Towards Rapid Calibration of Bioprocess Quantification Models Using Single Compound Raman Spectra

A Comparison of Four Approaches

Authors
Klaverdijk, Maarten; Smulders, Lisa A.; Ottens, Marcel; Klijn, Marieke E.

DOI
10.1002/bit.70092

Publication date
2026

Document Version
Final published version

Published in
Biotechnology and Bioengineering

Citation (APA)
Klaverdijk, M., Smulders, L. A., Ottens, M., & Klijn, M. E. (2026). Towards Rapid Calibration of Bioprocess Quantification Models Using Single Compound Raman Spectra: A Comparison of Four Approaches. *Biotechnology and Bioengineering*, *123*(2), 324-336. https://doi.org/10.1002/bit.70092

Biotechnology and Bioengineering

WILEY

BIOTECHNOLOGY AND BIOENGINEERING

ARTICLE OPEN ACCESS

Towards Rapid Calibration of Bioprocess Quantification Models Using Single Compound Raman Spectra: A Comparison of Four Approaches

Maarten Klaverdijk | Lisa A. Smulders | Marcel Ottens | Marieke E. Klijn

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

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

Received: 11 August 2025 | **Accepted:** 17 October 2025

Funding: This project is funded by the Department of Biotechnology at Delft University of Technology.

Keywords: bioprocessing | data augmentation | in-line measurements | process analytical technology | Raman spectroscopy | *Saccharomyces cerevisiae*

ABSTRACT

In-line Raman spectroscopy combined with accurate quantification models can offer detailed real-time insights into a bioprocess by monitoring key process parameters. However, traditional approaches for model calibration require extensive data collection from multiple bioreactor runs, resulting in process-specific models that are sensitive to operational changes. These challenges can be tackled by simplifying experimental data generation or implementation of computational methods to obtain synthetic and augmented Raman spectra. In this study, we utilized a small experimental dataset of 16 single compound spectra to calibrate quantification models by using partial least squares (PLS) and indirect hard modeling (IHM), leading to comparable rRMSEP values for glucose (4.8% and 4.2%), ethanol (11.6% and 6.3%), and biomass (16.2% and 10.0%) when applied to yeast batch and fed-batch bioprocesses. Subsequently, isolated spectral features extracted during IHM were used to generate fully synthetic spectral datasets for PLS model calibration, resulting in rRMSEPs of 3.2% and 14.5% for glucose and ethanol, respectively. Finally, spectra from a single batch process were augmented with the same isolated spectral features, and calibration with these augmented spectra reduced rRMSEP by 18.6% point (glucose) and 4.3% point (ethanol) compared to process-only calibrated models. This study demonstrates how different approaches may support robust development and rapid implementation of Raman spectroscopy-based models while minimizing experimental efforts, where even complete independence of process data can be achieved.

1 | Introduction

Monitoring metabolite, product, and biomass concentration during a bioreactor process is often based on labor-intensive manual sampling and off-line sample analysis. Development and implementation of novel process analytical technology (PAT) aims to automate quantitative data collection on these process parameters to achieve hands-free real-time monitoring. In recent decades, optical PAT has seen a rise in popularity as it allows for in-line measurements that can be combined with automated data analysis (Esmonde-White et al. 2021). Specifically Raman spectroscopy is highly suitable for bioreactor processes due to its low signal interference from water and the ability to provide specific fingerprint signals for many relevant compounds. Raman spectroscopy is successfully implemented across a wide range of bioreactor processes, from microbial to animal cell cultures, with the goal to quantify both simple and complex target compounds (Tanemura et al. 2023). However, as the complexity of the measured systems increases, spectral features of all compounds in the system overlap in the singular

spectrum, hindering direct interpretation of the raw signal. Therefore, multivariate modeling techniques are employed to translate the complex spectral signal into quantitative data.

The most popular multivariate technique is partial least squares (PLS) regression and it is extensively used for a wide range of bioprocesses (Zavala-Ortiz et al. 2022). During PLS model calibration, spectral and reference value datasets are provided, and the model defines latent variables (LVs) that capture the most relevant spectral variations correlated to the target compound identity and abundance (Wold et al. 2001). This makes PLS a powerful technique for Raman spectral decomposition to predict the target compound concentration, while little spectral knowledge of the system is required. These models are traditionally calibrated with extensive datasets for which multiple bioreactor runs have to be performed, leading to labor-intensive data collection procedure. Unfortunately, the required time and material investment delays the adoption of Raman spectroscopy as PAT to generate valuable process insights in early-stage development of new processes or in R&D-based environments. Moreover, repeating the same bioreactor process to collect calibration data leads to a limited design space, as the relationship between process compounds remains similar for every process run. As a consequence, the PLS model will learn to predict compound concentrations only under the circumstances occurring in that specific process. This means models can be trained to identify and predict abundance based on unspecific spectral features when compounds have strong cross-correlations (e.g., correlations between substrates, inverse correlations between substrates and product or biomass). Such models will perform poorly when process conditions or process operation disturbs this cross-correlation. To calibrate robust PLS models that can deal with process variations and transfers to related processes, the calibration dataset should include variation outside of the standard process evolution. However, collecting datapoints from the process itself by repeating bioreactor runs with different concentration settings is inefficient, especially when it is solely for the improvement of a Raman spectroscopy-based PLS model. Experimental methods for introducing these variations in the dataset, such as spiking the process with the compound of interest or creating custom samples in a cell culture matrix, can enhance a model's specificity for target compounds (Romann et al. 2022; Santos et al. 2018). However, these approaches are labor-intensive, requiring careful experimental design and sample preparation.

Computational approaches can also offer a solution to build spectral datasets for robust model calibration. When extensive process knowledge is available and spectral composition of most individual process compounds is obtainable, methods such as indirect hard modeling (IHM) can be used. IHM is a physics-based approach that incorporates known spectral properties to extract chemical information from a process spectra (Alsmeyer et al. 2004). It differs from implicit modeling techniques, such as PLS, by explicitly modeling known spectral features of a compound with individual peak functions. Spectra of pure or dissolved compounds are deconvoluted by fitting Pseudo-Voigt peak profiles to the spectra until the residuals between the fitted model and experimental spectra are minimized. For mixtures containing a single unknown compound, the unknown spectral variation can be characterized through complemental hard modeling, allowing the spectral composition of the unknown compound to be extracted (Kriesten et al. 2008). The defined individual models can be combined into a mixture model, which is calibrated on training spectra by weighing the intensity of each compound model to minimize the fit residuals. Complemental hard modeling was successfully applied to chemical processes and yeast bioprocesses (Alsmeyer et al. 2004; Echtermeyer et al. 2021; Müller et al. 2023). When unexplained spectral residuals remain after optimization, the model can be expanded by extracting the unknown contribution, use it to develop a new hard model, and include it in the mixture model (Müller et al. 2023). These applications demonstrate that the IHM approach offers a flexible and low calibration effort approach for quantification from spectral data. Nevertheless, IHM requires a high level of spectral knowledge of the process and the availability of isolated spectral measurements of the major process compounds.

As effective calibration data is labor-intensive to collect, alternative methods by which spectral data can be obtained computationally are highly desired. Methods to artificially generate Raman spectra or modify existing spectral data can alleviate current limitations, such as data scarcity and low variability of the compounds of interest. Several automated methods for generating synthetic spectra are developed for classification problems, such as synthetic minority over-sampling technology (SMOTE) and generative adversarial networks (GANs) algorithms (Hao et al. 2023). The SMOTE algorithm interpolates between existing spectra of a minority-class to reduce class imbalance and to increase the diversity of a data set (Chawla et al. 2002). The GAN approach consists of a generator and discriminator model that go through adversarial training, where the generator model learns to generate realistic spectra while the discriminator attempts to recognize synthetic data (Hao et al. 2023; Goodfellow et al. 2020). Both methods can be used to expand small or imbalanced calibration datasets to improve the performance of classification models (Wu et al. 2021). However, the use of these algorithms to generate spectra for quantification model calibration is limited, as a physically accurate relationships between spectral intensities and compound concentrations is not guaranteed. Examples for the generation of synthetic spectra to enhance quantification models are rare. Goldrick et al. generated synthetic Raman spectra simulating penicillin fermentation by combining empirical baseline spectra with simulated characteristic compound peaks in the form of Gaussian shapes (Goldrick 2019). Sulub & Small employed a similar method to simulate near-infrared spectra to calibrate a PLS model for the prediction of glucose in mixture measurements (Sulub and Small 2007). While these studies highlight the potential of augmenting spectral data for quantification problems, the application of synthetic Raman spectra remains largely unexplored.

This study compares four approaches for utilizing single compound spectra to calibrate Raman spectroscopy quantification models, applicable for bioprocess monitoring or control purposes. These approaches simulate scenarios where little or no process data is available before operating a bioreactor process, aiming to build robust models and enable availability of quantification models before a new process begins. The first approach uses a small experimental spectral data set (16 spectra) containing

single compound measurements of glucose, ethanol, and *Saccharomyces cerevisiae* biomass acquired under bioreactor conditions to calibrate PLS models directly. Secondly, a calibrated mixture model is obtained through IHM, calibrated with the same small experimental spectral data set. Both models are validated on a bioprocess data set of 4 batches and a single fed-batch (65 spectra total) to assess model performance. In the third approach, isolated spectral features extracted obtained from IHM are used to generate synthetic spectra that simulate bioprocess conditions. This yielded a full factorial data set of 125 synthetic spectra to maximize spectral variation, with which PLS models were calibrated. In the fourth approach, isolated spectral features of glucose and ethanol from IHM were employed to augment a small spectral process batch bioprocess data set (12 spectra). It was aimed to improve model specificity towards these targets by artificially increasing the spectral feature variability, providing a data augmentation method for situations where process data is limited. To conclude, four approaches to leverage little to no process data for Raman spectroscopy-based quantification model calibration are compared in terms of prediction accuracy, calibration effort, and flexibility towards new compounds. By investigating these calibration approaches we contribute to rapid development of flexible and robust quantification models that can be obtained without running (additional) bioprocesses.

2 | Materials and Methods

2.1 | Experimental Methods

2.1.1 | Bioreactor Settings and Reference Sampling

The *Saccharomyces cerevisiae* strain CEN. PK113-7D was used for all bioprocesses (Nijkamp et al. 2012), and cultures were grown on defined medium containing 5 g/L $(NH_4)_2SO_4$, 3 g/L KH_2PO_4, and 0.5 g/L $MgSO_4.7H_2O$ corrected to a pH of 6.0 with 2 M KOH (Verduyn et al. 1992). After medium sterilization 50% glucose (J.T. Baker, Philipsburg, NJ) solution (in-house) was added until 20 g/L, and vitamins and trace elements were added through 0.2 μM syringe filters (Whatman, Maidstone, UK). The medium was completed by adding 0.2 g/L sterile Antifoam-C (BASF, Ludwigshafen, Germany). Bioprocess data was collected by operating 4 batches and a single fed-batch in a 2 L bioreactor system (Applikon, Delft, the Netherlands) using a 1 L working volume. The cultures were maintained at 30℃, stirred at 800 rpm, and aerated with 0.5 L/min of air by a Biostat B bioreactor controller (Sartorius, Göttingen, Germany). The pH setpoint of 6.0 was maintained by the automatic addition of 2 M KOH. The batch bioprocesses were inoculated at an OD660 of 0.3 and sampled until glucose depletion. The fed-batch started as a batch culture operated at identical settings, and was bolus fed with 50% glucose solution three times whenever glucose depleted to extend the process. The bioprocesses were sampled every hour, and sample supernatants were analyzed for their glucose and ethanol concentrations with an Agilent 1260 infinity HPLC (Agilent Technologies, CA) equipped with a Bio-RAD Aminex HPX-87H (300 × 7.8 mm) cation-exchange column (Bio-Rad, Hercules, CA). The biomass concentration of each sample was determined by measuring the OD660 values using a Libra S11 spectrophotometer (Biochrom, UK), and dry-weight determination was performed by loading and drying 10 mL of culture broth on nitrocellulose membrane filters (pore size: 0.45 μm); Gelman Laboratory, MI). An overview of the reference measurements for each bioprocess is shown in Supporting Information S1: Figure S5.1.3.

2.1.2 | Single Compound Measurements

The glucose, ethanol, and biomass single compound spectra were acquired in the same 2 L bioreactor system operated under identical temperature and aeration settings as the bioprocess (Section 2.1.1). For each compound, the bioreactor was filled with 1 L of defined media, and 5 concentrations values were achieved by adding 50% glucose solution (described in Section 2.1.1, 96% ethanol, and biomass obtained from a batch bioprocess and subsequently washed in defined media. The final spectral dataset with a total of 16 spectra consisted of a single defined media spectra followed by 5 glucose (50–250 mM), 5 ethanol (50–250 mM), and 5 biomass (0.8–5 g/L) spectra (Supporting Information S1: Figure S5.1.1A The concentrations of each step were verified with HPLC and dry-weight determination as described in Section 2.1.1.

2.1.3 | Raman Spectral Acquisition

Raman spectra were acquired using a RXN2 analyzer (Kaiser Optical Systems Inc., Ann Arbor, MI) equipped with a 400 mW 785 nm laser, which acquired spectra in the bioreactor system using a bIO-Optic immersion probe. Spectra were collected over a range of 100–3400 cm^{-1} with a resolution of 4 cm^{-1}. The immersion probe was mounted through the headplate of the 2 L bioreactor, and was autoclaved with the bioreactor for the bioprocesses. The spectroscope was set to continuously acquire spectra of 60 s that yielded a detector saturation between 30% and 58% over all performed measurements. The bioprocess spectra were averaged from two 60-second acquisitions to achieve a high monitoring resolution, while the single compound measurements were averaged from a total of ten 60-second spectra to obtain high quality spectra for modeling.

2.2 | Computational Methods

2.2.1 | PLS Model Calibration

All PLS models were developing in PLS_Toolbox version 9.3.8 (Eigenvector Research Inc., WA) running on Matlab R2023a (MathWorks, WA). All spectra were pre-processed by reducing variables to the fingerprint region of 700–1800 cm^{-1}, Automatic Whittaker filter baseline correction ($\lambda = 10000$, $\alpha = 0.001$), sulfate peak normalization, and mean centering. The reference values for glucose, ethanol, and biomass were mean centered before calibration. An individual model was generated for each compound of interest, and Venetian blinds cross-validation was used. The number of latent variables for each model was selected based on the elbow point of the root mean square error of calibration (RMSEC) and cross-validation (RMSECV) plots, and by inspecting the loadings of each latent variable to prevent the inclusion of spectral noise. Model performance across calibration datasets was compared by using the relative root mean

square error of prediction (rRMSEP) based on the interquartile range (IQR) shown in Equation 1:

$$rRMSEP = \frac{RMSEP}{Q3 - Q1} \times 100, \quad (1)$$

where *Q1* and *Q3* represent the first and third quartiles, respectively.

2.2.2 | Indirect Hard Modeling

Single compound hard models and mixture models used for quantification were generated in the PEAXACT version 5.9 (Aachen, Germany) spectroscopy software. Single compound spectra were reduced to the fingerprint region (700–1800 cm^{-1}), and corrected by sulfate peak normalization. The Complemental Hard Modeling (CHM) (Kriesten et al. 2008) approach was utilized to generate a hard model consisting of 7 peaks for defined media (Supporting Information S1: Figure S5.2.1). The defined media hard model was fitted into the highest concentration measurements of glucose, ethanol, and biomass, and Pseudo-Voigt profiles were fitted sequentially at the location with the highest residual error. This procedure was continued until the newly fitted peaks could not be verified with literature references of their Raman spectra. This resulted in models with 20 peaks for glucose, 8 peaks for ethanol, and 9 peaks for biomass. The four generated models were combined in a single bioprocess mixture model that was subsequently calibrated on the 16 single compound measurements by fitting each component to minimize the spectral residuals. For the reference concentrations, the weight of each component was balanced according to Equation 2:

$$1 = \omega_{DefinedMedia} + \omega_{Glucose} + \omega_{Ethanol} + \omega_{Biomass}, \quad (2)$$

The calibration procedure generated linear correlations between component weight and concentration (Supporting Information S1: Figure S5.2.3). During calibration and application, the model was only allowed to change the weights of each hard model, without accounting for peak shifts and shape changes. The performance of the IHM model was expressed in RMSEP and rRMSEP values (Equation 1) to allow for comparison with the PLS models.

2.2.3 | Synthetic Spectra Generation From Single Compounds

Synthetic Raman spectra simulating bioprocess conditions were generated using the Pseudo-Voigt profiles obtained during the IHM steps (Section 2.2.2). Individual peak parameters and linear correlations between peak intensity and concentration were extracted and re-combined into bioprocess spectra using an in-house Python script. Concentration ratios between glucose, ethanol, and biomass were designed according to a full factorial design of experiments (DoE) approach with five concentrations per compound, leading to a total of 125 combinations (Supporting Information S1: Figure S5.3.1). The concentration ratios were inserted in Equation 2 to extract the weight of defined media, and the synthetic spectra were generated by multiplying the Pseudo-Voigt features with the weights corresponding to the desired concentration according to the calibration lines (Supporting Information S1: Figure S5.2.3). A detailed workflow of all steps is presented in Supporting Information S1: Figure S5.3.2.

2.2.4 | Spectral Augmentation of Batch Bioprocess Data

The augmentation of Raman spectra from a single batch bioprocess was performed using the same spectral features and weight versus concentration calibrations as used during the generation of synthetic spectra (Section 2.2.3). Two augmented datasets were generated by adjusting the concentrations of (1) glucose and (2) ethanol, where ±10 and ±20 mM, respectively, around the original values was generated. This was done by adding and subtracting the Pseudo-Voigt profiles, with a boundary at a concentration of 0 mM. This resulted in two datasets consisting of 60 spectra (12 original batch process spectra and 48 spectra augmented with Pseudo-Voigt profiles), see Supporting Information S1: Figure S5.4.2. The detailed workflow of these steps is presented in Supporting Information S1: Figure S5.4.1.

3 | Results and Discussion

In this study we compare four approaches using simple measurements to calibrate Raman spectroscopy quantification models for monitoring key compounds during bioprocessing (glucose, ethanol, and biomass) in scenarios where no or limited bioprocess data is available. A bioprocess setup with a simple broth composition was selected as the target process for quantification, with the main process components being: defined media, glucose, ethanol, biomass, and low amounts of glycerol and acetate (abundance for glycerol and acetate was considered insignificant for modeling). A small data set of 16 single compound spectra was generated, consisting of one defined media spectrum and five concentrations of glucose, ethanol, and biomass each, plus their reference measurements (Supporting Information S1: Figure S5.1.1).

For the first approach, the data set of 16 single compound measurements was used to directly calibrate three PLS models for the quantification of glucose, ethanol, and biomass (Supporting Information S1: Figure S1.1, Section 3.1). In the second approach, single compound spectra were used to generate compound hard models (HMs) for defined media, glucose, ethanol, and biomass. The HMs were combined in a single mixture model with the IHM method (Supporting Information S1: Figure S1.2, Section 3.2). In the third approach, the HMs were used to generate de novo synthetic mixture spectra of custom concentration ratios (Supporting Information S1: Figure S1.3, Section 3.3). This approach allowed for the simulation of bioprocess conditions across the entire design space defined by the concentration ranges of the single compound measurements. In the fourth approach, isolated spectral features of glucose and ethanol were used to augment a small data set of a single batch process (12 spectra), by which the spectral variability for these compounds could be increased (Supporting Information S1: Figure S1.4). The performance of these four modeling approaches was validated using a bioprocess data set consisting of

multiple batch bioprocesses and a single fed-batch bioprocess to investigate quantitative accuracy on process data. An overview of all experimental datasets and the four modeling approaches is shown in Figure 1.

3.1 | Approach 1: PLS Model Calibration With Single Compound Spectra

The performance of PLS models calibrated directly using single compound spectra (16 samples) obtained under standard bioreactor conditions was investigated for three targets: glucose, ethanol, and biomass. These models were assessed using the validation data set of four batch processes and a single fed-batch bioprocess (65 samples), and the corresponding model performances, regression coefficient vectors (RCVs), and high concentration single compound spectra are shown in Figure 2.

Quantitative analysis of model performance resulted in rRMSEP values of 4.8%, 11.6%, and 16.2% for glucose, ethanol, and biomass, respectively. Three latent variables were selected for each model, resulting in low RMSEC and RMSECV values by capturing the variation of glucose, ethanol, and biomass in separate components (Supporting Information S1: Figure S5.1.3). Glucose concentrations were accurately quantified across the batches and the fed-batch bioprocess. Qualitative model assessment indicates that the RCV of each model contains the key spectral features of their compound of interest, while correcting for overlapping spectral features. Glucose model specificity is reflected by the strong representation of the peaks for COH-bending (918 cm^{-1} and 1125 cm^{-1}), CO-stretching (1066 cm^{-1}), and CH-bending (1368 cm^{-1}) (Dudek et al. 2019) in the RCV, and overlapping peaks of ethanol are corrected (e.g., 879 cm^{-1} and 1085 cm^{-1}). The ethanol model showed good prediction accuracy on the batch bioprocesses (rRMSEP of 6.74%), but the predictions on the fed-batch data deviated from the 1:1 line, leading to the overall rRMSEP of 11.6%. The ethanol model RCV closely resembles the single compound spectrum of ethanol, indicated by high coefficients for the CC-stretching (879 cm^{-1}), CO-stretching (1046 cm^{-1}), CH_3-rocking (1085 cm^{-1}), CH_2-twisting (1277 cm^{-1}), and CH_3-deformation (1456 cm^{-1}) peaks (Boyaci et al. 2012; Pappas et al. 2016). An inspection of the residuals on the fed-batch samples revealed multiple regions where true and fitted spectra deviated (1250–1480 cm^{-1} and 1560–1660 cm^{-1}), but the pattern could not be directly related to a known compound.

Our previous work showed that spectral features associated with the molecular composition of biomass can be detected with in-line Raman spectroscopy after correcting for the extinction effect with a normalization step (Klaverdijk 2025a). This is reflected by the biomass model RCV that contains spectral features matching with Raman spectroscopy studies of *S. cerevisiae*, displaying positive coefficients for bands related to phenylalanine (1002 cm^{-1}), phospholipids (1084 cm^{-1}), CH-deformation of proteins (1344 cm^{-1}), CH_2-deformation of lipids and proteins (1448 cm^{-1}), and amide I stretching (1669 cm^{-1}) (Wang et al. 2023). In the work of Yang et al. (2024), PLS models for monitoring yeast bioprocesses were calibrated with single compound spectra of glucose, ethanol, peptone, yeast extract, and biomass, but no specific signal was found for the yeast cells (Yang et al. 2024). Instead, signal extinction caused by biomass was modeled by measuring mixtures of glucose and ethanol at varying concentrations of biomass, and the nonlinear relation was used to quantify biomass and correct predictions of the PLS models during bioprocessing. Other literature on monitoring of yeast with Raman spectroscopy mainly highlight the attenuation of spectral features with increasing biomass concentration, and there is little data on the detection of its protein and lipid signal during in-line measurements.

The results in Figure 2 show an accurate prediction of the biomass concentration in three out of four batches, while the predictions for one batch and the fed-batch bioprocess deviated from the 1:1 line. Further inspection of the residuals of the deviating batch data set highlighted large differences in the water peak at 1640 cm^{-1}, but the cause of these differences could not be determined. For the fed-batch bioprocess, the biomass model had to extrapolate, as the single biomass spectra only reached 5 g/L while the fed-batch process went up to 9.2 g/L. The two lowest concentration biomass spectra from the calibration data set were overpredicted during model development, and the need for extrapolation on the fed-batch data could have propagated this effect outside of the calibration concentration range.

The latent variable loadings of each model show that the third latent variable did not contain more than 0.45% of the spectral variation (Supporting Information S1: Figure S5.1.3). The vast majority of the spectral variation belonging to glucose and ethanol is explained in latent variables one and two (loadings of 28.6%–70.9%), which may result from the difference in signal strength between biomass and the metabolites. To investigate the impact of including biomass spectra during calibration on overall model performance, models calibrated with only glucose and ethanol spectra were tested and applied to the same validation data set (Supporting Information S1: Figure S5.1.4). This resulted in models with two latent variables, where the rRMSEP of the glucose model without biomass in the calibration data set increased to 13.4% (from 4.8%) and the rRMSEP of the ethanol model increased to 14.8% (from 11.6%). The RCVs show that including biomass spectra in the calibration data set allows the model to correct for overlapping spectral features (e.g., 1448 cm^{-1} and 1669 cm^{-1}). Moreover, the broad features of both glucose and biomass overlap over a large section of the spectra. Although these effects are less visible in the RCVs of the ethanol models due to narrow peaks, the inclusion of biomass spectra and the subsequent selection of an additional latent variable led to a higher predictive performance. Thus, despite the small magnitude of the biomass spectral signal, including biomass spectra improved spectral decomposition of the molecular features of yeast and increased prediction performance on bioprocess data.

The performance decrease seen for the fed-batch samples to quantify ethanol and biomass could also be related to the preprocessing strategy. Single compound spectra were acquired in individual bioreactor setups, which resulted in baseline offsets between the experiments (Supporting Information S1: Figure S5.1.1). An Automatic Whittaker baseline correction ($\lambda = 10000$, $\alpha = 0.001$) was utilized to achieve baseline alignment. After baseline correction, spectra needed to be normalized for intensity and a normalization based on the sulfate peak of the synthetic medium as an internal standard yielded the best

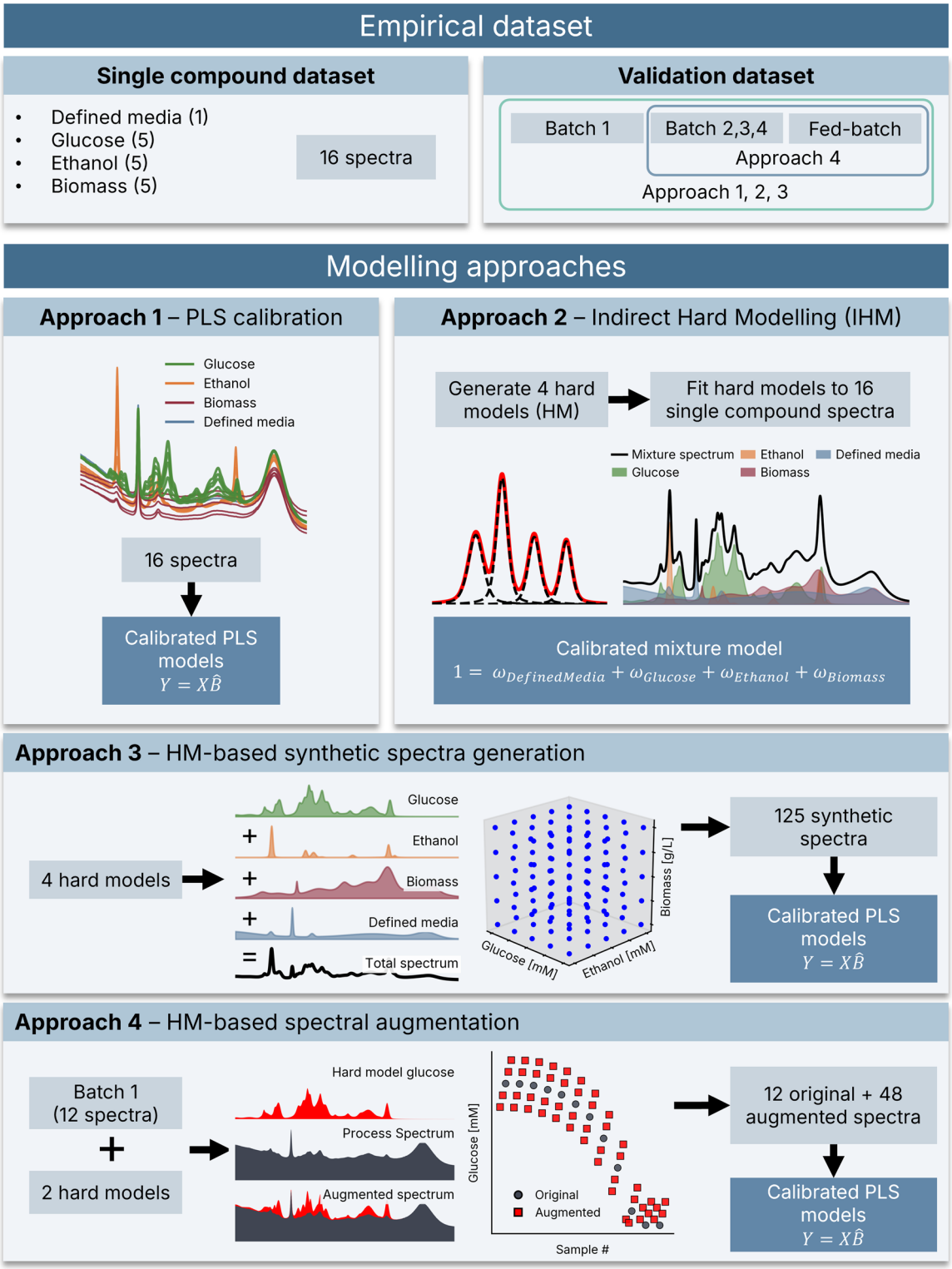

FIGURE 1 | An overview of the experimental datasets (top row) and the four modeling approaches performed in this study. For Approach 1: 16 single compound spectra were used to calibrate Partial Least Squares (PLS) regression models for glucose, ethanol, and biomass directly. Approach 2: the single compound spectrum of defined media, and the highest concentration spectra of glucose, ethanol, and biomass were used to generate four hard models (HM) by fitting Pseudo-Voigt features. These HMs were combined in a mixture model that was calibrated on the full single compound data set to establish relations between signal intensity and compound concentration. Approach 3: HMs of glucose, ethanol, biomass, and defined media and their respective intensity/concentration calibrations were used to generate synthetic Raman spectra with custom concentrations according to a full factorial Design of Experiments with 5 concentrations per quantification target. The resulting data set of 125 synthetic spectra was used to calibrate PLS models for glucose, ethanol, and biomass. Approach 4: a single batch spectral data set (Batch 1) was augmented with the HMs of glucose and ethanol to increase the spectral variability of each quantification target, resulting in two augmented datasets of 60 spectra used to calibrate PLS models for glucose and ethanol. All PLS models and the mixture model obtained with IHM were validated on a data set consisting of multiple batch bioprocesses and a single fed-batch bioprocess.

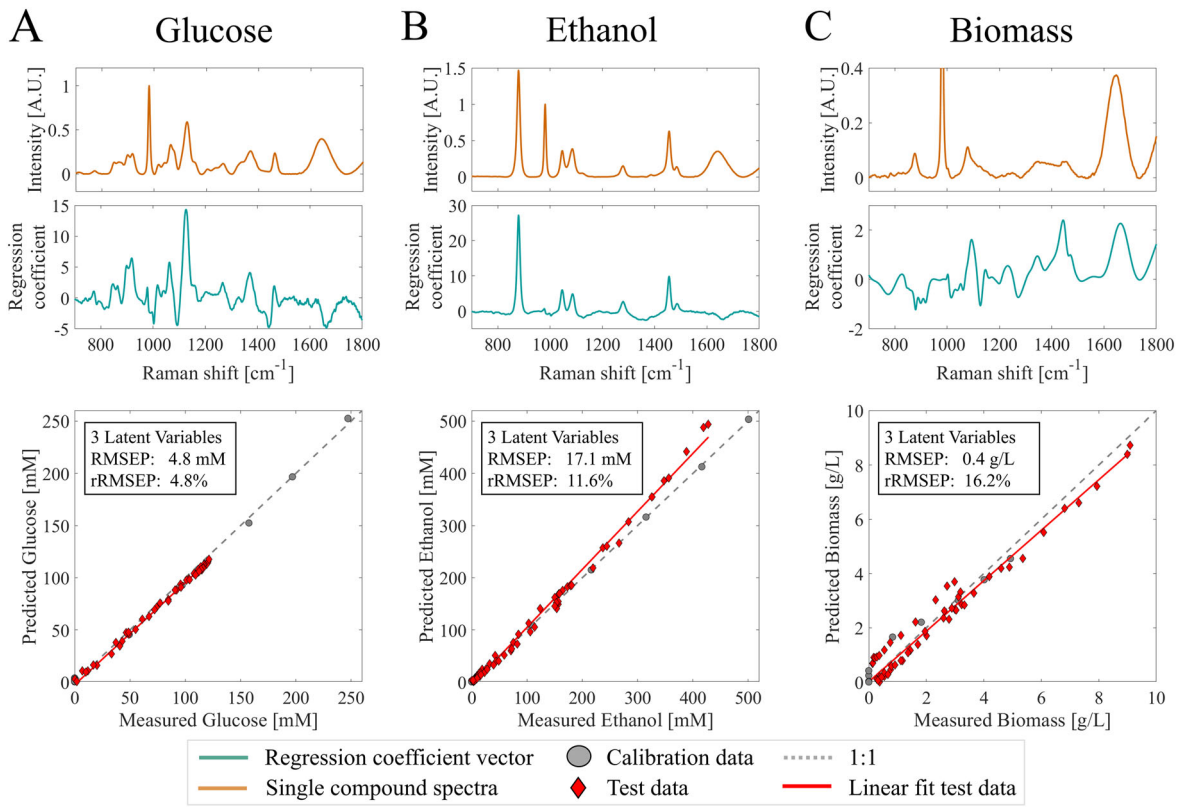

FIGURE 2 | Pre-processed spectra (before mean-centering, top), regression coefficient vector (middle), and measured versus predicted plots (bottom) of the glucose (A), ethanol (B), and biomass (C) Partial Least Squares (PLS) models, respectively.

results. However, the fed-batch was bolus fed with 50% sterile glucose solution three times, slightly diluting the sulfate peak, thereby compromising the intensity correction. This underlines the downside of utilizing internal standards for intensity normalization, as process adjustments can directly influence preprocessing accuracy. Despite these challenges, the sulfate peak normalization provided the most accurate models, and other normalization methods (e.g., standard normal variate) resulted in high prediction errors.

3.2 | Approach 2: IHM Calibrated With Single Compound Spectra

The 16 single compound spectra were used to generate HMs, where individual spectral features of each compound are modeled as Pseudo-Voigt profiles. To prevent the inclusion of noise into the HMs, fitted peaks were cross-checked with literature, resulting in 21 peaks for glucose (Dudek et al. 2019), 8 peaks for ethanol (Boyaci et al. 2012; Pappas et al. 2016), and 9 peaks for biomass (Wang et al. 2023). The fitted single compound models showed a high similarity to other work in literature using the IHM approach (Müller et al. 2023). The individual HMs were combined to form a mixture model, which was calibrated on the single compound data set. The calibrated mixture model was subsequently applied to evaluate performance with the bioprocess validation data set (Figure 3). A detailed overview of the workflow is provided in Supporting Information S1: Figure S5.2.2.

Quantitative model assessment resulted in rRMSEP values of 4.2%, 6.3%, and 10.0% for glucose, ethanol, and biomass, respectively. The predictions showed high linearity for glucose and ethanol, with a slight overprediction of glucose. The prediction accuracy for biomass was considered decent for the batch bioprocesses, but the predictions for the late fed-batch samples deviated from the 1:1 line, where the model had to extrapolate past the 5 g/L upper limit of the calibration data. Another factor leading to the decrease in prediction accuracy for biomass in the late fed-batch samples may be the broad spectral features with low specificity obtained using the complemental hard modeling approach, even though the absolute position of each fitted peak closely matched the features associated with *S. cerevisiae* reported in literature (Wang et al. 2023).

The IHM approach was successfully applied by Muller et al. to monitor glucose and ethanol concentrations during yeast bioprocessing in a 20 mL cuvette setup, where Raman spectra were acquired with a Raman microscope through the bottom of the glass cuvette (Müller et al. 2023, 2024). A total of 11 mixture spectra plus a single measurement of yeast suspension were used to calibrate their model, and glucose and ethanol concentrations of around 100 g/L and 50 g/L, respectively were successfully quantified. Our model was calibrated without the need for mixture spectra, and glucose and ethanol concentrations only reached 21 g/L and 23 g/L, respectively, thus resulting in a weaker Raman signal. Despite the inherent differences between experimental setups and lower concentration ranges, our mixture model accuracy is in the same order of magnitude, as our RMSEPs for glucose and ethanol were 0.74 mg/g and 0.43 mg/g versus their 3.68 mg/g and 1.70 mg/g. It should be noted that the use of an immersion probe inside the bioreactor led to high signal extinction by biomass, and despite the signal

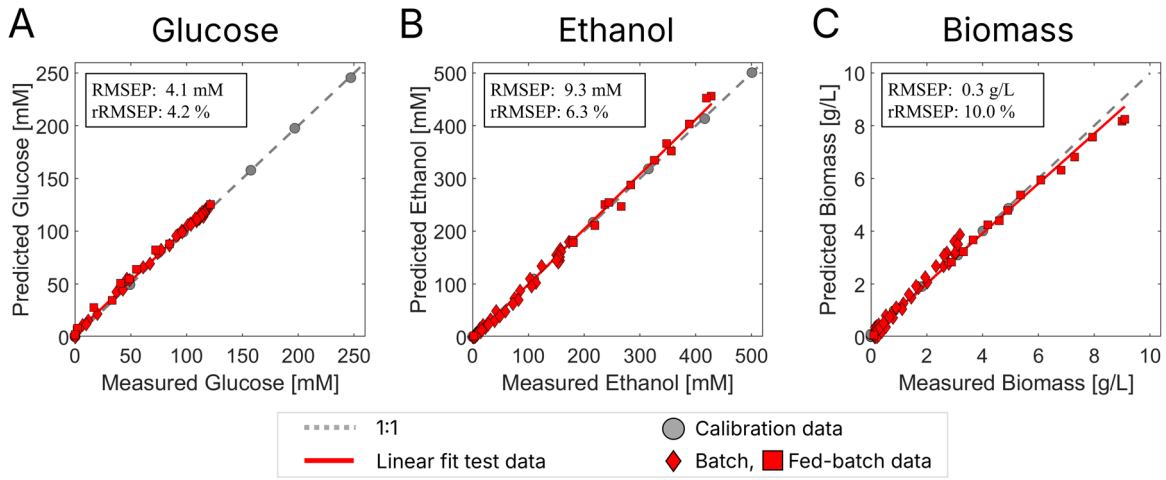

FIGURE 3 | The measured (*x*-axis) versus predicted (*y*-axis) plots of an Indirect Hard Modeling (IHM) model applied to a bioprocess data set consisting of 4 batches and a single fed-batch. The model quantified glucose (A), ethanol (B), and biomass (C).

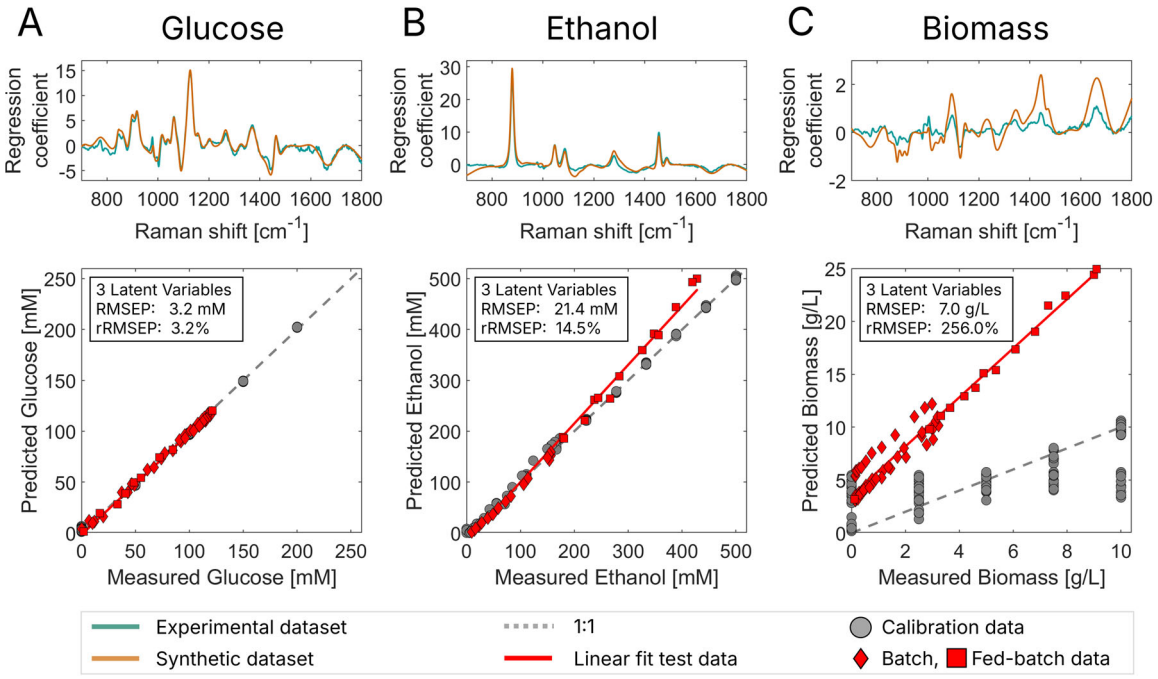

FIGURE 4 | The regression coefficient vectors (RCV, top) of the partial least squares (PLS) model calibrated with experimental data (orange), calibrated with synthetic spectra (blue), and the measured versus predicted plots (bottom) of the PLS models calibrated on synthetic spectra for glucose (A), ethanol (B), and biomass (C).

attenuation our model provided accurate predictions after a simple normalization step.

3.3 | Approach 3: PLS Model Calibration With Synthetic Spectra

In this section, we evaluate PLS model performance when calibrated with synthetic spectra. The HMs obtained in the previous section were extracted and utilized to generate de novo Raman spectra of custom concentration ratios. A total of 125 synthetic spectra were generated according to a full factorial design, with ranges of 0–200 mM for glucose, 0–500 mM for ethanol, and 0–5 g/L for biomass, including five concentration steps for each compound (Supporting Information S1: Figure S5.3.1). The synthetic data set was subsequently used to calibrate PLS models for the quantification of glucose, ethanol, and biomass, and applied to predict four batch and one fed-batch bioprocess datasets (Figure 4).

PLS models calibrated on the data set of 125 synthetic spectra resulted in rRMSEP values of 3.2%, 14.5%, and 256.0% for glucose, ethanol, and biomass, respectively. The synthetic spectra managed to simulate spectral variation of glucose closely, resulting in a more accurate prediction than direct calibration with the 16 experimental single compound spectra (1.6% point). The RCVs of the glucose models calibrated on experimental and synthetic data were similar, with the synthetic model

containing less noise due to the smooth nature of the Pseudo-Voigt profiles. The ethanol model calibrated on synthetic spectra had a slightly higher (2.9% point) prediction error than the model calibrated on the experimental single compound spectra, but the RCVs were still considered highly similar between the models. The RCV similarity seen for ethanol models also resulted in a comparable deviation for the prediction of ethanol concentrations in the fed-batch dataset.

The poor performance of the biomass model clearly indicates that synthetic spectra cannot properly replicate the spectral variation caused by biomass. The broad spectral features of biomass extracted from the complemental hard modeling method did not accurately represent the true spectral contribution of *S. cerevisiae*, and this propagated to the synthetic spectra. Furthermore, spectral variation caused by biomass was a small percentage of the total spectral variation, as seen in the PLS models calibrated with experimental single compound spectra directly (Section 3.1). The small magnitude of the biomass signal possibly caused a high sensitivity to small deviations in intensity, leading to difficulties of recreating the proper signal proportions. In addition, the broad shapes of the simulated biomass signal is sensitive to intensity changes by baseline correction steps.

This approach shows synthetic spectra allow for setting custom ratios between glucose and ethanol, by which the design space could be covered entirely without additional experimental effort. However, calibration with synthetic spectra did not lead to improved model performance for ethanol and biomass compared to calibration with 16 experimental single compound spectra. Furthermore, the use of synthetic spectra only led to very minor differences in the model RCVs for glucose and ethanol, supporting the lack of added benefit in terms of model performance or specificity. Nevertheless, this approach demonstrated the ability to expand a dataset with spectra highly similar to the process conditions, while maintaining the linear correlation between signal intensity and compound concentration. If this method can be expanded with HMs of additional compounds, it can generate high variability datasets without the need of collecting process data.

3.4 | Approach 4: PLS Model Calibration With Augmented Process Spectra

Calibrating quantification models with (repeated) process data limits the design space, which may lead to incorporation of cross-correlations and hinders model robustness (Klaverdijk et al. 2025b). However, generating process spectra de novo as discussed in Section 3.3 is limited by the availability of HMs for all process compounds. In many applications, process knowledge is minimal and single compound spectra can only be acquired for a few compounds. This section investigates the augmentation of a small dataset of process spectra with spectral features of the compound of interest obtained from HMs. This approach utilizes the standard spectral variation of a small process dataset (a single batch) while attempting to improve the specificity of models towards a compound of interest, without needing to define other process compounds. A batch data set of 12 spectra was expanded by synthetically modifying the concentration of either glucose or ethanol, up to a total of 60 spectra (Supporting Information S1: Figure S5.4.2). Augmenting spectra with the isolated biomass features was not considered because these features exhibited low specificity in Section 3.3. PLS models were calibrated with the standard (12 spectra) and augmented (60 spectra) data set, and applied to the reduced validation data set (3 batches, 1 fed-batch). The prediction performance of these models on bioprocess data is shown in Figure 5.

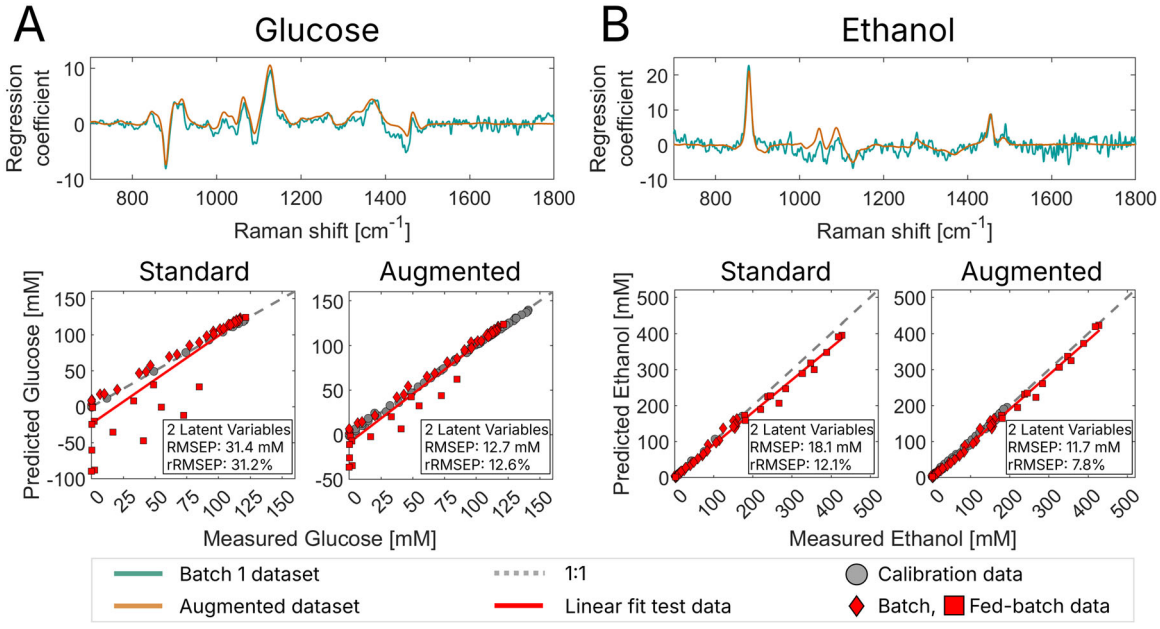

FIGURE 5 | The regression coefficient vectors (RCV, top) of the partial least squares (PLS) model calibrated with augmented data (orange), calibrated with batch process spectra (blue), and the measured versus predicted plots (bottom) of the PLS models. The performance of PLS models quantifying glucose (A) and ethanol (B) while calibrated on a standard batch data set (12 process spectra) and augmented batch spectra (12 process spectra and 48 augmented spectra). The calibration datasets were expanded by augmenting the spectral features of the quantification target by ± 10 and ± 20 mM.

A total of two latent variables was selected for all four models based on the RMSEC and RMSECV values and the noise levels present in the third latent variable (data not shown). Compared to the calibration with a single batch, increasing the number of calibration datapoints through augmentation reduced noise in the RCVs. Calibrating PLS models with the standard batch data set of only 12 spectra resulted in rRMSEPs of 31.2% and 12.1% for glucose and ethanol, respectively. The prediction performance of the standard models on the three validation batch bioprocesses was considered high (rRMSEP of 6.34% and 3.91% for glucose and ethanol, respectively), but accuracy decreased for the fed-batch process. Calibration with data from an exponential batch process led to strong cross-correlations between spectral features in the model, and performance was disrupted when the ratios between process compounds changed during the fed-batch mode of operation (Klaverdijk et al. 2025b). The performance for glucose and ethanol quantification improved by augmenting the calibration data set and increasing variability in the data set, improving the rRMSEPs for glucose and ethanol to 12.6% and 7.8%, respectively. Despite the overall reduction in prediction error, the augmented glucose model still performed poorly on the fed-batch data. Comparing the RCV with those of the PLS models from calibrated on experimental single compound (Section 3.1) and synthetic spectra (Section 3.2) indicates that the model lost its ability to compensate for the overlapping biomass signal. Including a third latent variable also did not improve the model's ability to extract spectral variation for biomass (data not shown). This means that correcting for overlapping biomass features seems essential for accurately predicting glucose concentration with high specificity (also highlighted in Supporting Information S1: Figure 5.1.4). The ethanol model seem less affected, as the augmented ethanol model did not suffer a similar performance decrease for the fed-batch data. This corresponds to earlier observations, where it was assumed that the fewer and narrower spectral features, as seen in the ethanol spectrum, are less affected by the biomass spectral features.

Despite the improvements in model performance and increased specificity of the RCVs, the loadings of latent variable 1 did not show large differences from those of the standard models and mainly captured batch evolution (Supporting Information S1: Figure 5.4.3). This is expected, as the augmented datasets followed an identical concentration trend to the standard batch bioprocess and the spectra were only slightly adjusted in concentration (Supporting Information S1: Figure 5.4.1). Modifying process spectra with concentrations at the extremes of the process design space to break cross-correlations (e.g., adjusting a true concentration of 20 mM to 140 mM) led to large prediction errors and nonlinear effects. The errors in these cases could be attributed to spectral normalization, where we did not manage to equally normalize process spectra and synthetic spectral features before combining the two components in augmented spectra. As a result, biomass extinction effects in the batch spectra propagate throughout the augmentation process, underlining how augmentation approaches are highly sensitive to small intensity changes not related to concentration changes, as these disrupt signal linearity.

This section showcased how Pseudo-Voigt profiles isolated from single compound spectra can be utilized to customize the concentration of specific compounds in process spectra. This allows for expansion of a data set's design space while minimizing the impact on other spectral features in the data. The augmentation approach can be used to artificially spike concentrations of compounds of interest without the need for extensive experimental setups where cell cultures cannot be recovered after spiking. However, to optimize augmentation methods, adaptive normalization techniques are necessary that transfer across spectra without the need for internal standards.

3.5 | Discussion on Modeling Approaches

This section discusses the four modeling approaches demonstrated in this study. A comparison of the quantitative performance of each model, the calibration data used, and the complexity of each modeling approach is shown in Table 1. Time-evolution plots of the model predictions from all four approaches are provided in Supporting Information S1: Figures S5.5.1–5.5.4.

Approach 1 showed that PLS models calibrated using 16 experimentally obtained single compound spectra can achieve decent prediction performance without the need of collecting mixture (process) spectra. This approach resulted in compound specific models, reflected by distinct peaks in the model RCVs. Moreover, models obtained with Approach 1 outperformed models calibrated on batch process data supplemented with single compound spectra for biomass prediction (rRMSEP of 27.0%), as demonstrated in previous work (Klaverdijk et al. 2025b). The observed improvement for biomass quantification resulted from a higher quality of single-compound biomass measurements, which were consistent with spectral features reported in literature (Klaverdijk 2025a; Wang et al. 2023). Despite these improvements, the implicit PLS models are not expected to perform well outside of their calibration design space, as the performance relies on empirical relationships learned from training data and the models do not contain physical understanding of the monitored process. From this perspective, the semi-explicit IHM used in Approach 2 has some inherent advantages over implicit modeling techniques. Once individual HMs of the main process compounds are available, a mixture model can perform predictions based on chemical principles. Moreover, the baseline itself is defined as a process component (consisting of mainly water), thereby reducing the dependency on spectral pre-processing. In addition, IHM also offers more flexibility in situations where a novel or unknown compound is present. For example, spectral variation of glycerol and acetate was assumed negligible for this application, but when such a compound becomes more abundant in the process, its single compound spectra could be used to generate a hard model which is subsequently added to the mixture model. An alternative route would be fitting Pseudo-Voigt features to the residuals between the old mixture model and new process spectra to generate the model of an additional component (Müller et al. 2023). Approach 2 is therefore considered highly flexible, as a database of hard models can be easily expanded with new compounds and can be calibrated based on simple measurements. In addition, where PLS models rely on learned weights at specific wavenumbers, the IHM approach can be tuned to allow peak shifts and shape changes

TABLE 1 | Comparison of the four modeling approaches discussed in this study. The relative root mean square error of prediction (rRMSEP) on the validation dataset is provided for each model and compound to directly compare quantitative performance between the methods.

Approach	Model type	Calibration data set		Experimental time	Computational time	Model flexibility	rRMSEP		
		Total spectra	Spectral source				Glucose	Ethanol	Biomass
1	PLS	16	Experimental	Low	Low	Medium	4.8%	11.6%	16.2%
2	IHM	16	Experimental	Low	Medium	High	4.2%	6.3%	10.0%
3	PLS	125	Synthetic	Low	High	Medium	3.2%	14.5%	256.0%
4	PLS	12 + 48	Process + Augmented	High	High	Low	12.6%	7.8%	—

during component fitting, leading to higher model robustness for changing measurement conditions or compound interactions. This tunability could also make quantification models perform better under extrapolation conditions as long as the number of process compounds does not change. However, these advantages come with increased modeling complexity, as HMs must be generated for all major process compounds, and the final mixture model must fit each HM to the process spectra for each prediction. This means that Approach 2 can become challenging for complex process mixtures where process knowledge and access to single compound spectra is limited. It is therefore important to note that the strength of PLS models lies in their simplicity and their ability to extract key spectral features of the target compound from complex spectra, thereby reducing the need for extensive process knowledge. This is considered beneficial for cell cultures with more complex media (e.g., for Chinese Hamster Ovary cells), where the number of relevant process compounds increases rapidly, and gaining complete spectral knowledge is challenging.

Calibrating PLS models with 125 synthetic mixture spectra during Approach 3 did not result in model improvements when compared to calibration with only the 16 experimental single compound spectra. Biomass prediction accuracy was particularly poor, likely due to the low specificity of the biomass HM. In addition, PLS models can benefit from calibration with experimental mixture spectra as interactions between compounds could influence the position and shape of their spectral features. The synthetic spectra generated in this study did not include these interactions, as each compound was modeled from single compound spectra. However, when compared to automated methods for spectra generation such as GANs or SMOTE, our approach can maintain physically accurate linear relationships between signal and concentration, provided that two key assumptions are met. First, signal intensity must change linearly with compound concentration withing the calibration range, supported by our calibration lines based on five single compound spectra. Second, scaled HMs for individual compounds must combine additively to represent mixture spectra (Equation 2). Under these conditions, our method allows synthetic spectra generation for any concentration ratio within the single compound spectra measured range.

Another aspect that was considered challenging was matching the intensity between synthetic, augmented, and process spectra. We exclusively utilized the sulfate peak at 981 cm^{-1} as an internal standard for normalization, which is also reported for the application of Raman spectroscopy for other yeast bioprocesses (Picard et al. 2007; Hirsch et al. 2019). However, using internal standards for normalization should be done with great caution as they are dependent on a single variable, and therefore sensitive to changes in measurement conditions (Yang et al. 2024). Furthermore, internal standards are not available in every measurement matrix, and alternative normalization methods should be explored when generating synthetic data. In this study, a robust method for normalizing the intensity of individual spectral components during spectra synthesis and augmentation was not found. Moreover, disruptions in signal linearity might occur at every spectral modification step, including normalization, and errors in intensity can propagate to the final spectrum.

The concept of Approach 4, which allows the artificial adjustment of spectral features related to target compounds, holds potential for complex processes where both process data availability and knowledge are limited. Since single compound spectra for common quantification targets (e.g., metabolites and products) can be easily measured, their key spectral features can be extracted to build HMs. The HMs can subsequently be used to enhance the spectral variation of this target compound within complex mixtures, even in situations where detailed process knowledge is missing. However, maintaining linearity between signal intensity and compound concentration is essential for linear regression techniques like PLS, but this is often disrupted by noise and scattering effects present in bioreactor spectra (Klaverdijk 2025a).

Despite the challenges highlighted in this study, the ability to generate synthetic and augmented spectra that accurately simulate process conditions can be valuable for calibrating quantification models for Raman spectroscopy. One of the largest hurdles for calibrating robust quantification models is the need for extensive data collection, especially capturing process states outside typical operational patterns, which can be crucial for improving model accuracy. The operation of bioreactor processes at different compound concentrations could provide valuable spectral information, but requires substantial time and material if solely performed for improving Raman spectroscopy quantification models. In addition, literature reports studies that investigated the effectiveness of compound spiking to generate this valuable data, but this typically leads to the loss of a cell culture (Santos et al. 2018). The options to generate these valuable conditions synthetically or to augment existing process spectra towards the edges of the desired design space could provide efficient and low-effort alternatives.

4 | Conclusion

Raman spectroscopy coupled with accurate quantification models serves as a powerful tool for monitoring bioreactor processes. Nevertheless, quantification model calibration is often labor-intensive and requires extensive experimental efforts. Furthermore, collecting large process datasets to calibrate these models often results in process-specific models with a narrow design space, highlighting the need for flexible methods to collect data and expand small process datasets. This study investigated four approaches by which a small dataset of 16 single compound measurements could be utilized to calibrate quantification models for glucose, ethanol, and biomass during *S. cerevisiae* bioprocesses.

The single compound dataset was used to calibrate quantification models using PLS (Approach 1) and IHM (Approach 2). Both modeling approaches showed similar performance when comparing the rRMSEP values for glucose (4.8% and 4.2%), ethanol (11.6% and 6.3%), and biomass (16.2% and 10.0%). The PLS approach demonstrated how isolated biomass measurements incorporate spectral features associated with the molecular composition of yeast, while the IHM approach proved to be a robust and flexible method that can be easily extended to accommodate new process compounds or conditions.

The compound hard models were also applied to synthetically generate Raman spectra to directly calibrate PLS models (Approach 3) and to augment experimental process data to increase model specificity (Approach 4). Direct calibration with synthetic spectra proved effective for glucose and ethanol quantification PLS models, with rRMSEP values of 3.2% and 14.5%, respectively. Due to difficulties in isolating sharp spectral features for biomass, calibration of PLS models with synthetic spectra did not result in accurate biomass quantification. Spectral augmentation of a single batch bioprocess data set led to rRMSEPs of 12.6% and 7.8% for glucose and ethanol, respectively, compared to 31.2% and 12.1% when calibrated solely on the standard batch data. The synthetic generation and augmentation of Raman spectra showed potential for the enhanced calibration of PLS models, but robust normalization steps are required to maintain signal integrity during these processes.

Overall, this study showcased multiple approaches by which simple spectral measurements can be applied to calibrate quantification models for bioprocesses, without the need for (additional) process data. This means that quantification models for yeast bioprocesses can be developed even before running the actual process, and models can be easily adapted to changes in process conditions or when transferring between processes. Furthermore, the possibility to augment existing spectra of complex processes enables model calibration improvement without extensive spectral knowledge of the system. The use of single compound, synthetic, or augmented Raman spectra supports efficient quantification model calibration, thereby simplifying the implementation of Raman spectroscopy for bioreactor monitoring.

Author Contributions

Maarten Klaverdijk: conceptualization, methodology, validation, formal analysis, investigation, software, writing – original draft, writing – review and editing, visualization. **Lisa Smulders:** investigation, software, writing - review and editing. **Marcel Ottens:** writing – review and editing, supervision. **Marieke Klijn:** conceptualization, resources, writing - original draft, writing – review and editing, supervision, administration, funding acquisition.

Acknowledgments

We would like to thank Christiaan Mooiman and Jeroen Schmidt for their technical support while setting up the lab and during experiments. This project is funded by the Department of Biotechnology at Delft University of Technology.

Conflicts of Interest

The authors declare no conflicts of interest.

References

Alsmeyer, F., H.-J. Koß, and W. Marquardt. 2004. "Indirect Spectral Hard Modeling for the Analysis of Reactive and Interacting Mixtures." *Applied Spectroscopy* 58, no. 8: 975–985.

Boyaci, I. H., H. E. Genis, B. Guven, U. Tamer, and N. Alper. 2012. "A Novel Method for Quantification of Ethanol and Methanol in Distilled Alcoholic Beverages Using Raman Spectroscopy." *Journal of Raman Spectroscopy* 43, no. 8: 1171–1176.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. "SMOTE: Synthetic Minority Over-Sampling Technique." *Journal of Artificial Intelligence Research* 16: 321–357.

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

Echtermeyer, A., C. Marks, A. Mitsos, and J. Viell. 2021. "Inline Raman Spectroscopy and Indirect Hard Modeling for Concentration Monitoring of Dissociated Acid Species." *Applied Spectroscopy* 75, no. 5: 506–519.

Esmonde-White, K. A., M. Cuellar, and I. R. Lewis. 2021. "The Role of Raman Spectroscopy in Biopharmaceuticals From Development to Manufacturing." *Analytical and Bioanalytical Chemistry* 414: 1–23.

Goldrick, S., et al. 2019. "Modern Day Monitoring and Control Challenges Outlined on an Industrial-Scale Benchmark Fermentation Process." *Computers & Chemical Engineering* 130: 106471.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, et al. 2020. "Generative Adversarial Networks." *Communications of the ACM* 63, no. 11: 139–144.

Hao, Y., X. Li, and C. Zhang. 2023. "Improving Prediction Model Robustness With Virtual Sample Construction for Near-Infrared Spectra Analysis." *Analytica Chimica Acta* 1279: 341763.

Hirsch, E., H. Pataki, J. Domján, et al. 2019. "Inline Noninvasive Raman Monitoring and Feedback Control of Glucose Concentration During Ethanol Fermentation." *Biotechnology Progress* 35, no. 5: e2848.

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

Klaverdijk, M., M. Ottens, and M. E. Klijn. 2025b. "Single Compound Data Supplementation to Enhance Transferability of Fermentation Specific Raman Spectroscopy Models." *Analytical and Bioanalytical Chemistry* 417: 1–12.

Kriesten, E., D. Mayer, F. Alsmeyer, C. B. Minnich, L. Greiner, and W. Marquardt. 2008. "Identification of Unknown Pure Component Spectra by Indirect Hard Modeling." *Chemometrics and Intelligent Laboratory Systems* 93, no. 2: 108–119.

Müller, D. H., M. Börger, J. Thien, and H. J. Koß. 2024. "Bioprocess In-Line Monitoring and Control Using Raman Spectroscopy and Indirect Hard Modeling (IHM)." *Biotechnology and Bioengineering* 121, no. 7: 2225–2233.

Müller, D. H., C. Flake, T. Brands, and H. J. Koß. 2023. "Bioprocess In-Line Monitoring Using Raman Spectroscopy and Indirect Hard Modeling (IHM): A Simple Calibration Yields a Robust Model." *Biotechnology and Bioengineering* 120: 1857–1868.

Nijkamp, J. F., M. van den Broek, E. Datema, et al. 2012. "De Novo Sequencing, Assembly and Analysis of the Genome of the Laboratory Strain *Saccharomyces cerevisiae* CEN. PK113-7D, a Model for Modern Industrial Biotechnology." *Microbial Cell Factories* 11, no. 1: 36.

Pappas, C., B. Marianthi, E. Konstantinou, et al. 2016. "Evaluation of a Raman Spectroscopic Method for the Determination of Alcohol Content in Greek Spirit Tsipouro." *Current Research in Nutrition and Food Science Journal* 4, no. Special Issue Nutrition in Conference October 2016: 01–09.

Picard, A., I. Daniel, G. Montagnac, and P. Oger. 2007. "In Situ Monitoring by Quantitative Raman Spectroscopy of Alcoholic Fermentation by *Saccharomyces cerevisiae* under High Pressure." *Extremophiles* 11, no. 3: 445–452.

Romann, P., J. Kolar, D. Tobler, C. Herwig, J. M. Bielser, and T. K. Villiger. 2022. "Advancing Raman Model Calibration for Perfusion Bioprocesses Using Spiked Harvest Libraries." *Biotechnology Journal* 17: 2200184.

Santos, R. M., J. M. Kessler, P. Salou, J. C. Menezes, and A. Peinado. 2018. "Monitoring Mab Cultivations With In-Situ Raman Spectroscopy: The Influence of Spectral Selectivity on Calibration Models and Industrial Use as Reliable PAT Tool." *Biotechnology Progress* 34, no. 3: 659–670.

Sulub, Y., and G. W. Small. 2007. "Spectral Simulation Methodology for Calibration Transfer of Near-Infrared Spectra." *Applied Spectroscopy* 61, no. 4: 406–413.

Tanemura, H., R. Kitamura, Y. Yamada, M. Hoshino, H. Kakihara, and K. Nonaka. 2023. "Comprehensive Modeling of Cell Culture Profile Using Raman Spectroscopy and Machine Learning." *Scientific Reports* 13, no. 1: 21805.

Verduyn, C., E. Postma, W. A. Scheffers, and J. P. Van Dijken. 1992. "Effect of Benzoic Acid on Metabolic Fluxes in Yeasts: A Continuous-Culture Study on the Regulation of Respiration and Alcoholic Fermentation." *Yeast* 8, no. 7: 501–517.

Wang, K., J. Chen, J. Martiniuk, et al. 2023. "Species Identification and Strain Discrimination of Fermentation Yeasts *Saccharomyces cerevisiae* and Saccharomyces Uvarum Using Raman Spectroscopy and Convolutional Neural Networks." *Applied and Environmental Microbiology* 89, no. 12: e01673-23.

Wold, S., M. Sjöström, and L. Eriksson. 2001. "PLS-Regression: A Basic Tool of Chemometrics." *Chemometrics and Intelligent Laboratory Systems* 58, no. 2: 109–130.

Wu, M., S. Wang, S. Pan, A. C. Terentis, J. Strasswimmer, and X. Zhu. 2021. "Deep Learning Data Augmentation for Raman Spectroscopy Cancer Tissue Classification." *Scientific Reports* 11, no. 1: 23842.

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

Zavala-Ortiz, D. A., A. Denner, M. G. Aguilar-Uscanga, A. Marc, B. Ebel, and E. Guedon. 2022. "Comparison of Partial Least Square, Artificial Neural Network, and Support Vector Regressions for Real-Time Monitoring of CHO Cell Culture Processes Using In Situ Near-Infrared Spectroscopy." *Biotechnology and Bioengineering* 119, no. 2: 535–549.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.
Supplementary Compound Raman Spectra a Comparison of Four Approaches.