Journal of Raman Spectroscopy

WILEY

Journal of RAMAN SPECTROSCOPY

RESEARCH ARTICLE OPEN ACCESS

Uncertainty Quantification Using Ensemble Learning and Monte Carlo Sampling for Performance Prediction and Monitoring in Cell Culture Processes

Thanh Tung Khuat[1] | Robert Bassett[2] | Ellen Otte[2] | Bogdan Gabrys[1]

[1]Complex Adaptive Systems Laboratory, The Data Science Institute, University of Technology Sydney, Ultimo, New South Wales, Australia | [2]CSL Innovation, Melbourne, Victoria, Australia

Correspondence: Thanh Tung Khuat (thanhtung.khuat@uts.edu.au)

Received: 16 August 2024 | **Revised:** 28 December 2024 | **Accepted:** 24 March 2025

Funding: This research was supported under the Australian Research Council's Industrial Transformation Research Program (ITRP) funding scheme (Project Number IH210100051).

Keywords: cell culture processes | machine learning | Raman spectroscopy | real-time monitoring | uncertainty quantification

ABSTRACT

Biopharmaceutical products, particularly monoclonal antibodies (mAbs), have gained prominence in the pharmaceutical market due to their high specificity and efficacy. As these products are projected to constitute a substantial portion of global pharmaceutical sales, the application of machine learning models in mAb development and manufacturing is gaining momentum. This paper addresses the critical need for uncertainty quantification in machine learning predictions, particularly in scenarios with limited training data. Leveraging ensemble learning and Monte Carlo simulations, our proposed method generates additional input samples to enhance the robustness of the model in small training datasets. We evaluate the efficacy of our approach through two case studies: predicting antibody concentrations in advance and real-time monitoring of glucose concentrations during bioreactor runs using Raman spectra data. Our findings demonstrate the effectiveness of the proposed method in estimating the uncertainty levels associated with process performance predictions and facilitating real-time decision-making in biopharmaceutical manufacturing. This contribution not only introduces a novel approach for uncertainty quantification but also provides insights into overcoming challenges posed by small training datasets in bioprocess development. The evaluation demonstrates the effectiveness of our method in addressing key challenges related to uncertainty estimation within upstream cell cultivation, illustrating its potential impact on enhancing process control and product quality in the dynamic field of biopharmaceuticals.

1 | Introduction

In recent years, biopharmaceutical products, including monoclonal antibodies (mAbs) and therapeutic proteins derived from biological organisms for the treatment or prevention of diseases, have emerged as top-selling drugs in the pharmaceutical market [1]. This trend is attributed to their numerous advantages, such as high specificity and activity [2]. As global pharmaceutical sales are projected to surpass $1 trillion by 2026, biopharmaceutical products are expected to contribute significantly, constituting 37% of the total sales, an increase from 30% in 2020 [3]. By 2026, over half of the top 100 best-selling medications are anticipated to be biologics. Within the realm of biological products, mAbs stand out as the forefront runners in the swiftly expanding market of high-value biologics [4]. The mAb products are manufactured through biotechnological processes within living systems, including microorganisms, plants, animals or human cells such as Chinese hamster ovary (CHO) cells, mouse myeloma (NS0), baby hamster kidney (BHK), human embryo kidney

(HEK-293) and human retinal cells [5]. Hence, the cultivation and harvesting of cells responsible for producing the active pharmaceutical ingredient [6] play a crucial role in facilitating the growth and reproduction of cells in quantities sufficient to meet production demands [7]. Continuing this trajectory, the utilisation of machine learning models across different phases of mAb development and manufacturing, including the prediction and monitoring of biophysical properties, cell growth, nutrient, metabolite and protein concentrations throughout bioreactor cell cultivation processes, is not only gaining popularity but also improving in accuracy [8].

According to Kelley [9], the attention in cell culture process development has been shifted from solely pursuing the elevated titres to emphasising the control of product quality and process consistency at every stage of the development and across all production scales. Therefore, it is crucial to monitor the changes in the culture operating parameters which include physical, chemical and biological parameters. Physical parameters encompass factors such as temperature, gas flow rate and agitation speed, while chemical parameters encompass dissolved oxygen and carbon dioxide levels, pH, osmolality, nutrient and metabolite concentrations. Biological parameters are employed to assess the physiological condition of the culture and include metrics such as viable cell concentration, viability and a range of intracellular and extracellular measurements [10]. Optimising culture operating parameters is essential to attain high product expression while maintaining acceptable product quality profiles. This purpose can be achieved by monitoring the relationships among process variables and extracting valuable knowledge from bioprocess data using machine learning models to gain novel insights into the interdependence between critical process parameters (CPPs) and critical quality attributes (CQAs) in biopharmaceutical process development and manufacturing. To construct bioprocess datasets and calibrate the machine learning models, it is essential to measure process parameters during the cell culture process within bioreactors. These process parameters can be measured either online or offline with operator intervention. Examples of offline measurements include pH (often for verification of online pH readings), cell counting, viability measurements, osmolality and specific metabolite and product concentrations. Metabolites in cell culture, such as glucose, lactate, glutamine and glutamate, are typically assessed offline using enzymatic biosensors designed for each specific analyte [10]. These measurements play a crucial role not only in sustaining substrate levels above critical thresholds through feeding strategies but also in formulating processes with minimised by-product formation.

Commercially available autosamplers and integrated multifunctional offline analysers, such as the BioProfile FLEX, have typically been used for offline monitoring of metabolite levels, osmolality, pH, dissolved gases and measurements of sodium, potassium and calcium. This is done as a replacement for manual sampling, which is often labour-intensive and can introduce operator-dependent errors into the process [11]. Although autosamplers and analysers can obtain a high accuracy, analyses in [12] showed that the coefficient of variation among different measurement times ranges from 3% to 8% for each process parameter. This fact reflects the uncertainty of input features and target variables when building predictive models. To assist in making control decisions based on the predictive outcomes of machine learning models, it is necessary to provide uncertainty levels associated with each predictive value. In the context of the regression problem addressed in this paper, the variance in predictive outcomes for an input query can serve as a meaningful indicator of uncertainty. A comprehensive review paper conducted by Hullermeier and Waegeman [13] categorised uncertainty sources into aleatoric (data dependent and noise induced) and epistemic (model dependent) uncertainties. Aleatoric uncertainty, often termed as the irreducible component of uncertainty, is associated with randomness, which is the variability in the outcome of an experiment due to inherently random effects. This type of uncertainty cannot be mitigated through model enhancements. Instead, reducing aleatoric uncertainty is achievable through improvements in the data, such as incorporating repeat measurements or eliminating erroneous entries [14]. On the contrary, epistemic uncertainty represents the reducible uncertainty stemming from insufficient knowledge about the optimal model, and it can be diminished through model enhancements [15]. This type of uncertainty can be further divided into uncertainties arising from the selection of the model (including architecture, representations and features) and the ambiguity in parameter optimisation once a model is selected.

Numerous methods for quantifying uncertainty in predictive outcomes of machine learning models are available in the literature, as outlined in [13]. Among them, the two most popular groups are ensemble learning and Bayesian methods [8]. While Bayesian methods such as Gaussian process (GP) regression focus mainly on quantification and reducing the epistemic uncertainty, ensemble methods aim to estimate the impact of the aleatoric uncertainty due to the use of sampling techniques on the input data.

This paper focuses on introducing a general framework for uncertainty quantification applicable to any regressors, especially in the context of small training data, using ensemble learning. In situations involving limited training data and the presence of noise in input features, the conventional approach of developing multiple base learners through bootstrap resampling from input spaces in ensemble learning becomes inefficient. To estimate the impact of the aleatoric uncertainty on the prediction accuracy in small training datasets, we propose the use of Monte Carlo simulations to generate additional input samples by considering available training features as mean values. The additional instances will be used to train base learners. The effectiveness of the proposed method will be assessed in the context of predicting and monitoring the performance of process parameters during upstream cell culture bioreactor runs. One of the inherent challenges in cell culture processes pertains to the limitation of available data, commonly referred to as the small data issue. This limitation arises from the scarcity of process data for emerging bioproducts, with instances where only one or two production runs are conducted for a novel product at manufacturing sites [16, 17]. The substantial cost associated with cell culture processes further exacerbates this issue, as conducting bioreactor runs for new cell lines or experimental variations (e.g., novel base media or feeding strategies) is

economically constrained. Additionally, the practical necessity of relocating products across different production sites to accommodate various products and their life cycles contributes to what is known as the small training data problem. This operational characteristics result in a limited number of historical experiments available at new manufacturing facilities. The adoption of process analytical technology (PAT) tools in bioprocess development and manufacturing steps has facilitated the real-time collection of extensive and diverse measurements and information. This can lead to the availability of thousands of input features, for example, each Raman spectrum from the spectrometer can contain thousands of spectrum variables (e.g., wave numbers) considered as input features. Meanwhile, the number of experiments (i.e., samples) is limited. Consequently, these circumstances give rise to a low-N problem, wherein the number of training samples is considerably smaller than the number of input data dimensions. This inherent disparity poses a considerable challenge for machine learning algorithms, when the number of training samples is inadequate relative to the number of input features.

By generating new values for input features and the target variable based on their actual values, along with a coefficient of variation associated with each input feature, we can overcome the shortage issue of training data when training an ensemble model of multiple base learners. In our proposed method, we will use the standard deviation value of the predictive outcomes of all base learners as an indicator of the uncertainty level for each predicted value. In short, our novel contribution can be summarised as follows:

1. We introduce a comprehensive framework for assessing the uncertainty level linked to each predictive value through the utilisation of ensemble learning of regressors in tandem with Monte Carlo sampling. Our proposed method is designed to address the challenge of limited training data, a factor that can affect the effectiveness of traditional ensemble learning approaches. Moreover, our method represents the general framework employing ensemble learning in conjunction with Monte Carlo simulations to quantify the uncertainty level associated with each predictive outcome, particularly in scenarios characterised by a shortage of training data.
2. The effectiveness of the proposed method is evaluated through its application to two prominent challenges in upstream cell cultivation within bioreactors. The first problem involves the early prediction of antibody concentrations 1 day in advance, utilising solely current offline measurements as input features. The second problem entails real-time monitoring of glucose concentrations throughout the bioreactor run of a cell culture process, employing Raman spectra as input features.

The subsequent sections of this paper are organised as follows. Section 2 provides an introduction to the key features of the proposed method, including its application in predicting mAb concentrations in a cell culture process and in real-time monitoring of glucose concentrations within bioreactors using Raman spectral data. Section 3 shows the results of the proposed approach in addressing the challenge of predicting process performance in cell culture bioreactors 1 day in advance, using solely offline process measurements of the bioreactors and in addressing the real-time monitoring problem of glucose concentrations during bioreactor runs, using Raman spectra data as input features. Finally, the concluding remarks of this paper will be presented in Section 4.

2 | Methodology

2.1 | General Framework for Uncertainty Estimation of Predictive Values

Because of the errors linked with offline measurements, as discussed in the Introduction section, relying solely on a single predicted value generated by machine learning models for each set of input features proves inadequate for making informed decisions during the cell culture process. This limitation becomes evident, for instance, in scenarios where it may not suffice to determine crucial actions such as deciding when to add glucose or terminate the cell cultivation process. It becomes challenging to assess the accuracy of a predicted value without considering the associated uncertainty range. With predictions that include uncertainty values, we gain knowledge of possible minimum and maximum values associated with each prediction. Hence, it is preferable to have results presented in the form of $\hat{y} \pm 2 \cdot \sigma$, where $\hat{y}$ represents the predicted value and σ signifies the standard deviation of the prediction. The prediction values may not follow a normal distribution, so $2 \cdot \sigma$ will be used to derive confidence limits, providing approximately 95% certainty that the actually observed values will fall within the prediction range. This section outlines a method to achieve this objective by employing an ensemble of regressors and Monte Carlo sampling on both input and output spaces to construct training sets.

Let $\mathbf{X} = [X_1, X2, \ldots, X_m]$ be a set of m input samples, where each sample $X_k = (x_{k1}, x_{k1}, \ldots, x_{kn})$ $(k \in [1, m])$ includes n input features, and $\mathbf{Y} = [Y_1, Y2, \ldots, Y_m]$ $(Y_k = \{y_k\}, y_k \in \mathbb{R}, k \in [1, m])$ be outputs corresponding to input samples. We need to build a regressor $\mathbb{F}(\mathbf{X}) \rightarrow \hat{\mathbf{Y}}$ such that minimises $||\mathbf{Y} - \hat{\mathbf{Y}}||$ value. The output of the regressor $\mathbb{F}$ for each unseen input sample X_T will be in the form of $\hat{y}_T \pm \sigma_T$. In this case, $\hat{y}_T$ is the average predictive value of N base learners within the ensemble model computed by Equation (1), while σ_T represents the standard deviation value of N predictive values given in Equation (2).

$$\hat{y}(X_T) = \frac{1}{N} \cdot \sum_{i=1}^{N} \hat{y}_i(X_T) \tag{1}$$

where $\hat{y}_i(X_T)$ is the predicted value of the i th base regressor for an unseen input sample X_T within the ensemble model, which comprises N base regressors. The standard deviation of the ensemble prediction for each data point (X_T) can be employed to establish confidence intervals and uncertainty bounds. The utilised standard deviation is the unbiased standard deviation derived from the predictions of an individual base model for each data point:

$$\sigma(X_T) = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i(X_T) - \hat{y}(X_T))^2}{N-1}} \tag{2}$$

To train N base regressors of an ensemble model, we need to generate N training sets, each for training a base regressor. We will not use resampling or random sampling without replacement from original training set to create N training subsets because the training dataset is small in size, so the traditional sampling techniques may not generate sufficient diversity in the training subsets for the base regressors within the ensemble model. In addition, each input value is subjected to the errors due to the variation in measurements resulting from the final accuracy of offline analysers. As a result, in our proposed method, the Monte Carlo sampling method is employed to generate N random values for each input feature x_{kj} ($k \in [1, m], j \in [1, n]$) and each target value y_k satisfying a Gaussian distribution requirement with mean being an actual value (x_{kj} or y_k) and given standard deviation value. In a mathematical form, we generate N random values from a Gaussian distribution $\{x_{kj}^{(1)}, \ldots, x_{kj}^{(N)}\} = \mathbb{G}(\mu_{kj} = x_{kj}, \sigma_{kj}, N)$ for each input feature x_{kj} and $\{y_k^{(1)}, \ldots, y_k^{(N)}\} = \mathbb{G}(\mu_k = y_k, \sigma_k, N)$ for each target value y_k, where σ_{kj} and σ_k are the standard deviations for each input feature x_{kj} and target value y_k, respectively. These standard deviation values are computed from corresponding actual values and coefficient of variation (δ) of each input feature as follow: $\sigma_{kj} = \delta_j \times x_{kj}$ and $\sigma_k = \delta_Y \times y_k$, where δ_j is the maximum coefficient of variation of feature j, while δ_Y is the maximum coefficient of variation of output variable $\mathbf{Y}$. After generating N samples for all input and output values, we will concatenate all values $x_{kj}^{(i)}$ and $y_k^{(i)}$ at the i th position ($i \in [1, N]$) to create the i th training set $\mathbf{X}^{(i)}$ and $\mathbf{Y}^{(i)}$ in order to train the i th base regressor within the ensemble model. The fundamental steps of the proposed framework are presented in Figure 1.

To optimise the hyperparameters of the ensemble models including the number of base regressors N and the hyperparameters of base regressors, we need a separate validation set and find the set of hyperparameters resulting in the minimum average value of errors over N_{val} samples in the validation sets. For instance, we can seek for a given set of hyperparameters to obtain the minimum mean absolute error (MAE) of the ensemble model with N base regressors:

$$MAE = \frac{\sum_{k=1}^{N_{val}} \left(\left|y(X_k) - \frac{1}{N} \cdot \sum_{i=1}^{N} \hat{y}_i(X_k)\right|\right)}{N_{val}} \quad (3)$$

where $y(X_k)$ is the target value of the k th sample in the validation set and $\hat{y}_i(X_k)$ is the prediction of the i th base regressor in the ensemble model for a validation sample X_k. Although our proposed framework can be used for any regression models as base learners, this work only illustrates the empirical outcomes for two typical types of regression models for small datasets. The first model is the support vector regression (SVR), which is one of the most often used machine learning algorithms for small datasets [18]. The second model is the partial least squares regression (PLSR), which is very popular for small data [19] and dominates in machine learning models applied for predicting a variety of process performance issues in cell culture processes [8].

2.2 | Performance Prediction of a Cell Culture Process

CHO cells are commonly employed for the production of mAbs. The cell culture procedure involves a sequence of scale-up and expansion stages designed to yield a sufficient cell mass for the inoculation of the production bioreactors. This process includes additional cell growth, mAb production, the elimination of cellular mass from the bioreactor material through centrifugation and a three-stage filtration process resulting in the acquisition of clarified material [2]. In the course of biopharmaceutical process development, it is crucial to enhance titre aiming to reduce manufacturing expenses but still maintaining consistent quality attributes, safety and efficacy of therapeutic proteins. Throughout this development phase, continuously enhancing upstream titre will play a critical role in increasing the output and reducing the costs of production [20]. The optimisation of nutrient concentrations, including amino acids, vitamins and trace metals, is widely recognised as a crucial factor for enhancing the protein production [21]. In addition, the cell concentration and viability play a pivotal role in the development of cell culture processes. These measurements are essential for assessing the culture

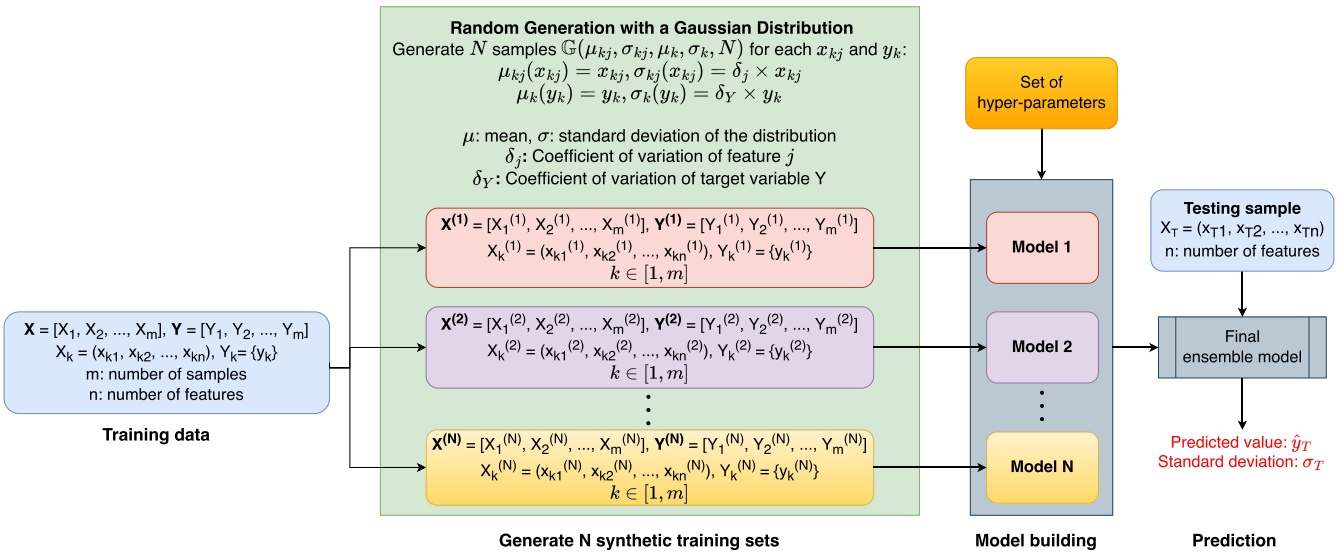


FIGURE 1 | General framework for estimating uncertainty levels of predictive values using ensemble learning and Monte Carlo sampling.

physiology in response to operating conditions, calculating growth rates, specific consumption/production rates of metabolites and determining cell-specific productivity [10]. Therefore, there is a high expectation of making early predictions of mAb concentration in the upcoming culture days based on current values of tens of offline measurements, such as operational conditions, nutrients and metabolite concentrations monitored over time. Accurate predictions will contribute to adjusting the culture environments and nutrient compositions to increase the product concentration. The effectiveness of the proposed framework presented in Section 2.1 will be evaluated in predicting mAb concentration 1-day-ahead using values of offline measurements at the current day.

2.2.1 | Dataset

The dataset utilised in predicting mAb concentration 1-day-ahead was derived from AstraZeneca's upstream process development and production databases, encompassing various antibody products employing CHO cell lines, as detailed in [22]. The dataset included information from 106 cultures, reflecting a diverse operational scale ranging from bench-top (5-L volume) to manufacturing (500-L volume), spanning a period of 7 years from 2010 to 2016. Each culture involved the recording of over 20 offline parameters for up to 17 days, including culture days; elapsed culture time (ECT); viable cell density (VCD); total cell density (TCD); pH; cell viability; elapsed generation number (EGN); average cell volume (ACV); osmolality; average cell compactness (ACC); average cell diameter (ACD); cumulative population doubling level (CPDL); concentrations of glucose, lactate, ammonium, glutamine, glutamate, sodium, potassium and bicarbonate; temperature; pCO_2; pO_2; monomer content of the final product; and product (mAb) concentration. The time-series dataset has been normalised to safeguard proprietary rights.

We have used all offline measurements in the original dataset as input features and created a new target variable, which is the mAb concentration of the next culture day. We have also removed the sample corresponding to the last culture day as there is no value in the target variable.

2.2.2 | Learning Procedures

To address the problem of 1-day-head mAb concentration prediction, we employed two types of regression models, namely, the PLSR and the SVR, as base learners within the ensemble model for small-sized datasets shown in Figure 1. The dataset, extracted from [22], lacks information regarding the coefficient of variation for each offline measurement. Consequently, a fixed value of $\delta_j = \delta_Y = 0.05$ was used for all offline measurements serving as input features and the target variable. In addition to the ensemble models of PLSR and SVR, we conducted separate training for PLSR, SVR and GP Regression as competing models for performance comparison. The GP regression has notable advantages such as lower data requirements and more importantly the possibility to assess the uncertainty of predictions [15]. The implementation of PLSR, SVR and GP was taken from the scikit-learn library [23].

We will assess the obtained performance of all learning models based on errors through a fivefold group cross-validation. The set of 106 cultures will be partitioned into five folds, with each fold encompassing data from all culture days within a specific culture. Four folds will serve for training, while the remaining fold will be designated as testing data. This process will be iterated five times, with each bioreactor used once in a testing fold. The average error of the trained models across the five testing folds will be employed for comparing the performance among competing models. For each training fold, we employed a hyperparameter optimisation procedure using the Bayesian optimisation approach within the Optuna library [24]. This optimisation involved 50 iterations and fivefold cross-validation to determine the optimal hyperparameter settings before training the model on the respective training fold. The potential range of hyperparameter values for each regression model is detailed in Table S1. It is noted that all individual regressors within the ensemble model will use the same hyperparameter setting.

2.3 | Real-Time Monitoring of Glucose Concentrations Within Bioreactors Using Raman Spectra Data

Currently, monitoring the cell culture profile during production involves taking small samples of medium components and metabolites at specific culture points, which are then quantified using a bioanalyser [25]. However, this sampling process presents challenges, including potential effects on the culture volume and the risk of microbial contamination. Additionally, the limited number of sampling points makes it challenging to acquire data at high frequencies [25]. As a result, various PAT methods have been developed to enable continuous analysis [8]. For instance, Raman spectrometers and near-infrared spectroscopy can offer information on components in the culture solution, while capacitance-based measurements allow for cellular concentration analysis [26]. The application of Raman spectrometers for the continuous acquisition of various culture data in cultivation processes enables real-time monitoring of cell growth, nutrient and metabolite concentrations. This marks a crucial step towards implementing feedback controls for culture conditions. For instance, a Raman-based glucose feedback control mechanism can enhance overall bioreactor health, product output and product quality [27]. Moreover, real-time monitoring of culture components may expedite faster medium development by continuously optimising a broader range of components [25].

In this study, we will evaluate the efficacy of various ML models in real-time monitoring of glucose concentrations within bioreactors throughout the cell culture process, using only Raman spectra data as input features. The actual outputs of the target variable corresponding to input Raman spectra will be based on offline glucose measurements. Periodically sampled offline glucose concentrations will be analysed by bioanalysers such as Nova Biomedical BioProfile FLEX Analyser. As a result, the target variable will exhibit a coefficient of variation. Meanwhile, the input Raman data are high dimensional, and the relationships and dependencies among input wave numbers (features) and between all input features and the output variable are complex and usually non-linear. Therefore, the value of the coefficient of variation for each input Raman feature is typically unknown.

Consequently, this scenario differs from the issue discussed in Section 2.2, so it requires modifications to the framework outlined in Section 2.1.

2.3.1 | Dataset

To address the problem of real-time monitoring of glucose concentrations within bioreactors using Raman spectral data, we used a dataset extracted from the cell culture process within CSL Innovation Pty Ltd. The dataset includes the historical culture data of three 5-L bioreactors (A1, A2 and A3) taking place in 2 weeks. All of the three bioreactors used the same culture media (base and feed), but A3 used a different cell line expressing a different product than A1 and A2, which were the same cell line and product.

Raman spectra were acquired within the bioreactor using a Kaiser Raman Rxn2 analyser equipped with a 785-nm excitation laser and a probe. The spectra were collected in the Raman shift range of 100–3425.0 cm^{-1}. On average, approximately four spectra were recorded per hour during bioreactor runs. In contrast, glucose concentrations were sampled and analysed twice daily using the BioProfile FLEX Analyser. To enhance the accuracy of ML models, offline measurements were also taken before and immediately after glucose feeding. In total, there are 100 offline values of glucose concentrations for all three bioreactors over the 2 weeks of cell culturing.

Due to the mismatch of time points at which online Raman spectra and offline glucose concentration measurements were collected, it is necessary to map the Raman spectra to the corresponding offline glucose concentration values for building training and testing datasets. In this study, we will associate each offline glucose concentration value with the closest Raman spectra collected after the timestamp of the offline measurement. As there are no Raman spectra collected during the feeding process, the offline glucose concentration value acquired immediately before glucose feeding will be mapped to the closest Raman spectra collected prior to that specific offline collection timepoint.

2.3.2 | Raman Data Preprocessing

Raman spectroscopy holds great promise as a real-time monitoring tool for key analytes in mammalian cell culture fermentations. However, significant challenges accompany this promising technology, including noise, strong background fluorescence and cocorrelations between multiple components. Preprocessing of the spectra is crucial to overcome these challenges [28] before employing multivariate regression analysis to extract relevant information and build a robust model. As affirmed by Poth et al. [29], preprocessing methods strongly influence the performance of machine learning models. Therefore, a typical pipeline for Raman-based machine learning models encompasses essential steps starting from Raman preprocessing methods, as illustrated in Figure 2.

Initially, the Raman spectra undergo several preprocessing steps, including wavelength selection (clipping), smoothing,

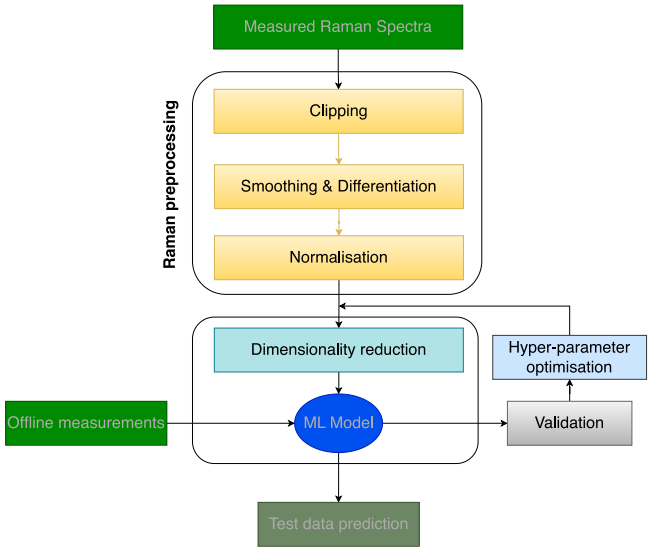


FIGURE 2 | A pipeline for Raman spectra modelling consists of two main procedures: preprocessing and model building. The preprocessing steps aim to standardise the data by removing noise and background-related contributions. At the end of the pipeline, statistical models or machine learning approaches are constructed. These models are then assessed, and parameter optimisation may be performed based on the model outcomes. All these steps together contribute to the creation of a robust prediction from the constructed model.

signal differentiation, normalisation and dimensionality reduction. Subsequently, the preprocessed Raman spectra, along with their corresponding offline measurements, will be employed for the development of a machine learning model. Throughout the model-building process, the hyperparameters of the models can be fine-tuned.

In this study, each Raman spectrum will be trimmed to the wavelength range of 500–3000 cm^{-1} to eliminate highly variable spectral slopes, window peaks, an artificial jump in the Raman signal caused by spectrograph mapping on Kaiser analysers and interference with water [29]. Additionally, this preprocessing step ensures the exclusion of information unrelated to glucose concentration [25]. In this study, we employed the Savitzky–Golay procedure [30] for the smoothing and differentiation step. This technique, based on least square fitting, was chosen for its effectiveness in preserving peaks from corruption. We used a moving average of 25 points, first-order differential and a polynomial order of 2 to fit the samples in the Savitzky–Golay smoothing. After performing the smoothing and differentiation, the Raman spectra are standardised, and in some cases, they can be directly analysed. However, variations in intensity between Raman spectra of different samples and even within spectral maps can be significant due to changes in focusing and other experimental factors. Therefore, the use of normalisation can help alleviate this effect. Each Raman spectrum in our experiment was normalised by first subtracting the mean and then dividing by its standard deviation.

Raman spectral datasets are typically characterised by a large number of variables, presenting challenges for statistical analysis in terms of generalisation performance and

computational effort. As a result, a dimensionality reduction should be conducted before the ML model building to find a lower-dimensional representation of the original dataset without significant loss of information. Several ML models, such as PLSR, can perform this step implicitly, while many other ML models may encounter challenges when learning from a dataset with limitations in the number of samples but high dimensionality. In this study, we employed kernel principal component analysis (KPCA) as a dimensionality reduction method. The number of principal components will be fine-tuned within the range of [3, 30]. The KPCA employs the radial basis function (RBF) as a kernel, and the kernel coefficient (γ) for RBF will be a floating-point number, logarithmically tuned within the range of $[10^{-6}, 10^{2}]$.

2.3.3 | Raman-Based Machine Learning Model Building

Unlike the general framework mentioned in Section 2.1, where input features (offline measurements) are associated with coefficients of variation, the coefficients of variation of input Raman features in this problem are usually unknown because of high dimension and complex relationships among input features. We assume that only the target variable (offline glucose concentration) exhibits the uncertainty in the obtained values. Therefore, we will modify the proposed framework as in Figure 3. In this modified framework, all base learners within the ensemble model will use the same input features but different values of the output variable. We will generate N training sets $(\mathbf{X}, \mathbf{Y}^{(i)})$ $(i \in [1, N])$ for N base learners by randomly generating N values for each output value using Monte Carlo method with Gaussian distribution. In a mathematical form, let $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_m]$ $(Y_k = \{y_k\}, y_k \in \mathbb{R}, k \in [1, m])$ be the m output values in the training set. For each target value y_k, N random values $\{y_k^{(1)}, \ldots, y_k^{(N)}\}$ will be generated from a Gaussian distribution $\mathbb{G}(\mu_k = y_k, \sigma_k, N)$, where σ_k is the standard deviation for each offline measurement y_k, calculated as follows:

$$\sigma_k = \begin{cases} \delta_Y, & \text{if } y_k \leq \beta \\ \delta_Y \times y_k, & \text{if } y_k > \beta, \end{cases} \tag{4}$$

where δ_Y is the maximum coefficient of variation of output variable $\mathbf{Y}$ and β is the threshold value used to compute the standard deviation for each offline measurement, depending on the measuring devices. For example, $\delta_Y = 0.07$ and $\beta = 1$ for glucose concentration measured by NOVA BioProfile Flex in our experiment.

After generating N samples for all output values in the training set, we will concatenate all values $y_k^{(i)}$ at the i th position $(i \in [1, N])$ to create the i th training set $(\mathbf{X}, \mathbf{Y}^{(i)})$ in order to train the i th base learner within the ensemble model. As all base learners utilise the same input features, it is not advisable to set identical best hyperparameters for all base learners, as indicated in the general framework in Figure 1. Instead, during the model-building process, we will conduct a hyperparameter tuning procedure to identify the specific optimal set of hyperparameters for each base learner, employing k-fold cross-validation or hold-out validation. In the case of using hold-out validation, a separate validation set $(\mathbf{X}^{val}, \mathbf{Y}_k^{val})$ needs to be prepared in the same manner as the training set $(\mathbf{X}, \mathbf{Y}_k)$. To facilitate fine-tuning for each base learner using the Optuna library [24], a fixed value of the number of base learners (N) needs to be used, as opposed to considering it as a tunable hyperparameter.

We performed an initial experiment to identify the suitable value of N using SVR as base learners in the ensemble model and KPCA as a dimensionality reduction method. The data from bioreactor A2 were used to train the ensemble model. If the fivefold cross-validation method is employed for hyperparameter tuning of the base estimators, the performance of the trained model will be tested on the data from bioreactor A1 (belonging to the same project as A2) and bioreactor A3 (from a different project). If the hold-out validation approach is used for hyperparameter tuning, the data from bioreactor A1 are used as a validation set. For this experiment, the number of base estimators considered includes 10, 30, 50, 70, 100, 200, 300, 400 and 500. The input Raman spectra were preprocessed as depicted in Section 2.3.2. The predicted performance of the trained models is presented in Figure 4. In the case of fivefold CV, given a fixed number of base estimators, the best

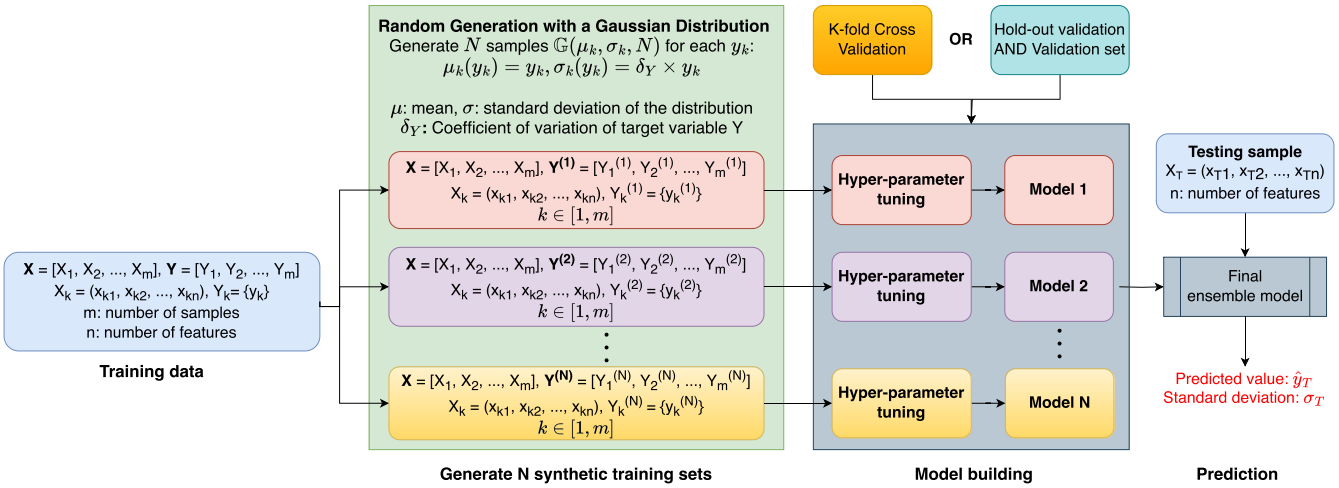
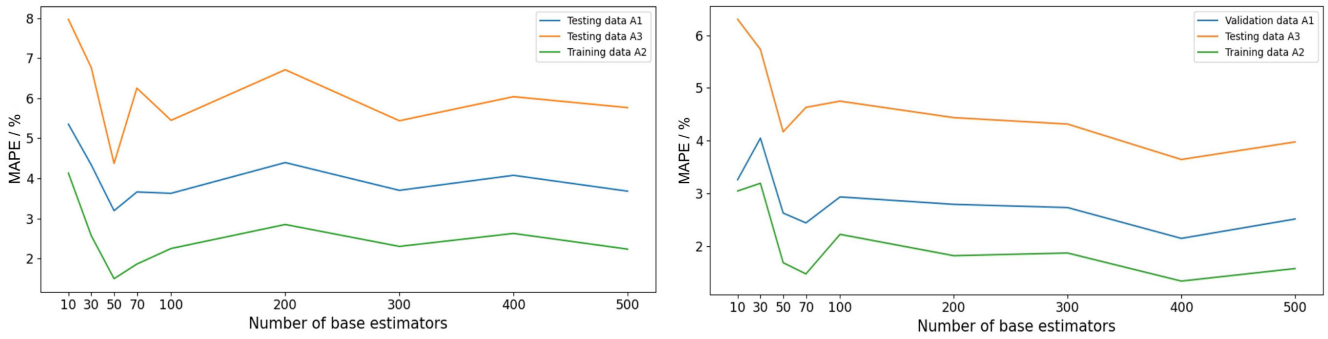


FIGURE 3 | The modified general framework for estimating uncertainty levels of predicted values from Raman input data using ensemble learning and Monte Carlo sampling.



(a) 5-fold CV for parameter tuning

(b) Hold-out validation for parameter tuning

FIGURE 4 | The impact of the number of estimators on the prediction performance of an ensemble model consisting of KPCA and SVR base estimators.

combination of hyperparameters for the base KPCA and SVR models that provided the smallest mean absolute percentage error (MAPE) value across all five training folds during the hyperparameter tuning process was used to train the ensemble model on the entire training set (A2). This trained ensemble model was then used to make predictions on the training and testing (A1 and A3) sets, recording the MAPE values for plotting the graph in Figure 4a. It should be noted that a fixed random seed value of 42 was used in this case to split the training data into five folds and to initialise the starting values for the hyperparameters. This random seed value was maintained across all experiments in Section 3. For the hold-out validation, given a fixed number of base estimators, the best combination of hyperparameters for the base KPCA and SVR models that provided the smallest MAPE value on the validation set (A1) was employed to train the ensemble model. This trained ensemble model was then used to make predictions on the training (A2), validation (A1) and testing (A3) sets for the specified number of base estimators, enabling the graph in Figure 4b to be constructed. Although this experimental approach can be applied to any Raman study, the results obtained here are dependent on the given experimental Raman training, validation and testing datasets, as well as the specified hyperparameter search ranges.

We can observe that for small numbers of base estimators (10 and 30 estimators), the testing error is high for both the outcomes of the fivefold CV and the hold-out validation methods. However, when the number of base estimators is equal to or greater than 50, increasing the number of base estimators does not significantly contribute to the reduction of prediction errors. Therefore, it can be concluded that 50 base estimators are sufficient to achieve good predictions without requiring a long training time in this case. As a result, we will use $N = 50$ to report the outcomes in the next parts.

For the model-building step, this study also uses the same ML algorithms for small datasets as the study presented in Section 2.2, including SVR, PLSR and GP. The hyperparameters of these algorithms will be fine-tuned using Bayesian optimisation methods within the Optuna library. Two validation methods are employed for hyperparameter tuning: the fivefold cross-validation and the hold-out validation. For the hold-out validation, the entire dataset of another bioreactor run would be used for validation. For the experiments, 100 iterations were used for hyperparameter optimisation. The ranges of hyperparameters for each ML model are provided in Table S2. It is noted that the single SVR model and the SVR used as a base learner within the ensemble model will use the same searching range of hyperparameter values.

3 | Results and Discussion

3.1 | Empirical Results for Predicting Cell Culture Process Performance

The public dataset given in Gangadharan et al. [22] was normalised to the range of 0 to 1. With the existence of values of 0, several metrics with the actual values in the denominator such as MAPE will not work. In this experiment, we will use MAE as a performance metric to compare the learning models:

$$MAE = \frac{\sum_{i=1}^{N_{test}} |\hat{y}_i - y_i|}{N_{test}} \tag{5}$$

where N_{test} is the number of testing samples, $\hat{y}_i$ is the prediction of the i th testing sample and y_i is the true value of the i th testing sample.

For the learning models which return the standard deviation associated with the predictive values, we will compute both MAE scores for the upper bound $(\hat{y} + 2\sigma)$ and the lower bound $(\hat{y} - 2\sigma)$ as follows:

$$MAE^{+} = \frac{\sum_{i=1}^{N_{test}} |(\hat{y}_i + 2\sigma_i) - y_i|}{N_{test}} \tag{6}$$

$$MAE^{-} = \frac{\sum_{i=1}^{N_{test}} |(\hat{y}_i - 2\sigma_i) - y_i|}{N_{test}} \tag{7}$$

where σ_i is the standard deviation associated with the prediction $\hat{y}_i$ of the i th testing sample. If the MAE value is small, while the values of MAE^{+} and MAE^{-} are high, the average of predictions provided by all individual base learners contributes to the reduction of variations among individual learners. This case

also indicates that the uncertainty level in the predictive outcomes is high. When the level of uncertainty is high, decision-making based on the prediction results should be approached with caution.

This section compares the average performance of the proposed ensemble models with individual models such as SVR, PLSR and GP. Table 1 summarises the mean MAE scores and standard deviation values over fivefold group cross-validation of predictions, upper bounds and lower bounds. Meanwhile, Figure 5 shows box-and-whisker plots of the compared ML models for MAE scores of all 106 cultures used as testing data over fivefold group cross-validation. It can be observed that the performance of the ensemble of SVRs outperforms that of using an individual SVR. However, the performance of the ensemble model of PLSRs is equal to the performance of a single PLSR model. When comparing the MAE scores of upper and lower bounds with the MAE score of predictive values generated by the average value of base regressors, we can see that the uncertainty level of predictions is small in this case. This is different from the case of using an individual GP model. In this experiment, although the GP can provide the best performance, the uncertainty level of predictions is higher than that using our proposed method. In addition, the predictive performance of all 106 cultures presented in Figure 5 illustrates that the ensemble of SVRs is competitive with the GP model.

Figure S1 depicts the culture exhibiting the most accurate predictive performance among the 106 cultures using various ML models. In contrast, Figure S2 showcases the culture with the least accurate predictive performance. It is evident that when the mAb concentration gradually increases throughout the culture time, the predictive performance of ML models tends to be high. Conversely, when the mAb concentration fluctuates, either increasing or decreasing suddenly during the cell culture process, the performance of ML models typically diminishes, and the associated uncertainty of predicted values increases. Additionally, even in the case of the best prediction, some experimental data points fall outside the range of uncertainty provided by the algorithm when using PLSR as base learners, as shown in Figure S1b. This occurs because the base PLSR models are less sensitive to small errors, which are limited to a maximum of 5% of the measured values, within a steady upward trend in antibody concentrations. As a result, the variation level among the base learners is small. Consequently, the low uncertainty in this best case reflects high confidence in the reliability of the predictive results provided by the PLSR models, particularly when the changing trends of mAb concentrations are easy to capture.

The outcomes presented in this section serve as a proof of concept for the proposed framework, operating under the assumption of a uniform coefficient of variation of 5% for all offline measurements, given that all input features are normalised to the range of 0 and 1. In practical scenarios, offline measurements may possess varying coefficients of variation, impacting the predictive performance of ML models. When dealing with different coefficients of variation for distinct input features, it becomes crucial to assess the importance of each input feature in determining predictive outcomes.

In the upcoming section, we will address another scenario where real-time measured input features are presumed accurate without the presence of a coefficient of variation. However, the target variable, being a specific offline measurement, exhibits variations in accuracy across different measuring times.

3.2 | Empirical Results for Real-Time Monitoring of Glucose Concentration Within Bioreactors Using Raman Spectra Data

In this experiment, we would like to assess the average percentage difference between predicted and actual values to compare with the maximum coefficient of variation of the actual values (about 7% for glucose concentrations). Therefore, the metric employed to assess the performance of ML models and determine the optimal parameter configurations in this section is the MAPE. Additionally, MAPE is scale independent, meaning it provides a percentage error and is not affected by the scale of the data. This makes it useful when comparing the performance of different ML models on testing data of bioreactors A1 and A3 with different scales. The MAPE metric will be computed as follows:

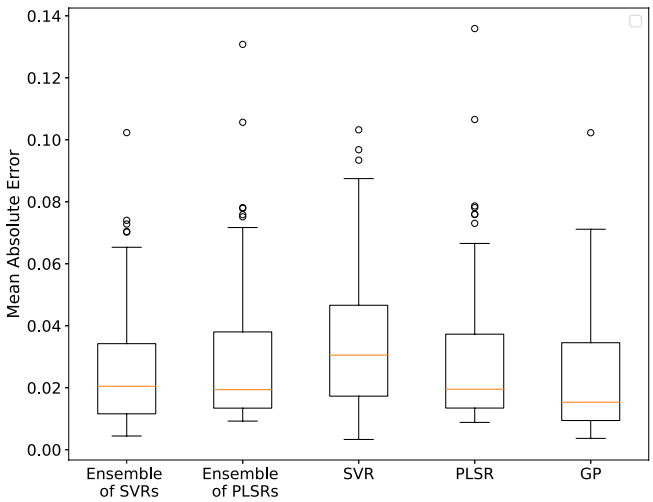


FIGURE 5 | Comparing the performance of different machine learning models in predictions of all 106 bioreactors used in the testing set over fivefold group cross-validation.

TABLE 1 | The mean and standard deviation of MAE scores over fivefold group cross-validation for different ML models.

Type	Ensemble of SVRs	Ensemble of PLSRs	SVR	PLSR	GP
Prediction	0.0256±0.0042	0.0288±0.0024	0.0344±0.0106	0.0285±0.0026	0.0233±0.0035
Upper bound	0.0301±0.0065	0.0293±0.0024	—	—	0.0573±0.0033
Lower bound	0.0287±0.0036	0.0293±0.0028	—	—	0.0583±0.0055

$$MAPE = \frac{1}{N_{test}} \cdot \sum_{i=1}^{N_{test}} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \quad (8)$$

where N_{test} is the number of testing samples, $\hat{y}_i$ is the prediction of the i th testing sample and y_i is the true value of the i th testing sample. For the ML models which are able to provide the standard deviation associated with the predictive values, we will compute both MAPE scores for the upper bound $(\hat{y} + 2\sigma)$ and the lower bound $(\hat{y} - 2\sigma)$ as follows:

$$MAPE^{+} = \frac{1}{N_{test}} \cdot \sum_{i=1}^{N_{test}} \left| \frac{(\hat{y}_i + 2\sigma_i) - y_i}{y_i} \right| \quad (9)$$

$$MAPE^{-} = \frac{1}{N_{test}} \cdot \sum_{i=1}^{N_{test}} \left| \frac{(\hat{y}_i - 2\sigma_i) - y_i}{y_i} \right| \quad (10)$$

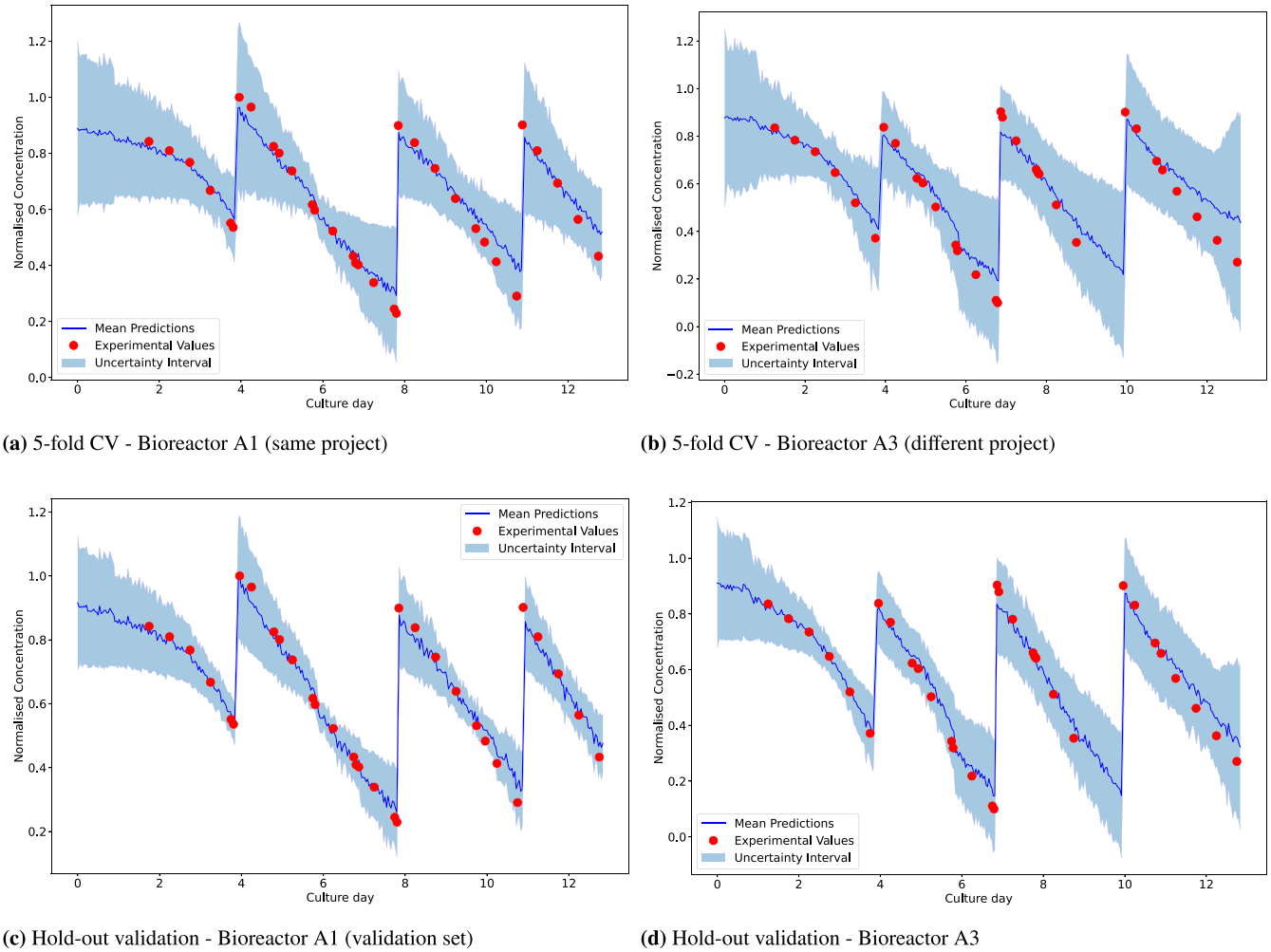
where σ_i is the standard deviation associated with the prediction $\hat{y}_i$ of the i th testing sample. If the $MAPE$ score is small, while the values of $MAPE^{+}$ and $MAPE^{-}$ are high, we can confirm that the average of predictions provided by all individual base learners contributes to the mitigation of variations among individual learners and increasing in the accuracy.

This section will present the practical outcomes of glucose concentration prediction employing various ML models. These models encompass the combination of KPCA and SVR, the combination of KPCA and GP, PLSR, ensemble of KPCA and SVRs and ensemble of PLSRs. The training dataset is derived from bioreactor A2, while datasets from bioreactors A1 and A3 serve as the testing data. In the case of using hold-out validation, the dataset from bioreactor A1 is utilised as the validation set.

3.2.1 | Comparing the Uncertainty Level in the Predicted Results With the Same and Different Projects

This section aims to evaluate the uncertainty level of real-time predictions made by the ensemble of KPCA and SVR base estimators. The predictions are based on Raman spectra that have undergone preprocessing steps as presented in Section 2.3.2.

Figure 6 illustrates the real-time predictions and uncertainty levels of glucose concentrations for different bioreactors within the same and different projects using an ensemble model. This is a typical use case in industry for a new project which will initially have very little specific data available for training, relying instead on data from previous projects. When considering both validation



(a) 5-fold CV - Bioreactor A1 (same project)

(b) 5-fold CV - Bioreactor A3 (different project)

(c) Hold-out validation - Bioreactor A1 (validation set)

(d) Hold-out validation - Bioreactor A3

FIGURE 6 | Real-time predictions of glucose concentrations within various bioreactors with the same and different projects using an ensemble of KPCA and SVRs.

methods, we observe that the uncertainty levels of predicted values within the bioreactors with the same project (Figure 6a,c) are smaller compared to those of different projects (Figure 6b,d). This discrepancy arises from the differences in metabolism and growth between cell lines which is reflected in the Raman spectra. These elements are quite similar between cell culture bioreactors within the same project, so the information recorded in the Raman spectra of the training data is more likely to be present in the testing data of the same project. In contrast, cell lines from different projects can have distinct interactions with the culture composition, and so the information recorded in the Raman spectra testing data may not be available if the training data are from a different project. Consequently, this disparity impacts the predicted values for the testing data and increases the uncertainty levels. To mitigate this, we can reduce the uncertainty in predicted values for testing data from a different project by validating and optimising the parameters of the trained model on the validation data within the same project as testing data.

Furthermore, as illustrated in Figure 6, the uncertainty associated with real-time predictions of the ensemble models fine-tuned using fivefold cross-validation (Figure 6a, b) is significantly higher than that of the ensemble model fine-tuned using hold-out validation (Figure 6c, d). This observation underscores that having a validation set encompassing data from all cell culture days enhances the reliability of the ensemble model, whereas validation on only a limited portion of cell culture days tends to amplify the uncertainty in predictive outcomes.

3.2.2 | Comparing the Predicted Performance Among the Different Tested ML Models

This section aims to compare the predicted performance of the proposed method with two different sets of base estimators: PLSR and the combination of KPCA and SVR, to the performance of single models such as PLSR, the combination of KPCA and GP and the combination of KPCA and SVR. Because the base estimators of the proposed ensemble method were trained on data with the target variable within a 7% deviation from the ground truth values, we do not expect the ensemble method to produce the best predicted performance compared to those trained on the ground truth values. However, we anticipate that the performance of the ensemble model will be comparable to that of the best single model. One of the strengths of the proposed method, compared to the single ML models, is its ability to provide the standard deviation of the predicted outcomes based on the coefficient of variation of bioanalysers.

In this experiment, we used 50 base estimators to build the ensemble model. The ML models were trained using the data from bioreactor A2. For the case of fivefold CV employed for the hyperparameter tuning, the trained models were tested on the data from bioreactor A1 (the same project as A2) and bioreactor A3 (a different project from A2). In the case of using hold-out validation for the hyperparameter tuning, the data from bioreactor A1 were used as a validation set, and the data from bioreactor A3 were used as a testing set.

Table 2 presents the predicted performance of various ML models on the data from bioreactor A1 (the same project as A2) and bioreactor A3 (a different project from A2) using the fivefold CV approach. Among the models, the combination of KPCA and SVR provided the best performance on predictions of glucose concentrations of bioreactor A3, while the combination of KPCA and GP yielded the best results on bioreactor A1 of the same project with the training data.

In this experiment, the ensemble of models did not consistently outperform the single models in terms of predicted values. While the ensemble of PLSRs can outperform a single PLSR model, the ensemble of KPCA and SVRs is not able to provide a better performance in comparison of the single combination of KPCA and SVR. However, the difference in predicted performance between the ensemble models and single models was generally within an error of around 4%. Nevertheless, the ensemble model provided additional benefits by generating an estimation of the uncertainty for each predicted outcome compared to the single models. It is worth noting that the prediction errors of the ML models in this experiment were often below 7%, which is also smaller than the maximum error of 7% associated with the bioanalyser for offline glucose concentrations. This observation suggests that the ML models can be effectively employed for developing soft sensors for real-time monitoring of glucose concentrations.

Table 3 presents the predicted performance of various ML models trained on data from bioreactor A2, validated on data from bioreactor A1 and tested on data from bioreactor A3. In this scenario, the combination of KPCA and SVR stands out as the top-performing model on the testing data (A3). Furthermore, it is noticeable that the testing errors of all five ML models fine-tuned by the hold-out validation are smaller than those fine-tuned by the fivefold cross-validation method. Additionally, the uncertainty levels on the testing data for the ensemble models using the hold-out validation are smaller than those using the fivefold cross-validation method.

TABLE 2 | The MAPE (%) values of various ML models trained and optimised by the fivefold CV method.

	Bioreactor A1			Bioreactor A3		
MAPE (%)	**Prediction**	**Upper bound**	**Lower bound**	**Prediction**	**Upper bound**	**Lower bound**
Ensemble of KPCA + SVRs	3.8280	18.1349	13.8381	6.1250	28.4885	19.5367
Ensemble of PLSRs	4.5388	14.0537	6.2497	4.3285	13.1566	11.2454
KPCA + SVR	1.6133	—	—	2.5731	—	—
PLSR	4.2982	—	—	4.2428	—	—
KPCA + GP	2.0874	2.0942	2.0806	4.1223	4.1400	4.1047

TABLE 3 | The MAPE (%) values of various ML models trained and optimised by the hold-out validation method.

	Bioreactor A1 (validation set)			Bioreactor A3		
MAPE (%)	**Prediction**	**Upper bound**	**Lower bound**	**Prediction**	**Upper bound**	**Lower bound**
Ensemble of KPCA + SVRs	2.1227	9.8612	8.9721	3.6011	18.1493	14.0127
Ensemble of PLSRs	3.8046	13.6509	7.3930	4.2195	13.4981	12.3385
KPCA + SVR	1.3246	—	—	2.3634	—	—
PLSR	4.2982	—	—	4.2428	—	—
KPCA + GP	1.8168	1.8142	1.8194	3.4109	3.4381	3.3837

TABLE 4 | Coverage of uncertainty bounds with respect to testing samples of the proposed ensemble framework and Gaussian process models.

	Fivefold cross-validation		Hold-out validation
Coverage (%)	**Bioreactor A1**	**Bioreactor A3**	**Bioreactor A3**
Ensemble of KPCA + SVRs	100	100	100
Ensemble of PLSRs	96.97	96.88	96.88
KPCA + GP	0	0	3.125

Figures S3 and S4 illustrate the real-time predictions of glucose concentrations using different ML models based on preprocessed Raman spectra. The demonstrations reveal that while single models can, on average, produce better predictions than the ensemble model of uncertainty-associated base learners, they fail to accurately capture the lowest glucose concentration values (bottom points) where decisions regarding glucose addition have been made or the highest glucose concentrations after glucose feeding. In contrast, the predicted values of the ensemble model, along with the standard deviation values, encompass these lowest glucose levels and the highest glucose concentrations immediately after glucose addition to bioreactors. These results affirm the strengths of the proposed methods in assessing the uncertainty of predicted values, which is crucial for developing control strategies for automated glucose feeding in bioreactors.

3.2.3 | Comparing the Uncertainty Levels of GP Models With the Proposed Ensemble Framework

From Figures S3 and S4, it can be observed that the standard deviation of the predicted responses generated by the GP models is nearly zero. Consequently, the prediction intervals of the GP models are very narrow, limiting their ability to assess the uncertainty levels of predicted values. This limitation arises from the fact that the GP models were trained solely on offline measurements without considering the uncertainty and errors associated with the bioanalysers. In contrast, our proposed method takes into account the uncertainty associated with each offline measurement and incorporates this information during the construction of the ML models. As a result, the predicted responses, along with the corresponding standard deviation values, provide a comprehensive coverage of practical observations. The wider prediction intervals obtained using our proposed method effectively assess the uncertainty of predicted values and facilitate informed decision-making for the control process.

Table 4 shows the coverage percentage of the uncertainty regions of our proposed ensemble models and the GP models with respect to actual observations in the testing bioreactors, using fivefold cross-validation and hold-out validation for model building and hyperparameter tuning. It can be seen that over 95% of the actual offline glucose concentration values fall within the uncertainty boundary of the proposed ensemble models, as expected when using $2 \cdot \sigma$ to estimate the uncertainty boundary of predictions. In contrast, the boundary of Raman-based GP models in this experiment usually does not contain the actual observations because the width of the boundary is very small. These results confirm that our proposed ensemble method estimates the uncertainty of the predictions better than the popular GP models for the real-time monitoring problem of glucose concentrations within cell culture bioreactors using Raman spectra as input features.

3.3 | Applicability of the Proposed Method Within the Industry and Potential Roadblocks

Our proposed method combines ensemble learning and Monte Carlo sampling to quantify uncertainty in the predictive results provided by ML models, particularly in the context of limited training data. Any single ML model, such as PLSR, SVR or GP, can be used as a base learner within the framework to estimate the uncertainty associated with the predictions of these individual models. Although single models, such as PLSR, are well adopted, straightforward to implement and frequently used in building predictive models for various problems in biopharmaceuticals, they cannot reliably estimate the uncertainty associated with their predictions. The uncertainty bounds of predictions are critical for making decisions related to control operations and assessing the error tolerance of learning models, especially when actual offline measurements in training data also exhibit uncertainty. Our proposed framework addresses this gap. The empirical outcomes illustrate the applicability of the framework in enhancing therapeutic manufacturing by providing reliable insights into

model predictions and facilitating decision-making under uncertainty. Although the performance of the ensemble models, in some cases, is not higher than the accuracy of single models due to the base learners being trained on simulated offline values derived from the variability and errors of actual offline values, the uncertainty bounds of the predicted values can still encompass all actual testing values, as illustrated in Figure 6.

It should be noted that the simulated data generated using Monte Carlo sampling, which is a rigorous and statistically valid method, for training base learners does not fabricate new trends but rather extends the variability inherent in actual offline data. The simulated data distribution mirrors actual measurements by using the obtained offline values as the mean and the coefficient of variation from offline analysers to calculate the standard deviation. This approach is employed to augment limited experimental data, aiming to construct more reliable learning models, while the performance of the proposed method is still evaluated using experimental data. Consequently, our method does not present any issues with regulatory bodies when applied to manufacturing processes.

The proposed framework could be integrated into real-time monitoring systems for bioreactors in cell culture processes or future value forecasting systems for measurements of interest. Uncertainty quantification enables operators to understand the confidence in predictions related to cell growth, metabolite concentrations or product titre, allowing for dynamic adjustments to process parameters. Current practices in the industry where decisions are made on measurements include consideration of the uncertainty of those measurements. It is imperative that ML models are able to provide similar estimates to support effective decision-making during manufacturing and process development. Predictions with uncertainty estimates could also improve process optimisation by identifying scenarios with lower risks, leading to more efficient resource usage such as media components and energy. The methodology can be a core component of biopharma digital twins, enabling simulations of 'what-if' scenarios with explicit uncertainty propagation. To facilitate the reproduction of our proposed method, we also provide pseudocode in Figures S5 and S6.

Potential roadblocks for the proposed method in the context of the biopharmaceutical industry include model interpretability, integration challenges, data availability and quality, as well as resistance to change. Regulatory bodies often favour interpretable models; however, while ensemble models are robust, their outputs may be less interpretable compared to simpler mechanistic models. Deploying predictive ML models in established biopharmaceutical manufacturing setups requires seamless integration with existing process control systems and data acquisition systems. Continuous data integration facilities may require costly upgrades, which are often not readily available in the context of biopharmaceutical manufacturing. Finally, bioprocess data are suboptimal for training machine learning models due to batch-to-batch variability, noise in sensor readings (such as Raman spectroscopy) and limited labelled data.

4 | Conclusions

This paper introduced a novel framework capable of integrating any regressors as base learners to estimate uncertainty associated with each predictive outcome, especially in situations with limited training data. The coefficients of variation from offline measurements are utilised to calculate the standard deviation of Gaussian distributions, which, in turn, are employed to generate synthetic samples complementing the available values in the dataset. All synthetic data contribute to the training of base learners. The effectiveness of the proposed method was evaluated through two case studies. The first case involves using obtained offline measurements on the current culture day to predict mAb concentrations on the next culture day. The second case uses real-time Raman spectral data as input features to predict glucose concentrations for real-time monitoring of bioreactor runs. Empirical results demonstrated the robust performance of the proposed framework in both case studies, with small testing errors. Notably, a key strength of the proposed method lies in its ability to provide the uncertainty level associated with each prediction. This uncertainty level is crucial for informed decision-making in control strategies to enhance cell culture process performance, such as adjusting glucose levels in bioreactors to sustain cell growth and productivity.

There are several potential directions for expanding the proposed framework. In the scenario where only offline measurements are used for early predictions regarding the future state of bioreactors, it becomes crucial to assess the impact of each input feature by assigning a specific coefficient of variation for each offline measurement. For online monitoring based on Raman spectral data, enhancing model accuracy could involve considering the incorporation of additional information beyond Raman spectra. This might include control variables, manual intervention data or domain knowledge derived from computational fluid dynamics models and the kinetics of each cultivation process [25]. Moreover, leveraging the predictive results along with the uncertainty levels provided by the proposed method could serve as a foundation for developing control strategies for real-time feedback control of bioreactors, particularly based on online glucose concentrations. In addition, there is a need to develop automated methods for combining and optimising hyperparameters of Raman data preprocessing techniques, moving beyond the fixed parameter setting used in the current work.

Acknowledgements

This research was supported under the Australian Research Council's Industrial Transformation Research Program (ITRP) funding scheme (Project Number IH210100051). The ARC Digital Bioprocess Development Hub is a collaboration between the University of Melbourne, University of Technology Sydney, RMIT University, CSL Innovation Pty Ltd, Cytiva (Global Life Science Solutions Australia Pty Ltd) and Patheon Biologics Australia Pty Ltd. Open access publishing facilitated by University of Technology Sydney, as part of the Wiley - University of Technology Sydney agreement via the Council of Australian University Librarians.

Ethics Statement

The authors have nothing to report.

Conflicts of Interest

Robert Bassett and Ellen Otte are employees of CSL Innovation Pty Ltd. Thanh Tung Khuat and Bogdan Gabrys declare no competing interests, including no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability Statement

Data used in Section 3.1 can be downloaded from https://ars.els-cdn.com/content/image/1-s2.0-S0098135421000041-mmc3.xlsx. However, the data used in Section 3.2 are the intellectual property of CSL Limited and are therefore not shared publicly.

References

1. R. M. Lu, Y. C. Hwang, I. J. Liu, et al., "Development of Therapeutic Antibodies for the Treatment of Diseases," *Journal of Biomedical Science* 27, no. 1 (2020): 1–30.

2. M. S. Hong, F. Mohr, C. D. Castro, et al., "Smart Process Analytics for the End-to-End Batch Manufacturing of Monoclonal Antibodies," *Computers & Chemical Engineering* 179 (2023): 108445.

3. Evaluate Pharma, *World Preview 2021, Outlook to 2026* (Evaluate Ltd, 2021).

4. M. M. Papathanasiou, B. Burnak, J. Katz, N. Shah, and E. N. Pistikopoulos, "Assisting Continuous Biomanufacturing Through Advanced Control in Downstream Purification," *Computers and Chemical Engineering* 125 (2019): 232–248.

5. F. M. Wurm, "Production of Recombinant Protein Therapeutics in Cultivated Mammalian Cells," *Nature Biotechnology* 22, no. 11 (2004): 1393–1398.

6. M. M. Papathanasiou, A. L. Quiroga-Campano, F. Steinebach, M. Elviro, A. Mantalaris, and E. N. Pistikopoulos, "Advanced Model-Based Control Strategies for the Intensification of Upstream and Downstream Processing in mAb Production," *Biotechnology Progress* 33, no. 4 (2017): 966–988.

7. A. C. Satheka, "Upscaling of Clinical Grade Stem Cell Production: Upstream Processing (USP) and Downstream Processing (DSP) Operations of Cell Expansion, Harvesting, Detachment, Separation, Washing and Concentration Steps, and the Regulatory Requirements," in *Stem Cell Production: Processes, Practices and Regulations* (Springer, 2022): 159–184.

8. T. T. Khuat, R. Bassett, E. Otte, A. Grevis-James, and B. Gabrys, "Applications of Machine Learning in Antibody Discovery, Process Development, Manufacturing and Formulation: Current Trends, Challenges, and Opportunities," *Computers & Chemical Engineering* 182 (2024): 108585.

9. B. Kelley, "Industrialization of mAb Production Technology: The Bioprocessing Industry at a Crossroads," *MAbs* 1, no. 5 (2009): 443–452.

10. F. Li, N. Vijayasankaran, A. Shen, R. Kiss, and A. Amanullah, "Cell Culture Processes for Monoclonal Antibody Production," *MAbs* 2, no. 5 (2010): 466–479.

11. G. E. Derfus, D. Abramzon, M. Tung, D. Chang, R. Kiss, and A. Amanullah, "Cell Culture Monitoring via an Auto-Sampler and an Integrated Multi-Functional Off-Line Analyzer," *Biotechnology Progress* 26, no. 1 (2010): 284–292.

12. M. McRae, J. McHale, S. Granger, B. Goulart, and E. Kilcoyne, "Monitoring Progress of Bioreactor Runs," *Genetic Engineering & Biotechnology News* 32, no. 12 (2012): 42–42.

13. E. Hüllermeier and W. Waegeman, "Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods," *Machine Learning* 110, no. 3 (2021): 457–506, https://doi.org/10.1007/s10994-021-05946-3.

14. E. Heid, C. J. McGill, F. H. Vermeire, and W. H. Green, "Characterizing Uncertainty in Machine Learning for Chemistry," *Journal of Chemical Information and Modeling* 63 (2023): 4012–4029.

15. C. Hutter, M. von Stosch, M. N. Cruz Bournazou, and A. Butté, "Knowledge Transfer Across Cell Lines Using Hybrid Gaussian Process Models With Entity Embedding Vectors," *Biotechnology and Bioengineering* 118, no. 11 (2021): 4389–4401.

16. A. Tulsyan, C. Garvin, and C. Ündey, "Advances in Industrial Biopharmaceutical Batch Process Monitoring: Machine-Learning Methods for Small Data Problems," *Biotechnology and Bioengineering* 115, no. 8 (2018): 1915–1924.

17. M. Banner, H. Alosert, C. Spencer, et al., "A Decade in Review: Use of Data Analytics Within the Biopharmaceutical Sector," *Current Opinion in Chemical Engineering* 34 (2021): 100758.

18. P. Kokol, M. Kokol, and S. Zagoranski, "Machine Learning on Small Size Samples: A Synthetic Knowledge Synthesis," *Science Progress* 105, no. 1 (2022): 00368504211029777.

19. D. L. Goodhue, W. Lewis, and R. Thompson, "Does PLS Have Advantages for Small Sample Size or Non-Normal Data?," *MIS Quarterly* (2012): 981–1001.

20. X. Jianlin, M. S. Rehmann, X. Xu, et al., "Improving Titer While Maintaining Quality of Final Formulated Drug Substance via Optimization of CHO Cell Culture Conditions in Low-Iron Chemically Defined Media," *MAbs* 10 (2018): 488–499.

21. D. Y. Kim, J. C. Lee, H. N. Chang, and D. J. Oh, "Effects of Supplementation of Various Medium Components on Chinese Hamster Ovary Cell Cultures Producing Recombinant Antibody," *Cytotechnology* 47 (2005): 37–49.

22. N. Gangadharan, D. Sewell, R. Turner, et al., "Data Intelligence for Process Performance Prediction in Biologics Manufacturing," *Computers & Chemical Engineering* 146 (2021): 107226.

23. F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research* 12 (2011): 2825–2830.

24. T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-Generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Association for Computing Machinery, 2019): 2623–2631.

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

26. C. Gillespie, D. P. Wasalathanthri, D. B. Ritz, et al., "Systematic Assessment of Process Analytical Technologies for Biologics," *Biotechnology and Bioengineering* 119, no. 2 (2022): 423–434.

27. L. Gibbons, F. Maslanka, N. Le, et al., "An Assessment of the Impact of Raman Based Glucose Feedback Control on CHO Cell Bioreactor Process Development," *Biotechnology Progress* 39, no. 5 (2023): e3371.

28. K. H. Liland, A. Kohler, and N. K. Afseth, "Model-Based Pre-Processing in Raman Spectroscopy of Biological Samples," *Journal of Raman Spectroscopy* 47, no. 6 (2016): 643–650.

29. M. Poth, G. Magill, A. Filgertshofer, O. Popp, and T. Großkopf, "Extensive Evaluation of Machine Learning Models and Data Preprocessings for Raman Modeling in Bioprocessing," *Journal of Raman Spectroscopy* 53, no. 9 (2022): 1580–1591.

30. A. Savitzky and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," *Analytical Chemistry* 36, no. 8 (1964): 1627–1639.

Supporting Information

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