



Universidad Autónoma
de Madrid

Biblos-e Archivo
Repositorio Institucional UAM

Repositorio Institucional de la Universidad Autónoma de Madrid
<https://repositorio.uam.es>

Esta es la **versión de autor** del artículo publicado en:
This is an **author produced version** of a paper published in:

Organic Geochemistry 143 (2020): 104012

DOI: <https://doi.org/10.1016/j.orggeochem.2020.104012>

Copyright: © 2020 Elsevier Ltd. This manuscript version is made available under the CC-BY-NC-ND 4.0 licence <http://creativecommons.org/licenses/by-nc-nd/4.0/>

El acceso a la versión del editor puede requerir la suscripción del recurso
Access to the published version may require subscription

Assessment of the molecular composition of humic acid as an indicator of soil carbon levels by ultra-high-resolution mass spectrometric analysis

Marco A. Jiménez-González ^{(a,c)*}, Gonzalo Almendros ^(a), Derek C. Waggoner ^(b), Ana M. Álvarez ^(c), Patrick G. Hatcher ^(b)

^a Department of Biogeochemistry and Microbial Ecology, National Museum of Natural Sciences, Madrid, Spain.

^b Department of Chemistry and Biochemistry, Old Dominion University, Norfolk (VA), USA.

^c Department of Geology and Geochemistry, Autonomous University of Madrid, Madrid, Spain.

* email: marco_jim@hotmail.com

Keywords: carbon sequestration; FTICR-MS; humic acid; soil organic matter

Highlights

- A relationship exists between SOC content and the composition of soil HAs.
- Unsaturated lipid and lignin structures prevail in high SOC content scenarios.
- Saturated lipid prevails in low SOC content scenarios.
- Humic acid structure keeps information on soil properties as SOC content.
- The SOC can be predicted from the molecular composition of HAs.

Abstract

Long-term stabilization of soil organic matter (SOM) plays an important role in the carbon cycle. Hence, understanding biogeochemical mechanisms of soil C sequestration is crucial to control its release to the atmosphere. This study aims at investigating the biogeochemical mechanisms of soil C sequestration. An exploratory assessment was carried out on the information about the soil C levels provided by the molecular composition of humic acids (HAs) analyzed by electrospray ionization (ESI) Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS). Significant PLS forecasting model for total soil C was obtained using as descriptors the 131 compounds in common in all the HAs detected by FTICR-MS, and its variable importance for projection (VIP) was plotted in the space defined by their atomic ratios using van Krevelen diagrams. The results indicated that significant relationship exists between the HAs molecular composition and the soil organic C levels. The VIP values for the different groups of compounds illustrate how HA contains information about the amounts of C stored in the soil: the HAs in the soils with high levels of organic C have significantly ($P < 0.1$) higher proportions of unsaturated lipid and lignin-derived compounds; on the other side, low soil organic C levels are associated to HAs with comparatively high proportions of saturated lipid compounds.

1. Introduction

Soil is the largest carbon reservoir on Earth (Batjes, 1996; Lal, 2004); for this reason, small variations in soil organic C (SOC) content may have a dramatic effect on the atmospheric CO₂ concentration. In fact, progressive land desertification of soils in Mediterranean areas associated to increasing emission of CO₂ to the atmosphere due to mineralization of the soil organic matter (SOM) has encouraged the research attempting to interpret the changes in molecular composition of SOM (Solomon et al., 2007; Faria et al., 2015). This molecular composition of SOM in particular its humic acid (HA) fraction can be a source of information to be related with the humification mechanisms involved in SOM recalcitrance, resulting in the soil potential for C storage (Jiménez-González et al., 2017, 2018, 2019). In consequence, the soil humic fractions more resistant to degradation, which presumably retain in their structure more environmental information are among the most investigated topics of soil science (Spaccini et al., 2002; Sutton & Sposito, 2005; Song et al., 2014). In particular, HAs include molecular components of plant and microbial origin, and show a different composition in each soil and a macromolecular complexity that increases over time (Panettieri et al., 2014; Miller et al., 2015; Tadini et al., 2015). Therefore HAs have been considered a suitable soil fraction to obtain information in environmental studies (Fernández-Getino et al., 2013; Jiménez-González et al., 2019). According to its solubility, humic substances can be operationally divided into three fractions: fulvic acids (alkali- and acid-soluble fraction), HAs (alkali-soluble and acid-insoluble fraction) and humin (alkali- and acid-insoluble fraction) (Stevenson, 1994). Some recent studies have suggested a clear relationship (irrespective to any causal relationship) between the HAs' molecular components, SOC levels and

resistance to biodegradation of the SOM (Almendros et al., 2018). The humification processes and the transformation of SOM depend on different chemical, physical and environmental factors (vegetation type, climate, pH, geological substrate, etc) (Parton et al., 1987) and this transformation is expressed by means of progressive changes in the molecular composition of the SOM (Stevenson, 1994). In consequence, the origin of SOM from plant or microbial biomass, as well as the extent of its biogeochemical transformation could be scrutinized from the analysis of molecular proxies such as a series of signature compounds that are part of the structure of the HAs. For instance guaiacyl and syringyl methoxyphenols from lignins of different origin, levoglucosan from carbohydrates, aliphatic hydroxyacids and diacids from plant polyesters e.g. cutins, or even cyclic lipid compounds such as diterpenoids, from gymnosperms, triterpenes from most angiosperms, etc (Derenne & Quénéa, 2015).

It is well known how soil mineralogy can play an important role in the SOM stabilization (Kögel-Knabner et al., 2008; Feng et al., 2014), the organo-mineral interactions and the different protection models of SOM have also been studied as important factors in SOM resilience (Spaccini et al., 2002; Spaccini & Piccolo, 2012; Simonetti et al., 2017). On the other hand, the composition of SOM can also be important in its stabilization and resistance to the degradation. In fact, the chemical composition of the HAs may be primarily or secondarily linked to the recalcitrance of this SOM fraction to the extent to which compositional differences in the molecular fractions of the HA structure are significantly reflected in the levels of organic C in the soil (Almendros et al., 2018; Piccolo et al., 2018).

A classical problem in the study of the molecular composition of the SOM fractions is the unavoidable structural alteration produced during the pretreatments

required for its isolation in the laboratory, e.g., extraction of SOM by strong acid and bases. Nonetheless, destructive techniques such as analytical pyrolysis and wet chemical degradation still provide useful information about the nature of the SOM composition despite in some cases the quantitative yields are very doubtful, and these techniques are mainly used for “fingerprinting” humic substances from different origins. In any case, accurate structural analysis cannot be carried out by analysis of the whole unfractionated SOM using classical approaches e.g., infrared spectroscopy (IR) or nuclear magnetic resonance spectroscopy (NMR), that in general do not provide structural information at the level of the units of the macromolecules. The application of “hybrid” techniques such as Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS) has shown a considerable potential for fast and direct analysis of molecular compounds in more detailed level (Sleighter and Hatcher, 2007). Although the fragments released by FTICR-MS may not be representative of the total components of the SOM, it is also true that they provide fine molecular-level information that varies significantly depending on the origin of the sample under study. This is especially helpful both in fingerprinting studies as well as for statistical approaches based on information provided by a limited number of compounds with value as source-specific tracers (Jiménez-Morillo et al., 2018). For this reason, FTICR-MS has been recently applied to the compositional research of humic substances (Ikeya et al., 2015; DiDonato et al., 2016) and even in studies on black carbon and alicyclic aliphatic compounds formation from lignin precursors (Waggoner et al., 2015). In particular, this approach has the advantage that can identify thousands of compounds with a detailed level of information about its empirical formulas. At first sight, this massive information may represent a limitation in the interpretation of the results, however using the van Krevelen diagrams (van

Krevelen, 1950) to display the molecules in the space defined by its H/C and O/C ratios, it is possible to compare the differentiating features of the SOM molecular structure even in whole soil samples. This graphical representation of the compounds in the H/C and O/C plane is helpful to obtain an insight on their origin, composition and chemical nature (van Krevelen, 1950; Kim et al., 2003; Kramer et al., 2004; Ikeya et al., 2015). After the study by Kim et al. (2003), using this diagram to display compounds detected by FTICR-MS from natural organic matter, a series of authors used this approach to compare different fractions of humic substances from contrasting environments (McKee and Hatcher, 2015; DiDonato et al., 2016; Kamjunke et al., 2017). For exploratory purposes, different regions of the van Krevelen diagram have been considered by several authors as corresponding to defined groups of organic compounds with characteristic H/C and O/C ratios (aromatic, condensed aromatic, lignin-derived, carbohydrate, etc) (Kramer et al., 2004; Ikeya et al., 2015). This combination of the FTICR-MS and the van Krevelen diagram has provided a very perceptual approach to study the natural transformation and evolution paths of SOM and fossil organic resources (Kim et al., 2003; Ikeya et al., 2015).

This research is based on the application of FTICR-MS to HA fractions isolated from organic matter of contrasting soils with a large variability in ecological, physical and chemical properties and specifically SOC content. The principal goal is to explore possible structural relationships between HA composition and the SOC level. In addition, and as high carbon content in soil could in some cases be associated to a particular resistance to the biodegradation of the corresponding SOM, this study also aims to describe which molecular constituents behave as SOM resilience descriptors. Moreover, the data obtained would be useful to infer

systematic structural relationships illustrating the extent to which HAs retain environmental information on soil properties reflected in the resulting SOC concentration.

2. Materials and methods

2.1. Sampling sites

Topsoil samples (0–10 cm) were collected from 35 different ecosystems mainly from continental Spain with variability in climatic conditions, vegetation and geological substrate (Table 1). The soils were selected intending to cover a wide range of variability in SOC content. The soils were classified according to the IUSS Working Group WRB (2014). Three sampling points were selected in each soil, the soil samples were collected from the A horizon after removing the litter layer in each point. In a second stage, a composite soil sample was prepared by mixing soil material from the three sampling points, then the resulting material was air-dried and homogenized to < 2 mm.

2.2. Laboratory analysis

The free organic matter (FOM, particulate soil organic fraction) was removed by flotation in 2M H₃PO₄. The extractable humic substances were isolated with standard method using 0.1M Na₄P₂O₇ and 0.1M NaOH in reducing atmosphere. The fraction of HAs was separated from the fulvic acids by precipitating the total humic extract with 6M HCl. The HAs were purified by high-speed centrifugation (46000 × g)

followed by dialysis (dialysis tubing-Visking size 6, Medicell International Ltd.) to remove the salts. The main HA structural groups were studied using solid-state ^{13}C NMR spectroscopy. The spectra were obtained in a Bruker Avance 400 MHz, operating at 100.63 MHz with zirconium oxide rotors of 4 mm o.d. with Kel-F caps. The cross polarization (CP) was used during magic-angle spinning (MAS) of the rotor at 12.5 kHz. Between 5000 and 6000 scans were accumulated with a pulse delay of 300 ms. Tetramethylsilane was used to calibrate the ^{13}C chemical shifts (0 ppm). To circumvent Hartmann-Hahn mismatches, a ramped ^1H -pulse was applied during the 1 ms contact time. The spectral ranges were chosen according to the body of classical literature (González-Vila et al., 1983; Knicker, 2011; De la Rosa et al., 2019): Alkyl + α -amino C (0–45 ppm); *N*-alkyl + OCH_3 C (45–60 ppm); *O*-alkyl C (60–110 ppm); aromatic C (110–160 ppm); carbonyl C (mainly carboxyl + amide, 160–220 ppm). The ^{13}C intensity distribution was determined by integrating signal intensity over the above-mentioned chemical shift regions. The elemental analyses of HAs were conducted on a LECO CHNS-932 instrument.

2.3. FTICR-MS analysis

A Bruker 12 Tesla Apex-Qe FTICR-MS instrument equipped with an Apollo II electrospray ionization (ESI) source, operating in negative ionization mode was used. For this analysis, 0.5 mg of purified HA was dissolved in 1% NH_4OH , and a blank sample without HA was prepared. Before injection, the sample was diluted with methanol:water 1:1 v:v to improve the ionization efficiency. The injection was in a flow rate of $120\ \mu\text{L}\cdot\text{h}^{-1}$ with a nebulizer gas pressure of 20 psi and a drying gas pressure of 15 psi. Peaks identified in the blank were subtracted from the sample

peak list prior formula assignment. The empirical molecular formulas were assigned in the range from 200 to 800 m/z, using an in-house Matlab code (The MathWorks, Inc., Natick, MA) according the following criteria: $^{12}\text{C}_{2-50}$, $^1\text{H}_{5-100}$, $^{14}\text{N}_{0-6}$, $^{16}\text{O}_{1-30}$, $^{32}\text{S}_{0-2}$ and $^{32}\text{P}_{0-2}$ within an error of 1 ppm, and using the rules outlined by Stubbins et al., 2010. The parameter number of double bond equivalents (DBE), which represent the number of double bond in an structure and is calculated according:

$$DBE = \frac{1}{2}(2C + N - H + 2)$$

The mass calibration was carried out based on naturally present fatty acids (Sleighter et al., 2008). In order to use appropriate data matrices for chemometric approaches, only molecular formulas identified in all HAs from the different soils were selected. For the selection of the common compounds, after the correct assignment of the formula, the exact mass of the different molecules was used, due to the high resolution of the FTICR-MS, which allows differentiating them.

For this study, the common compounds found in all the samples were selected. We observed that the distribution of both the H/C and O/C ratios and the relative abundance of these common compounds is very similar to those of the total compounds. For this reason, selecting the common compounds has the advantage that it allows working with a reduced data matrix that is comparable among all samples and convenient for subsequent multivariate statistical analysis. Compounds intensities were normalized as total abundances i.e., percentages of the total intensity. These compounds were represented in the van Krevelen diagram using the H/C and O/C ratios calculated from each empirical formula. Additional refinement is addressed with surface density plots simultaneously showing compound clusters of points in different regions of the van Krevelen diagram, as well as its total

abundances, i.e., percentages of the total intensity. With this purpose, and using authors' own ad hoc computer program, the original $z(x,y)$ data were transferred into a 50×50 matrix (suitable to reallocate the 131 compounds represented in the plane defined by the atomic ratios) by an agglomerative manner. When several compounds coincided in the same H/C and O/C range their intensities values were aggregated. From this matrix, an interpolated surface is obtained by applying the moving average algorithm. The modified aromaticity index (AI_{mod}), which assumed half of O participating in a double bond, was used to identify the aromatic condensed structures ($AI_{mod} > 0.67$) (Koch and Dittmar, 2006), this was represented by a diagonal line in the Fig. 1.

$$AI_{mod} = \frac{1 + C - 0.5O - S - 0.5H}{C - 0.5O - S - N - P}$$

2.4. Partial least squares regression

Partial least squares regression (PLS) using the ParLeS program (Viscarra-Rossel, 2008) was applied to explain the variance of the SOC content (as dependent variable) in terms of descriptors (independent variables) consisting of the 131 common compounds obtained by FTICR-MS. In particular, PLS was used to explore the utility of the common compounds as predictors of SOC concentration. Prior to statistical analyses, the total abundances of the compounds were managed as compositional data and subjected to the convenient centered log-ratio (CLR) transformation (Aitchison, 1986). To select the minimum number of factors or latent variables (LVs) of each PLS model a series of complementary criteria were considered during the cross-validation with the leave-one-out method i.e.; the root

mean squared error (RMSE) and the Akaike's (1974) information criterion (AIC). Finally, an additional more rigorous criterion was used to confirm the possible overfitting of the model, consisting of repeating the PLS study using fully random permutation of the SOC values as dependent variables (the models were discarded as overfitted if some significant ($P < 0.05$) model is also obtained with any random dependent variable. These different validation methods help to corroborate that these compounds, which are present systematically in all sets of samples, really can reflect a relationship with the SOC level in the case of obtaining a good prediction model.

2.5. Assessment of diagnostic compounds

After confirming the significance of the model explaining the SOC values in terms of the abundances of the common compounds, an evaluation about the individual contribution of these compounds in the prediction model was carried out. With this purpose, additional 2D van Krevelen plot was prepared where the variable importance for projection (VIP) scores calculated during the PLS regression was represented in the z axis. This plot is useful to compare the contribution of the independent variables to the prediction model. Nevertheless, by definition VIP values are always positive, not informing on whether the diagnostic compounds are those prevailing in the samples with high or low SOC content.

In order to illustrate compositional differences in terms of SOC levels, new van Krevelen plots were prepared representing the average composition of a number of samples. For this purpose, the samples were ordered according the SOC content and classified by quartiles. The average composition of the 131 common compounds from the HAs of soils in the 1st quartile (Q1, samples with highest SOC content) and

the average composition of the HAs from soils with C levels in the 4th quartile (Q4, samples with lowest SOC content) were calculated and normalized as total abundances. Finally, a subtraction of the spectroscopic arrays (with the 131 common compounds) from the HA samples of these quartiles was calculated to represent in z axis the differences between the compounds proportions. The resulting three-dimensional van Krevelen plot (with positive and negative values) intuitively illustrate the differences in the molecular composition of the HA in the soils according the SOC levels (Almendros et al., 2018; Jiménez-Morillo et al., 2018).

Finally, the Student's *t* test between concentrations of the compounds released from the fraction of HAs in soils with different SOC levels was calculated to check for significant differences using the Microsoft Excel (2016) function =T.TEST(array1, array2, 2, 3). In this research, differences ($P < 0.1$) were considered in the discussion of the differential features in the composition of the different groups of HAs with contrasting C levels.

3. Results

In Table 1 it is shown the high variability in SOC content (17–157 g C·kg⁻¹ soil), and the C/N ratio, which ranged between 8.9 and 31. Table 2 also displays large variability in the elemental composition of HA: C was the most abundant element, ranging from 50.5 to 59.1 g·100g⁻¹, the next was O with values between 31.2 and 41 g·100g⁻¹. Finally, the other minor elements were H (3.3 to 5.8 g·100g⁻¹), N (2.9 to 5.7 g·100g⁻¹) and S (0.3 to 0.9 g·100g⁻¹). The solid-state ¹³C NMR spectra (supplementary Fig. 1) of the HAs showed the proportions of its C-types, corresponding to different functional groups. The aromatic C content in HA presented

values between 14.3 to 35.1%. In the case of *O*-alkyl C, the HA sample with the lowest C content had a value of 14.3% whereas the highest one reached 27.4%. The content in *N*-alkyl + OCH₃ groups showed a low variation (7.9 to 11.8%). The alkyl C region was the major region of the ¹³C NMR spectrum, with values between 25.5 and 41.5%. Finally, the carbonyl region presented values ranging from 9.4 to 14.8%.

The FTICR-MS analysis identified between 1000 and 4000 different molecular formulas (> 70% of assigned formulas) for each HA (example showed in Fig. 1); most of these compounds were composed by C, H, O and N, some of them including S and P in their structures (Table 3). Although thousands of molecular formulas were identified, only the 131 compounds, with molecular weight ranging between 309 and 555 Da, which were systematically present in all HAs amounted to an average value of 21% of the total abundance of all compounds (Table 3). These common compounds consisted mainly of C, H and O atoms, only three of these compounds included N. When the common compounds were represented in a van Krevelen diagram (Fig. 1) according the H/C and O/C ratios, it was possible to observe that these common compounds can be classified into four groups: two major groups either corresponding to lipid- or lignin-derived compounds, and other two groups with comparatively low number of molecules in regions characteristic for condensed aromatic and protein-derived structures.

The comparison of the cross-validation plot obtained in the PLS model from the experimental SOC values with that using fully-randomized SOC values is shown in Fig. 2. In the best model the values of RMSE and AIC suggested selecting up to 5 LVs to generate the model. In the case of randomized values, the RMSE and AIC calculated with the same number of LVs did not present a progressive trend of the curve (Fig. 2 b,d), and the predicted vs observed SOC sets were not correlated. This

indicates the reliability and lack of overfitting of the prediction model. Figure 2 shows the observed vs predicted values of SOC obtained using as independent variables the common compounds. In the cross-validation plots for the PLS model using the experimental values of SOC there was a significant correlation ($P < 0.05$) between predicted and observed values, with $R^2 = 0.6799$, whereas in the models using randomized values there was no significant correlation ($R^2 = 0.0084$).

The Fig. 3 shows VIP values for each compound in the SOC prediction model. These values inform on the extent to which the different compounds contribute to explain SOC levels. According to the molecular mass (Fig. 3a), it was possible to observe how comparatively higher VIP values correspond to molecules with high molecular weight, the highest values corresponding to compounds with molecular weight between 425 and 555 Da. The Fig. 3b shows intensity of VIP values for the compounds represented in the van Krevelen diagram. This plot illustrates how the major VIPs correspond to compounds that were located in the lipid region ($H/C > 1.4$). Two well-differentiated subgroups were observed in the lipid region, these two subgroups can be classified into saturated ($H/C > 1.8$; $DBE < 3$) and unsaturated lipids ($H/C = 1.4–1.8$; $DBE = 4–9$).

An overall representation of the compositional differences between HAs from the soils representing C-sinks as regards to those with low SOC levels was obtained by subtracting the average values for the total abundances of the individual compounds present in the molecular assemblages of the sample sets with high and low content of SOC i.e., quartiles (Q1-Q4) as shown in Fig 4. This plot shows up to three independent clusters in the lipid region (i.e., saturated ($H/C = 1.8–2$ and $O/C = 0–0.15$), unsaturated/alicyclic ($H/C = 1.4–1.8$ and $O/C = 0–0.15$) and highly unsaturated/aromatic lipids ($H/C = 0.7–1.4$ and $O/C = 0–0.15$). Using the values of

the Student's test, the significant ($P < 0.1$) differences between the proportions of the 131 compounds in groups with high and low SOC content were calculated. The resulting values were plotted in the same van Krevelen diagram as a contour diagram superimposed on the values of the corresponding abundances. The Student's t test values indicate that the lipid region and the region for lignin-derived compounds are those which showed significant statistical differences between the composition of HAs from soils with C levels in quartiles Q1 and Q4. In particular, the region including compounds with atomic ratios similar to those of lignin structural units showed high values in the HA samples from soils with high SOC content. In the lipid region, it was observed that HAs from soils with low SOC content were comparatively richer in saturated lipid structures. Conversely, unsaturated lipid structures showed the highest values in HAs from soils with high SOC content.

4. Discussion

Although the HAs showed very similar elemental composition, the analysis by ^{13}C NMR showed differences in its structural components. In particular, ^{13}C NMR regions assigned to aromatic C, which is often considered as an indicator of the progress of the humification process (Tinoco et al., 2014), presented a large variability. This fact indicates variations in the humification process and the resulting HA maturity that are effectively reflected in the quantitative composition of the different C-types in the HAs.

The fact that only 131 compounds were common in all HAs, may be related to the large variability of the soils studied, developed under contrasting formation factors. In any case, previous comparisons of the relative abundances and distribution in the

van Krevelen plot of the 131 common compounds and of the total compounds showed practically the same patterns, suggesting that the former set is representative of the molecular composition reflected in the whole spectra. In addition, the exclusive use of common compounds has the additional advantage of acting as a convenient data reduction method suitable for multivariate data processing, based on a data matrix with an improved normal distribution and no missing values. In the van Krevelen diagram (Fig. 1), most molecules were located in the regions for lipid and lignin-derived compounds. The important extent of the lipid domain illustrates a significant preservation of aliphatic structures, which increase in terms of the humification degree (Stevenson, 1994; Jiménez-González et al., 2017, 2018; Tinoco et al., 2018). On the other hand, the information supplied by lignin-derived compounds suggests a major molecular domain consisting of structural units of biomass from vascular plants, a feature in common with fresh organic matter at early transformation stages and continuously deposited in the soil by vegetation inputs (Fernández-Getino et al., 2013; Miralles et al., 2015; Jiménez-González et al., 2019).

The principal goal of this work is to examine for possible relationships existing between carbon content of the soils and the chemical composition of the corresponding HAs, using the PLS as an exploratory chemometric data treatment. The significant PLS prediction model (Fig. 2) obtained for SOC content showed that a relationship exists between the SOC levels and the composition of HAs as reflected by the 131 common compounds analyzed by FTICR-MS. It is probable that this chemical composition and distribution of the abundances of diagnostic compounds as regards soil C level is reflected, on the one hand, by the presence of

comparatively biodegradable products from microbial metabolism and, on the other hand, by inputs of fresh organic matter from vascular plants.

It is observed that high VIP values for the common compounds showed a systematic preference for some type of molecules. In particular, compounds with comparatively higher molecular weight (> 425 Da) displayed higher VIP values (Fig. 3a), suggesting that they play an important role for the prediction model. The fact that the most diagnostic compounds in forecasting the SOC levels selectively correspond to those with high MW could suggest the existence of a threshold of complexity in the FTICR-MS biomarkers to retain sufficient structural information to behave as chemometric proxies of emergent soil properties such as its potential for C sequestration. As a whole, when comparing the chemical structure of the diagnostic molecules (Fig. 3b) located in the van Krevelen diagram, high VIP values are located in the lipid region. These compounds could be considered as surrogates for biodegradability and humification stages and may represent a repository of diagnostic molecules informing both on the structural constituents of higher plants (epicuticular waxes, plant tissues...) or microorganisms involved into soil C storage (Nip et al., 1986; Gocke et al., 2013).

As a whole, the PLS results support the idea that the composition of HAs still retains relevant information after alkali extraction, as evidenced by the relationships found with SOM formation processes. In addition, the PLS model confirms that the composition of the HAs is related to soil properties, in this case SOC content. This is summarized in Fig. 4, showing the pattern resulting from the subtraction of compound abundances of the two average spectra from HAs in extreme quartiles according the SOC distribution. The plot showed the compound groups most characteristic of the structure of HAs from soils behaving as C sinks (blue colour)

compared to those that showed greater C concentration in the soils comparatively behaving as hot spots (red color), the most significant difference according to Student's test corresponds to the abundances of structures derived from lipids and lignin. This corroborates the previous differences in the VIPs calculated by PLS, and suggested that these two regions in the van Krevelen diagram are the most diagnostic source of HA compositional indicators informing on SOC accumulation. As a whole, the predominance of unsaturated and alicyclic lipid was observed in HAs from soils with high SOC content, whereas mainly saturated alkyl molecules were in characteristically high proportions in the HAs from soils of low C content. In the case of lignin-derived compounds, its predominance is indicative of HAs from soils with high content of SOC. Similar result was found by Jiménez-González et al. (2019). This suggests that these compounds, indicating recent input of plant material, are useful to explain to large extent the variability of the SOC levels in our soils (Jiménez-González et al., 2017).

5. Conclusions

The chemometric assessment based on compounds detected by FTICR-MS depicts the HA structure as a fingerprint for SOC levels. A significant relationship exists between the SOC levels and the molecular composition of HAs studied by FTICR-MS, to the extent that it is possible to obtain a very significant prediction model using only some common compounds released by FTICR-MS from the HAs. The fact that the structural features of the HAs are linked to the SOC content of the corresponding soils indicated that HAs ought not to be considered artefacts generated in the alkali extraction, because its composition still reflects

biogeochemical information relevant about soil emergent properties as the SOC content even after structural changes inherent to its isolation and purification procedures.

The information window provided by FTICR-MS could be considered a valuable source of surrogates in the form of compounds reflecting the effects of the environmental factors controlling the different SOC levels. Soils with high SOC levels accumulate HAs with structures rich in unsaturated lipid and lignin-derived compounds, whereas HAs in soils with low SOC content present a composition in which saturated lipid compounds predominate.

Acknowledgements

Financial support by Spanish CICYT (grant CGL2013-43845-P) is gratefully acknowledged. Marco A. Jiménez-González thanks the Spanish Ministry of Economy and Competitiveness (MINECO) for funding his pre-doctoral fellowship (BES-2014-069238). The authors would like to thank Dr. Koegel-Knabner, Associate Editor of Organic Geochemistry and two anonymous reviewers for their valuable comments which helped to improve the manuscript.

References

Aitchison, J., 1986. The statistical analysis of compositional data. In: Monographs on Statistics and Applied Probability. London: Chapman & Hall.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716–723. DOI: 10.1109/TAC.1974.1100705

Almendros, G., Hernández, Z., Sanz, J., Rodríguez-Sánchez, S., Jiménez-González, M.A., González-Pérez, J.A., 2018. Graphical statistical approach to soil organic matter resilience using analytical pyrolysis data. *Journal of Chromatography A* 1533, 164–173. <https://doi.org/10.1016/j.chroma.2017.12.015>

Batjes, N.H., 1996. Total carbon and nitrogen in the soils of the world. *European Journal of Soil Science* 47, 151–163. <https://doi.org/10.1111/j.1365-2389.1996.tb01386.x>

De la Rosa, J.M., Jiménez-Morillo, N.T., González-Pérez, J.A., Almendros, G., Vieira, D., Knicker, H.E., Keizer, J., 2019. Mulching-induced preservation of soil organic matter quality in a burnt eucalypt plantation in central Portugal. *Journal of Environmental Management* 231, 1135–1144. <https://doi.org/10.1016/j.jenvman.2018.10.114>

Derenne, S., Quénéa, K., 2015. Analytical pyrolysis as a tool to probe soil organic matter. *Journal of Analytical and Applied Pyrolysis* 111, 108–120. <https://doi.org/10.1016/j.jaap.2014.12.001>

DiDonato, N., Chen, H., Waggoner, D., Hatcher, P.G., 2016. Potential origin and formation for molecular components of humic acids in soils. *Geochimica et Cosmochimica Acta* 178, 210–222. <https://doi.org/10.1016/j.gca.2016.01.013>

Faria, S.R., De la Rosa, J.M., Knicker, H., González-Pérez, J.A., Villaverde, J., Keizer, J.J. 2015. Wildfire-induced alterations of topsoil organic matter and their

recovery in Mediterranean eucalypt stands detected with biogeochemical markers. European Journal of Soil Science 66, 699–713. <https://doi.org/10.1111/ejss.12254>

Feng, W., Plante, A.F., Aufdenkampe, A.K., Six, J. 2014. Soil organic matter stability in organo-mineral complexes as a function of increasing C loading. Soil Biology and Biochemistry 69, 398–405. <https://doi.org/10.1016/j.soilbio.2013.11.024>

Fernández-Getino, A.P., Hernández, Z., Piedra Buena, A., Almendros, G., 2013. Exploratory analysis of the structural variability of forest soil humic acids based on multivariate processing of infrared spectral data. European Journal of Soil Science 64, 66–79. <https://doi.org/10.1111/ejss.12016>

Gocke, M., Kuzyakov, Y., Wiesenberger, L.B., 2013. Differentiation of plant derived organic matter in soil, loess and rhizoliths based on *n*-alkane molecular proxies. Biogeochemistry 112, 23–40. <https://doi.org/10.1007/s10533-011-9659-y>

González-Vila, F.J., Lüdemann, H-D., Martín, F., 1983. ¹³C-NMR structural features of soil humic acids and their methylated, hydrolyzed and extracted derivatives. Geoderma 31, 3–15. DOI: 10.1016/0016-7061(83)90080-0

Ikeya, K., Sleighter, R.L., Hatcher, P.G., Watanabe, A., 2015. Characterization of the chemical composition of soil humic acids using Fourier transform ion cyclotron resonance mass spectrometry. Geochimica et Cosmochimica Acta 153, 169–182. <https://doi.org/10.1016/j.gca.2015.01.002>

IUSS Working Group WRB (2014) World Reference Base for Soil Resources 2014. International soil classification system for naming soils and creating legends for soil maps. World Soil Resources Reports No. 106. FAO, Rome.

Jiménez-González, M.A., Álvarez, A.M., Carral, P., González-Vila, F.J., Almendros G., 2017. The diversity of methoxyphenols released by pyrolysis-gas chromatography as predictor of soil carbon storage. *Journal of Chromatography A* 1508, 130–137. <https://doi.org/10.1016/j.chroma.2017.05.068>

Jiménez-González, M.A., Álvarez, A.M., Hernández, Z., Almendros, G., 2018. Soil carbon storage predicted from the diversity of pyrolytic alkanes. *Biology and Fertility of Soils* 54, 617–629. <https://doi.org/10.1007/s00374-018-1285-6>

Jiménez-González, M.A., Álvarez, A.M., Carral, P., Almendros G., 2019. Chemometric assessment of soil organic matter storage and quality from humic acid infrared spectra. *Science of the Total Environment* 685, 1160–1168. <https://doi.org/10.1016/j.scitotenv.2019.06.231>

Jiménez-Morillo, N.T., González-Pérez, J.A., Almendros, G., De la Rosa, J.M., Waggoner, D.C., Jordán, A., Zavala, L.M., González-Vila F.J., Hatcher, P.G., 2018. Ultra-high resolution mass spectrometry of physical speciation patterns of organic matter in fire-affected soils. *Journal of Environmental Management* 225, 139–147. <https://doi.org/10.1016/j.jenvman.2018.07.069>

Kamjunke, N., von Tümpling, W., Hertkorn, N., Harir, M., Schmitt-Kopplin, P., Norf, H., Weitere, M., Herzsprung, P., 2017. A new approach for evaluating transformations of dissolved organic matter (DOM) via high-resolution mass spectrometry and relating it to bacterial activity. *Water Research* 123, 513–523. <https://doi.org/10.1016/j.watres.2017.07.008>

Kim, S., Kramer, R.W., Hatcher, P.G., 2003. Graphical method for analysis of ultra-high resolution broadband mass spectra of natural organic matter, the van Krevelen diagram. *Analytical Chemistry* 75, 5336–5344. DOI: 10.1021/ac034415p.

Knicker, H., 2011. Solid state CPMAS ^{13}C and ^{15}N NMR spectroscopy in organic geochemistry and how spin dynamics can either aggravate or improve spectra interpretation. *Organic Geochemistry* 42, 867–890. <https://doi.org/10.1016/j.orggeochem.2011.06.019>

Koch, B.P., Dittmar, T., 2006. From mass to structure: an aromaticity index for high-resolution mass data of natural organic matter. *Rapid Communications in Mass Spectrometry* 20, 926–932. <https://doi.org/10.1002/rcm.7433>

Kögel-Knabner, I., Guggenberger, G., Kleber, M., Kandeler, E., Kalbitz, K., Scheu, S., Eusterhues, K., Leinweber, P. 2008. Organo-mineral associations in temperate soils: integrating biology, mineralogy, and organic matter chemistry. *Journal of Plant Nutrition and Soil Science* 171, 61–82. <https://doi.org/10.1002/jpln.200700048>

Kramer, R.W., Kujawinski, E.B., Hatcher, P. G., 2004. Identification of black carbon derived structures in a volcanic ash soil humic acid by Fourier transform ion cyclotron resonance mass spectrometry. *Environmental Science & Technology* 38, 3387–3395. DOI: 10.1021/es030124m.

Lal, R., 2004. Soil carbon sequestration to mitigate climate change. *Geoderma* 123, 1–22. <https://doi.org/10.1016/j.geoderma.2004.01.032>

McKee, G.A., Hatcher, P.G., 2015. A new approach for molecular characterisation of sediments with Fourier transform ion cyclotron resonance mass spectrometry: Extraction optimisation. *Organic Geochemistry* 85, 22–31. <https://doi.org/10.1016/j.orggeochem.2015.04.007>

Miller, K.E., Lai, C.-T., Friedman, E.S., Angenent L.T., Lipson D.A., 2015. Methane suppression by iron and humic acids in soils of the Arctic Coastal Plain. *Soil Biology and Biochemistry* 83, 176–183. <https://doi.org/10.1016/j.soilbio.2015.01.022>.

Miralles, I., Piedra-Buena, A., Almendros, G., González-Vila, F.J., González-Pérez, J.A., 2015. Pyrolytic appraisal of the lignin signature in soil humic acids: Assessment of its usefulness as carbon sequestration marker. *Journal of Analytical and Applied Pyrolysis* 113, 107–115. <https://doi.org/10.1016/j.jaap.2014.11.010>

Nip, M., Tegelaar, E.W., de Leeuw, J.W., Schenck, P.A., Holloway P. J., 1986. A new non-saponifiable highly aliphatic and resistant biopolymer in plant cuticles. *Naturwissenschaften* 73, 579–585. <https://doi.org/10.1007/BF00368768>

Panettieri, M., Knicker, H., Murillo, J.M., Madejón, E., Hatcher, P.G., 2014. Soil organic matter degradation in an agricultural chronosequence under different tillage regimes evaluated by organic matter pools, enzymatic activities and CPMAS ^{13}C NMR. *Soil Biology and Biochemistry* 78, 170–181.

<https://doi.org/10.1016/j.soilbio.2014.07.021>.

Parton, W.J., Schimel, D.S., Cole C.V., Ojima, D.S., 1987. Analysis of factors controlling soil organic matter levels in Great Plains grasslands. *Soil Science Society of America Journal* 51, 1173–1179. doi:10.2136/sssaj1987.03615995005100050015x

Piccolo, A., Spaccini, R., Drosos, M., Vinci, G., Cozzolino, V., 2018. The molecular composition of humus carbon: recalcitrance and reactivity in soils, Garcia, C., Nannipieri, P., Hernandez, T. (Eds.), *The future of soil carbon*. Academic Press pp. 87–124. <https://doi.org/10.1016/B978-0-12-811687-6.00004-3>

Simonetti, G., Francioso, O., Dal Ferro, N., Nardi, S., Berti, A., Morari, F., 2017. Soil porosity in physically separated fractions and its role in SOC protection. *Journal of Soils and Sediments* 17, 70–84. <https://doi.org/10.1007/s11368-016-1508-0>

Sleighter, R.L., Hatcher, P.G., 2007. The application of electrospray ionization coupled to ultrahigh resolution mass spectrometry for the molecular characterization of natural organic matter. *Journal of Mass Spectrometry* 42, 559–574.

<https://doi.org/10.1002/jms.1221>

Sleighter, R.L., Mckee, G.A., Liu, Z., Hatcher, P.G., 2008. Naturally present fatty acids as internal calibrants for Fourier transform mass spectra of dissolved organic matter. *Limnology and Oceanography Methods* 6, 246–253.

<https://doi.org/10.4319/lom.2008.6.246>

Solomon, D., Lehmann, J., Thies, J., Schäfer, T., Liang, B., Kinyangi, J., Neves, E., Petersen, J., Luizão, F., Skjemstad, J., 2007. Molecular signature and sources of biochemical recalcitrance of organic C in Amazonian Dark Earths. *Geochimica et Cosmochimica Acta* 71, 2285–2298. <https://doi.org/10.1016/j.gca.2007.02.014>

Song, X-Y., Liu, S-T., Liu, Q-H., Zhang, W-J., Hu, C-G., 2014. Carbon sequestration in soil humic substances under long-term fertilization in a wheat-maize system from North China. *Journal of Integrative Agriculture* 13, 562–569.

[https://doi.org/10.1016/S2095-3119\(13\)60713-3](https://doi.org/10.1016/S2095-3119(13)60713-3)

Spaccini, R., Piccolo, A., Conte, P., Haberhauer, G., Gerzabek, M.H., 2002. Increased soil organic carbon sequestration through hydrophobic protection by humic substances. *Soil Biology and Biochemistry* 34, 1839–1851.

[https://doi.org/10.1016/S0038-0717\(02\)00197-9](https://doi.org/10.1016/S0038-0717(02)00197-9)

Spaccini, R., Piccolo, A., 2012. Carbon sequestration in soils by hydrophobic protection and In situ catalyzed photo-polymerization of soil organic matter (SOM): Chemical and physical–chemical aspects of SOM in field plots. In: Piccolo A. (eds) *Carbon Sequestration in Agricultural Soils*. Springer, Berlin, Heidelberg.

https://doi.org/10.1007/978-3-642-23385-2_4

Stevenson, F.J., 1994. Humus Chemistry: Genesis, Composition, Reactions. Wiley, New York.

Stubbins, A., Spencer, R.G.M., Chen, H., Hatcher, P.G., Mopper, K., Hernes, P.J., Mwamba, V.L., Mangangu, A.M., Wabakanghanzi, J.N., Six, J., 2010. Illuminated darkness: Molecular signatures of Congo River dissolved organic matter and its photochemical alteration as revealed by ultrahigh precision mass spectrometry. *Limnology and Oceanography* 55, 1467–1477.

<https://doi.org/10.4319/lo.2010.55.4.1467>

Sutton, R., Sposito, G., 2005. Molecular Structure in Soil Humic Substances: The New View. *Environmental Science & Technology* 39, 23, 9009-9015.

<https://doi.org/10.1021/es050778q>

Tadini, A.M., Constantino, I.C., Nuzzo, A., Spaccini, R., Piccolo, A., Moreira, A.B., Bisinoti, M.C., 2015. Characterization of typical aquatic humic substances in areas of sugarcane cultivation in Brazil using tetramethylammonium hydroxide thermochemolysis. *Science of the Total Environment* 518–519, 201–208.

<https://doi.org/10.1016/j.scitotenv.2015.02.103>

Tinoco, P., Almendros, G., González-Vila F.J., Sanz, J., González-Pérez, J.A., 2014. Revisiting molecular characteristics responsive for the aromaticity of soil humic acids. *Journal of Soils and Sediments* 15, 781–791. [https://doi.org/10.1007/s11368-014-](https://doi.org/10.1007/s11368-014-1033-y)

[1033-y](https://doi.org/10.1007/s11368-014-1033-y)

Tinoco, P., Almendros, G., Sanz, J., 2018. Soil perturbation in Mediterranean ecosystems reflected by differences in free-lipid biomarker assemblages. *Journal of Agricultural and Food Chemistry* 66, 9895–9906. DOI: 10.1021/acs.jafc.8b01483.

Van Krevelen, D.W., 1950. Graphical-statistical method for the study of structure and reaction processes of coal. *Fuel* 29, 269–284.

Viscarra Rossel, R.P., 2008. ParLeS: Software for chemometric analysis of spectroscopic data. *Chemometrics and Intelligent Laboratory Systems* 90, 72–83.
<https://doi.org/10.1016/j.chemolab.2007.06.006>

Waggoner, D.C., Chen, H., Willoughby, A.S., Hatcher, P., 2015. Formation of black carbon-like and alicyclic aliphatic compounds by hydroxyl radical initiated degradation of lignin. *Organic Geochemistry* 82, 69–76.
<https://doi.org/10.1016/j.orggeochem.2015.02.007>

Table 1 Location, classification and general characteristics of the soils

Sample No.	Geographical coordinates	Soil type IUSS Working Group WRB (2014)	Soil texture	Vegetation	SOC ^a (g·kg ⁻¹)	Soil C/N ratio
1	40°33'N 4°8'W	Dystric Cambisol (Humic)	Sandy loam	<i>Quercus pyrenaica</i>	41	11.3
2	41°7'N 3°34'W	Haplic Umbrisol (Hyperhumic)	Sandy loam	<i>Pinus sylvestris</i>	67	14.8
3	40°23'N 3°16'W	Calcaric Cambisol (Humic)	Silt loam	<i>Quercus ilex</i>	96	15.3
4	40°53'N 3°34'W	Gleyic Cambisol (Humic)	Sandy loam	<i>Fraxinus angustifolia</i>	87	13.3
5	40°53'N 3°34'W	Dystric Cambisol (Humic)	Sandy loam	<i>Paeonia coriacea</i> ,	48	13.1
6	42°36'N 3°11'W	Calcic Chernozem (Pachic)	Sandy clay loam	Pastureland with: <i>Micropyrum tenellum</i> , <i>Trifolium dubium</i> , <i>Trifolium campestre</i> , <i>Anacyclus clavatus</i>	17	13.9
7	40°44'N 3°42'W	Dystric Cambisol (Ochric)	Sandy loam	<i>Quercus rotundifolia</i>	18	16.0
8	40°47'N 2°57'W	Leptic Kastanozems (Hyperhumic)	Clay loam	<i>Quercus rotundifolia</i>	87	13.3
9	41°14'N 3°24'W	Leptic Podzol (Arenic)	Sandy loam	<i>Fagus sylvatica</i>	32	16.4
10	40°44'N 3°48'W	Dystric Cambisol (Humic)	Sandy loam	<i>Pinus sylvestris</i>	140	18.1
11	40°21'N 3°56'W	Dystric Cambisol (Humic)	Loamy sand	<i>Pinus pinea</i>	117	26.7
12	40°33'N 3°43'W	Dystric Cambisol (Loamic)	Loamy sand	<i>Quercus rotundifolia</i>	93	8.9
13	40°54'N 3°28'W	Eutric Cambisol (Humic)	Sandy loam	<i>Juniperus oxycedrus</i>	134	12.1
14	40°52'N 3°34'W	Eutric Cambisol (Humic)	Sandy loam	<i>Juniperus oxycedrus</i>	104	18.5
15	40°58'N 3°37'W	Dystric Cambisol (Humic)	Sandy loam	<i>Pinus pinaster</i>	81	16.9
16	40°54'N 3°53'W	Dystric Cambisol (Humic)	Sandy clay loam	<i>Quercus pyrenaica</i>	55	18.0
17	40°51'N 3°44'W	Dystric Cambisol (Colluvic)	Sandy loam	<i>Pinus sylvestris</i>	39	13.0
18	40°45'N 3°41'W	Leptic Cambisol (Humic)	Loam	<i>Quercus ilex</i>	105	17.0
19	43°15'N 2°51'W	Leptic Umbrisol (Loamic)	Loam	Pastureland for grazing: <i>Brachypodium retusum</i> , <i>Lolium perenne</i> , <i>Trifolium repens</i>	41	15.8
20	43°4'N 2°35'W	Haplic Luvisol (Humic)	Silty clay loam	<i>Fagus sylvatica</i>	44	13.4
21	42°34'N 2°38'W	Eutric Cambisol (Humic)	Clay loam	Pastureland for grazing: <i>Brachypodium retusum</i> , <i>Cynosurus cristatus</i> , <i>Trifolium repens</i>	57	13.9

22	43°15'N 2°51'W	Haplic Umbrisol (Loamic)	Loam	<i>Pinus radiata</i>	27	17.0
23	42°28'N 8°53'W	Leptic Regosol (Humic)	Sandy loam	<i>Pinus pinaster</i>	133	31.0
24	42°36'N 8°38'W	Leptic Regosol (Humic)	Sandy loam	<i>Pinus pinaster</i>	90	20.0
25	43°4'N 8°22'W	Leptic Umbrisol (Hyperhumic)	Loam	<i>Pinus pinaster</i>	132	18.0
26	28°22'N 16°39'W	Vitric Andosol (Hyperhumic)	Sandy loam	<i>Laurus canariensis</i>	18	22.0
27	28°26'N 16°29'W	Leptic Regosol (Arenic)	Clay loam	<i>Euphorbia canariensis</i>	22	12.0
28	28°14'N 16°28'W	Leptic Regosol (Arenic)	Sandy loam	Fallow: <i>Solanum tuberosum</i>	23	14.0
29	28° 9'N 16°38'W	Folic Umbrisol (Chromic)	Sandy loam	<i>Pinus canariensis</i>	105	27.0
30	40°13'N 4°29'W	Dystic Regosol (Arenic)	Sand	<i>Pinus pinea</i>	35	20.0
31	41°29'N 4°19'W	Eutric Cambisol (Humic)	Sand	<i>Pinus pinea</i>	99	25.9
32	40°18'N 4°38'W	Eutric Cambisol (Arenic)	Sandy loam	<i>Quercus rotundifolia</i>	46	16.8
33	41°1'N 3°12'W	Eutric Cambisol (Humic)	Silt loam	<i>Quercus rotundifolia</i>	89	23.7
34	40°58'N 3°44'W	Eutric Cambisol (Humic)	Sandy loam	<i>Juniperus thurifera</i>	157	21.6
35	40°56'N 3°41'W	Dystic Leptosol (Humic)	Loam	<i>Juniperus thurifera</i>	92	13.9

^aSOC: soil organic carbon

Table 2 Elemental composition and peak area integration values of ¹³C NMR spectra of humic acids

Sample No.	C (g·100g ⁻¹)	H (g·100g ⁻¹)	N (g·100g ⁻¹)	S (g·100g ⁻¹)	O (g·100g ⁻¹)	¹³ C NMR alkyl C (%)	¹³ C NMR <i>N</i> -alkyl + OCH ₃ C (%)	¹³ C NMR O-alkyl C (%)	¹³ C NMR arom C (%)	¹³ C NMR carbonyl C (%)
1	54.3	5.0	5.7	0.7	34.3	28.4	10.8	27.4	18.9	14.6
2	55.9	4.4	4.6	0.3	34.8	32.7	8.4	18.7	27.8	12.5
3	57.2	4.8	4.1	0.6	33.3	31.4	11.3	24.2	21.9	11.3
4	56.8	5.8	5.5	0.7	31.2	35.4	11.7	24.8	16.8	11.3
5	56.4	5.5	5.7	0.9	31.4	31.6	11.8	24.8	19.1	12.7
6	53.7	3.8	4.6	0.7	37.2	30.8	9.6	14.3	32.3	13.0
7	56.0	5.6	4.3	0.4	33.7	41.1	9.7	21.0	18.1	10.1
8	57.3	4.7	4.3	0.4	33.2	32.9	9.9	22.2	22.1	12.9
9	57.8	5.2	4.9	0.4	31.7	37.2	10.4	21.4	18.1	12.9
10	57.6	5.1	3.2	0.3	33.8	32.4	9.6	23.1	24.1	10.9
11	57.2	4.8	3.4	0.4	34.3	31.0	11.0	25.2	22.4	10.5
12	55.8	5.0	4.8	0.4	33.9	33.2	10.7	26.6	16.6	12.8
13	56.3	5.1	3.1	0.5	35.0	36.0	9.9	24.6	18.1	11.5
14	52.3	4.6	2.9	0.4	39.8	31.0	9.4	26.3	21.0	12.2
15	59.1	5.6	3.6	0.4	31.3	34.8	10.2	25.2	19.2	10.6
16	57.5	5.3	3.8	0.4	33.0	36.1	10.2	25.7	17.4	10.7
17	54.9	4.8	4.9	0.6	34.8	33.9	10.5	25.3	18.9	11.5
18	55.0	4.8	4.4	0.5	35.2	32.3	11.8	24.7	18.6	12.7
19	54.1	4.5	3.7	0.4	37.3	37.0	8.9	20.1	22.6	11.4
20	55.8	5.1	4.4	0.5	34.2	33.8	11.2	25.7	18.2	11.1
21	55.0	4.6	3.7	0.4	36.3	31.3	10.3	26.5	19.1	12.8
22	52.9	4.2	3.4	0.3	39.2	41.5	8.8	18.6	20.6	10.5
23	57.8	4.0	4.2	0.3	33.8	32.0	8.8	20.2	26.4	12.7
24	56.9	4.6	3.2	0.3	35.0	35.0	8.9	22.3	21.7	12.0
25	54.3	4.8	4.2	0.4	36.3	29.7	10.3	26.0	21.9	12.1
26	50.9	3.3	4.3	0.5	41.0	25.5	7.9	16.8	35.1	14.8
27	57.8	4.9	5.4	0.7	31.2	37.0	10.9	19.2	20.4	12.5
28	50.5	4.3	5.0	0.8	39.4	31.9	11.3	21.7	22.6	12.5
29	51.6	4.4	3.0	0.3	40.8	32.9	10.1	22.5	24.1	10.4
30	58.3	5.5	4.1	0.4	31.8	39.5	10.5	21.4	18.4	10.3
31	58.3	5.2	3.1	0.4	33.0	36.8	10.2	22.4	20.8	9.8
32	56.0	4.3	3.6	0.3	35.7	31.5	8.7	21.3	25.0	13.6
33	57.9	5.1	3.1	0.3	33.5	35.4	10.3	24.1	20.3	9.8
34	57.0	5.0	3.7	0.4	33.9	37.3	10.9	23.7	18.8	9.4
35	55.0	4.8	4.3	0.4	35.4	34.8	11.0	25.0	17.8	11.5

Table 3 FTICR results and peak assignments

Sample No.	Total peaks	Identified compounds	Identified compounds (%)	Common compounds intensity from total (%)	CHO	CHON	CHONS	CHOP	CHOPN	CHOPNS	CHOPS	CHOS
1	2546	2124	83	21	1639	403	29	11	20	5	2	15
2	2951	2490	84	12	1745	662	23	1	40	2	3	14
3	2396	2021	84	26	1477	474	26	3	20	0	0	21
4	2836	2419	85	36	1946	374	29	14	20	4	1	31
5	2758	2388	87	16	1884	419	19	14	24	3	0	25
6	1449	1149	79	20	808	302	16	8	3	5	1	6
7	3009	2518	84	22	2136	304	24	3	32	2	2	15
8	2968	2366	80	41	1978	239	39	16	78	1	1	14
9	3701	2915	79	34	2388	350	46	13	76	1	1	40
10	4476	3487	78	11	2849	478	31	6	91	1	4	27
11	4001	3076	77	12	2471	461	51	7	62	4	5	15
12	2814	2277	81	33	1819	312	47	9	58	1	0	31
13	3318	2639	80	20	2107	402	35	2	62	1	9	21
14	2642	2099	79	13	1648	345	22	5	37	0	5	37
15	3366	2710	81	17	2357	222	26	10	45	5	4	41
16	2347	1814	77	40	1619	63	32	14	68	2	0	16
17	2742	2116	77	24	1664	297	36	10	52	1	0	56
18	1745	1309	75	43	1048	163	27	14	28	1	1	27
19	4005	3130	78	23	2496	469	49	5	73	5	2	31
20	3008	2317	77	46	1928	262	41	17	41	2	2	24
21	2117	1723	81	35	1422	225	23	10	23	1	0	19
22	3426	2881	84	12	2360	412	31	3	60	6	2	7
23	3125	2594	83	10	1790	722	24	2	31	10	1	14
24	3609	3148	87	12	2551	529	22	2	23	5	0	16
25	3499	3004	86	11	2255	694	15	0	25	5	0	10
26	2222	1616	73	10	885	671	24	2	10	8	0	16
27	2248	2010	89	15	1490	473	9	7	9	2	1	19
28	1829	1555	85	16	1212	311	13	6	1	1	0	11
29	3330	2889	87	10	2557	250	22	1	26	4	3	26
30	3477	2954	85	10	2673	210	23	2	22	4	2	18
31	3696	3128	85	14	2861	187	15	2	34	6	2	21
32	3345	2857	85	13	2316	465	23	9	13	11	0	20
33	3584	3078	86	26	2708	263	29	6	40	4	1	27
34	3378	2776	82	12	2395	291	25	4	25	3	1	32
35	1684	1409	84	12	1304	48	11	8	6	3	0	29

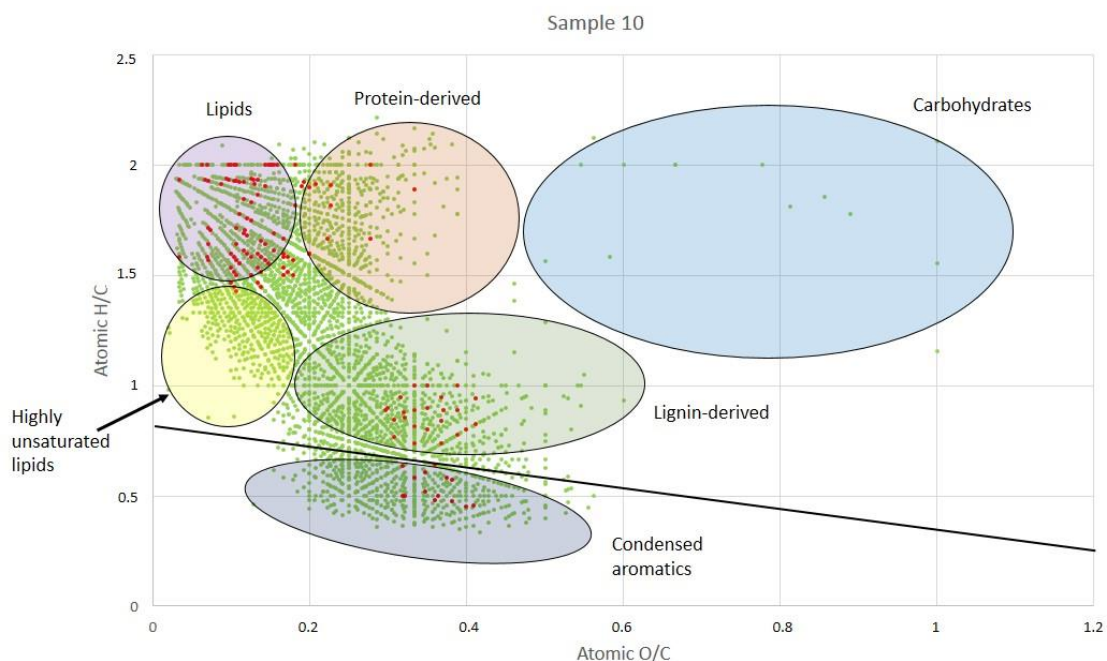


Figure 1 Van Krevelen diagram of total compounds detected by FTICR-MS in sample 10 (3487 compounds) with the 131 common compounds identified in all the HAs of different soils marked in red. The regions defined by the characteristic H/C and O/C atomic ratios of different chemical structures (lipid, protein- and lignin-derived, carbohydrates and condensed aromatics) are indicated in the diagram. The black solid line denotes the regression line for formulas with $AI_{mod}=0.67$ (condensed aromatic molecules are below the line)

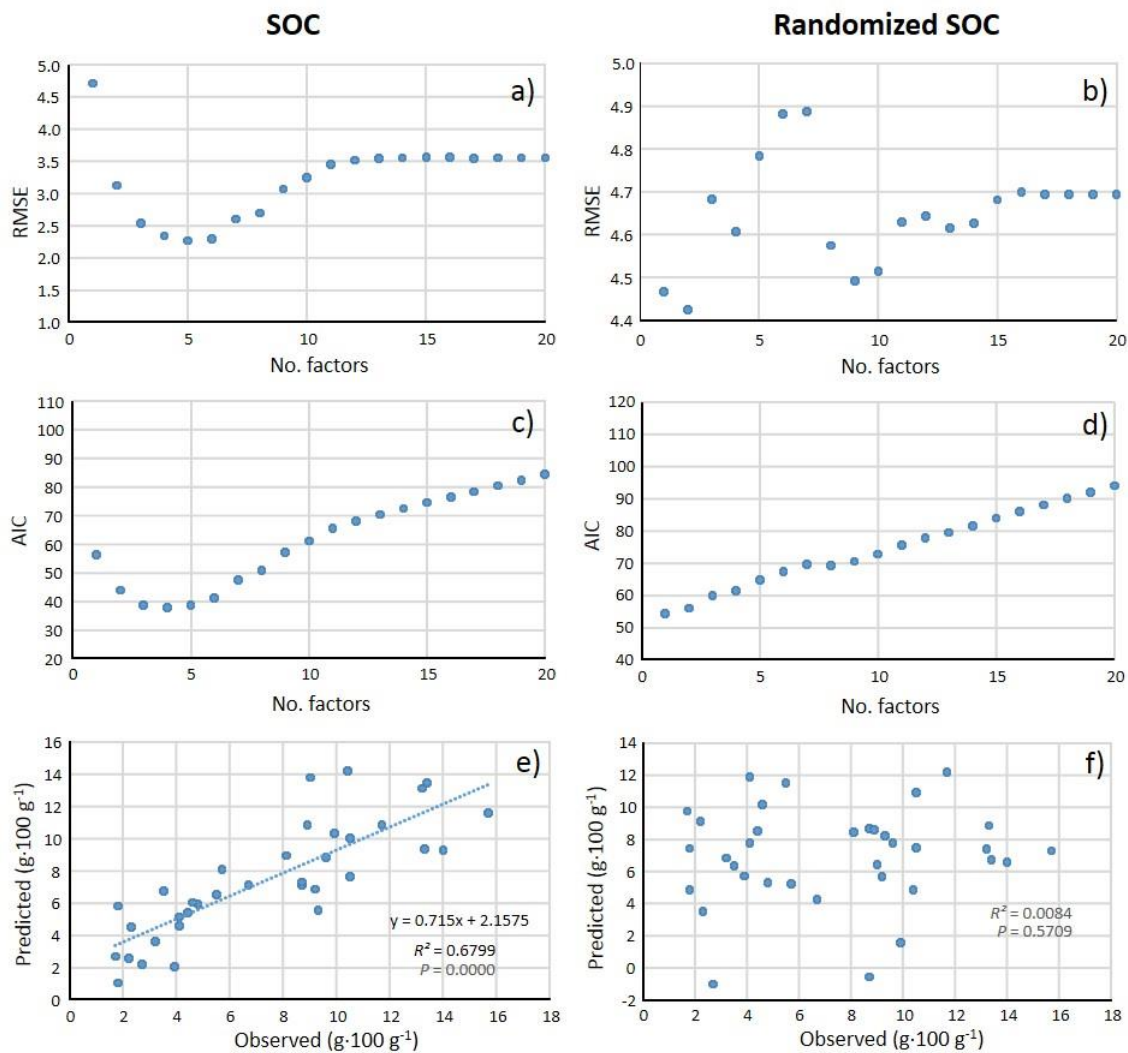


Figure 2 Cross-validation plots (experimental vs predicted values) corresponding to partial least squares (PLS) models to predict the concentration of soil organic carbon (SOC). Comparison of results of experimental values of SOC (a, c, e) and randomized SOC values (b, d, f). Root Mean Squared Error (RMSE) of SOC (a) and randomized SOC (b) respectively. Akaike Information Criterion (AIC) (c, d). Observed SOC values vs predicted values obtained by the PLS prediction model for SOC (e) ($R^2 = 0.6799$) and randomized values of SOC (f) ($R^2 = 0.0084$) using 5 latent variables or factors suggested by the RMSE and AIC values.

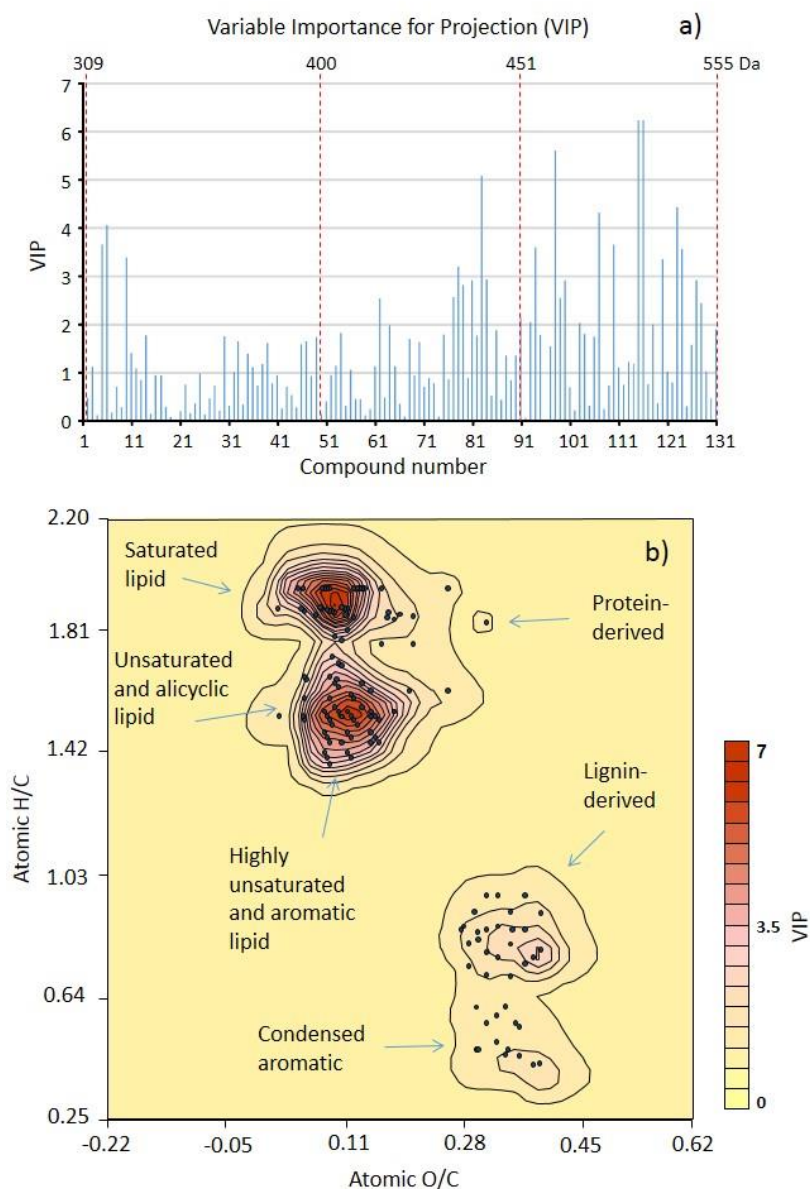


Figure 3 Variable importance for projection (VIP) for the 131 common compounds used as descriptors in the partial least squares (PLS) model to predict the soil organic carbon (SOC) content (ordered by molecular weight in x axis) in a); van Krevelen diagram simultaneously showing the location of the common compounds and the VIPs values shown with colour intensity as a contour diagram in b).

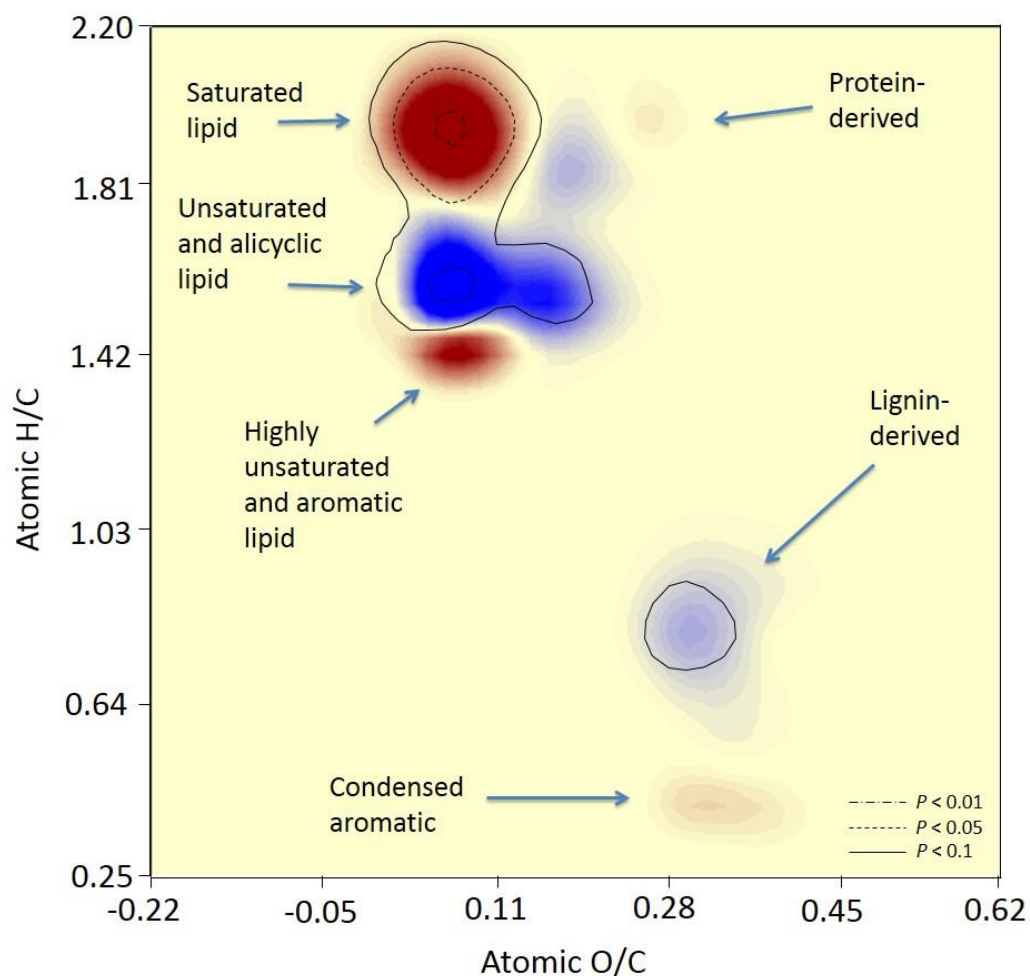


Figure 4. Comparison of abundances of common compounds obtained by FTICR in terms of the SOC levels of the corresponding soils. The van Krevelen diagram shows the subtraction values between average compounds composition in extreme quartiles as regard the SOC content of the corresponding soils. Q1–Q4 values with positive values are shown in blue (Q1, high SOC level) and negative values in red (Q4, low SOC content). The superimposed contour map shows the Student's test values showing values statistically different ($P < 0.1$) between the proportions of compounds in the soils with comparatively high C storage potential as regards the comparatively C depleted soils.