Exploring Image Processing Tools To Unravel Complex ^{1}H-^{13}C Heteronuclear Single-Quantum Correlation Nuclear Magnetic Resonance Spectra

Ghysels, Stef; Verwaeren, Jan; Heeres, Hero Jan; Rohrbach, Léon; Backx, Simon; Mangelinckx, Sven; Ronsse, Frederik

Published in:
Energy and Fuels

DOI:
10.1021/acs.energyfuels.2c04100

Publication date:
2023

energy&fuels

pubs.acs.org/EF

Article

Exploring Image Processing Tools To Unravel Complex $^{1}H-^{13}C$ Heteronuclear Single-Quantum Correlation Nuclear Magnetic Resonance Spectra: A Demonstration for Pyrolysis Liquids

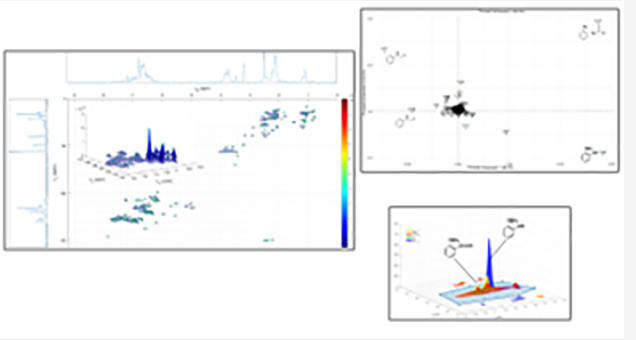
Stef Ghysels,* Jan Verwaeren, Hero Jan Heeres, Léon Rohrbach, Simon Backx, Sven Mangelinckx, and Frederik Ronsse

Cite This: *Energy Fuels* 2023, 37, 4446−4459

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Pyrolysis liquids are very complex and heterogeneous in composition. This makes them hard to comprehensively analyze, which is one of the hurdles that could hinder further advances in science and technology toward their valorization. Recently, renewed interest grew for quantitative recording of two-dimensional $^{1}H-^{13}C$ heteronuclear single-quantum correlation (HSQC) nuclear magnetic resonance (NMR). This makes $^{1}H-^{13}C$ HSQC NMR a valuable tool to fingerprint and quantitatively assess these complex liquids. However, data analysis of complex $^{1}H-^{13}C$ HSQC spectra lacks behind on these recent experimental developments. That is, $^{1}H-^{13}C$ HSQC spectra are often manually and *ad hoc* analyzed. This work, therefore, seeks to automate data analysis from $^{1}H-^{13}C$ HSQC spectra. We explored the use of image processing tools and identified their much underestimated potential. Indeed, many of the existing tools (often built-in software) were found to be applicable for noise detection/removal, generation/comparison of regions of interest, etc. Moreover, pseudo-Voigt peaks were fitted to the $^{1}H-^{13}C$ HSQC spectra, with an average R^2 of 0.94. These fitted spectral peaks allowed for the generation of a peak list, as an input for multivariate analysis. This allowed for pinpointing differences in the chemical composition of the samples. Overall, a new echelon for easy analysis of $^{1}H-^{13}C$ HSQC spectra has been explored and demonstrated.



INTRODUCTION

Biorefining goes in tandem with the application and processing of often very heterogeneous feedstocks and complex refinery intermediates. One example thereof is fast pyrolysis of lignocellulosic biomass. Indeed, fast pyrolysis generally leads to liquids, whose inherent complexity challenges subsequent advances for further valorization as a fuel precursor or pool for chemical building blocks.[1,2] Pyrolysis liquids contain a large number of different phenolics, anhydrosugars, carboxylic acids, aldehydes, and other volatile compounds. On top of that, these pyrolysis liquids contain water and high-boiling-point compounds that have a large molecular weight.[3] Gas chromatography-based techniques, coupled to mass spectrometry and/or flame ionization detection (GC−MS, GC−FID, and GC−MS/FID) dominate pyrolysis liquid analysis.[4] However, GC only allows for analysis of volatile compounds but not water (20−30 wt %) and the high-boiling-point compounds (20−40 wt %).[5] While water analysis is straightforward (Karl Fischer titration), analysis of the compounds within heavy liquids usually is not and requires separate supplementary analyses, like gel permeation chromatography and high-pressure liquid chromatography coupled to mass spectrometry, among others.[6] Recently, Dao Thi et al.[7] were able to analyze 34 wt % lignin monomers but, more importantly, 16 wt % dimers and 7 wt % trimers of a lignin oil from reductive catalytic fractionation using high-temperature-resistant column sets and sample derivatization upon GC analysis. However, this is one of the very few examples thereof.

Nuclear magnetic resonance (NMR) analysis of pyrolysis liquids offers the advantage that it renders analysis of both light and heavy compounds simultaneously, without extensive sample preparation.[8] ^{1}H and ^{13}C NMR are techniques that are useful to assess the structural hydrogencarbon framework of the pyrolysis constituents.[8] Especially, $^{1}H-^{13}C$ heteronuclear single-quantum correlation (HSQC) NMR of pyrolysis liquids has increasingly been applied.[9−13] Indeed, this technique can detect bond correlations between H and C and, therefore, result in a two-dimensional (2D) plot, where

Received: December 5, 2022
Revised: February 24, 2023
Published: March 7, 2023

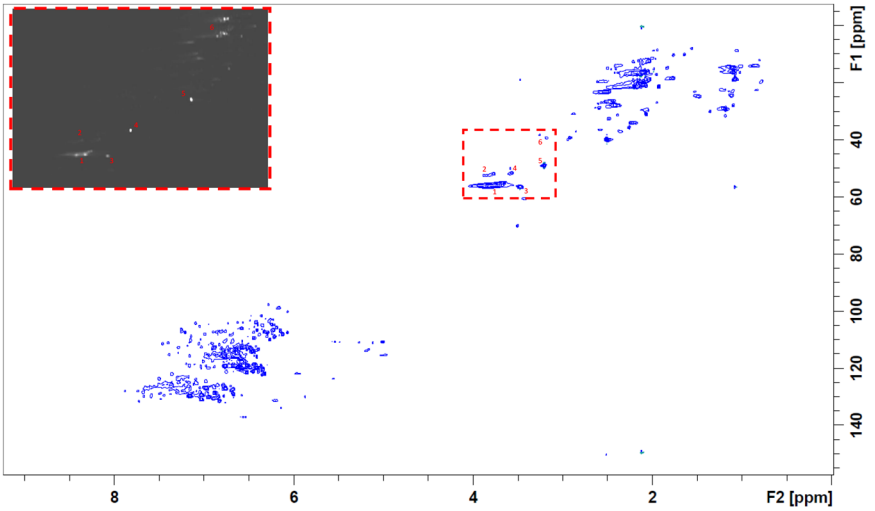


Figure 1. Illustrative $^{1}H-^{13}C$ HSQC NMR spectrum of pyrolysis liquids, obtained from lignin treated with $Ca(OH)_2$, with an embedded gray-scale representation of the methoxy region, indicated by a dashed square.

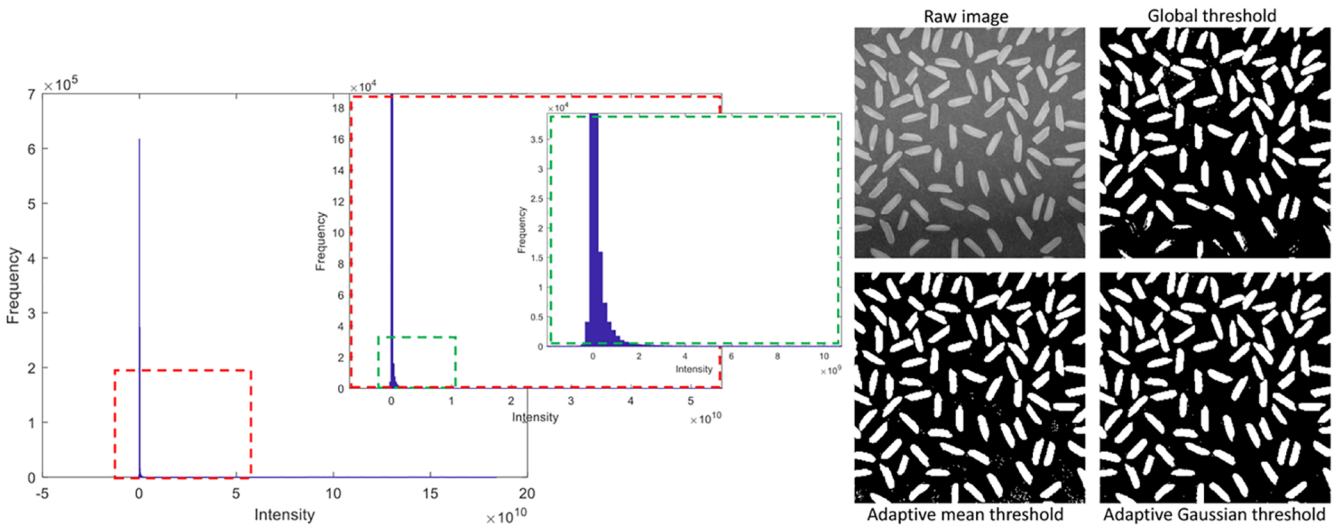


Figure 2. (Left) Histogram of the spectral intensities from the raw as-imported $^{1}H-^{13}C$ HSQC NMR spectrum, obtained from pyrolysis liquids of $Ca(OH)_2$-pretreated lignin. (Right) Illustration of various thresholding methods, using a gray-scale image of rice.

peak overlap is minimized and, thus, signals are better resolved. While single-compound detection is possible (to some extent), an additional advantage of NMR is that it probes the chemical environment of an atom (or H−C correlation in the case of $^{1}H-^{13}C$ HSQC NMR). Therefore, changes in chemical functionalities upon, e.g., hydrodeoxygenation, can be simply inferred, because intensities signify "functionalities". Practice however shows that $^{1}H-^{13}C$ HSQC NMR analysis is often not a standard choice for quantitative analysis of complex liquids. Typically, $^{1}H-^{13}C$ HSQC NMR analyses fulfill a role in qualitative assessment. This is because the intensities of peaks in the 2D spectrum are not directly proportional to the concentration of the compounds within the analyte as a result of resonance-specific signal attenuation during the coherence transfer periods. Nonetheless, valuable work has been conducted in recent years, which does enable quantitative $^{1}H-^{13}C$ HSQC NMR experiments.[14−18] One technique of increasing interest is the extrapolated time-zero $HSQC_0$. There, a 2D spectrum can be obtained, being $HSQC_0$, in which the signals are proportional to the concentration. This is because errors as a result of different relaxation times are compensated. The signal intensities within the $HSQC_0$ spectrum are obtained by extrapolating the peak heights derived from spectra collected at the end of the basic HSQC block, which is repeated two ($HSQC_2$) or three ($HSQC_3$) times, to time zero ($HSQC_0$).[15,17] Referencing to an internal standard of known concentration then allows for the derivatization of absolute concentrations. Validation of this technique is presented by Fardus-Reid et al.[14] and Talebi Amiri et al.,[16] among others.

These progresses prompt for the development of fast, elegant, and accurate data processing tools to fully unfold $^{1}H-^{13}C$ HSQC NMR as a comprehensive toolbox to analyze complex biorefinery streams, such as pyrolysis liquids. This work focuses on the data processing of recorded $^{1}H-^{13}C$ HSQC NMR spectra and puts forth a methodology that is largely automated (that is, which requires as few as possible *ad hoc* manual manipulations). It is indeed common in the literature that certain peaks of interest are manually exported prior to further (multivariate) analysis.[19]

https://doi.org/10.1021/acs.energyfuels.2c04100
Energy Fuels 2023, 37, 4446−4459

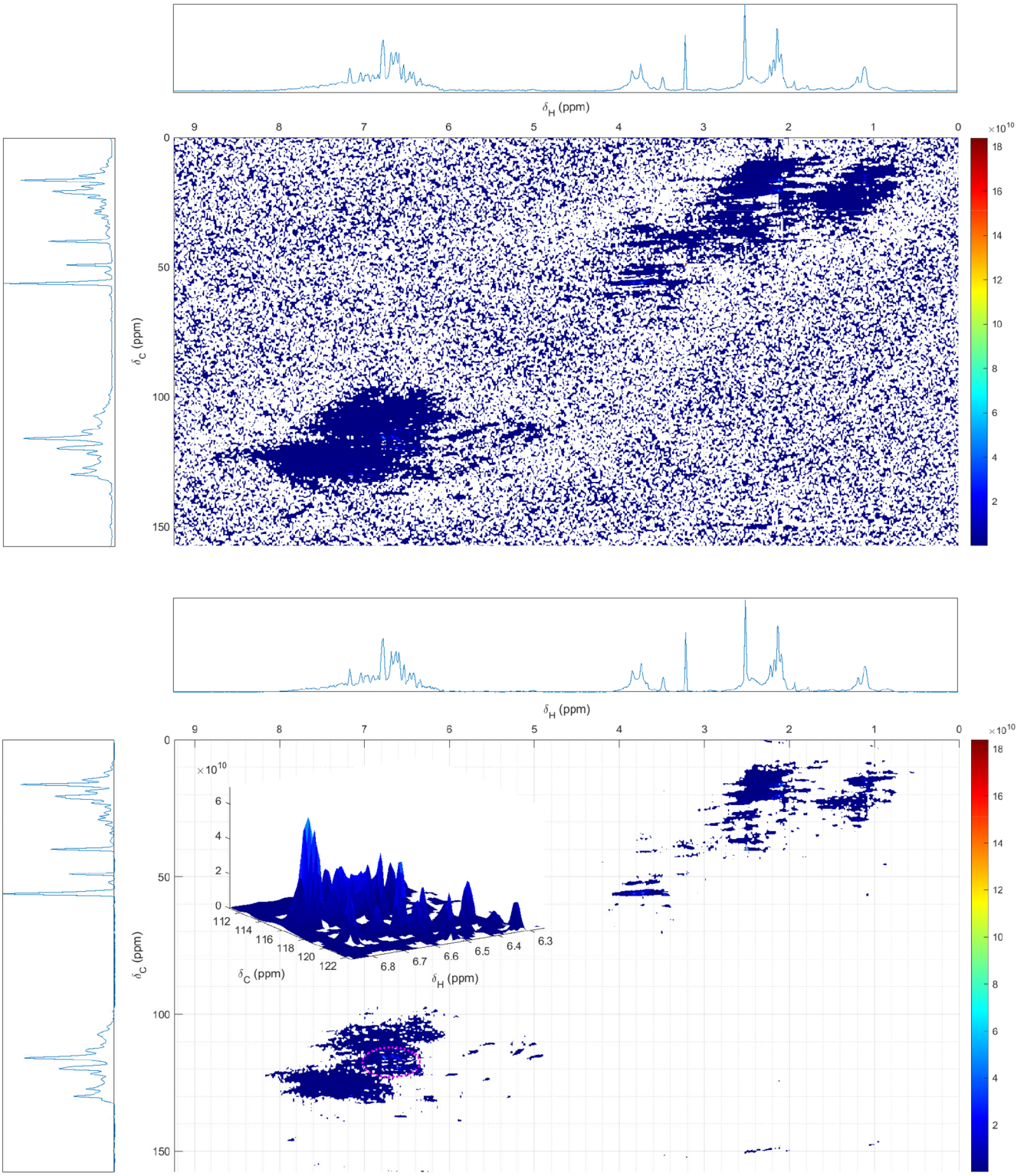


Figure 3. (Top) Raw $^{1}H-^{13}C$ spectrum from liquids of lignin, treated with $Ca(OH)_2$. Noise-related intensities that were negative were given a white color for better visibility of the overall spectrum. (Bottom) Same spectrum, using mean centering, involving global thresholding with an intensity of 5.16×10^7. A magnified part of the spectrum, indicated by the dashed ellipse, is embedded in the spectrum.

The application of image processing coupled to multivariate analysis of spectra is a very much underexplored research area. Only very recently, Kuhn et al.[20] conducted a pilot study, where substructures in 2D NMR spectra of mixtures were identified, using a tailored image-based convolutional neural network application. They were able to classify 2D HSQC

https://doi.org/10.1021/acs.energyfuels.2c04100
Energy Fuels 2023, 37, 4446−4459

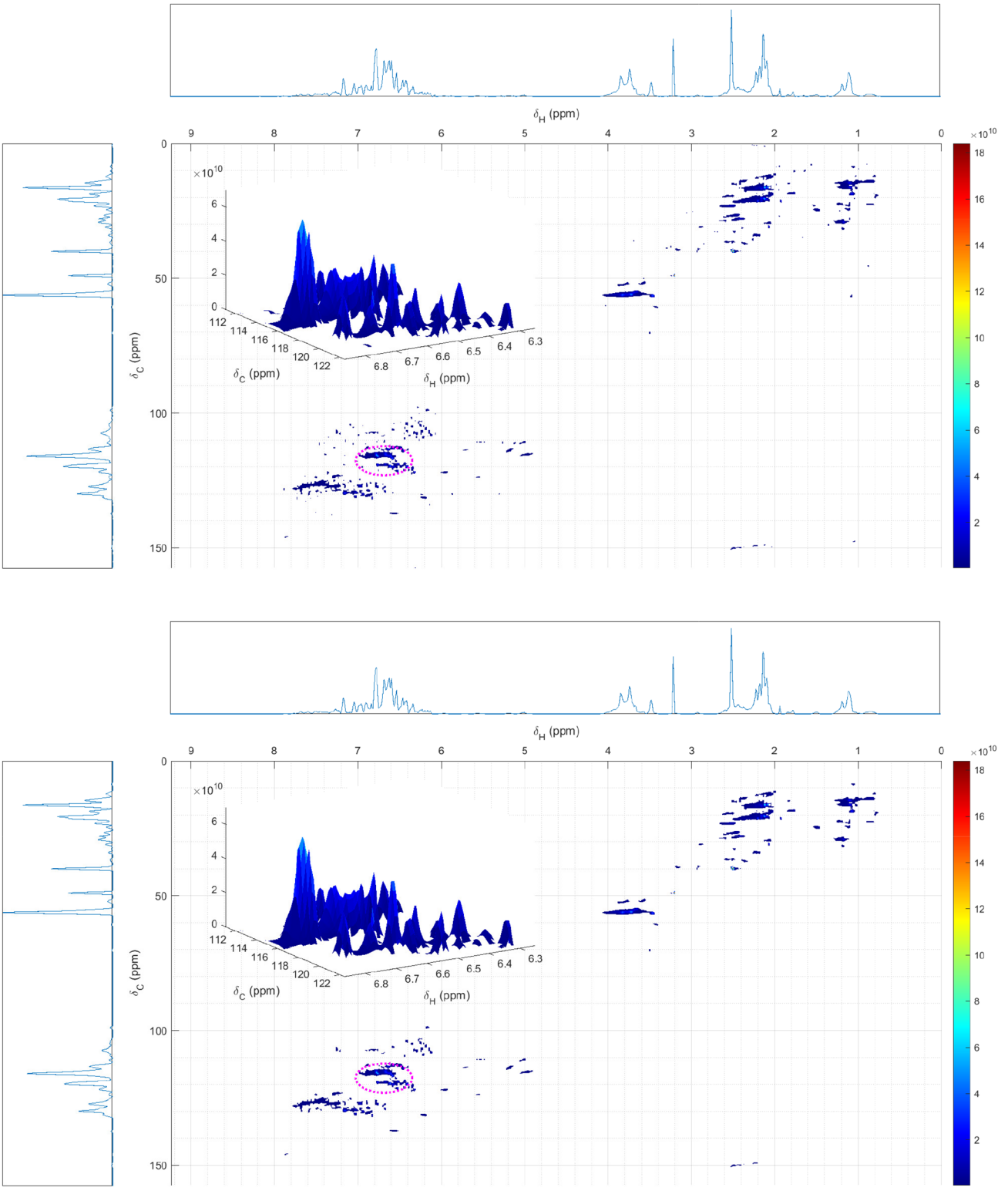


Figure 4. (Top) $^{1}H-^{13}C$ HSQC NMR spectrum from pyrolysis liquids from lignin, treated with $Ca(OH)_2$, subjected to adaptive thresholding using a Gaussian-weighted mean. (Bottom) Intersect of a median-filtered spectrum and the above adaptive-thresholded spectrum.

NMR of metabolite mixtures, according to the substructures that it would contain. However, the application of image processing tools to recognize and extract individual peak volumes for multivariate analysis has not been assessed before, to the best of our knowledge. Also, the use of very complex analytes rather than model mixtures is new to this context. However, the complex 2D NMR that results from herein used samples is expected to really challenge the image processing tools and, thus, assess their use for these liquids.

Classically, a $^{1}H-^{13}C$ HSQC NMR spectrum is be regarded as a data matrix, with the rows and columns being the carbon chemical shift (δ_C) and the hydrogen chemical shift (δ_H),

https://doi.org/10.1021/acs.energyfuels.2c04100
Energy Fuels 2023, 37, 4446−4459

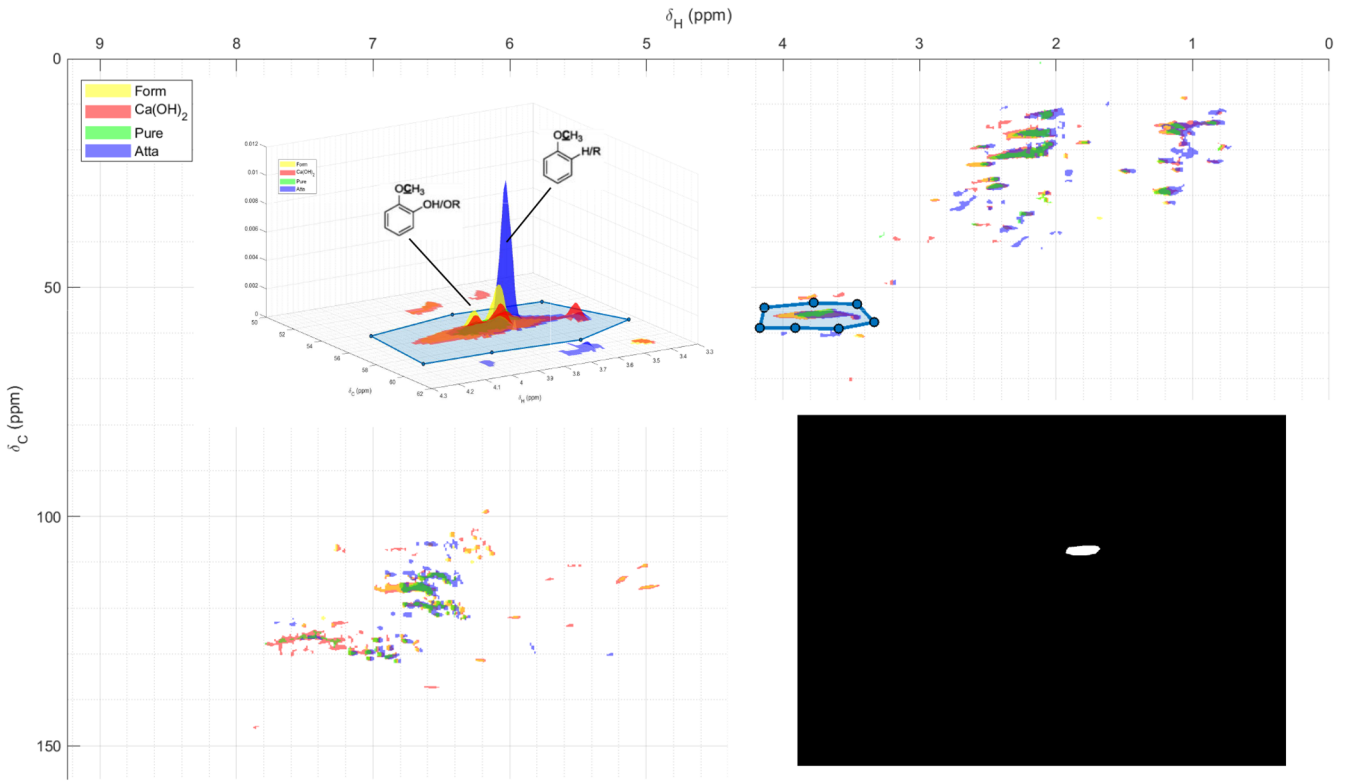


Figure 5. Overlay of the four investigated spectra in this work. They are derived from the heavy liquids from lignin pyrolysis, using no pretreatment ("Pure"), calcium hydroxide ["$Ca(OH)_2$"], sodium formate ("Form"), or attapulgite clay ("Atta") as pretreatment agents. The region of interest is magnified and embedded in this overlay as well as the binary mask for this region of interest.

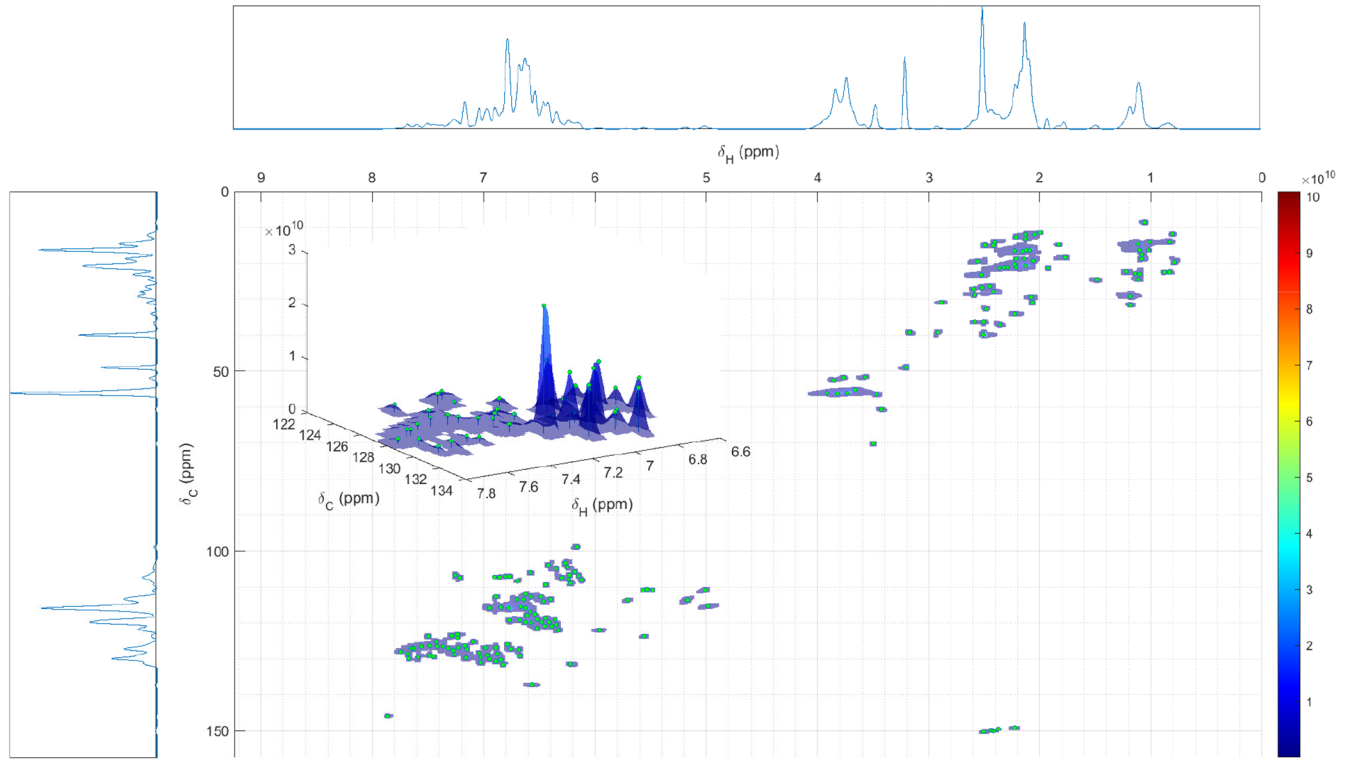


Figure 6. Regional maxima of the $^{1}H-^{13}C$ HSQC spectrum from lignin, treated with $Ca(OH)_2$, after applying a Gaussian filter.

respectively. The values in each matrix element then represent the recorded intensity. Alternatively, this data matrix can be regarded as a gray-scale image once the intensities are rescaled between 0 and 1. Each matrix element then represents a pixel

https://doi.org/10.1021/acs.energyfuels.2c04100
Energy Fuels 2023, 37, 4446−4459

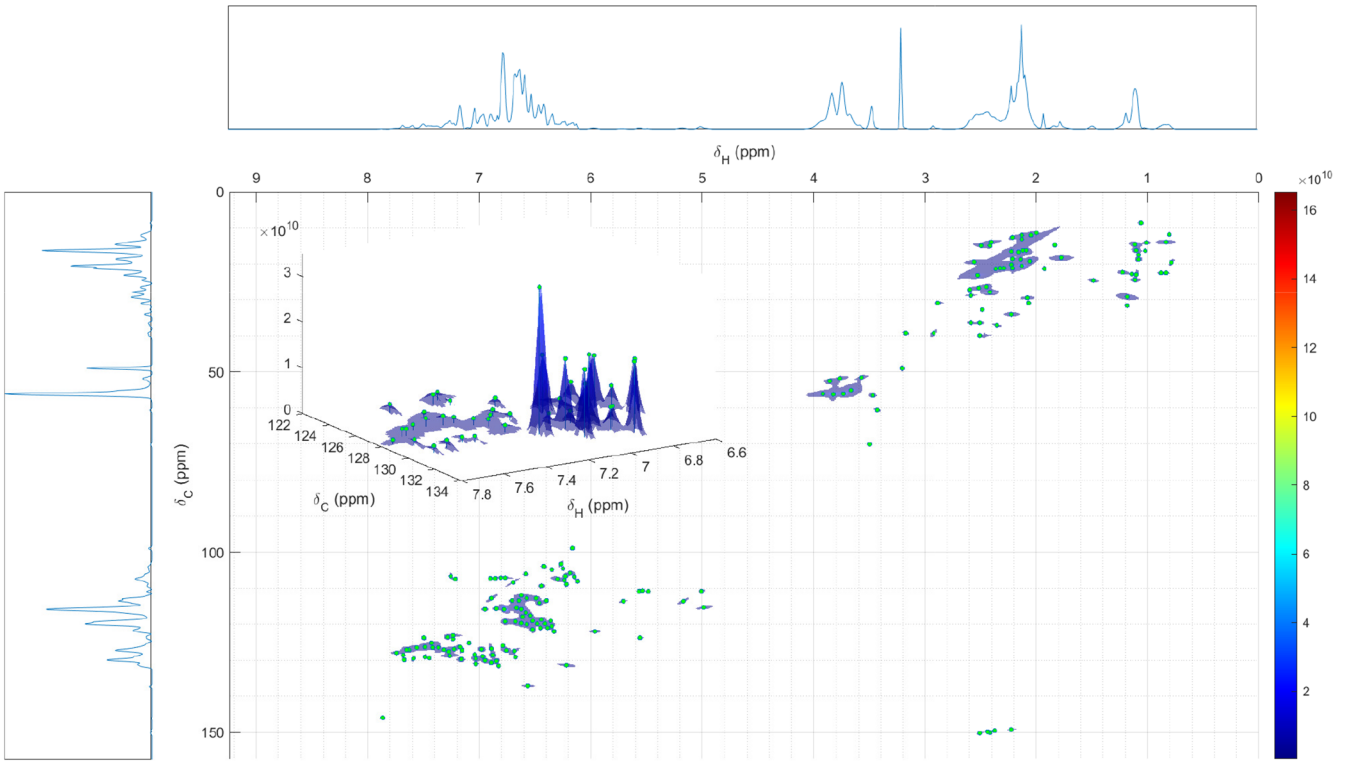


Figure 7. Modeled spectrum of heavy liquids from lignin, treated with $Ca(OH)_2$, involving the fitting of pseudo-Voigt peaks.

of a picture with a black (0), gray (between 0 and 1), or white (1) color. Figure 1 shows the classical representation of a HSQC spectrum, read by commercial Bruker TopSpin 4.0.7 software. Embedded in this figure is the gray-scale representation of the methoxy region. This insight opens a new avenue to explore image processing tools, like image segmentation (dividing images into constituent parts) and object detection (assigning labels to parts of the image), among others, and apply them for $^1H-^{13}C$ HSQC NMR analysis. The goals of this work were, thus, to investigate (i) if and how image processing tools can accommodate quantitative NMR analysis and (ii) whether image processing tools can be the cornerstone for peak-fitting algorithms and multivariate analysis.

METHOD DEVELOPMENT AND DISCUSSION

Recording of the HSQC Spectra of the Samples. Lignin and its derived pyrolysis liquids are known to be complex in nature and are, therefore, suitable samples to explore image-processing-based analyses. The applied lignin was an organosolv lignin, from which four different pyrolysis liquids were obtained using different pretreatment agents, as described by Ghysels et al.[1] The lignin stemmed from the organolsolv pulping process from Fraunhofer-Gesellschaft in Germany.[21] Lignin treatment with attapulgite clay encompassed the mixing of a 50/50 wt % mixture of clay/lignin with water, followed by filtration and drying. Wet mixing of calcium hydroxide with lignin and subsequent filtering and drying gave rise to a lignin to which calcium was bound. Sodium formate and lignin were also wet-mixed but not filtered; the mixture was dried to withhold sodium formate within the lignin. After fast pyrolysis of the (treated) lignins, a light aqueous phase and a heavy organic phase are obtained, which were subjected to $^1H-^{13}C$ HSQC NMR analyses and of which the heavy phase spectra were used in this work. The names of the samples, being "pure", "formate", "$Ca(OH)_2$", and "attapulgite", signify the applied pretreatment agents, which were sodium formate, calcium hydroxide, and attapulgite clay, respectively.[1]

$^1H-^{13}C$ HSQC NMR spectroscopy was performed using a Bruker Avance NEO 600, with a 600 MHz (14.1 T) UltraShield Magnet. The Bruker CryoProbe Prodigy was used, heaving a diameter of 5 mm and being at a temperature of 298 K for analysis. The used pulse program was hsqcetgpsi2. The number of scans was 8, and the number of dummy scans was 16. The spectral width of the ^1H NMR spectra varied and was either 9.2572, 10.4144, or 13.0179 ppm. ^{13}C NMR spectra were acquired using a fixed spectral width of 165 ppm. These different spectral widths caused the dimensions of the extracted data matrix to vary slightly (*vide infra*). As a result, acquisition times for the ^1H NMR dimension differed (0.0921600, 0.0819200, and 0.1310720 s) while that of the ^{13}C dimension remained the same (0.0051404 s). With regard to sample analysis, approximately 0.3 g of the heavy liquids was dissolved in 1.4 g of deuterated dimethyl sulfoxide (DMSO-d_6).

Table 1. R^2 of the Modeled Representations for the Spectra of Pyrolysis Liquids

origin of the heavy pyrolysis liquids, with recorded NMR spectra	coefficient of determination for the modeled peak representations, R^2
lignin treated with sodium formate	0.92
lignin treated with $Ca(OH)_2$	0.93
untreated lignin	0.96
lignin treated with attapulgite	0.94

https://doi.org/10.1021/acs.energyfuels.2c04100
Energy Fuels 2023, 37, 4446−4459

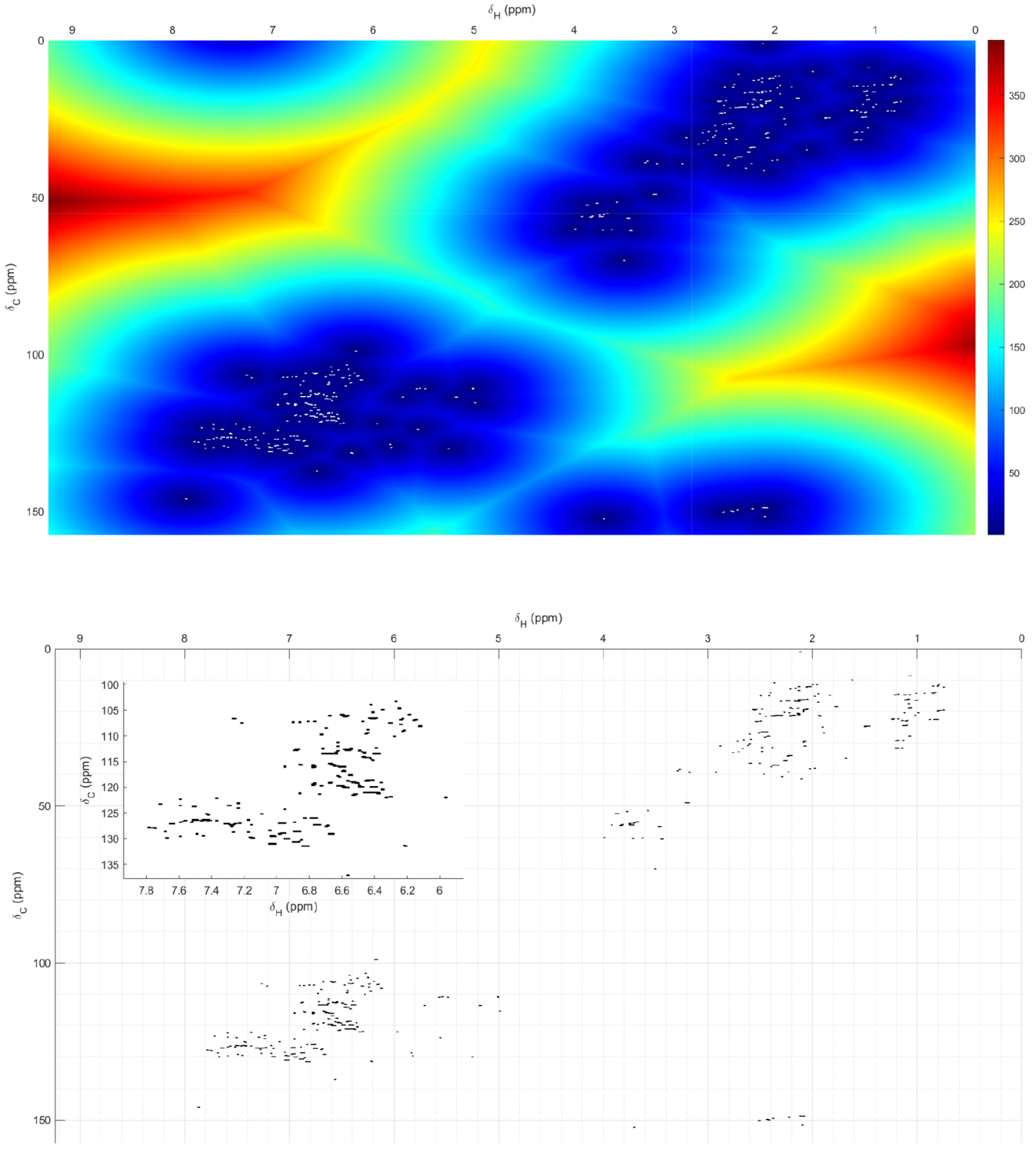


Figure 8. (Top) Distance transform of a matrix, containing the peak positions of the fitted pseudo-Voigt peaks of all four spectra. The color bar denotes the distance (in pixels) from a certain pixel to the position of the nearest fitted peak. (Bottom) Small aggregate of pixels that denote single peaks (or one chemical environment/moiety).

Data Import and Homogenization. The $^{1}H-^{13}C$ HSQC NMR spectra were first opened with Bruker TopSpin 4.0.7 software, from which a text file was exported that contained the intensities for every (δ_H, δ_C) couple. Those text files were imported to MATLAB R2021b and converted into a matrix. Although the number of rows was constant (1024), the number of columns (^{1}H dimension) varied among the different spectra (either 911, 1024, or 1457) as a result of the different spectral widths. This is equivalent to pictures having a different number of pixels. Further processing of the different spectra required matrices of the same size. Hence, the matrices were resized, using bicubic interpolation. That is, the output pixel

https://doi.org/10.1021/acs.energyfuels.2c04100
Energy Fuels 2023, 37, 4446−4459

Table 2. Fragment from a Peak List, Showing the Intensities Underneath the Peak(s) at Specific δ_H and δ_C, for All Four Samples

			relative peak volume (vol %)				
number	δ_H (ppm)	δ_C (ppm)	sodium formate	$Ca(OH)_2$	pure	attapulgite	annotation
106	3.72	55.6	1.02×10^{-1}	7.62×10^{-2}	3.61×10^{-2}	2.53×10^{-1}	methoxy
107	3.72	60.4	0	0	0	1.27×10^{-3}	methoxy
108	3.71	152.3	0	0	0	6.26×10^{-4}	not assigned
109	3.62	60.3	0	0	0	1.08×10^{-2}	methoxy
110	3.57	51.6	0	4.18×10^{-3}	0	0	methoxy
111	3.50	70.1	0	9.18×10^{-4}	0	0	methoxy
112	3.46	56.6	0	1.67×10^{-2}	0	8.05×10^{-3}	methoxy
113	3.43	60.6	3.98×10^{-3}	1.33×10^{-3}	0	0	methoxy

value is a weighted average of pixels in the nearest 4 × 4 neighborhood. This was implemented with the built-in MATLAB function imresize. This contrasts with bilinear interpolation, applied by Arbogast et al.,[22] where spectral matrices were resized to have the same size and chemical shift ranges as the first spectrum in the series.

Adaptive Thresholding. The $^1H-^{13}C$ HSQC NMR spectra contain many low-intensity values, which may be considered as background noise. This is apparent from the histogram in Figure 2 and the many dark pixels in the embedded gray-scale image representation in Figure 1. A previously reported method to discriminate the signal from noise in $^1H-^{13}C$ HSQC NMR spectra, prior to multivariate analysis, involves mean centering.[23] This comes down to subtracting every spectral intensity with the average intensity of the entire $^1H-^{13}C$ HSQC NMR spectrum. The average intensity of the spectrum behind Figure 2 is 5.16×10^7. The top spectrum in Figure 3 shows the raw $^1H-^{13}C$ HSQC NMR spectrum of the heavy pyrolysis liquids from lignin treated with $Ca(OH)_2$. Below is the same spectrum after mean centering. A disadvantage of mean centering is that one single value is used to discriminate the signal from noise. A direct consequence is that the noise-corrected spectrum contains large patches, of which a large part visually appears as noise (bottom panel of Figure 3). This is due to an elevated background noise for zones that contain many signal peaks. Similarly, Hedenström et al.[24] accounted for noise by thresholding against the average intensity of a "noise region". This gives a bit more flexibility but is equally *ad hoc* as mean centering.

One useful tool in image processing is thresholding, which aims to separate bright objects in a gray-scale image from a dark background. A couple of possibilities exist and are demonstrated in Figure 2, using a gray-scale image of rice as an illustrative example. Like mean centering, global thresholding applies a single threshold. This threshold value is deduced from Otsu's method,[25] which minimizes intraclass variance of the thresholded black and white pixels. Other techniques proposed in the literature, like taking the average signal from a noise window or the mean of the entire spectrum, have been tested.[24] A herein obtained new threshold value was also tested, by plotting the entirety of intensities against a normal distribution. The intensity of the data that started deviating from the normal distribution (assumed noise) was used as threshold value. However, visual inspection of the spectra however showed that global thresholding, regardless of how the threshold was deduced, was ineffective. This is also visualized in Figure 2, where some rice grains were incorrectly regarded as background. Adaptive thresholding applies variable threshold values for different locations in the picture based on the local mean, median, or Gaussian-weighted mean intensity in the neighborhood. From visual inspection of the spectra, adaptive Gaussian thresholding was recognized as most suitable, as also illustrated in Figure 2. Within a pixel window of 57 × 51 pixels, Gaussian-weighted mean intensities were calculated, which resulted in spatially different threshold values. Then, these local thresholds were used to binarize the spectra and, thus, set all background signals to zero. The built-in MATLAB functions adaptthresh and imbinarize allowed for elegant implementation of this procedure. The result of this thresholding is shown in the top panel of Figure 4. The difference between noise correction based on mean centering versus adaptive thresholding is clear. Adaptive thresholding recognizes the signal from noise more precisely, unlike mean centering (the bottom panel of Figure 3 versus the top panel of Figure 4).

Salt and Pepper Filtering. To obtain a spectrum with the least amount of noise-related peaks, an additional step was implemented, known as "salt and pepper" filtering. Indeed, the top panel of Figure 4 shows a remarkable number of small peaks with low intensity, which were not removed by the adaptive thresholding step. The binarized rice image in Figure 2 also shows a remaining number of small bright pixels, which are noise-related. Removing that small salt and pepper-like noise is a known practice in image processing and can be achieved by a median filter. With this technique, each pixel was replaced by the median intensity of the 6 × 6 neighborhood around that pixel, which is a relatively easy approach; in MATLAB, medfilt2 was used to this end. Figure S1 in the Supporting Information shows the result from median filtering of the top panel of Figure 4. Visual inspection of Figure S1 in the Supporting Information confirms the removal of small-area and low-intensity spots. However, edge distortion during median filtering also occurred. To combine the desired noise-removing capacity of median filtering and preserve the edges of signal-related intensities, the intersect of the median-filtered spectrum (Figure S1 in the Supporting Information) and adaptive-thresholded spectrum (the top panel of Figure 4) was taken. This intersect is shown in the bottom panel of Figure 4. Indeed, small-area and low-intensity signals were removed while retaining the shape of the remaining signal intensities.

Normalization. After noise correction with the help of image processing tools, analysis of the spectra follows. Two methods are put forth, with the second being more automated than the first. Initially, an *ad hoc* generation of spectral regions of interest is outlined. Because these regions can be quite spacious (*vide infra*), individual peaks were fitted by 2D pseudo-Voigt functions, to autogenerate and compare a list of small (even individual) peaks among the spectra. For any of

https://doi.org/10.1021/acs.energyfuels.2c04100
Energy Fuels 2023, 37, 4446−4459

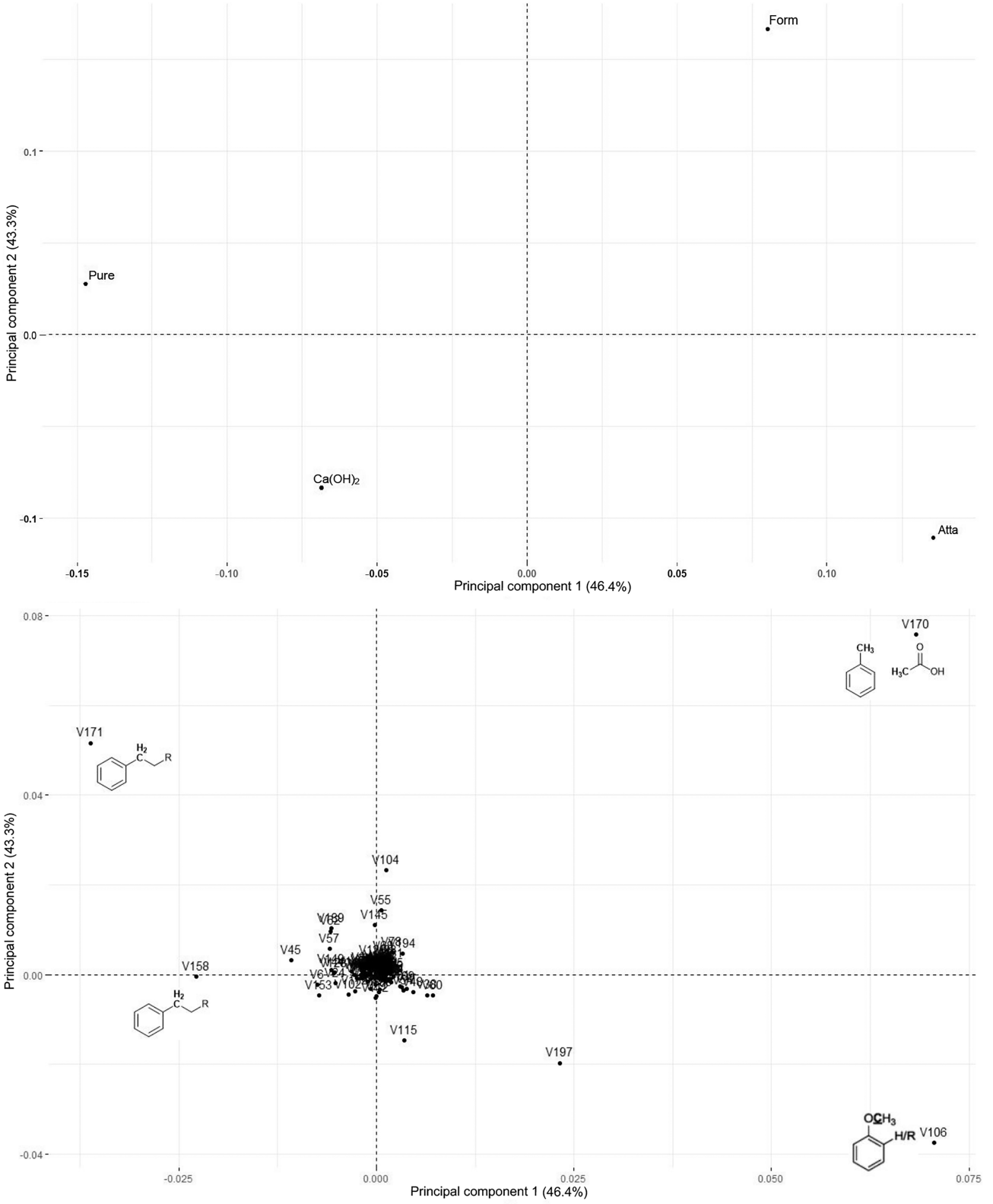


Figure 9. (Top) Position of individuals (samples) on a 2D plane, with the two principal components as axes. (Bottom) Location of the variables (peak intensities) on the PC1−PC2 plane.

these steps, normalization of all individual spectra was required. That is, the peak intensities for DMSO (the solvent) were removed, and the remaining peak intensities, solely stemming from the sample, were normalized so that the total spectral intensity equals unity.[19]

***Ad Hoc* Comparison of User-Defined Regions of Interest.** Dependent upon the intention of the analyses,

https://doi.org/10.1021/acs.energyfuels.2c04100
Energy Fuels 2023, 37, 4446−4459

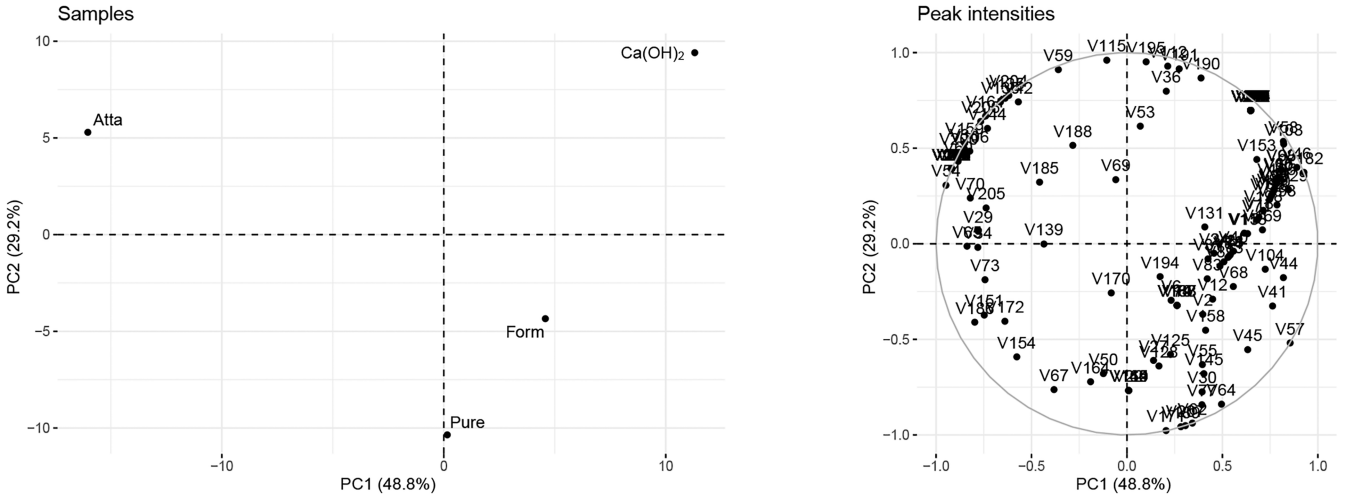


Figure 10. (Left) Position of individuals (samples) on a 2D plane, with the two principal components as axes. (Right) Location of the variables (peak intensities) on the PC1−PC2 plane.

there may be an interest to compare the total intensity of signals by a specific δ_H−δ_C zone. To do so, an overlay representation was established. This overlay representation visualizes each spectrum in a different color and highlights differences (Figure 5). From this overlay, specific regions of interest were identified and extracted. For instance, all spectra show methoxy moieties in the region[11] δ_H = 3.6−4.1 ppm and δ_C = 54−57 ppm. A very useful image processing tool to obtain the cumulative intensity of a user-defined area is by the creation of a so-called mask image. This mask image is binary, having zeroes (black) for masked image fragments and ones (white) for unmasked image fragments (see the inset in Figure 5). The unmasked image fragment, here the methoxy region (indicated by the blue connected dots in Figure 5), was created by drawing a polygon, using the built-in function drawpolygon and a computer mouse or stylus. Applying this mask over the noise-corrected ^{1}H−^{13}C HSQC NMR spectra sets all intensities behind the mask to zero. Then, the cumulative intensity within the remaining region of interest can be calculated and compared.

For instance and visually in accordance to the enlarged methoxy region in Figure 5, the cumulative intensity in the methoxy area was the largest (0.21 vol %) for the attapulgite-pretreated sample ("Atta"), while that of untreated lignin ("Pure") was the smallest (0.05 vol %). The relative intensities of the entire methoxy region for sodium-formate-pretreated lignin ("Form") and $Ca(OH)_2$-pretreated lignin "$Ca(OH)_2$" were similar (0.16 and 0.13 vol %, respectively).

Automated Fitting and Analysis of Individual Spectral Peaks. *Ad hoc* analysis is a valid and currently much used practice. However, it also is time-consuming and dependemt upon the executing scientist. Therefore, a more automated approach is desired. Here, individual peaks in the spectra were fitted to allow for the autogeneration of a peak list for multivariate analysis.

Finding Peak Positions. A first step was to find individual peaks within the spectra. That is, to identify pixel(s) that are surrounded by other pixels having lower intensity. However, two closely neighboring peaks do not necessarily represent different chemical functionalities and, thus, do not require separate peak fitting. Hence, a smoothing procedure was first performed. This is analogous to blurring a picture, which can be achieved by various image filters, of which a Gaussian filter proved effective for this purpose. A Gaussian filter with a standard deviation of 1 was applied to all ^{1}H−^{13}C HSQC spectra. This resulted in a reduced set of regional maxima (180 green dots in the bottom panel of Figure 6). Note that filtering, as expected, changed the shape of the blurred spectra. However, only the positions of regional maxima of the blurred image were used for subsequent analysis. The regional maxima were found using the built-in function imregionalmax, and Gaussian filtering was performed using the built-in function imgaussfilt.

Pseudo-Voigt Functions. Pseudo-Voigt peaks were then fitted, using the noise-corrected spectra at the positions where regional maxima were detected for the blurred spectrum. The aim was to resemble the original spectrum as close a possible and to obtain a peak list with intensities.

Pseudo-Voigt peak shapes can be viewed as a convolution of a Gaussian and Lorentzian peak shape. These peak shapes are particularly fit to resemble spectral peaks from NMR.[26] Indeed, the natural line shapes from NMR have a Lorentzian profile, while, e.g., apodization using commonly used window functions, such as the 2π-Kaiser window function,[27] adds a Gaussian component. Because 2D Lorentzian functions, 2D pseudo-Voigt functions, and peak-fitting procedures for 2D ^{1}H−^{13}C HSQC are generally not described in detail, an elaborate description is given below. Formally written, the pseudo-Voigt representation $\hat{y}$ of a peak i at a certain location (δ_H, δ_C) is

$$\hat{y}_i(\delta_H, \delta_C, \theta) = fG(\delta_H, \delta_C, \theta_G) + (1 - f)L(\delta_H, \delta_C, \theta_L) \quad (1)$$

where f gives the weight to how much the pseudo-Voigt function acquires a Gaussian or Lorentzian character, θ is the parameter for the pseudo-Voigt function and comprises of parameters for the Gaussian and Lorenztian components, with θ_G entailing the parameters for the Gaussian component (A, μ_{δ_H}, μ_{δ_C}, σ_{G,δ_H}, σ_{G,δ_C}, and ρ) and θ_L enclosing the parameters for the Lorentzian component (A, μ_{δ_H}, μ_{δ_C}, σ_{L,δ_H}, σ_{L,δ_C}, and ρ). The Gaussian component is shown in eq 2.

https://doi.org/10.1021/acs.energyfuels.2c04100
Energy Fuels 2023, 37, 4446−4459

$$G(\delta_H, \delta_C, \theta_G) = A \exp\left[\frac{-1}{2(1-\rho^2)}\left(\left(\frac{\delta_H - \mu_{\delta_H}}{\sigma_{G,\delta_H}}\right)^2 - 2\rho\left(\frac{\delta_H - \mu_{\delta_H}}{\sigma_{G,\delta_H}}\right)\left(\frac{\delta_C - \mu_{\delta_C}}{\sigma_{G,\delta_C}}\right) + \left(\frac{\delta_C - \mu_{\delta_C}}{\sigma_{G,\delta_C}}\right)^2\right)\right] \quad (2)$$

Equation 2 represents a Gaussian peak with a regional maximum at μ_H, μ_C, with an amplitude A, a standard deviation along the δ_H axis of σ_{G,δ_H}, a standard deviation along the δ_C axis of σ_{G,δ_C}, and ρ as the correlation between the δ_H axis and the δ_C axis (i.e., the rotation). The 2D Lorentzian component in eq 3 was derived from the Student distribution with 1 degree of freedom, for $k = 2$, being the number of dimensions.

$$L(\delta_H, \delta_C, \theta_L) = A\frac{1}{\left[1 + \left(\frac{\delta_H - \mu_{\delta_H}}{\sigma_{L,\delta_H}}\right)^2 - 2\rho\left(\frac{\delta_H - \mu_{\delta_H}}{\sigma_{L,\delta_H}}\right)\left(\frac{\delta_C - \mu_{\delta_C}}{\sigma_{L,\delta_C}}\right) + \left(\frac{\delta_C - \mu_{\delta_C}}{\sigma_{L,\delta_C}}\right)^2\right]^{3/2}} \quad (3)$$

Peak Fitting and Estimation of Parameters. For every regional maximum at the bottom panel of Figure 6, a pseudo-Voigt function was fitted. The parameter values were sought that minimized the objective function $J(\theta)$, with θ being the combination of θ_G and θ_L, along with the mixing parameter f. The objective function was the sum of squared errors between the measured intensity and the modeled intensity.

$$J(\theta) = \sum_i^n (\hat{y}_i - y_i)^2 \quad (4)$$

The spectra of heavy liquids from lignin treated with $Ca(OH)_2$ showed 179 peaks to be fitted (180 minus the DMSO peak), which translates to the optimization of $179 \times 9 = 1611$ parameters (A, δ_H, δ_C, σ_{G,δ_H}, σ_{G,δ_C}, σ_{L,δ_H}, σ_{L,δ_C}, ρ, and f). The simultaneous optimization of all 1611 parameters at once is computationally challenging but also not necessary. Indeed, changing parameter values to fit a certain peak in the aromatic region has virtually no effect on the shape of a fitted peak further away in the spectrum (e.g., in the aliphatic region). On the other hand, the subsequent one-by-one fitting of single peaks is also sub-optimal for overlapping neighboring signal peaks. We therefore considered one region of connected components (the top panel of Figure 6) at a time in which the (overlapping) signal peaks were fitted. This method is an intuitive optimum for computational speed and accuracy. The minimization problem in eq 4 was solved using fmincon, which uses an interior-point algorithm.

Eventually, for all four spectra, modeled peak representations were made based on a number of pseudo-Voigt peaks. The modeled representation for the spectrum of pyrolysis liquids from lignin, treated with $Ca(OH)_2$, is shown in Figure 7. As a measure of similarity between the real and modeled peak representations, coefficients of determination were calculated, as presented in Table 1. The equally high values in Table 1 indicate that the followed approach did result in good modeled representations, of equal quality for all spectra.

Generation of a Peak List. From each of the four spectra, a peak list was extracted, which contains the positions of its fitted peaks and the cumulative intensities thereof (i.e., those underneath the curves). Nevertheless, the creation of a general peak list is not simply the combination of the four individual peak lists. This is because peak positions do not necessarily precisely overlap the same positions. Instead, it was observed that the positions of fitted peaks formed dense clusters. The task to cross-identify similar peaks is called "peak matching" by Li et al.[28] While the latter authors developed an algorithm based on peak data, we herein used the image-processing-based distance transform, which is fairly easy implemented. The white pixels in the top panel of Figure 8 are those positions where a peak was fitted in either of the four spectra. The blue to red colors and associated values as shown in the color bar signify the distance from any position (δ_H, δ_C) to the nearest white pixel in the top panel of Figure 8. This visual representation is obtained using the distance transform (bwdist). This plot was further used to decide which peak positions belong to the same cluster (i.e., to the same chemical environment in the spectrum).

First, a Gaussian filter was applied to the rescaled distance transform, after which the image was binarized to result in small isolated patches, as shown in the bottom panel of Figure 8. Every fitted peak of which its peak maximum fell within the same collection of pixels in the bottom panel of Figure 8 was regarded to represent the same chemical moiety and was regarded as the same peak. If one spectrum had multiple peak maxima within one small aggregate of pixels, their corresponding cumulative intensities were summed. The result of this is a peak list, as partly presented in Table 2. The entire list is in Table S1 of the Supporting Information.

As a result, the methoxy region in Figure 5 is now deconvoluted into 13 individual peaks, of which those at δ_H = 3.83 ppm, δ_C = 56.1 and δ_H = 3.72 ppm, δ_C = 55.6 ppm govern the largest peak volume. The entire peak list consists of 208 peaks, of which 19 were attributed to primary alkyls, 45 were attributed to secondary alkyls, 4 were attributed to aliphatic OH, 3 were attributed to carbon at the α position of carbonyls, 4 were attributed to CH_2−Ar, 13 were attributed to CH_3−Ar, 13 were attributed to methoxys, 6 were attributed to unconjugated CH═CH, 91 were attributed to conjugated CH═CH (in aromatics), and 10 were not assigned. These data as partly presented in Table 2 allow for statistical analysis, because they represent narrow chemical moieties rather than encompassing a large area in the HSQC spectrum with different chemical functionalities.

Principal Component Analysis (PCA). The peak list as (partly) presented in Table 2 is a multivariate data set. These data were used for PCA, which is a recognized tool for chemometric analysis of spectra.[22,24,29] Often, PCA is used for pattern recognition and dimension reduction, among others. The dimensions are the 208 spectral peaks (in Table S1 of the Supporting Information). These dimensions are reduced to often two dimensions. These new dimensions are linear combinations of the original dimensions and are composed in such a way that the largest variance is retained in the first dimension and the second largest variance is retained in the second dimension. This allows for the plotting of high-dimensional data on a 2D plane with the two principal components as axes, with minimal loss of data variance. The peak list (i.e., Table S1 of the Supporting Information) was imported into RStudio, version 4.1.1. The packages FactoMineR and factoextra were used to perform and plot PCA.

Prior to performing PCA, the peak intensities can either be scaled or not. This implies that each row in Table 2 is standardized to have an average of 0 and a standard deviation

https://doi.org/10.1021/acs.energyfuels.2c04100
Energy Fuels 2023, 37, 4446−4459

of 1. In practice, scaling is required if the variables (here, the relative peak volumes) have different units and, hence, different orders of magnitude. As in this case, all units were the same (being volume percent) and scaled PCA and unscaled PCA were performed.[29] In the former, PCA will pinpoint the peaks that have a high intensity for certain spectra, while the latter rather shows differences within the spectra, regardless their intensity absolute values.

Figure 9 shows the result from unscaled PCA. In general, PCA using two dimensions explained ca. 90% of variance in the data. This was rather expected, given that four spectra were compared. More interestingly, however, is the analysis of the variables. This indeed allows for pinpointing which specific peaks characterize a certain pyrolysis liquid type. PCA with unscaled data highlights that pyrolysis liquids from attapulgite-pretreated lignin and sodium formate to a lesser extent are strongly and positively correlated to compound "V106", i.e., the relative intensity for peak 106 in Table S1 of the Supporting Information. With consultation of this peak list, it is learned that V106 corresponds to a peak in the methoxy region at δ_H = 3.72 ppm and δ_C = 55.6 ppm. This neatly confirms the previous findings, e.g., Figure 5, showing a dominant methoxy peak for attapulgite-pretreated lignin. On the other hand, the unpretreated lignin is strongly and negatively correlated to V106, meaning that it has the lowest methoxy content.

The formic-acid-pretreated sample is associated with large relative peak volumes for peak "V170", which is the peak with a maximum at δ_H = 2.07 ppm and δ_C = 19.49 ppm and falls within the aliphatic region. This peak could represent a combination of both α-methyl of acetic acid ($\mathbf{CH_3}$COOH) and/or the CH_3−aryl group. The acid content, determined in a previous work by Ghysels et al.,[1] could confirm the latter. The pyrolysis liquids from untreated and $CaOH_2$-pretreated lignin contained higher peak volumes for an alkyl peak ("V158") at δ_H = 2.16 ppm and δ_C = 21.1 ppm. This peak can be attributed, like "V171", to a R−CH_2−Ar function.

Figure 10 presents the results from scaled PCA. It is apparent that the scaled PCA and unscaled PCA result in approximately the same covered data variance. Moreover, it is immediately visible that scaling the peak volumes results in many more discernible signals, which contribute to differentiate the different samples. The two highest contributors to PC1 were peaks 54 and 182. Peak 54 corresponds to an aromatic peak (δ_H = 6.68 ppm and δ_C = 119.1 ppm), is located at the negative end of PC1 on the variable plot, and hence, is most characteristic for attapulgite-pretreated lignin and least characteristic for $Ca(OH)_2$-pretreated lignin. This can be verified from the peak list in Table S1 of the Supporting Information. The volume covered by peak 54 is 2.36 vol % for attapulgite-pretreated lignin, while being 0.97 vol % for $Ca(OH)_2$-pretreated lignin. Peak 182 corresponds to an (secondary) aliphatic peak (δ_H = 1.77 ppm and δ_C = 18.4 ppm), is at the positive end of PC1 on the variable plot, and hence, is least characteristic for attapulgite-pretreated lignin and most characteristic for $Ca(OH)_2$-pretreated lignin. The volume covered by peak 182 is 1.28 vol % $Ca(OH)_2$-pretreated lignin while 0.13 vol % for attapulgite-pretreated lignin. It is, thus, clear that PCA can render analysis of the major dominant peaks and also comparatively analyze all peaks, regardless of their dominance in the spectrum.

RELEVANCE AND OUTLOOK

In this work, a combination of image processing, peak fitting, and multivariate analysis of peaks was explored, using $^1H-^{13}C$ HSQC spectra from a previous study[1] that were not quantitative. However, it is important to mention that the exact same method as developed and shown in this work is directly applicable for quantitatively recorded spectra, using, e.g., the earlier mentioned extrapolated time-zero technique and adding an internal standard. From the obtained peak list, this study applied PCA to assess which peaks are characteristic for certain samples and which are not, similar to that by Li et al.[28] However, upon the further development of a spectral database, the peak list, generated in this work, could be subjected for compound identification or functionality pairing. Future work could, therefore, entail method development for database querying. Indeed, peaks within the list reflect H−C correlations rather than compounds. A PCA could be helpful in this step, because it could pinpoint to specific peak positions, which always increase/decrease in tandem and which could, therefore, stem from the same chemical compound. While this work focused on the generation of a peak list, using image processing software, advanced analysis thereof, e.g., using deep learning, is an opportune path that could follow this work. For instance, Kuhn et al.[20] applied convolutional neural networks as a classification technique to group spectra by fragments contained in the structure that they represent.

CONCLUSION

In conclusion, this work has demonstrated the large potential for image processing tools to aid in the automated analysis of complex HSQC spectra. These tools were deployed in various stages, from noise detection and removal, over the comparison of regions of interest, to the localization of spectral peaks to be fitted. Moreover, many of the analyses in this work were performed with built-in functions, which allow for an easy and practical implementation. However, other programming environments (like Python) are also well-equipped for this approach. Pseudo-Voigt functions were fitted to the different spectra, which led to the generation of a peak list. PCA with that peak list allowed for pinpointing interesting features of the spectra and highlighting mutual differences. Overall, image processing is an underexplored tool to analyze complex HSQC spectra, despite being easy to implement, elegant, and fast. The methods as described here can also be applied to other 2D spectra.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.energyfuels.2c04100.

(Top) $^1H-^{13}C$ HSQC NMR spectrum from pyrolysis liquids from lignin, treated with $Ca(OH)_2$, subjected to median filtering, using a 6 × 6 window and (Bottom) intersect of a median-filtered spectrum and the adaptive-thresholded spectrum (Figure S1) and peak list, showing the intensities underneath the peak(s) at specific δ_H and δ_C, for all four samples (Table S1) (PDF)

https://doi.org/10.1021/acs.energyfuels.2c04100
Energy Fuels 2023, 37, 4446−4459

AUTHOR INFORMATION

Corresponding Author

Stef Ghysels − *Thermochemical Conversion of Biomass Research Group (TCCB), Department of Green Chemistry and Technology, Ghent University, 9000 Ghent, Belgium;* orcid.org/0000-0002-6957-725X; Phone: +32-0-92646190; Email: stef.ghysels@ugent.be

Authors

Jan Verwaeren − *KERMIT Research Group, Department of Data Analysis and Mathematical Modelling, Faculty of Bioscience Engineering, Ghent University, 9000 Ghent, Belgium*

Hero Jan Heeres − *Department of Chemical Engineering (ENTEG), University of Groningen, 9747 AG Groningen, Netherlands;* orcid.org/0000-0002-1249-543X

Léon Rohrbach − *Department of Chemical Engineering (ENTEG), University of Groningen, 9747 AG Groningen, Netherlands*

Simon Backx − *SynBioC Research Group, Department of Green Chemistry and Technology, Faculty of Bioscience Engineering, Ghent University, 9000 Ghent, Belgium;* orcid.org/0000-0002-1136-0744

Sven Mangelinckx − *SynBioC Research Group, Department of Green Chemistry and Technology, Faculty of Bioscience Engineering, Ghent University, 9000 Ghent, Belgium;* orcid.org/0000-0002-9349-880X

Frederik Ronsse − *Thermochemical Conversion of Biomass Research Group (TCCB), Department of Green Chemistry and Technology, Ghent University, 9000 Ghent, Belgium;* orcid.org/0000-0002-3290-9177

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.energyfuels.2c04100

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Ghysels, S.; Dubuisson, B.; Pala, M.; Rohrbach, L.; Van den Bulcke, J.; Heeres, H. J.; Ronsse, F. Improving fast pyrolysis of lignin using three additives with different modes of action. *Green Chem.* **2020**, *22*, 6471−6488.

(2) Han, Y.; Paiva Pinheiro Pires, A.; Denson, M.; McDonald, A. G.; Garcia-Perez, M. Ternary Phase Diagram of Water/Bio-oil/Organic Solvent for Bio-oil Fractionation. *Energy Fuels* **2020**, *34*, 16250−16264.

(3) Pinheiro Pires, A. P.; Arauzo, J.; Fonts, I.; Domine, M. E.; Fernández Arroyo, A.; Garcia-Perez, M. E.; Montoya, J.; Chejne, F.; Pfromm, P.; Garcia-Perez, M. Challenges and Opportunities for Bio-oil Refining: A Review. *Energy Fuels* **2019**, *33*, 4683−4720.

(4) Staš, M.; Auersvald, M.; Vozka, P. Two-Dimensional Gas Chromatography Characterization of Pyrolysis Bio-oils: A Review. *Energy Fuels* **2021**, *35*, 8541−8557.

(5) Elliott, D. C.; Meier, D.; Oasmaa, A.; van de Beld, B.; Bridgwater, A. V.; Marklund, M. Results of the International Energy Agency Round Robin on Fast Pyrolysis Bio-oil Production. *Energy Fuels* **2017**, *31*, 5111−5119.

(6) Staš, M.; Auersvald, M.; Kejla, L.; Vrtiška, D.; Kroufek, J.; Kubička, D. Quantitative analysis of pyrolysis bio-oils: A review. *TrAC, Trends Anal. Chem.* **2020**, *126*, 115857.

(7) Dao Thi, H.; Van Aelst, K.; Van den Bosch, S.; Katahira, R.; Beckham, G. T.; Sels, B. F.; Van Geem, K. M. Identification and quantification of lignin monomers and oligomers from reductive catalytic fractionation of pine wood with GC × GC−FID/MS. *Green Chem.* **2022**, *24*, 191−206.

(8) Hao, N.; Ben, H.; Yoo, C. G.; Adhikari, S.; Ragauskas, A. J. Review of NMR Characterization of Pyrolysis Oils. *Energy Fuels* **2016**, *30*, 6863−6880.

(9) Mattsson, C.; Andersson, S.-I.; Belkheiri, T.; Amand, L.-E.; Olausson, L.; Vamling, L.; Theliander, H. Using 2D NMR to characterize the structure of the low and high molecular weight fractions of bio-oil obtained from LignoBoost™ kraft lignin depolymerized in subcritical water. *Biomass Bioenergy* **2016**, *95*, 364−377.

(10) Ren, X.; Wang, P.; Han, X.; Zhang, G.; Gu, J.; Ding, C.; Zheng, X.; Cao, F. Depolymerization of Lignin to Aromatics by Selectively Oxidizing Cleavage of C−C and C−O Bonds Using $CuCl_2$/Polybenzoxazine Catalysts at Room Temperature. *ACS Sustainable Chem. Eng.* **2017**, *5*, 6548−6556.

(11) Ben, H.; Ragauskas, A. J. Heteronuclear Single-Quantum Correlation-Nuclear Magnetic Resonance (HSQC-NMR) Fingerprint Analysis of Pyrolysis Oils. *Energy Fuels* **2011**, *25*, 5791−5801.

(12) Lyckeskog, H. N.; Mattsson, C.; Olausson, L.; Andersson, S.-I.; Vamling, L.; Theliander, H. Thermal stability of low and high Mw fractions of bio-oil derived from lignin conversion in subcritical water. *Biomass Convers. Biorefin.* **2017**, *7*, 401−414.

(13) Chen, W.; McClelland, D. J.; Azarpira, A.; Ralph, J.; Luo, Z.; Huber, G. W. Low temperature hydrogenation of pyrolytic lignin over Ru/TiO_2:2D HSQC and ^{13}C NMR study of reactants and products. *Green Chem.* **2016**, *18*, 271−281.

(14) Fardus-Reid, F.; Warren, J.; Le Gresley, A. Validating heteronuclear 2D quantitative NMR. *Anal. Methods* **2016**, *8*, 2013−2019.

(15) Hu, K.; Westler, W. M.; Markley, J. L. Simultaneous Quantification and Identification of Individual Chemicals in Metabolite Mixtures by Two-Dimensional Extrapolated Time-Zero $^{1}H-^{13}C$ HSQC (HSQC0). *J. Am. Chem. Soc.* **2011**, *133*, 1662−1665.

(16) Talebi Amiri, M.; Bertella, S.; Questell-Santiago, Y. M.; Luterbacher, J. S. Establishing lignin structure-upgradeability relationships using quantitative $^{1}H-^{13}C$ heteronuclear single quantum coherence nuclear magnetic resonance (HSQC-NMR) spectroscopy. *Chem. Sci.* **2019**, *10*, 8135−8142.

(17) Hu, K.; Ellinger, J. J.; Chylla, R. A.; Markley, J. L. Measurement of Absolute Concentrations of Individual Compounds in Metabolite Mixtures by Gradient-Selective Time-Zero $^{1}H-^{13}C$ HSQC with Two Concentration References and Fast Maximum Likelihood Reconstruction Analysis. *Anal. Chem.* **2011**, *83*, 9352−9360.

(18) Sette, M.; Lange, H.; Crestini, C. Quantitative HSQC analyses of lignin: a practical comparison. *Comput. Struct. Biotechnol. J.* **2013**, *6*, e201303016.

(19) Rios, S. M.; Barquín, M.; Nudelman, N. S. Characterization of oil complex hydrocarbon mixtures by HSQC-NMR spectroscopy and PCA. *J. Phys. Org. Chem.* **2014**, *27*, 352−357.

(20) Kuhn, S.; Tumer, E.; Colreavy-Donnelly, S.; Moreira Borges, R. A pilot study for fragment identification using 2D NMR and deep learning. *Magn. Reson. Chem.* **2022**, *60*, 1052−1060.

(21) Schulze, P.; Seidel-Morgenstern, A.; Lorenz, H.; Leschinsky, M.; Unkelbach, G. Advanced process for precipitation of lignin from ethanol organosolv spent liquors. *Bioresour. Technol.* **2016**, *199*, 128−134.

(22) Arbogast, L. W.; Delaglio, F.; Schiel, J. E.; Marino, J. P. Multivariate Analysis of Two-Dimensional ^{1}H, ^{13}C Methyl NMR Spectra of Monoclonal Antibody Therapeutics To Facilitate Assessment of Higher Order Structure. *Anal. Chem.* **2017**, *89*, 11839−11845.

(23) Rudd, T. R.; Mauri, L.; Marinozzi, M.; Stancanelli, E.; Yates, E. A.; Naggi, A.; Guerrini, M. Multivariate analysis applied to complex biological medicines. *Faraday Discuss.* **2019**, *218*, 303−316.

(24) Hedenström, M.; Wiklund-Lindström, S.; Öman, T.; Lu, F.; Gerber, L.; Schatz, P.; Sundberg, B.; Ralph, J. Identification of Lignin and Polysaccharide Modifications in Populus Wood by Chemometric Analysis of 2D NMR Spectra from Dissolved Cell Walls. *Mol. Plant* **2009**, *2*, 933−942.

(25) Otsu, N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Trans. Syst., Man, Cybern.* **1979**, *9*, 62−66.
(26) Happs, R. M.; Addison, B.; Doeppke, C.; Donohoe, B. S.; Davis, M. F.; Harman-Ware, A. E. Comparison of methodologies used to determine aromatic lignin unit ratios in lignocellulosic biomass. *Biotechnol. Biofuels* **2021**, *14*, 58.
(27) Li, D.-W.; Hansen, A. L.; Yuan, C.; Bruschweiler-Li, L.; Brüschweiler, R. DEEP picker is a deep neural network for accurate deconvolution of complex two-dimensional NMR spectra. *Nat. Commun.* **2021**, *12*, 5229.
(28) Li, D.-W.; Leggett, A.; Bruschweiler-Li, L.; Brüschweiler, R. COLMARq: AWeb Server for 2D NMR Peak Picking and Quantitative Comparative Analysis of Cohorts of Metabolomics Samples. *Anal. Chem.* **2022**, *94*, 8674−8682.
(29) Brinson, R. G.; Arbogast, L. W.; Marino, J. P.; Delaglio, F. Best Practices in Utilization of 2D-NMR Spectral Data as the Input for Chemometric Analysis in Biopharmaceutical Applications. *J. Chem. Inf. Model.* **2020**, *60*, 2339−2355.

https://doi.org/10.1021/acs.energyfuels.2c04100
Energy Fuels 2023, 37, 4446−4459