Journal of Biotechnology 377 (2023) 23–33

Contents lists available at ScienceDirect

Journal of Biotechnology

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

Soft sensor based on Raman spectroscopy for the in-line monitoring of metabolites and polymer quality in the biomanufacturing of polyhydroxyalkanoates

João Medeiros Garcia Alcântara [1], Francesco Iannacci [1], Massimo Morbidelli, Mattia Sponchioni [*]

Dept. of Chemistry, Materials and Chemical Engineering "Giulio Natta", Politecnico di Milano, via Mancinelli 7, Milano 20131, Italy

ARTICLE INFO

Keywords:
Raman spectroscopy
In-line monitoring
Perfusion
Polyhydroxyalkanoates
Multivariate data analysis
Cupriavidus necator

ABSTRACT

Polyhydroxyalkanoates (PHA) are among the most promising bio-based alternatives to conventional petroleum-based plastics. These biodegradable polyesters can in fact be produced by fermentation from bacteria like *Cupriavidus necator*, thus reducing the environmental footprint of the manufacturing process. However, ensuring consistent product quality attributes is a major challenge of biomanufacturing. To address this issue, the implementation of real-time monitoring tools is essential to increase process understanding, enable a prompt response to possible process deviations and realize on-line process optimization. In this work, a soft sensor based on *in situ* Raman spectroscopy was developed and applied to the in-line monitoring of PHA biomanufacturing. This strategy allows the collection of quantitative information directly from the culture broth, without the need for sampling, and at high frequency. In fact, through an optimized multivariate data analysis pipeline, this soft sensor allows monitoring cell dry weight, as well as carbon and nitrogen source concentrations with root mean squared errors (RMSE) equal to 3.71, 7 and 0.03 g/L, respectively. In addition, this tool allows the in-line monitoring of intracellular PHA accumulation, with an RMSE of 14 g_{PHA}/g_{Cells}. For the first time, also the number and weight average molecular weights of the polymer produced could be monitored, with RMSE of 8.7E4 and 11.6E4 g/mol, respectively. Overall, this work demonstrates the potential of Raman spectroscopy in the in-line monitoring of biotechnology processes, leading to the simultaneous measurement of several process variables in real time without the need of sampling and labor-intensive sample preparations.

1. Introduction

Plastic plays a fundamental role in our society, either as industrial or domestic material, replacing wood, glass and many other raw materials thanks to its incomparable thermomechanical properties, product stability, light weight, and durability, all combined with economic feasibility (Singh et al., 2017). These features have made plastic manufacturing reach over 390 million tons per year, most of which coming from fossil sources (https://plasticseurope.org/knowledge-hub/plastics-the-facts-2022/.). Nevertheless, over the recent years, the negative aspects related to this wide usage have been highlighted (Andreasi Bassi et al., 2021). In particular, one of the most dangerous sources of pollution is the incorrect disposal of these plastics (Meereboer et al., 2020). The complex structure and the high molecular mass extend the durability, allowing the various plastics to remain in water bodies, soil, and landfills for extremely long times (Raza et al., 2018). This persistence in the environment is source of the notorious concern of microplastics (MPs). The pervasive use of plastics in all facets of human life results in a daily exposure to MPs. This continual exposure through drinking water has sparked increasing worries regarding potential health risks for humans (Ivar do Sul and Costa, 2014; Kirstein et al., 2021). Even though recovering and recycling technologies are improving, we still cannot avoid sustained pollution. An alternative solution is the replacement of these traditional plastics with biopolymers, materials obtained from renewable sources, which ideally combine biodegradability with the mechanical and chemical properties of conventional polymers (Albuquerque and Malafaia, 2018). Even if this ultimate solution has not been reached yet, one of the most

* Corresponding author.
E-mail address: mattia.sponchioni@polimi.it (M. Sponchioni).
[1] J.M.G.A. and F.I. equally contributed to this work.

https://doi.org/10.1016/j.jbiotec.2023.10.005
Received 5 August 2023; Received in revised form 4 October 2023; Accepted 16 October 2023
Available online 23 October 2023

appealing green alternatives to fossil-based plastics is currently represented by polyhydroxyalkanoates (PHA), a class of renewable and biodegradable polyesters that shows the highest expected growth rate among the bio-based polymers (Andreasi Bassi et al., 2021; Medeiros Garcia Alcântara et al., 2020).

PHA are produced as a natural strategy for carbon and energy storage by several microorganisms. Their large-scale manufacturing can be then accomplished by cultivating such microorganisms either in discontinuous or continuous bioreactors (Medeiros Garcia Alcântara and Sponchioni, 2022). However, high operating costs, lot-to-lot variability of the produced PHA and availability of an appropriate downstream processing are still important bottlenecks to be considered. To ensure consistent product quality, robustness of the manufacturing, and a labor effective process optimization aimed at maximizing the PHA productivity, making it competitive with the traditional plastics, the development of reliable tools capable of monitoring in real time the biopolymer quality is crucial (Blunt et al., 2018).

Indeed, the Process Analytical Technology (PAT) initiative, enacted by the Food and Drug Administration (FDA) in 2004, has among its main goals the establishment of a comprehensive understanding of manufacturing processes through the definition of critical quality attributes and their relationship with critical process parameters (Hinz, 2006). Mainly applied in the pharmaceutical industry, the so-called PAT tools for the monitoring and control of manufacturing processes have been spreading to many other industries in the last decades, including the manufacturing of PHA (Gernaey et al., 2012). In fact, several works reported the possibility of monitoring, quantifying, and assessing the ongoing PHA synthesis via on-line measurements, exploiting different analytical techniques. Among all, Fourier-transformed infrared spectroscopy (FT-IR) combined with attenuated total reflection (ATR) is a very promising monitoring solution as demonstrated by different works (Jarute et al., 2004; Doppler et al., 2021), as well as the in-line photon density wave (PDW) technique, able to collect measurements in highly turbid biochemical processes (Gutschmann et al., 2019, 2023). In addition, different research groups focused on information from exhaust gas data in order to predict both substrate consumption and PHA production trends (Duvigneau et al., 2022; García et al., 2019; Ochoa et al., 2020). Despite their potential as monitoring tools, most of these technologies either require sampling of the culture broth or complex sample preparation policies, which limit the data acquisition rate as well as introduce the risk of culture contamination.

Among the different monitoring tools available, Raman spectroscopy is gaining much attention from both academia and industry since it can provide an in-depth characterization of the cell culture, measuring, at the same time, nutrients, metabolites, cell density and polymer quantity and quality (Samek et al., 2016). As such, this analytical technique has found applicability in many fields, such as material science, diagnostics, and biology (Mulvaney and Keating, 2000), producing a fingerprint of culture biochemical composition (Das and Agrawal, 2011). These fingerprint-like signatures can be translated into snapshots for identification of molecular reactions and biologics (Pezzotti, 2021). Given its potential, the interest in applying this spectroscopic technique to PHA fermentation processes is reflected in different scientific works. De Gelder et al. (De Gelder et al., 2008) and Samek et al. (Samek et al., 2016) explored the potential of Raman spectroscopy for PHA quantification directly from bacterial cultures. Tao and co-workers (Tao et al., 2016) focused on the dynamic changing of PHA-related wavelength intensity, estimating the precise periods of maximum productivity. Thanks to regression statistics, it is possible to train chemometric models via offline measurements, obtaining from Raman spectra a suitable methodology for quantification (Abu-Absi et al., 2011; Rowland-Jones and Jaques, 2019). In order to extract the maximum possible information from the measured spectra, we need to rely on suitable multivariate data analysis (MVDA) methodologies (Madden and Ryder, 2003). MVDA projection methods, such as principal component analysis (PCA), partial least squares (PLS), and orthogonal PLS (OPLS), are statistical techniques that have been developed and used to analyze data generated from more than one variable. As such, they fit suitably within the goals of PAT and Quality by Design (QbD) since they allow significant information mining from spectroscopic data (Beckett et al., 2018). Indeed, the objective is to obtain, with sufficient accuracy, the simultaneous quantification of several relevant variables. This approach requires the suitable formulation of a model as well as data preprocessing (DPP), which strongly affects the final prediction accuracy (Gonzalez Zelaya, 2019; Narayanan et al., 2022). Among the different options, PLS regression models have been demonstrated as the most promising methodology to extract information from Raman spectra, due to their relative simplicity compared to other statistical approaches and sufficient accuracy (Hisazumi and Kleinebudde, 2017; Radtke et al., 2020; Esmonde-White et al., 2022).

Despite the significant advances in the monitoring of PHA production from bacterial cultures through Raman spectroscopy, the reported approaches are still mainly based on at-line measurements, requiring sampling and labor-intensive sample preparations before being able of extracting quantitative information. These drawbacks erode the great potential of Raman spectroscopy for providing a fast response, and potentially a real-time characterization of the culture broth. Conversely, *in situ* monitoring offers great potential to overcome typical limitations associated to at-line and off-line process supervision. Indeed, *in situ* monitoring tools are in use in various fields, from crystalline solid formation (Pienack and Bensch, 2011) to battery development (Grey and Tarascon, 2017) to 3D metal printing (Colosimo and Grasso, 2020).

To bridge this gap between the biotechnology industry and other process industries, we developed a soft sensor enabling the real-time monitoring of PHA production from *Cupriavidus necator* through *in situ* Raman spectroscopy, avoiding any sampling and sample preparation. With this approach, based on the MVDA of the collected Raman spectra, we show the possibility of quantifying important cell culture parameters such as the viable cell density and metabolite concentration, but also the polymer quality and, specifically, its molecular weight distribution (MWD). To the best of our knowledge, this is the first report about the real-time assessment of the MWD of an intracellular product like PHA, without the need for sampling. In this direction, we first developed an optimal algorithm architecture. This was obtained by evaluating the best hyperparameters (HP), which are the characteristic factors of data preprocessing, through a Bayesian optimization loop and a K-Fold cross-validation (CV) procedure. We then applied this soft sensor for the in-line monitoring of PHA from perfusion cultures.

The in-line information provided by this tool can be advantageously exploited for the development of a process control algorithm for on-line disturbance rejection and process optimization, allowing to reduce the costs associated to PHA biomanufacturing.

2. Materials and methods

2.1. Experimental setup

The experimental setup is the high-density perfusion bioreactor schematically shown in Fig. 1. This bioreactor consists of a 2 L stirred tank vessel (Vaudaux-Eppendorf, Switzerland), connected to an alternating tangential flow filtration (ATF) unit, composed of a hollow-fiber membrane (0.5 μm, PES, 1570 cm^2 Repligen, USA) and a diaphragm pump connected to its controller (Repligen, USA). This ATF device allows to retain the cells inside the bioreactor, thus increasing the density. The cell line used in these experiments is *Cupriavidus necator* DSM 428 (DSMZ, Germany). The experimental runs were performed at 37 °C, pH 7, Dissolved Oxygen (DO) 20% v/v and stirring speed 500 rpm. By using DASGIP Control software (Eppendorf, Germany), these parameters were kept constant throughout the experiments. Temperature was controlled using a heating mantel and pH through the automatized addition of carbon dioxide and pH buffers. DO percentage was adjusted by oxygen flowrate and mixing was provided by a Rushton impeller installed at the

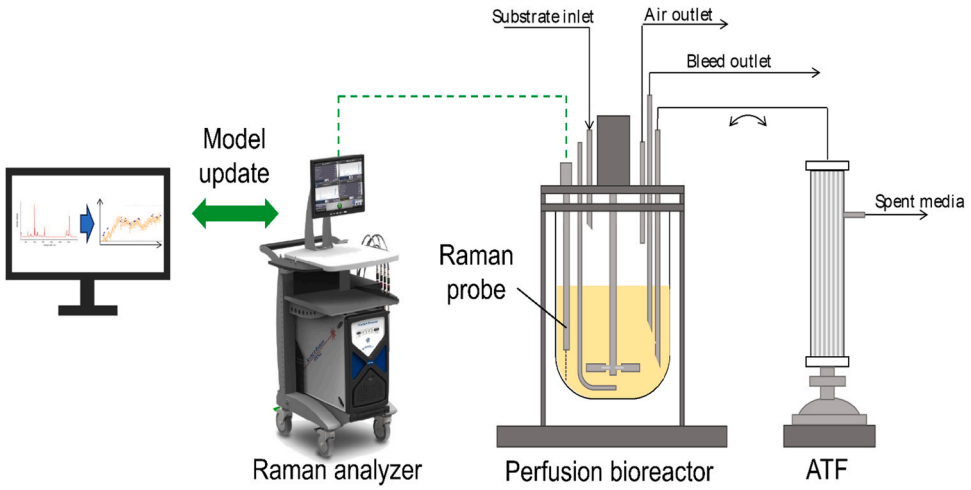

Fig. 1. Schematic representation of the high-density perfusion bioreactor coupled with the alternating tangential flow (ATF) filtration unit for cell retention used for PHA biomanufacturing. The reactor is equipped with a submersion probe for in situ Raman spectroscopy. The Raman probe was connected to a computer for spectra collection. The signal collected in-line is used for model training and update.

bottom of transmission shaft. The inlet air was filtered through an air filter (Midisart 2000, 0.2 µm PTFE, Sartorius, Germany) and inserted in the culture volume via a 7-hole sparger. Depending on oxygen uptake, the inlet volumetric concentration was in the range 0–30%, while air flowrate varied between 20 and 50 sL/h. Antifoam B emulsion at 5% in water (Sigma-Aldrich, USA) was added manually while two marine impellers were implemented above the liquid level, to prevent foam accumulation.

Three different media were used: a basal growth medium, a basal growth medium without nitrogen source, and a concentrated C-Fraction. The carbon source was gluconic acid and the nitrogen source ammonium phosphate dibasic. The growth medium formulation was adapted from (López-Abelairas et al., 2015), the basal medium without nitrogen had the same formulation but without any nitrogen source. The C-Fraction solution was composed of 5.0 g/L $MgSO_4 \cdot 7\,H_2O$ (Honeywell Fluka, Germany) and 324.0 g/L D-gluconic acid sodium salt (Sigma Aldrich, USA).

In all experimental runs, the growth medium was supplied at a constant flowrate (*i.e.* 42 mL/h). After the exponential phase, this medium was switched to the one without nitrogen in order to promote PHA accumulation. The C-fraction was added manually whenever the carbon source concentration dropped below 17 g/L.

Raman spectra were acquired from all bioreactor experiments using a Kaiser RamanRxn2 analyzer (Kaiser Optical Systems Inc., USA), which includes a 785 nm laser with a power of 400 mW and a cooled charged-coupled device (CCD) detector, measuring inelastic photon scattering across the range of 100–3425 cm^{-1}. This device was connected to a bIO-Optic immersion probe placed inside the perfusion bioreactor through a bioreactor port. The Raman probe, as all the other hardware together with the reactor unit, were sterilized via autoclave using a Systec DX65 (Systec GmbH & Co. KG, Germany). Spectra were collected every 15 min using an exposure time of 10 seconds and 75 cumulated scans, leading to a measurement time of 12.5 min per spectrum. In this way the final resulting spectrum came from the average of cumulated scans, with the aim of improving the signal-to-noise ratio.

2.2. Analytics

Samples were taken regularly and analyzed for optical density (OD) at 583 nm, cell dry weight (CDW), ammonia and gluconic acid concentrations, and PHA accumulation and molecular weight distribution.

The OD and ammonia concentration were determined using a Cedex Bio Analyzer (Roche Diagnostics, Switzerland). In addition, an estimation of cell concentration was obtained by measuring the CDW. To perform this, a precise cell culture volume (1 mL) was centrifugated by a Centrifuge 5415 (Eppendorf, Germany) at 13'000 rpm for 6 min and then the resulting pellet was dried in a vacuum concentrator (Concentrator Plus, Eppendorf, Germany) and finally weighted.

Concerning the gluconic acid, its concentration was determined via high performance liquid chromatography (HPLC) (Agilent Technologies 1200 series, Germany) at room temperature using an Aminex HPX-87 H Ion Exclusion Column (Bio-Rad, USA) and an isocratic elution at 0.7 mL/min with 14 mM H_2SO_4 as mobile phase. UV absorbance was carried with a diode array detector set at 210 nm.

PHA extractions were performed with chloroform using a Soxhlet apparatus at 61 °C for 5.5 h coupled with solvent evaporation under reduced pressure (Rotavapor© R-300, Buchi, Switzerland). The PHA concentration was then determined using the crotonic acid method (Karr et al., 1983). Briefly, PHA containing samples were digested in concentrated H_2SO_4 (98%) for 30 min at 95 °C before dilution by mixing with 100 folds milliQ water. The crotonic acid was measured via HPLC with UV detection at 210 nm (Agilent Technologies 1200 series, Germany) using an Aminex HPX-87 H Ion Exclusion Column. The elution was carried out at 0.7 mL/min using 14 mM H_2SO_4 as the mobile phase. PHA molecular weight distribution was determined by size exclusion chromatography (SEC) performed through an Agilent 1100 GPC/SEC (Agilent, USA) unit equipped with two columns (PSS PFG linear M columns) connected to a refractive index detector. The elution was performed in hexafluoro isopropanol (HFIP) at 1 mL/min at 35 °C. The instrument was calibrated with polymethyl methacrylate (PMMA) standards (Tan et al., 2014).

2.3. Multivariate data analysis pipeline

In order to develop a functioning soft sensor based on spectroscopic data, the Raman spectra collected during the cell culture need to undergo a multivariate data analysis pipeline. This comprises different stages. It starts with the collection and preparation of the data set, including both spectra and offline reference measurements, and the matching of the spectra to the reference values since normally spectra are collected every 15 min while the reference values are only available every few hours. After these two initial steps, the spectra undergo a series of preprocessing steps, in order to remove most of the noise and increase the relevant signal. Based on these preprocessed spectra, a model can be developed which correlates the spectra to the variables to be predicted. In this work, a partial least squares (PLS) model was

employed. Furthermore, a wavenumber selection step was added, initially using the so-called "bio importance regions" and afterwards using the variable importance as a selection metric, as described later in this paper.

2.3.1. Spectra matching

Since the spectra are acquired at specific time intervals, *i.e.* 15 min, while the variable reference values from offline analysis are only available at the end of the working day, a proper time matching has to be performed. Two possibilities were explored: single spectrum matching and multiple spectra matching.

This leads to a dataset where each sample has one or more Raman spectra attached. In fact, single spectrum matching corresponds to the matching of a single spectrum to each reference value. This is done via a list search algorithm which matches the timestamp of the two elements. The use of single spectrum matching leads to discarding a high quantity of spectra. In fact, the *in situ* Raman spectroscopy generated many more spectra than the samples collected for offline analysis. Therefore, with this approach, all the Raman spectra recorded throughout the day and not associated to any reference value are discarded.

This can be avoided through a multispectra matching. In this work, this was implemented by matching each reference sample data not only to the spectrum attributed in a single matching phase but also to the 4 temporarily closest spectra. Hence, to the timestamp corresponding to the j reference measurement, a single matched spectrum i, and the spectra $i-2$, $i-1$, $i+1$ and $i+2$ were related. The only exceptions are when the first and second spectra are involved. The two strategies are schematically illustrated in Fig. 2. The main idea behind this approach exploits the fact that biological systems are characterized by slow kinetics of cellular growth and substrate consumption, as well as product accumulation. In this way, the dataset is augmented by an approximate factor of 5, thus leading to a more accurate and robust model.

2.3.2. Preprocessing pipeline

The first step of the preprocessing pipeline is wavenumber (feature) selection. In this work, two different approaches were developed. The first one consists of the removal of the non-informative wavenumbers based on the so-called "bio importance regions", which are selected based on the concept behind the fingerprint-like signatures of specific biological and chemical species (Pezzotti, 2021). Here we should eliminate spectra corrupted by water and window material interferences, as well as non-informative regions dictated by bioprocess modeling experience. This leads to the selection of the spectral ranges of 450 – 1820, 1880 – 2530, and 2590 – 3100 cm^{-1}, as reported in (Feidl et al., 2019a, 2019b). The second approach consists in the use of the variable importance (VIP) (Andersen and Bro, 2010) as a feature selection metric. This metric can be calculated based on the PLS scores and, as a rule, the wavenumbers with VIP higher than 1 would be considered important and used as features for model calibration.

The next step of the pipeline consisted of a Savitzky-Golay (SG) filter (Savitzky and Golay, 1964), the goal of which is to smooth noisy data obtained from chemical spectra analyzers. Specifically, for Raman spectroscopy, this filter is needed due to the presence of high-frequency peaks (Schafer, 2011). This technique was selected for its ability to retain the signal shape and, compared to other filters, to provide better performance (Chen et al., 2004; Acharya et al., 2016). The implementation of SG smoothing is accomplished by choosing in a convenient way both the polynomial degree and the working frame size. Usually, their values are estimated through a try-and-error methodology (Acharya et al., 2016).

Since spectra may have been recorded under different analyzer conditions, it is important to equalize the impact of data on model accuracy. For Raman spectroscopy, normalization is a scaling technique applied on the intensity of each wavenumber throughout the entire sampling. In this case, the Standard Normal Variate (SNV) scaler was applied (Barnes et al., 1989). Furthermore, in order to avoid giving higher importance to different wavenumbers because of their characteristically higher intensities, it is important to normalize the dataset, hence mean centering was applied wavenumber-wise (van den Berg et al., 2006).

In this work, initially the hyperparameter values of the SG filter and the application or not of the normalization and scaling approaches were

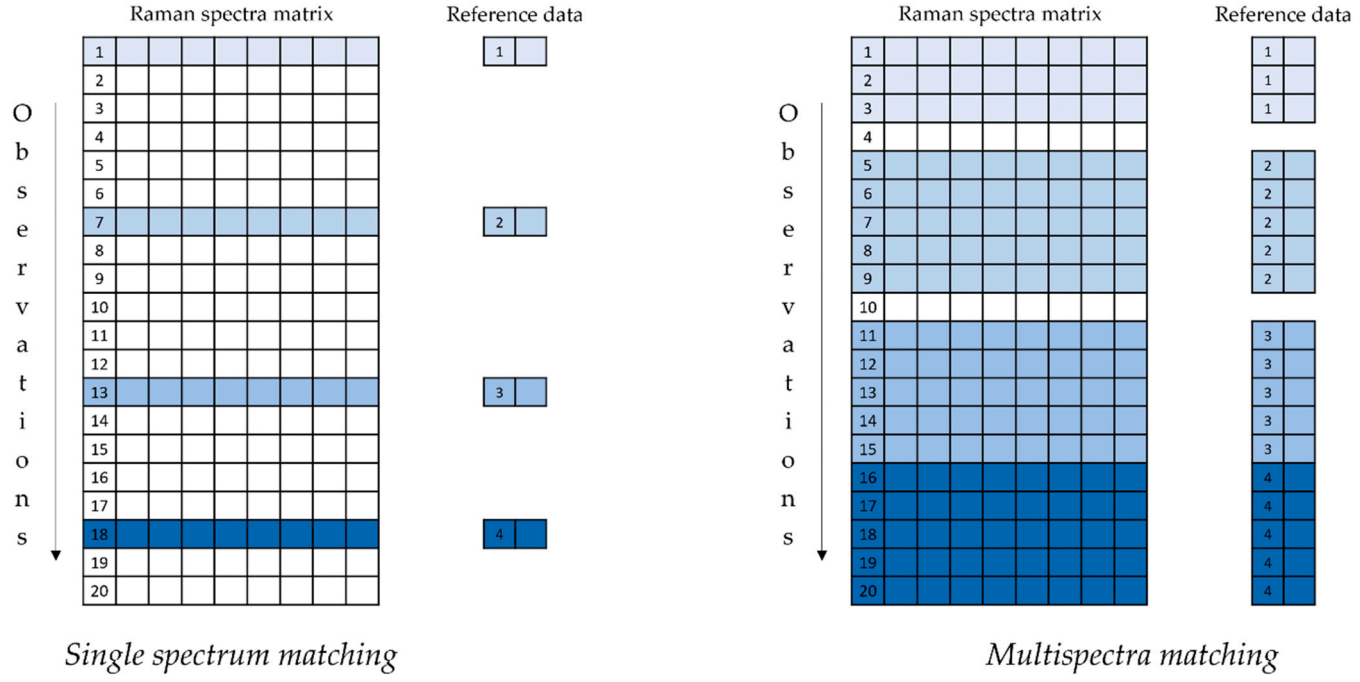

Fig. 2. Schematic comparison between the single and the multispectra matching approaches. In this scheme, the different colors correspond to different samples taken during the process and characterized offline for metabolite concentrations, PHA accumulation and MWD. Hence for single spectrum matching only one spectrum per reference measurement is selected while all others are discarded (blank lines). In multispectra matching instead, the majority of the spectra are retained.

selected based on expert guidance and, in a second approach, optimized based on the cross-validation error.

2.3.3. Partial least squares model

A partial least squares (PLS) model (Wold et al., 2001) was used as regressor and predictor in this work. This model is widely used for chemometrics applications (Feidl et al., 2019a, 2019b; Rajalahti and Kvalheim, 2011; Teixeira et al., 2009). Indeed, PLS is suitable for the analysis of data with strongly collinear and noisy predictor variables. In addition, due to the fact that it is a space reduction technique, it is particularly employed when the number of features (wavenumbers) highly surpasses the number of predictive variables. In a general way, the basic idea behind PLS is that it forms orthogonal score vectors (called latent vectors or components) by maximizing the covariance between different sets of variables. In the case of regression, the blocks of variables are the aforementioned predictors (factors) and responses (offline measurements). Thus, PLS is able to extract score vectors which are used as new predictors representation and regress responses on these new ones (Saunders et al., 2006). Therefore, in this model approach, the number of components to be used needs to be selected. In fact, too many components will lead to an overfitted model, while a too low number of them may lead to an imprecise model. Even if, in principle, it is possible to evaluate as many PLS components as the rank of predictors variables, just a part of them is used, because the measured data are rich in noise and some of the smaller components will describe only this disturbance. Therefore, a method must be employed to select a suitable number of components (Geladi and Kowalski, 1986). As reported by Kvalheim et al. (Kvalheim et al., 2018), there are several approaches to accomplish this selection: cross-validation, Monte Carlo and F-test (Geladi and Kowalski, 1986) are just a few of them. An easy and powerful approach is to use a minimum number of components to account for most of the cumulative variance expressed in response values, and this technique has been adopted in this work (Wold et al., 2001).

2.3.4. Hyperparameter optimization and cross-validation

In order to select the best set of hyperparameters for the preprocessing pipeline, an optimization algorithm combined with a cross-validation (CV) step was developed. Specifically, Bayesian optimization together with a 5-fold cross-validation approach was done. Named after Bayes' theorem (Brochu et al., 2010), Bayesian optimization is an iterative algorithm made up of two key factors: a probabilistic surrogate model and an acquisition function. The core idea of Bayesian optimization is that in each iteration, the surrogate model is fitted to all the observations of the target function collected until that point. Next, the acquisition function determines the next points to be calculated, using the prediction of the probabilistic surrogate model, automatically balancing the trade-off between exploration and exploitation (Yang and Shami, 2020; Lei, 2021). The acquisition function is selected so as to be cheap to compute and easy to be optimized (Lei, 2021), while traditionally, surrogate models employ Gaussian processes, due to their expressiveness, flexibility, and easy handling (Wu et al., 2019).

This Bayesian optimization algorithm was used to minimize the average root mean squared error (RMSE) in cross-validation. Specifically, 5-fold cross-validation was used, in which the training dataset is divided into 5 subsets, and each used as test set in a rotational manner, while the other 4 subsets are used for model training. This rotational approach leads to training and testing the model 5 different times, generating five different RMSE values. The RMSE in CV is then the average of these five errors.

Concluding, a schematic summary of the multivariate data analysis approach used in this work is shown in Fig. 3.

2.4. Algorithm implementation and error definition

All calculations were carried out on MATLAB 2021b (The Mathworks Inc., USA), using both built-in functions and in-house developed routines. Regarding the built-in functions, *bayesopt* was used for Bayesian optimization, *crossvalind* for determination of the k-fold subdivision, *plsregress* for PLS model development and *sgolayfilter* for the SG filter.

The RMSE and relative RMSE for variable *j* are defined as:

$$RMSE_j = \sqrt{\frac{1}{N_{data}} \sum_{i=1}^{N_{data}} (y_i - \hat{y}_i)^2}$$

$$RMSE_{rel} = 100\% \bullet \frac{\sqrt{\frac{1}{N_{data}} \sum_{i=1}^{N_{data}} (y_i - \hat{y}_i)^2}}{\sqrt{\frac{1}{N_{data}-1} \sum_{i=1}^{N_{data}} (y_i - \bar{y})^2}}$$

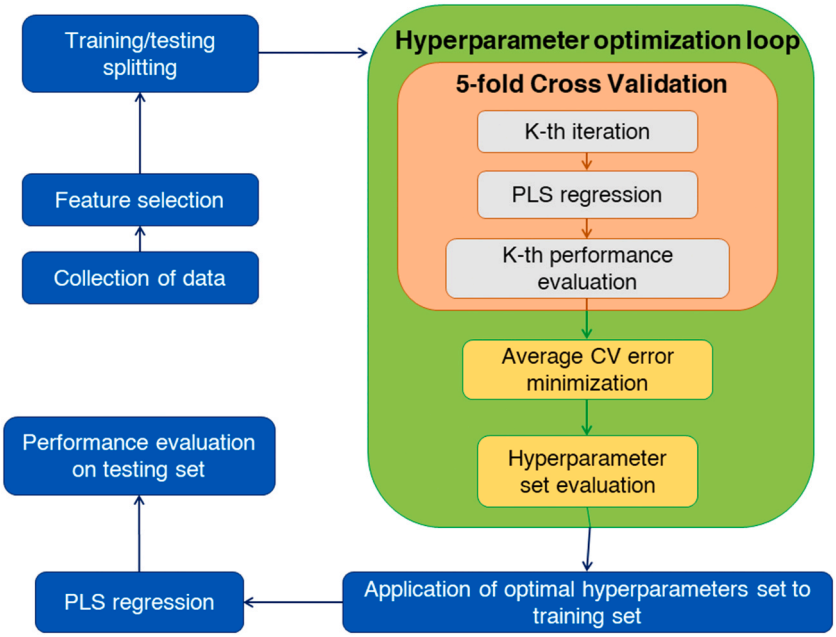

Fig. 3. Schematic representation of the multivariate data analysis pipeline employed in this work.

Where N_{data} is the number of samples of the selected variable, y_i, $\widehat{y}_i$ and $\overline{y}$ are the experimental, predicted, and mean values, respectively, and j represents the variable of interest.

3. Results and discussion

Raman spectroscopy is emerging as a powerful tool for the monitoring of biotechnology processes. Indeed, it combines a fast analysis with an in-depth characterization of the system. To access the full potential of this technique, we explored the possibility of applying it *in situ*, for the in-line characterization of PHA biomanufacturing from *Cupriavidus necator* without the need for sampling, which may introduce the risk of contamination, and labor-intensive sample preparations. The Raman spectra collected every 15 min from a perfusion bioreactor are stacked in Fig. 4.

The PHA accumulation during time can be tracked at 1735 cm^{-1} (García et al., 2019). However, the heterogeneity and complexity of the system prevents the unambiguous attribution of the bands to the different components. Hence, the extraction of quantitative information from this analysis requires the development of a suitable MVDA pipeline.

3.1. Single spectrum matching

The first task of model development was the creation of a correct dataset by establishing a suitable correspondence between Raman spectra (matrix of intensity values) and offline measurements of cell dry weight, metabolite concentration, and polymer molecular weight distribution.

The spectroscopy probe performed measurements with a specific time frequency (15 min, necessary to accumulate several scans and reduce the noise) while samples were taken with a lower frequency and analyzed offline. Therefore, in a preliminary approach based on a temporal scale, by comparing the two analysis times it was possible to correlate one spectrum to one sample following the approach of single spectrum matching.

After having matched the spectra and the samples, they were first analyzed without any preprocessing. In Table 1, the absolute and the relative RMSE values are shown for gluconic acid, CDW, ammonia, PHA accumulation, and molecular weight distribution, for both the training and testing set. As mentioned above, the entire available dataset, comprising 5 high-density perfusion runs, was divided using a random 80:20 split, in training and test datasets, respectively. The training dataset is used for model calibration, while the remaining data set is used for testing prediction accuracy. In this way, overfitting of the model can be detected, since the model was not calibrated on the test data set. These RMSE values show already the potential predictive power of Raman spectroscopy. However, they are still too high to be used as an in-line monitoring tool.

Table 1
RMSE and relative RMSE of prediction in training and test sets for all the variables using single spectrum matching and no preprocessing.

Variable	Train set		Test set	
	RMSE	Relative RMSE (%)	RMSE	Relative RMSE (%)
Gluconic acid	4.57 g/L	30.4	7.70 g/L	50.4
Cell dry weight	1.96 g/L	23.8	16.9 g/L	187
Ammonia	0.02 g/L	23.1	0.73 g/L	1071
PHA Accumulation	4.95%	22.5	21.5%	92.7
Mn	2.02×10^4 g/mol	18.5	9.27×10^4 g/mol	93.3
Mw	2.97×10^4 g/mol	21.52	1.20×10^5 g/mol	124

In order to improve the predictive performance of this tool, we inserted a data preprocessing step in the pipeline. With regards to spectroscopic data, several smoothing and scaling tools can be implemented. However, the most used set of tools are a Savitzky-Golay smoothing combined with SNV and mean centering as normalization methods. The main drawback of this step is the manual tuning of the architecture algorithm parameters, usually referred to as hyperparameters (HP). This procedure requires a high user experience combined with a great difficulty in managing different HP simultaneously, whose interactions are unknown a priori.

In this first preprocessing approach, the HP and combination of techniques were set via expert decision. In Table 2, the RMSE and relative RMSE values for this approach are shown. Here, it is possible to deduce that the preprocessing approach has brought a limited yet noticeable improvement in predictive potential for the CDW and ammonia concentration. In particular, for the latter, these preprocessing steps have reduced the relative RMSE from more than 1000% (indicating that this value could not be predicted at all) to 180%, which,

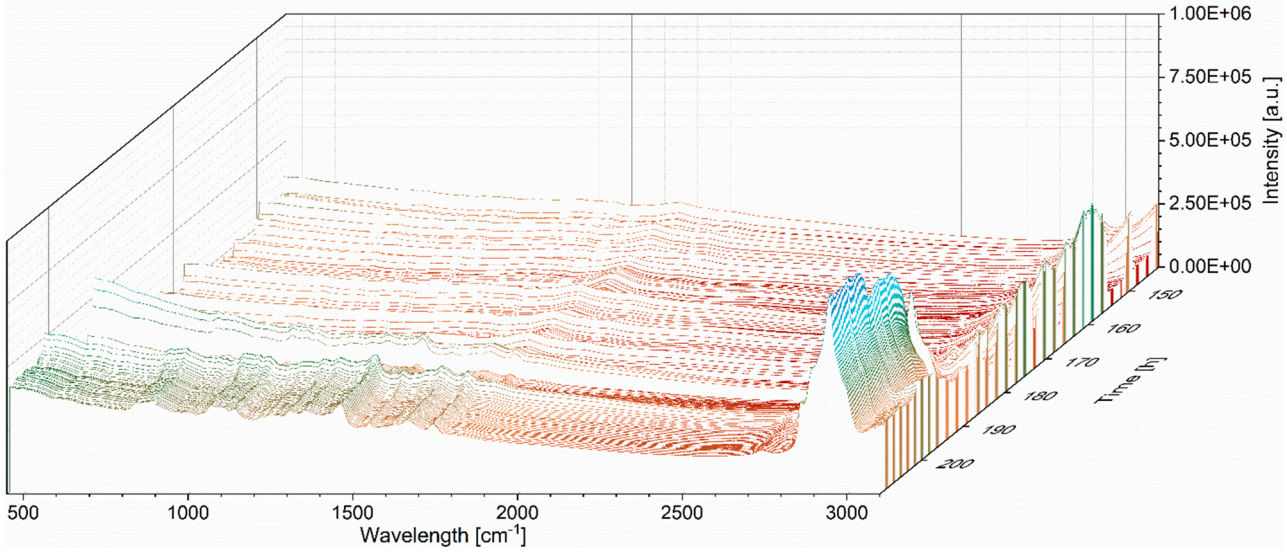

Fig. 4. Example of collection window of Raman spectra during a perfusion bioreactor experiment. Colors indicate the values of emission intensities: the lower intensities are in red, the higher in blue. The change in color of the signal at 1735 cm^{-1} indicates the increasing quantity of PHA produced for energy storage.

Table 2
RMSE and relative RMSE of prediction in training and test sets for all the variables using single spectrum matching and a default preprocessing approach.

Variable	Train set		Test set	
	RMSE	Relative RMSE (%)	RMSE	Relative RMSE (%)
Gluconic acid	6.68 g/L	44.4	21.7 g/L	142
Cell dry weight	2.46 g/L	29.8	14.5 g/L	162
Ammonia	0.02 g/L	30.6	0.12 g/L	181
PHA Accumulation	5.65%	25.7	35.6%	153
Mn	3.47×10^4 g/mol	31.8	8.99×10^5 g/mol	904
Mw	4.99×10^4 g/mol	36.2	9.41×10^5 g/mol	975

although still too high, indicates how Raman spectroscopy is able to detect it, even if not accurately. However, for all the remaining variables, the errors have increased. A potential cause for this aggravation could be a deficient choice of the hyperparameter values, emphasizing the need of an automated hyperparameter optimization.

In order to define an automatic preprocessing hyperparameter approach, an optimization loop was added to the predictive pipeline, using a Bayesian optimizer and 5-fold cross-validation approach. The main aim of cross-validation is to improve model robustness by exploiting the same amount of data while the objective of the Bayesian optimization is to minimize the average RMSE in cross-validation, defining an adequate pretreatment approach for the raw spectra. In the optimization loop, not only the Savitzky-Golay filter hyperparameters were considered, but also, in a Boolean fashion, the application of the scaling and centering methods.

In Table 3, the RMSE values obtained with this new approach are shown. Relative to cell dry weight and PHA recovery, the testing error dropped by 21% and 36% respectively, achieving a better prediction accuracy compared to the expert-guided preprocessing. The other variables estimations are also improved, with decreasing relative errors, although the number-average molecular weight Mn error remains significantly high. On the other hand, the weight-average molecular weight Mw predictions keep worsening, and despite the boost in preprocessing, the model continues to be somehow imprecise, and still not adequate for advanced control applications.

A possible justification might be related to the low number of data used to train the model: with respect to big data framework (Zhou et al., 2017), in this case the data volume is limited to a bench-scale biochemical configuration. An additional concern comes from the collected Raman intensities: each spectrum refers in fact to a small sample and therefore poorly representative of the whole reactor volume. To mitigate this drawback, the multispectra approach was adopted.

3.2. Multispectra matching

In order to have a more complete picture of the whole volume of the bioreactor, which is expected to compensate the fluctuations from point to point and improve model robustness, a larger number of Raman spectra associated to a given sample was used, adopting the multispectra matching procedure described above. Differently to the single spectrum approach, the main idea of multispectra matching is to create a correspondence between more than one spectrum to the same reference value. This obviously increases the size of data and introduces some kind of temporal-mean on Raman intensities. By combining this promising approach with the optimized preprocessing step discussed before, the procedure provided outstanding results, obtaining the lowest error for all the variables as reported in Table 4.

With this approach, it is possible to reliably predict important variables associated to the cell culture and to the polymer quality, thus providing a promising tool for in-line quality monitoring and control.

3.3. Feature selection using VIP

In order to increase the model robustness and reduce noise in the in-line implementation of the developed soft sensor, it is worthwhile to conduct a feature selection procedure. Here, a metric such as variable importance (Mehmood et al., 2020) can be used to select the wavenumbers which carry more useful information for each predicted variable instead of using all the variables contained in the so-called "bio importance regions". The reason behind choosing VIP as feature selection mainly lies on the use of supervised feature selection over unsupervised ones in labeled data problems. As already declared, this preference leads to lower model prediction errors (Cho et al., 2008).

In Fig. 5, the VIP variable importance plot is shown as an example for the case of PHA accumulation. It is seen that the different Raman wavenumbers are not equally important for the prediction of PHA accumulation. In particular, by selecting only the wavenumbers with a VIP score higher than one, a more robust model would be obtained, since the noise and non-PHA-related signal interference would be eliminated. In general, all variables with a VIP score closer (or greater) than 1 (in red in Fig. 5) are considered in developing the model. In this manner, VIP scores estimation is a suitable tool for feature selection when PLS regression is employed.

With this approach, it would then be possible to develop a lighter yet effective multivariate data analysis pipeline, using a reduced number of predictor variables. In Table 5, the RMSE values obtained for all measured variables using this approach are shown. Compared to the previous approach, it is evident how the results have slightly either improved or worsened, being quite similar in most of the variables: in the case of Mw, a weak yet important improvement is obtained (−4% on testing set error) while for PHA recovery and CDW the relative error has increased by 9% and 1%, respectively, which could be argued to be

Table 3
RMSE and relative RMSE of prediction in training and test set for all the variables using single spectrum matching and an optimized preprocessing step, via Bayesian optimization and cross-validation.

Variable	Train set		Test set	
	RMSE	Relative RMSE (%)	RMSE	Relative RMSE (%)
Gluconic acid	8.59 g/L	57.1	20.55 g/L	134
Cell dry weight	1.76 g/L	21.4	12.63 g/L	140
Ammonia	0.03 g/L	38.3	0.14 g/L	210
PHA Accumulation	8.22%	37.3	27.29%	117
Mn	2.77×10^4 g/mol	25.4	6.64×10^5 g/mol	667
Mw	6.17×10^4 g/mol	44.8	9.71×10^5 g/mol	1005

Table 4
RMSE and relative RMSE of prediction in training and test sets for all the variables using multispectra matching and an optimized preprocessing step, via Bayesian optimization and cross-validation.

Variable	Train set		Test set	
	RMSE	Relative RMSE (%)	RMSE	Relative RMSE (%)
Gluconic acid	5.46 g/L	36.8	6.27 g/L	40.3
Cell dry weight	3.78 g/L	45.0	3.63 g/L	43.7
Ammonia	0.03 g/L	35.5	0.03 g/L	42.8
PHA Accumulation	11.7%	51.2	12.5%	76.9
Mn	5.75×10^4 g/mol	54.9	7.44×10^4 g/mol	66.9
Mw	9.12×10^4 g/mol	73.5	1.23×10^5 g/mol	77.7

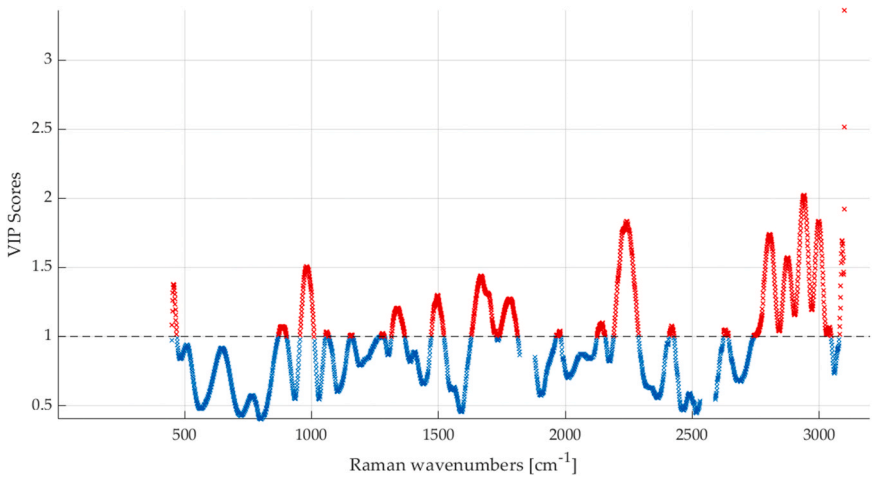

Fig. 5. Variable importance scores vs the Raman wavenumbers as predictor variables for PHA accumulation. Wavenumbers with a VIP score higher than one are shown in red.

Table 5
RMSE and relative RMSE of prediction in training and test sets for all the variables using multispectra matching and an optimized preprocessing step, via Bayesian optimization and cross-validation, together with variable importance as wavenumber selection tool.

Variable	Train set		Test set	
	RMSE	Relative RMSE (%)	RMSE	Relative RMSE (%)
Gluconic acid	5.54 g/L	37.3	7.00 g/L	45.0
Cell dry weight	3.67 g/L	43.7	3.71 g/L	44.8
Ammonia	0.03 g/L	41.5	0.03 g/L	46.2
PHA Accumulation	11.0%	48.1	14.1%	86.1
Mn	6.85×10^4 g/mol	65.3	8.72×10^4 g/mol	78.4
Mw	8.87×10^4 g/mol	70.8	1.16×10^5 g/mol	73.4

inside the margin of error of the analytical method.

On the other hand, this approach has several and important advantages, such as performing a higher-in-detail feature selection, employing a more robust and less prone-to-noise model, and highlighting the informative content of specific Raman wavelengths.

With this lighter model it is then possible to speed up the prediction stage, thus providing an important tool for the implementation of a prompt control strategy smoothing down the fluctuations at the manufacturing level.

In Fig. 6, the observed vs predicted plots (or parity plots) for all variable predictions in the test set are shown. As evident from these plots, the overall predictions are sufficiently in line with the observed values, underlying the ability of the developed algorithm for quantification of extra- and intracellular substances.

Overall, although the error in the prediction is larger than those reported from other examples in using Raman spectroscopy for the characterization of PHA production, it is worth highlighting that in these previous works samples taken from the bioreactor are pre-processed to reduce disturbance from the turbid medium and interference with signals not related to the polymer (Samek et al., 2016; De Gelder et al., 2008; Tao et al., 2016; Hermann et al., 2020). With our approach instead, the signal is collected *in situ* through the use of immersion probes located directly in the bioreactor. This leads to important advantages related to the reduced necessity of human intervention, avoided risk of contamination and spectra collected at higher frequency. Still, as confirmed by the parity plots, the model developed can reproduce the trend of most of the predicted variables, thus providing an efficient and useful method for a prompt reaction to process fluctuations.

3.4. Prediction of perfusion reactor variables

To demonstrate the predictivity of the soft sensor developed in this work, the experimental evolution for all the considered variables during a fermentation process is compared with the model predictions based on the signals collected by *in situ* Raman spectroscopy in Fig. 7.

The model accuracy varies depending on the variable of interest, with a RMSE as reported for each attribute in Table 5 (see test set). While the PHA accumulation can be poorly predicted, as also confirmed by the parity plot in Fig. 6, the predictions of the cell dry weight are in very good agreement with the experimental measurements, displaying the lag, exponential, and stationary phases expected during the cell growth. In addition, the high variance reported by this prediction underlines the robustness of the developed algorithm and the absence of model overfitting. The trend of gluconic acid consumption is also nicely predicted, although with an underestimation of the residual concentration. The monotonically decreasing concentration of gluconic acid is perceived as a complete depletion of the carbon source, although empirical values demonstrated a steady-state concentration of 5 g/L.

Very promising results were obtained in the prediction of the molecular weight distribution of PHA, which is reported for the first time in this work and is even more valuable considering the intracellular nature of the product. The slight overestimation observed for Mn must indeed be contextualized in the bacterial culture monitoring, which can easily create a noisy and disturbed signal. However, the reliable trend and proper quantification of Mw, which is the parameter considered for the final application of the polymer, confirm the possibility of employing the soft sensor developed in this work for the in-line prediction of the polymer quality and for a prompt intervention to keep this property at the desired set-point during a manufacturing campaign.

4. Conclusion

In this work, Raman spectroscopy has been demonstrated to be a valuable tool for the in-line monitoring of PHA manufacturing. In particular, thanks to a suitable multivariate data analysis pipeline, characterized by an optimized preprocessing step, it is possible to achieve sufficiently accurate predictions of several process variables, ranging from metabolite concentrations to cell dry weight to polymer quality (Mn and Mw). It is to be noted that it is quite reasonable to expect that this list could be prolonged by introducing more variables. This work underlines the importance of spectroscopic sensors for in-line

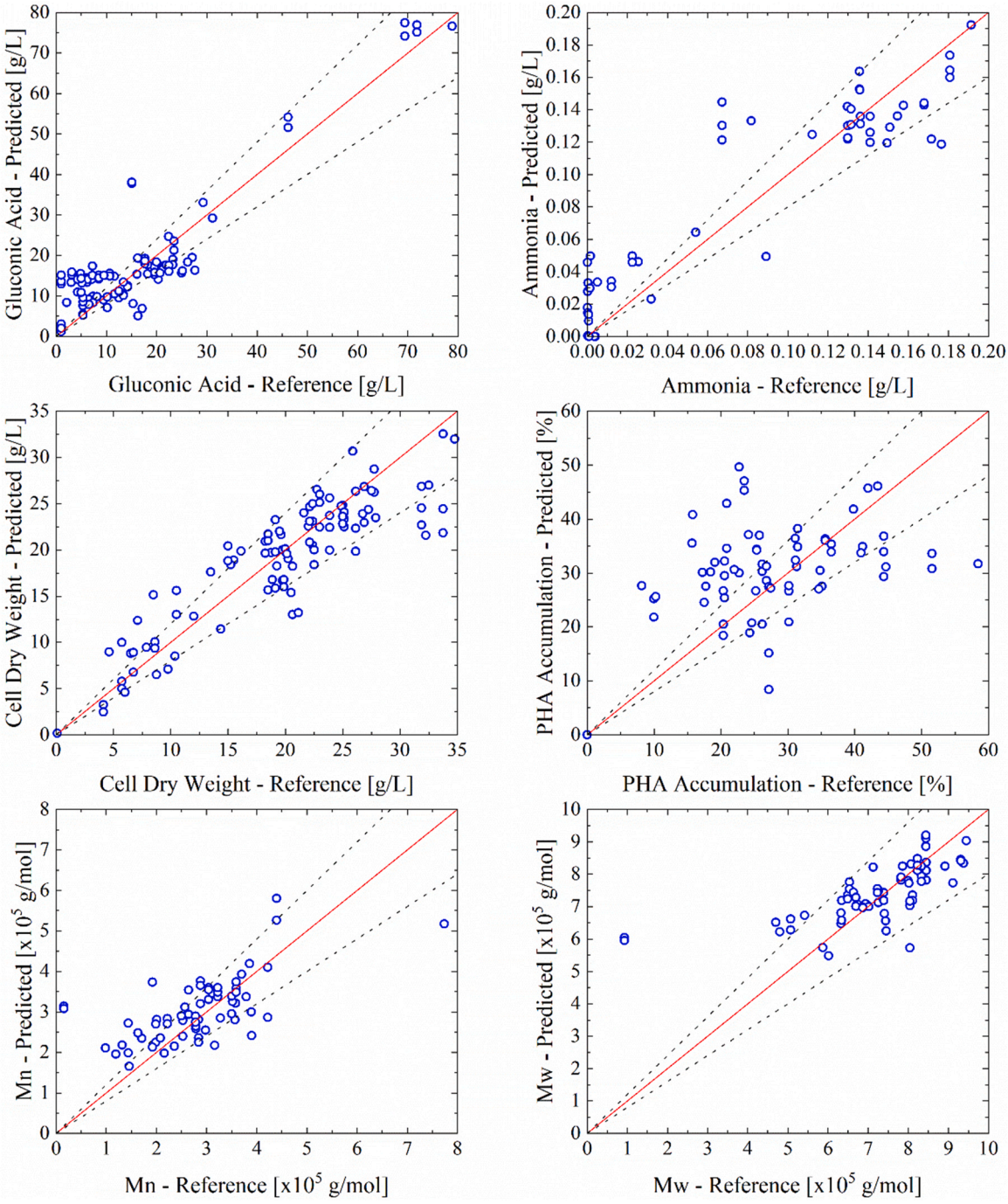

Fig. 6. Parity plots for the predictions in test set for all the variables using multispectra matching and an optimized preprocessing step, via Bayesian optimization and cross-validation, together with variable importance as wavenumber selection tool. The dashed lines represent an error of ±20% with respect to the reference values.

applications, which, although requiring higher investment costs, allow the simultaneous monitoring of multiple process variables. In this context, it should be underlined the remarkable result, reported here for the first time, to measure in-line through a spectroscopic method the molecular weight distribution of a biopolymer inside a bacterial cell. A further improvement of the performance of this sensor is expected when combining it with suitable mechanistic or hybrid models, for example in the frame of extended Kalman filters (Narayanan et al., 2020). Furthermore, the results reported in this work can be further extended and utilized for the monitoring of other processes, since the data analysis framework is agnostic to measured variables.

CRediT authorship contribution statement

João Medeiros Garcia Alcântara: Investigation, Data curation, Writing – original draft. **Francesco Iannacci:** Validation, Investigation, Writing – review & editing. **Massimo Morbidelli:** Supervision, Writing – review & editing. **Mattia Sponchioni:** Supervision, Project administration, Writing – review & editing.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

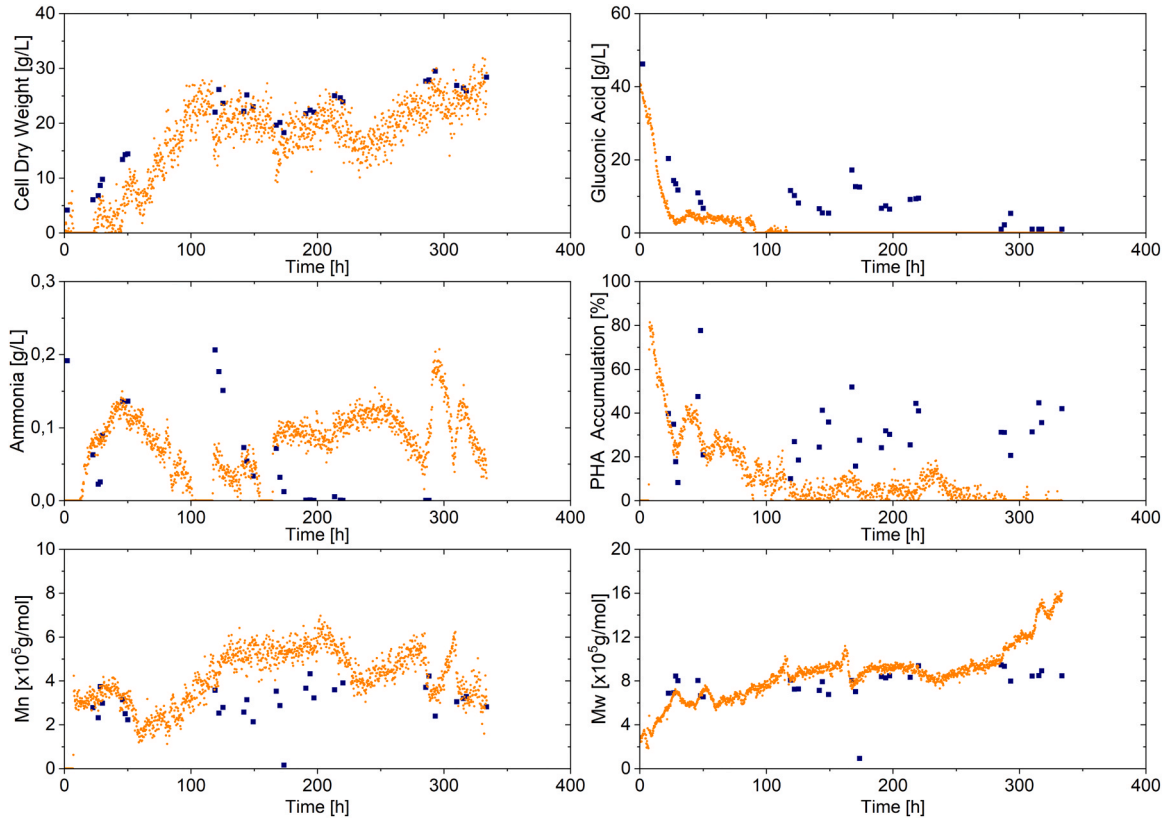

Fig. 7. Evolution of the main process variables during time for a reference bioreactor, comparing experimental results (blue squares) and model predictions (orange circles). The model was trained using a multispectra matching, and an optimized pre-processing and variable importance projection as feature selection.

Joao Medeiros Garcia Alcantara reports financial support was provided by Horizon Europe.

Data availability

Data will be made available on request.

Acknowledgments

JMGA acknowledges the financial support by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 812909 CODOBIO, within the Marie Skłodowska-Curie International Training Networks framework. The authors acknowledge Paolo Trotti and Francesco Bonfanti for their assistance in acquiring the experimental dataset used in this work.

References

Abu-Absi, N.R., et al., 2011. Real time monitoring of multiple parameters in mammalian cell culture bioreactors using an in-line Raman spectroscopy probe. Biotechnol. Bioeng. vol. 108 (5), 1215–1221. https://doi.org/10.1002/bit.23023.

Acharya, D., Rani, A., Agarwal, S., Singh, V., 2016. Application of adaptive Savitzky–Golay filter for EEG signal processing. Perspect. Sci. vol. 8, 677–679. https://doi.org/10.1016/j.pisc.2016.06.056.

Albuquerque, P.B.S., Malafaia, C.B., 2018. Perspectives on the production, structural characteristics and potential applications of bioplastics derived from polyhydroxyalkanoates. Int. J. Biol. Macromol. vol. 107, 615–625. https://doi.org/10.1016/j.ijbiomac.2017.09.026.

Andersen, C.M., Bro, R., 2010. Variable selection in regression-a tutorial. J. Chemom. vol. 24 (11–12), 728–737. https://doi.org/10.1002/cem.1360.

Andreasi Bassi, S., Boldrin, A., Frenna, G., Astrup, T.F., 2021. An environmental and economic assessment of bioplastic from urban biowaste. The example of polyhydroxyalkanoate. Bioresour. Technol. vol. 327, 124813 https://doi.org/10.1016/j.biortech.2021.124813.

Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Appl. Spectrosc. vol. 43 (5), 772–777. https://doi.org/10.1366/0003702894202201.

Beckett, C., Eriksson, L., Johansson, E., Wikström, C., 2018. Multivariate data analysis (MVDA). Pharmaceutical Quality by Design. Wiley, pp. 201–225. https://doi.org/10.1002/9781118895238.ch8.

van den Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A., Smilde, A.K., van der Werf, M.J., 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genom. vol. 7 (1), 142. https://doi.org/10.1186/1471-2164-7-142.

Blunt, W., Levin, D., Cicek, N., 2018. Bioreactor operating strategies for improved polyhydroxyalkanoate (PHA) productivity. Polymers vol. 10 (11), 1197. https://doi.org/10.3390/polym10111197.

Brochu, E., Cora, V.M., de Freitas, N., 2010. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning.

Chen, J., Jönsson, Per, Tamura, M., Gu, Z., Matsushita, B., Eklundh, L., 2004. A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky–Golay filter. Remote Sens. Environ. vol. 91 (3–4), 332–344. https://doi.org/10.1016/j.rse.2004.03.014.

Cho, H.W., et al., 2008. Discovery of metabolite features for the modelling and analysis of high-resolution NMR spectra. Int. J. Data Min. Bioinform. vol. 2 (2), 176. https://doi.org/10.1504/IJDMB.2008.019097.

Colosimo, B.M., Grasso, M., 2020. In-situ monitoring in L-PBF: opportunities and challenges. Procedia CIRP vol. 94, 388–391. https://doi.org/10.1016/j.procir.2020.09.151.

Das, R.S., Agrawal, Y.K., 2011. Raman spectroscopy: recent advancements, techniques and applications. Vib. Spectrosc. vol. 57 (2), 163–176. https://doi.org/10.1016/j.vibspec.2011.08.003.

Doppler, P., Gasser, C., Kriechbaum, R., Ferizi, A., Spadiut, O., 2021. In situ quantification of polyhydroxybutyrate in photobioreactor cultivations of synechocystis sp. using an ultrasound-enhanced ATR-FTIR spectroscopy probe. Bioengineering vol. 8 (9), 129. https://doi.org/10.3390/bioengineering8090129.

Duvigneau, S., Dürr, R., Wulkow, M., Kienle, A., 2022. Multiscale modeling of the microbial production of polyhydroxyalkanoates using two carbon sources. Comput. Chem. Eng. vol. 160, 107740 https://doi.org/10.1016/j.compchemeng.2022.107740.

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

Feidl, F., et al., 2019a. A new flow cell and chemometric protocol for implementing in-line Raman spectroscopy in chromatography. Biotechnol. Prog. vol. 35 (5) https://doi.org/10.1002/btpr.2847.

Feidl, et al., 2019b. Combining mechanistic modeling and Raman spectroscopy for monitoring antibody chromatographic purification. Processes vol. 7 (10), 683. https://doi.org/10.3390/pr7100683.

García, C., Alcaraz, W., Acosta-Cárdenas, A., Ochoa, S., 2019. Application of process system engineering tools to the fed-batch production of poly(3-hydroxybutyrate-co-3-hydroxyvalerate) from a vinasses–molasses Mixture. Bioprocess Biosyst. Eng. vol. 42 (6), 1023–1037. https://doi.org/10.1007/s00449-019-02102-z.

Geladi, P., Kowalski, B.R., 1986. Partial least-squares regression: a tutorial. Anal. Chim. Acta vol. 185, 1–17. https://doi.org/10.1016/0003-2670(86)80028-9.

De Gelder, J., et al., 2008. Monitoring poly(3-hydroxybutyrate) production in Cupriavidus necator DSM 428 (H16) with Raman spectroscopy. Anal. Chem. vol. 80 (6), 2155–2160. https://doi.org/10.1021/ac702185d.

Gernaey, K., Bolic, A., Svanholm, B., 2012. PAT tools for fermentation processes. Chim. Oggi Chem. Today vol. 30 (3), 38–43.

Gonzalez Zelaya, C.V., 2019. Towards explaining the effects of data preprocessing on machine learning. In: Proceedings of the IEEE 35th International Conference on Data Engineering (ICDE). IEEE, pp. 2086–2090, 10.1109/ICDE.2019.00245.

Grey, C.P., Tarascon, J.M., 2017. Sustainability and in situ monitoring in battery development. Nat. Mater. vol. 16 (1), 45–56. https://doi.org/10.1038/nmat4777.

Gutschmann, B., Schiewe, T., Weiske, M.T.H., Neubauer, P., Hass, R., Riedel, S.L., 2019. In-line monitoring of polyhydroxyalkanoate (PHA) production during high-cell-density plant oil cultivations using photon density wave spectroscopy. Bioengineering vol. 6 (3), 85. https://doi.org/10.3390/bioengineering6030085.

Gutschmann, B., et al., 2023. Continuous feeding strategy for polyhydroxyalkanoate production from solid waste animal fat at laboratory- and pilot-scale. Microb. Biotechnol. vol. 16 (2), 295–306. https://doi.org/10.1111/1751-7915.14104.

Hermann, D.R., et al., 2020. In situ based surface-enhanced Raman spectroscopy (SERS) for the fast and reproducible identification of PHB producers in cyanobacterial cultures. Analyst vol. 145 (15), 5242–5251. https://doi.org/10.1039/d0an00969e.

Hinz, D.C., 2006. Process analytical technologies in the pharmaceutical industry: the FDA's PAT initiative. Anal. Bioanal. Chem. vol. 384 (5), 1036–1042. https://doi.org/10.1007/s00216-005-3394-y.

Hisazumi, J., Kleinebudde, P., 2017. In-line monitoring of multi-layered film-coating on pellets using Raman spectroscopy by MCR and PLS analyses. Eur. J. Pharm. Biopharm. vol. 114, 194–201. https://doi.org/10.1016/j.ejpb.2017.01.017.

https://plasticseurope.org/knowledge-hub/plastics-the-facts-2022/.

Ivar do Sul, J.A., Costa, M.F., 2014. The present and future of microplastic pollution in the marine environment. Environ. Pollut. vol. 185, 352–364. https://doi.org/10.1016/j.envpol.2013.10.036.

Jarute, G., Kainz, A., Schroll, G., Baena, J.R., Lendl, B., 2004. On-line determination of the intracellular poly(β-hydroxybutyric acid) content in transformed Escherichia coli and glucose during PHB production using stopped-flow attenuated total reflection FT-IR spectrometry. Anal. Chem. vol. 76 (21), 6353–6358. https://doi.org/10.1021/ac049803l.

Karr, D.B., Waters, J.K., Emerich, D.W., 1983. Analysis of poly-β-hydroxybutyrate in *Rhizobium japonicum* bacteroids by ion-exclusion high-pressure liquid chromatography and UV detection. Appl. Environ. Microbiol. vol. 46 (6), 1339–1344. https://doi.org/10.1128/aem.46.6.1339-1344.1983.

Kirstein, I.V., Gomiero, A., Vollertsen, J., 2021. Microplastic pollution in drinking water. Curr. Opin. Toxicol. vol. 28, 70–75. https://doi.org/10.1016/j.cotox.2021.09.003.

Kvalheim, O.M., Arneberg, R., Grung, B., Rajalahti, T., 2018. Determination of optimum number of components in partial least squares regression from distributions of the root-mean-squared error obtained by Monte Carlo resampling. J. Chemom. vol. 32 (4), e2993 https://doi.org/10.1002/cem.2993.

Lei, C., 2021. Automated Machine Learning, pp. 245–284. https://doi.org/10.1007/978-981-16-2233-5_11.

López-Abelairas, M., García-Torreiro, M., Lú-Chau, T., Lema, J.M., Steinbüchel, A., 2015. Comparison of several methods for the separation of poly(3-hydroxybutyrate) from Cupriavidus necator H16 cultures. Biochem Eng. J. vol. 93, 250–259. https://doi.org/10.1016/j.bej.2014.10.018.

Madden, M.G., Ryder, A.G., 2003. In: Glynn, T.J. (Ed.), Machine Learning Methods For Quantitative Analysis of Raman Spectroscopy Data, p. 1130. https://doi.org/10.1117/12.464039.

Medeiros Garcia Alcântara, J., Distante, F., Storti, G., Moscatelli, D., Morbidelli, M., Sponchioni, M., 2020. Current trends in the production of biodegradable bioplastics: the case of polyhydroxyalkanoates. Biotechnol. Adv. vol. 42, 107582 https://doi.org/10.1016/j.biotechadv.2020.107582.

Medeiros Garcia Alcântara, J., Sponchioni, M., 2022. Evolution and design of continuous bioreactors for the production of biological products. Adv. Chem. Eng. 1–26. https://doi.org/10.1016/bs.ache.2022.03.001.

Meereboer, K.W., Misra, M., Mohanty, A.K., 2020. Review of recent advances in the biodegradability of polyhydroxyalkanoate (PHA) bioplastics and their composites. Green Chem. vol. 22 (17), 5519–5558. https://doi.org/10.1039/D0GC01647K.

Mehmood, T., Sæbø, S., Liland, K.H., 2020. Comparison of variable selection methods in partial least squares regression. J. Chemom. vol. 34 (6) https://doi.org/10.1002/cem.3226.

Mulvaney, S.P., Keating, C.D., 2000. Raman spectroscopy. Anal. Chem. vol. 72 (12), 145–158. https://doi.org/10.1021/a10000155.

Narayanan, H., Sponchioni, M., Morbidelli, M., 2022. Integration and digitalization in the manufacturing of therapeutic proteins. Chem. Eng. Sci. vol. 248, 117159 https://doi.org/10.1016/j.ces.2021.117159.

Narayanan, H., et al., 2020. Hybrid-EKF: hybrid model coupled with extended Kalman filter for real-time monitoring and control of mammalian cell culture. Biotechnol. Bioeng. vol. 117 (9), 2703–2714. https://doi.org/10.1002/bit.27437.

Ochoa, S., García, C., Alcaraz, W., 2020. Real-time optimization and control for polyhydroxybutyrate fed-batch production at pilot plant scale. J. Chem. Technol. Biotechnol. vol. 95 (12), 3221–3231. https://doi.org/10.1002/jctb.6500.

Pezzotti, G., 2021. Raman spectroscopy in cell biology and microbiology. J. Raman Spectrosc. vol. 52 (12), 2348–2443. https://doi.org/10.1002/jrs.6204.

Pienack, N., Bensch, W., 2011. In-situ monitoring of the formation of crystalline solids. Angew. Chem. Int. Ed. vol. 50 (9), 2014–2034. https://doi.org/10.1002/anie.201001180.

Radtke, J., Rehbaum, H., Kleinebudde, P., 2020. Raman spectroscopy as a PAT-tool for film-coating processes: in-line predictions using one PLS model for different cores. Pharmaceutics vol. 12 (9), 796. https://doi.org/10.3390/pharmaceutics12090796.

Rajalahti, T., Kvalheim, O.M., 2011. Multivariate data analysis in pharmaceutics: a tutorial review. Int. J. Pharm. vol. 417 (1–2), 280–290. https://doi.org/10.1016/j.ijpharm.2011.02.019.

Raza, Z.A., Abid, S., Banat, I.M., 2018. Polyhydroxyalkanoates: characteristics, production, recent developments and applications. Int. Biodeterior. Biodegrad. vol. 126, 45–56. https://doi.org/10.1016/j.ibiod.2017.10.001.

Rowland-Jones, R.C., Jaques, C., 2019. At-line Raman spectroscopy and design of experiments for robust monitoring and control of miniature bioreactor cultures. Biotechnol. Prog. vol. 35 (2), e2740 https://doi.org/10.1002/btpr.2740.

Samek, O., et al., 2016. Quantitative Raman spectroscopy analysis of polyhydroxyalkanoates produced by Cupriavidus necator H16. Sensors vol. 16 (11), 1808. https://doi.org/10.3390/s16111808.

Saunders, C., Grobelnik, M., Gunn, S., Shawe-Taylor, J., 2006. Subspace, Latent Structure and Feature Selection, vol. 3940. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/11752790.

Savitzky, Abraham, Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. vol. 36 (8), 1627–1639. https://doi.org/10.1021/ac60214a047.

Schafer, R., 2011. What is a Savitzky-Golay filter? [Lecture Notes]. IEEE Signal Process. Mag. vol. 28 (4), 111–117. https://doi.org/10.1109/MSP.2011.941097.

Singh, A.K., Sharma, L., Mallick, N., Mala, J., 2017. Progress and challenges in producing polyhydroxyalkanoate biopolymers from cyanobacteria. J. Appl. Phycol. vol. 29 (3), 1213–1232. https://doi.org/10.1007/s10811-016-1006-1.

Tan, G.-Y., et al., 2014. Start a research on biopolymer polyhydroxyalkanoate (PHA): a review. Polymers vol. 6 (3), 706–754. https://doi.org/10.3390/polym6030706.

Tao, Z., Peng, L., Zhang, P., Li, Y.Q., Wang, G., 2016. Probing the kinetic anabolism of poly-beta-hydroxybutyrate incupriavidus necator H16 using single-cell Raman spectroscopy. Sensors vol. 16 (8). https://doi.org/10.3390/s16081257.

Teixeira, A.P., Oliveira, R., Alves, P.M., Carrondo, M.J.T., 2009. Advances in on-line monitoring and control of mammalian cell cultures: supporting the PAT initiative. Biotechnol. Adv. vol. 27 (6), 726–732. https://doi.org/10.1016/j.biotechadv.2009.05.003.

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. vol. 58 (2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.

Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., Deng, S.-H., 2019. Hyperparameter optimization for machine learning models based on Bayesian optimization. J. Electron. Sci. Technol. vol. 18 (1), 26–40.

Yang, L., Shami, A., 2020. On hyperparameter optimization of machine learning algorithms: theory and practice. Neurocomputing vol. 415, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061.

Zhou, L., Pan, S., Wang, J., Vasilakos, A.V., 2017. Machine learning on big data: opportunities and challenges. Neurocomputing vol. 237, 350–361. https://doi.org/10.1016/j.neucom.2017.01.026.