Contents lists available at ScienceDirect

Bioresource Technology

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

Addressing raw material variability: *In-line* FTIR sugar composition analysis of lignocellulosic process streams

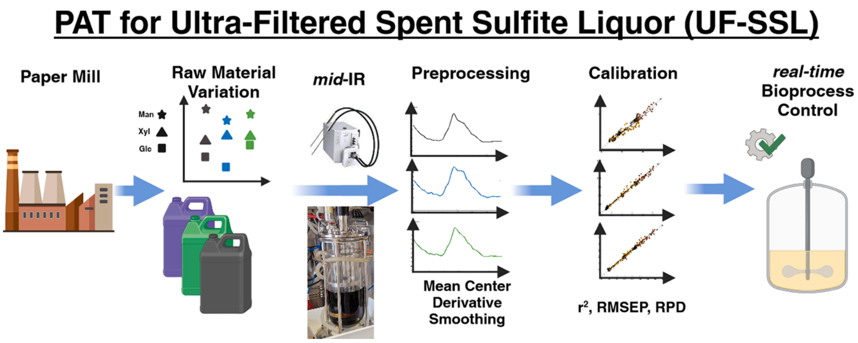
Daniel Waldschitz [a], Yannick Bus [a], Christoph Herwig [a,b], Julian Kager [c], Oliver Spadiut [a,*]

[a] Research Group Bioprocess Technology, Institute of Chemical, Environmental and Bioscience Engineering, TU Wien, Gumpendorferstraße 1A, Vienna A-1060, Austria
[b] Körber Pharma Austria GmbH, Mariahilferstraße 88A, Vienna A-1070, Austria
[c] Department of Chemical and Biochemical Engineering, Technical University of Denmark, Søltofts Plads 229, Kgs. Lyngby 2800, Denmark

HIGHLIGHTS

- Process analytical technology for analysis of raw material variations.
- mid-IR measurement of spent sulfite liquor.
- Simultaneous in-line quantification of multiple sugars.
- Comparable RMSEP reached for UF-SSL as for hydrolysates from agricultural residues.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Raw material variation
Lignocellulosic feedstock
Spent sulfite liquor
PAT
mid-IR

ABSTRACT

For a sustainable economy, biorefineries that use second-generation feedstocks to produce biochemicals and biofuels are essential. However, the exact composition of renewable feedstocks depends on the natural raw materials used and is therefore highly variable. In this contribution, a process analytical technique (PAT) strategy for determining the sugar composition of lignocellulosic process streams in *real-time* to enable better control of bioprocesses is presented. An *in-line mid*-IR probe was used to acquire spectra of ultra-filtered spent sulfite liquor (UF-SSL). Independent partial least squares models were developed for the most abundant sugars in the UF-SSL. Up to 5 sugars were quantified simultaneously to determine the sugar concentration and composition of the UF-SSL. The lowest root mean square errors of the predicted values obtained per analyte were 1.02 g/L arabinose, 1.25 g/L galactose, 0.50 g/L glucose, 1.60 g/L mannose, and 0.85 g/L xylose. Equipped with this novel PAT tool, new bioprocessing strategies can be developed for UF-SSL.

1. Introduction

As the largest renewable biomass feedstock on earth, lignocellulosic biomass has huge potential for a sustainable economy based on biorefineries (Demirbas, 2009). The most prominent fermentation process in lignocellulosic biorefineries is the production of ethanol, but also

[*] Corresponding author.
E-mail address: oliver.spadiut@tuwien.ac.at (O. Spadiut).

https://doi.org/10.1016/j.biortech.2024.130535
Received 24 November 2023; Received in revised form 10 February 2024; Accepted 4 March 2024
Available online 15 March 2024

other biofuels (methane, butanol), biomaterials (polyhydroxyalkanoates) and other value-added products, such as organic acids (itaconic acid, glutamic acid), and lipids are operational or at least under consideration (de Jong et al., 2012). Second generation feedstocks in particular are being prioritised by the authorities for future development (Hassan et al., 2019). The most commonly used are softwood, hardwood and grasses, each with its own structural properties and chemical composition based on its core components cellulose, hemicellulose and lignin (De Bhowmick et al., 2018). Unfortunately, lignocellulosic biorefineries suffer from low profitability, not only due to variable substrate availability and transport costs, but also due to process instabilities resulting from feedstock variability and consequently low yields. In paper production, either the sulfite or kraft process are commonly used. In the sulfite process, wood chips are treated with sulfurous acid at high temperatures and pressures to break down the lignin and hemicellulose matrix of the wood while preserving the cellulose. The remaining liquid fraction is called spent sulfit liquor (SSL) from the sulfite process (or black liquor (BL) from the Kraft process). The industrial pulp-based side streams are a complex matrix of lignin degradation products such as lignosulfonates, C5 sugars (xylose, arabinose), C6 sugars (glucose, mannose, galactose), acids (acetic acid, sulfurous acid, and sulfuric acid) and inhibitory compounds such as furfural and 5-hydroxymethylfurfural (HMF). In addition, the composition of fermentable sugars is not dominated by glucose as in enzymatic hydrolysis, but is also dependent on the raw material used, creating a challenging matrix for analysis. However, SSL (and BL) are available year-round in large quantities as by-products. The biggest sugar fraction for softwood is the pentose xylose, whereas the hexose mannose is dominant for hardwood (Fatehi and Ni, 2011). In further process steps, the water content is reduced and the SSL mostly incinerated. A desirable and stable operation of a bioprocess requires constant raw material quality and composition (Kenney et al., 2013). In the case of raw material variation, it is imperative to track any changes for further use in bioprocessing. The main variations in this feedstock arise from the chemical composition of hemicellulose, which varies between softwood and hardwood species affecting the composition of sugars, and the reduction in water content affecting the overall concentration of sugars. To tackle these process challenges, the importance of Quality by Design (QbD) principles based on Process Analytical Techniques (PAT), which can accurately measure the composition of the complex renewable feedstocks used, was highlighted (Rathore et al., 2016). Various techniques have been used to detect and quantify components in lignocellulosic feedstocks, such as high pressure liquid chromatography (HPLC) or nuclear magnetic resonance (NMR) (Gjersing et al., 2013). However, these techniques are limited to off-line use due to the extensive sample pre-treatment. Vibrational spectroscopic sensors, such as IR and Raman probes, have been proposed for lignocellulosic bioprocesses (Lopez et al., 2019). Several applications have been studied for analysis of raw material pre-treatment (Xu and Wang, 2013) or particle-free supernatant of fermentation samples (Lopez et al., 2021). All of the listed examples are from food crops or grasses using enzymatic hydrolysis or treatment with diluted acids, generating free sugars for fermentation. Thus, they present PAT solutions for seasonal agricultural residues with a low matrix complexity. To the best knowledge of the authors, no *real-time* ready spectroscopic method has been published to quantify key metabolites, such as sugars, for complex lignin-rich SSL (and BL) that are produced as a by-product of the paper industry. Thus, a cheap, readily available source of sugars is lacking suitable PAT solution necessary for developing QbD based bioprocesses. As transportation cost is a major driver of the total cost of lignocellulosic process, this study focused on SSL from a single paper mill within an integrated biorefinery concept (Rødsrud et al., 2012). The developed method should be capable of *in-line* quantification of multiple sugars simultaneously present in SSL for at least one working week (⩾120 h) without significant probe fouling. In this study, the performance of the monitoring system was tested for two process scenarios: a low complexity case, designed to resemble the process deviations, and a high complexity case, intended to explore the full capability of the newly developed PAT tool. Two analytical objectives were defined: to quantify only the three most abundant sugars with low background variation for the low complexity case and to explore the full capability of the PAT tool by quantification of five sugars in high concentration ranges with high background variation for the high complexity case. PLS models were generated for both cases and tested and compared for a SSL stream of varying composition and dilution for extended time periods to ensure long-term accuracy, precision, and stability of prediction.

2. Materials and methods

2.1. Ultra-filtered spent sulfite liquor

Multiple independent batches of low molecular weight permeate of ultra-filtered spent sulfite liquor (UF-SSL) from softwood pulping (stored at 4 °C) from a single biorefinery paper mill were used for all experiments. The UF-SSL batches had a composition of 132–143 g/L mannose, 56–65 g/L xylose, 41–47 g/L glucose, 29–34 g/L galactose, and 13–18 g/L arabinose, as well as several other mono- and disaccharides, such as rhamnose in low quantities (⩽5 g/L), which were not considered. To extend the range of sugar concentrations tested, highly concentrated solutions of each sugar in ultra-pure water (spike) were prepared and added to the dilutions of UF-SSL batches (matrix). The spikes ranged from +1 % to +28 % sugar relative to the amount of sugar in 100 % UF-SSL.

2.2. Spectra acquisition

The FT-MIR spectra were obtained using a Fiber MultiplexIR FT-IR system (ReactIR 45 m, Mettler Tolido, USA) equipped with a liquid N^2 MCT detector and an optical fiber immersion probe from silver halide, with 9.5 mm optical path length and a DiComp diamond probe tip (ReactIR 45 m, Mettler Tolido, USA) which was connected with a 1.5 m long fiber optic cable. Each spectrum ranged from 3000 cm^{-1} to 650 cm^{-1} and consisted of an average of 256 scans with a resolution of 4 cm^{-1}. The probe was inserted into a *lab-scale* stirred tank reactor glass vessel and agitation was performed to ensure homogeneity of the particle-free solution. The optical fiber was further stabilized to ensure that it did not move or bend between measurements (as recommended by the manufacturer). The equipment and setup remained the same during both model building and in-line application, with only the spectral acquisition methods differing. During model building, each spectrum was acquired manually once the media was well mixed, whereas during in-line application a spectrum was acquired every 3 min. The instrument and data collection were controlled by iC IR 7.0 software (Mettler Toledo, USA).

2.3. Off-line sample analysis

Glucose, Xylose, Arabinose, Galactose, and Mannose were measured using HPLC (Ultimate 3000, Thermo Fisher Scientific, USA) equipped with an RI detector (RI100, Shodex, USA) on a Pb-column (NUCLEOGEL SUGAR Pb 719530, Machery-Nagel, Germany) at 79 °C with an isocratic flow of 0.4 ml/min ultra-pure water with a runtime of 65 min. All samples were diluted 1:20 with ultra-pure water and filtered using a 0.22 μm filter before analysis.

2.4. Calibration and validation

Independent calibration and validation sets were prepared for the low and high complexity cases. Table A.1 provides a summary of the analytes, dilutions, and number of batches used for each set. The samples were prepared in a glass vessel with a working volume of 1 L under constant stirring. The vessel was filled with 500 mL of UF-SSL (diluted

Table A.1
Sample preparation for both case studies. Step sizes were set per sugar to increase sugar concentrations from +1 % to +28 % sugar relative to amount of that sugar in 100 % UF-SSL.

Set	Analytes	Dilutions	Batches	Samples
Low Complexity Calibration	C5: Xyl C6: Man, Glc	25%	1	12 steps per sugar
Low Complexity Validation	C5: Xyl C6: Man, Glc	25%	1	2 sugars per step 3x 12 steps
High Complexity Calibration	C5: Xyl, Ara C6: Man, Glc, Gal	25,50,75%	2	6 steps per sugar + dilution + batch
High Complexity Validation	C5: Xyl, Ara C6: Man, Glc, Gal	25,50,75%	50/50mix	2 sugars per step 3x 12 steps per dilution

with water) to serve as the first calibration sample. Spikes of glucose, xylose, arabinose, galactose, and mannose were added sequentially to generate a new calibration sample. When multiple dilutions (25/50/75 %) were used, spiking from +1 % to +28 % sugar, relative to the amount of each sugar in 100 % UF-SSL, ensured overlap between dilutions. After spiking UF-SSL with one sugar, the vessel was emptied and fresh UF-SSL was prepared and spiked with the next sugar. When multiple batches of UF-SSL were used, spikes were performed alternately between both batches. Validation samples were prepared similarly. To increase variability and minimize correlation between the sugars, the different enriched UF-SSL solutions were added together in random order. Each addition of the enriched UF-SSL solution generated a new sample, creating a set of validation samples. In cases where multiple batches of UF-SSL were used, a 50/50 mixture of both was utilized for the validation samples. Multiple independent sets were measured.

2.5. In-line application

During the *in-line* application of the models, a 3.5 L glass vessel was utilized to represent a feed tank. Diluted UF-SSL, with or without added sugars, was regularly added or removed from the vessel to mimic use and refilling of the tank. Almost every addition of new UF-SSL resulted in a different sugar composition by spiking with selected sugars. For the low complexity case, only step-wise fill/drain cycles were tested, while for the high complexity case, gradients were also included. In addition, multiple UF-SSL dilutions and batches were used to prepare the spike solutions to vary the background in addition to the analyte concentration in the high complexity case. An automated sampling system (Numera, Securecell, Switzerland) was used to take samples from inside the feed tank during the *in-line* application (2 ml, stored at 4 °C until analysis). Samples were taken every 1.5 h before and after step-wise addition as well as while gradual addition was performed, and every 3 h in phases where no changes were applied in between.

2.6. Data analysis

Spectra and reference data were imported into MATLAB R2021a (The MathWorks Inc., Natick, USA) for multivariate data analysis using PLS_Toolbox 8.9.2 (Eigenvector Research Inc., Manson, USA). Appropriate spectral regions were selected for analysis, taking into account the corresponding active regions for carbohydrates and instrument limitations (e.g. the diamond region). The spectra were pre-processed by mean centering, smoothing (with a Savitzky-Golay-Filter) and the application of derivative methods (also using the Savitzky-Golay method). The Model Optimizer tool of PLS_Toolbox was used to screen for optimal preprocessing and model settings [tested settings: derivative order 1–5, polynomial order 1–5, window size 1–100, latent variables 1–20] and minimized the RMSEP of the calibration samples. Optimal settings were calculated per analyte and each spectrum was analyzed for all analytes. Spectra were matched to reference data, using the spectrum acquired closest to the time of sampling for in-line application. Independent Partial Least Square (PLS) regression models were developed for each analyte based on FT-MIR spectra. Principal Component Analysis (PCA) was used to identify outliers and remove anomalous samples from the data sets. Detailed information on the properties of the PLS models and the preprocessing methods used can be found in the Supplementary material. The normalised root mean square error of prediction (NRMSEP) was calculated as RMSEP divided by the range of the calibration set (max sugar concentration – min sugar concentration) as measured by the reference analytic and further validation metrics calculated as described by Lotfollahi et al. (2023).

3. Results and discussion

Before evaluating the results, the acceptance criteria of the proposed tool for use in biotechnological processes were defined. To ensure selectivity between the analytes of interest, the PLS models were evaluated using coefficients of determination of predicted values r^2 of $\geqslant$0.95. Since the sugar concentration ranges within the UF-SSL were significant, the relative errors were calculated by normalizing to the size of the range analyzed (NRMSEP = RMSEP/ (max–min)). A target of $\leqslant$10 % NMRSEP as a comparable measure of variance between sugars was used and the ratio of performance to deviation (RPD) was used to validate that sufficiently large ranges were tested. In addition, the precision, accuracy and stability of the sensor performance needed to be evaluated to assess the ability of the proposed tool to be used as an *in-line* sensor in *real-time* bioprocesing (Rajamanickam et al., 2021). An error distribution over time within the interquartile range to check whether the error can be approximated as symmetrically distributed around the median (precision) and an interquartile spread of the relative error over time of $\leqslant$20 % (accuracy). Furthermore, the 25 & 75 % percentile boundaries of the relative error over time of $\leqslant$15 % (precision + accuracy) as a limit on the output variance from the measurement, as well as no drift or deterioration of the prediction in time frames of at least one working week of operation $\geqslant$120 h to validate its ability to be used *in-line* over extended periods of time (stability).

3.1. Choice of analytical technique

In order to develop a suitable PAT strategy for the assessment of feedstock variability within the complex lignin-rich matrix of UF-SSL, an appropriate *real-time* technique must first be selected. IR and Raman spectroscopy are the most widely used spectroscopic techniques in biotechnology for the detection of sugars, hence both were tested for use in this study. Raman spectroscopy is well known for the analysis of complex mixtures of biological compounds due to its wide spectral range, narrow peak widths, and high sensitivity and selectivity as water is invisible to it. However, the Raman signal can be overshadowed by fluorescence from matrix components excited by the laser light source. This is unfortunately the case for lignocellulosic applications where the highly conjugated aromatic groups of lignin are a major source of fluorescence and have been reported to be problematic for samples subjected to enzymatic hydrolysis (Ewanick et al., 2013). Since the primary goal of the pulping process is to remove the hemicelluloses and lignin from the cellulose, the expected lignin content is much higher in the UF-SSL. When the dark brown to black UF-SSL was analyzed by Raman, the fluorescence signal, caused by the aromatic rings of the lignin degradation products, led to saturation of the detector (a Raman spectrum of UF-SSL is provided in the Supplementary material). A reduction of the lignin content by extraction may not be desirable due to the additional cost and complexity of the process, which may make it technically or economically not feasible. Therefore, the use of time-gated Raman spectroscopy (Knorr et al., 2010), a Raman technique designed for highly fluorescent samples, would be necessary. However, because time-gated Raman systems are much more expensive than comparable IR systems, IR spectroscopy may be a more cost-effective solution. For IR spectroscopy, both *near*-IR (~ 800–2500 nm) and *mid*-IR (~ 2500–25000 nm) are commonly used as PAT tools for bioprocessing. The advantage of *mid*-IR is that it detects fundamental vibrations, which are much stronger than overtones, providing greater accuracy when analyzing a large number of components (Lantz et al., 2010).

To first assess the feasibility of *mid*-IR for this purpose, 25 % UF-SSL was spiked with increasing amounts of each of the three most abundant sugars in UF-SSL (mannose, xylose, and glucose) separately. Fig. A.1 shows the measured *mid*-IR spectra of the carbohydrate region (~ 950–1150 cm^{-1}) where the most prominent bands of the sugars were expected. The 2 most distinct *mid*-IR spectra of the 25 % UF-SSL batches used for this study showed a high degree of similarity, indicating that matrix differences between the batches were minimal when measuring the carbohydrate region with *mid*-IR. In addition, the addition of sugar to the 25 % UF-SSL resulted in peak sizes in the carbohydrate region that increased with the amount of sugar added, as well as different peak shapes depending on the specific sugar added. Therefore, the signal-to-background ratio was considered sufficient and *mid*-IR was selected as the technique of choice for this study.

3.2. Low complexity case study

In the low complexity case 1 C5 sugar (xyl) and 2 C6 sugars (man, glc) were analysed within 1 batch of UF-SSL at a constant dilution of 25 %.

3.2.1. Model generation

For each sugar, an independent PLS model was generated using all the calibration samples (including the spectra where a different sugar was spiked than the model was to predict) using the built-in pre-processing optimizer of the PLS toolbox. All sugars were predicted with high accuracy and precision within the matrix of UF-SSL (Fig. A.2). All 3 models have 3 latent variables (optimal for minimal RMSEP determined using PLS_Toolbox) with coefficients of determination r^2 of ⩾0.95 and RMSEP between 0.5 and 1.6 g/L, respectively, suitable for parallel quantification of sugars. The RPD metric is used to determine whether the measurement error is low enough to quantify the analyte within the concentration range being analyzed, with values above 2.5 considered excellent (Lotfollahi et al., 2023). RPD values of 6.41–8.32 were obtained, indicating that a viable calibration was performed for all models. Mannose had the highest RMSEP of all 3 sugars, but also the largest range. When the relative errors were calculated as RMSEP per range analyzed (=normalized RMSEP), comparable 4.5 ± 0.5 %NRMSEP were obtained, indicating no significant differences in error per range between all sugars. Therefore, all sugars had similar quantification accuracy per range. Furthermore, by including all calibration samples in the calibration of each individual sugar, a high selectivity was achieved even between very similar stereoisomers, glucose and mannose, as indicated by the small scatter between samples of the base sugar concentration (samples spiked with a different sugar than predicted).

3.2.2. In-line application

To test the performance of the sensor during prolonged *in-line* use, a different batch of UF-SSL compared to the calibration samples was used. This was done to explore the possibility of adapting the model for use cases outside its original designed scope (1 batch of UF-SSL). For the use of the model outside its original scope, 1/3 of the reference samples obtained during the *in-line* application (from another batch of UF-SSL) were included in the calibration, to update and adapt the model to the applied conditions.

Fig. A.3 shows the model prediction over time. The predicted value is shown surrounded by a shade corresponding to the RMSEP of the model, with the original unadapted model prediction in a light shade and the adapted model in a darker shade. In order to further test the ability of the models to discriminate between the different sugars, one analyte was held constant along with the background, while the other two were changed in a stepwise fashion at different ratios. For the constant analyte (xylose), the different background resulting from the use of a different batch of UF-SSL resulted in an offset in the unadapted model. This shifted the relative error toward overprediction. However, by including 1/3 of the reference samples in the model calibration, the offset could be corrected for the entire process, resulting in a relative

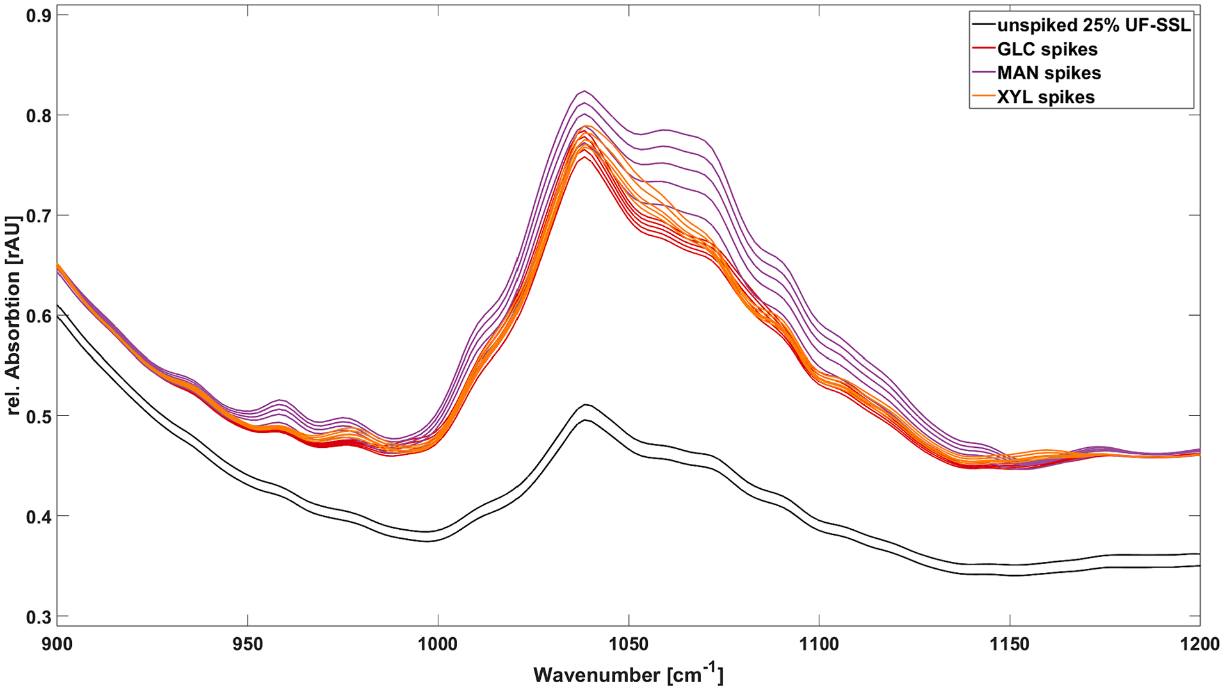


Fig. A.1. *mid*-IR raw spectra of the carbohydrate region. The 2 most distinct UF-SSL batches in terms of *mid*-IR spectra (black lines) at 25 % dilution are shown to provide an overview of the maximum matrix influence covered in this study. Spikes with increasing concentrations of individual sugars (colored lines) result in different peak shapes.

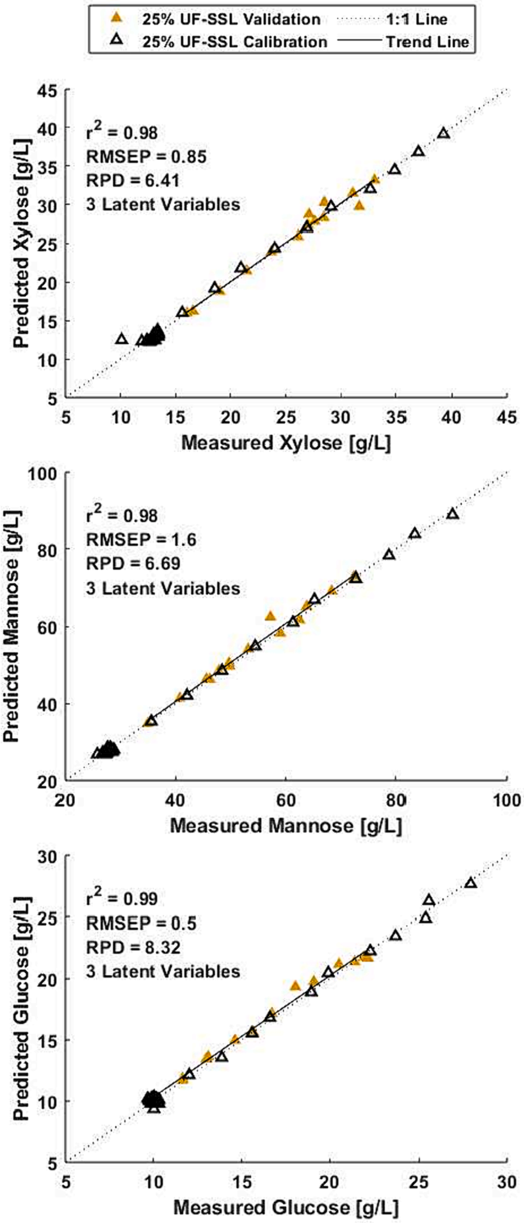


Fig. A.2. Low complexity case model calibration using 12 calibration concentrations per sugar (black outlines) as well as a test set with randomized spikes of all 3 sugars (filled shapes) with measured sugar concentrations on the X-axis. For generation of the 3 independent PLS models for xylose (top), mannose (middle) and glucose (bottom), all calibration samples (both spiked with the respective or other sugars, 36 samples total) were used, each with predictions on the Y-axis. The 1:1 line indicates where the prediction would perfectly match the reference measurements and the trend line represents the linear regression obtained from the calibration samples.

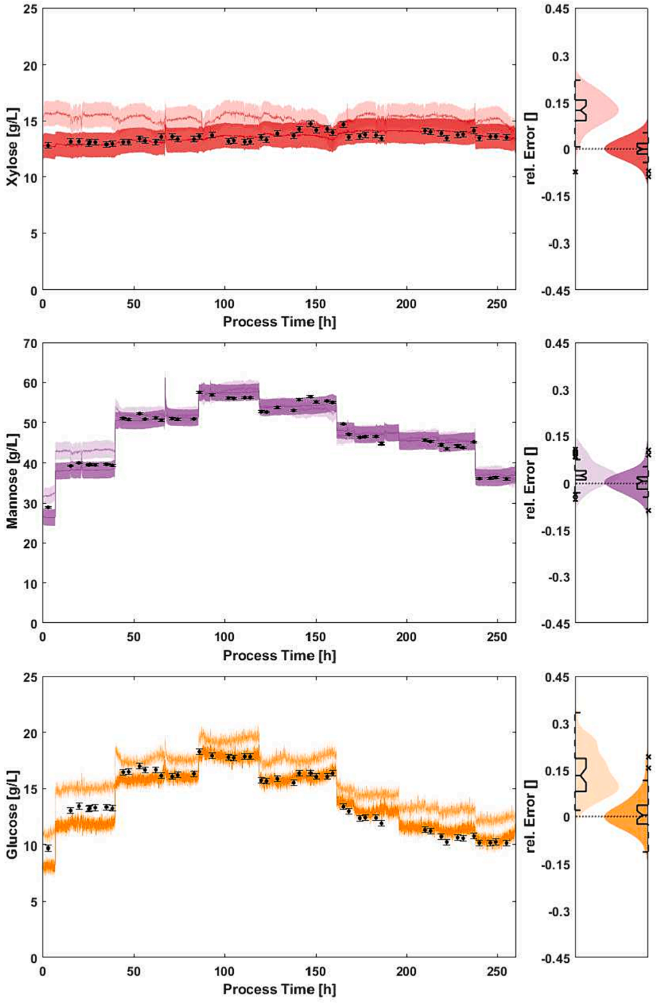


Fig. A.3. *In-line* application of the developed low complexity models over ~ 260 h. Throughout the experiment, 25% UF-SSL was used as the background. The original PLS model prediction (based on the calibration samples only – solid line with a lighter shade representing the model RMSEP) compared to the adapted model prediction (calibration +1/3 reference samples – solid line with a darker shade representing the model RMSEP) compared to the HPLC reference values (black dots) for xylose (top), mannose (middle) and glucose (bottom) including their relative error distribution for the whole experiment.

error with a median that matched the reference. For the most abundant analyte (mannose), most steps were already predicted within the error of the measurements without model adaptation. In addition, steps that were previously overestimated could be corrected with model adaptation, resulting in a median relative error that matched the reference. As glucose concentrations were significantly lower than those of mannose, the different background due to the use of a different UF-SSL batch caused a greater offset across all steps. Model adaptation was still able to correct the offset, resulting in a median relative error that matched the reference. Furthermore, no significant drift or degradation in prediction was observed for any analyte for ⩾250 h (20–25 °C, probe tip always fully submerged).

3.2.3. Summary: low complexity case study

Table A.2 contains a summary of all key performance indicators of the prediction. For the low complexity case study, 3 analytes were able to be quantified within a constant matrix with 3 latent variables meeting the desired acceptance criteria (coefficients of determination r^2 of ⩾0.95 and root mean square error of prediction (RMSEP) of ⩽10 % of the analyzed range). For comparison with similar results published in the literature: Xu and Wang (2013) analyzed glucose, xylose and arabinose in corn stover hydrolisate, an agricultural residue from diluted acid pretreatment, with *near*-IR spectroscopy. Although not a perfect comparison due to the different substrate, concentration ranges, and detector used, it provides the closest comparison in terms of the number of sugars analyzed in liquid lignocellulosic residues. They achieved r^2 between 0.77 and 0.92 (this study ⩾0.95), RMSEP of 0.39–1.28 g/L (this study 0.50–1.60 g/L) and RDP values of 2.03 to 3.53 (this study 6.41–6.69) in concentration ranges of 0.41–14 g/L (this study ≃ 10–90 g/L), with other authors reporting similar or worse values for sugars in lignocellulosic residues. Hence, both the quantity of simultaneously

Table A.2
Summary of the key performance indicators (coefficient of determination r^2, RMSEP, NRMSEP, Median, 25&75% percentile boundaries, interquartile Spread) of the original model for model generation and adapted model for *in-line* application. 3 analytes were measured with 36 calibration samples each.

Analyte	Model generation				*in-line* application				
	r^2	RMSEP	NRMSEP	RPD	Median	25th%	75th%	Spread	RMSEP
	[-]	[g/L]	[%]	[-]	[%]	[%]	[%]	[%]	[g/L]
Man	0.98	1.60	4.3	6.69	0.6	−2.0	2.0	4.0	1.47
Xyl	0.98	0.85	5.0	6.41	−0.1	−1.7	1.8	3.5	0.39
Glc	0.99	0.50	4.2	8.32	0.4	−2.5	3.6	6.1	0.72

measured sugars and the quality of prediction could be matched while measuring in an even more challenging matrix. *mid*-IR detects fundamental vibrations are observed whereas *near*-IR detects overtones an combination bands. However, fundamental vibrations are known to be much stronger than overtones (Lantz et al., 2010) which may contribute in the compensation of the more challenging matix of UF-SSL. After model adaptation for the *in-line* application, to adapt the model to the different batch of UF-SSL used, no significant difference was found in the RMSEP values for all three models between calibration and application. All medians of the prediction over time matched the reference (⩽1%), ensuring an evenly distributed error. Furthermore, an interquatrile spread of ±2.5% was achieved, well within the outlined acceptance criteria.

3.3. High complexity case study

For the high complexity case, within 2 batches of UF-SSL at 25/50/75% dilutions, 2 C5 sugars (xyl, ara) and 2 C6 sugars (man, glc, gal) were analyzed.

3.3.1. Model generation

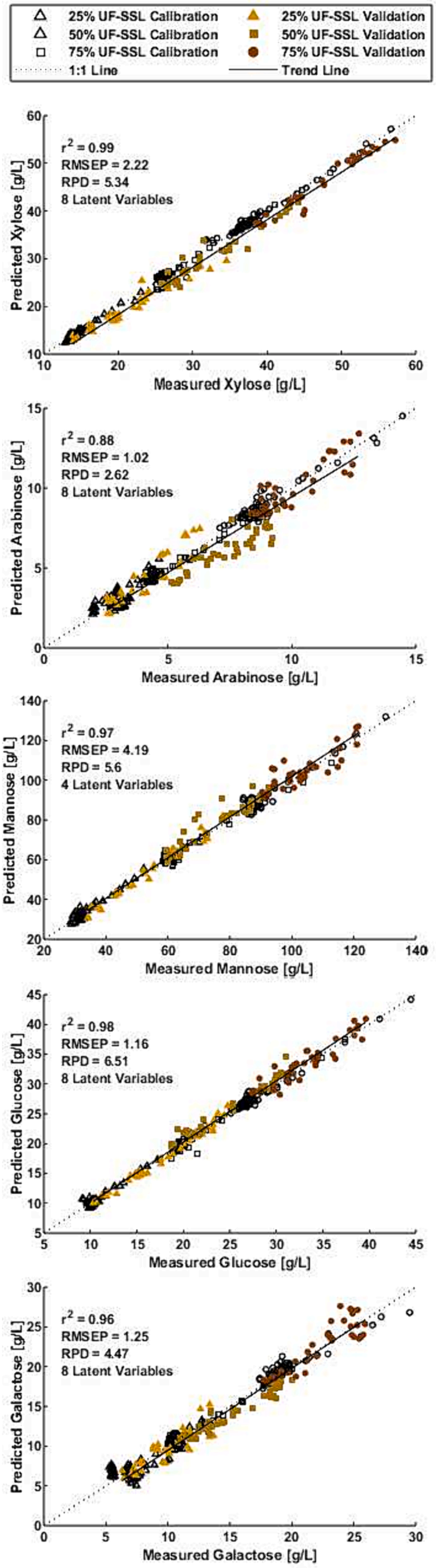
Fig. A.4 shows the results of the calibration for the three C6 sugars and two C5 sugars investigated. Here, an increased spread between measurements from higher UF-SSL% was a common trend for all sugars analyzed. The coefficients of determination r^2 remained comparable to the low complexity case of ⩾0.95, except for the scarcest analyte arabinose with 0.88. However, for arabinose, the trend of the measurements of the 50% dilution had a significant offset to the other dilutions, affecting the r^2. Interestingly, this shift occurred in the middle of the 3 dilutions, indicating that this effect was not solely due to increasing background. Fortunately, for all other sugars (with higher concentrations), no significant offset between dilutions was observed, demonstrating that the PLS model can accurately predict sugar concentrations from *mid*-IR measurements even within a variable matrix background.

In contrast to the low complexity case where all 3 models had 3 latent variables each, 4 of the models have 8 latent variables (optimal for minimal RMSEP determined using PLS_Toolbox), while only for the mannose model the lowest RMSEP was identified with 4 latent variables. The increase in complexity due to different dilutions, different batches of UF-SSL and more analytes may explain the increase in latent variables. Fewer latent variables may be needed for mannose than for any other analyte because its maximum concentration was more than twice that of the next most abundant analyte and about 10 times that of the least abundant analyte. With the inclusion of more analytes in total, higher sugar concentration ranges, multiple dilutions and batches, higher RMSEP values were expected. While the RMSEP for C5 xylose and C6 mannose approximately doubled compared to the low complexity case, the RMSEP for C6 glucose remained the same. This was a surprising result since glucose is neither the most abundant analyte nor the most abundant C6 sugar. However, glucose is the most important sugar for biotechnological applications, so the ability to correctly predict its concentration is essential for the application of the analytical technique in bioprocessing. RPD values ranging from 2.62 to 6.51 were obtained, with values greater than 2.5 indicating that the selected range of quantification was adequately large for the resulting errors. In order to better compare the results between the sugars, the RMSEP was calculated per analyzed range. Here, similar values of 5 ± 1.4 %NRMSEP were observed compared to the low complexity models, with the exception of arabinose at 10% NRMSEP. Arabinose was the least abundant analyte tested. This indicated that the influence of analyte concentration levels on the absorption spectra was already reduced compared to other analytes, to the point where background variations had a significantly higher impact on the prediction. However, a relative error of ≃ 10 % still met the outlined acceptance criteria for quantification.

3.3.2. In-line application

The *in-line* application of the low complexity model showed that there was no deterioration of the prediction (e.g. due to probe fouling) which affected the prediction of abundant analytes. The high complexity model included analytes present at up to 10-fold lower concentrations. Due to the lower concentration relative to background, these analytes are more susceptible to the effects of fouling on prediction over time. However, no deterioration in prediction of analytes with up to 10-fold lower concentrations (ara, gal) was observed over ≃ 240 h (Fig. A.5). Prediction accuracy varied for all sugars over time. Periods of higher and lower prediction accuracy occurred for all analytes regardless of analyte concentration. In addition, periods of decreased prediction accuracy for one analyte did not consistently coincide with decreased prediction accuracy for other sugars. For example, the prediction of the C6 sugar mannose was more accurate between 40 and 100 h than between 100 and 120 h. In contrast, the second most abundant C6 sugar, glucose, had high prediction accuracy before 60 h, lower accuracy between 60 and 100 h, and again high accuracy between 100 and 150 h. The accuracy trend for the third C6 sugar, galactose, did not follow the accuracy trend of either of the other two C6 sugars. Similarly, the accuracy trends of the two C5 sugars differed from each other as well as from all three C6 sugars. Although some background changes coincide with improved or worsened prediction accuracy of analytes (e.g., glacatose during the shift at ≃ 175 h), other analytes remain unaffected (e.g., arabinose during the same shift). Furthermore, offsets of similar magnitude were observed during concentration steps as well as background shifts for all analytes. Therefore, background shifts were not observed to be more problematic for model prediction than concentration steps. In the majority of the time frames with reduced prediction accuracy, the prediction still reflects the general trend of the *off-line* samples, albeit with an offset that did not occur in the time frames with higher accuracy. An example of this would be the prediction of glucose with an offset between 60 and 105 h, which was not present before or after. However, the prediction still followed the concentration change trend measured in the offline samples. Only in rare cases in the data set, e.g. the prediction of xylose between 110 and 175 h, did jumps in the prediction not coincide with jumps in the reference samples.

Looking at the relative error distribution, all models were able to predict the sugar concentration with the reference within the interquartile range. Additionally, two models (xylose and glucose) matched the reference with their median prediction, resulting in evenly distributed error around the reference. The relative errors of the three most common analytes met the acceptance criteria of 25&75% percentile limits over time of ⩽15 %. In addition, although the high complexity



(caption on next column)

Fig. A.4. High complexity case model calibration using 3 different dilutions from 2 batches each with spiked concentrations per sugar (black outlines) as well as a test set with randomized spikes of each 2 randomized sugars simultaneously covering all 5 sugars (filled shapes) with measured sugar concentrations on the X-axis (105 samples total). To generate the 5 independent PLS models for xylose (top), arabinose (middle top), mannose (middle), glucose (middle bottom), and galactose (bottom), all calibration samples (both where the respective or other sugars were spiked) were used, each with predictions on the Y-axis. The 1:1 line indicates where the prediction would perfectly match the reference measurements and the trend line represents the linear regression obtained from the calibration samples.

model of arabinose did not meet the acceptance criteria of $r^2 \geqslant 0.95$, its 25&75% percentile boundaries and interquartile range were within the target range. However, for galactose, the C6 sugar with the lowest concentrations, one side of the interquartile range was not within the target range, and its total interquartile spread was ⩾20%.

3.3.3. *Summary: high complexity case study*

Table A.3 contains a summary of all key performance indicators of the prediction. For the high complexity case, 4 out of 5 models met all acceptance criteria. The model for the scarcest C5 sugar, arabinose, had a coefficient of determination below the target value (0.88 instead of ⩾0.95), and barley met the second acceptance criterion with an NRMSEP of precisely 10 %. The generated models had higher RMSEP values compared to the low complexity method (see Table A.2). However, the concentration range over which quantification was attempted was greatly increased. When comparing the NRMSEP between the methods, similar values were obtained. Quantification of sugars at lower total concentrations resulted in higher NRMSEP values (RMSEP per range). Therefore, the overall prediction quality did not decrease significantly when a larger background range was analyzed. Compared to the low complexity model, the simultaneous quantification of more sugars at higher concentrations and background variations was successfully implemented, which already stacked up favorably compared to published values in literature for lignocellulosic residues. This further highlights the ability of *mid*-IR to differentiate between very similar analyses (multiple C5 and multiple C6 sugars) while being significantly less effected by changes in the sample matrix components such as lignosulfonates and organic acids (different UF-SSL batches in multiple dilutions). Furthermore, no significant differences in RMSEP values were found for all models between calibration and application. All models met the acceptance criteria that the relative error over time around the reference concentration was distributed within the predicted interquartile range. Two models even matched the reference with the median of the distribution over time. The low complexity models achieved both only when the reference measurement was included in the calibration. 4 out of 5 models had 25&75% percentile limits within the acceptance limits of ⩽15 % and a total interquartile range of ⩽20%. Only for the model of the scarcest C6 sugar did the interquartile range not meet these criteria.

4. Conclusion

An *in-line* applicable spectroscopic method for the simultaneous quantification of multiple sugars within the complex matrix of a process side stream from the pulping process was developed. Previously, similar spectroscopic methods have only been demonstrated for agricultural lignocellulosic residues subjected to gentler enzymatic or acid pretreatment. The results demonstrate that such spectroscopic quantification of sugars can be extended to more complex matrices, such as UF-SSL, with comparable RMSEP values. These results lay the groundwork for addressing raw material variations within pulp-based second generation feedstocks.

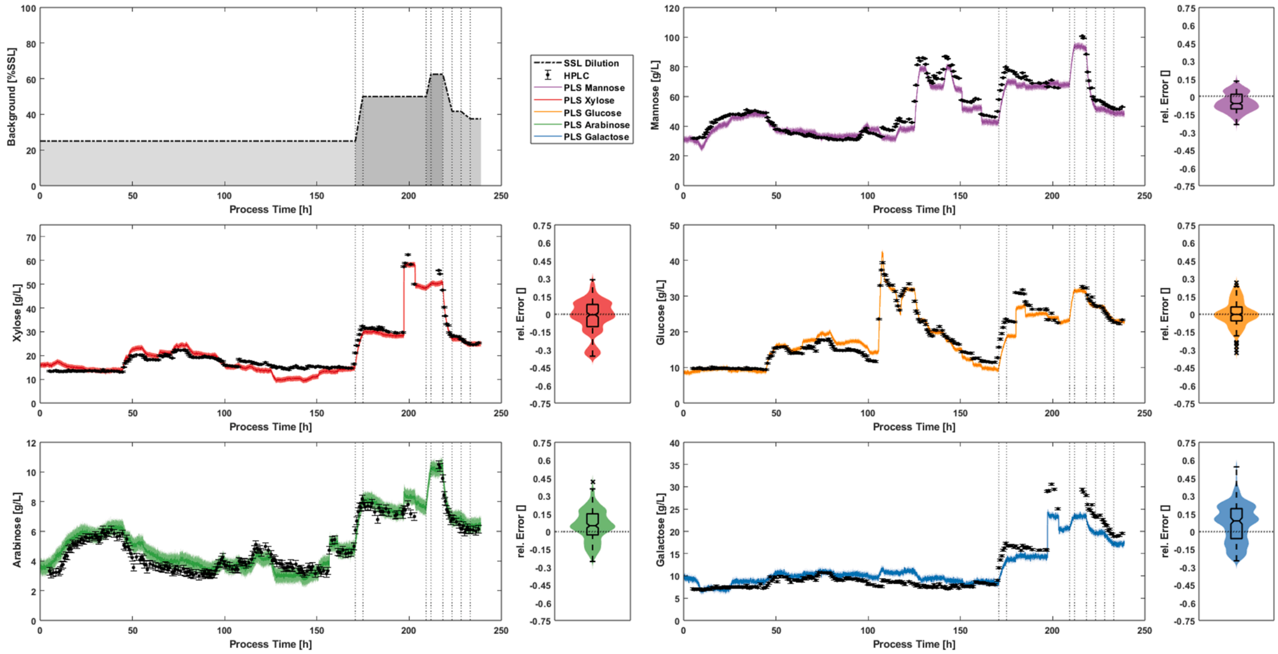


Fig. A.5. *In-line* application of the developed high complexity models over ~ 240 h. Different UF-SSL dilutions were tested as background variation, both in steps and gradients (top left). The PLS model predicted value (solid line with a shade representing the model RMSEP) compared to the HPLC reference values (black dots) for xylose (top middle), arabinose (top right), mannose (bottom left), glucose (bottom middle) and galactose (bottom right) including their relative error distribution for the whole experiment.

Table A.3
Summary of the key performance indicators (coefficient of determination r^2, RMSEP, NRMSEP, Median, 25&75% percentile boundaries, interquartile Spread) of the high complexity model. 5 analytes were measured with 105 calibration samples each.

Analyte	Model generation				*in-line* application				
	r^2 [-]	RMSEP [g/L]	NRMSEP [%]	RPD [-]	Median [%]	25^{th}% [%]	75^{th}% [%]	Spread [%]	RMSEP [g/L]
Man	0.97	4.19	4.8	4.19	−5.9	−10.4	2.0	12.4	4.90
Xyl	0.99	2.22	5.1	5.34	−0.7	−10.7	8.0	18.7	2.46
Glc	0.98	1.16	3.9	6.51	−0.5	−5.8	5.8	11.6	1.87
Gal	0.96	1.25	6.4	4.47	8.5	−6.1	19.1	25.2	2.08
Ara	0.88	1.02	10.0	2.62	4.7	−3.0	14.8	17.8	0.56

Data availability

All raw data used for this study is available at Mendeley Data with (DOI: 10.17632/62zrbmgh2v.1) an open-source online data repository.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Uploaded to Mendely Data, conformation pending, reserver DOI: 10.17632/62zrbmgh2v.1

Acknowledgments

This project has received funding from the Bio Based Industries Joint Undertaking under the European Union's Horizon 2020 research and innovation program under grant agreement No 790507 (iFermenter).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.biortech.2024.130535.

References

Cabaneros Lopez, P., Udugama, I.A., Thomsen, S.T., Roslander, C., Junicke, H., Iglesias, M.M., Gernaey, K.V., 2021. Transforming data to information: a parallel hybrid model for real-time state estimation in lignocellulosic ethanol fermentation. Biotechnol. Bioeng. 118, 579–591.

De Bhowmick, G., Sarmah, A.K., Sen, R., 2018. Lignocellulosic biorefinery as a model for sustainable development of biofuels and value added products. Bioresour. Technol. 247, 1144–1154.

Demirbas, A., 2009. Biofuels: Securing the Planet's Future Energy Needs. Springer, London.

Ewanick, S.M., Thompson, W.J., Marquardt, B.J., Bura, R., 2013. Real-time understanding of lignocellulosic bioethanol fermentation by raman spectroscopy. Biotechnol. Biofuels. 6, 1–8.

Fatehi, P., Ni, Y., 2011. Integrated forest biorefinery- sulfite process. In: Sustainable Production of Fuels, Chemicals, and Fibers from Forest Biomass. ACS Publications. pp. 409–441.

Gjersing, E., Happs, R.M., Sykes, R.W., Doeppke, C., Davis, M.F., 2013. Rapid determination of sugar content in biomass hydrolysates using nuclear magnetic resonance spectroscopy. Biotechnol. Bioeng. 110, 721–728.

Hassan, S.S., Williams, G.A., Jaiswal, A.K., 2019. Moving towards the second generation of lignocellulosic biorefineries in the eu: Drivers, challenges, and opportunities. Renew. Sust. Energ. Rev. 101, 590–599.

de Jong, E., Higson, A., Walsh, P., Wellisch, M., et al., 2012. Bio-based chemicals value added products from biorefineries. IEA Bioenergy Task42 Biorefinery 34, 1–33.

Kenney, K.L., Smith, W.A., Gresham, G.L., Westover, T.L., 2013. Understanding biomass feedstock variability. Biofuels 4, 111–127.

Knorr, F., Smith, Z.J., Wachsmann-Hogiu, S., 2010. Development of a time-gated system for raman spectroscopy of biological samples. Opt. Express 18, 20049–20058.

Lantz, A.E., Gernaey, K.V., Franzén, C.J., Olsson, L., 2010. Online monitoring of fermentation processes in lignocelluloses-to-bioalcohol production. Biofuel Res. J. Elsevier 315–339.

Lopez, P.C., Feldman, H., Mauricio-Iglesias, M., Junicke, H., Huusom, J.K., Gernaey, K. V., 2019. Benchmarking real-time monitoring strategies for ethanol production from lignocellulosic biomass. Biomass Bioenergy 127, 105296.

Lotfollahi, L., Delavar, M.A., Biswas, A., Fatehi, S., Scholten, T., 2023. Spectral prediction of soil salinity and alkalinity indicators using visible, near-, and mid-infrared spectroscopy. J. Environ. Manage. 345, 118854.

Rajamanickam, V., Babel, H., Montano-Herrera, L., Ehsani, A., Stiefel, F., Haider, S., Presser, B., Knapp, B., 2021. About model validation in bioprocessing. Processes 9, 961.

Rathore, A.S., Chopda, V.R., Gomes, J., 2016. Knowledge management in a waste based biorefinery in the qbd paradigm. Bioresour. Technol. 215, 63–75.

Rødsrud, G., Lersch, M., Sjöde, A., 2012. History and future of world's most advanced biorefinery in operation. Biomass Bioenergy 46, 46–59.

Xu, F., Wang, D., 2013. Rapid determination of sugar content in corn stover hydrolysates using near infrared spectroscopy. Bioresour. Technol. 147, 293–298.