Prediction of Cell Wall Properties and Response to Deconstruction Using Alkaline Pretreatment in Diverse Maize Genotypes Using Py-MBMS and NIR

Muyang Li [1,2] · Daniel L. Williams [2,3] · Marlies Heckwolf [4,5] · Natalia de Leon [4,5] · Shawn Kaeppler [4,5] · Robert W. Sykes [6] · David Hodge [2,3,7,8]

Published online: 4 October 2016

Abstract In this work, we explore the ability of several characterization approaches for phenotyping to extract information about plant cell wall properties in diverse maize genotypes with the goal of identifying approaches that could be used to predict the plant's response to deconstruction in a biomass-to-biofuel process. Specifically, a maize diversity panel was subjected to two high-throughput biomass characterization approaches, pyrolysis molecular beam mass spectrometry (py-MBMS) and near-infrared (NIR) spectroscopy, and chemometric models to predict a number of plant cell wall properties as well as enzymatic hydrolysis yields of glucose following either no pretreatment or with mild alkaline pretreatment. These were compared to multiple linear regression (MLR) models developed from quantified properties. We were able to demonstrate that direct correlations to specific mass spectrometry ions from pyrolysis as well as characteristic regions of the second derivative of the NIR spectrum regions were comparable in their predictive capability to partial least squares (PLS) models for *p*-coumarate content, while the direct correlation to the spectral data was superior to the PLS for Klason lignin content and guaiacyl monomer release by thioacidolysis as assessed by cross-validation. The PLS models for prediction of hydrolysis yields using either py-MBMS or NIR spectra were superior to MLR models based on quantified properties for unpretreated biomass. However, the PLS models using the two high-throughput characterization approaches could not predict hydrolysis following alkaline pretreatment while MLR models based on quantified properties could. This is likely a consequence of quantified properties including some assessments of pretreated biomass, while the py-MBMS and NIR only utilized untreated biomass.

Keywords Analytical pyrolysis · Plant cell wall property prediction · Py-MBMS · NIR · Chemometrics · PLS regression

Introduction

Lignocellulose biorefining technologies, whereby non-food plant biomass is utilized to produce biofuels, have the potential to yield liquid fuels that are sustainable, contribute to rural agricultural economies, and have a significantly lower CO_2 footprint than fossil-derived fuels [1]. Biorefineries employing a conversion pathway that involves a chemical pretreatment and enzymatic hydrolysis, followed by fermentation of the cellulose- and hemicellulose-derived sugars to

Electronic supplementary material The online version of this article (doi:10.1007/s12155-016-9798-z) contains supplementary material, which is available to authorized users.

✉ David Hodge
hodgeda@msu.edu

[1] Department of Plant Biology, Michigan State University, East Lansing, MI, USA

[2] DOE Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, MI, USA

[3] Department of Chemical Engineering and Materials Science, Michigan State University, East Lansing, MI, USA

[4] Department of Agronomy, University of Wisconsin, Madison, WI, USA

[5] DOE Great Lakes Bioenergy Research Center, University of Wisconsin, Madison, WI, USA

[6] National Renewable Energy Laboratory, Golden, CO, USA

[7] Department of Biosystems and Agricultural Engineering, Michigan State University, East Lansing, MI, USA

[8] Division of Sustainable Process Engineering, Luleå University of Technology, Luleå, Sweden

ethanol, have begun to be deployed commercially in the USA, Europe, and Brazil [2]. These processes utilize mostly graminaceous agricultural residues such as corn stover, wheat straw, and sugarcane bagasse with future plans to potentially employ perennial grasses such as switchgrass and giant reed (*Arundo donax*) among others. The properties of the biomass feedstock are an obvious and well-recognized contributor to both the economics and technological feasibility of lignocellulosic biorefining technologies. As such, in parallel with improvements in processing technologies, developing biomass cultivars with attractive agronomic traits (e.g., yield, growth, drought tolerance) and/or processing traits (e.g., "recalcitrance" or response to a conversion process, generation of fermentation inhibitors, sugar yields) is an important target for the developing cellulosic biofuel industry. As a consequence, there is clearly an important need for biomass characterization techniques that can provide meaningful information about a particular phenotypic trait of a plant such as cell wall composition, cell wall organization, or processability. Plant biomass is comprised primarily of cell walls by mass, which are, in turn, comprised primarily of the biopolymers that include cellulose, hemicellulose, and lignin. A wide range of wet chemical approaches are available for determining the content of structural polysaccharides and lignins in plants [3–9]. Select compositional and structural features of the plant cell wall are known to exert a positive or negative impact on the plant's response to specific conditions for pretreatment and enzymatic hydrolysis [10, 11], with lignin content and its alteration during pretreatment as an obvious example [12]. Saponifiable acetate and hydroxycinnamates may influence the response to hydrolysis only, pretreatment and hydrolysis, and ruminant digestibility [10, 13, 14] and, furthermore, can act as fermentation inhibitors [15]. Additionally, differences in pectic polysaccharides [16] and even the content of extractives [17] have been linked recently to differences in the cell wall's response to pretreatment and enzymatic hydrolysis. Other work has attempted to link structural properties, including how plant cell walls and their surfaces sorb water [18], to the cell wall's response to enzymatic hydrolysis. However, determining the sugar yields and potentially hydrolysate fermentability of a particular biomass feedstock following a pretreatment and enzymatic hydrolysis based solely on compositional or structural data is not sufficiently informative and typically requires the collection of empirical data under relevant conditions.

In the context of lignocellulosic biorefineries, extracting information about plant cell wall composition or processability utilizing high-throughput characterization approaches is important in at least two relevant applications. In the first application, high-throughput characterization can be applied as a tool for screening large panels of plant biomass samples representing differences in genotype, environment, or plant development stage to identify phenotypes that are promising for a specific conversion process or set of processing conditions. It should be noted that a "reduced recalcitrance" phenotype and the cell wall properties contributing to this phenotype differ depending on whether a pretreatment is performed, the type and severity of the pretreatment, and the conditions used for enzymatic hydrolysis [10, 19–21]. In a lignocellulosic biofuel process, it is expected that even when utilizing a single feedstock such as corn stover, substantial variability will be encountered as a consequence of differences in the feedstock's genotype, growth conditions and inputs, environment, harvest time, and storage history. Therefore, a second important potential application for high-throughput biomass characterization would be to assess variability in feedstock quality and to use this information to determine necessary processing conditions (e.g., pretreatment conditions, enzyme loadings, and/or hydrolysis times) to reach a target sugar yield.

Wet chemical methods have been developed and employed that are high-throughput and automated and utilize small sample masses for composition analysis [7, 8, 22], and which can be used to assess the response to pretreatment and hydrolysis [23–26]. However, these techniques may still require time-consuming sample preparation and often may lack relevant information that is provided by other lower-throughput methods. While wet chemical data are preferable, a number of high-throughput biomass characterization approaches are available based on indirect characterization methods. These high-throughput approaches for characterizing plant cell wall composition, properties, or the response to processing include non-destructive techniques such as infrared spectroscopy [27–30], Raman scattering spectroscopy [27], NMR spectroscopy [31], and destructive characterization techniques such as analytical pyrolysis [32, 33]. While direct correlations between the data obtained from these techniques and the plant cell wall property of interest have been established [34], these techniques are typically coupled to chemometric models. These include principal component analysis (PCA) that can be applied to reduce the dimensionality and remove linear correlations within the data set and partial least squares (PLS) regression models that can be applied to correlate the principal components to easily identifiable cell wall features. However, using either approach requires accurate wet chemical data to calibrate and validate the models.

Relevant to the current work, both infrared spectroscopy and analytical pyrolysis have been widely applied as high-throughput tools for characterizing plant biomass and have been the subject of a recent review [35]. As examples, infrared spectroscopy has been applied to predict composition [28, 30, 36, 37] including the content and monomer residue abundance for structural polysaccharides, lignin, and extractives as well as minor compositional differences (e.g., hemicellulose sugars, uronic acids, acetate, hydroxycinnamates), water-extractable sugars and starch, nitrogen content [29], and lignin properties such as syringyl-to-guaiacyl (S/G) ratio [38] or total lignin content [39]. Infrared spectral assignments for lignins are made either empirically or based on known assignments for model compounds [40]. Analytical pyrolysis has been used as a tool for characterizing macromolecules, including plant cell wall polymers by profiling the major pyrolysis-derived cell wall fragments by mass spectrometry [41, 42]. Although lignin macromolecules are relatively difficult to be depolymerized by traditional wet chemistry methods, lignin properties such as S/G ratio and hydroxycinnamic acid content have been assessed by pyrolysis gas chromatography and mass spectrometry (py-GC/MS) of these pyrolyzable lignin-derived fragments [43]. Pyrolysis molecular beam mass spectrometry (py-MBMS) profiles the entire set of molecular ion peaks without GC separation, which increases the throughput. As examples, py-MBMS coupled to PLS regression has been applied to predict composition [33] or lignin content in grasses [44, 45] and in diverse herbaceous plants subjected to diverse pretreatments [32]. These techniques have been extended to predict the response of the plant cell wall to a deconstruction and conversion process. For example, PLS models have been used to predict enzymatic hydrolysis yields or in vitro ruminant digestibility from infrared spectra for diverse feedstocks including untreated grasses [46], dilute acid-pretreated grasses (based on feedstock analysis prior to pretreatment) [47], alkali-pretreated diverse grasses and hardwoods (based on feedstock analysis following pretreatment composition) [48], and AFEX pretreatment of rice straw [49], although with poor results. As a final example of relating hydrolysis yields to cell wall properties, other types of models have been developed to relate the cell wall properties including cellulose crystallinity [50], lignin content, and *p*-coumaric and ferulic acid content [51] to hydrolysis yields by either neural network [52] or regression models [51].

In our previous work, we identified correlations between responses to pretreatment and enzymatic hydrolysis yields and cell wall properties such as composition, lignin content, S/G ratio, and the hydroxycinnamic content in a maize diversity panel [10]. Using the same sample set, in the present study, py-MBMS and near-infrared (NIR) characterization are applied and several modeling approaches are employed to predict cell wall properties and the cell wall's response to alkaline pretreatment and enzymatic hydrolysis. Specifically, in the first part of this work, PCA is applied to the py-MBMS spectra obtained from the complete maize diversity panel in addition to several other diverse grasses to identify cell wall properties that contribute most strongly to the variance in the principal components and how these can be used to distinguish between the plant species and diverse maize genotypes. Next, py-MBMS and NIR spectra from the maize diversity panel are used to develop PLS models to predict cell wall properties and hydrolysis yields for untreated and mild alkali-pretreated biomass. The performance of these PLS models for predicting hydrolysis yields based on indirect characterization is next compared to the performance of multiple linear regression (MLR) models based on the direct quantification of biomass properties. Conclusions are drawn about the capability of the various characterization and modeling techniques to extract information relevant to predicting the deconstruction and conversion behavior of this sample set.

Materials and Methods

Generation of Biological Materials

The maize diversity panel comprising biological replicates of 27 lines was described in our previous work [10]. Briefly, the set of lines was grown in four row plots in Arlington, Wisconsin, during the summer of 2012. Rows are 6.8 m long and 0.7 m between rows and planted at a target density of approximately 70,000 plants per ha. Plants were harvested at grain physiological maturity using a modified single path combine, which allows harvest and separation of grain and biomass simultaneously. A sub-sample of approximately 1 kg per plot was obtained and dried at 50 °C for approximately 7 days. Samples of *Miscanthus* (*Miscanthus* × *giganteus*), sideoats grama (*Bouteloua curtipendula* Michx.), and big bluestem (*Andropogon gerardii*) were obtained from the Michigan State University Crop and Soil Science Research Farm as reported previously [53]. Two cultivars of switchgrass (*Panicum virgatum* L.) were employed in this work. The upland cultivar "Shawnee" was grown at the Arlington Agricultural Research Station in Arlington, Wisconsin, as reported in previous work [54] and harvested in Fall 2011. The lowland cultivar "Alamo" was grown in Ardmore, Oklahoma, and provided the Samuel Roberts Noble Foundation and was harvested in November 2007. The dried material for all samples was further ground to pass through a 1-mm mesh prior to analysis.

The approach for alkaline pretreatment, enzymatic hydrolysis, and estimation of yields was also reported in our previous work [10]. Briefly, for alkaline pretreatment, approximately 2.0 g of air-dried biomass of known moisture content was added to 20 mL of 0.8 % (wt/wt) NaOH aqueous solution and incubated in a water bath at 80 °C for 1 h. After pretreatment, the liquid was removed via filtration and the residual biomass was washed by deionized water until the wash water was neutral. The mass yields for pretreatment was determined by measuring the difference between the mass of the original and air-dried pretreated materials on a dry basis. The enzymatic hydrolysis was performed at pH 5.0 using 50 mM Na-citrate buffer, 50 °C, with orbital shaking at 180 rpm, and an enzyme loading of 30 mg protein/g glucan (CTec2, Novozymes A/S, Bagsværd, Denmark) for 6 h (untreated) or 72 h (pretreated). The glucan yield was determined as the amount of glucan released after enzymatic hydrolysis as a fraction of the glucan content in the samples following pretreatment.

Physical Characterization of Biomass Samples

The characterization methods for structural carbohydrates, acetate, hydroxycinnamates, and water retention values (WRVs) for the maize diversity panel were reported in our previous work [10]. The hydroxycinnamates were determined by saponification of 0.5 g of dry biomass samples in 25 mL of 3 M NaOH at 120 °C for 1 h. After cooling to room temperature, 250 μL of 10 mg/mL *o*-coumaric acid in methanol was added as an internal standard. The mixture was transferred to 1.5-mL centrifuge tubes and centrifuged at 13,000 rpm for 10 min. The pH of the supernatant was adjusted to 2.0 using concentrated HCl, and the samples were then stored overnight at 4 °C. These samples were subsequently analyzed by HPLC (Agilent 1100 Series) equipped with a C18 column (Discovery, 5 μm particle size, 500 × 2.1 mm; Sigma-Aldrich). The binary solvent gradient consisted with detection at 280 nm. The binary solvent gradient consisted of a 1 % acetic acid in water (solvent A) and 1 % acetic acid in 50 % aqueous methanol (solvent B). The flow of solvent B was increased by 5 % per min until 80 % solvent B was reach, and this was held for 3 min followed by increasing to 100 % solvent for 5 min and returning to solvent A for 5 min. Quantitative thioacidolysis was performed according to the procedure described in previous work [43]. A complete property data set is available in supplemental materials (Supplemental Table S1).

The py-MBMS analysis was performed according to the methods outlined previously [44] with the complete py-MBMS spectra for this data set that are available as supplemental materials (Supplemental Table S2). NIR reflectance spectra over the range of 800–2500 nm were obtained using a Foss NIRS DS2500 (Foss North America, Eden Prairie, MN). Each dried and ground (<1 mm) sample was scanned in replicates, and the average of three to six measurements was used in the analysis. The complete NIR spectra for this data set are available as supplemental materials (Supplemental Table S3).

Model Development and Analysis

The PCA for between-species comparison was performed on duplicate biomass samples and py-MBMS spectra in the range of 51–450 *m*/*z* without autoscaling using the *princomp* function in MATLAB (MathWorks, Natick, MA). PLS models were developed using the *plsregress* function in MATLAB, which employs SIMPLS as proposed by de Jong [55] whereby the data are mean-centered but not rescaled with respect to the standard deviation. A range of 30–450 *m*/*z* was used for the py-MBMS data and second derivatives of the NIR spectra in the range of 1400 to 2500 nm.

Combinations of the 12 previously described cell wall properties were employed in MLR models to predict pretreated or untreated hydrolysis yields. Several model selection algorithms were compared, including Akaike information criterion (AIC) and Bayesian information criterion (BIC). Stepwise selection was performed using the R software environment and the *stepAIC* function in the MASS package and *regsubsets* function in LEAPS package, respectively. The multiple R^2 values were compared for the selected models, and the model with the least number of variables and the highest value of R^2 was selected to predict the hydrolysis rate of untreated maize and hydrolysis yield of the pretreated maize.

Results and Discussion

Principal Component Analysis of Py-MBMS Spectra

The use of chemometric techniques (e.g., principal component or discriminant analysis) to classify or distinguish phylogenetically diverse plants is relatively straightforward when significant compositional differences may contribute to variability in the high-dimensionality descriptor data set. There are abundant examples in the literature that demonstrate that principal components (PCs) obtained from py-MBMS spectra can be utilized for classification of grasses [56], diverse plants and cell wall polymers [57], and diverse pretreatments [35]. When applied to more closely related species or even within species diversity, extracting information that distinguishes phenotypic differences may be more challenging. In this part of the work, the principal components derived from the py-MBMS spectra of a maize diversity panel consisting of 27 maize lines described previously plus four other grass species including two cultivars of switchgrass, *Miscanthus*, sideoats grama, and big bluestem. These results demonstrate firstly that the diverse maize lines can be segregated from the other grasses only in the third PC (PC_3) (Fig. 1a), with additional PCs not demonstrating clear differences between biomass classes (data not shown).

These diverse grasses exhibit substantial structural differences as well as substantial differences in composition. Compositional differences include differences in the relative abundance of the main cell wall biopolymers (i.e., lignins, hemicelluloses, and cellulose), the monomer makeup of these biopolymers, and the relative abundance and spectrum of extractives in the plants. Lastly, minor structural components including acyl substitutions on lignins and hemicelluloses (i.e., acetyl and *p*-coumaryl esters), ferulate ester and ether cross-links in lignins and hemicelluloses, and pectic polysaccharide content may have important implications for the higher-order structure of plant cell walls and, furthermore, may represent substantial diversity between plants. It should be noted that rather than differences between a single chemical structural feature of cell wall polymers and extractives of the different biomass samples, PCA of py-MBMS spectra can identify multiple features simultaneously. Quantifiable compositional differences between the maize diversity panel and the diverse

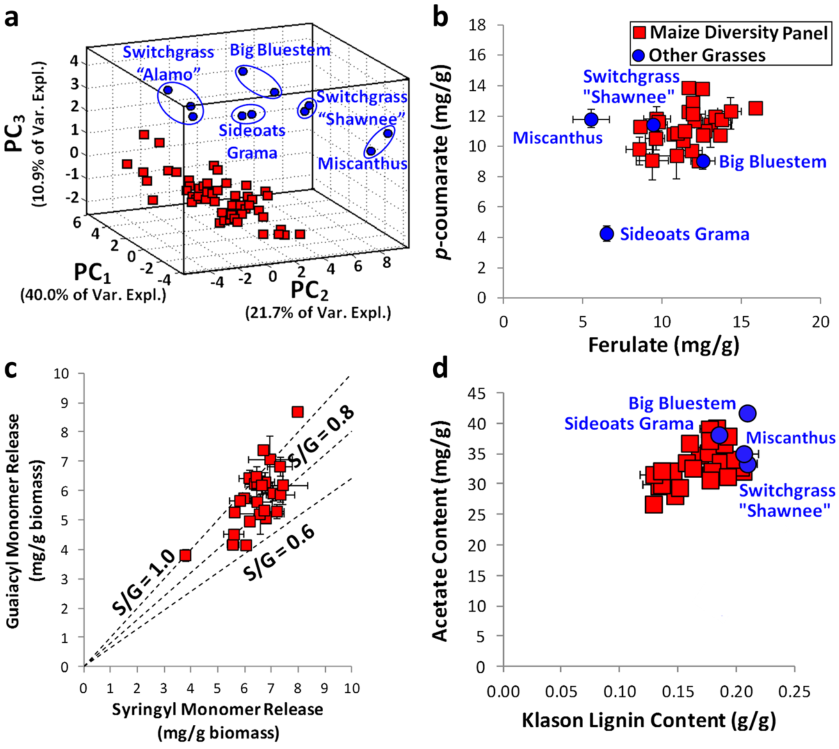


Fig. 1 Observable differences in the **a** first three principal components representing 72.6 % of the variance in the py-MBMS spectra for the maize diverse panel plus five other grasses together with the properties that may contribute to these differences including **b** the hydroxycinnamic acid content of these samples, **c** the lignin-derived monomer yields and S/G ratios of the maize diversity panel as determined by thioacidolysis, and **d** the Klason lignin and saponifiable acetate content of these samples

grasses that could be manifested in differences in py-MBMS spectra are the hydroxycinnamic acid content (i.e., *p*-coumaric acid and ferulate) as shown in Fig. 1b, the syringyl and guaiacyl monomer yield and ratio as determined by thioacidolysis (Fig. 1c), and the lignin and acetate content (Fig. 1d).

To assess individual molecular ion contributions to the differences in the PCs, PC loadings are plotted (Fig. 2). A number of molecular ion peaks that may contribute to the differences between maize panel and the other grasses are highlighted. Among other peaks, this shows that molecular ion peaks at 120 and 150 *m*/*z* can be linked to differences in PC_3. These two peaks have previously been demonstrated to be derived from 4-vinylphenol and 4-vinylguaiacol, respectively, which, in turn, are the pyrolysis products of cell wall biopolymer-associated *p*-coumarate and ferulate and have been implicated in the past as contributing to substantial differences in the pyrolytic products of graminaceous monocots versus dicots [42, 58]. While clearly there are other significant differences within this sample set, differences in hydroxycinnamate abundance can be clearly implicated as a substantial contributor to the variance in the py-MBMS spectra and as well as a property capable of discriminating between diverse grasses.

Prediction of Cell Wall Properties in Diverse Maize Lines from Py-MBMS and NIR Spectra

As just demonstrated, discriminating differences in the pyrolytic products in taxonomically diverse plants or between taxonomically related species may be possible due to substantial differences in composition. However, in comparing closely related samples (e.g., within-species diversity, association panels, or isogenic lines grown under different conditions), it may be more difficult to confidently assess minor phenotypic differences. For the next part of this work, the goals were to

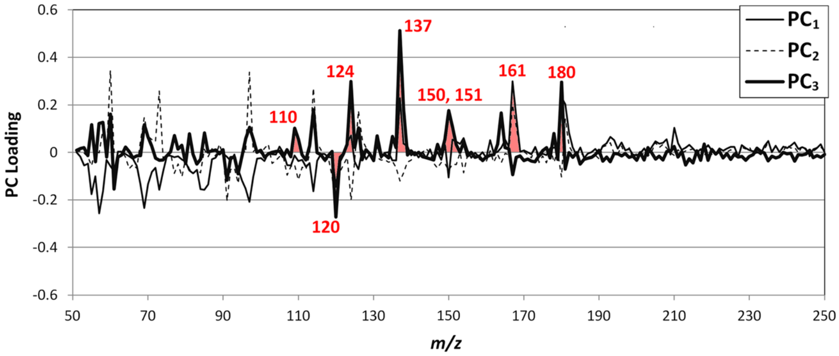


Fig. 2 Contribution of individual molecular ions from the py-MBMS spectra to the first three principal components for the PCA of the maize diversity panel plus five other grasses

understand how variance within the py-MBMS and NIR spectra for the maize diversity panel could be related to minor compositional differences including Klason lignin content, syringyl and guaiacyl lignin monomer release by thioacidolysis, hydroxycinnamate content, and acetate content through either direct correlations to specific molecular ions (py-MBMS) or wavelengths (NIR) or through application of partial least squares (PLS) regression models.

The example spectra for py-MBMS and NIR are presented in the top panels of Fig. 3(a, b), respectively, while the corresponding heat maps showing the strength of the linear correlation between regions of the py-MBMS or NIR spectra and quantified cell wall properties or composition (glucan, xylan, Klason lignin, guaiacyl and syringyl monomer yields from thioacidolysis, hydroxycinnamate content, and acetate content) from the complete maize diversity panel are presented in the bottom panels of Fig. 3(a, b). The NIR spectra were normalized as the second derivative of reflectance with respect to wavelength in order to remove baseline differences prior to analysis. Select individual mass ion peaks from the py-MBMS spectra or wavelengths from NIR data exhibit strong correlations to individual properties may act as diagnostic peaks for these properties (Figs. 4 and 5). From these results, a number of important observations can be made. In comparing the correlations between maize cell wall properties and the py-MBMS spectra, substantial differences can be observed between the first three categories in the heat map (glucan, xylan, and Klason lignin). These show that regions of the py-MBMS spectra characteristic for each of the cell wall biopolymer fractions exhibit positive correlations (i.e., these are more green in the heat map) and that regions correlated to another biopolymer fraction exhibit negative correlations (i.e., these are more red in the heat map). Regions of the spectra that may derive from pyrolysis products of hexosans, including 60, 73, and 114 *m*/*z* as identified in previous work [56], are positively correlated to glucan and exhibit negative correlations to lignin and xylan (Fig. 3(a)). This relationship is reasonable as samples with elevated glucan contents would simultaneously exhibit lower lignin and xylan contents. A similar phenomenon is observed for the NIR data (Fig. 3(b)) whereby glucan and lignin exhibit opposite correlation strengths for some regions of the spectra as a consequence of higher lignin contents corresponding to lower polysaccharide contents. As another example of this phenomenon, clusters of peaks centering on 68, 81, and 95 *m*/*z* in the py-MBMS spectra (Fig. 3(a)) have been identified as originating from extractives [42] and correspond to strong negative correlations to most of the other properties.

Another important observation is that many of the properties presented exhibit similar profiles in the heat map, including Klason lignin, guaiacyl and syringyl monomer release by thioacidolysis, *p*-coumarate, and acetate for both the py-MBMS spectra (Fig. 3(a)) and the NIR spectra (Fig. 3(b)). This is not surprising, as some of these properties were found to be strongly positively correlated to each other for this maize diversity panel in our previous work [10]. As such, it should be stressed that specific molecular ion peaks from the py-

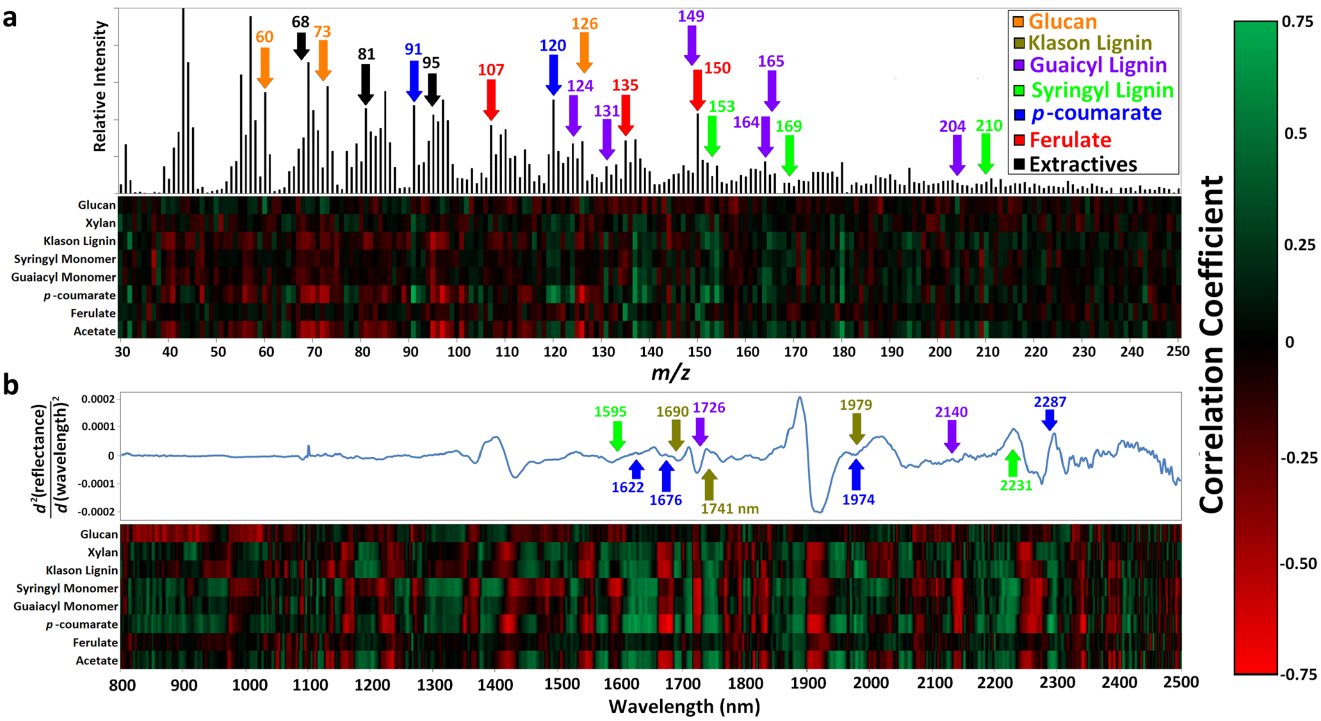
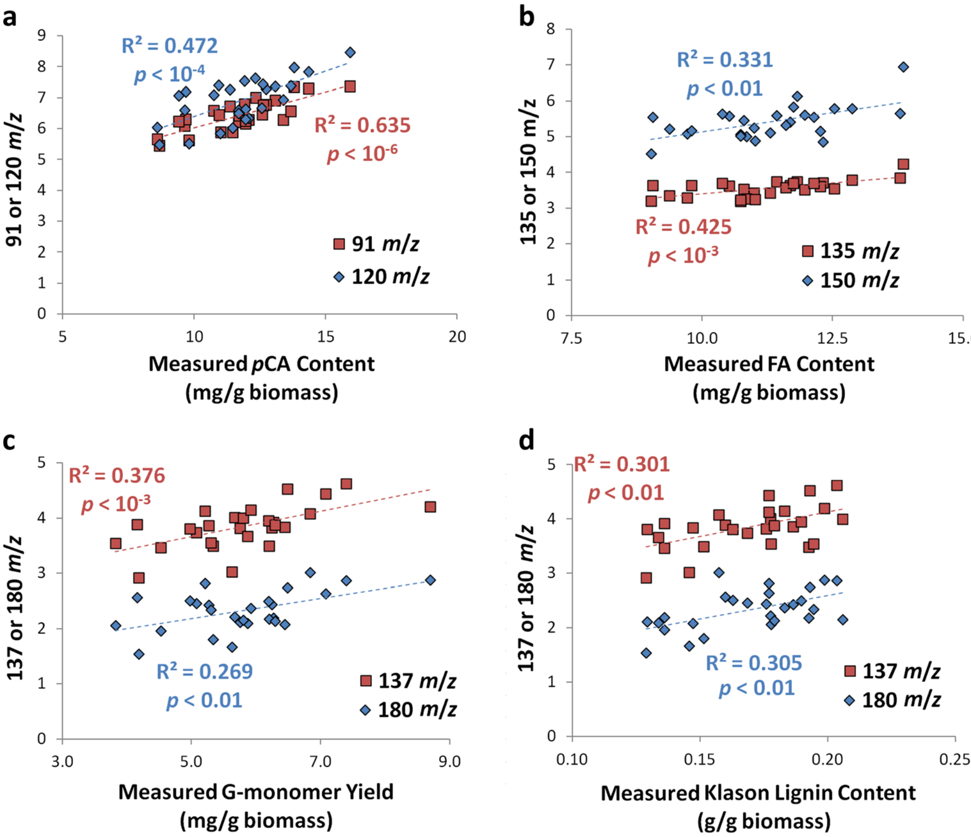


Fig. 3 Characteristic spectra for *a* py-MBMS (*top panel*) and *b* NIR (*top panel*) and heat maps of Pearson's correlation coefficients (*bottom panels*) between quantified cell wall properties from the complete maize diversity panel and the respective set of spectral data

Fig. 4 Example correlations between individual molecular ion peaks from py-MBMS spectra biomass properties for the maize diverse panel including **a** *p*-coumarate, **b** ferulate, **c** guaiacyl monomer release by thioacidolysis, and **d** Klason lignin



MBMS spectra or regions of the NIR spectra exhibiting strong correlations to a compositional property do not necessarily imply that these are diagnostic for that compositional property. For example, while there are many molecular ion peaks that are strongly correlated to acetate content (Fig. 3(a)), it is not possible to ascertain whether these correlations are independent of other compositional contributions. Our previous work with this sample set found acetate to be strongly positively correlated to both lignin content and *p*-coumarate content [10]; therefore, it is reasonable that strong positive correlations to molecular ion peaks with masses greater than that of acetate (137, 152, 180, and 210 *m*/*z*) are not actually diagnostic peaks but are due to the correlation of acetate content to other properties such as lignin content. It is notable that the ferulate exhibits distinct correlation profiles from any of the other aromatics in both the py-MBMS spectra and the NIR spectra. This indicates that this property can be individually differentiated from the other aromatics as in the py-MBMS spectra (Fig. 3(a)) or cannot be uncoupled from the signal of other aromatics in the NIR (Fig. 3(b)), resulting in a poor correlation across the entire measured spectra.

In the py-MBMS spectra, a number of peaks can be clearly identified as exhibiting strong positive correlations to the *p*-coumarate (*p*CA) content (91 and 120 *m*/*z*; Fig. 4a) and ferulate (FA) content (107, 135, and 150 *m*/*z*; Fig. 4b). These molecular ion peaks have previously been linked to 4-vinylguaiacol and 4-vinylphenol, respectively [42], representing the pyrolytic products of these two hydroxycinnamic acids and provide strong evidence that this correlations identified are indeed authentic. Interestingly, Penning et al. found no correlation between 120 and 150 *m*/*z* molecular ion peaks from py-MBMS and the *p*CA and FA content in diverse switchgrass samples [45], although the authors employed a different analytical method for both saponifying and quantifying alkali-solubilized hydroxycinnamate monomers. An additional difference is that the biomass samples in the present study did not have extractives removed prior to either py-MBMS or wet chemical analysis. Other strong individual correlations were identified from the analysis and diagnostic peaks for both guaiacyl monomer and total Klason lignin at mass ion peaks of 137 and 180 *m*/*z* (Fig. 4c, d). Previous work employing py-MBMS of 282 maize lines identified strong ($R^2 > 0.85$), significant correlations between the abundance of guaiacyl monomers released by CuO oxidation and the sum of mass ion peaks at 124, 137, 138, and 151 *m*/*z* and between the syringyl monomers released by CuO oxidation and the sum of mass ion peaks at 154, 167, 168, and 198 *m*/*z* [45]. These combinations of mass ion peaks did not yield significant correlations in the present data set for guaiacyl or syringyl monomer released by thioacidolysis (data not shown). Reasons for this may be differences in the fraction of lignin monomers quantified by each method, potentially with CuO oxidation yielding monomers in a similar proportion to what is obtained by pyrolysis.

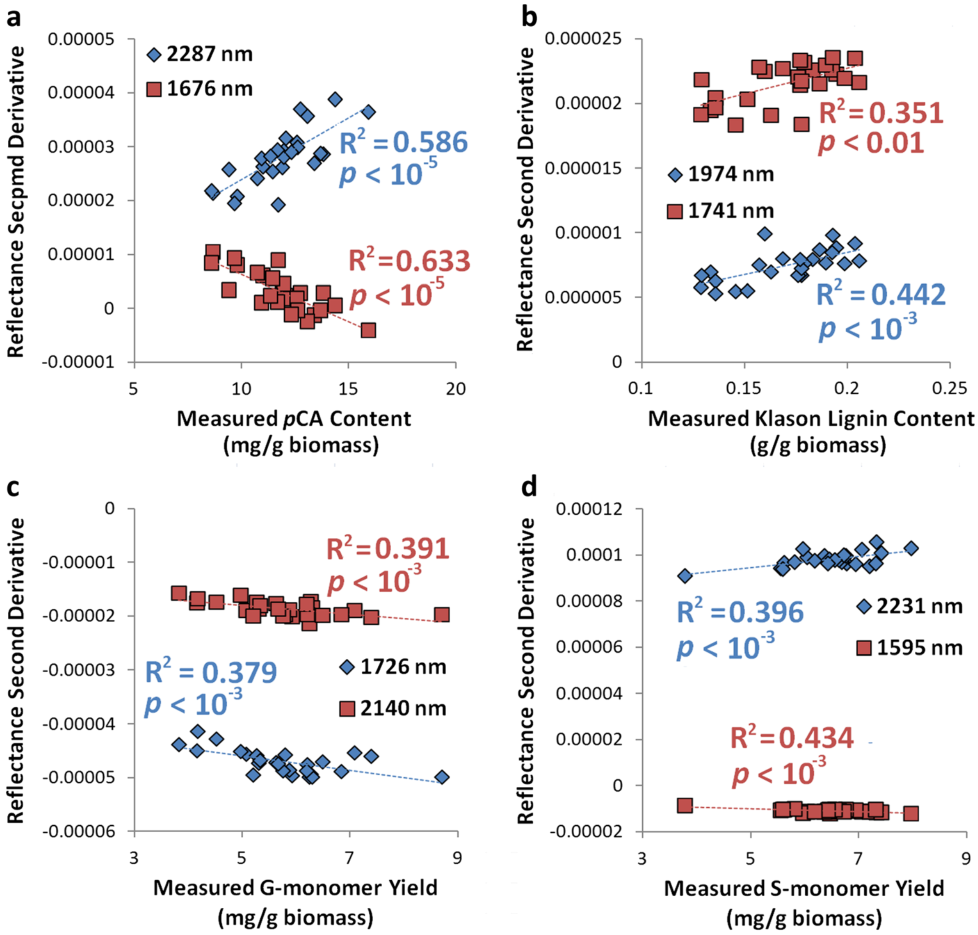


Fig. 5 Example correlations between regions of the NIR spectra and select cell wall properties including **a** *p*-coumarate, **b** Klason lignin, **c** guaiacyl, and **d** syringyl monomer release by thioacidolysis

Notable regions of the NIR spectra that exhibit the strongest correlations to select compositional properties (Fig. 3(b)) are replotted in Fig. 5. From this plot, it can be observed that *p*CA exhibits the strongest correlations of any of the properties including regions of the spectra at 1620, 1674, 2285, and 2336 nm (Fig. 5a). While slightly less significant than the *p*CA correlations, the NIR spectra in the regions of 1668, 1686, 1741, and 1974 nm can be correlated to the cell wall Klason lignin content (Fig. 5b). Previous studies have proposed that regions of the NIR spectra corresponding to 1682 nm can be correlated to “C=O stretching” and 1688 nm to “aromatics” in wood lignins [27]. While exhibiting similar patterns to both Klason lignin and *p*CA, distinct regions of the NIR spectra can be identified exhibiting significant correlations to both syringyl and guaiacyl monomers released by thioacidolysis (Fig. 3(b)). These include 1726 and 2140 nm for guaiacyl monomers (Fig. 5c) and 1595 and 2231 nm for syringyl monomers (Fig. 5d). Previous empirical correlations have been identified at 1595 nm with the assignment “guaiacyl-syringyl” [34], while 1715 and 1722 has been linked to C=O stretching in hardwood and softwood lignins [59].

General trends observed for both the direct correlations from the py-MBMS and NIR spectra are that *p*CA content gives the strongest correlations (slightly higher for py-MBMS) followed by both guaiacyl monomer release by thioacidolysis (again, slightly higher for py-MBMS) and Klason lignin content (slightly higher for NIR). Notable differences in the direct prediction capability of the two analytical methods are that unlike the NIR data set, the py-MBMS data is not able to provide a robust prediction of syringyl monomer released by thioacidolysis. A second notable difference is that, unlike the model developed from the py-MBMS data set, the NIR data set is unable to predict ferulate content. This could be a consequence of the difficulty in distinguishing ferulate and guaiacyl monomers using non-destructive infrared characterization, whereas py-MBMS releases volatilizes the ferulate and its pyrolytic products can be quantified.

PLS models were next applied to predict plant cell wall properties (Klason lignin content, hydroxycinnamate content, and acetate content) within the maize diversity panel to determine whether multivariate combinations of the py-MBMS molecular ion spectra or NIR spectra could provide robust predictions of these properties. Importantly, in developing these models, cross-validation was applied to assess model generalizability to a validation set to prevent over-fitting of the model. For these models, the py-MBMS molecular ion abundance from 30 to 450 *m*/*z* or the NIR second derivative of reflectance from 1400 to 2500 nm was utilized without scaling. The model performance as a function of the number of PCs utilized is presented in Fig. 6. It can observed that the R^2 for the model prediction in the calibration data set (R^2_{Cal}) increases with the number of PCs (Fig. 6a, c) and will reach 1.00 if enough PCs are employed. This R^2_{Cal} is identical to the variance in the prediction variable explained by the model, and it can be observed that utilizing eight PCs or more captures more than 99.5 % of the variance in the prediction variables for the py-MBMS data (Fig. 6a) and, with the exception of ferulate, utilizing 14 PCs or more captures more than 99.5 % of the variance in the prediction for the NIR data (Fig. 6c). However, utilizing this, many PCs would likely over-fit the models as a substantial amount of this variance may not be explainable by the measurements or may be associated with measurement noise or random error. For this reason, cross-validation is applied to assess the robustness of the models and determine how many PCs should be utilized in a prediction model.

For PLS regression models, R^2 for cross-validation (R^2_{CV}) can be utilized to assess the model goodness-of-fit [60] and can include utilizing combinations of subsets of data and/or application of an independent validation set. Assessment of PLS model performance is often through the root-mean-square error (RMSE) associated with the model prediction and is sometimes defined as a predicted residual sum of squares (PRESS) when applied to a validation data set and provides an absolute assessment of the error associated with a predicted variable [61]. In the present work, R^2 terms are employed consistently as this metric can provide comparisons between variables of differing magnitudes (e.g., Fig. 6). As the sample set was relatively small in the present work (i.e., 27 samples with duplicate measurements for py-MBMS or three to six replicates for NIR), the "leave-one-out" approach for cross-validation was applied. The "one" in the leave-one-out approach included all replicate measurements; otherwise, the cross-validation would be irrelevant since R^2_{CV} as well as R^2_{Cal} would approach unity as the number of PCs is increased. The results for R^2_{CV} show that, at most, three or four PCs from the py-MBMS spectra (Fig. 6b) and six to ten PCs from the NIR spectra (Fig. 6d) could be utilized for making property predictions with the PLS models without decreasing the R^2_{CV}.

A number of trends in the property prediction capability are shared for direct correlation approaches (Figs. 4 and 5) and the PLS model predictions (Fig. 6). Specifically, the cross-validation results (Fig. 6b) show that the py-MBMS spectra can be utilized to generate the best prediction for *p*CA content (maximum R^2_{CV} of 0.60) followed by predictions for acetate (maximum R^2_{CV} of 0.43) and guaiacyl monomer released by thioacidolysis (maximum R^2_{CV} of 0.32). The predictions for cross-validation were lower for ferulate followed by Klason

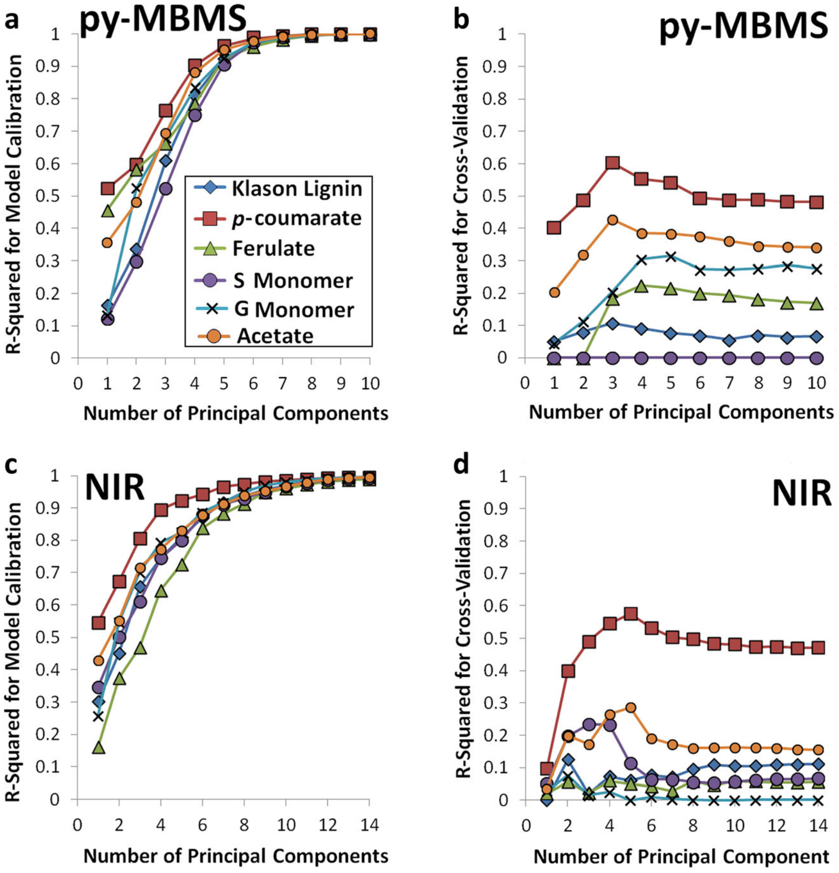


Fig. 6 PLS model development and cross-validation for estimating maize properties from py-MBMS and NIR spectra including **a**, **c** the R^2_{Cal} as a function of PCs used in the PLS model and **b**, **d** the R^2_{CV} for leave-one-out cross-validation using pooled replicates

lignin (0.22 and 0.11, respectively), indicating no predictive capability for these variables. Potential reasons for the poor prediction capability for ferulate and Klason lignin as revealed by the cross-validation may be due to either the difficulty in deconvoluting the individual contributions to each of these components by the pyrolysis products [62] or potentially due to deficiencies in the quantification method, particularly for lignin in graminaceous monocots [9]. Compounding this, it is possible that the range of some of the properties within the data set may have been insufficient to distinguish signal from noise. While select pyrolysis products have been linked to syringyl lignin monomer content (e.g., 154, 167, 168, 182, 194, 208, and 210 *m*/*z*) and used in the estimation of S/G ratios [35, 63], the current work found no strong, significant correlations to any of the individual molecular ion peaks and the syringyl monomer yield by thioacidolysis (Fig. 3(a)). This may explain why the cross-validation prediction for the syringyl monomer released by thioacidolysis was non-existent (Fig. 6b). Compared to the PLS model predictions from the py-MBMS data set, the PLS models derived from the NIR data set exhibited comparable prediction capability as assessed by cross-validation, exhibiting maximum values of R^2_{CV} of 0.58 for *p*CA, 0.29 for acetate, 0.23 for syringyl monomer yield, and less than 0.17 for all the other properties. Overall, a key finding from this analysis is that, except for *p*CA content, individual linear correlations developed directly from either the py-MBMS (Fig. 4) or from the NIR data (Fig. 5) that provided superior predictive capability to the PLS models when assessed by R^2_{CV} (Fig. 6), potentially indicating that some relevant information from the original data set is lost when transformed into the PC space and applied in PLS models.

Prediction of Enzymatic Hydrolysis Yields in Diverse Maize Lines from Quantified Properties and PLS Models Using Py-MBMS and NIR Spectra

Rather than determining structural or compositional features of plant cell walls, high-throughput analytical approaches that are derived from several contributing traits such as the cell wall's response to particular conditions for pretreatment and hydrolysis may be preferred to time-consuming wet chemical analysis. Applying a similar methodology as in the previous section, PLS models for predicting hydrolysis yields were developed and assessed based on the py-MBMS (Fig. 7a) and NIR (Fig. 7b) spectra for the maize diversity sample set. It should be noted that all models were based on analysis performed on the untreated biomass, which may contribute to weaker predictions for pretreated biomass behavior. Based on the cross-validation results (solid data points), it is clear that the model prediction capability is substantially more robust for predicting enzymatic hydrolysis yields of untreated biomass (maximum R^2_{CV} of 0.60 for the py-MBMS data and 0.62 using the NIR data) compared to the plant's response to pretreatment and hydrolysis (maximum R^2_{CV} of less than 0.05 for the py-MBMS data and 0.14 using the NIR data).

In the next part of this study, multiple linear regression (MLR) models were applied to predict these same hydrolysis yields from quantified properties for the maize diversity panel both to yield fundamental insight into the factors impacting hydrolysis yields and to use as a basis of comparison for the chemometric models. Some of the quantified properties and correlations between properties and hydrolysis yields for this data set were presented in our previous work [10]. Exceptions include the lignin syringyl and guaiacyl monomer release by thioacidolysis and the S/G ratio determined from these two terms. These correlations between quantified properties and hydrolysis yields for the 27-sample maize diversity panel are presented in Table 1. Not surprisingly, these results demonstrate that multiple properties exhibit correlations to enzymatic hydrolysis yields. According to Table 1, seven quantified properties exhibited significant correlations ($p < 0.05$) to untreated enzymatic hydrolysis yields (initial xylan, initial Klason lignin, initial *p*CA content, guaiacyl monomer released by thioacidolysis, and initial WRV), while six

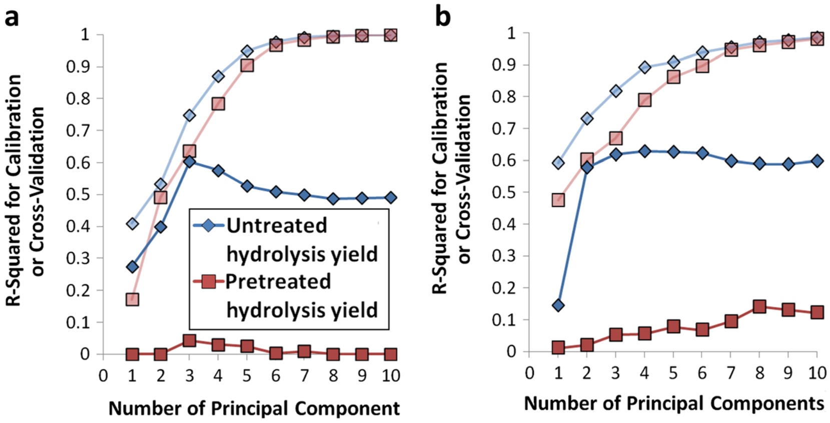


Fig. 7 PLS model development and cross-validation for estimating maize hydrolysis yields from **a** py-MBMS and **b** NIR spectra as a function of PCs used in the PLS model. *Solid data points* represent R^2_{CV} for leave-one-out cross-validation using pooled replicates and transparent data points represent R^2_{Cal}

Table 1 Summary of Pearson's correlation coefficients between individual cell wall properties and hydrolysis yields

	Initial biomass composition or property									Pretreated biomass composition or property				
	Xylan	Klason lignin	Acetate	Ferulate	*p*-Coumarate	*S*-Lignin monomer	*G*-Lignin monomer	Lignin S/G ratio	WRV	Xylan (final)	Klason lignin (final)	*p*-Coumarate release	Ferulate release	WRV release
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}
Untreated biomass hydrolysis yields	−0.45*	−0.40*	−0.53**	−0.11	−0.58**	0.19	−0.31*	0.24	0.40*	−0.31	−0.11	−0.61**	0.25	0.00
Pretreated biomass hydrolysis yields	0.13	−0.16	0.20	0.32	0.05	−0.37*	−0.58**	0.45*	0.06	0.23	−0.47*	0.40*	0.49**	−0.12

*$p < 0.05$; **$p < 0.01$

properties demonstrated significant correlations ($p < 0.05$) to the hydrolysis yields following mild alkaline pretreatment (syringyl and guaiacyl monomer released by thioacidolysis, S/G ratio, final Klason lignin, *p*CA, and FA released by pretreatment). As might be expected, the properties correlated to untreated hydrolysis yields were primarily initial biomass properties or composition, while half of the properties strongly correlated to pretreated hydrolysis yields were pretreated biomass properties. It is notable that the only properties associated with the initial cell wall composition that could predict pretreated enzymatic hydrolysis yields were associated with the thioacidolysis data (e.g., aromatic monomer yields and S/G ratio). This may explain the difficulty in predicting this yield from the py-MBMS spectra as only guaiacyl monomer release was able to be predicted from this data set (Figs. 4c and 6b).

Clearly, combinations of these variables provide more powerful predictions relative to individual variables (Table 1). However, contrasted to PLS models, utilization of highly correlated, high-dimensional data sets may be problematic for MLR models and may necessitate the selection of a subset of variables to develop robust correlation models. This variable down-selection serves the purpose of both minimizing linear redundancies and avoids the introduction of additional error into MLR models if low signal-to-noise ratios are present in some of the property measurements. A wide range of algorithms have been proposed for predictor variable selection in MLR models [64]. In the present work, the AIC [65] and BIC [66] as well as identifying the largest R^2_{Cal} for a set number of predictor variables were applied and compared to select the properties for incorporation into the MLR models that could be used for prediction of untreated and pretreated hydrolysis yields. Comparable to the approach employed for assessing PLS models, model validation was performed by fitting the MLR model parameters using the leave-one-out cross-validation approach for each set predictor and an R^2_{CV} for each of the model forms was determined. This metric was chosen to provide a basis of comparison with other prediction approaches. The selected predictor variables for MLR models along with model performance are presented in Table 2. From these results, it can be observed that the best case MLR models provide comparable performance (although slightly lower with respect to both R^2_{CV} and R^2_{Cal}) to the best case py-MBMS model for untreated biomass, while providing substantially better predictions for the hydrolysis yields of pretreated biomass relative to both the py-MBMS and NIR data sets. Specifically, it is shown that important variables for predicting untreated hydrolysis yields include S/G ratio (X_8), initial WRV (X_9), *p*CA release (X_{12}), and final WRV (X_{14}), while predictor variables chosen in linear models for alkali-pretreated corn stover include initial xylan content (X_1), initial acetate content (X_3), syringyl (X_6) and guaiacyl (X_7) monomer release by thioacidolysis, and ferulate release (X_{13}).

Table 2 Example multiple linear regression (MLR) models identified based on differing selection criteria

Selection criteria	MLR models for prediction of untreated hydrolysis yields			MLR models for prediction of pretreated hydrolysis yields		
	Model variables	R^2_{Cal}	R^2_{CV}	Model variables	R^2_{Cal}	R^2_{CV}
AIC	X_1 X_2 X_9 X_{11} X_{12} X_{14}	0.74	0.200	X_1 X_3 X_7 X_{13}[a]	0.59	0.43
BIC	X_8 X_9 X_{12} X_{14}[a]	0.67	0.509	X_1 X_3 X_7 X_{13}[a]	0.59	0.43
R^2 $(n \leq 4)$	X_8 X_9 X_{12} X_{14}[a]	0.67	0.509	X_1 X_3 X_7 X_{13}[a]	0.59	0.43
R^2 $(n \leq 5)$	X_1 X_2 X_9 X_{11} X_{12}	0.71	0.188	X_1 X_3 X_6 X_7 X_{13}	0.62	0.44

[a] Model used in validation comparison (Fig. 8)

Model predictions of the untreated and pretreated hydrolysis yields based along with cross-validation performance are shown for the PLS regression models using the py-MBMS spectra (Fig. 8a) and the NIR spectra (Fig. 8b) together with the predictions of the MLR models based the measured cell wall properties (Fig. 8c). From this, it is clear that predictions for hydrolysis yields of untreated biomass as assessed by R^2_{Cal} and R^2_{CV} are slightly better for the PLS models, while the MLR model provides a substantially better prediction (based on R^2_{CV}) for alkali-pretreated biomass. This is not surprising as the MLR models use cell wall properties following pretreatment as predictor variables while the py-MBMS and NIR PLS models are based only on the initial composition, indicating that predicting the cell wall's response to pretreatment and enzymatic hydrolysis from only initial cell wall properties is problematic. Notably, three pretreated biomass properties (final Klason lignin content, *p*CA release, and FA release) were all found to be significant predictors of pretreated hydrolysis yields (Table 1) and one of these properties (FA release) is necessary in all of the MLR models (Table 2). The implications of this are, at least for the pretreatment-biomass combinations employed in this work, that initial biomass properties are insufficient to predict the response to pretreatment and enzymatic hydrolysis and that knowledge of the cell wall's response to the pretreatment is necessary.

Conclusions

This study compared two high-throughput characterization approaches as applied to a maize diversity panel, py-MBMS and NIR, for their capability of the data obtained with these techniques to predict either plant cell wall properties or the plant cell wall's response to deconstruction using alkaline pretreatment and enzymatic hydrolysis. Cross-validation using pooled replicates is performed in order to prevent overfitting of PLS models and to provide a basis for comparison of model suitability. Furthermore, it should be noted that the sample set was relatively small for this study and that increasing the total sample size may yield improved models.

Overall, a number of key conclusions can be made. Firstly, the application of principal component analysis to py-MBMS spectra demonstrated that principal components could be used to discriminate between the diverse maize samples and other grasses and that these differences may be attributed to compositional differences, including differences in hydroxycinnamic acids. Next, the capability of py-MBMS and NIR for prediction of within-species plant cell wall properties and the cell wall's response to enzymatic hydrolysis or alkaline pretreatment and hydrolysis was assessed. Both direct correlations to characteristic region spectra and PLS models were found to be able to provide strong predictions for *p*-coumarate content, while

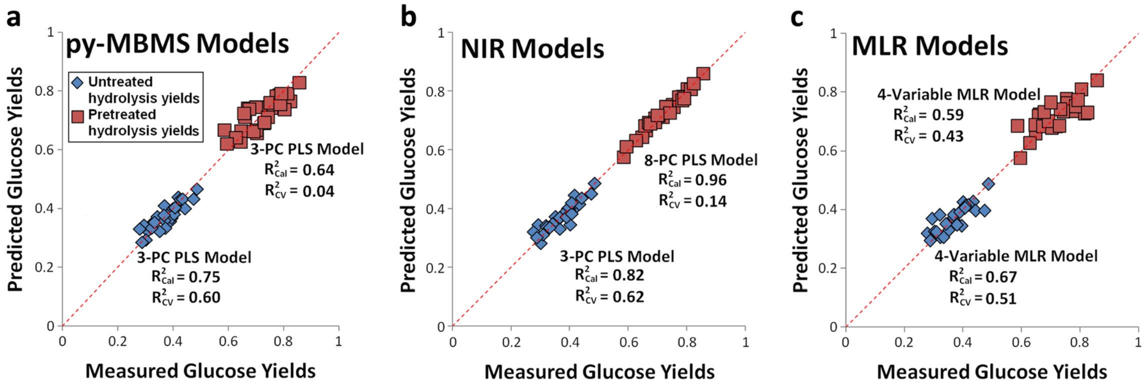


Fig. 8 Correlation between the predicted yields and the measured enzymatic hydrolysis yields for untreated maize and hydrolysis yields for NaOH-pretreated maize using **a** three-PC PLS models from py-MBMS spectra, **b** three- and eight-PC PLS models from NIR spectra, and **c** four-variable MLR models from quantified properties

guaiacyl monomer released by thioacidolysis and Klason lignin content was found to have comparably strong correlations to select regions of the NIR or py-MBMS spectra. Notably, suitable correlations for ferulate could only be determined using the py-MBMS data, while suitable correlations for syringyl monomer released by thioacidolysis could only be determined with the NIR data. As a possible indication of the signal-to-noise ratio in the experimental measurements of the response variables (properties and yields), several characterization/modeling approaches could independently yield similar predictive capabilities (as assessed by R^2_{CV}) with examples including *p*CA content and hydrolysis yields of untreated biomass. Furthermore, not all quantified properties exhibited strong links to py-MBMS and NIR data sets or to the cell wall's response to deconstruction. While this could indicate a genuine lack of correlation, this may also be a consequence of a limited range of these properties in the data set relative to the accuracy of the quantification method. Finally, it was determined for the data set tested, and the MLR models using quantified properties (including initial properties and properties following pretreatment) could predict hydrolysis yields following alkaline pretreatment while neither of the high-throughput characterization approaches (using unpretreated biomass only) could predict these yields, highlighting the difficulty in assessing and predicting a plant's response to pretreatment using only the initial properties.

AIC, Akaike information criterion; BIC, Bayesian information criterion; FA, ferulic acid; MLR, multiple linear regression; PC, principal component; PCA, principal component analysis; PLS, partial least squares; *p*CA, *p*-coumaric acid; py-MBMS, pyrolysis molecular beam mass spectrometry; RMSE, root mean square error; S/G, syringyl-to-guaiacyl ratio

Funding This work was funded by the DOE Great Lakes Bioenergy Research Center (DOE BER Office of Science DE-FC02-07ER64494). Robert Sykes was funded by the BioEnergy Science Center (BESC), a US Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science under contract no. DE-AC36-08-GO28308 with the National Renewable Energy Laboratory. Muyang Li was supported in part by a grant from the NSF (NSF CBET 1336622).

References

1. Souza GM, Victoria R, Joly C, Verdade L (eds) (2015) Bioenergy & Sustainability: bridging the gaps. SCOPE, Paris
2. Davison B, Brandt C, Guss A, Kalluri U, Palumbo A, Stouder R, Webb E (2015) The impact of biotechnological advances on the future of US bioenergy. Biofuels Bioprod Biorefin 9(5):454–467. doi:10.1002/bbb.1549
3. Sluiter A, Hames B, Ruiz R, Scarlata C, Sluiter J, Templeton D, Crocker D (2011) Determination of structural carbohydrates and lignin in biomass. Technical Report NREL/TP 10–42618.
4. Theander O, Åman P, Westerlund E, Andersson R, Pettersson D (1995) Total dietary fiber determined as neutral sugar residues, uronic acid residues, and Klason lignin (the Uppsala method): collaborative study. J AOAC Int 78(4):1030–1044
5. Van Soest P, Robertson J, Lewis B (1991) Methods for dietary fiber, neutral detergent fiber, and nonstarch polysaccharides in relation to animal nutrition. J Dairy Sci 74(10):3583–3597
6. Sluiter J, Ruiz R, Scarlata C, Sluiter AD, Templeton D (2010) Compositional analysis of lignocellulosic feedstocks. 1. Review and description of methods. J Agric Food Chem 58(16):9043–9053
7. Foster CE, Martin TM, Pauly M (2010) Comprehensive compositional analysis of plant cell walls (lignocellulosic biomass). Part I: lignin. J Vis Exp 37:e1745. doi:10.3791/1745
8. Foster C, Martin TM, Pauly M (2010) Comprehensive compositional analysis of plant cell walls (lignocellulosic biomass). Part II: carbohydrates. J Vis Exp 37:1837
9. Hatfield RD, Fukushima RS (2005) Can lignin be accurately measured? Crop Sci 45:832–839
10. Li M, Heckwolf M, Crowe JD, Williams DL, Magee TD, Kaeppler SM, de Leon N, Hodge DB (2015) Cell-wall properties contributing to improved deconstruction by alkaline pre-treatment and enzymatic hydrolysis in diverse maize (*Zea mays* L.) lines. J Exp Bot 66(14):4305–4315. doi:10.1093/jxb/erv016
11. Penning BWSR, Babcock NC, Dugard CK, Held MA, Klimek JF, Shreve JT, Fowler M, Ziebell A, Davis MF, Decker SR, Turner GB, Mosier NS, Springer NM, Thimmapuram J, Weil CF, McCann MC, Carpita NC (2014) Genetic determinants for enzymatic dgestion of lignocellulosic biomass are independent of those for lignin abundance in a maize recombinant inbred population. Plant Physiol 165: 1475–1487
12. Yu Z, Jameel H, H-m C, Park S (2011) The effect of delignification of forest biomass on enzymatic hydrolysis. Biores Technol 102(19): 9083–9089. doi:10.1016/j.biortech.2011.07.001
13. Jung HG, Mertens DR, Phillips RL (2011) Effect of reduced ferulate-mediated lignin/arabinoxylan cross-linking in corn silage on feed intake, digestibility, and milk production. J Dairy Sci 94(10):5124–5137. doi:10.3168/jds.2011-4495
14. Grabber JH, Hatfield RD, Ralph J (1998) Diferulate cross-links impede the enzymatic degradation of non-lignified maize walls. J Sci Food Agric 77(2):193–200
15. Sato TK, Liu T, Parreiras LS, Williams DL, Wohlbach DJ, Bice BD, Ong IM, Breuer RJ, Qin L, Busalacchi D (2014) Harnessing genetic diversity in *Saccharomyces cerevisiae* for fermentation of xylose in hydrolysates of alkaline hydrogen peroxide-pretreated biomass. Appl Environ Microbiol 80(2):540–554
16. Wang Y, Huang J, Li Y, Xiong K, Wang Y, Li F, Liu M, Wu Z, Tu Y, Peng L (2015) Ammonium oxalate-extractable uronic acids positively affect biomass enzymatic digestibility by reducing lignocellulose crystallinity in *Miscanthus*. Biores Technol 196:391–398
17. Yeh T, Chang M, Chnag W (2014) Comparison of dilute acid and sulfite pretreatments on *Acacia confusa* for biofuel application adn the influence of its extractives. J Agric Food Chem 62(44):10768–10775
18. Williams D, Hodge D (2014) Impacts of delignification and hot water pretreatment on the water induced cell wall swelling behavior

of grasses and its relation to cellulolytic enzyme hydrolysis and binding. Cellulose 21(1):221–235. doi:10.1007/s10570-013-0149-3

19. Zhang T, Wyman CE, Jakob K, Yang B (2012) Rapid selection and identification of Miscanthus genotypes with enhanced glucan and xylan yields from hydrothermal pretreatment followed by enzymatic hydrolysis. Biotechnol Biofuels 5(1):56
20. Lindedam J, Andersen SB, DeMartini J, Bruun S, Jørgensen H, Felby C, Magid J, Yang B, Wyman C (2012) Cultivar variation and selection potential relevant to the production of cellulosic ethanol from wheat straw. Biomass Bioenergy 37:221–228
21. Vandenbrink JP, Delgado MP, Frederick JR, Feltus FA (2010) A sorghum diversity panel biofuel feedstock screen for genotypes with high hydrolysis yield potential. Ind Crop Prod 31(3):444–448. doi:10.1016/j.indcrop.2010.01.001
22. DeMartini JD, Studer MH, Wyman CE (2011) Small-scale and automatable high-throughput compositional analysis of biomass. Biotechnol Bioeng 108:306–312
23. Decker SR, Brunecky R, Tucker MP, Himmel ME, Selig MJ (2009) High-throughput screening techniques for biomass conversion. BioEnergy Res 2(4):179–192
24. Santoro N, Cantu SL, Tornqvist C-E, Falbel TG, Bolivar JL, Patterson SE, Pauly M, Walton JD (2010) A high-throughput platform for screening milligram quantities of plant biomass for lignocellulose digestibility. BioEnergy Res 3(1):93–102
25. Zhang H, Fangel J, Willats G, Selig M, Lindedam J, Jørgensen H, Felby C (2014) Assessment of leaf/stem ratio in wheat straw feedstock and impact on enzymatic conversion. GCB Bioenerg 6(1): 90–96
26. Lindedam J, Bruun S, Jørgensen H, Decker SR, Turner GB, DeMartini JD, Wyman CE, Felby C (2014) Evaluation of high throughput screening methods in picking up differences between cultivars of lignocellulosic biomass for ethanol production. Biomass Bioenerg 66:261–267. doi:10.1016/j.biombioe.2014.03.006
27. Xu F, Yu J, Tesso T, Dowell F, Want D (2013) Qualitative and quantitative analysis of lignocellulosic biomass using infrared techniues: a mini-review. Appl Energ 104:801–809
28. Liu L, Ye XP, Womac AR, Sokhansanj S (2010) Variability of biomass chemical composition and rapid analysis using FT-NIR techniques. Carb Polym 81:820–829
29. Labbé N, Ye XP, Franklin JA, Womac AR, Tyler DD, Rials TG (2008) Analysis of switchgrass characteristics using near infrared spectroscopy. BioRes 3:1329–1348
30. Hames B, Thomas SR, Sluiter A, Roth C, Templeton D (2003) Rapid biomass analysis: new tools for composition analysis of corn stover feedstocks and process intermediates from ethanol production. Appl Biochem Biotechnol 105(1–3):5–16
31. Gjersing E, Happs RM, Sykes RW, Doeppke C, Davis MF (2013) Rapid determination of sugar content in biomass hydrolysates using nuclear magnetic resonance spectroscopy. Biotechnol Bioeng 110(3):721–728. doi:10.1002/bit.24741
32. Kelley SS, Rowell RM, Davis M, Jurich CK, Ibach R (2004) Rapid analysis of the chemical composition of agricultural fibers using near infrared spectroscopy and pyrolysis molecular beam mass spectrometry. BIomass Bioenerg 27:77–88
33. Sykes R, Gjersing E, Doeppke C, Davis M (2015) High-throughput method for determining the sugar content in biomass with pyrolysis molecular beam mass spectrometry. Bioenerg Res 8(3):964–972. doi:10.1007/s12155-015-9610-5
34. Herget HL (1971) Infrared spectra. In: Sarkanen KV, Ludwig CH (eds) Lignins: occurence, formation, structure and reactions. Wiley-Interscience, New York
35. Xiao L, Wei H, Himmel ME, Jameel H, Kelley SS (2014) NIR and py-MBMS coupled with multivariate data analysis as a high-throughput biomass characterization technique: a review. Front Plant Sci 5. doi:10.3389/fpls.2014.00388
36. Hiukka R (1998) A multivariate approach to the analysis of pine needle samples using NIR. Chemom Intell Lab Syst 44(1–2):395–401. doi:10.1016/S0169-7439(98)00067-7
37. Kelley SS, Jellison J, Goodell B (2002) Use of NIR and pyrolysis-MBMS coupled with multivariate analysis for detecting the chemical changes associated with brown-rot biodegradation of spruce wood. FEMS Microbiol Lett 209(1):107–111
38. Robinson AR, Mansfield SD (2009) Rapid analysis of poplar lignin monomer composition by a streamlined thioacidolysis procedure and near-infrared reflectance-based prediction modeling. Plant J 58(4):706–714
39. Zhou G, Taylor G, Polle A (2011) FTIR-ATR-based prediction and modelling of lignin and energy contents reveals independent intra-specific variation of these traits in bioenergy poplars. Plant Met 7(9):9–19
40. Schultz TP, Glasser WG (1986) Quantitative structural analysis of lignin by diffuse reflectance Fourier transform infrared spectroscopy. Holzforschung 40:37–44
41. Obst JR (1983) Analytical pyrolysis of hardwood and softwood lignins and its use in lignin-type determination of hardwood vessel elements. J Wood Chem Technol 3:377–397
42. Ralph J, Hatfield RD (1991) Pyrolysis-GC-MS characterization of forage materials. J Agric Food Chem 39(8):1426–1437. doi:10.1021/jf00008a014
43. Li M, Foster C, Kelkar S, Pu Y, Holmes D, Ragauskas A, Saffron C, Hodge D (2012) Structural characterization of alkaline hydrogen peroxide pretreated grasses exhibiting diverse lignin phenotypes. Biotechnol Biofuel 5(1):38
44. Mann DGJ, Labbe N, Sykes RW, Gracom K, Kline L, Swamidoss IM, Burris JN, Davis M, Stewart CN (2009) Rapid assessment of lignin content and structure in switchgrass (*Panicum virgatum* L.) grown under different environmental conditions. Bioenerg Res 2(4):246–256. doi:10.1007/s12155-009-9054-x
45. Penning B, Sykes R, Babcock N, Dugard C, Klimek J, Gamblin D, Davis M, Filley T, Mosier N, Weil C, McCann M, Carpita N (2014) Validation of PyMBMS as a high-throughput screen for lignin abundance in lignocellulosic biomass of grasses. Bionerg Res 7(3):899–908. doi:10.1007/s12155-014-9410-3
46. Nousiainen J, Ahvenjärvi S, Rinne M, Hellämäki M, Huhtanen P (2004) Prediction of indigestible cell wall fraction of grass silage by near infrared reflectance spectroscopy. Anim Feed Sci Technol 115(3–4):295–311. doi:10.1016/j.anifeedsci.2004.03.004
47. Payne CE, Wolfrum EJ (2015) Rapid analysis of composition and reactivity in cellulosic biomass feedstocks with near-infrared spectroscopy. Biotechnol Biofuels 8
48. Sills D, Gossett J (2012) Using FTIR to predict saccharification from enzyamtic hydrolysis of alkali-pretreated biomasses. Biotechnol Bioeng 109(2):353–362
49. Gollapalli L, Dale B, Rivers D (2002) Predicting digestibility of ammonia fiber explosion (AFEX)-treated rice straw. Appl Biochem Biotechnol 98:23–35
50. Zhu L, O'Dwyer JP, Chang VS, Granda CB, Holtzapple MT (2008) Structural features affecting biomass enzymatic digestibility. Biores Technol 99(9):3817–3828. doi:10.1016/j.biortech.2007.07.033
51. Zhang Y, Culhaoglu T, Pollet B, Melin C, Denoue D, Barrière Y, Sp B, Vr M (2011) Impact of lignin structure and cell wall reticulation on maize cell wall degradability. J Agric Food Chem 59(18): 10129–10135. doi:10.1021/jf2028279
52. O'Dwyer PJ, Zhu L, Granda CB, Chang VS, Holtzapple MT (2008) Neural network prediction of biomass digestibility based on structural features. Biotechnol Prog 24(2):283–292. doi:10.1021/bp070193v
53. Kelkar S, Li Z, Bovee J, Thelen KD, Kriegel RM, Saffron CM (2014) Pyrolysis of North-American grass species: effect of feedstock composition and taxonomy on pyrolysis products. Biomass Bioenerg 64:15

54. Shinners KJ, Boettcher GC, Muck RE, Weimer PJ (2010) Harvest and storage of two perennial grasses as biomass feedstocks. Trans ASABE 53(2):359–370

55. de Jong S (1993) SIMPLS: an alternative approach to partial least squares regression. Chemometr Intell Lab 18:251–263

56. Agblevor FA, Evans RJ, Johnson KD (1994) Molecular-beam mass-spectrometric analysis of lignocellulosic materials: I. Herbaceous biomass. J Anal Appl Pyrol 30(2):125–144

57. Evans RJ, Milne TA (1987) Molecular characterization of the pyrolysis of biomass. Energ Fuel 1(2):123–137. doi:10.1021/ef00002a001

58. Sykes R, Kodrzycki B, Tuskan G, Foutz K, Davis M (2008) Within tree variability of lignin composition in *Populus*. Wood Sci Technol 42(8):649–661. doi:10.1007/s00226-008-0199-0

59. Faix O (1991) Classification of lignins from different botanical origins by FT-IR spectroscopy. Holzforschung 45:21–27

60. Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst 58:109–130

61. Lupoi JS, Singh S, Davis M, Lee DJ, Shepherd M, Simmons BA, Henry RJ (2014) High-throughput prediction of eucalypt lignin syringyl/guaiacyl content using multivariate analysis: a comparison between mid-infrared, near-infrared, and Raman spectroscopies for model development. Biotechnol Biofuels 7:94

62. Del Río JC, Gutiérrez A, Rodríguez IM, Ibarra D, Martínez AT (2007) Composition of non-woody plant lignins and cinnamic acids by py-GC/MS, py/TMAH and FT-IR. J Anal Appl Pyrol 79:39–46

63. Hu Z, Sykes R, Davis MF, Charles Brummer E, Ragauskas AJ (2010) Chemical profiles of switchgrass. Biores Technol 101(9): 3253–3257. doi:10.1016/j.biortech.2009.12.033

64. Sauerbrei W, Royston P, Binder H (2007) Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Stat Med 26(30):5512–5528. doi:10.1002/sim.3148

65. Bozdogan H (1987) Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. Psychometrika 52(3):345–370. doi:10.1007/BF02294361

66. Burnham KP, Anderson DR (2004) Multimodel inference: understanding AIC and BIC in model selection. Soc Met Res 33(2):261–304. doi:10.1177/0049124104268644