Predictive Modeling of Lignin Content for the Screening of Suitable Poplar Genotypes Based on Fourier Transform–Raman Spectrometry

Wenli Gao, Ting Shu, Qiang Liu, Shengjie Ling, Ying Guan, Shengquan Liu,* and Liang Zhou*

ABSTRACT: The quick and non-invasive evaluation of lignin from biomass has been the focus of much attention. Several types of spectroscopies, for example, near-infrared (NIR) and Fourier transform–Raman (FT–Raman), have been successfully applied to build quantitative predictive lignin models based on chemometrics. However, due to the effect of sample moisture content and ambient humidity on its signals, NIR spectroscopy requires sophisticated pre-testing preparation. In addition, the current FT–Raman predictive models require large variations in the independent value inputs as restrictions in the corresponding mathematical algorithms prevent the effective biomass screening of suitable genotypes for lignin contents within a narrow range. In order to overcome the limitations associated with the current methods, in this paper, we employed Raman spectra excited using a 1064 nm laser, thus avoiding the impact of water and autofluorescence on NIR signals. The optimal baseline correction method, data type, mathematical algorithm, and internal reference were selected in order to build quantitative lignin models based on the data with limited variation. The resulting two predictive models, constructed through lasso and ridge regressions, respectively, proved to be effective in assessing the lignin content of poplar in large-scale breeding and genetic engineering programs.

1. INTRODUCTION

Lignocellulosic material, considered as one of the most promising substitutes for fossil energy, originates from the cell wall of natural plants and has extensive applications in biofuels and bio-based chemicals.1 To measure the chemical content of lignocellulose (i.e., cellulose, lignin, and hemicellulose), wet chemical methods (WCMs) such as the acetyl bromide, Klason, or acid-insoluble lignin and Van Soest methods are commonly applied.2 However, the majority of such methods are costly, time-consuming, and employ corrosive reagents (e.g., acetyl bromide and H₂SO₄).2 The assaying of a large number of lignocellulosic samples is typically required in genetic or breeding programs, resulting in low throughput and thereby hindering the working process of these WCMs. Furthermore, the minimum sample weight requirement of WCMs often exceeds the breeding program resources. Thus, quick and non-invasive methods have been established to replace the WCMs. For example, vibration spectroscopy, including near-infrared (NIR) and Raman spectroscopies, has been integrated with chemometrics to obtain the chemical content of lignocellulose.1,4–7

NIR accompanied by specific mathematical models can predict most chemical components in biomass. However, the absorbance signal at the NIR region is around 10–100 times weaker than that of the mid-infrared (MIR) region.9 Moreover, the moisture content of the samples and the humidity of the ambient atmosphere dramatically influence the NIR signal by forming “water peaks” at ~1450 and ~1890 cm⁻¹, corresponding to OH bonds.9,10 A relatively high sample moisture content can result in overwhelming peaks. Therefore, in terms of the chemical content assay of lignocellulose, the results obtained by NIR often need to be verified with the additional characterization techniques.5,7

Raman spectroscopy can complement NIR and has also been successfully applied to evaluate the lignin content.4,5,7 Unlike NIR, Raman spectroscopy is not sensitive to the sample moisture content as it is a water-insensitive technique.35 Suitable mathematical models have recently been employed to determine the prediction ratios of syringyl-to-guaiacyl (S/G) units in lignin as well as the classification of lignin.11,12 However, the auto-fluorescence originating from lignin or
other chemicals in lignocellulosic materials can lead to a broad signal in the resultant Raman spectrum.\textsuperscript{13} Furthermore, the conjugated lignin substructures can initiate the additional scattering contribution of Raman signals, resulting in a deviation from the prediction.\textsuperscript{3} To overcome these limitations, employment of NIR excitation using a 1064 nm laser has been proposed to significantly reduce the spectral background caused by the fluorescence emission.\textsuperscript{14} Besides, mathematical techniques including spectral data processing and anti-collinearity algorithms have proved to be effective for dealing with combination and the higher-order overtones of the fundamental signals in quantitative NIR models.\textsuperscript{1} Such mathematical techniques are expected to subdue the negative effect of the conjugated aromatic structures for quantitative Raman modeling.

Poplar has the largest plantation area in the world and has recently been identified as a key potential resource for biomass and biofuels.\textsuperscript{1} The high variation in lignin content among different poplar genotypes requires a quick and non-invasive method to detect lignin content in massive genetic or breeding programs.\textsuperscript{17} Thus, in the current paper, wood samples were obtained from several poplar clones in order to simultaneously determine their lignin contents via traditional WCMs and the Raman spectra of the samples. Different models were then organized according to the baseline correction strategy, data type, mathematical algorithm, and internal reference peak. These models were evaluated using the Pearson correlation coefficient ($R$) and root-mean-square error (RMSE) in order to determine the optimal model. This work evaluates the ability of Raman spectroscopy to predict the lignin content in genetic breeding selecting programs.

### 2. RESULTS AND DISCUSSION

#### 2.1. Lignin Content of Poplar Samples

Table 1 reports the WCM-derived lignin contents of the poplar samples ranging from 20.31 to 25.69 wt % and an average content of 22.79 wt %. This is in agreement with the previous reports of similar poplar genotype.\textsuperscript{1,18}

#### 2.2. Evaluating Parameters

$R$ and RMSE were selected to evaluate the quality of the mathematical models in the training and test sets. $R$ and $R_\text{p}$ generally represent the correlation between the measured and predicted values in a model, with values closer to 1 or $-1$, indicating a higher accuracy. Here, $R'$ and $R_{p}'$ were used to indicate the absolute value of $R$ and $R_p$, respectively. The value of $R'$ and $R_{p}'$ ranged from 0.13 to 0.95 and 0.01 to 0.90, respectively (Figures 1, 2).

| training set sample | lignin content (wt %) | training set sample | lignin content (wt %) | test set sample | lignin content (wt %) |
|---------------------|-----------------------|---------------------|-----------------------|----------------|-----------------------|
| S1                  | 20.31 ± 0.73\textsuperscript{b} | S9                  | 22.37 ± 0.16          | T1             | 21.49 ± 0.01          |
| S2                  | 20.72 ± 0.26          | S10                 | 22.47 ± 0.20          | T2             | 22.43 ± 1.27          |
| S3                  | 21.22 ± 1.68          | S11                 | 22.86 ± 0.74          | T3             | 22.86 ± 0.81          |
| S4                  | 21.28 ± 0.69          | S12                 | 22.97 ± 0.63          | T4             | 22.88 ± 0.34          |
| S5                  | 21.74 ± 0.53          | S13                 | 23.12 ± 0.58          | T5             | 22.97 ± 1.22          |
| S6                  | 22.04 ± 0.97          | S14                 | 23.90 ± 0.46          | T6             | 24.40 ± 0.07          |
| S7                  | 22.10 ± 0.22          | S15                 | 24.99 ± 0.78          | T7             | 24.64 ± 0.98          |
| S8                  | 22.16 ± 0.13          | S16                 | 25.69 ± 0.82          | T8             | 25.45 ± 0.78          |

\textsuperscript{a}Mean (% w/w). \textsuperscript{b}Standard deviation (% w/w).

![Figure 1](https://DOI.org/10.1021/acsomega.1c00400)
The evaluation results demonstrate the considerable variation of $R'$ and $R_p'$ across different baseline correction methods, data types, and mathematical algorithms. This indicates the ability of $R'$ and $R_p'$ to monitor the quality of the predictive models, as is the consensus in the literature.\(^{19-21}\) RMSE and root-mean-square error of prediction (RMSEP) denote the standard deviation of the residuals and are commonly used in regression analysis to verify experimental results. RMSE and RMSEP values close to 0 theoretically represent a perfect match between predicted and measured values. In the current paper, we calculated RMSE and RMSEP using the Raman peak intensities. The RMSE and RMSEP values ranged from 0.47 to 2.52 and 0.75 to 2.97, respectively (Figures 1, 2, S7, and S8). The values varied with the baseline correction methods, mathematical algorithms, and data types, indicating the effectiveness of RMSE and RMSEP to describe the quality of predictive models.\(^{22}\)

**2.3. Selection of the Baseline Correction Strategy.**

The baseline correction process is commonly applied to the Raman spectra to reduce the disturbing signals from baseline drift. In general, the number of baseline set points can alter each peak’s extracted value, irrespective of the intensity and area. The greater the number of baseline set points, the less impact the background will have on the extracted value. However, it will also reduce the signal originating from the molecular vibration of lignocellulosic materials.\(^{23}\) Therefore, a suitable baseline correcting method is required in order to balance these effects. We tested two baseline correction strategies to select the most optimal. Figure 3 presents the evaluation parameters corresponding to each of the strategies. The average and maximum values of $R'$ for the B2 strategy were higher than those for B1, while the opposite was observed for $R_p'$. Moreover, the average and minimum values of RMSE and RMSEP for B2 were lower than those for B1. This indicates a shorter distance between the predicted and measured values.
observed values for the models with six baseline set points. In addition, the relationship between the predicted and observed values were enhanced (weakened) in the training (test) set. The maximum value of $R' (0.87)$ for B2 was close to that of B1 (0.90). These results indicate B2 to be a more favorable baseline correction strategy for reliable and accurate predictive models of lignin content based on the Raman spectra compared to B1.

### 2.4. Selection of Data Type.

The predictive models used to describe the algebraic relationship between the lignin content and its specific scattering peaks were established based on the intensity (I type) and area (A type) of the specific Raman peaks. Figure 4 compares the models based on $R'$, $R_p'$, RMSE, and RMSEP.

The average values of $R'$ and $R_p'$ for the A type models exceeded those of the I type models (Figure 4); however the latter exhibited the highest $R'$ value (0.95). Moreover, the average values of RMSE and RMSEP were similar for both model types, with minimum values lower for the I type models compared to the A type models. The results of the model parameters suggest that the I type model is suitable for the analysis of lignin content in lignocellulosic materials. This may be attributed to the complex chemical components of the lignocellulosic materials, resulting in the overlap of Raman peaks. These overlapping Raman peaks cause uncertainty in the indeterminacy of the results after the deconvolution of the peak areas.

### 2.5. Selection of the Mathematical Algorithm.

PCR, PLSR, RR, and LR were applied to fit the models predicting the relationship between the lignin content and specific Raman peaks in the lignocellulosic material (Figure 5). The RR algorithm exhibited the highest average (0.74 and 0.31) and maximum (0.95 and 0.90) $R'$ and $R_p'$ values among the four models (Figure 5a). The maximum $R'$ value (0.94) of the LR algorithm was similar to that of the RR algorithm (0.95). However, the average $R'$ (0.62) value of the LR algorithm was much lower than that of the RR algorithm (0.74). The PCR and PLSR algorithms exhibited low maximum $R'$ values at 0.78 and 0.88, respectively. Figure 5a suggests that the RR and LR algorithm models are more suitable for the fitting of the lignin content in the lignocellulosic material compared to PCR and PLSR. The RMSE and RMSEP values of the four algorithm models in Figure 5b confirmed this result.

All four algorithms are considered effective for the analysis of multiple regression data with multi-collinearity. PCR and PLSR depress the collinearity by reducing the number of variables via the principal component extraction with the threshold value set as 90% in the current paper. In contrast, RR and LR adopt different techniques to solve the collinearity. A penalty term controlled by a regularization penalty ($\lambda$, see Section 2.6 in the Supporting Information) was arranged to substitute the role of principal component analysis used by PCR and PLSR in reducing the model bias without reducing the number of variables. The sum of the absolute penalty terms in LR varies until the Mallows’ $C_p$ value reaches a minimum. Ridge parameter $K$ plays the same role in controlling the penalty term in RR, iteratively reaching its optimal value. Therefore, at the same bias level, both RR and LR can potentially hold more reliable information from the variables compared to PCR and PLSR. Our results reveal higher maximum values of $R'$ and $R_p'$ for the RR and LR algorithm compared to the PCR and PLSR algorithms.

Ill-posed and ill-conditioned problems, which describe an overestimation of the output value of the function for small changes in the input, are often reported when building a predictive model. The problem is generally evaluated by...
the condition number, which is calculated from the independent data. In the current paper, we determined the condition number by extracting data from the Raman spectra (Figure S2). Values higher than 1 suggest a problem in the model.28,35 In our case, the predicted value of lignin content will be exaggerated by a minimal alteration in the Raman spectra. However, research has suggested the ability of the RR and LR algorithms to depress the negative effect caused by an ill-posed or ill-conditioned problem.30–32 This explains the lower maximum values of RMSE and RMSEP in RR and LR compared to PCR and PLSR.

2.6. Selection of the Internal Standard Peak in the Raman Spectra. The sample roughness, the auto-fluorescence of the lignocellulosic materials, and the input power of the laser source all affect the intensity of the Raman signals. Therefore, a standard internal peak was selected to normalize the Raman spectra prior to the quantitative or semi-quantitative analysis of the chemical components in the lignocellulosic materials.7,33,34 Raman bands used for the analysis are listed in Table 2. Figure 6 demonstrates that the

Table 2. Raman Peak Position and Band Assignments.1,4,16,35,36

| wavenumber/cm⁻¹ | component                        | assignment                                      |
|-----------------|----------------------------------|------------------------------------------------|
| 2939            | lignin, glucomannan, and cellulose| CH stretching in OCH₃ asymmetric                 |
| 2895            | cellulose                        | CH and CH₂ stretching                           |
| 1660            | lignin                           | ring conjugated C=C str. of coniferyl alcohol; C=O str. of coniferaldehyde |
| 1602            | lignin                           | ary1 ring stretching, symmetric.                |
| 1462            | lignin and cellulose             | lignin methoxy deformation, CH₂ scissoring, CH₂ scissoring |
| 1427            | lignin                           | OCH₁ deformation; CH₂ scissoring; guaiacyl ring vibration |
| 1378            | cellulose                        | HCC, HCO, and HOC bending                       |
| 1331            | cellulose                        | HCC and HCO bending, aliphatic OH stretch       |
| 1275            | lignin                           | Aryl-O of aryl OH and aryl O–CH₂ guaiacyl ring (with C=O group) mode |
| 1149            | cellulose                        | stretching and HCC and HCO bending, CC, CO ring breathing, asymmetric |
| 1122            | cellulose, xylan, and glucomanann | CC and CO stretching, COC, glycosidic; ring breathing, symmetric |
| 1095            | cellulose, xylan, and glucomanann | CC and CO stretching, COC, glycosidic; ring breathing, symmetric |
| 1043            | lignin                           | OC stretching; ring deformation; CH₂ wagging    |

data array corresponding to the internal standard with a peak intensity of 2895 cm⁻¹ (Iₛ/2895) has the highest maximum and average values for R’ and Rp’ among all groups. This implies the suitability of the 2895 cm⁻¹ peak intensity as an internal standard for standardizing the Raman spectra prior to quantitative or semi-quantitative analysis. This was also confirmed via the RMSE and RMSEP values, whereby Iₛ/2895 exhibits the lowest minimum and average values among all combinations.

Previous work has demonstrated the ability of the Raman peak centered at 1096 cm⁻¹, which is assigned to the vibration of CC and CO stretching in carbohydrates, to normalize the scattering peak of lignin at 1600 cm⁻¹. Moreover, Agarwal (2011) established a linear model based on the relationship between the standardized relative intensity (intensity of peak 1600 cm⁻¹/intensity of peak 1096 cm⁻¹) and the lignin content determined by the WCMs. However, the relative intensity needs to be further “corrected” by normalizing with the carbohydrate content in order to reduce the impact of scattering contribution at 1096 cm⁻¹ assigned to CC and CO stretching of the carbohydrate. The peak located at approximately 2895 cm⁻¹ in the Raman biomass spectra is generally assigned to the vibration of C–H in CH and CH₂ in cellulose. This peak was observed to disappear for the Raman spectra of spruce milled wood lignin (MWL). Therefore, the peak of 2895 cm⁻¹ is irrelevant to the lignin content and can be employed to standardize the lignin-related band intensity/area. Note that a “self-absorption” in the C–H stretching vibration of the Raman spectra will occur when water is added to the biomass sample. This may have a negative impact when the peak of the C–H stretching is used to quantify the cellulose content. However, this negative effect is minimal for predictive models of lignin content as the numerator data array includes the peak assigned to the C–H stretching of −OCH₁ in lignin. Research has speculated that the “self-absorption” effect will be depressed by the 2939 cm⁻¹ peaks with the help of anti-collinearity algorithms.26 These observations suggest the 2895 cm⁻¹ peak intensity to be the best internal standard option for the predictive model.

2.7. Description of the Optimal Predictive Models. Following the optimization of the baseline correction strategy, data type, mathematical algorithm, and internal standard peaks, two predictive models were established using the RR and LR mathematical algorithms. The R’, Rp’, RMSE, and RMSEP of these models indicate that both models are suitable for the prediction of lignin content in lignocellulosic materials via their Raman spectra (Table 3). Figure 7 depicts the lignin content determined by the WCMs and Raman spectra predictive models. The data points of the RR and LR models are distributed close to the line of equality in Figure 7a,b. This is attributed to the similar techniques applied by both models to reduce the multi-collinearity between the actual and predicted values of the lignin content.

The evaluation parameters determined for the training and test sets of the two models agree with those reported for the lignin content of wood samples.1 However, our observed parameters are inferior to those of models predicting the lignin content of kraft pulps or pre-treated biomass.4,35 Despite this, the aforementioned models do not meet the requirements of genetic breeding scientists as they take the independent value (i.e., namely, lignin content) as the input, with larger variations than that of biomass across genotypes. Quantitative models are generally considered as more reliable for predicted values that exhibit a variation equivalent to that of the independent value inputs. This suggests that the proposed model is more suitable than those in the literature for the prediction of lignin content in lignocellulosic materials from different genotypes.

With the inclusion of the optimal coefficients and constants, the Lasso regression model is described as follows:

\[
C_{\text{lignin}} = -30.7401 \times I_{1043/2895} - 21.0667 \times I_{1275/2895} - 63.4372 \times I_{1427/2895} + 6.6674 \times I_{1462/2895} + 2.5937 \times I_{1602/2895} + 30.7602 \times I_{1660/2895} + 11.6868 \times I_{2895/2895} + 42.5718 \] (1)
and the Ridge regression model is determined as

\[ C_{\text{lignin}} = -31.1652 \times I_{043/2895} - 22.2815 \times I_{1275/2895} - 65.7096 \times I_{1427/2895} + 5.0224 \times I_{1462/2895} + 3.0286 \times I_{1602/2895} + 35.2813 \times I_{1660/2895} + 10.6813 \times I_{2939/2895} + 42.8357 \]

where \( C_{\text{lignin}} \) is the lignin content and \( I_{043/2895} \) is the intensity ratio between the 1043 and 2895 cm\(^{-1} \) peaks. The remaining parameters follow the same definitions.

Our results demonstrate the ability of the two models to predict the lignin content for the screening of poplar genotypes based on their Raman spectra, particularly for the optimized building process detailed in the previous sections.

### 2.8. Leave-One-Out Cross-Validation for Evaluating the Optimal Predictive Models

For evaluating the stability and robustness of the quantitative models, leave-one-out cross-validation (LOOCV) was performed to estimate the performance of the optimal predictive models described in Section 2.7. In each iteration of LOOCV, one data group was set aside and the remaining data groups were used for training. The statistical results of LOOCV are listed in Table 3.

The ridge regression model and the lasso regression model for lignin content showed a good performance in cross-validation, the \( R^2 \) value is 0.95 and the root-mean-square error of validation (RMSEV) values are 0.41 and 0.48, respectively.

**Table 3. Statistical Results for the Training, Cross-Validation, and Test Sets of the RR and LR Models**

| data group       | algorithm | \( R^2 \) | RMSE  | \( R^2 \) | RMSEV | \( R^2 \) | RMSEP |
|------------------|-----------|-----------|-------|-----------|-------|-----------|-------|
| B2-1lp/2895      | RR        | 0.95      | 0.47  | 0.95      | 0.41  | 0.86      | 0.88  |
|                  | LR        | 0.94      | 0.52  | 0.95      | 0.48  | 0.87      | 0.75  |

**Figure 6.** Comparative Lasso and ridge regression results of 12 models based on different internal standards using the B2-1 data groups. (a) \( R^2 \) of the models; (b) \( R^p \) of the models; (c) RMSE of the models; and (d) RMSEP of the models.

**Figure 7.** Lignin content determined from the WCMs and Raman spectra based on the B2-1lp/2895 dataset (a) in the RR model and (b) in the LR model.
The values calculated from the validation datasets are nearly equivalent to that calculated from the training sets. Therefore, the results of LOOCV demonstrated that two optimal models are stable and robust. Then, the reliable predictive performance of lignin quantitative models was further verified.

3. CONCLUSIONS

In the current paper, we established an effective predictive model of lignin content for the screening of poplar genotypes based on their Raman spectra. We employed the intensity of deconvoluted Raman peaks to establish mathematical prediction models for the lignin content of lignocellulosic materials, with the Raman peak centered at 2895 cm⁻¹ considered as the standard internal peak. We reveal the capability of the RR and LR models to fit the lignin content in the lignocellulosic material via the collected Raman spectra. This highlights the potential of these models in assessing the lignin content in large-scale breeding and genetic engineering programs.

4. MATERIALS AND METHODS

4.1. Materials. Eight poplar clones, namely, Populus deltoides Cl. 55/65, Populus euramericana cv. “Zhonglin46”, P. euramericana cv. “Guierento”, P. deltoides Cl. “2KEN8”, Populus nigra Cl. “N179”, P. deltoides Cl. “Danhong”, P. euramericana Cl. “Sangia”, and P. deltoides Cl. “Nanyang”, were planted at the Jiaozuo Forest Farm in Jiaozuo City, Henan Province, China (113°13′E, 35°14′N), in 2010. Three trees of each clone were harvested in November 2018, and the discs at breast height were cut and processed into particles in order to study the chemical components. A total of 24 different wood samples were obtained. All samples were stored at room temperature for more than 1 month at a moisture content of approximately 9.5%. The wood particles were ground into a fine powder using a pulverizer. The 40–60 mesh wood powder was sieved out to measure the lignin content via the WCMs, while the 100–120 mesh wood powder was reserved for the collection of Raman spectra.

4.2. Methods. 4.2.1. Lignin Content Measurements. Lignin is typically divided into two parts during the compositional analysis of lignocellulosic materials with WCMs, namely, Klasson lignin and acid-soluble lignin. The Klason lignin measurements in this paper employed the official TAPPI (Technical Association of the Pulp and Paper Industry) T222 om-06 test method, while acid-soluble lignin measurements were based on TAPPI UM-250. The lignin content was then taken as the sum of Klasson and acid-soluble lignins. The average value of three technical replicates was registered as the lignin content of each sample. Therefore, the dataset consisted of 24 lignin content values.

4.2.2. Collection of Raman Spectra. A Fourier transform—Raman spectrometer (Bruker, VERTEX 70-RAMII) was used to collect Raman spectra of the poplar samples. The spectrometer was equipped with a 1064 nm, 500 mW Nd:YAG diode laser. During the test, the ambient relative humidity and temperature were 60 ± 10% and 20 ± 2 °C, respectively. Approximately 2 mg of wood powder was pressed into a sample cell and the collection was repeated three times for each sample. The average spectrum was regarded as the sample spectrum. During measurements, the fluorescence was excited at 1064 nm, the laser defocus mode was used, and the number of scans was increased to obtain a good signal-to-noise ratio. Spectra were collected with 64 scans and a spectral resolution of 4 cm⁻¹ across 200–3600 cm⁻¹.

4.2.3. Processing of Raman Spectra. The Raman spectra of the poplar samples exhibit an excellent resolution (Figure S3). Table 4 details the assignments of the specific peaks. Peak-fit 4.12 was employed to process the spectral data, including smoothing, background correction, and band peak-fitting. First, Savitzky–Golay (level 1.0%) smoothing was applied prior to the processing of the spectra. Second, the second derivation zero algorithm was used for baseline correction by using peakfit 4.12 software. The sets of baseline points were confirmed according to the second derivative spectra. The baseline correction was then executed via two different strategies by changing the set points: (i) Four set points located at 940, 1750, 2750, and 3050 cm⁻¹ divided the spectra into five sections (B1) and (ii) six set points located at 940, 1210, 1540, 1750, 2750, and 3050 cm⁻¹ divided the whole spectra into seven sections (B2). Regions 940–1750 and 2750–3050 cm⁻¹ include the signals closely related to the molecular vibration in lignocellulosic materials and also exhibit a high resolution in all spectra. Thus, these two regions were selected in strategy B1 for the baseline correction and further peak-fitting (Figures S4 and S5). In addition, regions 940–1210, 1210–1540, 1540–1750, and 2750–3050 cm⁻¹ were used for further processing in strategy B2 (Figure 8). Third, the peak-fitting of overlapping bands was performed for each
peak based on Gauss and Lorentz peak-fitting. In order to ensure the effectiveness of the separated peaks, they were determined according to the second-derivative spectra, while the peak positions were fixed during the peak-fitting process.47

Figure S6 presents the results of the Raman spectra for lignocellulosic materials.

4.2.4. Selection and Normalization of Input Data from the Raman Spectra. The intensity (I) and area (A) of each peak in the Raman spectra were determined following the peak-fitting process. As both of these parameters are suitable inputs for predictive models,4,34 they were compared in order to determine the most suitable based on the model evaluating parameters. A selection of internal standards was then executed to normalize these inputting values.4 A data array of the peak digital values related to lignin (e.g., 1043, 1275, 1427, 1462, 1602, 1660, and 2939 cm$^{-1}$), denoted as $l_p$, was assigned to the numerator.2,3 Digital values of each prominent peak in the Raman spectra were allocated as the denominators of the intensity ($I$) and area ($A$) of each peak based on Gauss and Lorentz peak-fitting. In order to ensure the effectiveness of the separated peaks, they were determined according to the second-derivative spectra, while the peak positions were fixed during the peak-fitting process.47

In the current paper, we employed all four algorithms to build the mathematical model based on the original lignin content data and to estimate the lignin content from the Raman spectra. The threshold value for the contribution of all principle components was set at 90% for PCR and PLSR. Based on the model grouping, 16 samples were randomly selected as the training set (S1−S16), with the remaining 8 samples as the test set (T1−T8) (Table 3). Furthermore, LOOCV was used to validate the performance of specified models. In addition, the validation set were split from the training set according to the LOOCV rules.53

4.2.6. Model Evaluation. The Pearson correlation coefficient ($R$) and RMSE were used to evaluate the model qualities of the training sets, while the validation set and the test set model qualities were evaluated via the Pearson correlation coefficient of validation ($R_v$), the RMSEP, the Pearson correlation coefficient of prediction ($R_p$), and the RMSEP.

The Pearson correlation coefficient ($R$), also known as the Pearson R statistical test, measures the strength of the relationship between the different variables. $R$ is determined as follows:

$$ R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} $$

where $X$ and $Y$ are two different variables and $n$ is the sample size. RMSE measures the error between two datasets and is an indicator of how far apart the predicted values are from the observed values.54−56 RMSE is described as follows:

$$ \text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(P_i - M_i)^2}{n}} $$

where $P_i$ is the predicted value for the $i$th observation in the dataset, $M_i$ is the measured value for the $i$th observation in the dataset, and $n$ is the sample size.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.1c00400.

Definition and cost function of variance inflation factor (VIF), ridge regression, lasso regression, Mallows’ $C_p$, and condition number; value of VIF; condition number; Raman spectra in different processing; and evaluation of models based on the B1 baseline correction strategy (PDF)

## AUTHOR INFORMATION

### Corresponding Authors

Shengquan Liu – School of Forestry and Landscape Architecture, Anhui Agriculture University, Hefei 230036, Anhui, China; Key Lab of State Forest and Grassland Administration on Wood Quality Improvement & High Efficient Utilization, Hefei 230036, Anhui, China; Email: liusq@ahau.edu.cn

Liang Zhou – School of Forestry and Landscape Architecture, Anhui Agriculture University, Hefei 230036, Anhui, China; Key Lab of State Forest and Grassland Administration on Wood Quality Improvement & High Efficient Utilization, Hefei 230036, Anhui, China; Email: mcyjs1@ahau.edu.cn

### Authors

Wenli Gao – School of Forestry and Landscape Architecture, Anhui Agriculture University, Hefei 230036, Anhui, China; Key Lab of State Forest and Grassland Administration on Wood Quality Improvement & High Efficient Utilization, Hefei 230036, Anhui, China

ACS Omega 2021, 6, 8578−8587
Ting Shu — School of Physical Science and Technology, Shanghai Tech University, Shanghai 201210, China
Qiang Liu — School of Physical Science and Technology, Shanghai Tech University, Shanghai 201210, China
Shengjie Ling — School of Physical Science and Technology, Shanghai Tech University, Shanghai 201210, China

Ying Guan — School of Forestry and Landscape Architecture, Anhui Agriculture University, Hefei 230036, Anhui, China

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.1c00400

Author Contributions
WL.G. contributed to the experimental data analysis, writing, and editing; T.S., Q.L., and Y.G. performed the experimental data analysis; SJ.L. involved in article revision; SQ.L. supervised the experiments; and L.Z. involved in article revision and supervision.

Notes
The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program (2017YFD0600201), the National Natural Science Foundation of China (no. 31770596), the Scientific Research Project of Anhui Educational Committee (grant no. KJ2016A83), and the International Science and Technology Cooperation Program of Anhui Province (1704a1002226).

REFERENCES

(1) Gebreselasie, M. N.; Ader, K.; Boizot, N.; Millier, F.; Charpentier, J-P.; Alves, A.; Simões, R.; Rodrigues, J. C.; Bodineau, G.; Fabbrini, F.; Sabati, M.; Bastien, C.; Segura, V. Near-infrared spectroscopy enables the genetic analysis of chemical properties in a large set of wood samples from populus nigra (l.) natural populations. Ind. Crops Prod. 2017, 107, 159–171.
(2) Lupoi, J. S.; Singh, S.; Parthasarathi, R.; Simmons, B. A.; Henry, R. J. Recent innovations in analytical methods for the qualitative and quantitative assessment of lignin. Renewable Sustainable Energy Rev. 2015, 49, 871–906.
(3) Agarwal, U. P. Analysis of cellulose and lignocellulose materials by raman spectroscopy: a review of the current status. Molecules 2019, 24, 1659.
(4) Agarwal, U. P. Lignin quantitation by FT-Raman spectroscopy. Proceedings 16th International Symposium on Wood, Fiber and Pulping Chemistry, 2011; pp 170–174.
(5) Agarwal, U. P.; Ralph, S. A.; Padmakanth, D.; Liu, S.; Foster, C. E. Estimation of syringyl units in wood lignins by FT-Raman spectroscopy. J. Agric. Food Chem. 2019, 67, 4367–4374.
(6) Pereira, A. A.; Martins, G. F.; Antunes, P. A.; Conrado, R.; Pasquini, D.; Job, A. E.; Curvelo, A. A. S.; Ferreira, M.; Rial, A.; Constantino, C. J. L. Lignin from sugar cane bagasse: extraction, fabrication of nanostructured films, and application. Langmuir 2007, 23, 6652–6659.
(7) Agarwal, U. P.; Ralph, S. A. Quantitation of lignin in grasses by Near-IR FT-Raman Spectroscopy. Abstracts of Papers of the American Chemical Society; American Chemical Society, 2012; Vol. 243.
(8) Porep, J. U.; Kammerer, D. R.; Carle, R. On-line application of near infrared (NIR) spectroscopy in food production. Trends Food Sci. Technol. 2015, 46, 211–230.
(9) Wenz, J. J. Examining water in model membranes by near infrared spectroscopy and multivariate analysis. Biochim. Biophys. Acta 2018, 1860, 673–682.
(10) Rambo, M. K. D.; Amorim, E. P.; Ferreira, M. M. C. Potential of visible-near infrared spectroscopy combined with chemometrics for analysis of some constituents of coffee and banana residues. Anal. Chim. Acta 2013, 775, 41–49.
(11) Ding, D.; Zhou, X.; You, T.; Zhang, X.; Zhang, X.; Xu, F. Exploring the mechanism of high degree of delignification inhibits cellulose conversion efficiency. Carbohydr. Polym. 2018, 181, 931–938.
(12) Sun, L.; Varanasi, P.; Yang, F.; Loqué, D.; Simmons, B. A.; Singh, S. Rapid determination of syringyl/guaiacyl ratios using FT-Raman spectroscopy. Biotechnol. Bioeng. 2012, 109, 647–656.
(13) Zhang, X.; Ji, Z.; Zhou, X.; Ma, J.-F.; Hu, Y.-H.; Xu, F. Method for automatically identifying spectra of different wood cell wall layers in Raman imaging data set. Anal. Chem. 2015, 87, 1344–1350.
(14) Lupoi, J. S.; Gersing, E.; Davis, M. F. Evaluating lignocellulosic biomass, its derivatives, and downstream products with raman spectroscopy. Front. Bioeng. Biotechnol. 2015, 3, 50.
(15) Meyer, M. W.; Lupoi, J. S.; Smith, E. A. 1064nm dispersive multichannel Raman spectroscopy for the analysis of plant lignin. Anal. Chem. 2011, 706, 164–170.
(16) Agarwal, U. P.; Ralph, S. A. FT-Raman spectroscopy of wood: identifying contributions of lignin and carbohydrate polymers in the spectrum of black spruce (picea mariana). Appl. Spectrosc. 1997, S1, 1648–1655.
(17) Sannigrahi, P.; Raguaskas, A. J.; Tuskan, G. A. Poplar as a feedstock for biofuels: a review of compositional characteristics. Biofuels, Bioprod. Biorefin. 2010, 4, 209–226.
(18) Sun, Q.; Foston, M.; Meng, X.; Sawada, D.; Pingali, S. V.; O'Neill, H. M.; Li, H.; Wyman, C. E.; Langan, P.; Raguaskas, A. J.; Kumar, R. Effect of lignin content on changes occurring in poplar cellulose ultrastructure during dilute acid pretreatment. Biotechnol. Biofuels 2014, 7, 150.
(19) Hauke, J.; Kossowski, T. Comparison of Values of Pearson’s and Spearman’s Correlation Coefficients on the Same Sets of Data. Quaest. Geogr. 2011, 30, 87–93.
(20) Adler, J.; Parmryd, I. Quantifying colocalization by correlation: the pearson correlation coefficient is superior to the mander’s overlap coefficient. Cytometry, Part A 2010, 77, 733–742.
(21) Boslaugh, S.; Watters, P. A. Statistics in a nutshell; O'Reilly Media, Inc., 2008.
(22) Agarwal, U. P.; Reiner, R. S.; Ralph, S. A. Cellulose I crystallinity determination using ft-raman spectroscopy: univariate and multivariate methods. Cellulose 2010, 17, 721–733.
(23) Vickers, T. J.; Wambles, R. E.; Mann, C. K. Curve fitting and linearity: data processing in raman spectroscopy. Appl. Spectrosc. 2001, 55, 389–393.
(24) Quinn, G. P.; Keough, M. J. Experimental design and data analysis for biologists; Cambridge University Press, 2002; pp 127–130.
(25) Haaland, D. M.; Thomas, E. V. Partial least-squares methods for spectral analyses. I. relation to other quantitative calibration methods and the extraction of qualitative information. Anal. Chem. 1988, 60, 1193–1202.
(26) Januaviani, T. M. A.; Gusriani, N.; Joebaedi, K.; Supian, S.; Subiyanto, S. The best model of lasso with the lars (least angle regression) algorithm using mallow’s cp. World Sci. News 2019, 116, 245–252.
(27) Lee, T.-S.; Callaghan, K. J.; Chen, J. Revisiting the collinear data problem: selecting the optimum k in ridge regression. Commun. Stat. Theor. Methods 1985, 14, 1589–1604.
(28) Callaghan, K. J.; Chen, J. Revisiting the collinear data problem: an assessment of estimator “ill-conditioning” in linear regression. Practical Assess. Res. Eval. 2008, 13, 5.
(29) Mu, X.; Cheng, H.; Xu, Q.; Liu, G. Stably numerical solving method and the extraction of qualitative information. Anal. Chem. 1980, 60, 1355–1359.
(30) Khalaf, G.; Iguernane, M. Ridge regression and ill-conditioning. Comput. Appl. Math. 2018, 37, 1941–1958.
(32) Matthiaou, M.; McKay, M.; Smith, P.; Nossek, J. On the condition number distribution of complex wishart matrices. *IEEE Trans. Commun.* 2010, 58, 1705−1717.
(33) Agarwal, U. P.; Kawai, N. “Self-Absorption” Phenomenon in Near-Infrared Fourier Transform Raman Spectroscopy of Cellulosic and Lignocellulosic Materials. *Appl. Spectrosc.* 2005, 59, 385−388.
(34) Agarwal, U. P.; Weinstock, I. A.; Atalla, R. H. FT-Raman spectroscopy for direct measurement of lignin concentrations in Kraft pulp. *Tappi J.* 2003, 2, 22−26.
(35) Gierlinger, N.; Schwanninger, M. Chemical Imaging of Poplar Wood Cell Walls by Confocal Raman Microscopy. *Plant Physiol.* 2006, 140, 1246−1254.
(36) Wiley, J. H.; Atalla, R. H. Band assignments in the raman spectra of cellulos. *Carbohydr. Res.* 1987, 160, 113−129.
(37) Agarwal, U. P.; McSweeny, J. D.; Ralph, S. A. FT-Raman investigation of milled-wood lignins: softwood, hardwood, and chemically modified black spruce lignins. *J. Wood Chem. Technol.* 2011, 31, 324−344.
(38) Dence, C. W. *The Determination of Lignin*; Springer, 1992; pp 33−61.
(39) Brumano, G. C.; Colodette, J. L.; Fernandes, S. A.; Barbosa, B. M.; Borges Gomes, F. J. Investigation of eucalypt and pine wood acid-soluble lignin by Py-GC-MS. *Holzforschung* 2020, 74, 149−155.
(40) Studer, M. H.; DeMartini, J. J.; Davis, M. F.; Sykes, R. W.; Davison, B.; Keller, M.; Tuskan, G. A.; Wyman, C. E. Lignin content in natural populus variants affects sugar release. *Proc. Natl. Acad. Sci. U.S.A.* 2011, 108, 6300−6305.
(41) Gomes, D. I.; Detmann, E.; Valadares Filho, S. d. C.; Fukushima, R. S.; de Souza, M. A.; Valente, T. N. P.; Paulino, M. F.; de Queiroz, A. C. Evaluation of lignin contents in tropical forages using different analytical methods and their correlations with degradation of insoluble fiber. *Anim. Feed Sci. Technol.* 2011, 168, 206−222.
(42) Campbell, M. M.; Sederoff, R. R. Variation in lignin content and composition (mechanisms of control and implications for the genetic improvement of plants). *Tappi.* 1996, 110, 3−13.
(43) Tappi. *Lignin in wood and pulp.* T222 Om-02, 2011; pp 1−7.
(44) Yang, S.; Akkus, O. Fluorescence background problem in Raman spectroscopy: is 1064 nm excitation an improvement of 785 nm?, Wasatch Photonics, 2013.
(45) Cao, S.; Wang, Y.; Xing, L.; Zhang, W.; Zhou, G. structure and physical properties of gelatin from bovine bone collagen influenced by acid pretreatment and pepsin. *Food Bioprod. Process.* 2020, 121, 213−223.
(46) Hu, H.; Bai, J.; Xia, G.; Zhang, W.; Ma, Y. Improved Baseline Correction Method Based on Polynomial Fitting for Raman Spectroscopy. *Photonics Sens.* 2018, 8, 332−340.
(47) Ling, S.; Dinjaski, N.; Ebrahimi, D.; Wong, J. Y.; Kaplan, D. L.; Buehler, M. J. Conformation transitions of recombinant spidroins via integration of time-resolved FTIR spectroscopy and molecular dynamic simulation. *ACS Biomater. Sci. Eng.* 2016, 2, 1298−1308.
(48) Gómez, R. S.; Pérez, J. G.; Martin, M. D. M. L.; García, C. G. Collinearity diagnostic applied in ridge estimation through the variance inflation factor. *J. Appl. Stat.* 2016, 43, 1831−1849.
(49) García, C.; Gómez, R. S.; García, C. B. Choice of the ridge factor from the correlation matrix determinant. *J. Stat. Comput. Simulat.* 2019, 89, 211−231.
(50) Hastie, T.; Tibshirani, R.; Wainwright, M. *Statistical learning with sparsity: the lasso and generalizations*; CRC Press, 2015; pp 92−95.
(51) Hoerl, A. E.; Kennard, R. W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 2000, 42, 80−86.
(52) Meier, L.; Van De Geer, S.; Bühlmann, P. The group lasso for logistic regression. *J. R. Stat. Soc* 2008, 70, 53−71.
(53) Wong, T.-T. Performance Evaluation of Classification Algorithms by K-Fold and Leave-One-out Cross Validation. *Pattern Recogn.* 2015, 48, 2839−2846.