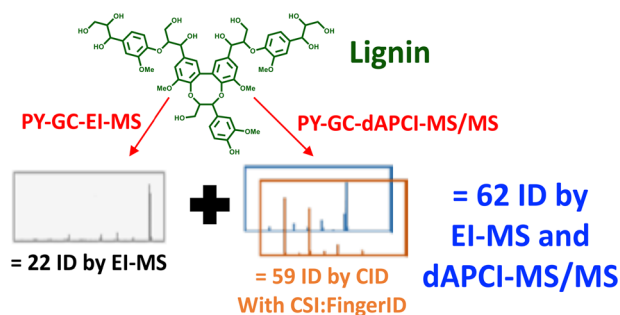RESEARCH ARTICLE

Gas Chromatography-Tandem Mass Spectrometry of Lignin Pyrolyzates with Dopant-Assisted Atmospheric Pressure Chemical Ionization and Molecular Structure Search with CSI:FingerID

Evan A. Larson, Carolyn P. Hutchinson, Young Jin Lee

Department of Chemistry, Iowa State University, 35A Roy J Carver Co-Lab, 1111 WOI Road, Ames, IA 50011-3650, USA

Lignin

PY-GC-EI-MS / PY-GC-dAPCI-MS/MS

= 22 ID by EI-MS

= 59 ID by CID
With CSI:FingerID

= 62 ID by EI-MS and dAPCI-MS/MS

Abstract. Dopant-assisted atmospheric pressure chemical ionization (dAPCI) is a soft ionization method rarely used for gas chromatography-mass spectrometry (GC-MS). The current study combines GC-dAPCI with tandem mass spectrometry (MS/MS) for analysis of a complex mixture such as lignin pyrolysis analysis. To identify the structures of volatile lignin pyrolysis products, collision-induced dissociation (CID) MS/MS using a quadrupole time-of-flight mass spectrometer (QTOFMS) and pseudo MS/MS through in-source collision-induced dissociation (ISCID) using a single stage TOFMS are utilized. To overcome the lack of MS/MS database, Compound Structure Identification (CSI):FingerID is used to interpret CID spectra and predict best matched structures from PubChem library. With this approach, a total of 59 compounds were positively identified in comparison to only 22 in NIST database search of GC-EI-MS dataset. This study demonstrates the effectiveness of GC-dAPCI-MS/MS to overcome the limitations of traditional GC-EI-MS analysis when EI-MS database is not sufficient.

Keywords: Lignin, Pyrolysis, Tandem mass spectrometry, Dopant assistance, Atmospheric pressure chemical ionization, GC-MS, CSI:FingerID, Molecular structure prediction

Introduction

Gas chromatography-mass spectrometry (GC-MS) analysis is commonly conducted by searching electron ionization (EI)-MS spectra against the NIST database [1]. The NIST database has over 200,000 entries accumulated in the last 40 years and allows for the identification of many volatile compounds. When a compound is not in the database, however, it cannot be identified by this approach. This becomes a serious problem for natural product or pyrolysis product analysis, with more than half the compounds often not identified.

Recently, GC-high-resolution mass spectrometry instruments (HRMS) with MS/MS capabilities have become commercially available to overcome such limitations including GC-quadrupole time-of-flight mass spectrometer (QTOFMS) [2] or GC-quadrupole Orbitrap [3]; however, this requires additional investment on specialized instruments. Additionally, these instruments use EI or chemical ionization (CI) as an ionization source, which is not as soft as atmospheric pressure ionization (API).

The use of API sources for GC-MS has a long history since the 1970s [4, 5], but has been rarely used until about 15 years ago when sensitive API sources became readily available for LC-MS [6]. The use of GC-APCI-QTOFMS has been reported for the environmental analysis of pesticides and other pollutants in water [7] and food [8, 9] as well as forensic drug studies [10]. It has also been shown to be comparable to LC-MS/MS for metabolite identification [11]. Studies have been made to

Correspondence to: Young Lee; *e-mail:* yjlee@iastate.edu

compare GC-EI-MS and GC-APCI-MS and demonstrated better performance of GC-APCI-MS identifying many more compounds [12, 13]. APCI in gas phase, however, produces significant fragmentation that is comparable to EI and has limitations for the analysis of fragile compounds [14].

Addition of a dopant in APCI has been used to reduce fragmentation and improve sensitivity in LC-MS or direct infusion analysis [15, 16]. However, the use of dopant-assisted APCI (dAPCI) in gas phase is very rare and has mainly been limited to fundamental studies with no extensive application [14, 17]. Previously, we have coupled pyrolysis (Py)-GC with a high-resolution TOFMS via dAPCI to study the pyrolysis products of carbohydrates [18]. Specifically, the use of dilute ammonia in helium as a dopant produces almost no fragmentation while enhancing ionization efficiencies through protonation and/or formation of an ammonium ion adduct. With this approach, we could study the effectiveness of in situ catalytic deoxygenation from the molecular composition of all the pyrolysis products. We also have successfully adopted this approach to study the reaction mechanism of glucose pyrolysis by tracing each pyrolyzate compound from isotopically labeled glucose [19]. Our previous work, however, determined only molecular formulae and could not confidently identify the compounds.

To address this previous shortcoming, in the current study, we added tandem mass spectrometry to Py-GC-dAPCI to obtain structural information of pyrolysis products and applied it to the analysis of Kraft lignin pyrolysis. In addition to true MS/MS using QTOFMS, we have also used in-source collision-induced dissociation (ISCID) to perform pseudo MS/MS on a single stage GC-dAPCI-TOF instrument. One critical limitation is the lack of MS/MS database, especially for volatile compounds. In order to identify compounds despite the lack of experimental MS/MS database, Compound Structure Identification (CSI):FingerID is adopted in the current study [20–22], which was one of the best automatic MS/MS analysis tools according to the 2016 Critical Assessment of Small Molecule Identification (CASMI) competition [23]. CSI:FingerID uses a machine learning algorithm that analyzes fragmentation trees to predict molecular fingerprints. A fragmentation tree is created from unknown MS/MS spectrum; then, predicted molecular fingerprints are compared to the molecular structure database (e.g., PubChem) to determine the best matches. This is the first time not only GC-dAPCI is coupled to tandem mass spectrometry, but also GC-MS/MS dataset of a complex mixture is systematically analyzed by computational model without MS/MS database.

In order to test the usefulness of our method, we have applied it to analyze a complex mixture of lignin pyrolysis products. Lignin is the most abundant aromatic biopolymer in the natural world and its better utilization is an important subject in bio-renewable research [24, 25]. Although lignin conversion is a difficult problem, there is currently a large effort to research better methods to valorize lignin and improve analytical techniques to characterize these methods. Pyrolysis of lignin produces a very complex mixture, of which the analysis is very challenging since many of them are not in the EI-MS or MS/MS database. Here, we demonstrate that our approach can effectively analyze this challenging system by complementing traditional GC-EI-MS analysis.

Experimental Methods

Materials

Kraft, low sulfonate lignin (Sigma Aldrich, St. Louis, MO) was used as the lignin feedstock for all experiments. This lignin was processed from a softwood source and is composed of mainly guaiacol-type monomer units originating from coniferyl alcohol. Commercially available standard compounds were purchased to confirm identifications with MS/MS and retention time. Vanillin and guaiacol were purchased from Alfa Aesar (Haverhill, MA) and 2-methoxy-4-methylphenol, 4-vinylphenol, syringol, eugenol, 4-propyl-2-methoxyphenol, isoeugenol, apocynin, and coniferyl aldehyde were purchased from Sigma Aldrich (Saint Louis, MO) and dihydroconiferyl alcohol was purchased from TCI America (Portland, OR).

Py-GC-dAPCI-TOF Instrumentation

A furnace type drop tube micropyrolyzer (PY-3030S; Frontier Laboratories, Fukushima, Japan) preheated at 500 °C was interfaced with a GC (7890A; Agilent Technologies, Santa Clara, CA). Deactivated stainless steel pyrolysis cups containing 150 µg of lignin powder were dropped into the micropyrolyzer furnace using an auto shot sampler (AS-1020E; Frontier Lab). The pyrolyzates were separated using an Agilent DB-5MS column. Oven programing was from 75 to 310 °C in 28 min (75 °C for 1 min, 20 °C min^{-1} to 150 °C, 7.5 °C min^{-1} to 310 °C, hold at 310 °C for 1.92 min), and the interface temperature was set at 310 °C. An Agilent GC-APCI source (G3212A) was used to interface the GC with an Agilent 6200 TOFMS for the ISCID experiments. Humidity was controlled in the GC-APCI source using a 1:1 mixture of dry nitrogen and humidified nitrogen with a total flow rate of 10 mL min^{-1}. Ammonia dopant (500 ppm in helium) was added to the source as a parallel sheath gas around column at a flow rate of 1 mL min^{-1} to improve ionization efficiency [18, 19]. Two independent Py-GC-MS runs were made at separate APCI source skimmer cone voltages (fragmentation parameters): 95 V to limit fragmentation and 200 V to achieve in-source CID. Data acquisition was made in positive mode for the *m/z* range of 40 to 1000.

Py-GC-dAPCI-QTOF Instrumentation

The same Py-GC-dAPCI setup as above was coupled to an Agilent 6540 QTOF to perform tandem mass spectrometry. The pyrolyzer, GC, and APCI source parameters were kept the same and the source fragmentor voltage of 95 V is used to minimize ISCID. The ammonia dopant was introduced to the source with the same concentration and flow rate as above. Humidity control was not made in this setup; however, the

result was barely affected due to sufficient ambient humidity. Data dependent MS/MS acquisition was used for the most abundant precursor ion with the collision energy calculated on-the-fly for each precursor using a formula, $0.03 \times m/z + 8$ (in eV). An exclusion mass list for common backgrounds was used to avoid unnecessary MS/MS acquisition. In addition to external calibration, post-calibration was performed using known peaks when necessary. Data acquisition was made in positive mode for the m/z range of 40 to 400.

Py-GC-EI-MS Instrumentation

A double shot drop tube micropyrolyzer (PY-3030D; Frontier Lab) was operated in single shot mode with the same parameters as above. The GC column and temperature parameters were kept the same as above. The GC was coupled to a Waters GCT mass spectrometer operated in electron ionization mode (70 eV) for the m/z range of 40 to 650.

Data Analysis

AMDIS software (NIST, Gaithersburg, MD) [26, 27] was used for deconvolution and database search of Py-GC-EI-MS data with 2012 NIST library. MassHunter (Agilent) was used for the data analysis obtained by Py-GC-dAPCI-TOFMS and Py-GC-dAPCI-QTOFMS. The precursor mass spectra from TOFMS and/or QTOFMS were used to determine the molecular formula. CSI:FingerID (v. 4.0 Build 15) was used to interpret CID spectra and search molecular fingerprints against the PubChem library (updated 08/28/17). Top ten best matching structures are inspected to avoid already identified compounds and choose most lignin like structure. Compounds with the matching score below − 100 is ignored to minimize false positive. Manual inspection was also made on a set of CID spectra to confirm the identification.

Results and Discussion

Overview: Comparison of Mass Spectrometry Instrumentation

Three pyrolysis-GC-MS systems were used and compared in this study: Py-GC-dAPCI-TOFMS, Py-GC-dAPCI-QTOFMS, and Py-GC-EI-MS. Pyrolysis was performed with the same type of drop tube micropyrolyzer for all three instrumentations. The first two instrumentations used soft ionization at atmospheric pressure, dAPCI, with virtually the same setup, which minimizes the fragmentation of GC eluents and ionizes them as molecular ions ($M^{+\bullet}$), protonated molecules ($[M+H]^+$), and/or ammonium ion adducts ($[M+NH_4]^+$). Although three forms of ions can be formed for each analyte, most of the time, only one or two ion forms are present in high abundance. The ratios between the three ionized forms of analytes were slightly different in the two mass spectrometers due to minor differences in the source conditions. For Py-GC-dAPCI-TOFMS, fragmentation is induced in the ion source region as ISCID which relies on GC for the separation of precursor ions, similar to GC-EI-MS. ISCID enables CID by increasing the skimmer cone voltage at the instrument entrance [28–30]. As "true" MS/MS can be performed in Py-GC-dAPCI-QTOFMS, it has several advantages: (1) MS and MS/MS can be obtained in a single experiment, (2) precursors can be selected for MS/MS with a narrow mass window (± 1 Da), and (3) collision energy can be automatically adjusted depending on the mass of the precursor ion (see experimental). The third instrumentation, Py-GC-EI-MS, is most widely used owing to its wide availability and the ability to perform database search against the NIST library that contains over 200,000 EI-MS spectra [1, 26]. However, Py-GC-EI-MS may not be able to identify compounds that are not in the library such as low abundance natural products or some pyrolysis compounds. Additionally, some compounds fragment too extensively to produce characteristic fragments that can be used for identification.

The gas chromatograms from each instrument are compared in Figure 1 for the pyrolysis of Kraft lignin. Kraft lignin is a side product of pulping process and already severely degraded. The goal of the current work, however, is to develop a methodology to characterize lignin pyrolysis products rather than the feedstock itself. The GC profiles for QTOFMS and TOFMS look quite similar since they utilize the same ionization method (dAPCI). The slight difference between the two chromatograms is mostly due to minor differences in the ion source humidity; however, it did not significantly affect the outcome. The Py-GC-EI-MS chromatogram looks quite different from the others due to the difference between EI and dAPCI. Early eluting compounds are seen only in Py-GC-EI-MS because they are not detected by dAPCI-MS. According to the EI-MS database match, these gases and volatile compounds include carbon dioxide, hydrogen sulfide, and dimethyl disulfide. The sulfur compounds most likely have come from the residues in the lignin extraction process [31]. In contrast, late eluting compounds have higher ion signals in the chromatograms with dAPCI, suggesting Py-GC-dAPCI-MS might be more efficient for the ionization of large pyrolyzates. The dead time for each instrument was slightly different but calibrated with a few known compounds.

Overview: Data Analysis

Data analysis was performed for each MS spectrum that was extracted from GC-MS datasets with the chromatographic peak abundance of at least 0.01% of the base peak. EI-MS data was deconvoluted using AMDIS and searched against the NIST database. The handbook of NIST MS Search Program suggests a matching score above 900 as an excellent match, 800–900 as a good match, 700–800 as a fair match, and below 600 as a very poor match [32]. In this study, a matching score of 800 or higher was considered as positive match whereas 600–800 was considered as potential match. Those with the score below 600 were considered as random incorrect matching. Among the 58 EI-MS spectra, 22 of them are positively identified and 17 are potentially identified.
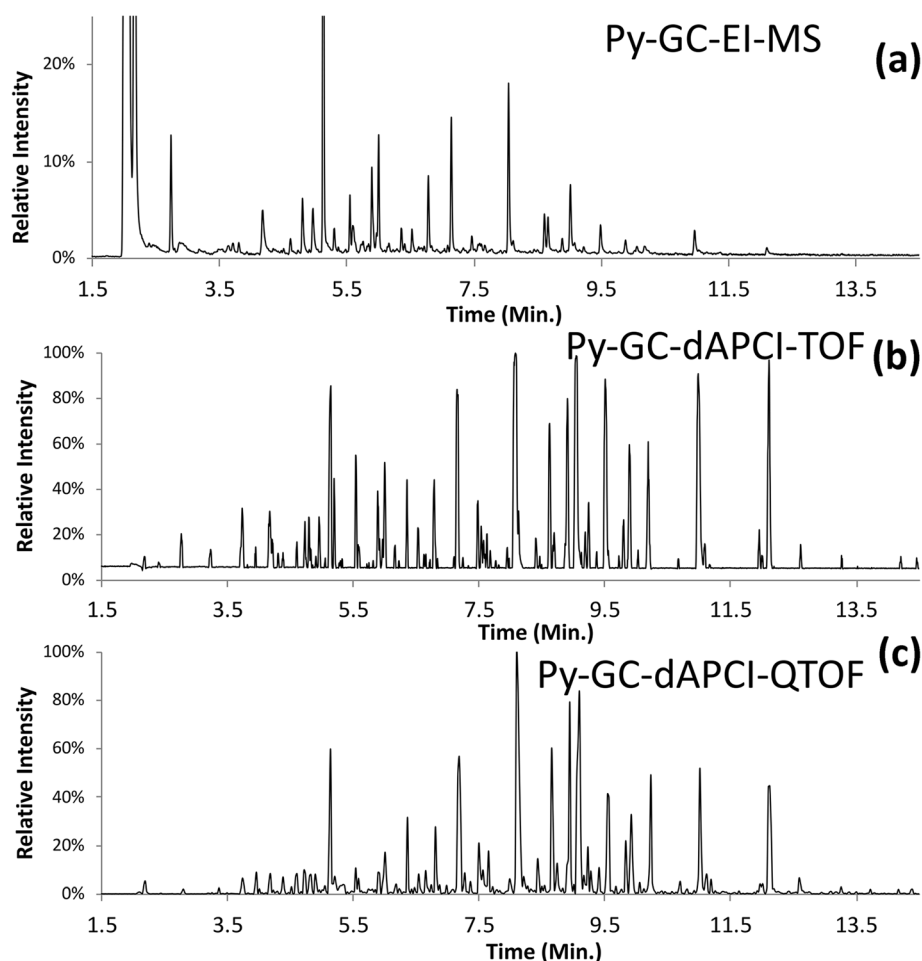
Figure 1. Chromatograms of Kraft lignin pyrolysis using **(a)** Py-GC-EI-MS, **(b)** Py-GC-dAPCI-TOFMS with low fragmentor voltage, and **(c)** Py-GC-dAPCI-QTOFMS. The *y*-axis is the total ion chromatogram for EI-MS and base ion chromatogram for TOFMS and QTOFMS

In order to analyze the CID spectra, data analysis was performed using CSI:FingerID program searched against more than 95 million molecular structures in PubChem library [33]. A database search was also made for the Metlin MS/MS database (https://metlin.scripps.edu); however, only one compound (vanillin) was identified using their search software. In Category 2 CASMI competition (automatic structure identification—in silico fragmentation only) [23], CSI:FingerID shows the best identification rate in positive mode, 55% for the top 1 rank and 79% among the top 10 ranks. According to Category 3 CASMI competition (automatic structural identification with full information), which CSI:FingerID did not participate, structural prediction can be much more accurate when metadata was used to assist identification, as high as 88% among the top ten matches. In such spirit, we adopted the expert's judgment to choose most likely structures among the top ten matches, assuming it would result in ~ 90% identification rate. This step of manual selection is mostly straightforward because after the removal of already identified compounds, there are only one or two lignin-like structures which are typically left, among which the one with the highest score is chosen as the correct match. Computational MS/MS

prediction cannot replace experimental MS/MS database; however, when there is not a broad enough MS/MS library, it can be a good alternative. In order to objectively determine a reasonable cutoff value in CSI:FingerID search with minimal false positive and false negative rates, the publicly available CASMI dataset (www.casmi-contest.org) was re-analyzed using CSI:FingerID. From the search result of the CASMI dataset, a value of -100 was chosen as an effective cutoff (negative scores closer to zero are better matches). This cutoff value gave a false positive rate of 7.4% and false negative rate of 15.7% for the top ten matches of CASMI dataset. In this calculation, false positive is defined when all top ten matches are incorrect above this threshold whereas false negative is defined when the correct match has the score below this cutoff.

A total of 72 CID spectra were extracted from GC-ISCID-TOFMS data and 70 CID spectra were extracted from GC-QTOFMS data; 59 and 43 of them are positively identified, respectively, with CSI:FingerID search against PubChem library, many more than 22 positive matches in EI-MS. ISCID-TOFMS showed better performance due to better signals. Therefore, most of the data analysis below is based on ISCID spectra. Three positive matching in EI-MS could not be

identified in CSI:FingerID search of CID spectra due to limited CID fragmentation. Apocynin and 2-methoxy-4-vinylphenol have scores below − 100 and eugenol has the rank of 16. All these compounds are well-known Kraft lignin pyrolyzates with high EI scores (> 800). Identities of apocynin and eugenol were also confirmed with standards for their retention times and CID spectra. Combining EI-MS and CID spectra, a total of 62 compounds are positively identified. Figure 2 shows the Venn diagram comparing NIST database search of EI-MS dataset vs CSI:FingerID search of CID spectra. The number in parenthesis indicates low score EI-MS matching confirmed to be correct according to CSI:FingerID search of CID spectra (see discussion below).

The search result of 72 CID spectra was summarized in Table 1, grouping into six categories in comparison to EI-MS database search result. A total of 19 compounds with a NIST matching score greater than 800 in EI-MS agreed well with CSI:FingerID search of CID spectra (Category 1), adding confidence to their identifications. Among the 17 potential identifications in EI-MS data (NIST score between 600 and 800), eight of them were supported by CID spectra and considered to be correct (Category 2); however, nine of them were disagreed with CID spectra and corrected based on CSI:FingerID analysis (Category 3). Additionally, a total of 23 compounds were identified solely from CSI:FingerID analysis of CID spectra (Category 4). Three positive matchings in NIST search of EI-MS spectra were not identified in CID spectra (Category 5). Finally, identification could not be made for ten compounds in either EI-MS or CID spectra, but their molecular formulas could be obtained from accurate mass information in TOFMS/QTOFMS data (Category 6). The full list of tentatively identified compounds are shown in Table 2 with the details of the CID and EI fragmentation in Table S1. A selected set of standard compounds were chosen from each
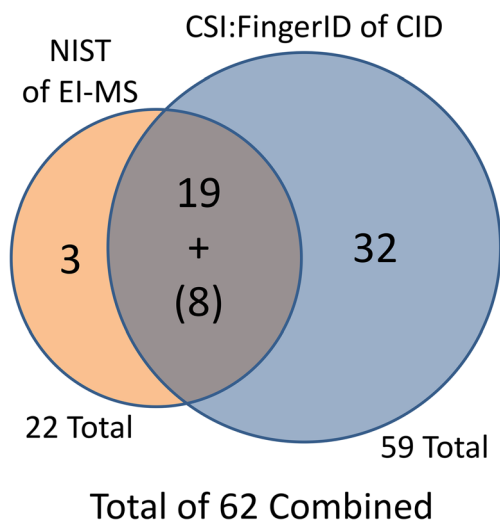
Table 1. Summary of Compounds Characterized by EI-MS and CID Spectral Search

Category	NIST database search of EI-MS spectra[a]	CSI:FingerID search of CID spectra[b]	Number of compounds[c]
1	Positive (> 800)	Positive	19 (5)
2	Potential (600–799)	Positive	8 (2)
3	Incorrect (600–799)	Positive	9 (4)
4	Unidentified (< 600)	Positive	23 (2)
5	Positive (> 800)	Unidentified	3 (2)
6	Unidentified (< 600)	Unidentified	10

[a]NIST score higher than 800 is considered as positive identification and the score between 600 and 800 is considered as potential identification, while the score lower than 600 is considered as incorrect random matching. Potential identifications in Category 3 are determined to be incorrect according to CSI:FingerID search
[b]Top ten matches with CSI:FingerID score above −100 were considered as positive matching, among which most lignin like structure not identified in other retention times is chosen as the correct one
[c]The numbers in parenthesis are confirmed with the retention time and CID spectra of standards

category and compared with GC-dAPCI-TOFMS or GC-dAPCI-QTOFMS data for their retention times and CID spectra to confirm their identities. Additionally, manual interpretation of a set of CID spectra was made to further support CSI:FingerID identification. The following sections illustrate a few examples in the first four categories that are carefully inspected manually and validated with standard compounds (Figure S1).

Category 1: Positive Identification in both EI-MS and CID

As discussed above, high EI-MS scores (> 800) in the NIST database search agree well with the CSI:FingerID analysis of CID spectra. These results serve as a proof of concept for the usefulness of CID spectra in the identification of lignin pyrolyzates. The first example of one such high scoring compound is vanillin, as shown in Figure 3 comparing the EI-MS, ISCID, and CID MS/MS spectra. The NIST database match score for vanillin is 950, which is high enough to be confident of the identification and further confirmed with the standard for the corresponding retention time and CID fragmentation pattern. The precursor ion in dAPCI, shown in the inset MS spectra of Figure 3, is mostly the protonated molecule, $[M+H]^+$, although some ammonium ion adduct, $[M+NH_4]^+$, is also observed in dAPCI-TOFMS. The minor difference in precursor ions between the two dAPCI systems is most likely due to a minor difference in the source humidity. The two CID spectra are almost identical except for some signal differences due to the collision energy difference. The EI-MS and CID spectra have slightly different but distinct fragmentation patterns: m/z 137 (M-CH$_3$), 123 (M-CHO), and 109 (M-C$_2$H$_3$O) in EI-MS; m/z 138 (MH-CH$_3$), 125 (MH-CO), and 110 (MH-C$_2$H$_3$O) for CID. This difference is attributed to have come from the fragmentation of molecular ion (EI-MS) vs protonated molecule (dAPCI-MS).

Figure 2. Venn diagram of NIST search of EI-MS and CSI:FingerID search of CID spectra for lignin pyrolysis products. The number in parenthesis is low score matching in EI-MS but confirmed by CID

NIST of EI-MS

CSI:FingerID of CID

19 + (8)

3

32

22 Total

59 Total

Total of 62 Combined

Table 2. List of Identified Compounds in Lignin Pyrolysis Combining Three Py-GC-MS Datasets

RT[a]	Molecular mass[b]	Formula	Compound name[c]	Category[d]	CSI:FingerID rank
3.92	108.057	C_7H_8O	Anisole	1	2
4.16	94.042	C_6H_6O	Phenol	1	1
14.28	102.032	$C_4H_6O_3$	N/A	6	N/A
4.36	112.052	$C_6H_8O_2$	N/A	6	N/A
4.58	122.073	$C_8H_{10}O$	1-Methoxy-2-methylbenzene	1	5
4.70	110.073	$C_7H_{10}O$	N/A	6	N/A
4.77	108.058	C_7H_8O	2-Methylphenol	1	3
4.81	126.068	$C_7H_{10}O_2$	N/A	6	N/A
4.88	124.089	$C_8H_{12}O$	N/A	6	N/A
4.93	108.058	C_7H_8O	4-Methylphenol	1	2
5.02	110.073	$C_7H_{10}O$	N/A	6	N/A
5.11	124.052	$C_7H_8O_2$	*2-Methoxyphenol*	1	4
5.18	116.047	$C_5H_8O_3$	N/A	6	N/A
5.20	124.052	$C_7H_8O_2$	N/A	6	N/A
5.27	122.073	$C_8H_{10}O$	2,4-Dimethylphenol	1	8
5.52	138.068	$C_8H_{10}O_2$	1,2-Dimethoxybenzene	1	7
5.56	122.073	$C_8H_{10}O$	3,5-Dimethylphenol	1	3
5.70	122.073	$C_8H_{10}O$	4-Ethylphenol	2	1
5.73	122.073	$C_8H_{10}O$	1-Methoxy-4-methylbenzne	2	8
5.79	152.084	$C_9H_{12}O_2$	3,4-Dimthyoxytoluene	2	3
5.87	138.068	$C_8H_{10}O_2$	2-Methoxy-5-methylphenol	1	1
5.94	138.068	$C_8H_{10}O_2$	Isocreosol	3	4
5.98	138.068	$C_8H_{10}O_2$	*2-Methoxy-4-methylphenol*	1	5
6.15	120.058	C_8H_8O	*4-Vinylphenol*	4	1
6.34	152.084	$C_9H_{12}O_2$	2,3-Dimethoxytoluene	1	3
6.50	124.052	$C_7H_8O_2$	3-Methyl-1,2-benzenediol	1	1
6.60	140.047	$C_7H_8O_3$	2,5-Dihydroxyanisole	4	1
6.64	152.084	$C_9H_{12}O_2$	1,2-Dimethoxy-4,5-dimethylbenzene	4	1
6.70	136.089	$C_9H_{12}O$	2,4,6-Trimethylphenol	2	5
6.76	152.084	$C_9H_{12}O_2$	4-Ethyl-2-methoxyphenol	1	1
6.82	152.084	$C_9H_{12}O_2$	5-Ethyl-2-methoxyphenol	4	2
7.09	152.084	$C_8H_{10}O_2$	4-Hydroxy-methylanisol	3	3
7.14	152.084	$C_9H_{10}O_2$	2-Methoxy-4-vinylphenol	5	8
7.21	152.084	$C_9H_{10}O_2$	2-Methoxy-5-vinylphenol	4	2
7.45	154.063	$C_8H_{10}O_3$	*Syringol*	3	10
7.51	122.037	$C_7H_6O_2$	4-Hydroxybenzaldehyde	4	1
7.60	164.084	$C_{10}H_{12}O_2$	*Eugenol*	5	16
7.65	166.099	$C_{10}H_{14}O_2$	*4-Propyl-2-methoxyphenol*	2	3
7.74	138.068	$C_8H_{10}O_2$	4,5-Dimethylcatechol	3	6
7.79	138.068	$C_8H_{10}O_2$	2,5-Dimethylhydroquinone	4	3
7.92	140.047	$C_7H_8O_3$	2-Methoxy-1,4-benzenediol	2	1
7.95	164.084	$C_{10}H_{12}O_2$	2-Methoxy-4-vinylanisole	4	5
8.05	152.047	$C_8H_8O_3$	*Vanillin*	1	1
8.10	164.084	$C_{10}H_{12}O_2$	*Cis-Isoeugenol*	3	1
8.19	154.063	$C_8H_{10}O_3$	1,2-Dimethoxy-4-hydroxybenzene	4	4
8.38	152.047	$C_8H_8O_3$	Isovanillin	4	2
8.45	152.084	$C_9H_{12}O_2$	2,6-Dimethoxytoluene	4	4
8.60	164.084	$C_{10}H_{12}O_2$	*Trans-Isoeugenol*	1	2
8.66	166.063	$C_9H_{10}O_3$	2-(4-Hydroxy-3-methoxyphenyl) Ethanal	4	3
8.85	154.063	$C_8H_{10}O_3$	4-Methoxy-5-methylbenzene-1,2-Diol	4	3
8.89	166.063	$C_9H_{10}O_3$	3,4-Dimethoxybenzaldehyde	1	1
9.03	166.063	$C_9H_{10}O_3$	*Apocynin*	5	8
9.06	178.099	$C_{11}H_{14}O_2$	*Trans-Methylisoeugenol*	2	2
9.17	166.063	$C_9H_{10}O_3$	N/A	6	N/A
9.22	180.079	$C_{10}H_{12}O_3$	Coniferyl alcohol	3	3
9.36	180.079	$C_{10}H_{12}O_3$	Methyl vanillate	4	8
9.48	180.079	$C_{10}H_{12}O_3$	Guaiacylacetone	1	9
9.70	178.099	$C_{11}H_{14}O_3$	4-Allyl-2,6-dimethoxyphenol	4	3
9.77	166.063	$C_9H_{10}O_3$	3-Hydroxy-4-methoxyacetophenone	2	2
9.87	180.079	$C_{10}H_{12}O_3$	3,4-Dimethoxyacetophenone	1	1
10.01	194.094	$C_{11}H_{14}O_3$	Syringolpropene	4	8
10.16	180.079	$C_{10}H_{12}O_3$	2,6-Dimethoxy-4-vinylphenol	4	5
10.65	192.079	$C_{11}H_{12}O_3$	(E)-4-(4-Hydroxy-3-methoxyphenyl) But-3-Enal	4	1
10.96	182.094	$C_{10}H_{14}O_3$	*Dihydroconiferyl alcohol*	3	1
11.07	182.058	$C_9H_{10}O_4$	Syringaldehyde	4	1
11.93	196.074	$C_{10}H_{12}O_4$	3-(4-Hydroxy-3-methoxyphenyl) propanoic acid	4	1
11.97	196.074	$C_{10}H_{12}O_4$	*Acetosyringone*	4	1
12.10	178.063	$C_{10}H_{10}O_3$	*Coniferyl aldehyde*	3	1
18.20	234.141	$C_{18}H_{18}$	11-Propan-2-ylidene-5,6-dihydrodibenzo[2,1-b:2′,1′-f][7]annulene	3	1

Table 2 (continued)

RT[a]	Molecular mass[b]	Formula	Compound name[c]	Category[d]	CSI:FingerID rank
18.26	240.115	$C_{16}H_{16}O_2$	4,4′-Dimethoxystilbene	4	1
22.66	272.105	$C_{16}H_{16}O_4$	3′,5-Dimethoxystilbene-3,4′-diol	4	1
23.53	314.115	$C_{18}H_{18}O_5$	N/A	6	N/A

[a]RT is retention time in minutes
[b]Compound mass is the neutral monoisotopic mass
[c]Compounds in italics were confirmed with standards
[d]Categories are defined in Table 1

The CID spectra also show a few more characteristic fragments, such as m/z 93 (MH-$C_2H_4O_2$) and 65 (MH-$C_3H_4O_3$). The loss of $C_2H_4O_2$ can be attributed to the loss of both the aldehyde (CHO) and methoxy (CH_3O) substituents leaving a phenolic fragment, and the loss of $C_3H_4O_3$ corresponds to additional CO loss from the phenolic fragment. The loss of both $C_2H_4O_2$ and $C_3H_4O_3$ can be also found in other aldehyde containing guaiacol compounds (e.g., coniferyl aldehyde).

Another example in this category is 2-methoxy-4-methylphenol (creosol), shown in Figure S2. The ESI-MS database match score for creosol was 940, and it is also a well-known product in lignin pyrolysis. Many fragments are identical among all three spectra such as m/z 123 (M-CH_3), m/z

95 (M-C_2H_3O), and m/z 77 (M-$C_2H_5O_2$), suggesting these fragmentations mostly occur from the molecular ion precursor. Some fragments only present in CID spectra but not in EI-MS, such as m/z 107, 93, and 79, are attributed to the fragmentation of the protonated molecule.

Category 2: Low Score EI-MS Match Confirmed by CID

Category 2 includes the compounds that are low scoring NIST database matches (600–800) but considered to be correct assignments according to CSI:FingerID analysis of CID spectra. There are several possibilities for low scores including poor deconvolution by AMDIS, low
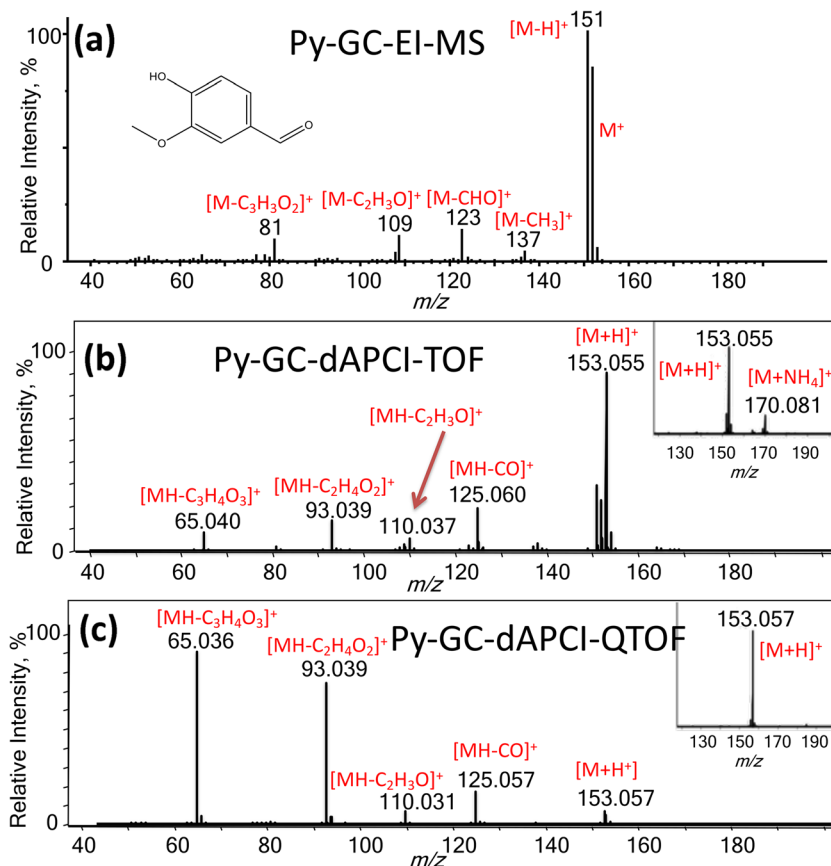

Figure 3. (a) EI-MS, (b) ISCID, and (c) CID MS/MS spectra of vanillin. The precursor spectra for the corresponding CID spectra are both shown as inset figures

signal intensity, or similarity to other compounds in the database. Normally, a low score match cannot be trusted unless validated with standards. The first example in this category is 4-propyl-2-methoxyphenol, which has low signal in GC-EI-MS and therefore a low score, 690. The spectra for this compound are shown in Figure 4. The EI-MS and ISCID spectra have many major fragments in common such as m/z 137 (M-C_2H_5), 135 (M-OCH_3), 107 (M-C_3H_7O), 94 (M-C_4H_8O), and 79 (M-$C_4H_7O_2$) corresponding to fragmentation of the molecular ion. In contrast, CID MS/MS spectrum is dominated by fragments from the protonated molecule such as m/z 139 (MH-C_2H_4), 125 (MH-C_3H_6), 107 (MH-C_3H_8O), 93 (MH-$C_4H_{10}O$), and m/z 79 (MH-$C_4H_8O_2$), which are also found in the CID spectrum of standard (Figure S1C).

Another example in this category is methylisoeugenol which had a NIST score of 660. The spectra for this compound are shown in Figure S3. This is an example of a low abundance compound without clear baseline chromatographic separation. As shown in the inset figures, there are contamination peaks at m/z 167 corresponding to the tailing of the overlapping apocynin peak (EI score of 900) in addition to the precursor peaks for methyl isoeugenol (m/z 179 for [M+H]$^+$ and m/z 196 for [M+NH$_4$]$^+$). Even with a few overlapping peaks,

CSI:FingerID was able to analyze the ISCID spectrum and match the correct compound with a good score ($-$ 49.56) and high rank (2). CID MS/MS spectrum is much clearer isolated from contamination peaks, and fragments are almost exclusively from protonated molecule.

Category 3: Incorrect Potential Match in EI-MS Corrected by CID

An example of low EI-MS score which was incorrectly identified compounds in the NIST database search is coniferyl aldehyde (Figure 5) which was originally identified as 4-methoxy cinnamic acid with EI-MS score of 690. The two compounds are isomers, and the EI-MS spectra in the NIST database are quite similar, leading to the misidentification. Due to the similarity in EI-MS spectra, confirmation with a standard for both retention time and CID was necessary (Figure S1E). Some major peaks are common in all three spectra (Figure 5) such as m/z 161, m/z 147, and m/z 107 while some others are distinct such as m/z 119 and 91 in CID. The CID spectra are also available for both the compounds in the NIST MS/MS library and that of coniferyl aldehyde matches well with our CID spectra. The NIST MS/MS spectrum
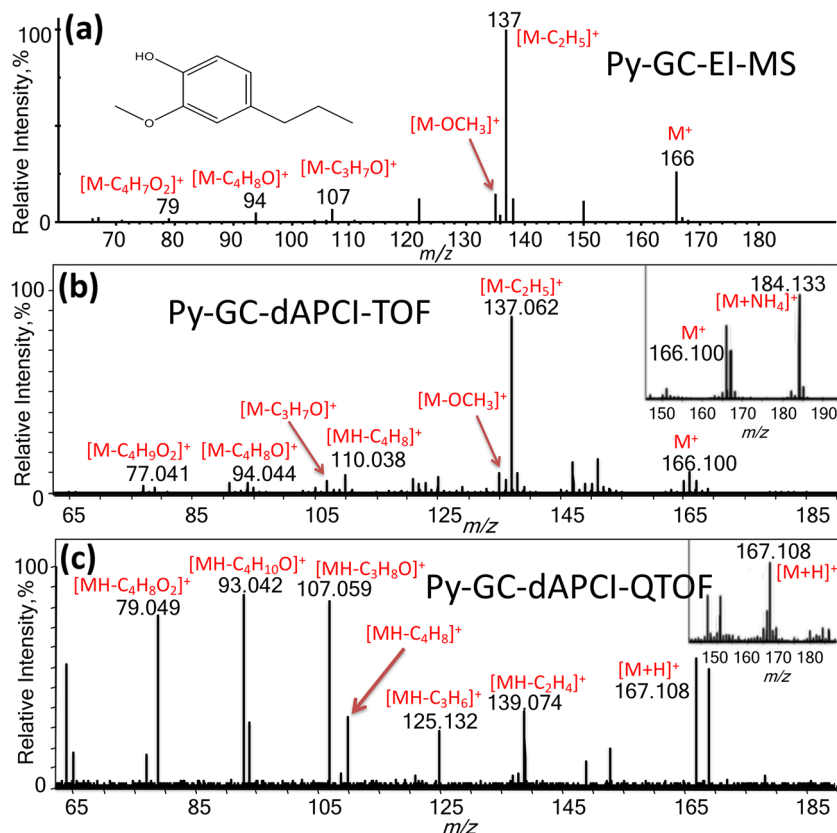

Figure 4. (a) EI-MS, (b) ISCID, and (c) CID MS/MS spectra of 4-propyl-2-methoxyphenol. The precursor spectra for the corresponding CID spectra are both shown as inset figures
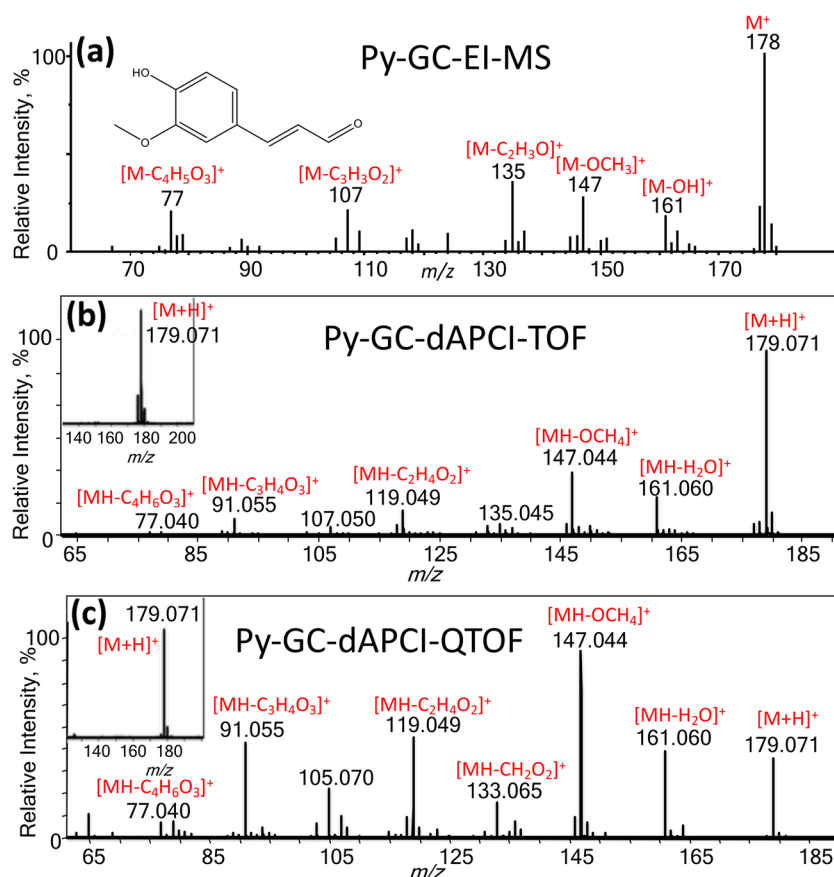
Figure 5. (a) EI-MS, (b) ISCID, and (c) CID MS/MS spectra of coniferyl aldehyde. The precursor spectra for the corresponding CID spectra are both shown as inset figures

for 4-methoxy cinnamic acid has only two major fragments and clearly distinguishable from the MS/MS of coniferyl aldehyde. CSI:FingerID gave a high score (− 62.31) and top rank match for this compound.

Another example is the peak at RT of 11.0 min (Figure S4). This peak was initially assigned as homovanillic acid ($C_9H_{10}O_4$) with a matching score of 760 in the NIST database search but the elemental composition does not agree with the molecular formula calculated from dAPCI-TOFMS: $C_{10}H_{14}O_3$. This is a good example where elemental composition from accurate mass can correct misassignment in NIST database search of nominal mass EI-MS spectra. Using the correct molecular formula and CID spectral search, we propose the structure as dihydroconiferyl alcohol (CSI:Finger ID score of − 54.92 and top rank). The EI-MS or CID spectra for this compound are not available in NIST, Metlin, or MassBank (www.massbank.jp) database. However, the EI-MS spectrum is available for coniferyl alcohol which has a similar spectrum, especially the base peak at m/z 137, supporting our assignment. Furthermore, we were able to purchase the reference standard, which produced CID MS/MS spectrum exactly matching with our experimental spectrum (Figure S1F). The most abundant

fragment at m/z 137 (M-C_2H_5O) is common to most guaiacol-type compounds with alkyl side chains. Additionally, the EI-MS spectrum of the standard is also matching as shown in Figure S5A.

Category 4: Identification Solely from CID

Category 4 compounds are those that do not have a reliable database match (EI score < 600) but identified in CSI:FingerID search of CID spectrum. They are mostly compounds missing from the NIST database or have very low EI-MS signals. New structures were proposed for these 23 compounds. They generally contain guaiacol lignin monomer structure with a variety of side chains. An example of such compound is 4-vinylphenol as shown in Figure 6, which is a high abundance pyrolyzate from p-coumaryl lignin (commonly found in herbaceous biomass) but is in lower abundance in guaiacol-rich lignin like the one pyrolyzed in this study. Although this compound is a common lignin pyrolyzate, it does not have an EI-MS spectrum in the NIST database. There are two major fragments present in both CID spectra but not in EI-MS: m/z 103 (-H_2O) and m/z
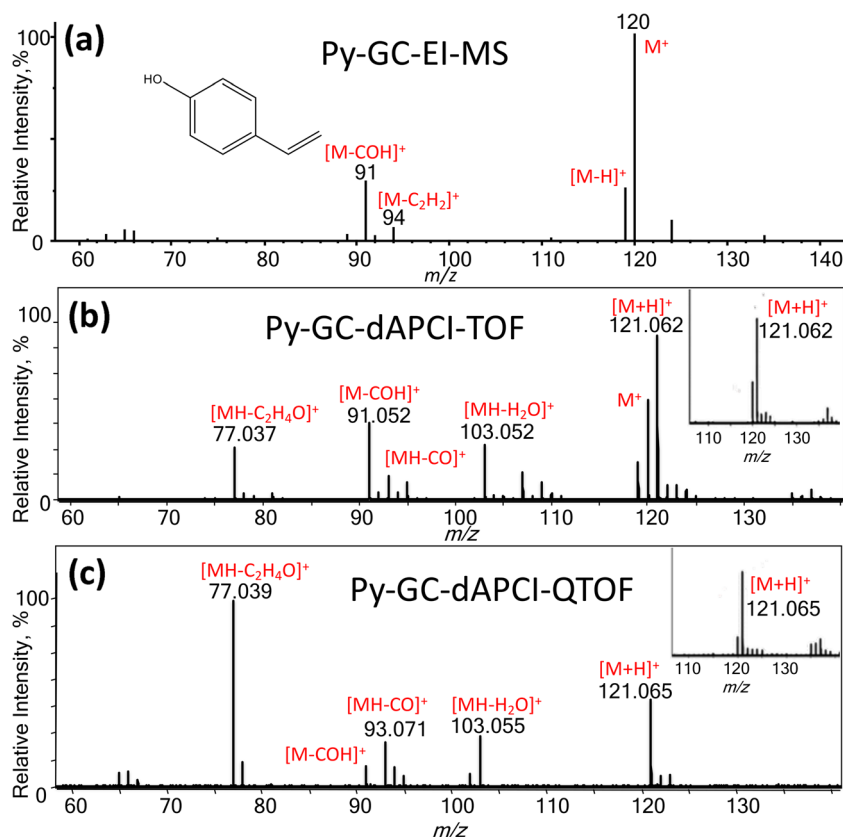
Figure 6. **(a)** EI-MS, **(b)** ISCID, and **(c)** CID MS/MS spectra of 4-vinylphenol. The precursor spectra for the corresponding CID spectra are both shown as inset figures

77 (-C_2H_4O). These fragments provide valuable information not available from EI-MS as they confirm the presence of a phenol as well as the identity of the side chain. This compound was confirmed with CID and EI-MS of a standard (Figures S1G and S5B) and retention time.

Conclusion

We have successfully demonstrated a novel method to identify volatile lignin pyrolyzates with GC-dAPCI-MS/MS and computational algorithm for MS/MS interpretation. Identification of these compounds has been typically accomplished by GC-EI-MS followed by NIST database search, but the identification rate is very low because many of them are absent in EI-MS library. In an application to Kraft lignin, we have positively identified 82% (59 out of 72) of the chromatographic peaks in CSI:FingerID search of CID spectra, whereas it is only 38% in NIST search of EI-MS spectra (22 out of 58). This technique also provides a means to confirm or correct any low EI-MS score matching. The success of this approach has come from the combination of (1) soft ionization of gas phase analytes in atmospheric pressure assisted by ammonia dopant, (2) molecular formula analysis with high-resolution mass spectrometry, (3) structural information in tandem mass spectrometry, and (4) computational algorithm to identify molecular structures without MS/MS database.

CSI:FingerID used in this study is a very effective algorithm but has a limitation to find a correct one among the top matches. Albeit this computational approach will be continuously improved by training with more experimental MS/MS spectra over the coming years, it is still expected to be challenging to determine the correct match among structurally similar compounds. Regardless, this approach turned out to be quite effective in the current application to lignin pyrolysis products when combined with GC-EI-MS because an expert can choose most lignin likely structures among the top matches. Overall, we believe the proposed approach could be very useful to complement GC-EI-MS analysis of a very complex mixture when the EI-MS library is not sufficient, such as volatile pyrolysis products or low abundance natural products.

References

1. Stein, S.E., Scott, D.R.: Optimization and testing of mass spectral library search algorithms for compound identification. J. Am. Soc. Mass Spectrom. **5**, 859–866 (1994)

2. Abushareeda, W., Lyris, E., Kraiem, S., Wahaibi, A.A., Alyazidi, S., Dbes, N., Lommen, A., Nielen, M., Horvatovich, P.L., Alsayrafi, M., Georgakopoulos, C.: Gas chromatographic quadrupole time-of-flight full scan high resolution mass spectrometric screening of human urine in antidoping analysis. J. Chromatogr. B. **1063**, 74–83 (2017)

3. Weidt, S., Haggarty, J., Kean, R., Cojocariu, C.I., Silcock, P.J., Rajendran, R., Ramage, G., Burgess, K.E.V.: A novel targeted/untargeted GC-Orbitrap metabolomics methodology applied to Candida albicans and Staphylococcus aureus biofilms. Metabolomics. **12**, (2016). https://doi.org/10.1007/s11306-016-1134-2

4. Dzidic, I., Carroll, D.I., Stillwell, R.N., Horning, E.C.: Comparison of positive ions formed in nickel-63 and corona discharge ion sources using nitrogen, argon, isobutane, ammonia and nitric oxide as reagents in atmospheric pressure ionization mass spectrometry. Anal. Chem. **48**, 1763–1768 (1976)

5. Horning, E.C., Horning, M.G., Carroll, D.I., Dzidic, I., Stillwell, R.N.: New picogram detection system based on a mass spectrometer with an external ionization source at atmospheric pressure. Anal. Chem. **45**, 936–943 (1973)

6. Li, D.-X., Gan, L., Bronja, A., Schmitz, O.J.: Gas chromatography coupled to atmospheric pressure ionization mass spectrometry (GC-API-MS): review. Anal. Chim. Acta. **891**, 43–61 (2015)

7. Portolés, T., Mol, J.G.J., Sancho, J.V., Hernández, F.: Use of electron ionization and atmospheric pressure chemical ionization in gas chromatography coupled to time-of-flight mass spectrometry for screening and identification of organic pollutants in waters. J. Chromatogr. A. **1339**, 145–153 (2014)

8. Portolés, T., Mol, J.G.J., Sancho, J.V., López, F.J., Hernández, F.: Validation of a qualitative screening method for pesticides in fruits and vegetables by gas chromatography quadrupole-time of flight mass spectrometry with atmospheric pressure chemical ionization. Anal. Chim. Acta. **838**, 76–85 (2014)

9. Nácher-Mestre, J., Serrano, R., Portolés, T., Berntssen, M.H.G., Pérez-Sánchez, J., Hernández, F.: Screening of pesticides and polycyclic aromatic hydrocarbons in feeds and fish tissues by gas chromatography coupled to high-resolution mass spectrometry using atmospheric pressure chemical ionization. J. Agric. Food Chem. **62**, 2165–2174 (2014)

10. Mesihää, S., Ketola, R.A., Pelander, A., Rasanen, I., Ojanperä, I.: Development of a GC-APCI-QTOFMS library for new psychoactive substances and comparison to a commercial ESI library. Anal. Bioanal. Chem. **409**, 2007–2013 (2017)

11. Ruttkies, C., Strehmel, N., Scheel, D., Neumann, S.: Annotation of metabolites from gas chromatography/atmospheric pressure chemical ionization mass spectrometry data using an in silico generated compound database and MetFrag. Rapid Commun. Mass Spectrom. **29**, 1521–1529 (2015)

12. Portolés, T., Mol, J.G.J., Sancho, J.V., Hernández, F.: Advantages of atmospheric pressure chemical ionization in gas chromatography tandem mass spectrometry: pyrethroid insecticides as a case study. Anal. Chem. **84**, 9802–9810 (2012)

13. Portolés, T., Cherta, L., Beltran, J., Hernández, F.: Improved gas chromatography–tandem mass spectrometry determination of pesticide residues making use of atmospheric pressure chemical ionization. J. Chromatogr. A. **1260**, 183–192 (2012)

14. McEwen, C.N., McKay, R.G.: A combination atmospheric pressure LC/MS:GC/MS ion source: advantages of dual AP-LC/MS:GC/MS instrumentation. J. Am. Soc. Mass Spectrom. **16**, 1730–1738 (2005)

15. Song, L., Cho, D.S., Bhandari, D., Gibson, S.C., McNally, M.E., Hoffman, R.M., Cook, K.D.: Liquid chromatography/dopant-assisted atmospheric pressure chemical ionization mass spectrometry for the analysis of non-polar compounds. Int. J. Mass Spectrom. **303**, 173–180 (2011)

16. Covey, T.R., Thomson, B.A., Schneider, B.B.: Atmospheric pressure ion sources. Mass Spectrom. Rev. **28**, 870–897 (2009)

17. Habib, A., Usmanov, D., Ninomiya, S., Chen, L.C., Hiraoka, K.: Alternating current corona discharge/atmospheric pressure chemical ionization for mass spectrometry: alternating current corona discharge/APCI-MS. Rapid Commun. Mass Spectrom. **27**, 2760–2766 (2013)

18. Cole, D.P., Lee, Y.J.: Effective evaluation of catalytic deoxygenation for in situ catalytic fast pyrolysis using gas chromatography–high resolution mass spectrometry. J. Anal. Appl. Pyrolysis. **112**, 129–134 (2015)

19. Hutchinson, C.P., Lee, Y.J.: Evaluation of primary reaction pathways in thin-film pyrolysis of glucose using 13C labeling and real-time monitoring. ACS Sustain. Chem. Eng. **5**, 8796–8803 (2017)

20. Shen, H., Dührkop, K., Böcker, S., Rousu, J.: Metabolite identification through multiple kernel learning on fragmentation trees. Bioinformatics. **30**, i157–i164 (2014)

21. Dührkop, K., Shen, H., Meusel, M., Rousu, J., Böcker, S.: Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proc. Natl. Acad. Sci. **112**, 12580–12585 (2015)

22. Heinonen, M., Shen, H., Zamboni, N., Rousu, J.: Metabolite identification and molecular fingerprint prediction through machine learning. Bioinformatics. **28**, 2333–2341 (2012)

23. Schymanski, E.L., Ruttkies, C., Krauss, M., Brouard, C., Kind, T., Dührkop, K., Allen, F., Vaniya, A., Verdegem, D., Böcker, S., Rousu, J., Shen, H., Tsugawa, H., Sajed, T., Fiehn, O., Ghesquière, B., Neumann, S.: Critical assessment of small molecule identification 2016: automated methods. J. Cheminformatics. **9**, (2017). https://doi.org/10.1186/s13321-017-0207-1

24. Bioenergy Technologies Office Publications and Product Library, https://www1.eere.energy.gov/library/viewdetails.aspx?productid=7854&Page=1

25. Zhou, S., Xue, Y., Sharma, A., Bai, X.: Lignin valorization through thermochemical conversion: comparison of hardwood, softwood and herbaceous lignin. ACS Sustain. Chem. Eng. **4**, 6608–6617 (2016)

26. Stein, S.E.: An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. J. Am. Soc. Mass Spectrom. **10**, 770–781 (1999)

27. Dromey, R.G., Stefik, M.J., Rindfleisch, T.C., Duffield, A.M.: Extraction of mass spectra free of background and neighboring component contributions from gas chromatography/mass spectrometry data. Anal. Chem. **48**, 1368–1375 (1976)

28. Bencsath, F.A., Field, F.H.: Ion retardation and collision-induced dissociation in the thermospray ion source. Anal. Chem. **60**, 1323–1329 (1988)

29. Abrankó, L., García-Reyes, J.F., Molina-Díaz, A.: In-source fragmentation and accurate mass analysis of multiclass flavonoid conjugates by electrospray ionization time-of-flight mass spectrometry. J. Mass Spectrom. **46**, 478–488 (2011)

30. Parcher, J.F., Wang, M., Chittiboyina, A.G., Khan, I.A.: In-source collision-induced dissociation (IS-CID): applications, issues and structure elucidation with single-stage mass analyzers. Drug Test. Anal. n/a-n/a. https://doi.org/10.1002/dta.2249

31. Han, T., Sophonrat, N., Evangelopoulos, P., Persson, H., Yang, W., Jönsson, P.: Evolution of sulfur during fast pyrolysis of sulfonated Kraft lignin. J. Anal. Appl. Pyrolysis. (2018). https://doi.org/10.1016/j.jaap.2018.04.006

32. NIST MS Search Program, https://chemdata.nist.gov/mass-spc/ms-search/

33. Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B.A., Wang, J., Yu, B., Zhang, J., Bryant, S.H.: PubChem substance and compound databases. Nucleic Acids Res. **44**, D1202–D1213 (2016)