TU Delft

Delft University of Technology

Document Version
Final published version

Licence
CC BY

Citation (APA)
Sfondalmondo, G. L., & Klijn, M. E. (2026). Balancing Prediction Performance and Spectral Interpretability: Multimetric Guided Design of Raman Spectrum Preprocessing Pipelines to Transfer Single Compound Models to Bioprocess Data. *Journal of Raman Spectroscopy*. https://doi.org/10.1002/jrs.70223

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

Check for updates

Journal of Raman Spectroscopy

WILEY

Journal of RAMAN SPECTROSCOPY

RESEARCH ARTICLE OPEN ACCESS

AI and Chemometrics for Raman Spectroscopy: Best Practice and Latest Applications

Balancing Prediction Performance and Spectral Interpretability: Multimetric Guided Design of Raman Spectrum Preprocessing Pipelines to Transfer Single Compound Models to Bioprocess Data

Gianmarco Luchetti Sfondalmondo | Marieke E. Klijn

Department of Biotechnology, Delft University of Technology, Delft, the Netherlands

Correspondence: Marieke E. Klijn (m.e.klijn@tudelft.nl)

Received: 30 June 2026 | **Revised:** 19 September 2026 | **Accepted:** 6 October 2026

Keywords: chemometrics | fermentation | optimization | preprocessing | process analytical technology

ABSTRACT

Raman spectroscopy is a widely explored in-line process analytical technology in bioprocessing due to its high chemical specificity and multiplexed readouts through chemometric modelling. However, developing robust chemometric models requires investment in bioreactor runs for data collection, making single-compound spectra a cost-effective calibration alternative, in which signal preprocessing is crucial for transferring models to bioprocess data. This work presents a multimetric framework for selecting preprocessing pipelines to support model transfer, balancing prediction performance, chemical interpretability, and flexibility to generalize across process conditions. It considers four aspects of signal quality: signal preservation, noise reduction, spectral resolution, and model performance. Approximately 65,000 preprocessing combinations were systematically screened to identify pipelines exhibiting balance across all four aspects. Selected pipelines were used to calibrate partial least squares (PLS) regressions and predict ethanol and glucose concentrations in four batch and one fed-batch aerobic *Saccharomyces cerevisiae* fermentation datasets. Compared to preprocessing pipelines selected based solely on prediction error, the multimetric framework yielded more robust and chemically interpretable models. Ethanol models matched literature performance using a single latent variable PLS. Moreover, this was achieved without a process-specific internal standard or biomass spectra, supporting flexibility and applicability to other processes. However, inherent weaker spectral features of glucose reduced performance, highlighting the limitations of a single-compound model developed solely on preprocessing optimization. Overall, this work shows that multiobjective preprocessing selection, combined with single-compound calibration, reduces costs, improves interpretability and flexibility in chemometric model development, while highlighting the value of analyte-specific strategies.

Abbreviations: AsLS, asymmetric least squares; EMSC, extended multiscattering correction; FWHM, full width at half maximum; HPLC, high-performance liquid chromatography; HTB, higher-the-better; IQR, interquartile range; LTB, lower-the-better; LV, latent variable; MAD, median absolute deviation; MPLS, morphological penalized least squares; PAT, process analytical technology; PDP, partial dependence plot; PLS, partial least squares; PLS-DA, partial least squares–discriminant analysis; RF, random forest; RMSECV, root mean square error of cross-validation; RMSEP, root mean square error of prediction; rRMSECV, relative root mean square error of cross-validation; $rRMSECV_{wq}$, relative root mean square error of cross-validation (worst quartile); rRMSEP, relative root mean square error of prediction; SNR, signal-to-noise ratio; VIP, variance importance for projection.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Journal of Raman Spectroscopy* published by John Wiley & Sons Ltd.

1 | Introduction

In recent years, bioprocess monitoring progressively shifted from off-line approaches, characterized by low frequency manual sampling and long analysis times, toward automated, (near-)real-time strategies, in alignment with Industry 4.0 principles [1]. This shift facilitates the collection of a greater number of data points, thereby minimizing temporal gaps during which the process remains unobserved, while minimizing operator-induced bias. This enhances reproducibility, data interoperability across processes, and overall interpretability of process dynamics [2, 3]. Among available process analytical technology (PAT), Raman spectroscopy is widely investigated for (near-)real-time bioreactor monitoring due to its high chemical specificity and compatibility with in-line probe configurations, which enable fully automated spectral collection [4]. The chemical specificity of this technique is well-suited for chemometric model implementation, both quantitative and qualitative, such as partial least squares (PLS) regression and partial least squares–discriminant analysis (PLS-DA), respectively [5]. However, although Raman instruments are relatively accessible in terms of purchase, installation, and operation costs, developing robust chemometric models requires significant resources in the field of bioprocessing [4, 6]. Model calibration necessitates dedicated bioreactor runs, rendering the development of models both costly and time-consuming. Calibration strategies based on single-compound spectra of the main fermentation broth components, including substrate, product, and biomass, represent a rapid and cost-effective alternative [7]. This enables time-efficient model development and facilitates model transfer across different bioreactor operation modes, such as batch and fed-batch, without the need to recalibrate. Nevertheless, a crucial enabling factor for single-compound-based calibration strategies, and their subsequent application to in-process spectra, is spectral data preprocessing [7].

Selecting an appropriate preprocessing pipeline is critical to the performance of chemometric models [8]. Insufficient preprocessing may obscure relevant spectral features, and overly aggressive pipelines may introduce mathematical artefacts that distort the shape of the signal and corrupt the interpretability of the results, potentially leading to erroneous conclusions [8, 9]. This sensitivity to preprocessing design is particularly pronounced in spectroscopic techniques applied to real-time bioprocess monitoring, such as in-line Raman spectroscopy in bioreactors, which operate without sample preparation or physical separation steps. Thus, spectra are exposed to a broad range of distortion sources, including spectral contributions from broth components, such as suspended biomass, as well as process-related physical interferences, such as gas bubbles [10]. As a result, this exposes recorded spectra to the simultaneous influence of fluorescence background, particle scattering contributions, and overlapping spectral bands. These challenges take on even greater significance in single-compound-based calibration workflows. Because PLS models cannot project or describe variance not expressed in the calibration dataset, model training based on single-compound data cannot mathematically compensate for unmodeled process disturbances, such as those arising from biomass or secondary metabolites [10]. The role of preprocessing is to suppress systematic physical interferences and reduce setup- and run-specific biases and variability. Applying appropriate preprocessing can improve the signal-to-interference ratio, highlighting the relevant spectral features related to the analyte. However, because different analytes have distinct Raman scattering cross sections and varying degrees of spectral overlap, preprocessing must be tailored for each target analyte [11].

Traditionally, the setup- and run-specific variability is addressed with signal standardization using the peak intensity of an internal standard that is neither consumed nor produced during the bioprocess [7, 12]. This strategy improves calibration model prediction performance but introduces process-specific conditions into preprocessing, compromising model transferability and making preprocessing and calibration sensitive to the specific measurement conditions under which the standard was established [13]. Instead, preprocessing should be designed to standardize variability, highlight relevant spectral features, and suppress disturbance sources, while avoiding the introduction of process-specific conditions. This would preserve the cost-effectiveness of single-compound calibration and transferability across processes. Alternatively, increasing model complexity by adding more latent variables (LVs) in PLS can capture unaddressed residual variance. However, this increases the risk of overfitting to variance specific to the calibration. This emphasizes the importance of selecting preprocessing methods based on principles to create robust and deployable chemometric models.

Regardless of its recognized importance, preprocessing pipeline selection in chemometrics historically relied on trial-and-error approaches, guided by visual inspection of spectra and shaped by knowledge and methodological habits of individual operators [8, 14]. In practice, manual selection can be effective, but it is time-consuming and susceptible to operator bias. It also often results in limited exploration of the preprocessing space, which includes the correction steps involved, the order of these steps, the selection of the algorithm, and the tuning of the parameters. More recently, systematic approaches to preprocessing optimization have emerged [15]. These approaches aim to navigate the preprocessing design space in a consistent and automated manner, reducing the human intervention during selection, targeting an objective function to be optimized [15]. Such approaches include computation and design methods such as design of experiment [16], Bayesian optimization [17], genetic algorithm [18], and grid search.

Despite the recent advancement in systematic screening, a fundamental challenge in preprocessing optimization lies in the definition of suitable objective functions. Cross-validation and external prediction error, such as mean squared error (MSE) [18] and root mean square error (RMSE) [16, 17], are conventionally used as primary optimization criteria, with pipeline configurations selected to minimize prediction error. However, minimizing prediction error alone is not always a sufficient proxy for model quality [8, 19]. A preprocessing pipeline selected based on prediction error may exploit chance correlations or anticorrelations within the calibration set. This can negatively impact chemical interpretability, which is widely recognized as an essential model property for robustness and reliable deployment [20]. Recognition of this

10974555, 0, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jrs.70223 by Technical University Delft, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

limitation motivated recent efforts to incorporate robustness metrics alongside prediction error in optimization frameworks, to balance prediction performance, interpretability, and model robustness simultaneously [19]. This issue is especially pertinent in the context of single-compound calibration strategies, where preprocessing must concurrently fulfil two partially conflicting objectives: enhancing the chemical features of the single-compound spectra, while preserving sufficient flexibility for the model to generalize to process spectra acquired in a complex bioprocess matrix. In this setting, optimizing exclusively for the lowest root mean square error of prediction (RMSEP) creates a dependency on process data. This dependency is fundamentally at odds with the rapid, abiotic calibration paradigm and risks incorporation of chance correlations that artificially reduce prediction error. Conversely, exclusive optimization for root mean square error of cross-validation (RMSECV) alone risks overfitting the single-compound calibration set, resulting in models that fail to transfer to in-process spectra within the complex bioprocess matrix [21]. Taken together, these limitations highlight the need for a balanced set of quantitative criteria to steer spectral preprocessing optimization.

To this end, a multimetric selection framework for Raman spectroscopy preprocessing pipelines is proposed in this work. A set of metrics was developed to evaluate preprocessed spectra across multiple dimensions commonly associated with spectral quality, including signal preservation, noise reduction, spectral resolution, and calibration performance. These metrics were systematically applied to assess approximately 65,000 preprocessing pipeline combinations via an automated grid search performed on single-compound spectra. The framework was developed and validated on glucose and ethanol spectra, two of the most analytically relevant compounds in microbial bioprocess monitoring, representing the primary carbon source and a key compound in overflow metabolism, respectively. This approach simulates a rapid calibration scenario in which biomass and process data are unavailable, and calibration resides solely on single-compound spectra acquired in the process medium. Rather than prioritizing a single metric or quality class, a pipeline selection logic was designed to simultaneously balance all metric classes, thereby identifying a pipeline configuration that best satisfies the full ensemble of spectral quality and calibration performance criteria. The multimetric balancing approach targets the following three model properties essential for practical deployment: prediction performance, interpretability, and flexibility to generalize across process conditions. By jointly enforcing these properties, the framework inherently reduces the risk of selecting pipelines that exploit spurious correlations or overfit single-compound features, known failure modes when optimizing solely on calibration or prediction error [22]. The proposed methodology aims to lower the barrier for rapid deployment of robust, transferable Raman spectroscopy-based PAT solutions across different bioreactor operation modes, thereby contributing toward a more systematic and interpretability-aware standard for preprocessing selection in bioprocess spectroscopy.

2 | Materials and Methods

All data processing, metrics development, calibrations, and analysis were performed in Python (Version 3.11.8). Spectral preprocessing routines were constructed using NumPy, SciPy, Pybaselines, and Chemotools. The implementation of PLS regression and random forest (RF) analysis was conducted using the Scikit-Learn library. The complete code is available on the authors' GitHub repository.

2.1 | Spectral Data and Reference Measurements

The Raman spectra were sourced from previous work, where in-line measurements were acquired during four batch and one fed-batch fermentations of *Saccharomyces cerevisiae* CEN.PK113-7D, as well as separate single-compound stepwise addition experiments in a bioreactor using a Kaiser Optical Systems RXN2 analyzer (785-nm excitation) coupled to a bIO-Optic immersion probe [23]. Spectra were recorded over the 100 to 3400 cm^{-1} range using continuous exposure, co-added into 1-min acquisition and subsequently cropped to the fingerprint region of 700–1800 cm^{-1}. In the present study, the five fermentation runs (60 spectra in total), and the single compound ethanol (31 samples, 0–500 mM, mean step 16.7 ± 2.6 mM) and glucose (26 samples, 0–247 mM, nonlinear spacing with mean step 9.9 ± 10.6 mM) stepwise addition series were used. For each single compound concentration, 10 consecutive 1-min acquisitions were averaged. The averaged single-compound spectra served as the reference dataset for data preprocessing pipeline evaluation and PLS calibration. Concentration values were determined by HPLC (Agilent 1260 Infinity, Amines HPX-87H column) using the same method for both the single-compound and in-line process samples. These concentration values were then used as reference labels for all calibration models.

2.2 | Metrics Implementation

In the absence of reference spectra for pure ethanol and glucose, a direct assessment of the quality of spectra post-preprocessing was not possible. To address this issue, a set of spectral quality metrics was developed for indirect evaluation of spectral quality after preprocessing. The metrics were grouped into the following four categories, each targeting a distinct aspect of spectral quality relevant to chemometric models: (I) preservation of spectral features, (II) noise reduction, (III) spectral resolution, and (IV) calibration performance (Figure 1). These are described in more detail in the following sections.

The metrics listed below were developed based on two assumptions. Firstly, Raman modes must manifest as positive peaks with identifiable positions that can be compared to a reference list. Consequently, metrics based on spectral shape are not applicable to spectra after the application of the first and second derivatives, where the peak position, the full width at half-maximum (FWHM), and the Voigt profile are no longer defined. The second assumption is that the signal intensity is proportional to the concentration of a single analyte. This makes metrics such as selectivity and signal-to-noise ratio (SNR) not applicable in spectra containing mixtures of compounds. However, in the case of mixtures, metrics based on peak shape remain applicable, provided that the positions of the target analyte peaks are known.

10974555, 0, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jrs.70223 by Technical University Delft, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

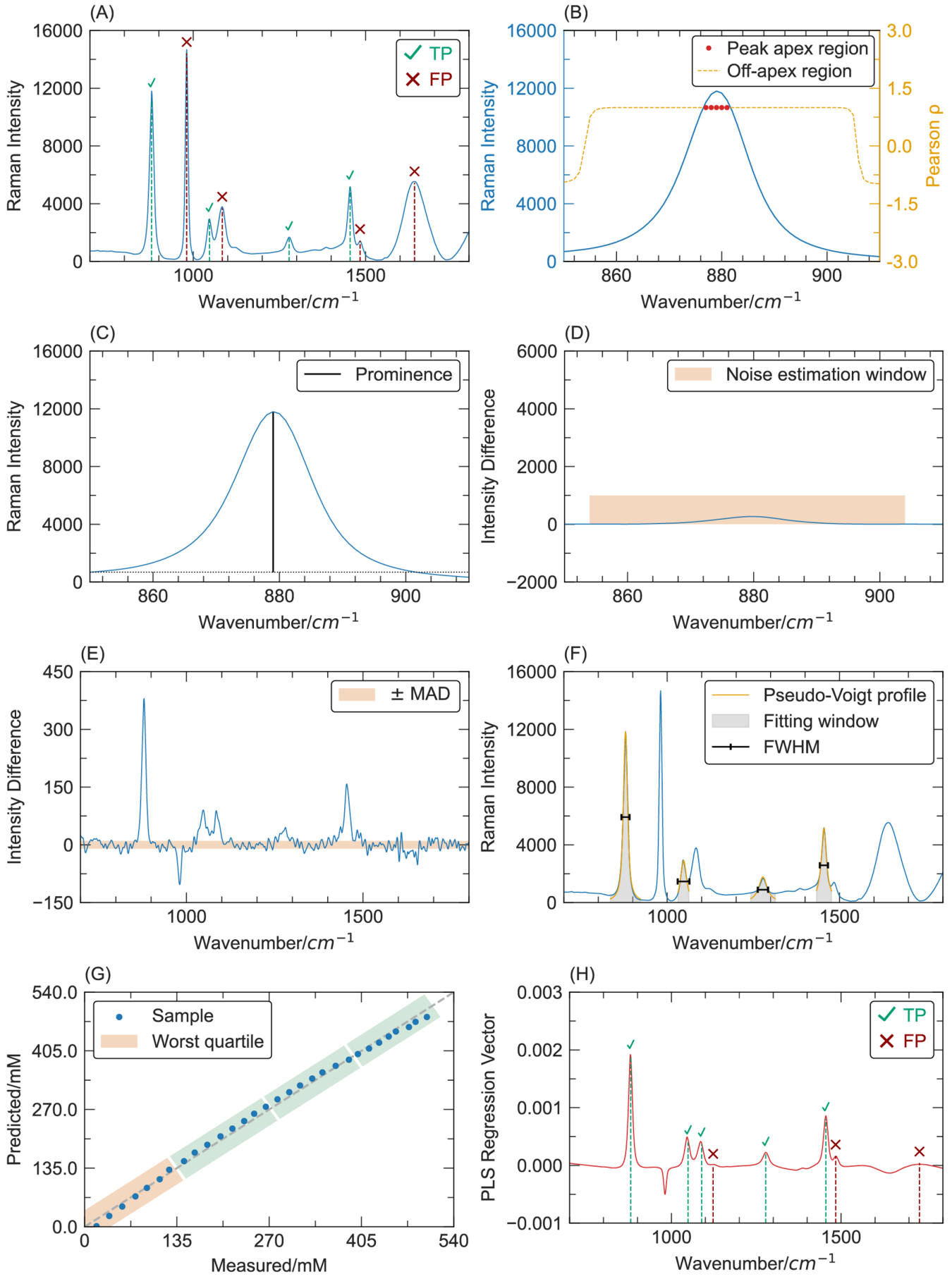


FIGURE 1 | Legend on next page.

10974555, 0, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jrs.70223 by Technical University Delft, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

FIGURE 1 | Graphical representation of the quality metrics calculated from the preprocessed ethanol single-compound spectra. (A) F1 score of the reference peaks on the mean single-compound spectrum. (B) Selectivity: the calculated Pearson correlation between the ethanol peak at 879 cm^{-1} and concentration. (C) SNR: measurement of the prominence of the ethanol peak at 879 cm^{-1}. (D) SNR: calculation for the noise in the ethanol difference spectrum in the 854-904 cm^{-1} region. (E) MAD calculation for the baseline of the ethanol difference spectrum. (F) Fitting of the pseudo-Voigt profile and calculation of the FWHM on the mean single-compound spectrum. (G) Determination of the quartile with the highest rRMSECV value following calibration. (H) F1 score of the reference peaks on the regression vector.

2.2.1 | Preservation of Spectral Feature Metrics

Two spectral preservation metrics were developed to assess the introduction of spectral artefacts or removal of chemically informative features after preprocessing, namely, the F1 score of reference peaks detection (Figure 1A) and selectivity (Figure 1B). For the F1 score, peaks were identified using the peak finder function of Scipy, which detects local maxima by comparison of neighboring values, with a minimum prominence threshold of 100 arbitrary units to exclude minor peaks arising from noise or baseline fluctuations. The identified peaks were compared with a set of reference peak positions that was compiled from literature sources (Table S1). Peaks that were matched within a tolerance of ± 2 cm^{-1} were counted as true positives (TP), reference peaks not identified in the preprocessed spectrum were counted as false negatives (FN), and detected peaks that were absent from the reference list were counted as false positives (FP). The ± 2 cm^{-1} tolerance was introduced to accommodate minor spectral shifts arising from differences in measurement conditions between the present dataset and literature reference values. The F1 score was subsequently determined by calculating the harmonic mean of precision and recall. The final value was obtained by calculating the median of the F1 scores from each spectrum (Equation (1)).

$$F1_{dataset} = median\left(2 \cdot \frac{TP_i}{TP_i + FP_i + FN_i}\right) \text{ for } i = 1, \ldots, N \tag{1}$$

where N is the total number of spectra. The F1 score was selected as it penalizes undersmoothed spectra, in which residual noise and baseline structures can generate numerous local maxima, and oversmoothed spectra, which retain only few well-defined peaks. In the first case, the high number of detected peaks increases the likelihood that some will coincide with reference positions within the tolerance range, thereby increasing TP number while concurrently resulting in a substantial increase of FP. In the second case, weak peaks risk being suppressed by overly aggressive preprocessing, leading to a low number of TP and high number of FN.

A selectivity metric was implemented based on the Pearson correlation coefficient (ρ) to ensure preprocessing does not distort the correlations between spectral intensity and concentration. For each wavenumber, ρ was calculated between the signal intensity and the known concentration label across all spectra in the single-compound datasets. Around the peaks identified as TP in the previous step, a consensus mask was constructed by gathering all data points within a window of ± 2 cm^{-1} (5 cm^{-1} in total). The narrow window was selected to isolate the peak apex where the signal yields the strongest correlation with concentration, and to exclude the peak's flanks, which may include overlapping bands and residual baseline. The median correlation coefficient was computed within the masked peaks regions ($\boldsymbol{\rho_{peaks}}$) and across the remaining wavenumbers ($\boldsymbol{\rho_{baseline}}$), excluding the peak regions. The selectivity (S) of a dataset was defined as Equation (2):

$$S = median(\boldsymbol{\rho_{peaks}}) - median(\boldsymbol{\rho_{baseline}}) \tag{2}$$

The quantity S captures how strongly peak regions correlate with concentration relative to nonpeak regions. S was calculated independently for both raw and preprocessed spectra. The ratio of these spectra was expressed as a symmetric log-ratio, bounded between 0 and 1 (Equation (3)):

$$S_{score} = exp\Big(-\big|log(S_{processed} / S_{raw})\big|\Big) \tag{3}$$

A S_{score} of 1 indicates that preprocessing fully preserves the original selectivity structure, while values approaching 0 indicate substantial alteration of the spectra-concentration relationship, penalizing both artificial enhancements and degradation equally.

2.2.2 | Noise Reduction Metrics

Two noise reduction metrics were implemented to evaluate the removal of noise artefacts, scattering, and baseline effects. The first metric in this category is the SNR improvement (Figure 1C,D) at the TP peak positions, estimated in accordance with the method described by McCreery et al. (2000) [24]. Local noise was estimated within a ±25-cm^{-1} window around each peak, since Raman shot noise is signal-dependent and an off-peak estimate would lead to an underestimation [24]. For each pair of adjacent spectra $(j, j+1)$ in the single-compound sets ($j = 1, \ldots, N-1$), the pointwise difference was calculated to remove the common signal component. The standard deviation within a ±25-cm^{-1} window around the k-th TP peak position ($k = 1, \ldots, K$) was calculated and divided by $\sqrt{2}$ to correct for variance inflation introduced by the subtraction of two independent measurements (Equation (4)).

$$\sigma_{j,k} = \frac{std(\boldsymbol{\delta s_j})}{\sqrt{2}} \tag{4}$$

where $\boldsymbol{\delta s_j} = \boldsymbol{s_{j+1}} - \boldsymbol{s_j}$ is the pointwise difference between consecutive spectra. This yields one noise estimate per peak per spectrum pair and compose the matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{((N-1)\times K)}$, with rows indexed by spectrum pair and columns indexed by peaks. Entries corresponding to a peak not detected in one of the two spectra of a given pair were left undefined (NaN). The representative noise for each TP group was then obtained as the column-wise median of $\boldsymbol{\Sigma}$, computed ignoring NaN entries (Equation (5)), giving the noise vector $\boldsymbol{\sigma} \in \mathbb{R}^{(K)}$. The SNR per peak group was defined as the ratio of the median peak prominence to the median noise (Equation (6)):

10974555, 0, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jrs.70223 by Technical University Delft, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$$\boldsymbol{\sigma} = median_j(\boldsymbol{\Sigma}) \; for \; J = 1, \; \ldots \; , N-1 \tag{5}$$

$$\mathbf{SNR} = \frac{\boldsymbol{h}}{\boldsymbol{\sigma}} \tag{6}$$

where $\boldsymbol{h}$ is the vector of median peak prominences, with h_k its k-th element. The overall SNR of the dataset was taken as the median SNR across all K TP peak groups, thereby producing a single scalar for the entire dataset (Equation (7)).

$$SNR_{dataset} = median_k(SNR) \text{ for } k = 1, \; \ldots \; , K \tag{7}$$

The final SNR improvement was calculated as the ratio between the SNR of the preprocessed spectra and that of the raw spectra.

Applied to a dilution series, the difference spectrum derived from the pointwise subtraction is not a pure noise measurement. The difference spectrum also includes contributions from the concentration increment between adjacent spectra. Because the same spectra are used to evaluate the candidate pipelines, this contribution is common to all configurations compared and should not affect their relative ranking. Anyway, this contribution is proportional to the analyte band, so an aggressive pipeline that attenuates the band can also reduce the noise estimation. Nevertheless, the attenuation acts proportionally on the prominence of the peak in the numerator and on the estimated noise in the denominator, which normalizes this effect. It is important to note that overly aggressive preprocessing may also distort the shape of the band, rather than uniformly resizing it. This may affect the two terms differently. This potential estimation error is compensated for by metrics based on peak shape, such as the pseudo-Voigt profile, which penalize unrealistic distortions in peak shape.

A second metric was used to counterbalance chance SNR overestimation caused by residual baseline or scattering artefacts increasing peak prominence. This was done through the median absolute deviation (MAD) of the baseline (Figure 1E). As outlined above, signal components were removed by computing the pointwise difference for each pair of consecutive spectra. The MAD was calculated on each difference spectrum, yielding a 1D array where each value is the MAD of a specific spectrum in the dataset. The overall MAD was taken as the median MAD across all spectra, thereby producing a single scalar for the entire dataset. Lower values indicate a flatter, more uniform baseline after preprocessing.

2.2.3 | Spectral Resolution Metrics

Spectral resolution metrics (Figure 1F) were implemented to assess enhancements in peak shape and separation after preprocessing. Even though only a single analyte is used, peak overlap can still occur between the vibrational modes of the same molecule. For example, bending and stretching modes of glucose within the range of approximately 900–1000 cm^{-1}[25]. Furthermore, residual baseline curvature beneath the peaks can alter both their apparent width and position. Consequently, these metrics serve to evaluate the recovery of the band profile and the presence of potential distortions introduced by overly aggressive preprocessing.

This was achieved by fitting individual peaks to a model line shape. The Voigt profile is the theoretical line shape for liquid-phase Raman scattering, arising from the convolution of Gaussian and Lorentzian components. The pseudo-Voigt approximation was adopted as a computationally inexpensive alternative, with the mixing parameter η and total FWHM derived from the individual Gaussian (f_G) and Lorentzian (f_L) widths following the polynomial approximation of Thompson et al. (1987) [26].

For each peak, an adaptive symmetric window was determined around the peak center to isolate the band from nearby spectral overlaps. Starting from the apex of all the peaks identified (Section 2.2.1), including those not matched to a reference position, the peaks were traversed outward in both directions until the intensity increased, signaling the beginning of a neighboring band. To exclude adjacent features, the narrower of the two identified half-widths was mirrored around the apex, and two points of zero intensity were appended one wavenumber beyond each edge. The zero anchors constrain the fitted profile to decay within the observed band extent, preventing convergence on broad, low-amplitude solutions that match the peak apex but overestimate the band width. Initial estimates for f_G and f_L were set by dividing the empirical FWHM equally between the two components. The fit was performed using the trust region reflective least squares optimization, with the peak center constrained to $\pm 5\,cm^{-1}$ of the detected position and individual width parameters bounded between 0.001 and $30\,cm^{-1}$. This was done to prevent convergence on broad baseline features rather than individual Raman peaks.

Subsequently, the goodness of fit was evaluated through the coefficient of determination (R^2). For each spectrum, the median FWHM and median R^2 across all fitted peaks were calculated, yielding one median FWHM and one median R^2 per spectrum. These values were then aggregated as the median across all spectra to produce a single scalar per dataset. Narrower FWHM values following preprocessing are indicative of enhanced spectral resolution, while R^2 values approaching 1 confirm that the pseudo-Voigt model adequately describes the peak shape and that the peak has not been distorted by the preprocessing.

2.2.4 | PLS Calibration Metrics

PLS-model metrics were used to assess the performance of calibration models trained on preprocessed single-compound spectra with their corresponding HPLC concentration labels. A PLS regression model with a single LV was calibrated and evaluated through Venetian blind cross-validation with five splits. Model accuracy was measured using a stratified relative root MSE of cross-validation (rRMSECV) expressed as percentage of the calibration interquartile range (IQR), as defined by Klaverdijk et al. (2025) [23]. The concentration range was divided into quartile bins, and the rRMSECV was computed independently within each bin. The rRMSECV of the worst performing quartile ($rRMSECV_{wq}$) was taken as the final metric, ensuring the value accurately reflects prediction accuracy across the entire concentration range, not just well-populated areas (Figure 1G).

To quantify model interpretability, the regression vector ($\boldsymbol{\beta}$) from the PLS model was analyzed. Peaks were detected in the positive portion of the regression vector and matched against the reference peak position within a tolerance of $\pm 5\,cm^{-1}$. The F1 score was calculated from the number of matched (TP), unmatched detected (FP), and missed reference peaks (FN), quantifying

10974555, 0, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jrs.70223 by Technical University Delft, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

how well the model relies on known spectral features of the target analyte (Figure 1H). The F1 score was further penalized by the fraction of negative energy in the regression vector, defined as the ratio of the squared sum of negative coefficients to the total squared sum (Equation (8)):

$$R_{\text{neg}} = \frac{\sum(\beta_i^2) \text{ for } \beta_i < 0}{\sum(\beta_i^2) \text{ for all } \beta_i} \tag{8}$$

where β_i denotes the regression coefficient at the i-th wavenumber. The final interpretability score was computed as (Equation (9)):

$$\text{F1}_{\text{score}} = \text{F1} \times (1 - R_{\text{neg}}) \tag{9}$$

This penalization prevents the selection of pipelines that produce models with large negative coefficient lobes, which in single-compound calibration indicate reliance on anticorrelated spectral regions rather than direct analyte features.

2.3 | Grid Search

The metrics described in Section 2.2 were used to screen spectral preprocessing configurations through a grid search. The preprocessing pipelines were structured as a sequential chain of three operations, which were executed in the following order: (I) scattering correction, (II) baseline removal, and (III) noise filtering, consistent with common practice in chemometric preprocessing of spectral data [27]. A rescaling step was appended to the end of each pipeline, ensuring that the minimum spectral intensity was shifted to zero if preprocessing introduced negative values. For each operation, a range of algorithms and parameterizations were considered, including a "none" condition to test configurations, where one or more preprocessing steps were excluded.

Baseline correction and noise filtering were selected as common preprocessing steps for Raman spectra and reflect established practice for this analytical technique. Scattering correction, on the other hand, is less commonly applied in Raman spectroscopy. It was included based on previous work, using extended multiplicative scattering correction (EMSC) with a quadratic term and the average spectrum as a reference to correct for the scattering effect [23]. The EMSC algorithm was implemented using the Chemotools library, and first-, second-, and third-order polynomials were included in the search space, whereas the zeroth order corresponds to the standard multiplicative scattering correction (MSC) [28]. The baseline correction algorithms were derived from the Pybaselines library, and the noise filters from SciPy. As each operation could also take the value "none," the necessity of each step was determined by the calculated metrics.

The pipeline screening was automated in the following way: 65,000 unique combinations were generated from a YAML configuration file and applied sequentially to single compound ethanol and glucose datasets independently (Table S2). Each preprocessed dataset was then passed to the evaluation routine, which returned a 1D array of the calculated metrics. The results matrix was assembled by means of concatenating the metric arrays across all pipeline configurations. The grid search was executed in parallel using the joblib module on a computer cluster that was managed through SLURM.

2.4 | Selection Framework

A selection framework was designed to identify the best preprocessing pipeline based on metrics calculated during the grid search. The framework consists of two steps: (I) hard filter and (II) desirability-based ranking (Figure 2).

The hard filter was implemented to exclude spectral preprocessing pipelines that fail to meet minimum requirements for both calibration accuracy and spectral interpretability. This step ensures that pipelines producing well-fitting, but chemically meaningless models, or vice versa, are discarded before ranking. One metric from each category was incorporated into the filter, namely, selectivity (spectral preservation), SNR improvement (noise reduction), Pseudo-Voigt profile (spectral resolution), and $\text{rRMSECV}_{\text{wq}}$ (calibration performance), with a threshold and direction defined for each (Table 1). Pipelines failing to satisfy at least one criterion were excluded from the subsequent selection step. The same thresholds were applied to both analytes, except for the pseudo-Voigt profile, for which glucose spectra yielded systematically lower values than ethanol due to the broader and more overlapping nature of glucose Raman bands. The selection of metrics and threshold values for the hard filter was informed by two principles: (1) each metric category should contribute at least one constraint, thus preventing single-criterion optimization and (2) the threshold must be permissive enough to retain a diverse pool of candidates while excluding clearly defective configurations. The SNR improvement threshold excludes preprocessing pipelines that reduce rather than increase the ratio. The $\text{rRMSECV}_{\text{wq}}$ threshold of 5% was adopted as a commonly used benchmark for adequate calibration performance in chemometric applications. Finally, the selectivity threshold ensures that

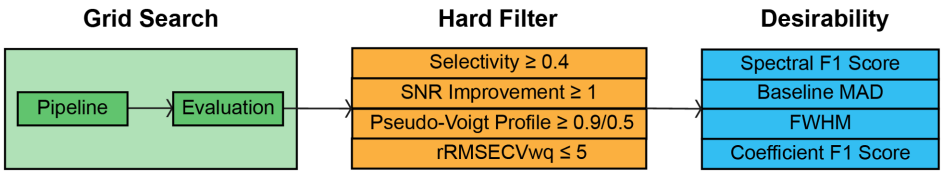


FIGURE 2 | Scheme of selection process for preprocessing pipelines. The evaluation of the candidates is executed through a grid search process, using the quality metrics previously defined. Subsequently, the hard filter selects pipelines based on selectivity, SNR improvements, pseudo-Voigt profile, and $\text{rRMSECV}_{\text{wq}}$ (in %). If the pipeline does not satisfy at least one of the established thresholds, it is rejected. The final selection step involves the ranking of the most effective pipelines using a desirability scoring metric, which is calculated for spectral F1 score, the baseline MAD (in intensity units), the FWHM (in cm^{-1}), and the coefficient F1 score.

10974555, 0, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jrs.70223 by Technical University Delft, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 | Hard filter thresholds and directions applied to each metric for the glucose and ethanol preprocessing grid search. Pipelines failing at least one criterion were excluded from the subsequent ranking. The direction indicates whether values above (≥) or below (≤) the threshold are retained.

	Glucose		Ethanol	
Metric	**Threshold**	**Direction**	**Threshold**	**Direction**
Selectivity	0.4	≥	0.4	≥
SNR improvement	1	≥	1	≥
Pseudo-Voigt profile	0.5	≥	0.9	≥
$rRMSECV_{wq}$ (%)	5	≤	5	≤

the preserved signal remains analytically meaningful, retaining only pipelines that maintain the linear relationship between spectral features and analyte concentration without substantially distorting the original spectral structure.

The pipelines selected by the hard filter were ranked using the following metrics: spectral F1 score (spectral preservation), baseline MAD (noise reduction), FWHM (spectral resolution), and regression vector F1 score (calibration performance). As these metrics differ in both magnitude and optimization direction, a desirability scaling was applied following the one-side transformation function described by Derringer et al. (1980) [29]. To transform each metric value (y_i) into a dimensionless desirability score in the range [0, 1], where 0 indicates an undesirable value and 1 a fully desirable value, the lower-the-better (LTB) or higher-the-better (HTB) function was used as appropriate (Equations (10) and (11)).

$$\mathrm{LTB} = \begin{cases} 0 \text{ if } y_i < L_i \\ \left(\dfrac{U_i - y_i}{U_i - L_i}\right) \text{ if } L_i \le y_i \le U_i \\ 0 \text{ if } y_i > U_i \end{cases} \tag{10}$$

$$\mathrm{HTB} = \begin{cases} 0 \text{ if } y_i < L_i \\ \left(\dfrac{y_i - L_i}{U_i - L_i}\right) \text{ if } L_i \le y_i \le U_i \\ 1 \text{ if } y_i > U_i \end{cases} \tag{11}$$

For each metric, upper (U_i) and lower (L_i) bounds were defined to set the desirability design space (Table 2). The lower bound for FWHM was derived from the maximum spectral resolution of the instrument, while the upper bound was set to the median FWHM of the raw spectra. For baseline MAD, the lower bound was set to the theoretical minimum of the statistic, while the upper bound was set to the MAD of the raw spectra prior to the application of the scattering correction and baseline removal. For the LTB metrics, any value that falls below the lower bound receives a desirability of 0 rather than 1, because the lower bound for FWHM and baseline MAD represents a physical limit. Therefore, a value below it shows mathematical overcorrection rather than improvement, and the resulting spectra cannot be considered valid.

The overall desirability score per spectrum was obtained as the geometric mean of the four individual desirability scores. The geometric mean was selected as it penalizes pipelines that yield a very low score on any single metric, thereby favoring configurations that perform robustly across all four criteria rather than excelling on one at the expense of the others.

TABLE 2 | Desirability function bounds and optimization direction for each metric used in the ranking step.

Metric	Lower bound (L_i)	Upper bound (U_i)	Direction
Spectral F1 score	0	1	HTB
Baseline MAD (intensity)	0	Raw value	LTB
FWHM (cm^{-1})	4	Raw value	LTB
Coefficient F1 score	0	1	HTB

2.5 | Validation Methods

2.5.1 | PLS Validation

To evaluate the efficacy of the proposed selection framework, a PLS model with a single LV was calibrated on single-compound spectra preprocessed with the selected pipeline, using Venetian blind cross-validation (five splits). A distinct model was calibrated for each analyte, using solely the target compound's dilution series together with a medium spectrum. Single-compound spectra of the other analyte and biomass were excluded from the calibration set. The calibrated models were subsequently applied to predict target analyte concentration in bioprocess spectra preprocessed with the selected pipelines. To assess transferability from single compound to process condition, the R^2 and the relative root MSE of prediction were calculated, with the latter being normalized by the IQR of process data (rRMSEP). This validation step checks if the selected preprocessing pipeline for single compound data can lead to calibration models that generalize to the complex spectral environment of a fermentation broth.

2.5.2 | Benchmark Dataset

To benchmark the performance of the proposed multimetric selection framework, a comparison was made between the single-compound PLS models obtained and those reported by Klaverdijk et al. (2026) [7]. While the PLS models were calibrated in the

10974555, 0, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jrs.70223 by Technical University Delft, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

present study using single-compound spectra of only the target analyte (glucose or ethanol), the benchmark study used a calibration set combining five glucose, five ethanol, and five biomass spectra, plus the medium spectrum (hereinafter referred to as biomass-inclusive dataset), selected from the same single-compound datasets used for the calibration of the present study. The benchmark study also reported results obtained by calibrating PLS models with a dataset containing only five glucose and five ethanol spectra, plus the medium spectrum (hereinafter referred to as the biomass-free dataset). The calibration sets in the present study are larger in terms of the number of samples but are limited to a single compound. Furthermore, the models obtained by the benchmark study incorporated three LVs for the biomass-inclusive dataset and two LVs for the models calibrated with the biomass-free dataset, whereas our PLS models were calibrated using only one LV. Moreover, our analysis did not use sulfate band normalization. For comparison, the rRMSEP was calculated over the same concentration range based on the same test dataset.

2.5.3 | Random forest Analysis

An RF analysis was performed to identify which metric and selection elements are relevant to the prediction error obtained on process spectra. For the 65,000-pipeline dataset, the rRMSEP was calculated for each configuration using the method described in Section 2.5.1. Two RF models were trained for each analyte: one using the raw spectral quality metrics as features, and the other using the selection elements of the framework, namely, the hard filter binary flags (0 for a failed criterion and 1 for a passed criterion) and the desirability scores. In both cases, the target variable was the rRMSEP. Following training and evaluation on a 70:30 train-test split (Tables S3 and S4), permutation importance was calculated to identify which metrics and selection elements are most predictive of calibration transfer. The importance analysis was supplemented by the calculation of partial dependence plots (PDPs) to emphasize the direction and shape of each metric's relationship with rRMSEP.

3 | Results and Discussion

3.1 | Effect of Selection Criterion on Calibration Quality

During chemometric model building, it is common practice to evaluate potential preprocessing configurations by selecting the pipeline that yields the lowest prediction error. To benchmark the lowest rRMSEP selection against the multimetric approach designed in this work, the rRMSEP was calculated for glucose and ethanol for each of the 65,000 preprocessing configurations. The preprocessing pipelines yielding the lowest rRMSEP were selected and used to train PLS models using the protocol described in 2.5.1. The same procedure was followed for the pipelines selected using the multimetric selection framework. The single-compound preprocessed spectra are reported in Figures S1 and S4, and the comparison between the raw average spectrum of bioprocess and the preprocessed average spectrum of bioprocess is shown in Figures S2 and S5.

The preprocessing pipeline selected based on only the lowest rRMSEP for ethanol was: (I) scattering correction through EMSC method (polynomial order = 3); (II) baseline removal through morphological penalized least squares (MPLS) method ($\lambda = 10^3, p = 10^{-1}$, differential order = 2); (III) smoothing through median filter (size = 7). The selected glucose preprocessing pipeline was: (I) scattering correction through EMSC method (polynomial order = 3); (II) baseline removal through MPLS method ($\lambda = 10^5, p = 10^{-3}$, differential order = 3); (III) smoothing through Savitzky–Golay filter (window length = 15, polynomial order = 1). The two selected pipelines demonstrated transfer capability from single-compound spectra to bioprocess spectra for both analytes, resulting in prediction errors below 10% and prediction R^2 values exceeding 0.97 (Figure 3A,C and Table 3).

However, single-compound preprocessed spectra (Figure S1) and model regression coefficients (Figure 3B,D) reveal distortions in spectral features and amplification of spurious correlations. When the preprocessing pipeline selected based on the lowest rRMSEP for ethanol is applied to bioprocess spectra, it removes all broad spectral features, retaining only sharp peaks on a flat baseline, with the water bending band and matrix-related features suppressed (Figure 3B). The aggressive suppression of broad features is also reflected in the regression vector, which shows oscillations and negative lobes concentrated in the 800-1100 cm^{-1} region (Figure 3B). In the calibration dataset, 73% of wavenumbers of the band located between 840 and 920 cm^{-1}, centered at about 879 cm^{-1} [30], exhibit a Pearson correlation coefficient with ethanol concentration ≤ -0.9 (Figure S1A). At the same time, the 47% of the regression coefficients of the same band have negative sign with variance importance for projection (VIP) score below 1 (values above the threshold of VIP > 1 are considered significant). In raw spectra, the corresponding values are 30% and 0%, respectively. The presence of strong negative correlations is unexpected, as the calibration spectra are collected for increasing analyte concentrations. However, negative correlations in raw data are located at the lobes of the band, suggesting medium or water features being diluted as ethanol is added. Furthermore, the sulfate peak located in the 960-1000 cm^{-1} region, centered at approximately 981 cm^{-1}, shows a VIP = 1.54 when the rRMSEP-optimized preprocessing is applied, with respect to the raw value of about 0.41. This indicates that the model considers this spectral feature to be a significant predictor of ethanol concentration, even though it is not specific for the target.

For glucose, the regression vector retains positive lobes at known glucose positions (~1063 and ~1126 cm^{-1}), but these are counterbalanced by equal negative lobes across the entire spectral range (Figure 3D). In single-compound spectra, within the region between 1000–1500 cm^{-1}, where eight characteristic peaks are located [25], about 35% of wavenumbers demonstrate a Pearson correlation coefficient with glucose concentration ≤ -0.9 (Figure S1B), and 23% exhibit regression coefficients with negative sign and VIP > 1, while the corresponding values in raw spectra are both 0%. The observed symmetry between positive and negative contributions in the glucose regression vector suggests that this model achieves its predictive performance through the cancellation of opposite spectral

10974555, 0, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jrs.70223 by Technical University Delft, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

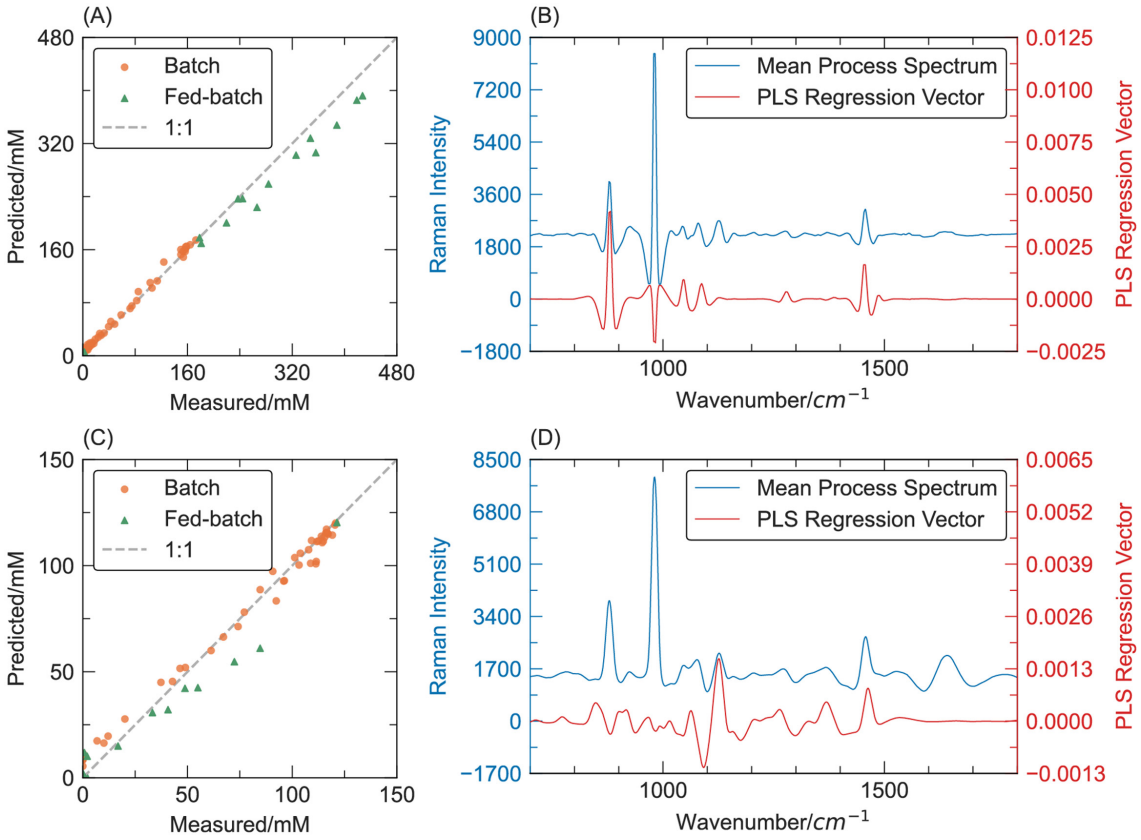


FIGURE 3 | Application of calibration models built with preprocessing selected based on lowest rRMSEP to bioprocess data. (A,B) Ethanol: (A) measured versus predicted values and (B) mean preprocessed bioprocess spectrum compared with regression coefficients. (C,D) Glucose: (C) measured versus predicted values and (D) mean preprocessed bioprocess spectra compared with regression coefficients.

TABLE 3 | Cross-validation and prediction statistics for the calibration model built with preprocessing pipeline selected through best rRMSEP.

	R^2 CV	RMSECV (mM)	rRMSECV (%)	R^2 P	RMSEP (mM)	rRMSEP (%)
Ethanol	0.98 ± 0.00	17.9 ± 2.6	6.94	0.97	14.1	9.39
Glucose	0.99 ± 0.00	3.7 ± 2.2	4.47	0.98	6.9	6.90

contributions rather than through a response to analyte-specific spectral features. Furthermore, bioprocess spectra after application of selected preprocessing show a baseline offset of about 2000 arbitrary units (Figure 3B,D). This distortion originates from the baseline removal step, when the fitted baseline crosses the characteristic peaks of the analytes rather than passing beneath them (Figure S3). Therefore, subsequent subtraction removes part of the bands, creating regions of negative intensity, and the following rescaling step introduces the observed offset. The aggressive subtraction also distorts the shape of the peaks and their correlation with concentration, as observed above, causing the pipelines to be excluded by the hard filter due to the selectivity and peak shape (pseudo-Voigt fitting).

Taken together, these data show how preprocessing that is solely optimized to reduce the rRMSEP can introduce anticorrelations where they are not present (i.e., glucose), boost them when they are already there (i.e., ethanol), and confer predictive relevance to features not related to the target. As previously demonstrated by other studies, when preprocessing is driven exclusively by minimizing prediction error, aggressive mathematical transforms can radically alter the significance of spectral variables, and chemometric modelling may not be interpretable or robust [19, 20]. Overly aggressive preprocessing can generate artefacts that displace chemically meaningful information and replace it with spurious correlations arising from the transforms themselves, rather than from the underlying chemistry [9].

To overcome the challenge of balancing spectral correction with the preservation of chemical information and predictive accuracy, the developed multimetric framework was applied rather than a single-metric approach. The preprocessing pipeline selected based on the multimetric framework for ethanol comprises the following stages: (I) scattering correction through EMSC method (polynomial order = 1); (II) baseline removal through asymmetric least squares (AsLS) method ($\lambda = 10^6, p = 10^{-1}$, differential order = 2); (III) smoothing through Savitzky–Golay filter (window length = 11, polynomial order = 2). The glucose preprocessing pipeline involves: (I) scattering correction through EMSC

10974555, 0, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jrs.70223 by Technical University Delft, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

method (none); (II) baseline removal through AsLS method ($\lambda = 10^6, p = 10^{-2}$, differential order = 2); (III) smoothing through Savitzky–Golay filter (window length = 7, polynomial order = 2). The two pipelines yielded poorer prediction statistics compared to the rRMSEP-optimized selection, particularly for glucose, where the RMSEP increased from 6.9 to 18.9 mM and the R^2 decreased from 0.98 to 0.83 (Figure 4A,C and Table 4).

However, the regression vectors of the ethanol and glucose models now align with known spectral features of their target analytes, enabling chemical interpretation (Figure 4B,D). The ethanol regression vector exhibits smooth and well-defined positive correlations with characteristic ethanol peaks (Figure 4B), such as symmetric C-C-O stretching (879 cm^{-1}), C-C stretching (1046 cm^{-1}), asymmetric C-C-O stretching (1086 cm^{-1}), scissoring vibration of -CH_2- (1278 cm^{-1}), and asymmetric deformation of -CH_3 (1455 cm^{-1}) [30]. The glucose regression vector (Figure 4D) presents a more complex and noisier structure due to more overlapping Raman-active modes in the 700-1800 cm^{-1} region. Nevertheless, unlike the previous model (Figure 3D), the regression vector obtained with the multimetric selected pipeline does not exhibit broad negative lobes and the main spectral features are identifiable. For glucose, these include C-C stretching (850 cm^{-1}) associated with the α-anomer, C-O-H bending (917 and 1126 cm^{-1}), C-O stretching (1063 cm^{-1}) of the β-anomer, -CH- bending (1369 cm^{-1}), and -CH_2- deformation (1462 cm^{-1}) [25]. Furthermore, both models assign negative regression coefficients at the sulfate band (981 cm^{-1}) [23]. Nevertheless, this peak's contribution to concentration prediction is minor, as the average VIP score in the 960-1000 cm^{-1} region is 0.70 for ethanol and 0.34 for glucose. This below the significance threshold and close to raw-spectra value (0.41) for ethanol, in contrast to the rRMSEP-optimized model discussed above. The ethanol model also assigns a negative regression coefficient at the water band (1640 cm^{-1}) [23], with a VIP score of 0.34.

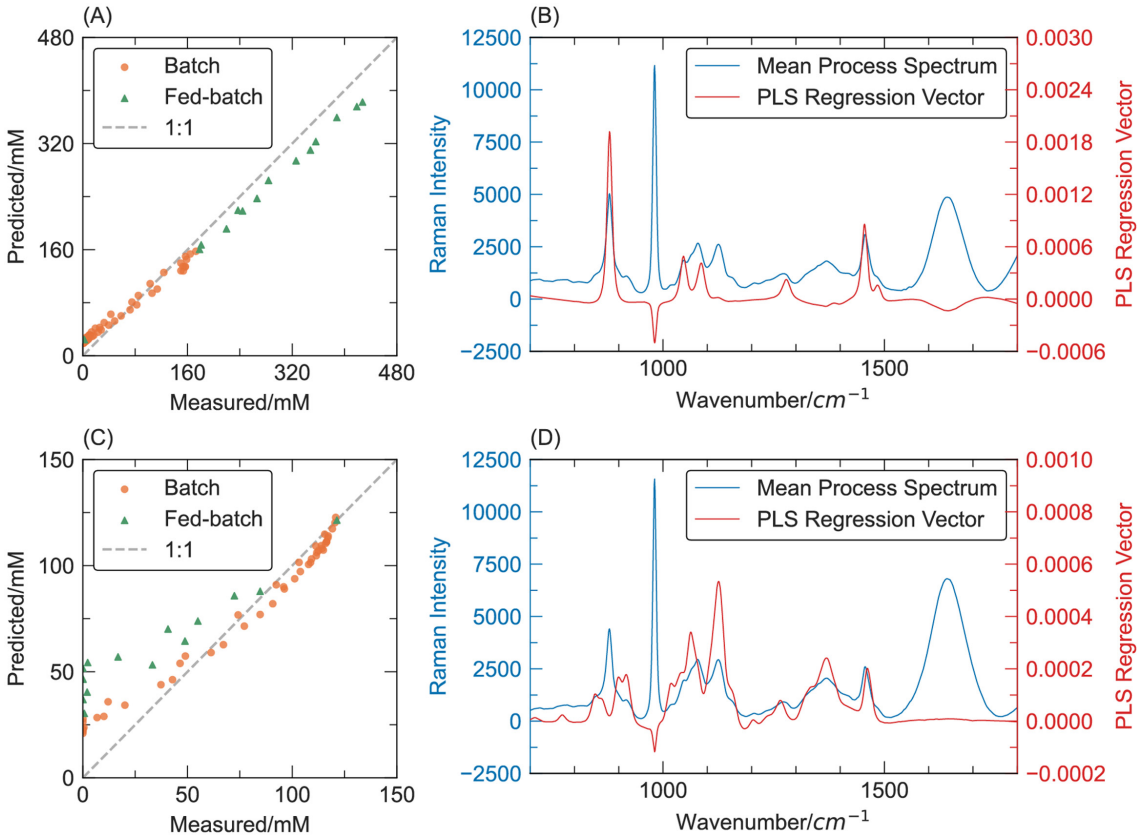


FIGURE 4 | Application of calibration models built with preprocessing selected through the multimetric framework to bioprocess data. (A,B) Ethanol: (A) measured versus predicted values and (B) mean preprocessed bioprocess spectrum compared with regression coefficients. (C,D) Glucose: (C) measured versus predicted values and (D) mean preprocessed bioprocess spectra compared with regression coefficients.

TABLE 4 | Cross-validation and prediction statistics for the calibration model built with preprocessing pipeline selected through multimetrics framework.

	R^2 CV	RMSECV (mM)	rRMSECV (%)	R^2 P	RMSEP (mM)	rRMSEP (%)
Ethanol	0.99 ± 0.00	9.7 ± 1.3	3.75	0.98	20.1	13.39
Glucose	0.99 ± 0.00	3.1 ± 0.3	3.65	0.83	18.9	18.93

10974555, 0, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jrs.70223 by Technical University Delft, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

To establish if the anticorrelations identified above are also introduced when selection is not driven by prediction error, the same analysis was applied to single-compound spectra and regression coefficients obtained after application of metric-selected pipelines. The Pearson correlation coefficient with concentration, the regression coefficient sign, and the VIP scores were calculated for wavenumbers in the 840-920 and 1000-1500 cm^{-1} regions, for ethanol and glucose, respectively. For ethanol spectra, the application of preprocessing resulted in 31% of wavenumbers with Pearson correlation coefficient ≤ -0.9, compared to 30% in the raw data, and 0% of negative regression coefficients with VIP > 1. For glucose, 3.4% of the wavenumbers exhibited a negative correlation with concentration, compared to 0% in the raw data, and 0% of negative regression coefficients with VIP > 1. This highlights how, for both analytes, neither the negative correlations nor chemically uncorrelated features carry weight in concentration prediction after preprocessing. Despite the lower predictive performance, the multimetric-selected pipelines preserve and highlight analyte-specific features, so the resulting models predict concentrations based on chemically relevant, analyte-specific variables, rather than from preprocessing-induced anticorrelations.

3.2 | Comparison With Established Single-Compound Calibration Approaches

PLS models calibrated with the multimetric-selected pipelines were compared with literature results reported by Klaverdijk et al. (2026) [7], as described in Section 2.5.2. The comparison results are reported in Table 5.

For ethanol, the PLS model calibrated on ethanol single-compound spectra alone, without sulfate peak normalization, and with a single LV yielded results comparable to Klaverdijk et al. (2026), within approximately 2 percentage points of error (Table 5). With ethanol-only calibration, the optimized preprocessing configuration enables a single LV to capture 99.91% of spectral variance of the calibration set (Table S5). Adding a second LV degrades rRMSEP sharply to 88.24%, as additional LVs fit spectra specific variations rather than genuine analyte signal (Figure S6A–C). The same preprocessing was applied to the biomass-inclusive calibration set of Klaverdijk et al. (2026) (Figure 5A,B and Table 6). With this calibration set, the first LV of PLS model explains 69.81% of the variance (Table S6), and the X-loading plot shows the structure of ethanol spectra, with the primary peaks identified at 879, 1046, 1087, 1279, and 1455 cm^{-1} (Figure S7A). The addition of two LVs increases the explained variance to 99.69% and 99.94% (Table S6), respectively. In contrast to the ethanol-only case, the additional components describe matrix structure (Figure S7B,C). For instance, the water band in LV2 at approximately 1639 cm^{-1} and yeast cell-related peaks in LV3 at approximately 1449 and 1663 cm^{-1}. With LV = 3, the model reaches an rRMSEP of 11.4%, against 11.6% reported by Klaverdijk et al. (2026). Therefore, the single-LV model calibrated using ethanol single-compound spectra alone, after selected preprocessing, achieves a comparable transfer performance with an error margin of 2 percentage points, using a simpler calibration set and a lower model rank.

For glucose, the PLS model calibrated on glucose single-compound spectra alone, without sulfate peak normalization, and with a single LV yielded a prediction error 5.5 and 14.1 percentage points higher than biomass-free and biomass-inclusive models, respectively (Table 5). As in the ethanol case, the first LV captures 99.73% of the calibration set variance, and the main peaks in the region 1000–1500 cm^{-1} are identifiable (Table S5 and Figure S6D). On the other hand, adding a second LV to the glucose-only model lowers the error to 14.54% (Table S5), approaching the biomass-free model. Additionally, visual inspection of the parity plot (Figure 4C) shows systematic overpredictions at low glucose concentrations in both batch and fed-batch data. This phenomenon is likely due to the absence of biomass in the calibration dataset. Peaks characteristic of yeast can be identified in the 1300-1500 cm^{-1} region [7], overlapping with peaks characteristic of glucose.

The same preprocessing was then applied to the biomass-inclusive calibration set (Figure 5C,D and Table 6) of Klaverdijk et al. (2026). With this calibration set, the bias at low concentrations inverts to underprediction (Figure 5C). In this case, the first LV explains 37.14% of the variance (Table S6). The X-loading shows the main glucose peaks at about 846, 917, 1063, 1368, and 1466 cm^{-1}, even though it is dominated by a negative lobe at 879 cm^{-1}, corresponding to the main ethanol peak (Figure S7D). Therefore, the first LV represents a glucose-ethanol contrast, rather than glucose alone. The inversion of the misprediction is likely due to the contrast between the two analytes. While the glucose is consumed, the ethanol is produced and accumulates in the media. Both effects are particularly evident in the fed-batch data, where biomass and ethanol reach higher concentrations. LV2 captures most of the variance, which brings the cumulative value to 93.44% and the X-loading reproduces the ethanol structure (879, 1046, 1085, and 1278 cm^{-1}), with a minor glucose contribution at 1064 cm^{-1} (Figure S7E). Furthermore, LV2 also shows contributions from the matrix, such as the water band at 1636 cm^{-1} and a yeast cell-related signal at 1455 cm^{-1}. The addition of LV3 increases explained variance to 99.93%, with a further water contribution at 1638 cm^{-1} and reduces the rRMSEP to 11.3%. Adding sulfate-band normalization to the three-LV model leads to the best calibration result, decreasing

TABLE 5 | Comparison of rRMSEP (%) for glucose and ethanol between the multimetric preprocessing selection framework (LV = 1, no sulfate normalization, single-analyte calibration set) and Approach 1 of Klaverdijk et al. (2026) in its biomass-free (LV = 2) and biomass-inclusive (LV = 3) configurations.

	Multimetric selection (LV = 1)	**[7] Biomass-free (LV = 2)**	**[7] Biomass-inclusive (LV = 3)**
rRMSEP ethanol (%)	13.4	14.8	11.6
rRMSEP glucose (%)	18.9	13.4	4.8

10974555, 0, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jrs.70223 by Technical University Delft, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

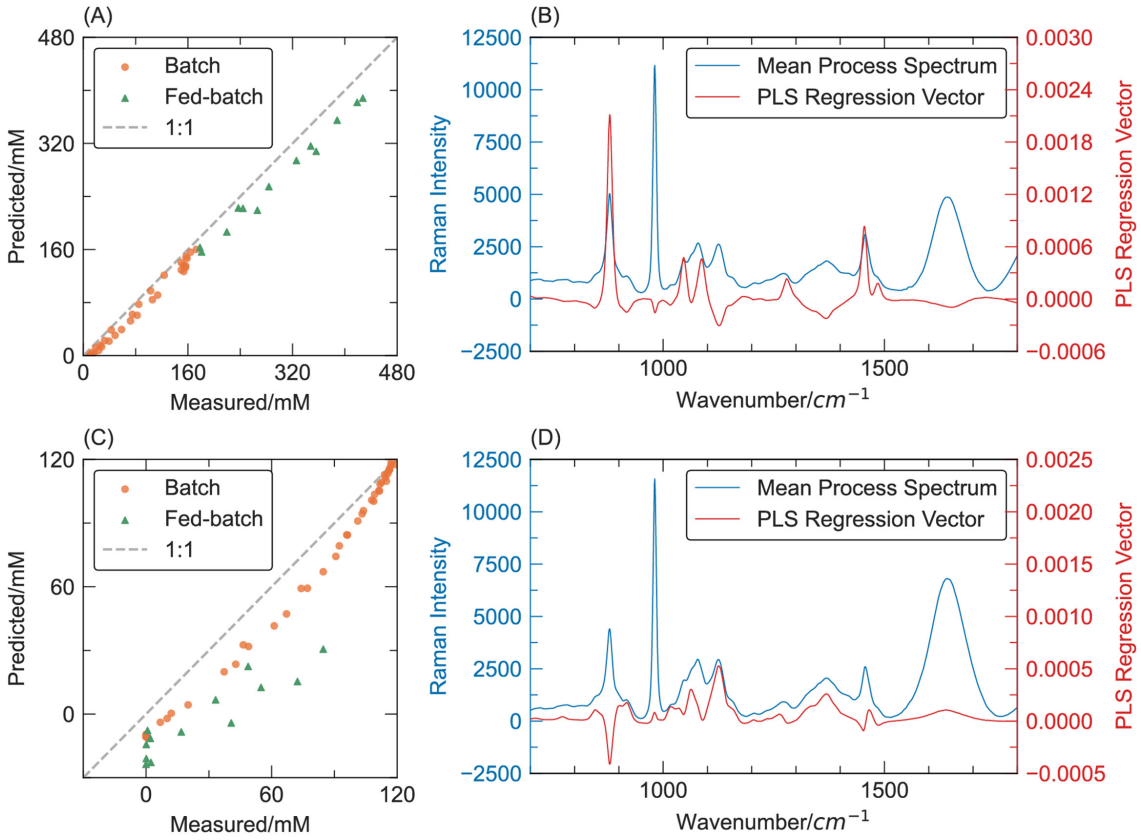


FIGURE 5 | Application to bioprocess data of calibration models trained on the biomass-inclusive calibration dataset from Klaverdijk et al. (2026), using preprocessing selected through the multimetric framework. (A,B) Ethanol: (A) measured versus predicted values and (B) mean preprocessed bioprocess spectrum compared with regression coefficients. (C,D) Glucose: (C) measured versus predicted values and (D) mean preprocessed bioprocess spectra compared with regression coefficients.

TABLE 6 | Cross-validation and prediction statistics for the calibration model trained on the biomass-inclusive calibration dataset from Klaverdijk et al. (2026), using preprocessing selected through the multimetric framework.

	R^2 CV	RMSECV (mM)	rRMSECV (%)	R^2 P	RMSEP (mM)	rRMSEP (%)
Ethanol	0.98 ± 0.02	13.7 ± 4.3	10.04	0.97	19.9	13.23
Glucose	0.86 ± 0.24	12.6 ± 8.4	20.55	0.85	18.1	18.09

the error to 4.3%, comparable with 4.8% reported by Klaverdijk et al. (2026).

Taken together, ethanol and glucose delineate both the power and limits of systematic preprocessing optimization within single-compound calibration workflows. For spectrally resolved analytes with intense, narrow peaks (e.g., ethanol), preprocessing optimization alone can compensate the need for greater model complexity, internal standard normalization, or multi-compound calibration, highlighting relevant spectral features and suppressing sources of disturbance. The simplest model, using the optimized preprocessing, matched the published results, where adding the other analyte and biomass to the calibration set, and increasing the number of LVs yielded only a small improvement. In contrast, for analytes with weak, broad bands (e.g., glucose), preprocessing alone is not sufficient. Matching the published results requires both ethanol and biomass in the calibration set, a three-LV model, and sulfate-band normalization. This shows that for strongly overlapping compounds, simple models cannot account for cross-analyte interactions. In presence of biomass and other analyte spectra, the first LV describes a glucose-ethanol contrast rather than glucose alone, while higher-rank LVs progressively resolve the spectral contributions of the other analyte and the spectral matrix. Consequently, a single-LV model calibrated on glucose alone cannot separate the variance of the target analyte from its interferents, even if the spectra are corrected accurately.

The analyte-specificity of the selected pipelines was further confirmed by a cross-pipeline swap, where the glucose pipeline was applied to ethanol, and vice versa. This increased the rRMSEP to approximately 46% in both cases. A single unified pipeline is therefore not feasible for the two analytes, given the spectral differences. These findings suggest that preprocessing optimization is both necessary and often sufficient for spectrally resolved analytes, enabling simpler, more interpretable models as

10974555, 0, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jrs.70223 by Technical University Delft, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

advocated by Borg et al. (2026) [20], but not for overlap-limited targets, where more complex models are required to separate the analyte from interferents.

3.3 | Mechanistic Analysis of the Selection Framework

Through an RF regression analysis, it was investigated to which extent components of the multimetric selection framework contributed to calibration transfer performance, and whether the contribution is analyte-dependent (Figure 6). Each metric was ranked based on the permutation importance to prediction accuracy. The importance analysis was supplemented by the calculation of partial dependence plots (PDPs) to emphasize the direction and shape of each metric's relationship with rRMSEP (Figures S8–S12). Specifically, it determines whether higher or lower values of a given metric result in reducing error (beneficial), increasing error (detrimental), or exhibiting no influence.

The importance analysis highlights similarities and differences between the analytes. For ethanol and glucose, the regression vector's F1 score is the primary selection criterion, the baseline MAD desirability is the second most important feature, and the $rRMSECV_{wq}$ is the most important hard filter threshold (Figure 6A). This outcome was anticipated as a well-calibrated model with coefficients aligned to the analyte's known peaks is a prerequisite for an effective chemometric model. In a similar manner, clean baselines are critical for both analytes, given the susceptibility of weak spectral features to background interference. The reduced gap between baseline MAD and regression vector's F1 score importance for glucose (gap ~0.16) in comparison to ethanol (gap ~0.48) is indicative of a higher sensitivity to background interference for glucose. This is because glucose's weak Raman scattering cross-section is more easily obscured by spectral background.

The importance analysis also shows analyte-dependent behavior during the selection. In the selection process for ethanol, desirability metrics were found to be of greater significance than hard filter flags. However, this is not the case for glucose, where the $rRMSECV_{wq}$ hard filter flag is ranked third in terms of overall importance. This reflects a more stringent constraint on calibration quality for this analyte. The pseudo-Voigt profile flag of glucose demonstrates a higher level of importance than the spectral F1 score, suggesting that peak shape preservation is more critical for glucose than peak detection. This is consistent with the broad, overlapping nature of glucose peaks, where maintaining the integrity of peak profile is a necessary condition to fit the relevant spectral features. For ethanol, well-defined peaks yield consistently high pseudo-Voigt R^2 values across preprocessing configurations, reducing the discriminant power of this metric.

The direction and shape of each metric's effect on transfer performance was further assessed through a PDP analysis. The U-shape shown by the PDP of the spectral F1 score, selectivity, and pseudo-Voigt profile of ethanol (Figure 6B–D) and the selectivity, baseline MAD, and SNR improvement of glucose (Figure 6F–H) indicate that the optimal configuration lies in a modest interval. Permissive preprocessing fails to highlight relevant features or remove disturbance sources (i.e., undersmoothing), while overly aggressive preprocessing removes relevant spectra features (i.e., oversmoothing). Similarly, a local minimum was observed for the $rRMSECV_{wq}$ of ethanol and the regression vector's F1 score of glucose (Figure 6E,I). This suggests that overoptimization of calibration metrics on single-compound spectra produces models that overfit pure compound spectral information and lose the ability to recognize concentration-relevant features when transferred to the more complex bioprocess matrix. Together, these findings show the necessity of a multimetric evaluation framework that balances different aspects of spectra quality to assure prediction performance and model interpretability.

The hard filter was designed to prevent optimization from converging on physically implausible solutions, as shown in an ablation experiment where the filter was removed and pipelines were selected solely based on desirability optimization. For ethanol, the same pipeline was selected regardless, confirming that the desirability optimization is the dominant selection mechanism for this analyte (Figure 6A). For glucose, removal of the hard filter selected a pipeline with lower rRMSEP, passing from 18.93% to 14.54% (Figure S13A and Table S7). Nevertheless, inspection of the regression vector profile obtained without hard filter revealed high-frequency oscillations across the entire spectral range, suppressed minor peak intensities (1369 and 1462 cm^{-1}), and absence of the water band around 1640 cm^{-1} (Figure S13B). These characteristics indicate spectral overcorrection, fitting of chance correlations, and loss of interpretability. This means the hard filter does not optimize transfer performance directly but excludes chemically meaningless pipelines. To further characterize the hard filter mechanism, the individual metric failure rates, the number of filters failed per pipeline, and the relationship between rRMSEP and spectral quality metrics were analyzed, distinguishing passing and failing pipelines.

The hard filter is more restrictive for glucose in comparison to ethanol (Figure 7A). Of the approximately 65,000 pipelines that were tested, only seven satisfied all the hard filter requirements for glucose, compared to approximately 12,000 for ethanol. For ethanol, 40% of pipelines failed one or two metrics, with $rRMSECV_{wq}$ showing the highest failure rate (~61%), followed by pseudo-Voigt R^2 (~30%) (Figure 7A,B). For glucose, most pipelines fail three or four metrics simultaneously (Figure 7A). The $rRMSECV_{wq}$ failure rate is comparable to that observed for ethanol, but all other metrics show failure rates above 70%, with selectivity being the most restrictive (Figure 7B). This result reflects a fundamental challenge of preprocessing glucose Raman spectra, namely, achieving well-defined peak shapes (high pseudo-Voigt R^2), a clean baseline (low baseline MAD), and preserving spectral selectivity simultaneously given the abundance of weak, overlapping peaks in the region 700–1800 cm^{-1}. The peak profiles of glucose result in a heightened susceptibility to distortion of the spectral features by preprocessing operations that are generally well-tolerated by stronger scatters [31, 32].

The scatter plots show that selection metrics not involved in the hard filter have a common distribution pattern for analytes (Figure 7C–J). Preprocessing pipelines that pass the hard filter cluster in the central region of the metric value distribution. The passing pipelines do not have the highest or the lowest value for any single metric and therefore fall within the moderate range. This indicates the hard filter selects pipelines that

10974555, 0, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jrs.70223 by Technical University Delft, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

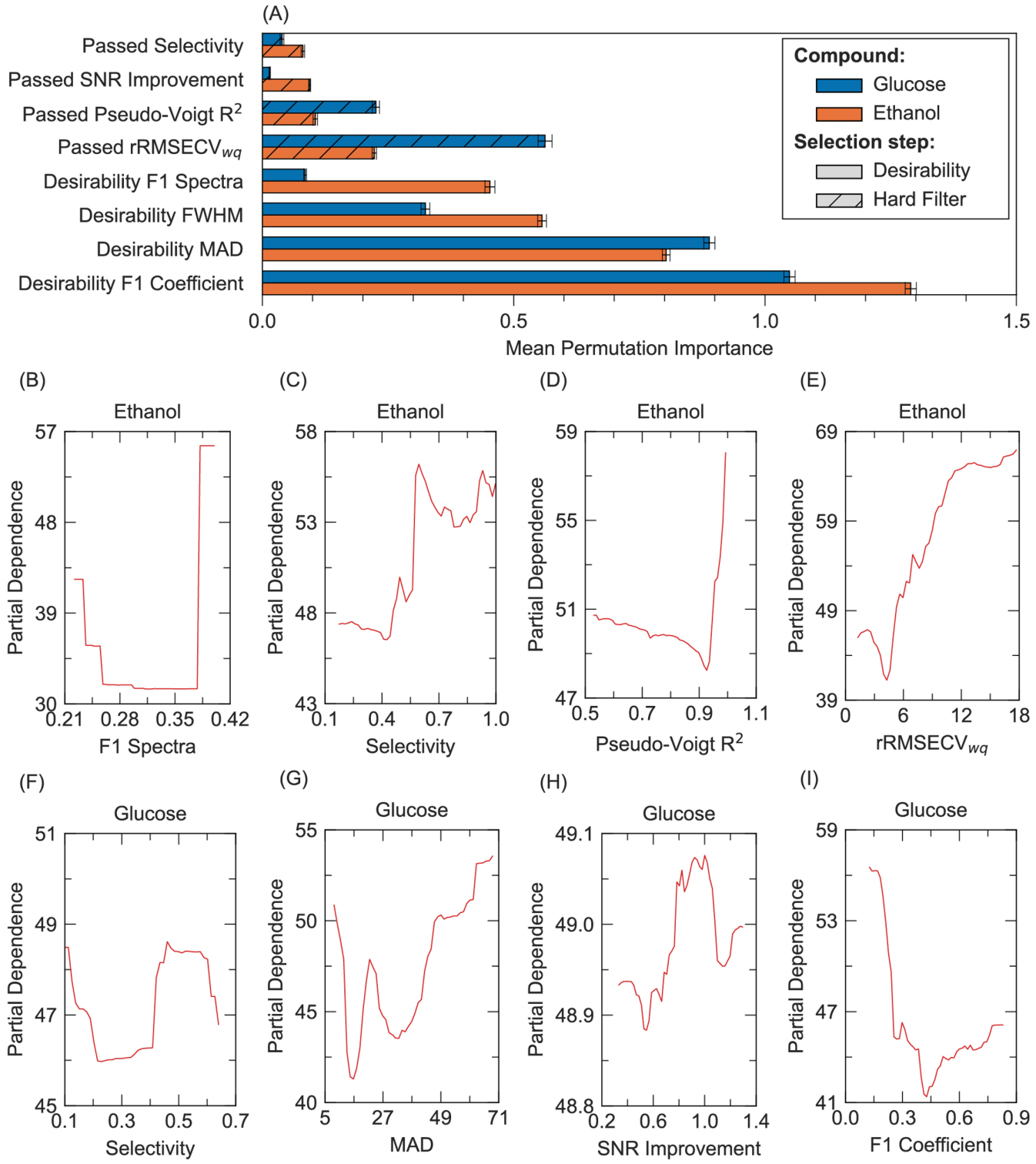


FIGURE 6 | Results of the random forest analysis. (A) Permutation importance of selection framework elements for predicting rRMSEP (in %), sorted by ethanol importance. (B–E) Raw metrics PDP for ethanol: F1 spectra (B), selectivity (C), pseudo-Voigt R^2 (D), and $rRMSECV_{wq}$ in % (E). (F–I) Raw metrics PDP for glucose: selectivity (F), baseline MAD in intensity units (G), SNR improvement (H), and regression coefficient vector F1 score (I). The PDP plots shown in the panel were selected because they represent specific cases of U-shaped curves.

achieve a balanced trade-off across multiple spectral qualities. Conversely, the pipelines that pass the filter cluster in the region of low rRMSECV values with rRMSEP values ranging from low to moderate (Figure 7K,L). Furthermore, all passed pipelines lie above the 1:1 line, meaning that the prediction error never falls below the calibration error. This suggests that the hard filter effectively prevents overfitting and excludes spurious cases where the model appears to predict better than it calibrates.

To quantify the overall effect of the hard filters on calibration transfer performance, a comparison was made based on the rRMSEP distributions of passing and failing pipelines. To reduce

10974555, 0, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jrs.70223 by Technical University Delft, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

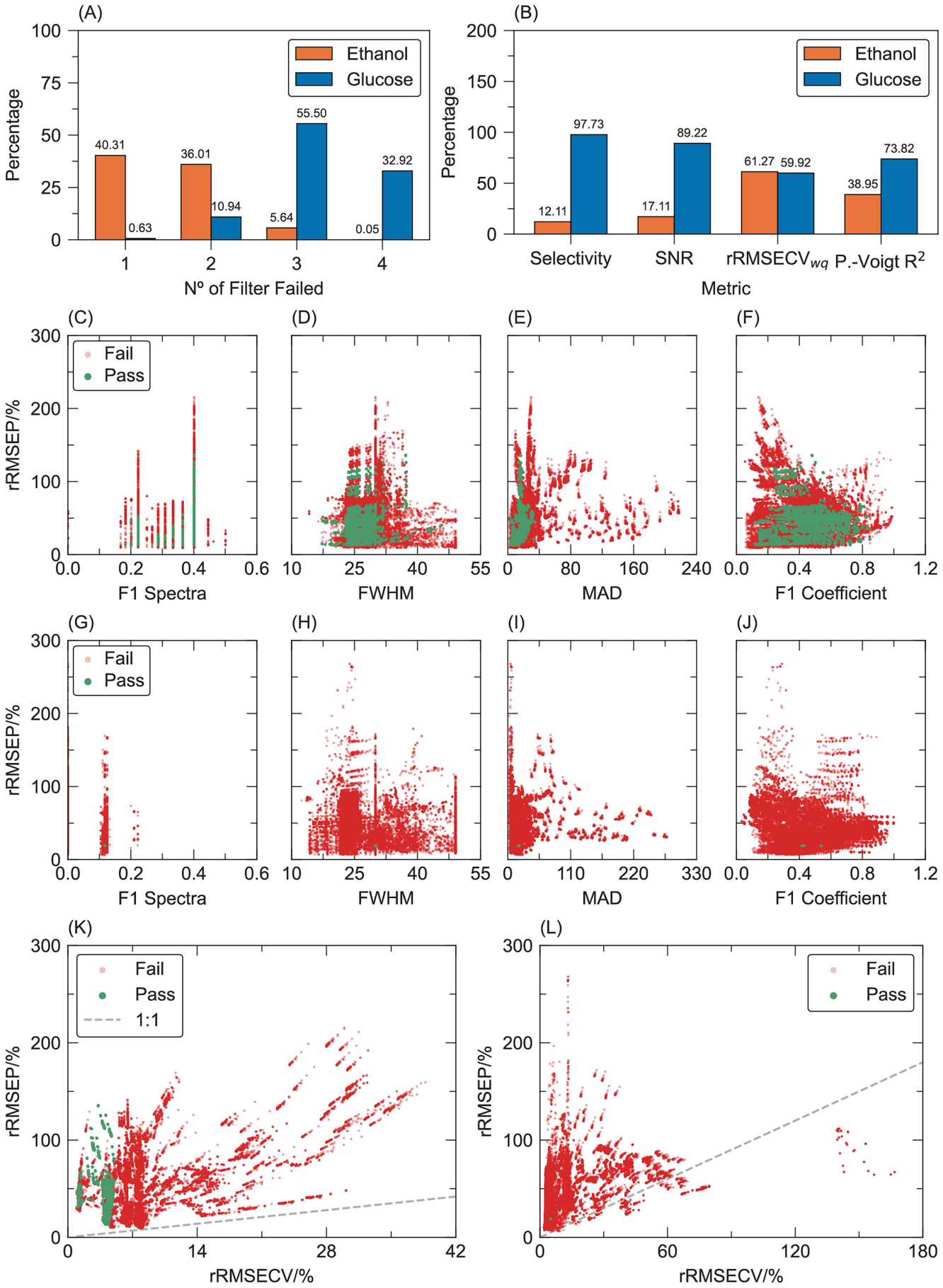


FIGURE 7 | Legend on next page.

10974555, 0, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jrs.70223 by Technical University Delft, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

FIGURE 7 | Hard filter analysis. (A) Percentage of pipelines failing the hard filter by number of failed thresholds (1–4) for ethanol and glucose. (B) Individual failure rate of each hard filter metric for ethanol and glucose. (C–F) Ethanol: scatter plots of nonfiltered spectral quality metrics versus rRMSEP, passing pipelines shown in green, failing pipelines in red. The y-axis is common to all graphs. (G–J) Glucose: scatter plots of nonfiltered spectral quality metrics versus rRMSEP, passing pipelines shown in green, failing pipelines in red. The y-axis is common to all graphs. (K) Ethanol: calibration consistency plots of rRMSECV versus rRMSEP; the dashed line indicates the 1:1 reference. (L) Glucose: calibration consistency plots of rRMSECV versus rRMSEP; the dashed line indicates the 1:1 reference.

TABLE 7 | Comparison of rRMSEP distribution between pipelines that pass the hard filter and those that fail. The best median and IQR values for ethanol and glucose are highlighted in bold.

Analyte	Group	No. of pipelines	Median rRMSEP (%)	P25 rRMSEP (%)	P75 rRMSEP (%)	IQR (%)
Ethanol	Pass	11,644	**45.93**	35.02	52.05	**17.03**
Ethanol	Fail	53,080	57.21	35.54	71.52	35.98
Glucose	Pass	7	**18.99**	18.93	19.02	**0.09**
Glucose	Fail	64,721	44.56	34.58	60.73	26.15

Note: The bold values represent the best statistics obtained for the comparison.

the influence of skewed and nonnormal distributions, the median, the 25th percentile (P25), the 75th percentile (P75), and the IQR were reported for each analyte (Table 7). Median captures the central tendency, while the IQR quantifies the spread of prediction errors within each group. Together these metrics reveal whether the hard filters shift the distribution toward lower errors and whether the hard filter reduces the variance of transfer performance, indicating more reliable and consistent spectral preprocessing pipeline selection.

For both analytes, the passing group shows a lower median rRMSEP than the failing group. This is expected as the hard filter excludes physically implausible spectra, leading to spectra with $rRMSECV_{wq}$ exceeding the 5% threshold. Chemically meaningful spectra with well-controlled cross-validation error should transfer more reliably to process data. For ethanol, the passing group shows a ~53% narrower IQR, indicating that the hard filter enforces consistency of transfer performance across a large and diverse set of pipelines. Nevertheless, the passing set remains sufficiently large for the desirability to identify the same optimal pipeline regardless of whether the hard filter is applied, as confirmed by the ablation experiment. This suggests that for analytes with intense, narrow peaks, the desirability ranking alone is sufficient to ensure model transferability, interpretability, and robustness. The hard filter reduces the candidate pool and enforces consistency but does not alter the final pipeline selection. For glucose, the IQR decreased by 99.7%, from 26.150 to 0.087, reflecting the restrictiveness of the hard filter for this analyte. The seven retained pipelines for glucose form a near-identical cluster, with almost no variance in rRMSEP. The restrictiveness simultaneously enforces both statistics and reduces the candidate set to a narrow cluster of pipelines that balance prediction performance and spectral interpretability. For glucose, whose signal is weak, broad, and overlapped with biomass, these quality dimensions lack the discriminative power to exclude implausible pipelines, making the hard filter a necessary gating mechanism. This is consistent with the permutation importance analysis (Figure 6A), where hard filter flags retained relative importance for glucose, while for ethanol all hard filter flags were ranked below the desirability metrics. Together, these results show how the hard filter plays an analyte-dependent role, namely, enforcing consistency for spectrally resolved compounds and acting as a critical gating mechanism for weak, overlapped Raman scatterers.

4 | Conclusions

Although single-compound calibration offers a cost-effective workflow for calibrating Raman chemometric models, realizing its full potential requires preprocessing strategies that ensure prediction performance, chemical interpretability, and transferability across processes without recourse to process-specific normalization factors or increased model complexity. To solve these challenges, this work introduces a multimetric framework for preprocessing pipeline selection across four quality dimensions: signal preservation, noise reduction, peak resolution, and calibration accuracy. It aims to balance model performance, chemical interpretability, and flexibility to generalize across process conditions. In this study ~65,000 preprocessing pipelines were screened for PLS (LV = 1) calibration of glucose and ethanol on single-compound spectra and validated on batch and fed-batch fermentation data. The results showed analyte-dependent outcomes in terms of model performance and effectiveness of each of the two layers in our selection framework. This was attributed to the fundamental difference in spectral features between ethanol and glucose, highlighting different needs between strong and weak scatters for calibration model building. For model performance, ethanol matched published results within two percentage points, while glucose performance was poorer, due to the fundamental limitation of developing single-compound calibrations for weak, overlap-limited analytes. Although glucose model showed lower predictive performance, the model developed using the two-layer multimetric selection strategy exhibited high chemical interpretability, with regression coefficients aligned with known relevant spectral peaks. In summary, the developed framework enabled the calibration of simple and

10974555, 0, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jrs.70223 by Technical University Delft, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

interpretable models, without the need for internal standard normalization or prior knowledge on other compounds present in a complex matrix. The presented framework showed that the three objectives, performance, interpretability, and flexibility, can be jointly achieved for analytes with sharp and intense spectral signatures. Nevertheless, for overlap-limited analytes, such as glucose, integration with multicompound calibration strategies and adoption of more complex modelling strategies represent a natural extension to address the remaining performance gap. In general, systematic selection of spectral preprocessing pipelines to highlight relevant spectral features reduces the necessity of process-specific steps or datasets for calibration, such as internal standards and biomass spectra, respectively. The findings in this work support the establishment of robust, transparent, and maintainable chemometric workflows in the field of bioprocessing.

Author Contributions

Gianmarco Luchetti Sfondalmondo: conceptualization, methodology, validation, formal analysis, investigation, software, writing – original draft, writing – review and editing, visualization. **Marieke Klijn:** conceptualization, resources, writing – original draft, writing – review and editing, supervision, administration, funding acquisition.

Acknowledgements

This project is funded by Schmidt Sciences.

Funding

This work was supported by Schmidt Sciences.

Disclosure

The authors used the large language model Claude AI (Sonnet 4.6 model, by Anthropic) in the preparation of this manuscript to rephrase sentences, improve the readability and flow of the text, and verify grammar. The generated text was reviewed by the author and manually integrated into the manuscript; it was never copied and pasted directly. The author is responsible for the content of this manuscript.

Data Availability Statement

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

References

1. C. L. Gargalo, I. Udugama, K. Pontius, et al., "Towards Smart Biomanufacturing: A Perspective on Recent Developments in Industrial Measurement and Monitoring Technologies for Bio-Based Production Processes," *Journal of Industrial Microbiology & Biotechnology* 47 (2020): 947–964.

2. A. Christler, E. Felföldi, M. Mosor, et al., "Semi-Automation of Process Analytics Reduces Operator Effect," *Bioprocess and Biosystems Engineering* 43 (2020): 753–764.

3. T. Habich and S. Beutel, "Digitalization Concepts in Academic Bioprocess Development," *Engineering in Life Sciences* 24 (2024): 2300238.

4. K. A. Esmonde-White, M. Cuellar, and I. R. Lewis, "The Role of Raman Spectroscopy in Biopharmaceuticals From Development to Manufacturing," *Analytical and Bioanalytical Chemistry* 414 (2022): 969–991.

5. R. S. Senger, H. J. Flick, and J. L. Robertson, "Raman Spectroscopy and Chemometrics: A Potential Method for Life Science Applications," in *Applied Raman Spectroscopy*, ed. V. K. Singh (Elsevier, 2025), 123–147.

6. A. Graf, A. Woodhams, M. Nelson, et al., "Automated Data Generation for Raman Spectroscopy Calibrations in Multi-Parallel Mini Bioreactors," *Sensors* 22 (2022): 3397.

7. M. Klaverdijk, L. A. Smulders, M. Ottens, and M. E. Klijn, "Towards Rapid Calibration of Bioprocess Quantification Models Using Single Compound Raman Spectra: A Comparison of Four Approaches," *Biotechnology and Bioengineering* 123 (2026): 324–336.

8. J. Engel, J. Gerretzen, E. Szymańska, et al., "Breaking With Trends in Pre-Processing?," *TrAC Trends in Analytical Chemistry* 50 (2013): 96–106.

9. P. Oliveri, C. Malegori, R. Simonetti, and M. Casale, "The Impact of Signal Pre-Processing on the Final Interpretation of Analytical Outcomes – A Tutorial," *Analytica Chimica Acta* 1058 (2019): 9–17.

10. M. Klaverdijk, M. Nemati, M. Ottens, and M. E. Klijn, "Impact of Bioreactor Process Parameters and Yeast Biomass on Raman Spectra," *Biotechnology Progress* 41 (2025): e70050.

11. S.-H. Han, J. Park, K.-B. Lee, C.-H. Park, and D.-Y. Lee, "Variable-Specific Preprocessing Enables Cross-Scale Transferable Raman Models From High-Throughput Bioreactor Data," *Bioresources and Bioprocessing* 13 (2026): 122.

12. J. A. Iversen, R. W. Berg, and B. K. Ahring, "Quantitative Monitoring of Yeast Fermentation Using Raman Spectroscopy," *Analytical and Bioanalytical Chemistry* 406 (2014): 4911–4919.

13. N. Yang, C. Guerin, N. Kokanyan, and P. Perré, "In-Line Monitoring of Bioreactor by Raman Spectroscopy: Direct Use of a Standard-Based Model Through Cell-Scattering Correction," *Journal of Biotechnology* 396 (2024): 41–52.

14. H. Jonsson and J. Gabrielsson, "Evaluation of Preprocessing Methods," in *Comprehensive Chemometrics*, vol. 2, ed. S. D. Brown, R. Tauler, and B. Walczak (Elsevier, 2009), 199–206.

15. R. Houhou and T. Bocklitz, "Trends in Artificial Intelligence, Machine Learning, and Chemometrics Applied to Chemical Data," *Analytical Science Advances* 2 (2021): 128–141.

16. J. Gerretzen, E. Szymańska, J. J. Jansen, et al., "Simple and Effective Way for Data Preprocessing Selection Based on Design of Experiments," *Analytical Chemistry* 87 (2015): 12096–12103.

17. H. A. Babatunde, O. M. McDougal, and T. Andersen, "Automated Spectral Preprocessing via Bayesian Optimization for Chemometric Analysis of Milk Constituents," *Food* 14 (2025): 2996.

18. B. Brunel, F. Alsamad, and O. Piot, "Toward Automated Machine Learning in Vibrational Spectroscopy: Use and Settings of Genetic Algorithms for Pre-Processing and Regression Optimization," *Chemometrics and Intelligent Laboratory Systems* 219 (2021): 104444.

19. C. D. Kappatou, J. Odgers, S. García-Muñoz, and R. Misener, "An Optimization Approach Coupling Preprocessing With Model Regression for Enhanced Chemometrics," *Industrial & Engineering Chemistry Research* 62 (2023): 6196.

20. A. Borg, M. Elhabiri, S. L. Calvé, and V. Portaluri, "Less Is More: Practical Insights Into Multivariate Regression Models for Raman Spectroscopy in Bioprocess Monitoring," *Biotechnology and Bioengineering* 123 (2026): 544.

21. A. Umprecht, V. Fonseca Diaz, B. Hüpfl, et al., "Unsupervised Optimization of Spectral Pre-Processing Selection to Achieve Transfer of Raman Calibration Models," *Measurement* 255 (2025): 117906.

22. J. Ezenarro, D. Schorn-García, O. Busto, and R. Boqué, "ProSpecTool: A MATLAB Toolbox for Spectral Preprocessing Selection," *Chemometrics and Intelligent Laboratory Systems* 247 (2024): 105096.

10974555, 0, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jrs.70223 by Technical University Delft, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

23. M. Klaverdijk, M. Ottens, and M. E. Klijn, "Single Compound Data Supplementation to Enhance Transferability of Fermentation Specific Raman Spectroscopy Models: M. Klaverdijk et al," *Analytical and Bioanalytical Chemistry* 417 (2025): 1873.

24. R. McCreery, *Raman Spectroscopy for Chemical Analysis* (John Wiley & Sons, Ltd, 2000).

25. M. Dudek, G. Zajac, E. Szafraniec, et al., "Raman Optical Activity and Raman Spectroscopy of Carbohydrates in Solution," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 206 (2019): 597–612.

26. P. Thompson, D. E. Cox, and J. B. Hastings, "Rietveld Refinement of Debye–Scherrer Synchrotron X-Ray Data From Al2O3," *Journal of Applied Crystallography* 20 (1987): 79–83.

27. X. Bian, K. Wang, E. Tan, P. Diwu, F. Zhang, and Y. Guo, "A Selective Ensemble Preprocessing Strategy for Near-Infrared Spectral Quantitative Analysis of Complex Samples," *Chemometrics and Intelligent Laboratory Systems* 197 (2020): 103916.

28. N. K. Afseth and A. Kohler, "Extended Multiplicative Signal Correction in Vibrational Spectroscopy, a Tutorial," *Chemometrics and Intelligent Laboratory Systems* 117 (2012): 92–99.

29. G. Derringer and R. Suich, "Simultaneous Optimization of Several Response Variables," *Journal of Quality Technology* 12 (1980): 214–219.

30. C. Pappas, P. Taranilis, M. Basalekou, et al., "Evaluation of a Raman Spectroscopic Method for the Determination of Alcohol Content in Greek Spirit Tsipouro," *Current Research in Nutrition and Food Science* 4 (2016): 01–09.

31. X. Sun, "Glucose Detection Through Surface-Enhanced Raman Spectroscopy: A Review," *Analytica Chimica Acta* 1206 (2022): 339226.

32. M. Kizilov, V. Cheburkanov, J. Harrington, and V. V. Yakovlev, "Modular and Automated Workflow for Streamlined Raman Signal Analysis," *Journal of Raman Specroscopy* 57 (2026): 360–377.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Table S1:** Reference peaks used in the evaluation policy. **Table S2:** Grid search configuration for spectral preprocessing pipeline selection, listing the algorithms and associated parameter ranges evaluated at each preprocessing step, together with the Python library implementing each algorithm. **Table S3:** Cross-validation metrics of random forest model (RF) trained with raw metrics as features to predict rRMSEP (%) of the calibration model applied to bioprocess spectra. The target value is expressed as a percentage, so the RF prediction error is reported in percentage points (pp), to avoid ambiguity between the two scales. **Table S4:** Cross-validation metrics of random forest (RF) model trained with selection steps as features to predict rRMSEP (%) of the calibration model applied to bioprocess spectra. The target value is expressed as a percentage, so the RF prediction error is reported in percentage points (pp), to avoid ambiguity between the two scales. **Figure S1:** Single-compound spectra preprocessed with preprocessing selected based on lowest rRMSEP to bioprocess data: (**A**) ethanol and (**B**) glucose. **Figure S2:** Mean raw bioprocess spectrum compared with the mean preprocessed bioprocess spectrum, using the preprocessing selected based on lowest rRMSEP for: (**A**) preprocessing selected for ethanol and (**B**) preprocessing selected for glucose. **Figure S3:** Mean raw process spectrum and mean fitted baseline for the preprocessing pipelines selected on the lowest rRMSEP, for (**A**) ethanol and (**B**) glucose. **Figure S4:** Single- compound spectra preprocessed with preprocessing selected through the multimetric selection framework to bioprocess data: (**A**) ethanol and (**B**) glucose. **Figure S5:** Mean raw bioprocess spectrum compared with the mean preprocessed bioprocess spectrum, using the preprocessing selected through the multimetric selection framework for (**A**) preprocessing selected for ethanol and (**B**) preprocessing selected for glucose. **Figure S6:** PLS loading vectors (LV1–LV3) for models trained on single-analyte calibration spectra containing ethanol alone (**A–C**) or glucose alone (**D–F**), preprocessed using the pipeline selected through the multimetric selection framework. **Table S5:** rRMSEP (%) and explained variance (%) per latent variable (LV) for ethanol and glucose calibration models trained with one analyte at a time, using preprocessing selected through the multimetric framework. **Table S6:** rRMSEP (%) and explained variance (%) per latent variable (LV) for ethanol and glucose calibration models trained on the biomass-inclusive calibration dataset from Klaverdijk et al., using preprocessing selected through the multimetric framework. **Figure S7:** PLS loading vectors (LV1–LV3) for models trained on the biomass-inclusive calibration dataset from Klaverdijk et al. preprocessed using the pipeline selected through the multimetric selection framework: ethanol (**A–C**) and glucose (**D–F**). **Figure S8:** Permutation importance of raw metrics for predicting rRMSEP, sorted by ethanol importance. **Figure S9:** Partial dependence plots (PDPs) of ethanol raw metrics. (**A**) F1 spectra, (**B**) selectivity, (**C**) FWHM, (**D**) pseudo-Voigt R^2, (**E**) baseline MAD, (**F**) SNR improvement, (**G**) $rRMSECV_{wq}$, and (**H**) F1 coefficient. **Figure S10:** Partial dependence plots (PDPs) of glucose raw metrics. (**A**) F1 spectra, (**B**) selectivity, (**C**) FWHM, (**D**) pseudo-Voigt R^2, (**E**) baseline MAD, (**F**) SNR improvement, (**G**) $rRMSECV_{wq}$, and (**H**) F1 coefficient. **Figure S11:** Partial dependence plots (PDPs) of ethanol selection steps. (**A**) passed selectivity, (**B**) passed SNR improvement, (**C**) passed pseudo-Voigt R^2, (**D**) passed $rRMSECV_{wq}$, (**E**) desirability F1 spectra, (**F**) desirability FWHM, (**G**) desirability MAD, and (**H**) desirability F1 coefficient. **Figure S12:** Partial dependence plots (PDPs) of glucose selection steps. (**A**) passed selectivity, (**B**) passed SNR improvement, (**C**) passed pseudo-Voigt R^2, (**D**) passed $rRMSECV_{wq}$, (**E**) desirability F1 spectra, (**F**) desirability FWHM, (**G**) desirability MAD, and (**H**) desirability F1 coefficient. **Figure S13:** Application of calibration models built with preprocessing selected through the multimetric framework without hard filter to bioprocess data. (**A,B**) glucose: (**A**) measured versus predicted values and (**B**) mean preprocessed bioprocess spectrum compared with regression coefficient. **Table S7:** Cross-validation and prediction statistics for the calibration model built with preprocessing pipeline selected through multimetrics framework without hard filter.

10974555, 0, Downloaded from https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jrs.70223 by Technical University Delft, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License