Determination of aliphatic and phenolic hydroxyl groups in lignin by chemometric analysis of FTIR spectroscopic data

MOHAMMAD NASHIR UDDIN, M. NUR ALAM LIKHON, M. MOSTAFIZUR RAHMAN, YANGCAN JIN, BO JIANG, M. TUSHAR UDDIN, AND M. SARWAR JAHAN

ABSTRACT: Lignin's potential as a source of sustainable aromatic compounds is significant, but its utilization is currently limited by its chemical reactivity. Chemical reactivity of lignin depends on the present functional groups, such as hydroxyl, methoxy, and carbonyl groups. Therefore, in this study, multivariate analysis-based chemometric models have been developed for rapid determination of aliphatic hydroxyl (Alp-OH) and phenolic hydroxyl (Ph-OH) groups in lignin samples.

Two chemometric models, principal component regression (PCR) and partial least squares regression (PLSR), were established with Fourier transform infrared spectroscopy (FTIR) spectral data of 28 lignin samples. Both the models were developed based on raw and pretreated spectroscopic data with Savitky-Golay (S-G) filtering and standard normal variate (SNV) and multiplicative scatter correction (MSC). The predictive performance of the PLSR model is better for predicting Alp-OH (R^2 = 0.94%), syringyl-OH (R^2 = 0.96%), guaiacyl-OH (R^2 = 0.98%), *p*-hydroxyphenyl (R^2 = 0.93%), and total Ph-OH groups (R^2 = 0.97%) with the data pretreated by MSC. Finally, the predicted results of these parameters for three new samples for the developed models are found to be very close to the estimated values by NMR.

Application: Rapid determination of aliphatic and phenolic hydroxyl groups in lignin was achieved using FTIR spectroscopy and chemometric modeling (PCR and PLSR). The PLSR model, particularly when applied to MSC-pretreated data, provided accurate predictions for various hydroxyl types (R^2 > 0.93), closely aligning with NMR-derived values.

Lignin, sourced from lignocellulosic biomass, is the second most abundant natural biopolymer on Earth, making it a significant forest-derived material. Chemically, lignocellulose is a composite of three biopolymers — cellulose, hemicelluloses, and lignin — along with small amounts of extractives and minerals [1]. Lignin itself is a complex phenolic polymer, created through the polymerization of three monolignols: coniferyl alcohol, sinapyl alcohol, and *p*-coumaryl alcohol [2].

Lignin is isolated as a byproduct from various industrial processes, primarily in the pulp and paper industry and biorefineries. Presently, conventional pulp mills are the major suppliers of lignins, which is called technical lignin. Technical lignin is derived from processes such as kraft pulping, soda pulping, sulfite pulping, organosolv processes, and steam explosion [3]. Each method yields lignin with different properties [4]. The structure varies depending on the plant source and extraction method. The chemical reactivity of technical lignin depends on the functional groups present, such as hydroxyl, methoxy, and carbonyl groups [5]. It can be a source of aromatic compounds for producing biofuels, phenols, and other chemicals. Technical lignin is a valuable byproduct with significant potential for sustainable applications, though its utilization requires further innovation to overcome technical and economic barriers.

A more detailed characterization of technical lignin and its functional properties is essential to convert it into a high-value product for the industry [6]. The phenolic hydroxyl (Ph-OH) groups within lignin's aromatic rings are highly reactive due to lignin's formation via oxidative polymerization of phenolic radicals. These groups play a significant role in determining lignin's chemical reactivity and its potential applications [7-9]. As such, accurately assessing the phenolic -OH group content in technical lignin is crucial. The Ph-OH groups are key determinants of technical lignin's chemical reactivity, as well as its solubility and antioxidant behavior [10,11].

To determine Ph-OH group content in lignin, a spectrum of analytical techniques is available [12,13]. Ranging in complexity, sensitivity, and specificity, these methods allow researchers to select the most suitable option for their objectives. Techniques such as potentiometric or conductometric titration, aminolysis, ultraviolet visible (UV–vis) spectrophotometry, Fourier transform infrared (FTIR) spectroscopy, nuclear magnetic resonance (NMR), high-performance liquid chromatography (HPLC), and gas chromatography (GC) are frequently employed. The UV–vis

spectroscopy is recognized as a straightforward, fast, and economical analytical technique. As a result, it is extensively employed for both qualitative and quantitative analysis of lignin [14-17].

Argyropoulos [13] quantified various hydroxyl groups in lignin, including aliphatic, phenolic, and carboxyl groups using ^{31}P NMR spectroscopy. These functional groups represent key lignin functionalities, making their measurement crucial for understanding lignin's structural properties. The measured values of total phenolic hydroxyl groups and total hydroxyl groups showed good agreement with results from other laboratories using different analytical methods. Since then, most of the investigators used this method for the quantification of hydroxyl groups in lignin [18,19].

Recently, chemometric modeling has been widely investigated in forest products, especially in quantification of lignocellulosic components, pulp properties, and lignin properties, etc. [20]. Our group has developed robust chemometric models based on multivariate analysis of spectroscopic data to analyze lignocellulose and pulp. These models effectively quantify key chemical components in agricultural residues ($R^2 \approx 0.99$, FTIR, artificial neural networks [ANN]) and predict pulp properties from jute (FTIR, PLSR). Specific models achieved R^2 values of 0.83% for lignin (PLSR, Savitzky-Golay) and 0.94% for α-cellulose (PLSR, mean normalization) [21-23].

Several chemometric models have been established for determining kraft pulps, utilizing various spectroscopic techniques, such as UV resonance Raman spectroscopy and FTIR [24-27]. These models have been successfully applied to both unbleached and bleached chemical pulps from diverse wood species, including *Eucalyptus globulus*, *Pinus radiata*, and *Picea abies*. Presently, hexenuronic acid (HexA) content in pulp is determined by UV spectrophotometry [24]. Recently, Uddin et al. [24] developed a rapid FTIR-based method for quantifying HexA in nonwood pulp, achieving a high prediction accuracy (R^2 = 94.24%) using a PLSR model with Savitzky-Golay filtering, derivatives, and leverage correction. The model's validity was further confirmed by its excellent performance in predicting HexA in unknown nonwood samples (R^2 = 99.30%). The guaiacyl and syringyl proportion in lignin was also quantified by Uddin et al. [25]. The PLSR achieved the highest predictive accuracy (R^2 = 99.90%) using FTIR data that was preprocessed with Savitzky-Golay (S-G) filtering, including its derivatives, and adjusted with leverage correction.

However, to date, no research has reported the quantification of Ph-OH groups in lignin using multivariate analysis and chemometric modeling. Therefore, this study aims to develop a feasible, rapid, and environmentally friendly method for quantifying both phenolic hydroxyl (Ph-OH) groups and aliphatic hydroxyl (Alp-OH) groups in lignin, employing chemometric modeling with FTIR spectroscopic data.

MATERIALS AND METHODS

Based on the availability of lignocellulosic materials in Bangladesh, fifteen raw materials were collected from the Northern part of the country. These were selected. They are nonwoods, namely, jute stick (Corchorus olitorius), sugarcane bagasse (*Saccharum officinarum*), mulberry plant (*Morus alba*), banana tree (*Musa acuminata*), corn stalk (*Dracaena fragrans*), chia (*Salvia hispanica*), rice straw (*Oryza sativa*), kash (*Saccharum spontaneum*), zara (*Pennisetum purpureum*), bamboo (*Bambusa vulgaris*), coconut husk (*Cocos nucifera*), jute fiber (*Corchorus olitorius*), hogla (*Typha elephantina*), bringle (*Solanum melongena*), and dhaincha (*Sesbania bispinosa*). From these lignocellulosic materials, 28 lignin samples were prepared.

The collected samples were cleaned, air dried, ground in a Wiley mill, and screened in 40/60 mesh. The sample passed through 40 mesh, and the remainder on 60 mesh was collected and stored for subsequent experiments. The ground nonwood samples were extracted with ethanol-toluene (2:1) in Soxhlet [26].

Pulping

The pulping of these lignocellulosic materials was performed using the soda process under the conditions specified in **Table I** to isolate technical lignin. Pulping was conducted in small bombs of 100-mL capacity immersed in an electrically heated oil bath. Bombs, containing raw materials and cooking chemicals, were placed in the bath at room temperature. The temperature was then raised to either 150°C or 170°C, taking 40 min and 60 min, respectively. After the desired cooking time, the bombs were immediately cooled in flowing tap water. Following cooling, the bombs were opened, and the cooked mass was filtered in a Büchner funnel to separate the liquor from the pulp.

Lignin isolation from lignocelluloses and soda liquor

The most common lignin extraction methods from the raw material are milled wood lignin (MWL), acidic dioxane, and cellulolytic methods. Among these methods, the acidic dioxane method is the simplest and provides a higher yield. Therefore, the acidic dioxane method was selected for isolating lignin from the raw material. The extract-free samples were subjected to reflux with an acidic dioxane solution containing 0.2 N hydrochloric acid (HCl), and the dioxane-to-water ratio was 9:1. A dioxane-to-sample ratio of 8 was maintained. The lignocellulosic materials were refluxed separately in the dioxane solution for approximately 1 h under a nitrogen (N_2) atmosphere, with a constant N_2 flow rate of 50 mL/min. After refluxing, the meal-dioxane mixture was filtered using a Büchner funnel with a No. 2 filter (pore size 8 µm). The residue was washed with dioxane solution, and the filtrate was concentrated using a vacuum evaporator at 40°C. The concentrated dioxane solution was then added dropwise to deionized water under

Raw Materials	NaOH (% based on raw material)	Material:Liquor Ratio	Temperature, °C
Jute stick	18	1:10	170
Sugarcane bagasse	18	1:10	150
Mulberry	18	1:10	170
Banana tree	18	1:10	150
Corn stalk	18	1:10	150
Chia	18	1:10	170
Rice straw	18	1:10	150
Kash	18	1:10	150
Zara	18	1:10	150
Bamboo	18	1:10	170
Coconut husk	18	1:10	170
Jute fiber	18	1:10	150
Hogla	18	1:10	150
Bringle	18	1:10	170
Dhaincha	18	1:10	170

I. Conditions for extraction of lignin from nonwood lignocellulosic materials. The reaction time was 120 min for all raw materials (NaOH: sodium hydroxide).

constant stirring to precipitate the lignin. The precipitate was centrifuged, washed until a neutral pH was achieved, and dried under vacuum over phosphorus pentoxide (P_2O_5). Further purification was carried out by dissolving the lignin in an aqueous dioxane solution (9:1) and re-precipitating it in diethylether. Finally, the lignin was dried in a desiccator containing P_2O_5 under vacuum.

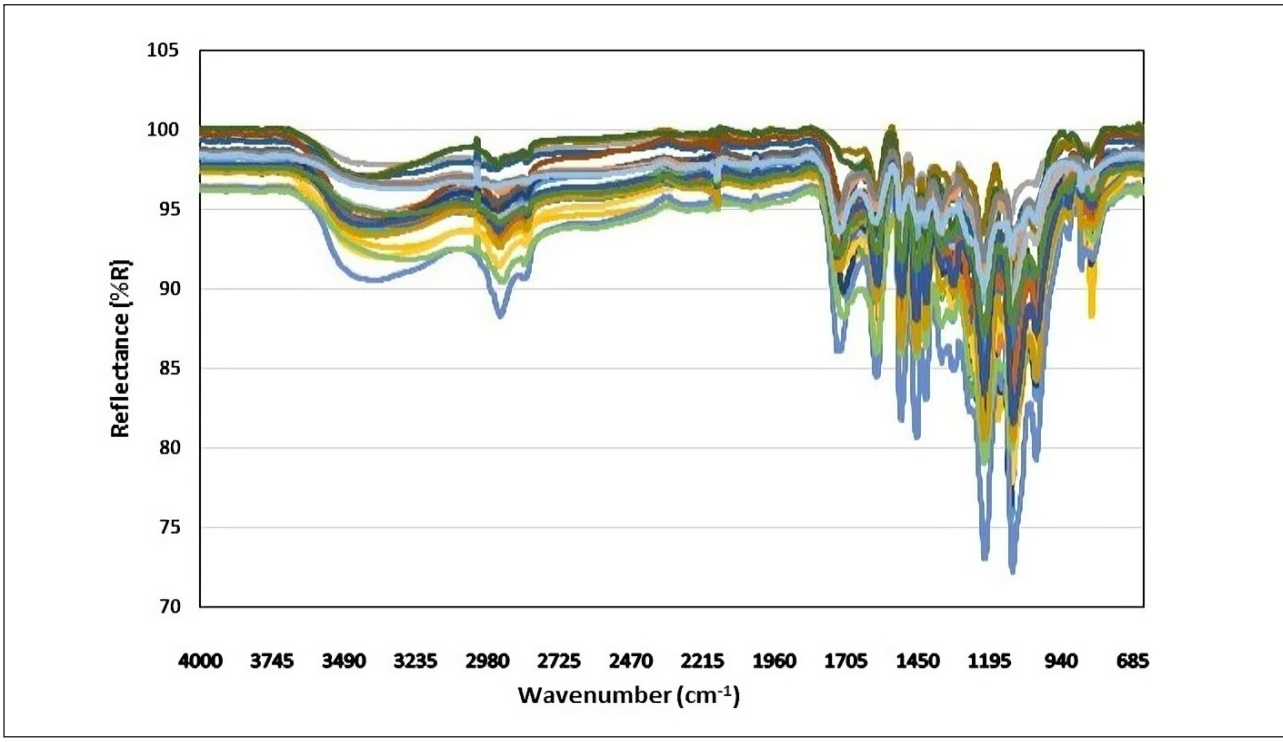
Lignins from the soda liquor of lignocellulosic materials were precipitated separately by adjusting the pH to 2 using dilute sulfuric acid. The acid precipitation method is the most common method for lignin isolation from the black liquor. The lignin isolated from the soda liquor was purified by dissolving it in a dioxane solution (9:1), precipitating it in diethylether, and then drying it under vacuum over P_2O_5.

Quantification of Ph-OH and Alp-OH groups by ^{31}P NMR

Quantitative ^{31}P NMR was used for the determination of hydroxyl content in lignin. A total of 40 mg of samples was dissolved in 500 µL of an anhydrous pyridine/deuterated chloroform ($CDCl_3$) mixture (1.6/1, v/v). Then, 200 µL of an endo-N-hydroxy-5-norbornene-2,3-dicarboximide (9.23 mg/mL, internal standard) and 50 µL of chromium (III) acetylacetonate (5.6 mg/mL, relaxation reagent), prepared using the anhydrous pyridine/$CDCl_3$, were added. The lignin solution was reacted with 100 µL of phosphitylating reagent (2-chloro-4,4,5,5-tetramethyl-1,2,3-dioxaphospholane) for 15 min and then transferred into an NMR tube for ^{31}P NMR analysis using a Bruker AVANCE III HD 600 MHz spectrometer (Bruker; Billerica, MA, USA). The 31P NMR spectra were measured using the following conditions: irradiation frequency, 202 MHz; acquisition time, 5.453 s; probe temperature, 30°C; spectral width, 80 ppm; free induction decay (FID) data points, 65,536; spinning, 15 Hz; dummy scans, 4; 1 H decoupling, Waltz16; pulse angle, 90°; and pulse width, 14.8 µs. The Alp-OH group was determined by integrating the ^{31}P-NMR signal at 145–149 ppm, while the phenolic hydroxyls were identified based on the following chemical shift ranges: 141–143 ppm for syringyl hydroxyl groups, 138–140 ppm for guaiacyl hydroxyl groups, and 136–138 ppm for *p*-hydroxyphenyl groups [27]

FTIR spectroscopic data acquisition

A PerkinElmer Frontier FTIR spectrometer (PerkinElmer; Waltham, MA, USA), fitted with a GAAS detector, was employed to collect spectra. Each sample underwent 32 scans within the spectral range of 4000 to 650 cm^{-1}, using a 16 cm^{-1} resolution and 1 cm^{-1} interval. The acquired scans were averaged, and PerkinElmer Spectrum software (Version 10.4.4) was used to process the data, yielding reflectance percentage (%R) values.



1. Presentation of spectroscopic data.

In a spectrum, each wavenumber is associated with a reflectance value, which plays a crucial role in spectral analysis. Due to the large number of reflectance values across different wavenumbers, the resulting spectroscopic variables are both numerous and highly interrelated. As a result, dimensionality reduction and pre-processing are essential to effectively manage this complexity.

The chemometric model for prediction of Alp-OH and Ph-OH groups of lignin was developed by using the whole spectra of FTIR ranging 4000–650 cm^{-1} instead of any specific peak to get a better predictive model. The whole spectra of all 28 lignin samples were plotted in **Fig. 1** to show the pattern of all the spectra together at a glance. Moreover, this figure also demonstrates whether there is any abrupt spectrum which might be an outlier from this study.

Preprocessing of spectral data

Following FTIR spectral data acquisition, the data underwent preprocessing involving various transformations. Specifically, Savitzky-Golay (S-G) smoothing with first and second derivatives, standard normal variate (SNV), and multiplicative scatter correction (MSC) were applied, and their effectiveness was evaluated.

Savitzky-Golay (S-G) smoothing

A more effective approach than merely averaging data points is to apply a least squares fit to a small set of consecutive points using a polynomial function, then use the computed central value of the fitted curve as the new smoothed data point. Savitzky and Golay demonstrated that a specific set of integers could be derived as weighting coefficients for this smoothing process. These coefficients, known as convolution integers, are mathematically equivalent to fitting the data to a polynomial, but they offer a more computationally efficient and significantly faster method [28-31]. Polynomial order of 2 and symmetric kernel with three smoothing points were used in both 1st and 2nd order derivatives of the Savitzky-Golay algorithm in this study, as this condition showed the best predictive efficiencies of the models.

Standard normal variate (SNV)

The SNV is another commonly used pretreatment in FTIR spectroscopy to eliminate scatter. It is applied individually to each spectrum. The mean and standard deviation of all data points are calculated, and each data point is then mean-centered by subtracting the average and scaling by the standard deviation [32].

Multiplicative scatter correction (MSC)

The MSC is a widely used normalization technique designed to adjust spectra so they closely align with a reference spectrum, typically the mean of the dataset. This is achieved by modifying both the scale and offset of the spectra [33]. The MSC compensates for additive and multiplicative effects in spectral data through a row-oriented transformation, meaning each data point is influenced by its neighboring values. By applying MSC, physical effects such as particle size variations and surface scatter — which do not contain relevant chemical or physical information — are effectively removed from the spectra [34].

Chemometric model calibration and validation

Principal component regression (PCR)

Principal component regression (PCR) is a technique used to analyze multiple regression data affected by multicollinearity. It follows a two-step multivariate calibration approach. In the first step, principal component analysis (PCA) is applied to the spectral data matrix, transforming the original measured variables (e.g., reflectance percentage at different wavenumber) into new scores based on latent variables. In the second step, multiple linear regression (MLR) is performed, where the scores obtained from PCA are used as predictors to model the viscosity index measured by the ASTM method [35].

Partial least squares regression (PLSR)

A PLSR identifies components from the independent variables (X) that most effectively predict the dependent variable (Y). It does so by extracting a set of latent vectors that simultaneously decompose both X and Y while maximizing their covariance. This process extends the principles of principal component analysis (PCA). Following this decomposition, a regression step is performed, where the latent vectors derived from X are used to predict Y. The PLSR models both X and Y as the product of a shared set of orthogonal factors and corresponding loadings [36, 37].

The PCR and PLS were evaluated for quantification of Alp-OH and Ph-OH groups of lignin with FTIR spectroscopic data, and the best conducting one was chosen for method development. Developed models were verified by creating a 6-fold cross validation dataset in each case. A total of 28 lignin samples were considered as calibration data for developing the models, and a new set of three lignin samples was used as the test dataset for predicting Alp-OH and Ph-OH groups using the best-performing model.

Five principal components (latent variables) in PCR and a maximum of seven factors in PLSR were considered. Random cross validation was used, and the model inputs' algorithm is singular value decomposition (SVD) in the case of PCR. The same cross validation technique was used in the case of PLSR. As input model, nonlinear iterative partial least squares (NIPALS) with maximum iteration of 100 was used here, as it showed the best predictive efficiency.

To prevent overfitting and ensure accurate calibration evaluation, six-fold cross-validation was employed in all analyses.

Following the acquisition of spectroscopic data from lignin samples via FTIR spectrophotometry, CAMO Unscrambler software (version 10.3, Camo Analytics/AspenTech; Bedford, MA, USA) was employed for PCR and PLSR calibration and validation.

Next, the data were pretreated with S-G smoothing with their first and second derivatives, SNV and MSC. Both raw and preprocessed data were used to construct chemometric models. The predictive capabilities of the PCR and PLSR models were subsequently assessed.

Finally, a new chemometric-assisted spectroscopic method is being proposed, offering better alternatives for the quantification of Alp-OH and Ph-OH groups (Syringyl-OH, Guaiacyl-OH, **p**-Hydroxyphenyl) in protolignin and technical lignin, which will reduce cost, time, and chemical usage.

RESULTS AND DISCUSSION

The Alp-OH, Syringyl-OH, Guaiacyl-OH, p-Hydroxylphenyl and total Ph-OH groups in mmol/g of proto lignin (13) and technical lignin (15) were determined by ^{31}P NMR (**Table II**). Figure 1 shows the reflectance (%R) against wavenumber (cm^{-1}) of FTIR data in the range from 4000 to 650 cm^{-1}.

Principal component analysis (PCA)

Lignin samples, when analyzed via FTIR spectroscopy and subjected to PCA, show a dispersed, non-patterned distribution in the score plot (**Fig. 2**). This lack of autocorrelation confirms their suitability for multivariate model development [38].

Principal components (PCs) are newly constructed variables formed by transforming a large set of variables into a smaller one that still contains most of the information in the large set. They are constructed in such a way that the first PC (PC1) contains maximum variation among the data, and then the second component (PC2), and so on. Here, the first seven components express 99.32% of the total variations in the data.

To ensure model robustness and detect any potential outliers that could compromise its performance, influence plots were generated within the PCA. The resulting influence plot, displaying F-Residuals against Hotelling's T2, revealed no outliers in the dataset, as clearly illustrated in **Fig. 3**.

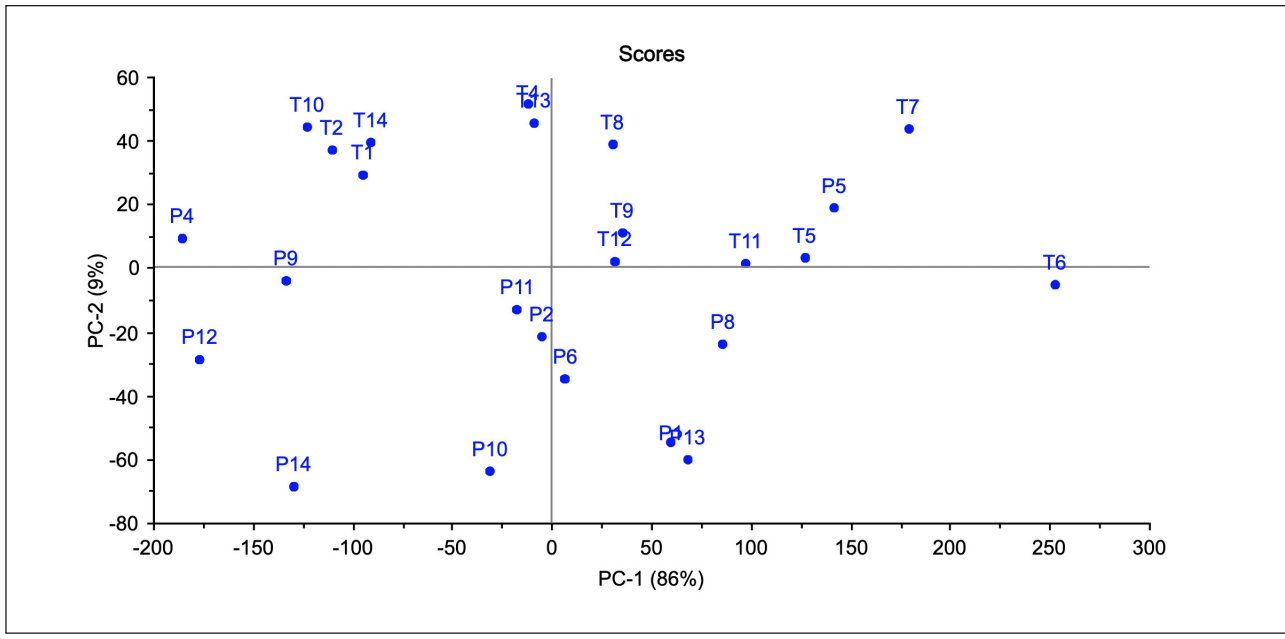
Calibration models

The PCR and PLSR are the most popular and efficient calibration techniques used in lignocellulosic materials analysis [19,21,39]. This study initially established PCR and PLSR models to quantify Alp-OH and Ph-OH groups using unprocessed FTIR data from lignin. For these models, two key parameters — slope and offset — were derived from the calibration dataset, while 6-fold cross-validation (CV) datasets were also utilized as model parameters. Additionally, the models' effectiveness was evaluated using two efficiency metrics: root mean square error (RMSE) and coefficient of determination (R^2). Model parameters (slope and offset) and model efficiency parameters (RMSE and R^2) are presented both for calibration (blue) and validation (red) datasets in **Figs. 4–8**.

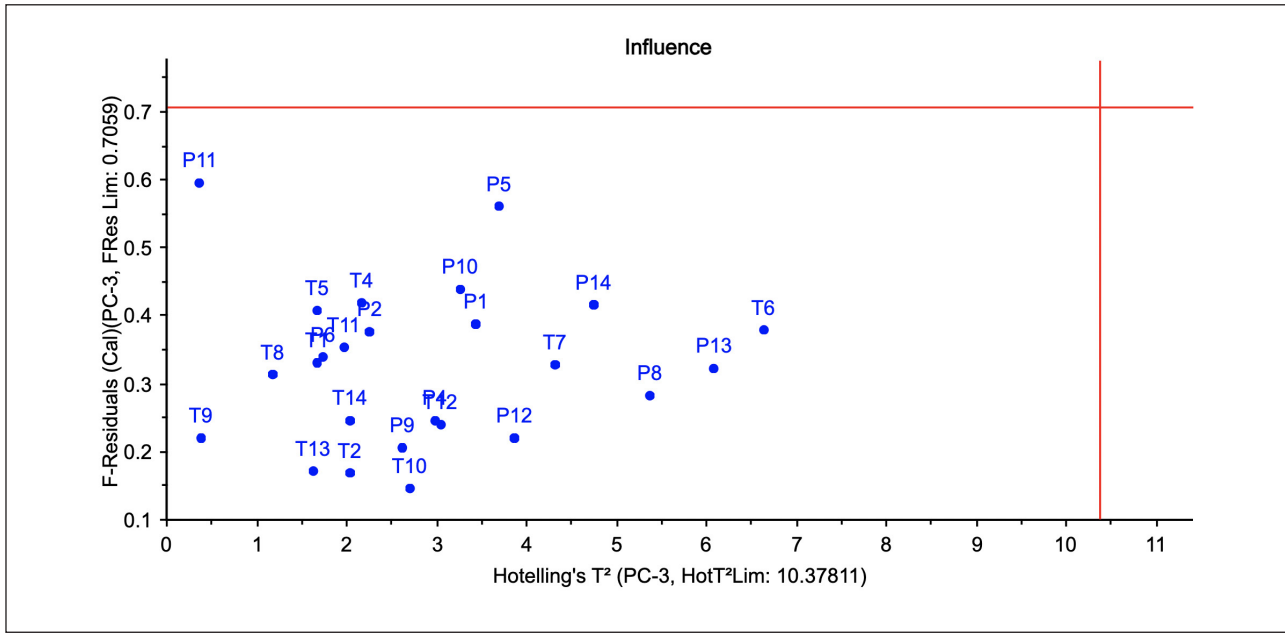
Here, the model parameters and the efficiency of the PCR and PLSR models with raw FTIR data are moderately good, making them suitable for prediction (**Table III**), though the overall performance of PLSR is better than that of PCR with raw FTIR spectral data. The moderate perfor-

SID	Samples	Alp-OH, mmol/g	Syringyl-OH, mmol/g	Guaiacyl-OH, mmol/g	p-Hydroxyphenyl, mmol/g	Total Ph-OH Group, mmol/g
PL1	Jute stick	2.700.96	0.48	0.30	1.73	
PL2	Sugarcane	1.99	0.38	0.34	0.91	1.63
PL3	Banana tree	2.07	0.30	0.37	0.61	1.28
PL4	Corn stalk	1.82	0.46	0.46	0.27	1.19
PL5	Chia	2.46	0.56	0.52	0.28	1.36
PL6	Rice straw	1.43	0.15	0.38	0.32	0.84
PL7	Kash	1.85	0.29	0.64	0.52	1.45
PL8	Zara grass	2.00	0.39	0.44	0.88	1.72
PL9	Bamboo	2.71	0.60	0.59	0.48	1.67
PL10	Coconut husk	1.96	0.22	1.28	0.71	2.20
PL11	Jute fiber	3.56	0.36	0.09	0.14	0.31
PL12	Hogla	2.71	0.43	0.50	0.59	1.53
PL13	Bringle	3.05	0.64	0.63	0.17	1.44
TL1	Jute stick	1.33	2.34	0.58	0.34	3.26
TL2	Sugercane	1.83	1.12	0.73	0.52	2.37
TL3	Mulberry	1.38	1.65	0.94	0.81	3.40
TL4	Banana tree	1.49	1.01	0.51	0.60	2.12
TL5	Corn stalk	1.65	1.32	0.94	0.56	2.82
TL6	Chia	1.53	1.55	0.87	0.77	3.18
TL7	Rice straw	1.13	0.63	0.84	0.22	1.69
TL8	Kash	1.61	0.81	1.17	0.44	2.42
TL9	Zara grass	1.96	1.14	0.81	0.51	2.46
TL10	Bamboo	1.05	1.61	0.84	0.69	3.13
TL11	Coconut husk	1.59	1.60	0.84	0.43	2.87
TL12	Jute fiber	2.36	1.74	0.41	0.44	2.59
TL13	Hogla	1.97	1.05	0.70	0.43	2.18
TL14	Bringle	0.91	0.81	0.86	0.76	2.43
TL15	Dhaincha	1.25	2.12	0.46	0.47	3.06

II. Alp-OH and Ph-OH groups in lignin samples (SID: sample ID) by ^{31}P nuclear magnetic resonance (^{31}P NMR).



2. Score plot of first two principal components (PC1, PC2).



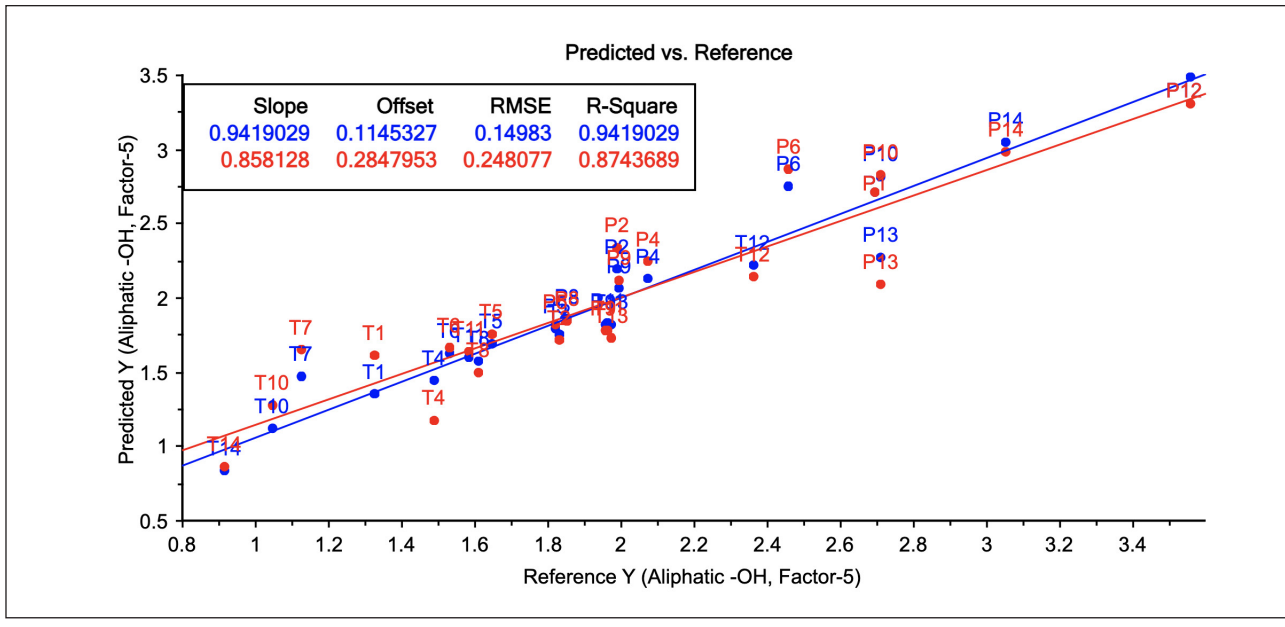
3. Loading plot of principal component analysis (PCA).

mance of the PCR model can be attributed the raw spectral data of lignin samples from mostly different heterogeneous nonwood agricultural residues. Assessing the pretreatment of spectral data to achieve better results with these two calibration techniques is needed here.

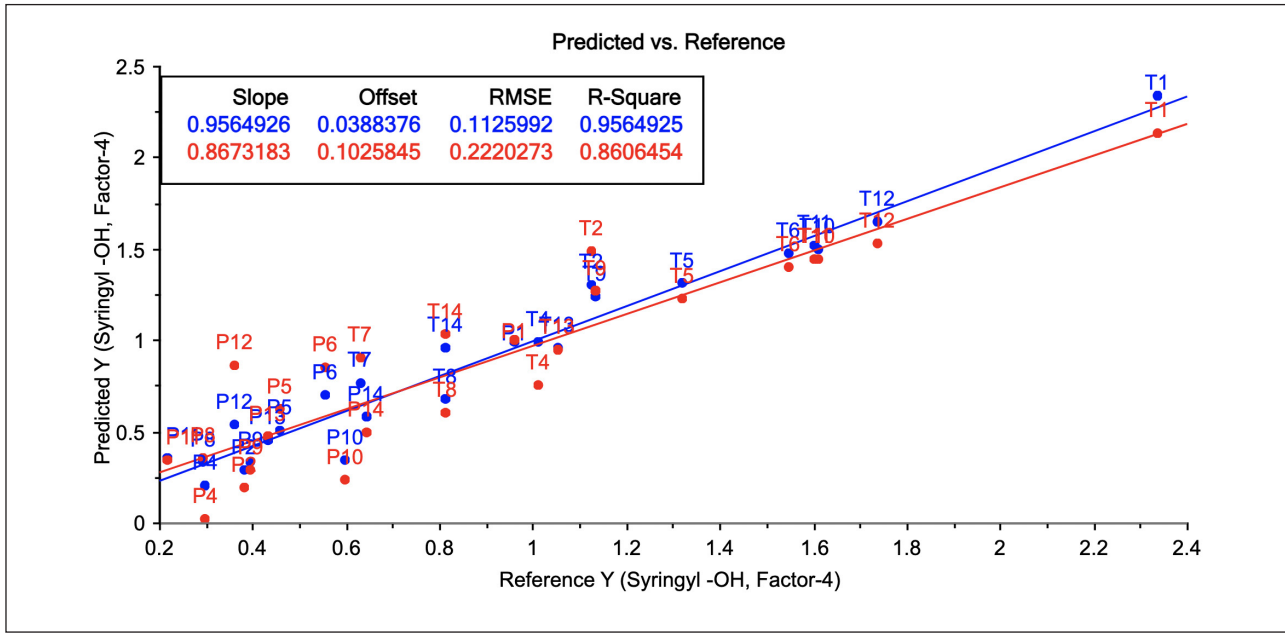
The PCR constructs components based solely on the variation in the predictor variables, without incorporating any information from the response variable. In contrast, PLSR takes the response variable into account when forming components, which often results in models that can achieve a good fit using fewer components. In the present study, the PLSR model performed better than PCR for predicting Alp-OH and Ph-OH groups of lignin with FTIR spectral data both in calibration and validation (**Table III**).

Effects of pretreatment of spectral data

To enhance predictive accuracy, the FTIR data were subjected to mathematical preprocessing to denoise them through smoothing by several techniques. To enhance predictive accuracy, Savitzky-Golay (S-G) filtering with a 1st derivative was initially performed on the raw FTIR data. The PCR and PLSR models were then constructed using the



4. The partial least squares regression (PLSR) model for determination of aliphatic OH with calibration (blue) and validation (red) datasets (RMSE: root mean square error; R-Square: coefficient of determination).

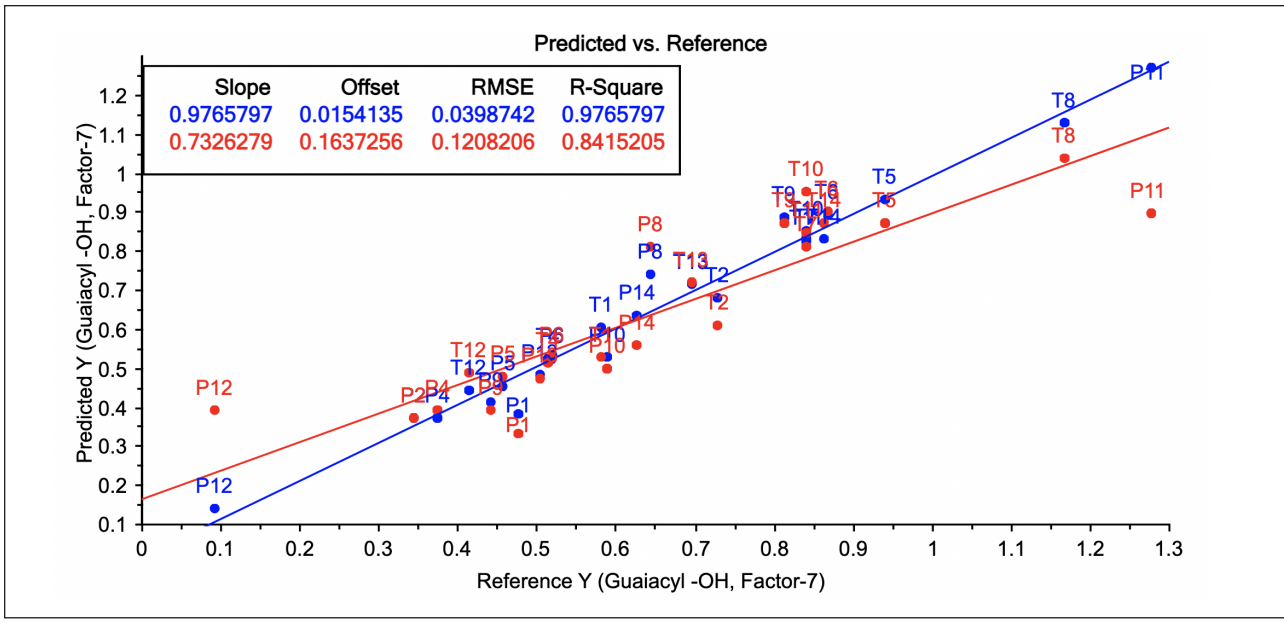


5. The PLSR model for determination of syringyl OH with calibration (blue) and validation (red) datasets.

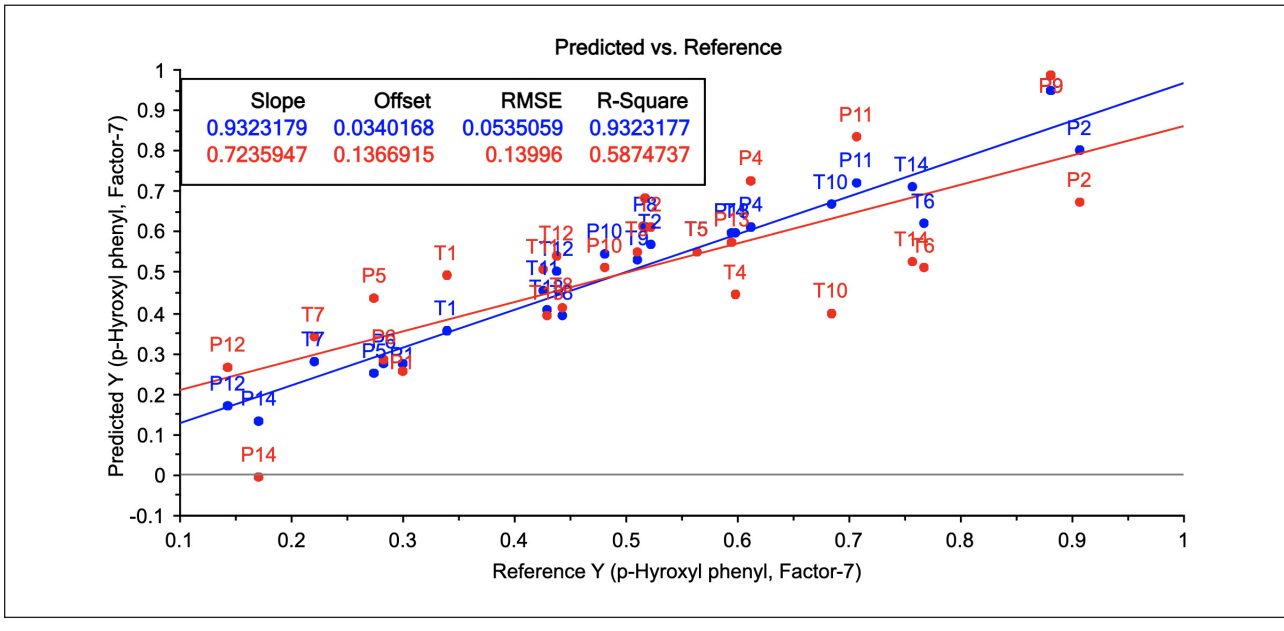
resulting filtered data. Unfortunately, the performance of these models was unsatisfactory (Table II). Consequently, in an effort to achieve better predictions, the FTIR data underwent a second round of processing, this time using S-G filtering with a 2nd derivative. New PCR and PLSR models were generated from these treated data, with the results summarized in Table III. Prediction performance of PCR and PLSR are still not up to the mark with both 1st and 2nd derivatives S-G filtering.

To further enhance predictive performance, the raw FTIR data were subjected to SNV transformation. This treatment dramatically improved both the PCR model and the PLSR model, with PLSR showing particularly significant gains (Table III). With this SNV treated data, the coefficient of determination (R^2) of the PLSR model improved satisfactorily from 93% to 97.6% for predicting Alp-OH, Syringyl-OH, Guaiacyl-OH, p-Hydroxyphenyl, and total Ph-OH groups in lignin.

Finally, MSC was applied to the FTIR spectroscopic data to enhance the predictive efficiency of the PCR and the PLSR models. The results are presented in Table III. While MSC pretreatment yielded only marginal improvement for



6. The PLSR model for determination of guaiacyl OH with calibration (blue) and validation (red) datasets.



7. The PLSR model for determination of p-hydroxyphenyl with calibration (blue) and validation (red) datasets.

the PCR model, it resulted in a dramatic increase in PLSR model efficiency, with the R^2 value for calibration data rising from 93.2% to 97.7%. The validation dataset also demonstrated satisfactory R^2 values. The findings demonstrate that PLSR models are highly effective and stable for predicting Alp-OH, Syringyl-OH, Guaiacyl-OH, p-Hydroxyphenyl, and total Ph-OH groups in lignin from lignocellulosic samples (Figs. 4–8).

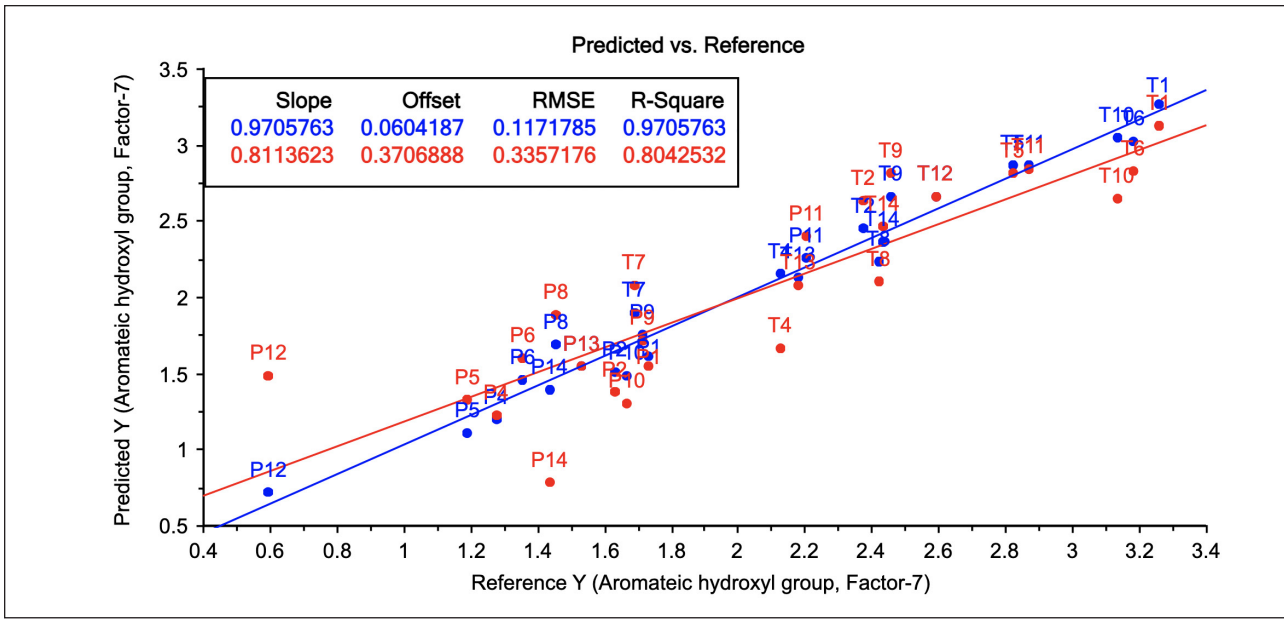
Prediction efficiency of the development models

Finally, with the best performing PLSR model of FTIR data pretreated by MSC, Alp-OH, Syringyl-OH, Guaiacyl-OH, p-Hydroxyphenyl and total Ph-OH group contents in lignin of three samples were predicted, and the results are presented in **Table IV**.

These predicted values of Alp-OH, Syringyl-OH, Guaiacyl-OH, p-Hydroxyphenyl, and total Ph-OH groups are very close to their calculated values, which manifests a very good predictive efficiency for the models developed in this study.

CONCLUSION

Neither the PCR model or the PLSR model demonstrate satisfactory predictive performance when applied to raw FTIR data for quantifying the Alp-OH, Syringyl-OH, Guaiacyl-



8. The PLSR model for determination of total Ph-OH group with calibration (blue) and validation (red) datasets.

Parameters	Datasets	Raw Data		First Derivative with 2nd Polynomial		Second Derivative with 2nd Polynomial		Standard Normal Variate (SNV)		Multiplicative Scatter Correction (MSC)	
		PCR	PLSR	PCR	PLSR	PCR	PLSR	PCR	PLSR	PCR	PLSR
Aliphatic-OH	Calibration	0.696	0.718	0.738	0.800	0.733	0.966	0.868	0.930	0.869	0.942
	Validation	0.671	0.600	0.649	0.676	0.510	0.480	0.745	0.841	0.792	0.874
Syringyl-OH	Calibration	0.714	0.800	0.683	0.979	0.568	0.906	0.844	0.955	0.858	0.956
	Validation	0.561	0.653	0.571	0.622	0.501	0.481	0.753	0.842	0.787	0.860
Guaiacyl-OH	Calibration	0.894	0.925	0.681	0.925	0.586	0.936	0.859	0.976	0.890	0.977
	Validation	0.484	0.596	0.453	0.547	0.193	0.405	0.514	0.818	0.793	0.841
p-Hydroxyphenyl	Calibration	0.542	0.824	0.383	0.981	0.00	0.823	0.347	0.934	0.350	0.932
	Validation	0.068	0.333	0.264	0.187	0.00	0.311	0.257	0.544	0.219	0.587
Total Ph-OH groups	Calibration	0.629	0.828	0.687	0.758	0.691	0.858	0.611	0.943	0.735	0.971
	Validation	0.312	0.492	0.619	0.654	0.489	0.502	0.502	0.806	0.542	0.804

III. Comparison of predictive efficiencies of the models in terms of coefficient of determination (R^2) with raw and pretreated spectral data (PCR: principal component regression; PLSR: partial least squares regression).

OH, p-Hydroxyphenyl, and total Ph-OH groups in nonwood lignin samples. To address this issue, mathematical preprocessing techniques, such as Savitzky-Golay (S-G) filtering with first and second derivatives, were employed. While these preprocessing steps improved model performance for the same parameters, they remained insufficient for reliable predictive use.

To further enhance predictive accuracy, SNV and MSC were applied to the preprocessed FTIR data. This transformation of spectral data had minimal impact on the PCR model; however, it led to a significant improvement in the PLSR model, making it highly efficient and stable for predicting the Alp-OH, Syringyl-OH, Guaiacyl-OH, p-Hydroxyphenyl, and total Ph-OH groups in nonwood lignin. By strictly comparing the efficiency of SNV and MSC, the latter showed the best predictive performance. The model

Parameters	Prediction Sample ID	Values Determined by ^{31}P NMR mmole/g	Predicted Values Determined by Developed Models mmole/g	Deviation
Alp-OH	PS1	2.71	2.28	0.43
	PS2	3.05	3.02	0.03
	PS3	1.33	1.55	-0.22
Syringyl-OH	PS1	0.43	0.45	-0.02
	PS2	0.64	0.58	0.06
	PS3	2.34	2.34	0
Guaiacyl-OH	PS1	0.50	0.48	0.02
	PS2	0.63	0.63	0
	PS3	0.58	0.60	-0.02
p-Hydroxyphenyl	PS1	0.59	0.59	0
	PS2	0.17	0.13	0.04
	PS3	0.34	0.35	-0.01
Total Ph-OH	PS1	1.53	1.55	-0.02
	PS2	1.44	1.38	0.06
	PS3	3.24	3.27	-0.03

IV. Prediction of Alp-OH and Ph-OH Groups of three new samples by using developed models.

achieved high predictive efficiency for predicting Alp-OH (R^2 = 94.2%), Syringyl-OH (R^2 = 95.6%), Guaiacyl-OH (R^2 = 97.7%), *p*-Hydroxyphenyl (R^2 = 93.2%), and total Ph-OH groups (R^2 = 97.1%) in lignin.

Ultimately, the Alp-OH, Syringyl-OH, Guaiacyl-OH, p-Hydroxyphenyl, and total Ph-OH groups were successfully quantified using the PLSR model with FTIR data preprocessed by MSC, and the predicted values are very close to the values measured by NMR. This approach enables Alp-OH, Syringyl-OH, Guaiacyl-OH, p-Hydroxyphenyl, and total Ph-OH groups quantification in non-wood lignin samples solely based on FTIR spectral data, eliminating the need for chemical reagents or labor-intensive procedures. As a result, this method offers a non-destructive, rapid, and cost-effective alternative for determining the Alp-OH, Syringyl-OH, Guaiacyl-OH, p-Hydroxyphenyl, and total Ph-OH groups in lignin from similar sources. **TJ**

LITERATURE CITED

1. Rowell, R.M., *Handbook of Wood Chemistry and Wood Composites*, CRC Press, Boca Raton, FL, USA, 2005. https://doi.org/10.1201/9780203492437.
2. Akpan, E.I., "Chemistry and structure of lignin," in *Sustainable Lignin for Carbon Fibers: Principles, Techniques, and Applications* (E.I. Akpan and S.O. Adeosun, Eds.), Springer International Publishing, Cham, 2019. https://doi.org/10.1007/978-3-030-18792-7_1.
3. Li, T. and Takkellapati, S., *Biofuels Bioprod. Biorefin.* 12(5): 756(2018). https://doi.org/10.1002/bbb.1913.
4. Eraghi Kazzaz, A. and Fatehi, P., *Ind. Crops Prod.* 154: 112732(2020). https://doi.org/10.1016/j.indcrop.2020.112732.
5. Matsushita, Y., *J. Wood. Sci.* 61: 230(2015). https://doi.org/10.1007/s10086-015-1470-2.
6. Gosselink, R.J.A., Van Dam, J.E.G., De Jong, E., et al., *Holzforschung* 64(2): 193(2010). https://doi.org/10.1515/hf.2010.023.
7. Hany, L.J.S., Rahman, M.M., Popy, R.S., et al., *Eur. J. Wood Prod.* 82: 1357(2024). https://doi.org/10.1007/s00107-024-02094-y.
8. Rahman, M.M., Jahan, E., Jahan, M.S., et al., *Biomass Convers. Biorefin.* 14: 30515(2024). https://doi.org/10.1007/s13399-023-04901-7.
9. Sutradhar, S., Arafat, K.M.Y., Nayeem, J., et al., *Cellul. Chem. Technol.* 54(5-6): 463(2020). https://doi.org/10.35812/CelluloseChemTechnol.2020.54.47.
10. Duan, X., Wang, X., Chen, J., et al., *RSC Adv.* 12: 24242(2022). https://doi.org/10.1039/D2RA02085H.

11. Tagami, A., Gioia, C., Lauberts, M., et al., *Ind. Crops. Prod.* 129: 123(2019). https://doi.org/10.1016/j.indcrop.2018.11.067.
12. Zakis, G.F., *Functional Analysis of Lignins and Their Derivatives* (T. Joyce and R. Brezny, Trans. Eds.), Trans. edn., TAPPI Press, Atlanta, 1994, p. 102.
13. Argyropoulos, D., *J. Wood Chem. Technol.* 14(1): 65(1994). https://doi.org/10.1080/02773819408003086.
14. Meng, X., Crestini, C., Ben, H., et al., *Nat. Protoc.* 14: 2627(2019). https://doi.org/10.1038/s41596-019-0191-1.
15. Stücker, A., Podschun, J., Saake, B., et al., *Anal. Methods* 10: 3481(2018). https://doi.org/10.1039/C8AY01272E.
16. Uddin, M.N., Rahman, M.M., Likhon, M.N.A., et al., *Nord. Pulp Pap. Res. J.* 40(1): 71(2025). https://doi.org/10.1515/npprj-2024-0060.
17. Uddin, M.N., Ahmed, S., Ray, S.K., et al., *Nord. Pulp Pap. Res. J.* 34(1): 1(2019). https://doi.org/10.1515/npprj-2018-0018.
18. Uddin, M.N., Ferdous, T., Islam, Z., et al., *J. Bioresour. Bioprod.* 5(3): 196(2020). https://doi.org/10.1016/j.jobab.2020.07.005.
19. Uddin, M.N., Nayeem, J., Islam, M.S., et al., *Biomass Conver. Biorefin.* 9: 585(2019). https://doi.org/10.1007/s13399-019-00383-8.
20. Saariaho, A.-M., Jääskeläinen, A.-S., Nuopponen, M., et al., *Appl. Spectrosc.* 57(1): 58(2003). https://doi.org/10.1366/000370203321165214.
21. Hoang, V., Bhardwaj, N.K., and Nguyen, K.L., *Carbohydr. Polym.* 61(1): 5(2005). https://doi.org/10.1016/j.carbpol.2004.12.007.

ABOUT THIS PAPER

Cite this article as:
Uddin, M.N., Likhon, M.N.A., Rahman, M.M., et al., *TAPPI J.* 24(10): 479(2025). https://doi.org/10.32964/TJ24.10.479

DOI: https://doi.org/10.32964/TJ24.10.479

ISSN: 0734-1415

Publisher: TAPPI Press

About this journal

22. Monrroy, M., Mendonça, R., Baeza, J., et al., *J. Near Infrared Spectrosc.* 16(2) :121(2008). https://doi.org/10.1255/jnirs.766.
23. Chai, X.S., Zhu, J.Y., and Li, J., *J. Pulp. Pap. Sci.* 27: 165(2001).
24. Uddin, M.N., Ferdous, T., Likhon, M.N.A., et al., *J. Appl. Polym. Sci.* e57698(2025). https://doi.org/10.1002/app.57698.
25. Uddin, M.N., Ferdous, T., Jin, Y., et al., *Anal. Sci. Adv.* 6(1): e70005(2025). http://dx.doi.org/10.1002/ansa.70005.
26. "Solvent extractives of wood and pulp," TAPPI Standard Test Method T 204 cm-17, TAPPI Press, Atlanta, 2017.

ABOUT THE AUTHORS

We studied this topic to develop an easy and environmentally friendly method to determine hydroxyl groups in lignin. Finding a stable and robust model was a major challenge, but this was solved by different treatments of data.

During this study, Jahan was responsible for idea generation and draft finalization; M.N. Uddin performed all analysis, modeling, and draft preparation; Likhon and Rahman were responsible for laboratory analysis; Jin and Jiang performed NMR analysis; and M.T. Uddin performed laboratory work and proof reading.

We eventually discovered an efficient and stable model for determining hydroxyl groups in lignin that was simple, non-destructive, and environmentally friendly. The model could be easily used by mill technicians.

M.N. Uddin

Likhon

Rahman

Jin

Jiang

M.T. Uddin

Jahan

M.N. Uddin, Rahman, and M.T. Uddin are principal scientists and Likhon is scientist in the Pulp and Paper Division at the Bangladesh Council of Scientific and Industrial Research (BCSIR) in Dhaka, Bangladesh. Jin and Jiang are professors at the Nanjing Forestry University in Nanjing, China. Jahan is retired director of BCSIR in Dhaka, Bangladesh. Email Jahan at sarwar2065@hotmail.com and M.N. Uddin at nashirbcsir@gmail.com.

27. Jiang, B., Zhang, Y., Guo, T., et al., *Polymers* 10(7): 736(2018). https://doi.org/10.3390/polym10070736.
28. Brereton, R.G., *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*, John Wiley & Sons, 2003.
29. Massart, D.L., Vandeginste, B.G.M., Deming, S.N., et al., *Chemometrics: A Textbook*, Data Handling in Science and Technology Series, Elsevier, Vol. 2, 1988.
30. Einax, J.W., Brown, S.D., Tauler, R., et al., (Eds.), "Comprehensive chemometrics. Chemical and biochemical data analysis," *Anal. Bioanal. Chem.* 396: 551(2009). https://doi.org/10.1007/s00216-009-3284-9.
31. Windig, W., Shaver, J., and Bro, R., *Appl. Spectrosc.* 62(10): 1153(2008). https://doi.org/10.1366/000370208786049097.
32. Maleki, M.R., Mouazen, A.M., Ramon, H., et al., *Biosyst. Eng.* 96(3): 427(2007). https://doi.org/10.1016/j.biosystemseng.2006.11.014.
33. Jolliffe, I.T., "Principal component analysis for special types of data," in *Principal Component Analysis*, Springer Series in Statistics, Springer-Verlag, New York, 2002, pp. 338-372.
34. Colares, C.J.G., Pastore, T.C.M., Coradin, V.T.R., et al., *J. Braz. Chem. Soc.* 26: 1297(2015). https://doi.org/10.5935/0103-5053.20150096.
35. He, W. and Hu, H., *J. Wood Chem. Technol.* 33(1): 52(2013). https://doi.org/10.1080/02773813.2012.731463.
36. Li, X., Sun, C., Zhou, B., et al., *Sci. Rep.* 5: 17210(2015). https://doi.org/10.1038/srep17210.
37. Luo, Q. and Zhu, J., *J. Sci. Technol. For. Prod. Processes* 5: 6(2015).
38. Shukla, S.R., Shashikala, S., and Sujatha, M., *J. Near Infrared Spectrosc.* 29(3): 168(2021). https://doi.org/10.1177/0967033521999118.
39. Uddin, M.N., Ray, S.K., Islam, M.S., et al., *J. Sci. Technol. For. Prod. Processes* 6: 22(2017).