 6, 8, 10, 20, 30, 40, 50, 75, and 100 mM). Citrate and glycine buffer datasets were prepared in the same manner. For citrate, the same eleven concentration levels ranging from 2 to 100 mM were adjusted to six different pH values: 3.0, 4.0, 5.0, 5.5, 6.0 and 6.5. Glycine buffers were titrated to pH values of 3.0, 3.5, 4.0, 5.0 and 6.0, and prepared at twelve concentration levels, ranging from 2 to 300 mM (2, 4, 6, 8, 10, 20, 50, 100, 150, 200, 250, 300 mM). All pH adjustments were performed using sodium hydroxide (NaOH) or hydrochloric acid (HCl) (Aug. Hedinger GmbH & Co. KG), which are not Raman-active.

2.1.2. Raman spectroscopy

In this study, multiple acquisition setups and Raman measurement parameters were intentionally varied to introduce spectral diversity and evaluate the robustness of preprocessing methods. Raman spectra were collected using a HyperFlux Pro Plus spectrometer (Tornado Spectral Systems, Mississauga, Ontario, Canada), operated via SpectralSoft 3.4 software. The instrument utilized a 785 nm excitation laser, covering a wavenumber range of 200–3300 cm^{-1} with a resolution of 1 cm^{-1} . Two acquisition setups were implemented: static and dynamic.

In the static mode, a Hudson Probe equipped with a 45 µL Micro Flow Cell (MFC) was connected to the spectrometer for fluidic measurements. The flow cell was mounted on a Tecan Fluent 780 Liquid Handler (Tecan, Männedorf, Switzerland) to enable automated sample handling, as described previously [8,18]. Here, the generated dilution series of mAb and buffers (buffers also at different pH) were measured under a constant condition with an exposure time of 500 ms per spectrum for 2 min at maximum laser power 495 mW. Concentration and pH difference introduce multiple simultaneous variations, affecting Raman spectra by altering absolute intensity and causing peak shape changes. For mAb specifically, exposure time (400, 500, 800 ms) and laser power (300, 400, 495 mW) were additionally varied systematically to induce intensity offsets, simulating fluctuations that commonly occur across instruments, manufacturing sites, and operational scales. These variations were applied to selected mAb concentrations (low: 5 g/L; high: 50 g/L) to represent potential changes across the full concentration range. Spectral noise was modulated by adjusting the number of averaged scans, resulting in total acquisition times of either 8 s or 2 min. This mimics scenarios in which rapid measurements are required for real-time process monitoring versus prolonged acquisitions suitable for processes with slow dynamics or offline analysis. Including noise variability challenges preprocessing algorithms to handle spectra of differing quality.

Dynamic acquisition was performed on an ÄKTA Avant 25 system (Cytiva, Uppsala, Sweden) controlled by UNICORN™ 7.5 software. A second flow cell was positioned between the conductivity and pH sensors for Raman detection. Flow rates were varied (2, 3, 5, 10, and 15 mL/min), including measurements without flow for 5 g/L and 50 g/L mAb to replicate conditions encountered during different chromatography or filtration operations.

This approach ensured that the datasets reflected realistic variability, supporting the development of robust preprocessing methods for chemometric modelling. The ranges of all acquisition parameters for corresponding datasets and their impact on spectral characteristics are summarized in Table 1.

Table 1
Overview of datasets: parameter settings and usage in the study.

Component	Parameter	Range	Spectral variability	Usage
mAb	Concentration	0.1–50 g/L	Baseline, intensity, peak morphology (Fig. 1a)	Sequence investigation, Quality metric evaluation
	Exposure time, Laser power	300, 400, 495 mW 400, 500, 800 ms per spectrum at 5, 50 g/L	Intensity, baseline (Fig. 1b, c)	Sequence investigation
	Flow rate	0–15 mL/min at 5, 50 g/L	Baseline (Fig. 1d)	Sequence investigation
	Measurement time (Averaging)	8 s, 2 min	Noise level (Fig. 1e)	Sequence investigation
Glycine	Concentration, pH	0–300 mM pH 3.0–6.0	Baseline, intensity, peak morphology	Sequence investigation, Quality metric evaluation
Acetate	Concentration, pH	0–100 mM pH 4.5–7.0	Baseline, intensity, peak morphology	Sequence investigation, Quality metric evaluation
Citrate	Concentration, pH	0–100 mM pH 3.0–6.5	Baseline, intensity, peak morphology	Sequence investigation, Quality metric evaluation

2.2. Data analysis and computation

All data analysis and computation were performed in Python 3.12.

2.2.1. Spectra preprocessing

Cropping, smoothing, normalization, and baseline correction were applied to construct preprocessing pipelines, kept intentionally simple as these steps were already necessary for the investigated dataset. The primary focus of preprocessing optimization was improving baseline correction. Baseline correction was implemented in the *pybaselines (v1.1.0)* package, employing the Whittaker filter -based algorithms, commonly referred to in the literature as AsLS. Spectral cropping was applied to 700–1800 cm^{-1} for covering the fingerprint region, and normalization was performed using the sapphire peak at 752 cm^{-1} . Smoothing utilized a Whittaker smoother according to [22] with a penalty parameter of 10 and a constraint order of 2, chosen based on prior experience for optimal performance.

The ordering of preprocessing steps is a viable contributor to the overall performance. Therefore, different sequences were systematically tested through manual iterative trials and evaluated by visual inspection, as the resulting changes were clearly discernible.

2.2.2. Quality metric

To characterize baseline correction quality, Guo *et al.* proposed a metric R^{12}, which describes two ideal properties of a baseline-corrected spectrum: (i) maximal retention of intensity in regions containing Raman peaks, and (ii) minimal yet non-negative residual intensity in regions without expected peaks. If an optimal baseline correction could fulfill both criteria simultaneously, the ratio between the integrated area of non-peak regions and that of peak regions should be minimized. R^{12} quantifies and compares the area contribution of these two types of regions. The exact mathematic formulation of R^{12} is provided in [39]. In principle, a lower R^{12} value indicates a better baseline estimation.

Inspired by R^{12}, our study introduces an additional metric, C, representing concentration correlation. The aim is to extend the utility of R^{12} by incorporating correlation-based evaluation to better support quantitative regression tasks required for downstream process monitoring. In this study, C is used in combination with R^{12} to guide the selection of optimal baselines during preprocessing optimization (see Section 2.2.4).

The rationale of C is rooted in quantitative Raman analysis, where the ultimate goal is accurate prediction of analyte concentration from spectral data. This relies on the fundamental physical principle that Raman scattering intensity (i.e analyte-related band inetenisty) varies

approximately proportional to analyte concentration. When baseline correction introduces distortions such as residual curvature, artificial peaks, or negative intensities, the relationship between concentration and corrected intensity becomes nonlinear. This non-linearity can severely degrade the performance of regression models, leading to biased predictions and reduced robustness.

The metric C accounts for the linearity of the spectra by maximizing linearity in peak regions while suppressing correlation in non-peak regions and the normalization peak (at 752 cm^{-1}). Due to the complexity of water- and amide-related spectral features, the peak region (around 1640 cm^{-1}) is excluded from C calculation to preserve authenticity. To achieve this assessment, Pearson correlation coefficient r^2, indicating linearity across concentrations, was calculated at each wavenumber in the preprocessed spectra. These r^2 values were aggregated by region to form C, defined as:

$$C_p = \sum_{i=1}^{n_p} \left(w_{p,i} \cdot r_{p,i}^2 \right), \quad w_{p,i} = \frac{|I_{p,i}|}{\sum I_{p,i}}$$

$$C_n = \frac{\sum_{i=1}^{n_n} \left(\left| r_{n,i}^2 \right| \cdot \sigma_{n,i} \right)}{n_n}$$

$$C_z = \frac{\sum_{i=1}^{n_z} \left(\left| r_{z,i}^2 \right| \cdot \sigma_{z,i} \right)}{n_z}$$

$$C = \frac{(1 - C_p) + C_n + C_z}{3}$$

Here, n denotes the number of wavenumber positions, and subscripts p, n, z, i represent peak regions (excluding water), normalization regions, non-peak regions, and individual wavenumbers, respectively. C consists of three components—C_p, C_n, C_z. The first component C_p, represents weighted r^2 values in peak regions, where the weights are based on the normalized intensity of the highest concentration spectrum ($I_{p,i}$). This weighting ensures that noise from non-peak positions within peak regions does not influence the metric and only true peaks contribute. When peak features exhibit perfect linearity (r^2=1), C_p approaches its maximum value of 1. The second component, C_n, captures correlation in the normalization region, where spectra should ideally overlap completely, resulting in r^2= 0. Deviations from this, whether positive or negative correlations, indicate artefacts and consequently increase C_n. To penalize random intensity fluctuations, each r^2 value is multiplied by the standard deviation at the respective wavenumber ($\sigma_{n,i}$), ensuring that cases with near-zero correlation but high variability are properly distinguished. C_n will thereby become larger when such scenarios exist. The sum of these penalized values is then divided by n_n, the total number of wavenumber positions in the normalization region, to normalize their contribution to C and prevent bias from region size. In the current implementation, C_n is defined for normalization against a specific reference peak. Full-spectrum normalization would not provide the same concentration-invariant reference region and would require C_n to be adapted or omitted. The third component, C_z is computed using the same logic for non-peak regions. Finally, the overall metric C is computed as the average of *(1 - C_p)*, C_n and C_z, where C_p is inverted to ensure that an optimal baseline minimizes C. Thus, the final value of C should be minimal for an optimal baseline correction.

R^{12} and C serve complementary roles in evaluating baseline correction quality. R^{12} measures the spectral-shape outcome of baseline removal within each corrected spectrum, whereas C evaluates whether the corrected intensity pattern across spectra series remains chemically plausible. R^{12} detects residual background and peak retention, whereas C assesses whether peak intensity remains concentration-related and non-peak and normalization regions acquire concentration-correlated distortions. Coupling R^{12} and C provides a joint control of baseline-removal artifacts and preservation of intensity-concentration linearity. Notably, strongly nonlinear but chemically valid responses, for example those caused by detector saturation or concentration-dependent interactions can be penalized by the present metric formulation. Such systems would require an adapted association term and are outside the scope of the current study.

Peak and non-peak regions were identified using single-component dilution series to eliminate interfering features. Spectra corresponding to the highest concentrations of the investigated component were baseline-corrected using a simple, independent *asls* algorithm (not used for baseline optimization study). The hyperparameters were manually adjusted to keep the estimated baseline rigid, preventing the correction from cutting into possible peaks. The highest-concentration spectrum was deliberately selected for peak-region identification because it provided the strongest analyte signal and the highest signal-to-noise ratio within the dilution series. After smoothing and normalization (same parameter setting as before), a threshold of 0.002 was applied: intervals above were classified as peak regions, and those below as non-peak regions. In this manner, the peak regions are optimistically segmented so that even weak or broad peaks are retained, thereby minimizing the risks of incomplete coverage of analyte-related spectral features. This procedure was implemented for every buffer and mAb. The defined regions for respective components were consistently used for calculating both R^{12} and C for evaluating preprocessing candidates.

Unlike model-performance metrics, R^{12} and C evaluate the spectral characteristics of the processed spectra directly. The metric was calculated independently of any regression model and served solely as a criterion for selecting preprocessing configurations.

2.2.3. Regression modelling

After spectra were preprocessed using the selected optimal preprocessing configurations, PLS regression models were implemented to compare the impact of metric-driven or model-driven preprocessing strategies on subsequent concentration regression outcomes. The models were developed using the *scikit-learn* library (version 1.7.2) for the quantification of mAb and buffer concentration. The dataset in question was divided into training (70%) and test sets (30%) using a train-test split to ensure representative distribution of concentrations. Cross-validation was performed on the training set using k-fold cross-validation (k = 5) to optimize model parameters. The optimal number of latent variables (PLS components) was determined by minimizing the RMSECV across candidate models (up to 10 components). Model accuracy was assessed using RMSECV for the training set, and RMSEP for the test set together with coefficient of determination (R^2) for the test set. When an external validation dataset was available (for buffer datasets), root mean square error of validation (RMSEV) was calculated.

2.2.4. Automated baseline correction optimization

The optimization process was carried out using the *Optuna* framework. Three baseline correction algorithms based on Whittaker-filter were evaluated: *arpls* [25], *aspls* [24], and *derpsalsa* [42]. Each algorithm includes a regularization parameter, λ, which controls the smoothness of the estimated baseline, with higher λ values producing smoother baselines. In addition, *derpsalsa* introduces an extra parameter, k, which influences the derivative penalty term. The parameter ranges investigated for each algorithm are summarized in Table 2, and these ranges were deliberately set wide to increase the complexity of parameter optimization.

Table 2
Baseline methods and parameters in automated baseline correction screening.

Method	Parameter	Range
arpls	λ	[1e1, 1e9]
aspls	λ	[1e1, 1e9]
derpsalsa	λ,	[1e1, 1e9],
	k	[1e-4, 1e3]

To explore the parameter space, a heuristic Tree-structured Parzen Estimator (TPE) sampler was employed, with the number of trials fixed at 1000. In each trial, a baseline correction algorithm and its corresponding parameter values were selected by the sampler. The selected baseline algorithm and parameters were then applied to the spectra, while all other preprocessing steps remained unchanged. For each candidate preprocessing configuration, an objective value J was computed based on the resulting spectra. Two optimization strategies were considered. For metric-driven optimization, the objective function was defined as the sum of the mean of R^{12} and mean of C of the spectra series:

$$J = \text{mean}(R^{12}) + \text{mean}(C)$$

Because both terms are minimized (see Section 2.2.2), the best preprocessing method and parameter set was the trial with the smallest J. For model-driven optimization, the objective function was defined as the sum of RMSECV and RMSEP:

$$J = \text{RMSECV} + \text{RMSEP}$$

This objective combines cross-validation and test-set prediction errors from the PLS model, and the preprocessing configuration with the lowest J was selected. In the case of model-driven optimization, a new train-test split was performed in each trial to avoid bias introduced by a fixed split. The overall goal of the optimization algorithm was to minimize the objective function values when targeting improved preprocessing quality or model predictive performance.

3. Results and discussion

3.1. Sequence investigation

Baseline correction is a critical step in spectral data processing, yet its effectiveness can be strongly influenced by preceding transformations such as cropping, normalization and smoothing. To establish a robust sequence of preprocessing steps for subsequent parameter optimization, we first assessed the dependency of baseline correction on other preprocessing operations through systematic manual trials across various datasets. These datasets were designed to encompass a wide range of variability in spectral characteristics, including baseline slope, intensity fluctuation, peak morphology, and noise level. These spectral variabilities were stimulated by differences in sample composition and measurement parameters (see variability summarized in Table 1 and Fig. 1). By introducing controlled yet realistic shifts in these spectral features, we evaluated how the order of preprocessing steps impacts baseline correction performance.

Fig. 2 highlights the most representative outcomes, demonstrating that an optimized preprocessing order significantly improves correction accuracy and overall data quality.

The sequence analysis begins with the cropping step, which involves selecting a specific spectral range and discarding the remainder. Baseline correction algorithms aim to fit and remove slowly varying background signals. Limiting the data to the region of interest simplifies baseline estimation by eliminating problematic areas such as the steep fluorescence-dominated segment (200–400 cm^{-1}). This allows algorithms to focus on chemically relevant regions like the fingerprint domain (800–1800 cm^{-1}). As a result, baseline correction accuracy can be improved and computational burden can be reduced, which is advantageous for large datasets or real-time applications. However, cropping alters the curvature context of the spectrum, potentially affecting penalty parameters in methods such as AsLS and leading to under- or over-correction. Within the evaluated datasets, the impact of cropping placement varies across algorithms. *arpls* exhibits a region-dependent behavior. For instance, in the mAb dilution dataset shown in Fig. 2a, cropping prior to baseline correction improves the spectral linearity in the 1150–1350 cm^{-1} region, whereas it impairs linearity in the 1400–1800 cm^{-1} region (Fig. 2a). By comparison, *aspls* is more susceptible to cropping order. When cropping was performed after baseline correction, pronounced artifacts, including peak loss (around 1130 cm^{-1}), are consistently observed in multiple glycine datasets (Fig. 2b). In contrast, *derpsalsa* displays minimal dependence on cropping placement, with negligible differences observed whether cropping was applied before or after baseline correction (data not shown). Taken together, these results indicate that cropping should generally be performed before baseline correction to minimize artifacts within the tested algorithms.

Normalization is intended to remove intensity differences that arise from measurement-related factors rather than true chemical variation. These factors can include changes in laser power, exposure time, detector sensitivity. By rescaling spectra to a common intensity range or reference value, normalization facilitates direct comparability across measurements. In principle, after effective normalization, spectra originating from the same sample but collected under different measurement conditions should overlap closely, allowing downstream analyses to concentrate on meaningful spectral shape differences rather than artificial intensity shifts. In this study, spectra were normalized to the intensity of sapphire peak at 752 cm^{-1} . As sapphire peak originates from the probe, its signal is expected to remain constant under a defined measurement condition regardless of sample composition. In practice, the normalization strategy depends on the analytical objective, instrument configuration, and source of multiplicative variability. Across the investigated datasets and baseline correction algorithms, the order of sapphire normalization has little influence on the results, with spectra remaining largely comparable regardless of whether normalization preceded or followed baseline correction. As illustrated in Fig. 2c—where exposure time and laser power were varied for mAb samples at a fixed concentration of 50 g/L—the preprocessed spectra are similar across most spectral regions, with deviations primarily occurring in the 1400–1800 cm^{-1} range. Nevertheless, altering the position of the normalization step within the preprocessing sequence produces nearly identical spectra in both the left and right panels of Fig. 2c, indicating that normalization order does not materially affect the final spectral outcome. This observation suggests that amplitude offsets do not compromise baseline estimation, as the tested algorithms primarily model low-frequency trends rather than absolute intensity. An exception occurred in the mAb dataset with varying flow rates, where normalization before baseline correction distorts the baseline shape, altering peak prominence relative to the background and consequently reducing the comparability among corrected spectra (Figure S1).

Smoothing is commonly employed in Raman spectral preprocessing to reduce high-frequency noise and improve signal clarity. However, as Fig. 2d points out, when smoothing is applied to spectra that are already relatively smooth, it can introduce unwanted variance that destabilizes the subsequent baseline correction. The spectra in Fig. 2d were acquired with long integration time (2 min) and are inherently smooth. In this case, additional smoothing causes the low-concentration spectra (< 1 g/L) to demonstrate inadequately compensated baselines in the 1250–1450 cm^{-1} region (right panel), which is not noticeable in the corresponding pipeline without extra smoothing beforehand (left panel). This effect likely stems from subtle changes in signal-to-noise ratio introduced by smoothing, which can modify the perceived local curvature and influence peak interpretation during baseline estimation. The phenomenon is evident for *arpls* (Fig. 2d) and *aspls* (data not shown), whereas *derpsalsa* remains insensitive to smoothing order (data not shown). Conversely, in noisy datasets, the presence of noise markedly undermines baseline correction consistency, particularly for *arpls* (Fig. 2d). In such cases, applying smoothing prior to baseline correction supports more reliable baseline correction. The spectra in Fig. 2e were acquired with short integration time (8 s) and are inherently noisy. Without preliminary smoothing (left panel), *arpls* fails to converge to coherent baseline estimates across the dilution series. This manifests as nonlinear intensity-concentration responses which are most noticeably

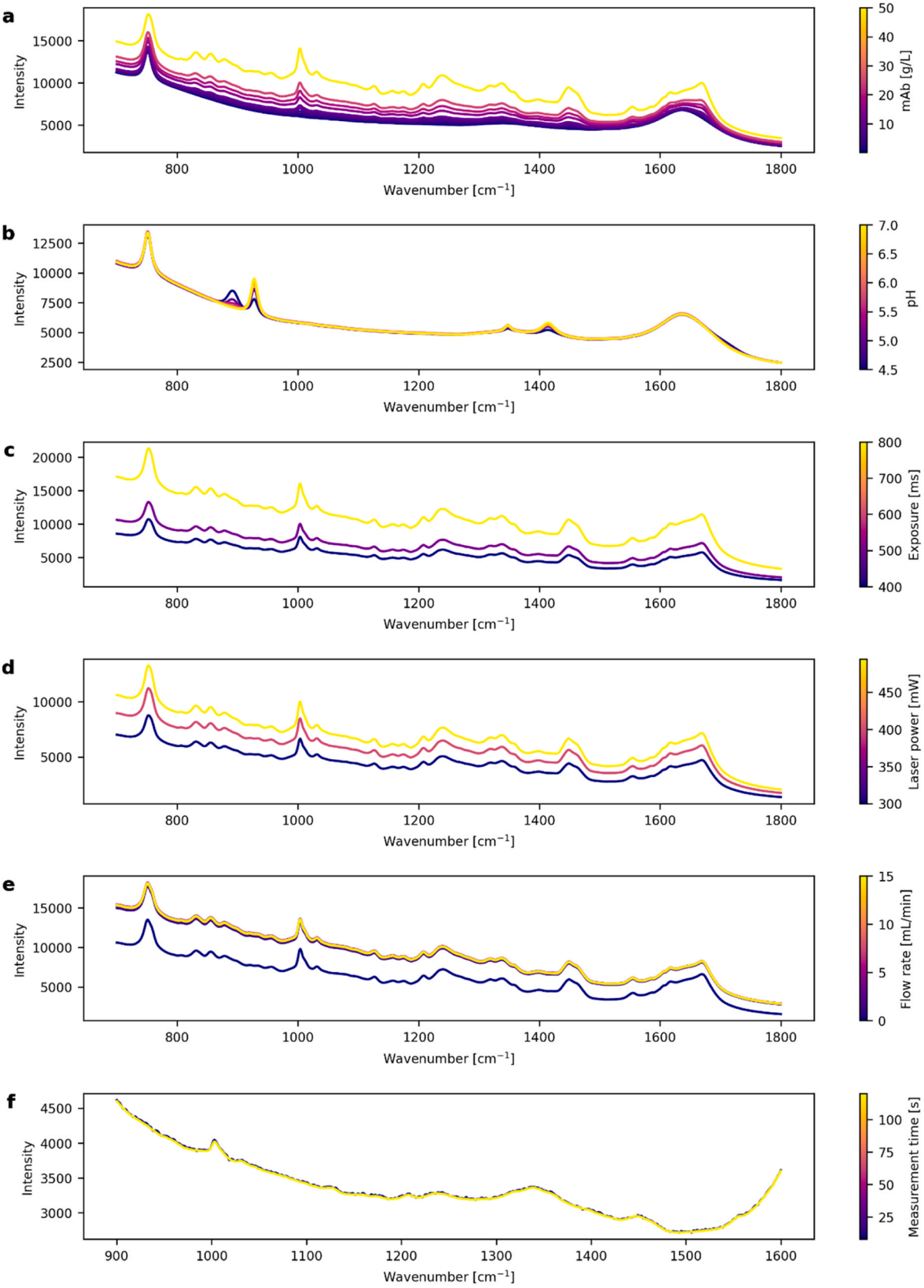

Fig. 1. Spectral variability observed under different experimental conditions:a. concentration variation of mAb; b. pH variation at 100 mM acetate; c. exposure variation at 50 g/L mAb; d. laser power variation at 50 g/L mAb; e. flow rate variation at 50 g/L mAb; f. measurement time (averaging) variation at 5 g/L mAb, 500 ms exposure time, 400 mW laser power.

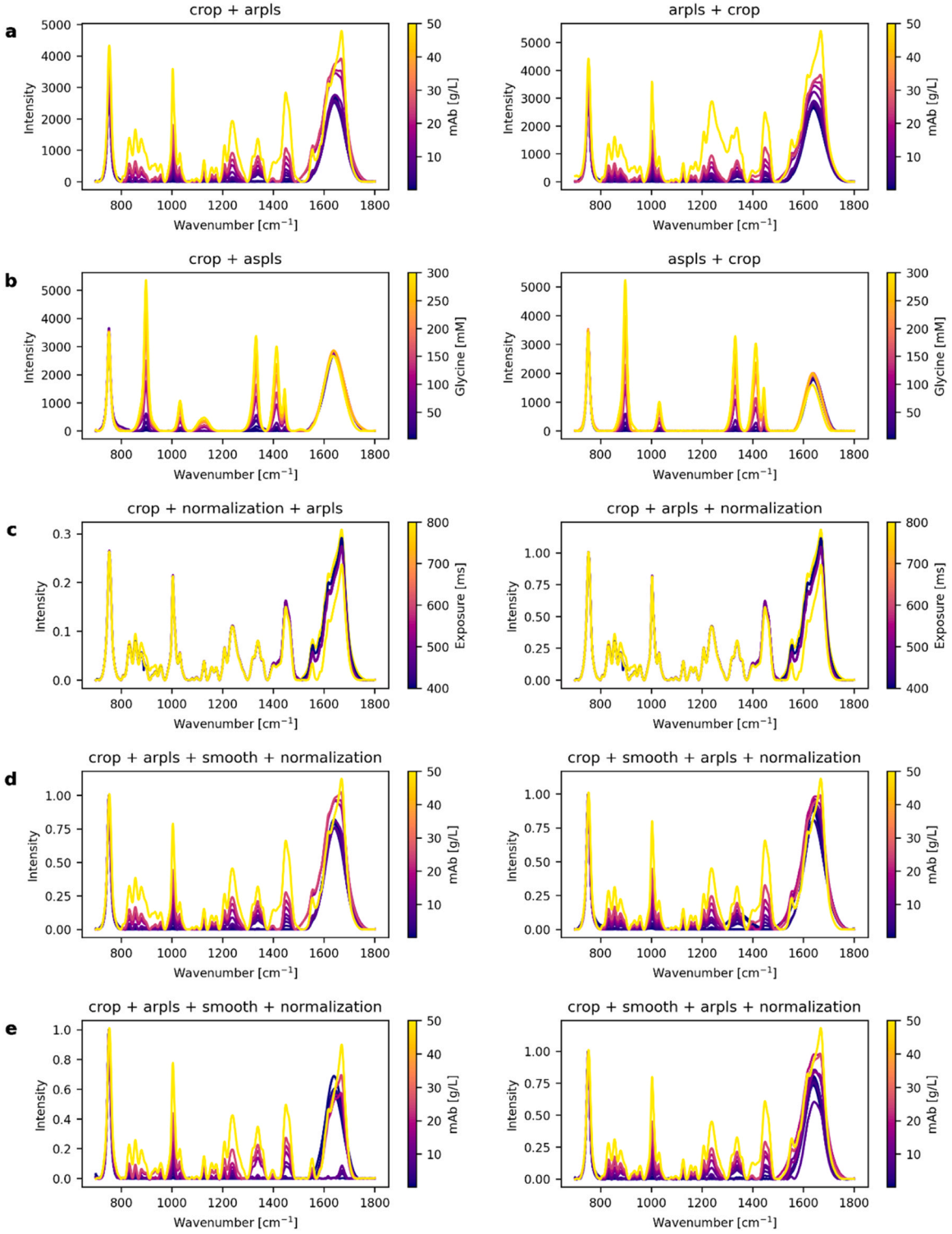

Fig. 2. Exemplary effects of preprocessing step sequence on baseline correction performance. a. cropping before (left) or after (right) baseline correction for dataset of mAb dilution series with a total measurement time of 2 min; b. cropping before (left) or after (right) baseline correction for dataset of glycine dilution series with a total measurement time of 2 min; c. normalization before (left) or after (right) baseline correction for dataset of mAb at 50 g/L with varying exposure time and laser power and with a total measurement time of 2 min; d. smoothing after (left) or before (right) baseline correction for dataset of mAb dilution series with a total measurement time of 2 min; e. smoothing after (left) or before (right) baseline correction for dataset of mAb dilution series with a total measurement time of 8 s.

in the 1300–1800 cm^{-1} region. In comparison, smoothing before baseline correction (right panel) largely restores the performance, yielding preprocessed spectra that approach the quality achieved for inherently smooth spectra shown in Fig. 2d (left panel). Both *aspls* and *derpsalsa* demonstrate greater resilience to noise, exhibiting minimal degradation in correction quality even without prior smoothing (data not shown).

Beyond the interplay of preprocessing steps, a more critical observation is the sensitivity of baseline correction under varying experimental conditions. Notably, even under ideal, controlled acquisition conditions, a mere shift in concentration range—the primary variable of

interest in regression analysis—can lead to markedly different correction outcomes. A certain algorithm and parameter value that performs well for weak signals at low concentrations often underfits strong peaks at high concentrations, while aggressive settings suited for high-intensity spectra at high concentrations risk erasing subtle features at low concentrations. Artifacts like peak distortion or residual baseline arise from correction over a broad concentration range, generating nonlinear spectral behaviors. Hence, a systematic evaluation of baseline correction performance in DSP environments is necessary, particularly when experimental parameters vary widely, such as in ultrafiltration/diafiltration (UF/DF) processes. Baseline correction settings must be robust and adaptive to accurately reject the background signal across diverse signal regimes. Such adaptability is essential for ensuring data integrity and preserving predictive accuracy in chemometric models.

Drawing on the findings from this analysis, the preprocessing sequence of cropping-baseline correction-smoothing-normalization shows robustness under variability perturbation in low-noise spectra. This sequence is further applied in the next section to select baseline correction algorithms and refine their parameters.

3.2. Quality metric validation

The instability of baseline correction algorithms to variations in signal characteristics underscores the importance of comprehensive quality checks across the full experimental range, as artifacts derived from improper baseline correction can create nonlinear signals and propagate through chemometric workflows. Therefore, we introduce a quantitative metric C, used in conjunction with the established R^{12} metric (hereafter collectively referred to as the quality metric), for judging preprocessing quality. By incorporating criteria such as the peak-to-baseline area ratio and peak linearity, the metric minimizes residual baseline, enhances the true chemical signals, as well as avoids artifact formation. It offers a more direct and realistic characterization of preprocessing performance than relying on the model error rates alone. The quality metric is embedded in an automated baseline correction optimization workflow, enabling efficient identification of optimal configurations. In the following sections, we validate the functionality of the proposed quality metric using two exemplary datasets selected from the four tested (Table 1) and benchmark its effectiveness against a model-driven optimization approach. Results for the remaining datasets are provided in the Supplementary Information (Figure S3-Figure S6, Table S1-Table S2).

3.2.1. mAb dataset

Preprocessing results of the mAb dilution series with different optimization approaches are presented in Fig. 3. In Fig. 3b-c, spectra preprocessed by a top-ranked, medium-ranked, worst-ranked preprocessing configuration according to the quality metric approach are demonstrated respectively. As anticipated, baseline correction outcome improves as the metric value decreases. The lowest metric value (Fig. 3b) corresponds to spectra with comparably best preprocessing quality—characterized by flat baselines, complete background subtraction, and preserved peak fidelity. Peaks appear visually clean and well-defined, with improved intensity–concentration linearity, except for the broad water and protein amide I band (1550–1750 cm^{-1}), which is deliberately not considered in the metric calculation (Section 2.2.2). Interestingly, a medium metric value (Fig. 3c) still delivers a comparable correction. While the result is generally satisfactory, minor artifacts are present in the 1300–1400 cm^{-1} region. This implies the sensitivity of the metric to subtle distortions. In contrast, the highest metric value (Fig. 3d), corresponding to the worst-ranked results, features significant baseline slope and correction inconsistency across concentrations. The strong background intensity masks the mAb peak signatures. These observations confirm that the metric effectively ranks preprocessing quality as intended. To evaluate the specific contribution of the C term to preprocessing selection, Fig. 3a displays spectra optimized using R^{12} alone. Remarkably, optimization by R^{12} (without C) produces artifacts such as negative intensities at several spectral positions and inconsistent baseline estimates across concentrations.

In addition to visual assessment, numeric readouts of spectral linearity at selected feature peaks are summarized in Table 3. Spectral linearity was quantified using the Pearson correlation coefficient r^2 between peak maximum intensity and concentration. The peaks were selected to distribute across the spectral range to provide representative coverage. Notably, the combined-metric optimization by $R^{12} + C$ improves r^2 from 0.8683 to 0.9969 at 1157 cm^{-1} and from 0.9356 to 0.9744 at 1338 cm^{-1} compared with R^{12}-only selection (Table 3). At 1004 cm^{-1} and 1449 cm^{-1} , all strategies retain high r^2 values (all above 0.99), indicating that the advantage of C is concentrated in spectral regions susceptible to distortion rather than uniformly distributed. Thus, augmenting R^{12} with metric C can further enhance the correction quality.

So far, the metric has demonstrated strong performance at the spectral level. Next, we examine whether optimal baseline correction—defined by the minimal metric value—also translates into optimal statistical results in regression modelling. To address this question, a PLS model was constructed to quantify mAb concentrations based on the preprocessed spectra after optimized baseline correction in Fig. 3b. In parallel, a second PLS model was built using preprocessed spectra where the baseline correction configuration was optimized to minimize RMSECV and RMSEP, namely the model-driven optimization. The predictive performance of the two models is compared in Fig. 4. As illustrated in Fig. 4, both approaches provide relatively accurate predictions across the wide concentration range (up to 50 g/L). However, the metric-driven optimization results in higher RMSECV and RMSEP values of 0.8883 and 0.3866, respectively, and a lower R^2 of 0.9971, compared with the model-driven approach, which achieves RMSECV and RMSEP values of 0.1557 and 0.0867, respectively, and an R^2 of 0.9999. Meanwhile, Table 3 shows that the metric-driven pipeline preserves high peak-wise linearity across the four reported bands, with r^2 values ranging from 0.9744 to 0.9982, comparable to those obtained with the model-driven pipeline. The results suggest the metric-driven solution may be limited in calibration prediction accuracy in this dataset while still maintaining the targeted spectral quality criteria.

Despite the superior prediction accuracy, the model-driven approach achieves this performance through fundamentally different spectral characteristics, as revealed by Fig. 3e. While spectra optimized for model performance appear well aligned across most Raman regions, a considerable amount of baselines remain after preprocessing. Fig. 1a indicates that the background signals of raw spectra in the current dataset already exhibit some degrees of protein concentration dependence, likely attributable to sample autofluorescence [13] and the relatively simple sample composition of the dataset. Consequently, the model-driven optimization approach may leverage these baseline trends and favor preprocessing configurations that retain portions of the baseline signals, because this can benefit prediction performance within the current dataset. However, modelling concentration with baseline information can introduce significant drawbacks. Baseline offsets are not analyte-specific and can be influenced by buffer composition and external factors such as measurement setup and instrument drift [17,18,35], which are independent of protein concentration. This means, the baseline may rise or fall without any real change in protein level. Raman spectra with residual baseline variations bear the risk that chemometric models may incorporate this information during the calibration process and misinterpret changes caused by non-protein factors as variations in analyte concentrations. While this relationship can improve model performance within the calibration dataset, it may not persist under different operating conditions. Although not directly demonstrated in this experiment, it can be deduced that models that rely partly on baseline variation may exhibit reduced robustness when transferred to new instruments, measurement setups, or sample matrices. In contrast, the quality metric-driven strategy focuses on removing baseline

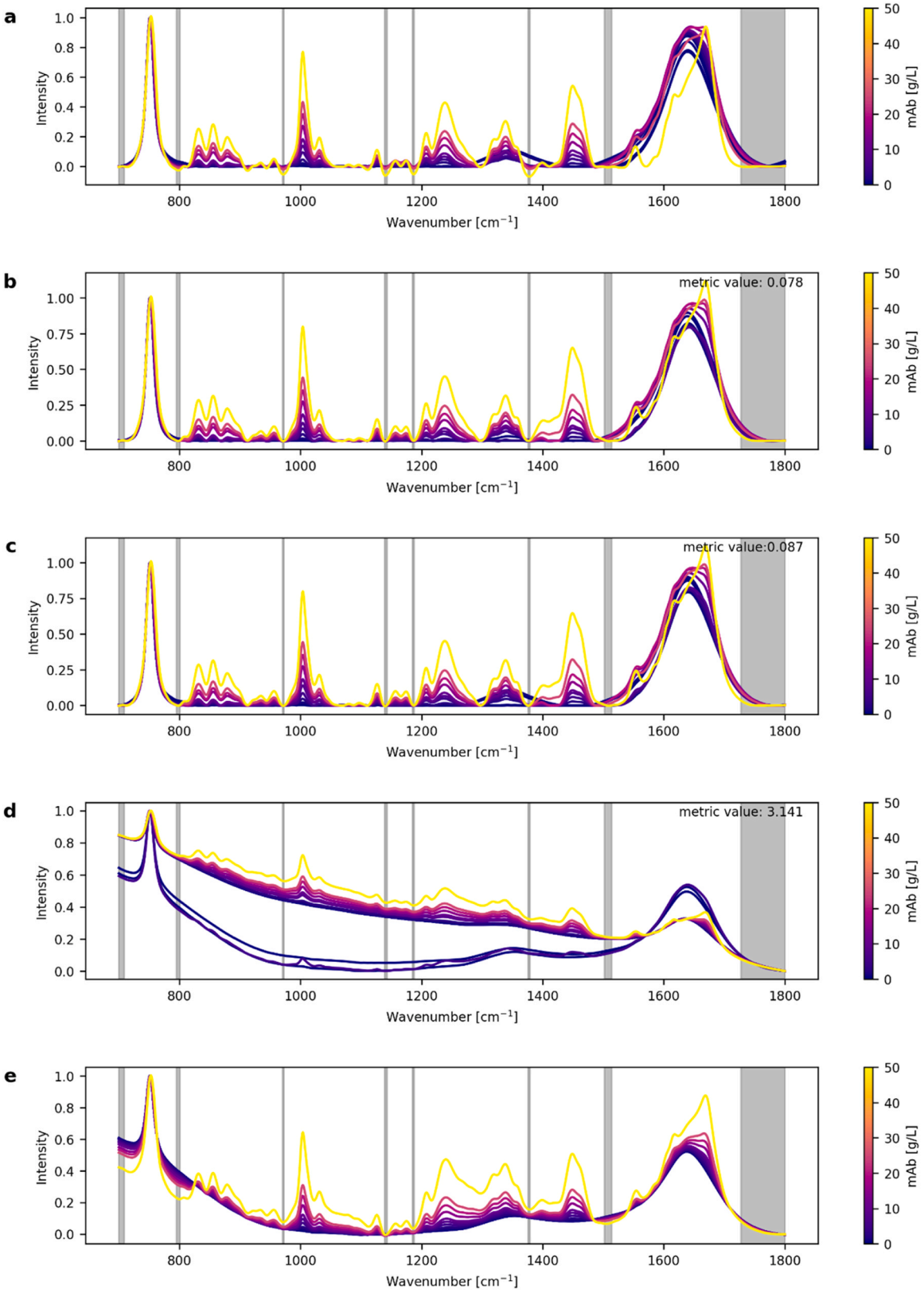

Fig. 3. Preprocessed spectra of the mAb dilution dataset using different baseline correction optimization approaches. a. optimal preprocessing result selected via R^{12}; (b-d. optimal, medium, worst preprocessing results selected via the combined R^{12} & C metric; e. optimal preprocessing result selected via RMSECV& RMSEP. Grey shaded areas indicate non-peak regions, unshaded areas represent peak regions.

Table 3
Comparison of spectral linearity at selected wavenumbers for the mAb dilution series. The Pearson correlation coefficient r^2 was used to measure spectral linearity. Values were computed for both raw spectra and spectra preprocessed using optimal configurations derived from different preprocessing optimization approaches.

	1004 cm^{-1}	1157 cm^{-1}	1338 cm^{-1}	1449 cm^{-1}
Optimal preprocessing selected via R^{12}	0.9974	0.8683	0.9356	0.9994
Optimal preprocessing selected via R^{12} & C	0.9982	0.9969	0.9744	0.9977
Optimal preprocessing selected via RMSECV & RMSEP	0.9983	0.9712	0.9983	0.9987

contributions and isolating chemically specific Raman features before any modelling steps. Although it results in lower predictive accuracy for the present dataset, it may offer advantages in future applications involving more complex datasets, as it provides spectra that are more representative of analyte-specific information.

3.2.2. Acetate dataset

When applied to a buffer system, the benefits of the quality-based preprocessing approach are proven again. Different from the mAb dataset, the acetate dilution series was replicated across multiple pH values (pH 4.5–7). The protonation state of acetate varies in this pH range, altering vibrational modes and inducing noticeable changes in spectral peak representation. For example, the ratio of peaks at 890 cm^{-1} and 930 cm^{-1} varies with increasing pH (Fig. 1b). A PLS model was trained to predict acetate concentrations regardless of pH, by including spectra from multiple pH conditions in the training set. The pH 6.5 data were left out and used as external validation dataset to examine model robustness and generalization.

In Fig. 5, preprocessed spectra obtained using different preprocessing strategies are displayed. Similar to previous findings, the metric effectively rates the preprocessing performance, with lower metric values linked to visually improved baseline removal, including clearer peak separation and more consistent baseline estimates (Fig. 5b-d). Compared to spectra optimized solely by R^{12}, the adopted $R^{12} + C$ metric produces better-resolved peaks with fewer artifacts, particularly in the regions of 780–900 cm^{-1} and 1270–1500 cm^{-1} (Fig. 5a-b). Table 4 complements the visual comparison by quantifying peak-wise linearity across selected acetate features. Compared with R^{12} alone, $R^{12} + C$ increases mean r^2 from 0.9896 to 0.9901 at 928 cm^{-1} and from 0.9804 to 0.9896 at 1347 cm^{-1} , while the change at 1415 cm^{-1} is small (0.9884–0.9887). The model-driven result is highest at 928 cm^{-1} (0.9925), comparable at 1347 cm^{-1} (0.9890), and lower at 1415 cm^{-1} (0.9748). Accordingly, no single strategy dominates at every band. Overall, however, the combined metric consistently preserves relatively high linearity across the reported acetate features.

Interestingly, the model-driven optimized spectra demonstrate a similar visual appearance (Fig. 5e) and peak linearity (Table 4) to the metric-driven approach regarding the acetate bands, but the water peak between 1550 and1750 cm^{-1} is largely omitted in the preprocessed spectra. One possible explanation is that the model-driven optimization selects the baseline-correction parameter λ according to its ability to maximize prediction of acetate concentration. Under this condition, the optimizer could ultimately arrive at a λ value that made the baseline overly flexible to follow the curvature of broad spectral features. Consequently, the broad water peak—which shows negligible variance with acetate concentration across the training set—is treated as part of the baseline and eliminated.

Fig. 6 compares the predictive performance of the models generated using the two preprocessing strategies. During model calibration (Fig. 6, left), the model-driven approach shows higher prediction accuracy relative to the metric-driven approach, reducing RMSECV from 5.2505 to 3.7478 and RMSEP from 4.6399 to 3.1596. The R^2 value also increases from 0.9780 to 0.9898. These results suggest that the model-driven approach yields improved prediction within the model development dataset, although the prediction errors of both methods remain low relative to the overall concentration range (up to 100 mM). However, this improvement is accompanied by a loss of spectral integrity because a genuine chemical feature, rather than only the baseline, is removed. That means, in spite of a higher accuracy, the baseline correction is from a spectroscopic point of view suboptimal. The water peak, or peaks from other matrix components that may be present in more complex samples, should ideally remain during baseline correction procedure, even if they do not directly enhance analyte prediction. Any exclusion of such features should be treated as an explicit variable-selection decision, rather than occurring implicitly as a consequence of the baseline correction.

In this dataset specifically, the removal of water peak also introduces transferability challenges. While the water peak may not be directly predictive of acetate concentration, it encodes information related to the surrounding chemical environment. Specifically, water signals can reflect changes in hydrogen bonding and solvent structure, which vary with matrix properties such as ionic strength and pH [43]. As depicted in Figure S2, the water peak exhibits a distinct shoulder located at the 1700–1750 cm^{-1} region depending on pH. Removing this region eliminates potentially informative variation related to the sample's chemical context and may therefore reduce the model's ability to accommodate changes in matrix conditions. This interpretation is supported by the external validation at pH 6.5. As shown in Fig. 6, when deploying the

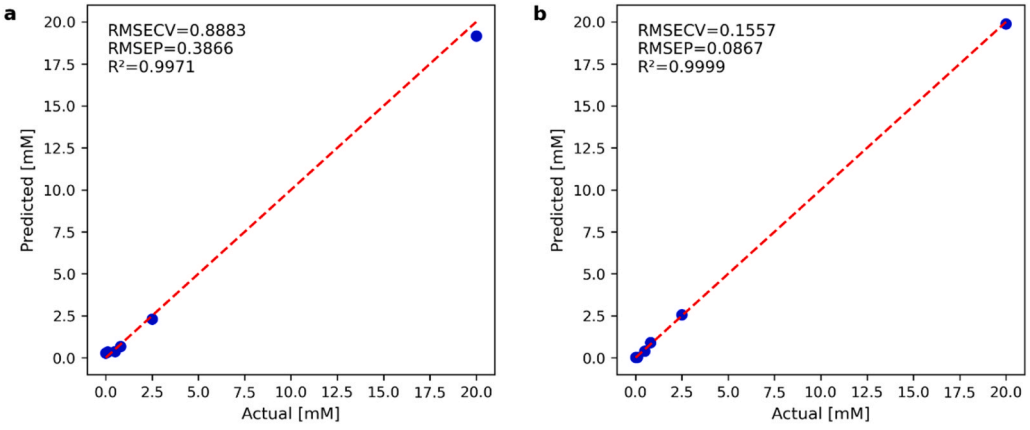

Fig. 4. Comparison of PLS model performance based on the preprocessed spectra of the mAb dilution dataset using a. quality metric-based approach; b. model error-based approach.

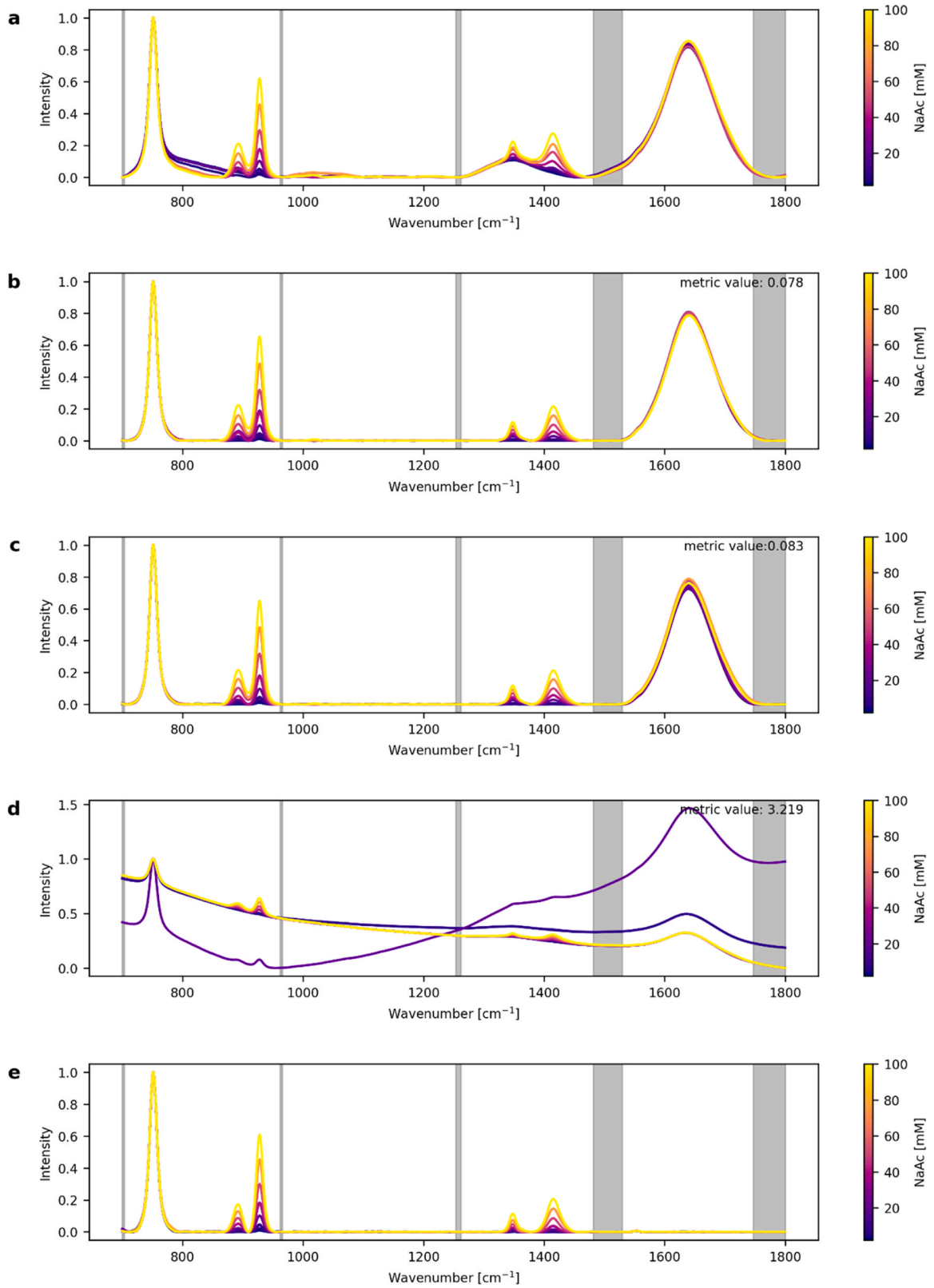

Fig. 5. Preprocessed spectra of the acetate dilution dataset at pH 5.0 using different baseline correction optimization approaches. a. optimal preprocessing result selected via R^{12}; b-d. optimal, medium, worst preprocessing results selected via the combined R^{12} & C metric; e. optimal preprocessing result selected via RMSECV& RMSEP. Grey shaded areas indicate non-peak regions, unshaded areas represent peak regions.

Table 4
Comparison of spectral linearity at selected wavenumbers for the acetate dilution series. The mean of Pearson correlation coefficient r^2 at all pHs except 6.5 was used to measure spectral linearity. Values were computed for both raw spectra and spectra preprocessed using optimal configurations derived from different preprocessing optimization approaches.

	928 cm^{-1}	1347 cm^{-1}	1415 cm^{-1}
Optimal preprocessing selected via R^{12}	0.9896	0.9804	0.9884
Optimal preprocessing selected via R^{12} & C	0.9901	0.9896	0.9887
Optimal preprocessing selected via RMSECV & RMSEP	0.9925	0.9890	0.9748

model on spectra from an unseen pH, RMSEV and R^2 worsen for both preprocessing approaches compared to their respective calibration set. However, the deterioration is more substantial for the model-driven approach. Relative to RMSEP in calibration, RMSEV increases by 21.5% for the metric-driven pipeline but by 85.4% for the model-driven pipeline. In the external validation, metric-driven pipeline consequently achieves slightly better performance, with an RMSEV of 5.6386 compared with 5.8570 and an R^2 of 0.9665 compared with 0.9638. Although these results do not establish water-band preservation as the sole cause of improved generalization, they indicate that the model-driven pipeline learns relationships that were less stable under the changed pH condition. In contrast, the metric-driven strategy better maintains chemically meaningful spectral features and provides greater robustness to the matrix variation evaluated in this study.

4. Conclusion

This study presents a quality-guided methodology for constructing robust and interpretable preprocessing pipelines tailored to quantitative Raman PAT applications in downstream mAb purification.

Given the high sensitivity of baseline correction to process variability, both its placement within the preprocessing pipeline and the selection of algorithmic parameters are critical. Using a spectra library that captures realistic variability in process parameters and measurement conditions for both mAb and buffers systems, we first systemically revised the order interdependencies among preprocessing steps. Within the tested datasets and AsLS-family methods, cropping and smoothing emerged as the most impactful steps affecting baseline correction consistency. A sequence of cropping, baseline correction, smoothing, and sapphire-peak normalization was identified to consistently stabilize baseline estimation for low-noise spectra. Establishing this sequence provided a robust starting point for the subsequent baseline correction optimization, eliminated unfeasible combinations and avoided exhaustive searches. Nevertheless, this order should be regarded as conditional. Smoothing before baseline correction may be preferable for noisy spectra, and cropping must retain sufficient and continuous baseline context. Normalization before baseline correction might be the safer default when scaling affects baseline variation or when normalization methods other than peak normalization are used. The appropriate order of preprocessing should be verified for individual applications and datasets.

The introduced metric $R^{12} + C$ combines artifacts assessment and concentration-intensity linearity evaluation. Validation across mAb and buffer datasets proved that this metric was able to reliably filter out suboptimal configurations and guide the automated optimization towards preprocessing results with minimized residual artifacts and preserved peak integrity, thereby ensuring analyte-specific spectral behaviors. For benchmarking, metric-driven optimization was compared with the conventional model-error-driven optimization at both spectral and model outcome levels. Our results revealed a critical limitation of model-driven optimization. While this approach occasionally achieved higher calibration accuracy, the improvement in prediction error was accompanied by background retention or removal of chemically relevant spectral features. This observation underscores

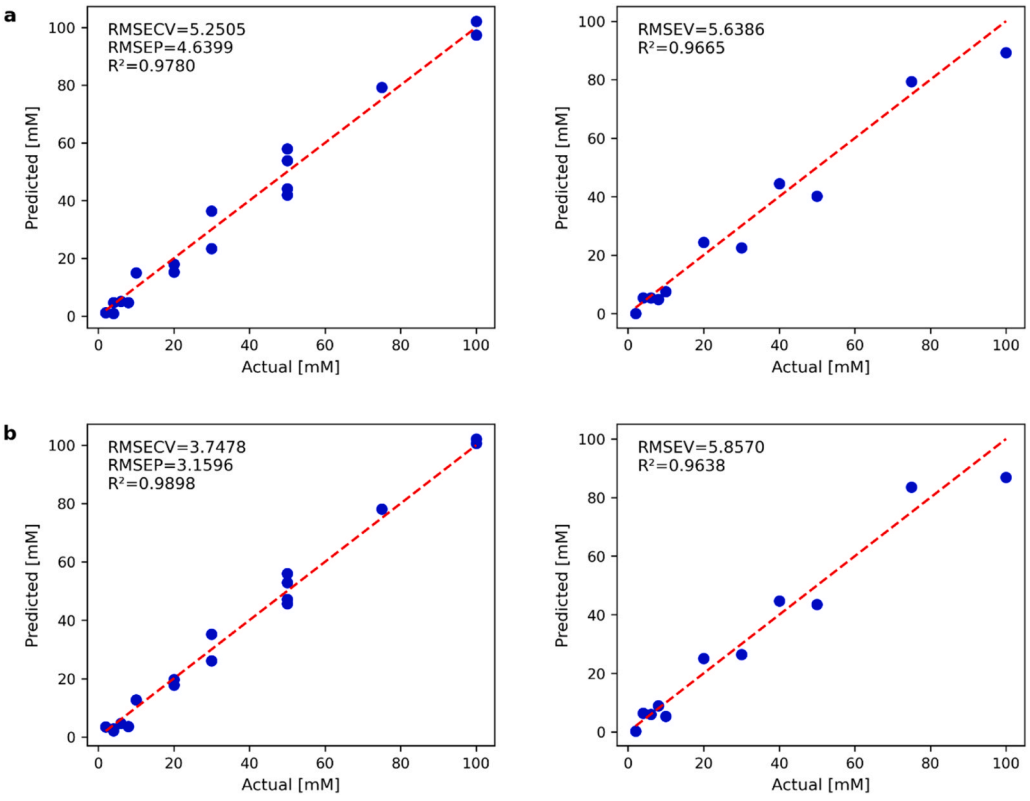

Fig. 6. Comparison of PLS model performance based on the preprocessed spectra of the acetate dilution dataset using a. quality metric- based approach; b. model error-based approach. Left: model training results at pH 4.5–7.0 except pH 6.5; Right: model validation at pH 6.5.

the value of incorporating a spectral quality metric into preprocessing decisions. In comparison, the metric-driven approach demonstrated better preservation of chemically related spectral information and showed improved robustness under changing sample conditions. Additional advantages include independence from modeling choice and applicability prior to model construction.

We acknowledge that single-component datasets represent relatively simple scenarios. They allow for a controlled evaluation of preprocessing but may not reflect the complexity of real bioprocess streams involving mixture composition. Nevertheless, the workflow establishes a solid foundation that can be extended for more challenging preprocessing tasks in mixture systems. Future work could also evaluate broader applicability to other baseline correction algorithm families. In principle, the same correction objective applies, but extension requires empirical validation. Overall, this study demonstrates that the quality metric serves as a viable alternative for directing preprocessing design. By balancing statistical criteria with chemical principles, the proposed approach supports the development of reliable and scalable Raman-based PAT, helping to realize real-time monitoring and adaptive control in dynamic downstream bioprocessing environments.

CRediT authorship contribution statement

Yue Guo: Writing – original draft. **Robin Schiemer:** Supervision. **Gang Wang:** Supervision. **Joey Studts:** Project administration. **Jürgen Hubbuch:** Project administration.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Yue Guo 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 per-sonal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.vibspec.2026.103960.

Data availability

The data that has been used is confidential.

References

[1] M. Kesik-Brodacka, Progress in biopharmaceutical development, Biotechnol. Appl. Biochem. 65 (3) (2018) 306–322, https://doi.org/10.1002/bab.1617.

[2] G. Walsh, E. Walsh, Biopharmaceutical benchmarks 2022, Nat. Biotechnol. 40 (12) (2022) 1722–1760, https://doi.org/10.1038/s41587-022-01582-x.

[3] Administration F and D. PAT- A Framework for Innovative Pharmaceutical Development, Manufacturing and Quality Assurance. Published online 2004.

[4] P. Sathiyapriyan, S. Mukherjee, T. Vogel, et al., Current PAT landscape in the downstream processing of biopharmaceuticals, Anal. Sci. Adv. 6 (1) (2025) e70013, https://doi.org/10.1002/ansa.70013.

[5] K.A. Esmonde-White, M. Cuellar, I.R. Lewis, The role of Raman spectroscopy in biopharmaceuticals from development to manufacturing, Anal. Bioanal. Chem. 414 (2) (2022) 969–991, https://doi.org/10.1007/s00216-021-03727-4.

[6] 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. 121 (5) (2024) 1729–1738, https://doi.org/10.1002/bit.28679.

[7] B. Wei, N. Woon, L. Dai, et al., Multi-attribute Raman spectroscopy (MARS) for monitoring product quality attributes in formulated monoclonal antibody therapeutics, mAbs 14 (1) (2022) 2007564, https://doi.org/10.1080/19420862.2021.2007564.

[8] 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 1718 (2024) 464721, https://doi.org/10.1016/j.chroma.2024.464721.

[9] N. Nitika, B. Keerthiveena, G. Thakur, A.S. Rathore, Convolutional neural networks guided raman spectroscopy as a process analytical technology (PAT) tool for monitoring and simultaneous prediction of monoclonal antibody charge variants, Pharm. Res. 41 (3) (2024) 463–479, https://doi.org/10.1007/s11095-024-03663-9.

[10] L. Rolinger, J. Hubbuch, M. Rüdt, Monitoring of ultra- and diafiltration processes by Kalman-filtered Raman measurements, Anal. Bioanal. Chem. 415 (5) (2023) 841–854, https://doi.org/10.1007/s00216-022-04477-7.

[11] N. Kuhar, S. Sil, T. Verma, S. Umapathy, Challenges in application of Raman spectroscopy to biology and materials, RSC Adv. 8 (46) (2018) 25888–25908, https://doi.org/10.1039/c8ra04491k.

[12] K. Buckley, A.G. Ryder, Applications of Raman spectroscopy in biopharmaceutical manufacturing: a short review, Appl. Spectrosc. 71 (6) (2017) 1085–1116, https://doi.org/10.1177/0003702817703270.

[13] C.A. Lieber, A. Mahadevan-Jansen, Automated method for subtraction of fluorescence from biological Raman spectra, Appl. Spectrosc. 57 (11) (2003) 1363–1367, https://doi.org/10.1366/000370203322554518.

[14] P.A. Mosier-Boss, S.H. Lieberman, R. Newbery, Fluorescence rejection in Raman spectroscopy by shifted-spectra, edge detection, and FFT filtering techniques, Appl. Spectrosc. 49 (5) (1995) 630–638, https://doi.org/10.1366/0003702953964039.

[15] N. Pavurala, C.N. Madhavarao, J. Lee, J. Das, M. Ashraf, T. O'Connor, Cell culture media and Raman spectra preprocessing procedures impact glucose chemometrics, J. Chemom. 39 (2) (2025), https://doi.org/10.1002/cem.70005.

[16] R. teja Vulchi, V. Morgunov, R. Junjuri, T. Bocklitz, Artifacts and anomalies in raman spectroscopy: a review on origins and correction procedures, Molecules 29 (19) (2024) 4748, https://doi.org/10.3390/molecules29194748.

[17] S. Guo, C. Beleites, U. Neugebauer, et al., Comparability of Raman spectroscopic configurations: a large scale cross-laboratory study, Anal. Chem. 92 (24) (2020) 15745–15756, https://doi.org/10.1021/acs.analchem.0c02696.

[18] J. Wang, J. Chen, J. Studts, G. Wang, In-line product quality monitoring during biopharmaceutical manufacturing using computational Raman spectroscopy, mAbs 15 (1) (2023) 2220149, https://doi.org/10.1080/19420862.2023.2220149.

[19] S. Guo, J. Popp, T. Bocklitz, Chemometric analysis in Raman spectroscopy from experimental design to machine learning–based modeling, Nat. Protoc. 16 (12) (2021) 5426–5459, https://doi.org/10.1038/s41596-021-00620-3.

[20] J. Engel, J. Gerretzen, E. Szymańska, et al., Breaking with trends in pre-processing? TrAC. Trends Anal. Chem. 50 (2013) 96–106, https://doi.org/10.1016/j.trac.2013.04.015.

[21] T. Bocklitz, A. Walter, K. Hartmann, P. Rösch, J. Popp, How to pre-process Raman spectra for reliable and stable models? Anal. Chim. Acta 704 (1-2) (2011) 47–56, https://doi.org/10.1016/j.aca.2011.06.043.

[22] P.H.C. Eilers, A Perfect Smoother, Anal. Chem. 75 (14) (2003) 3631–3636, https://doi.org/10.1021/ac034173t.

[23] C.D. Kappatou, J. Odgers, S. García-Muñoz, R. Misener, An optimization approach coupling preprocessing with model regression for enhanced chemometrics, Ind. Eng. Chem. Res. 62 (15) (2023) 6196–6213, https://doi.org/10.1021/acs.iecr.2c04583.

[24] F. Zhang, X. Tang, A. Tong, et al., Baseline correction for infrared spectra using adaptive smoothness parameter penalized least squares method, Spectrosc. Lett. 53 (3) (2020) 222–233, https://doi.org/10.1080/00387010.2020.1730908.

[25] S.J. Baek, A. Park, Y.J. Ahn, J. Choo, Baseline correction using asymmetrically reweighted penalized least squares smoothing, Analyst 140 (1) (2014) 250–257, https://doi.org/10.1039/c4an01061b.

[26] Z.M. Zhang, S. Chen, Y.Z. Liang, Baseline correction using adaptive iteratively reweighted penalized least squares, Analyst 135 (5) (2010) 1138–1146, https://doi.org/10.1039/b922045c.

[27] S. He, W. Zhang, L. Liu, et al., Baseline correction for Raman spectra using an improved asymmetric least squares method, Anal. Methods 6 (12) (2014) 4402–4407, https://doi.org/10.1039/c4ay00068d.

[28] R. Gautam, S. Vanga, F. Ariese, S. Umapathy, Review of multidimensional data processing approaches for Raman and infrared spectroscopy, EPJ Tech. Instrum. 2 (1) (2015) 8, https://doi.org/10.1140/epjti/s40485-015-0018-6.

[29] T. Chen, Y. Son, A. Park, S.J. Baek, Baseline correction using a deep-learning model combining ResNet and UNet, Analyst 147 (19) (2022) 4285–4292, https://doi.org/10.1039/d2an00868h.

[30] T. Chen, Y. Son, C. Dong, S.J. Baek, Baseline correction of Raman spectral data using triangular deep convolutional networks, Analyst 150 (12) (2025) 2653–2660, https://doi.org/10.1039/d5an00253b.

[31] Y. Hu, T. Jiang, A. Shen, W. Li, X. Wang, J. Hu, A background elimination method based on wavelet transform for Raman spectra, Chemom. Intell. Lab. Syst. 85 (1) (2007) 94–101, https://doi.org/10.1016/j.chemolab.2006.05.004.

[32] X. Ning, I.W. Selesnick, L. Duval, Chromatogram baseline estimation and denoising using sparsity (BEADS), Chemom. Intell. Lab. Syst. 139 (2014) 156–167, https://doi.org/10.1016/j.chemolab.2014.09.014.

[33] Y. Guo, W. Jin, W. Wang, Y. He, S. Qiu, Baseline correction for Raman spectra using a spectral estimation-based asymmetrically reweighted penalized least squares method, Appl. Opt. 62 (18) (2023) 4766, https://doi.org/10.1364/ao.489478.

[34] 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 (11) (2021) 4255–4268, https://doi.org/10.1002/bit.27894.

[35] K.H. Liland, A. Kohler, N.K. Afseth, Model-based pre-processing in Raman spectroscopy of biological samples, J. Raman Spectrosc. 47 (6) (2016) 643–650, https://doi.org/10.1002/jrs.4886.
[36] D. Vaskó, D.K. Hajdú, E. Márkus, et al., Implementation of inline Raman spectroscopy for buffer exchange monitoring in monoclonal antibody formulation, Int. J. Pharm. 684 (2025) 126129, https://doi.org/10.1016/j.ijpharm.2025.126129.
[37] F. Feidl, S. Garbellini, S. Vogg, et al., A new flow cell and chemometric protocol for implementing in-line Raman spectroscopy in chromatography, Biotechnol. Prog. 35 (5) (2019) e2847, https://doi.org/10.1002/btpr.2847.
[38] K. Denny, M. Znidarsic, H. Hamedi, B. Stanley, S. Khattak, Development of a Raman Model for Continuous Monitoring of Detergent Viral Inactivation, Biotechnol. Bioeng. Publ. Online (2025), https://doi.org/10.1002/bit.70087.
[39] S. Guo, T. Bocklitz, J. Popp, Optimization of Raman-spectrum baseline correction in biological application, Analyst 141 (8) (2016) 2396–2404, https://doi.org/10.1039/c6an00041j.
[40] M. Poth, G. Magill, A. Filgertshofer, O. Popp, T. Großkopf, Extensive evaluation of machine learning models and data preprocessings for Raman modeling in bioprocessing, J. Raman Spectrosc. 53 (9) (2022) 1580–1591, https://doi.org/10.1002/jrs.6402.
[41] B. Barton, J. Thomson, E.L. Diz, R. Portela, Chemometrics for Raman Spectroscopy Harmonization, Appl. Spectrosc. 76 (9) (2022) 1021–1041, https://doi.org/10.1177/00037028221094070.
[42] V.I. Korepanov, Asymmetric least-squares baseline algorithm with peak screening for automatic processing of the Raman spectra, J. Raman Spectrosc. 51 (10) (2020) 2061–2065, https://doi.org/10.1002/jrs.5952.
[43] K. Furić, I. Ciglenečki, B. Ćosović, Raman spectroscopic study of sodium chloride water solutions, J. Mol. Struct. 550 (2000) 225–234, https://doi.org/10.1016/s0022-2860(00)00388-4.