Chemometric exploration of APPI(+)-FT-ICR MS data sets for a comprehensive study of aromatic sulfur compounds in gas oils

Julie Guillemant†, Florian Albrieux†, Marion Lacoue-Nègre†*, Luis Pereira de Oliveira†, Jean-François Joly† and Ludovic Duponchel‡

† IFP Energies nouvelles, Rond-point de l'échangeur de Solaize, BP 3, 69360 Solaize, France

‡ Univ. Lille, CNRS, UMR 8516 - LASIR – Laboratoire de Spectrochimie Infrarouge et Raman, F-59000 Lille, France

ABSTRACT: Sulfur content in gas oils is strictly regulated by legal specifications for environmental reasons. Gas oils are composed of various aromatic sulfur compounds and some of them are known to be very refractory for the sulfur removal reactions. Thus, an accurate analysis of the sulfur compounds is important to find the appropriate operating conditions of the gas oil hydrotreating processes. Aromatic sulfur compounds contained in 23 gas oils samples were analyzed using APPI(+)-FT-ICR MS considering six replicates. Significant differences were spotted within several processed gas oils. A comparison of one feed and its corresponding effluents also confirmed the well-known refractory character of sulfur compounds such as poly alkylated dibenzothiophenes. To go deeper in the molecular exploration, chemometric tools were applied on this spectral dataset including Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA). A unique data re-arrangement was performed directly inspired on DBE vs carbon number plots that are systematically used in petroleomics studies. Then, these chemometric tools provided a successful classification of each type of gas oils. The PCA model has been also validated on mixed blends allowing us to conclude that it could be applied to unknown samples in order to identify the process used to produce them. Moreover, the exploration of the generated loadings revealed key types of molecules driving the classification such as C3-DBT which is a dibenzothiophene core with three additional carbon atoms. Indeed, it is known to remain mainly in deeply hydrotreated samples, validating previous observations regarding its potential refractory character. The ability of chemometric tools to extract specific molecular information from ultra-high resolution MS spectra reveals its huge potential for an exhaustive study of highly complex mixtures such as crude oils.

Petroleum market trends are focused on light products such as gasoline, kerosene and gas oil (GO), whereas their proportions obtained from atmospheric distillation of the crude oils (gas oils cuts called straight run and noted as SRGO) are smaller and smaller[1]. Increasingly heavy crude oils imposed refiners to improve conversion processes in order to produce lighter oils. Gas oils can be produced with several conversion processes such as fluid catalytic cracking (FCC, producing gas oils cuts called LCO), coking (producing gas oil cuts called GOCK) and catalytic hydroconversion (producing gas oil cuts called FBGO[2] with fixed-bed technology or EBGO[3] samples with ebullating-bed technology). However gas oils produced by these industrial processes are considered of poor quality[4,5] due to high sulfur content. Therefore, refiners undergo a hydrodesulfurization process (called HDS) to remove the sulfur compounds of the gas oils in order to meet the environmental legislations of the commercial on-road diesel[6]. Nevertheless, some of these sulfur-compounds are very refractive to hydrodesulfurization like dibenzothiophenes (DBT) which are not fully converted and remain in hydrotreated samples[6–8]. An exhaustive characterization of these refractory compounds could give more information about preferential reactivity pathways and potentially improve hydrotreatment catalysts efficiency as well as hydrotreatment process modeling[9,10]. To do this, we propose to use high-resolution mass spectrometry (FT-ICR MS)[11,12] which is a very helpful tool for petroleum characterization at the molecular level largely reported in the petroleomics literature[13–16]. In petroleomics studies, aromatic sulfur compounds are often characterized by combining atmospheric pressure photo-ionization (APPI) source in positive ion mode and FT-ICR MS[17–19] generating thousands of unequivocally identified molecular species.

Big data sets generated by FT-ICR MS analysis strongly limit the efficiency of classical univariate data processing tools when numerous samples need to be compared. As a consequence, we propose to use without a priori multivariate data analysis tools for understanding and highlighting differences or similarities between several processed gas oils. Among available chemometric methods, principal component analysis (PCA) is a very useful exploratory method[20,21]. PCA was previously used to identify the defective reservoir after an oil spill based on ESI(-)-FT-ICR MS data[22] or to evaluate the proper additivity of several crude oils mixtures analyzed by APPI(+)-FT-ICR MS[23]. Hierarchical cluster analysis (HCA) is another multivariate method used to compare samples and give a complementary view of the data set[24–26]. It is clear that multivariate analysis of FT-ICR MS spectra are more and

Table 1: Properties of gas oils samples used in this study. (-) indicates that analysis was not available. The ASTM standard used for analysis is mentioned for each property.

Sample	Type (*)	Reference in text	Geographical origin, hydrotreatment feed or mix compositions	Total sulfur (ppm) *Ref. method: ASTM D2622*	Total nitrogen (ppm) *Ref. method: ASTM D4629*	Basic nitrogen (ppm) *Ref. method: ASTM D2896*	Boiling point range (°C) *Ref. method: ASTM D86*
GO 1	SRGO		Middle-East	13555	115	47	219-386
GO 2	SRGO		North Europe	7044	254	100	258-396
GO 3	SRGO		North Europe	10979	350	129	244-396
GO 4	SRGO		Middle East	8892	114	42	221-381
GO 5	SRGO		-	4189	96	48	186-392
GO 6	LCO		-	9496	928	91	199-386
GO 7	LCO		Lybia	11074	1170	49	248-390
GO 8	LCO		-	2231	496	141	166-304
GO 9	GOCK		-	14796	893	404	148-358
GO 10	GOCK		-	12723	838	390	163-371
GO 11	GOCK		-	15314	1200	449	173-375
GO 12	GOCK		-	24270	1260	569	188-401
GO 13	EBGO		-	1248	1719	855	199-429
GO 14	FBGO		-	344	195	121	180-359
GO 15	MIX		65% SRGO (*GO 5*)+35% LCO (*GO 6*)	6400	380	63	189-391
GO 16	MIX		67% GOCK+33% LCO	14004	988	436	151-351
GO 17	HDT		GO 16	190	93	14	184-383
GO 18	HDT		GO 16	261	140	23	187-386
GO 19	MIX	Feed	55% SRGO (*GO5*) +30% LCO (*GO7*) +15% GOCK (*GO11*)	14162	586	122	218-390
GO 20	HDT	HDT 2	GO 19	2813	464	107	211-388
GO 21	HDT	HDT 3	GO 19	626	205	38	209-387
GO 22	HDT	HDT 1	GO 19	3656	723	330	210-389
GO 23	MIX		50% LCO+50% LCO	9125	925	98	206-368

(*): SRGO = Straight Run Gas oil; LCO = Light Cycle Oil; GOCK = Coker Gas Oil; EBGO = gas oil from ebullating bed reactor; FBGO = gas oil from fixed bed reactor; MIX = blended gas oil.

more reported in literature but such studies are mostly focused on the prediction of macroscopic properties and even more they often consider rather small databases or few replicates[23,27]. The proposed work constitutes the exploration of the most exhaustive gas oils database from many different industrial processes. High resolution mass spectra will be used to highlight differences or similarities between gas oils focusing on the sulfur compound characterization. In this study, several gas oils produced from different processes (LCO, GOCK, FBGO and EBGO), obtained directly from atmospheric distillation (SRGO), mixed blends and even several hydrotreated effluents will be analyzed by APPI(+)-FT-ICR MS. In view of further chemometric data processing, every sample will be analyzed considering six technical replicates to assess analysis repeatability. Samples will be compared with each other according to relative abundances of identified families and then by focusing on S1 family depending on types of aromatic sulfur compounds identified. The evolution of aromaticity and number of carbon atoms over increasing hydrodesulfurization severity will be followed. Finally, chemometric methods (PCA and HCA) will be applied to mass spectrometry data to evaluate their ability to classify samples according to their process origin and to identify specific families of molecules explaining this discrimination.

MATERIAL AND METHODS

Sample preparation and FT-ICR MS measurements

23 gas oils obtained from 6 different industrial processes were used in this study, including 5 SRGO, 3 LCO, 4 GOCK, 1 EBGO, 1 FBGO, 5 HDT and 4 blends (MIX). The properties of these samples are shown in Table 1. The ionization (level of dilution, mix of solvents used, vaporization temperature and infusion flow rate) and ions transfer (capillary voltage and tube lens) conditions were optimized with a Design of Experiments (DoE) approach. More details about this procedure can be found in previous work. Thus, it has been possible to find robust, optimal and common conditions for all gas oil samples, maximizing simultaneously the number of S1 identified peaks, the sum of S1 peaks intensities and the m/z ranges. The ratio between S1 (radical ion) and S1[H] (protonated ion) families was also considered to evaluate the ionization percentage of sulfur molecular ions compared to protonated ones[29,30]. To sum up, all samples were diluted to 1% v/v in a mixture of 75% Toluene - 25% Methanol.

Mass spectrometry (MS) analyses were carried out using a LTQ FT Ultra™ system (ThermoFisher Scientific, Bremen Germany) equipped with a 7T magnet (Oxford Instruments) and APPI source (Syagen Technology, Tustin CA, USA) used in positive mode. Mass range was set to *m/z* 98-1000. Spectra were acquired considering 4 µscans, 70 scans, an initial resolution set to 200,000 (transient length of 1.6s) at *m/z*=300

(center of average gas oil mass distribution) and transient signal was recorded. AGC (Automatic Gain Control) target value was set to 500,000 ions and injection time varied between few ms to 100 ms depending on the considered sample. Ionization and ions transfer conditions were optimized considering different gas oils samples in the DoE. Therefore, tube lens, capillary voltage and vaporization temperature were finally fixed to 70 V, 30 V and 250°C respectively. Sheath gas was 20 a.u. and auxiliary gas was 5 a.u. Nitrogen was used in both cases. Mass tuning was first performed using Calmix® (ThermoFisher Scientific, Bremen Germany). External mass calibration was then performed with a home-made sodium formiate clusters solution (sodium formiate from VWR, Fontenay-sous-Bois, France) covering the entire selected mass range (90-1000 Da).

Spectral data processing

Spectral data were processed using several softwares. Firstly, the PeakbyPeak® software (SpectroSwiss SARL, Lausanne, Switzerland) has been used to add the obtained transients. Then, phase correction has been managed using Autophaser® (DPAK apodization algorithm, zero pads set to 2, order of fit set to 3)[31,32]. Absorption-mode spectra were then loaded into PeakbyPeak® for signal-dependent noise thresholding and peak picking generating two output files (.txt and .h5)[33]. For each spectrum, the text file was then submitted to home-made software written in the Matlab environment (called *Kendrick Inside*) to get access to the identification of the different compounds, their molecular formula assignments and generate the corresponding Kendrick mass defect plot of the considered sample[34]. Molecular formula assignment conditions were the following ones: $C_{0-50}H_{0-100}O_{0-2}N_{0-2}S_{0-2}$ with maximum content of heteroatoms in one molecular formula set to 3 and maximum error between theoretical and experimental masses set to 5 ppm in the first round of attribution. Iterative mass recalibration was then processed on the samples with PeakbyPeak® considering S1 family with a maximum mass error set to 1 ppm using the .h5 file previously generated[35]. The workflow used is schematically described in Figure S1 in Supporting Information.

It was assumed that S1 family (family within compounds identified were composed of only one atom of sulfur ionized in radical ion $M^{+\cdot}$ form) was supposed to contain all elementary sulfur. Relative intensities were calculated by multiplying the compound absolute intensity by 100 and divided by the sum of all S1 absolute intensities. Pseudo-concentrations in sulfur were obtained by multiplying relative intensities by the amount of sulfur in the sample. Families were attributed regarding values of double bond equivalent (DBE). The mean of all relative standard deviation (RSD) obtained for the several sulfur families pseudo-concentrations was between 0.3% and 11% for the different gas oils enabling further chemometric data processing.

Chemometric approaches

Two different methods were assessed in this study: PCA as a descriptive method and HCA for clustering purposes.

The key principle of Principal Component Analysis (PCA) is that all information is contained in the study of similarity between samples. Variance is often used to measure this similarity and to characterize variables dispersion between several samples. Thus, variance is described by principal components (PCs) which are linear combinations of initial variables, reflecting samples variance projection. The efficiency of PCA relies on variance maximization between samples and is evaluated by explained variance percentages. Each linear combination contains original variables weighted by coefficients called loadings. Thus, every sample can be described by a linear combination of PCs with new coefficients called scores. At the end of a principal component analysis, each sample is defined by a small number of scores for a given number of significant principal components. A possible comparison of samples is obtained in a second step by plotting scores of samples along given principal components. Moreover, Q residual plots can be used to show how much each variable contributes to the overall Q statistic for each sample. Such contributions can be useful in identifying the variables which contribute most to a given sample's sum-squared residual error (i.e. variables not correctly explained by the considered model). HCA is an unsupervised classification method aimed at creating a natural grouping of samples without prior knowledge about their class membership. A dendrogram is generated to visualize the resulting grouping. Positions of samples in this dendrogram are directly related to the level of similarity or difference at which clusters they belonged to.

Only molecular formulas containing a single atom of sulfur (S1 family) were considered to focus the study on identified sulfur compounds. Replicates were considered as single samples instead of considering the mean of 6 relative intensities for each sample. The aim was to evaluate the replicates repeatability over PCA grouping for a single sample as a complement to calculated RSD values. Figure 1 shows the MS data transformation pipeline. In a first step, a DBE vs carbon number plot (i.e. a 2D representation) is generated for each MS spectrum. The next step consists in the unfolding of each previous plot in order to give a 1D representation, variables being given DBE/carbon number pairs. If no peak is observed for a given DBE/carbon number pair, a zero value is considered in the generated unfolded matrix. As a consequence, prior to chemometric explorations, APPI-FT-ICR MS data was re-arranged into a 138x1250 matrix where 138 correspond to the 23 gas oils samples times 6 replicates and 1250 to the possible combinations of DBE (from 1 to 25) and carbon number (from 1 to 50). Moreover, the matrix has been mean-centered prior statistical analysis.

All models were developed with the PLS_Toolbox version 8.6 for Matlab version R2018b (Eigenvector Research Inc, Wenatchee, WA, USA). Mixed blends samples (GO 15, 16, 19 and 23, see Table 1) were used for validation and all other samples were used for performing PCA. For HCA method, EBGO and FBGO samples were not considered as only one sample was available for each process and no classification was needed. The optimization of the models was performed with venitian-blinds cross-validation (10 data splits, 20 samples per blind, 20 maximum principal components). Optimal number of components for PCA was chosen based on % explained cumulative variance and log(eigenvalues) values. Ward's group method was used for HCA classification[36,37].

For each of these methods, several intensity corrections have been tested such as using peaks absolute intensities, peaks relative intensities or peaks sulfur pseudo-

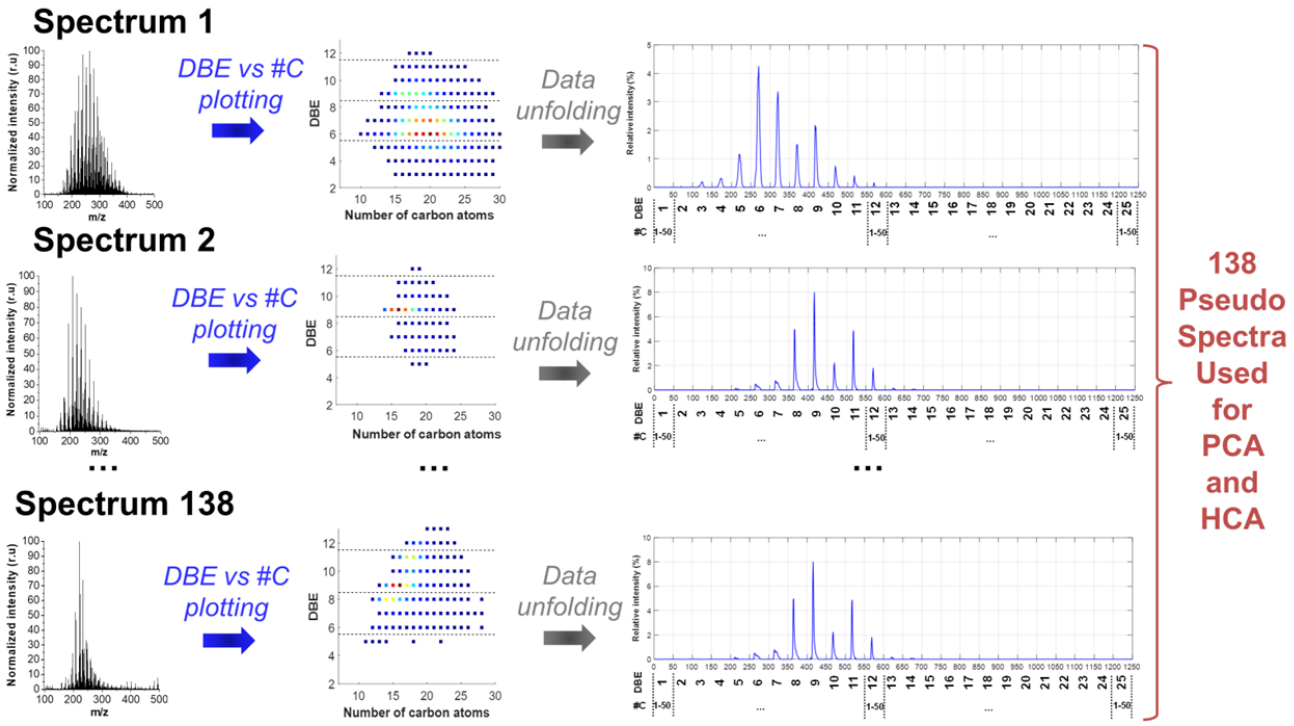


Figure 1: Mass spectrometry data transformation pipeline for chemometric approaches. Spectrum 1 corresponds to mass spectrum obtained from GO 4 (SRGO), spectrum 2 to mass spectrum from GO 22 (HDT) and spectrum 138 to mass spectrum from GO 8 (LCO).

concentrations. However, only relative intensities were finally used for all methods because better groupings between replicates from a given sample were observed reflecting best analysis repeatability.

RESULTS AND DISCUSSION

Part 1: classical mass spectrometry data analysis

Complex spectra were obtained via APPI(+)-FT-ICR MS. About 40k to 50k peaks were identified for each sample and resolution of peaks after phase correction was about 800k to 900k at *m/z* 300. Differences between theoretical masses and experimental masses (mass error) were lower than 50 ppb. An example of mass spectra from three different gas oils is shown in Figure 1 with their corresponding DBE vs carbon number plots.

The considered ionization source produces both molecular ion $M^{+\cdot}$ or protonated ion $[M+H]^+$ making identification tricky as molecular ion are identified in X families while protonated ions are identified as X[H] families. Besides, since ionization does not depend anymore on polarity of the compounds but rather on proton affinity and ionization energy of the molecules, greater number of families are more likely to be identified[30]. In this study, the most abundant heteroatomic classes identified have been selected for comparisons.

One sample from each gas oil class has been randomly selected: GO 4 [SRGO], 6 [LCO], 9 [GOCK], 13 [EBGO], 19 [MIX] and 21 [HDT] (see Table 1) to extract similarities or differences between gas oils production processes. As

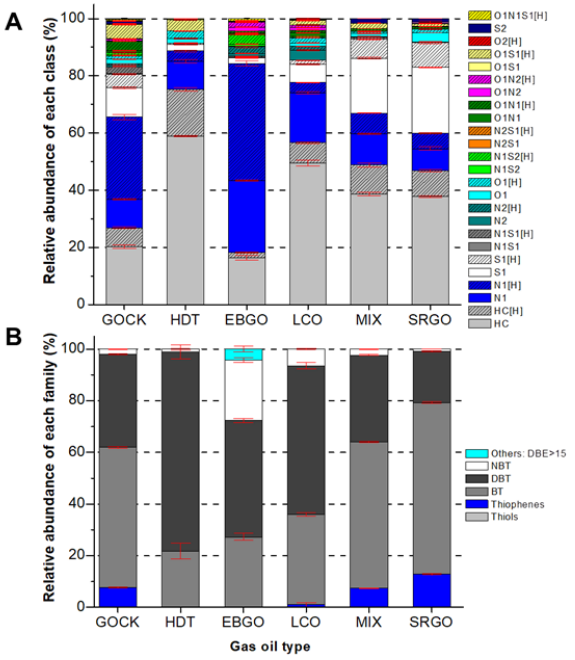


Figure 2: (A) Relative abundance of heteroatomic classes identified in APPI(+) depending on the type of gas oil. Radical ions are identified in X families (full color) and protonated ions are identified in X[H] families (dashes). (B) Relative abundance distribution of sulfur classes as a function of the type of gas oil. NBT = Naphtobenzothiophenes, DBT = Dibenzothiophenes, BT = Benzothiophenes. The standard deviation error bars colored in red have been added on both figures.

confirmed in Figure 2A, APPI(+) is known to be very efficient for hydrocarbons ionization that are of HC and HC[H] classes. The sum of both classes represents up to 50% for HDT, LCO, MIX and SRGO samples. EBGO sample, that has the highest content in basic nitrogen among all samples has the highest abundance in N1[H] class, and similar trends are observed for other gas oils except LCO, with N_{basic} EBGO>N_{basic} GOCK>N_{basic} MIX>N_{basic} LCO>N_{basic} SRGO>N_{basic} HDT. Same ranking is observed for neutral nitrogen classes (class N1) with $N_{neutral}$ EBGO>$N_{neutral}$ LCO>$N_{neutral}$ GOCK> $N_{neutral}$ MIX>$N_{neutral}$ HDT>$N_{neutral}$ SRGO.

Focusing on sulfur-containing classes, it can be stated that S1 relative abundances are lower than expected. Furthermore, no ranking can be observed regarding the sample sulfur content, on the contrary of nitrogen classes. APPI(+) ionizes aromatic sulfur compounds rather than saturated compounds[38] and high proportion of saturated compounds among total sulfur content could explain these incoherencies. Very low relative abundance of S1 class for EBGO sample can be related to relatively low content in sulfur (1248 ppm) and high content in nitrogen (1719 ppm) compared to other samples, resulting in competitive ionization between nitrogen and sulfur compounds.

Deeper insight into identified S1 families within samples gave supplementary information, as seen in Figure 2B. Very low DBE values (thiophenes family) with 13% and low DBE values (Benzothiophenes, BT family) with 67% were mainly observed for SRGO sample. Besides, SRGO sample also contains the lowest abundances in intermediate DBE values (Dibenzothiophenes, DBT family) and high DBE values (Naphtobenzothiophenes, NBT family). Very similar profiles are obtained for GOCK and MIX samples, with about 8% of thiophenes, 55% of BT, 35% of DBT and 1% of NBT. Few compounds with very high DBE (NBT or other classes families) are identified. LCO sample, according to its heavy aromatic character (see Table 1), contains more DBT and NBT compounds than GOCK and MIX samples. HDT sample mainly contains DBT and BT that are known to be refractory. Surprisingly, despite very low abundance of S1 class for EBGO sample, consistent relative abundance distribution of sulfur families is observed. This sample has the highest points range (199-429°C) which is specific to heavy gas oils that contain heavy compounds such as NBT or other compounds with DBE superior to 15.

Catalytic tests were performed with the same catalyst, same pressure and same temperature but at different severity levels or different H2/HC ratios. GO 20 and GO 22 samples were collected at VVH^{-1}=10 and GO 21 at VVH^{-1}=2. Drastic decrease in sulfur compounds is observed even at low severity level (HDT 1) as shown in Figure 3A. At highest severity (HDT 3) no more NBT or Thiophenes are identified, proving hydrotreatment efficiency. Refractory character of BT and DBT is highlighted and especially DBT one[39] reportedly attributed to inhibitor effects from H_2S, basic nitrogen compounds and aromatics[6,40]. Relatively high nitrogen contents in HDT 1 and HDT 2 could also explain hydrodesulfurization lack of efficiency[40,41]. Hence, about 500 ppm of DBT is still quantified in HDT 3 while BT concentration reaches 150 ppm, whereas initial concentration of BT is almost twice higher than DBT one. In Figure 3B, the evolution of sulfur pseudo-concentration as a function of DBE shows that DBT with DBE equals to 9 were major non-converted products within DBT family at highest severity.

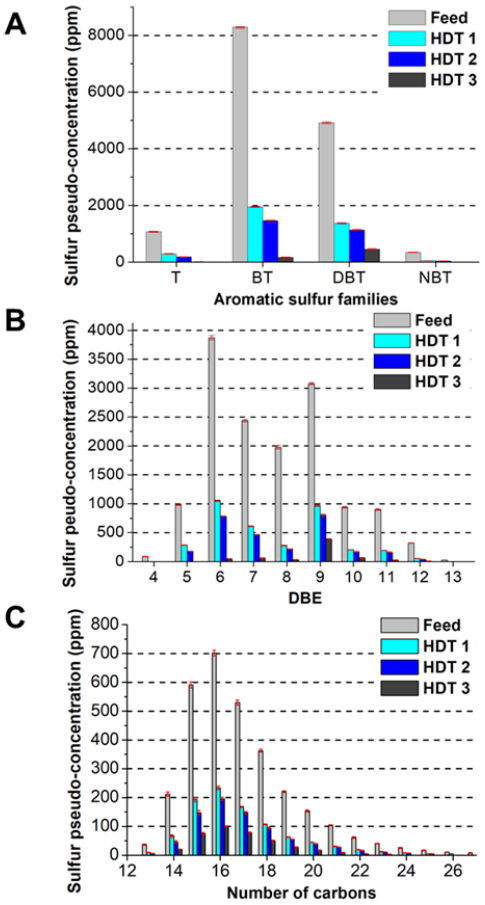


Figure 3: (A) Comparison of sulfur compounds pseudo-concentration evolution within different families. (B) Evolution of sulfur pseudo-concentration as a function of DBE. (C) Evolution of sulfur pseudo-concentration for DBE 9-DBT as a function of number of carbon. The standard deviation error bars colored in red have been added for all figures.

Considering DBT family with DBE equal to 9 and following the evolution of sulfur-compounds as a function of the number of carbon atoms in Figure 3C, it allows to conclude on compounds that are most likely to be refractory. Among them, C2-DBT (dibenzothiophene core with 2 additional carbon atoms that is a number of carbons atoms equal to 14 and so on for the other designations), C3-DBT (C15), C4-DBT (C16) and poly-alkylated DBT (C5, C6, C7, C8-DBT that are C17, C18, C19 and C20) can be mentioned. Thus, poly-alkylated DBT are found to be more resistant to hydrotreatment, especially at medium degree of alkylation (C3-DBT). C2-DBT, that might correspond to a 4,6-dimethyl dibenzothiophene was expected to be very refractory[8] but data shows that more alkylated compounds are also very refractory to hydrotreatment.

Part 2: Chemometric exploitation of FT-ICR MS data

PCA

As discussed earlier, PCA has been used for exploratory purposes. In this case, pure samples (i.e SRGO, LCO, GOCK, FBGO, EBGO and HDT) are used to build the PCA model and mixed samples (i.e MIX) are projected along principal components in order to validate it. 6 PCs have been considered

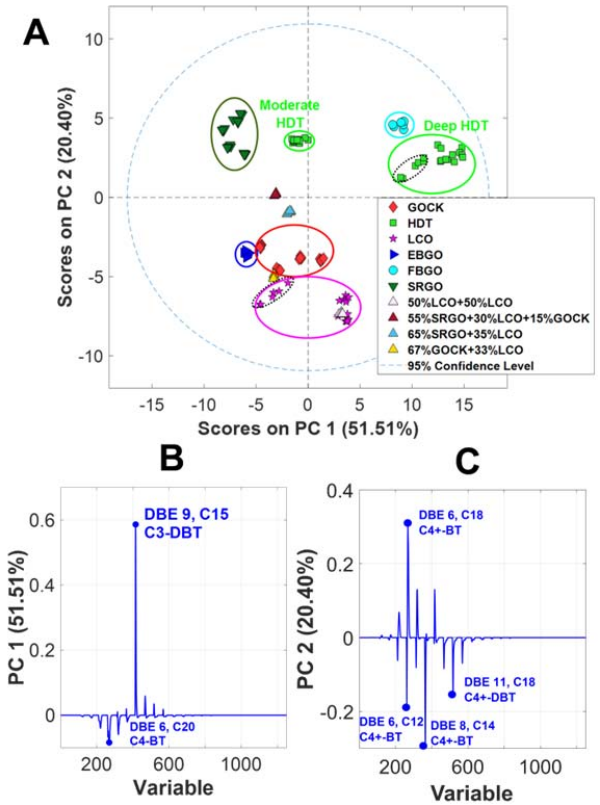


Figure 4: (A) Score plot of PC1 over PC2 obtained for APPI(+)-FT-ICR MS data. (B) Loadings from PC1. (C) Loadings from PC2.

as significant, explaining 96.7% of the total variance. The score plot obtained over PC1 and PC2 is shown in Figure 4A. PC1 explained 51.51% of the total variance when PC2 explained 20.40%. The repeatability of measurements can be evaluated from such scores plots as described elsewhere[23,27]. Close proximity of points is achieved for the six different replicates for most of samples. Moderate proximity is observed for sample 17 (deep HDT) and sample 8 (LCO), indicated in dash black circle. Indeed, sample 17 has the lowest content of sulfur (190 ppm, see Table 1) and APPI(+) is not very sensitive to sulfur compounds. As a consequence, repeatability decreases with decreasing sulfur content of gas oils. Moreover, a supplementary PCA analysis has been performed specifically on LCO samples to investigate this lack of repeatability and potentially identify variation source (see Figure S2 in Supplementary Information). The loadings analysis shows that significant concentrations changes are mainly observed along PC2 containing information about several compounds such as C14, C15 and C16-DBT molecules (Figure S2B). These concentrations decrease gradually over replicates time analysis, indicating that a memory effect might be encountered from previous sample analyzed (Sample 7 which contains very high amounts of C14, C15 and C16-DBT). This might has distorted actual concentrations of these compounds and has induced more variability but it remains extremely small.

From a general point of view, good results are obtained in terms of gas oils grouping with this basic method. As shown in PC1-PC2 score plot (Figure 4A), a clustered projection of samples is observed according to their process origins. Mixed blends used for validation are also well clustered and well projected with respect to mix proportions. Indeed, sample GO 15 (65% SRGO + 35% LCO) is projected between SRGO and LCO clusters, slightly closer to SRGO one. Sample GO 16 (67% GOCK + 33% LCO) is projected right in the middle of LCO and GOCK clusters. Finally, sample GO 23 which is a blend of two different LCO is logically located into the LCO area.

Along PC1, separation seems to be correlated to the amount of sulfur in the samples as SRGO contain much more sulfur than deep HDT and FBGO samples and moderate HDT is located in the middle of these two clusters (see Table 1). Among each cluster, the separation depending PC1 is also driven by the amount of sulfur contained by individual sample. For example, when considering GOCK family, GO 10 which has the lowest content in sulfur is located on the upper left, then comes GO 9, 11 and 12 whose contents in sulfur increase with respect to their projection over PC1. Regarding PC2, the samples projections could be correlated to the amount of elementary nitrogen (see Table 1). LCO and GOCK samples that contain the highest quantities of nitrogen are projected on the lower part while HDT and SRGO samples that contain few nitrogen compounds are projected on the upper part of the score plot. This could be due to the ionization of nitrogen compounds using APPI(+)-FT-ICR MS that could compete with sulfur compounds ionization. Enhanced nitrogen compounds ionization could decrease sulfur compounds ionization efficiency.

Samples with similar characteristics (i.e leading to similar products) are grouped together, such as EBGO, GOCK and LCO that are known to contain heavy compounds such as DBT or BT. FBGO replicates are logically grouped close to HDT samples as FBGO is a very efficient process to produce good quality gas oils (with few sulfur content).

Another interesting aspect observed in the score plot is a separation observed within HDT samples according to hydrotreatment severity level. The samples for which hydrotreatment severity is the highest are projected on the right side of the score plot, while moderate HDT samples are located right in the middle. Hence, PCA could be very useful to evaluate hydrotreatment severity for unknown samples.

The loadings plot corresponds to the visualization of variables distribution over one principal component. It can give clues on the variables that are more likely to explain the score plot obtained. The loadings plot obtained for PC1 is shown in Figure 4B. One major identified variable is C3-DBT, a dibenzothiophene core molecule containing 3 supplementary carbon atoms (variable number 415) corresponding to DBE 9 and Carbon 15). It could be stated that this compound has already been put forward by FT-ICR MS classical data analysis in Part 1 among other compounds due to its refractory character (Figure 3C) and the analysis of PCA model reveals that it could also explain classification between samples.

A closer look at this particular variable shows that projection of the samples along PC1 is actually well correlated to the amount of C3-DBT in samples. Deep HDT and FBGO samples have highest relative intensities while SRGO and EBGO samples have lowest relative intensities in C3-DBT (Figure S3 in Supplementary Information). To go further, this highlights the very refractory character of this compound as it exhibits highest relative intensities for deep HDT samples confirming results obtained in Part 1. In literature, 4,6-DBT is known to be very refractory to hydrotreatment[8,39]. The results obtained here are consistent with these observations since 3 supplementary carbon atoms could correspond to two additional substituents on DBT core structure on 4 and 6

positions[8,42]. However, information about substituents location on molecule cannot be obtained with FT-ICR MS as two equal molecular masses are not separated in the mass spectra. Ion mobility, that can separate isomers as a function of their drift time, could be very useful in this case to give clues on possible structure. It can also be stated that variable 270 that corresponds to a benzothiophene core with four supplementary carbon atoms (C4-BT) is anti-correlated with variable 415. Thus, it might be assumed that C4-BT conversion is easier to achieve than C3-DBT as there is less C4-BT in deep HDT samples than in native or moderate HDT samples.

Explained variance for PC2 is more equally split between several variables as shown in Figure 4C. Especially, poly-alkylated benzothiophenes (DBE 6, C12 and C18, variables number 262 and 268), poly-aromatic benzothiophenes (DBE 8, C14, variable 364) and poly-aromatic dibenzothiophenes (DBE 11, C18, variable 518) are among interest. Again, loadings obtained are consistent, as quite light samples such as SRGO, FBGO and HDT are expected to contain less aromatics and high number of carbons-containing molecules than heavier products (BT DBE 6 with C18 and DBT DBE 11 with C18) like LCO and GOCK. This is actually observed as light samples are negatively correlated with "heavy" aromatics.

About 30% of variance is still retained in other PCs. An interesting point from their analysis is that PC3 (12.10%), PC4 (7.24%) and PC5 (4.27%) variances are individually mainly explained by one or two special samples. The PC3-PC4 and PC5-PC6 score plots are available in Figure S4 along with their corresponding PC3, PC4, PC5 and PC6 loadings plots in Figure S5 Supplementary Information. For example, PC3 variance is mostly explained by FBGO and short-cut LCO samples and the corresponding loadings are related to “light” variables such as C4-BT or C2-DBT. FBGO sample is quite light and is expected to contain less aromatics and smaller alkylation levels than other GO which is in accordance with most expressed variables. Short-cut LCO (sample 8) is much lighter than other GO so its correlation to light variables is also in line with what could be expected. Regarding PC4, short-cut LCO sample and FBGO again explain most variance, and C4-BT and C2-DBT are also identified as most expressed variables. Finally, considering PC5, EBGO sample is revealed to explain most of 4.27% remaining variance. The variables identified in the loadings plots are C4+-BT and C4+-DBT that are negatively correlated to EBGO sample. EBGO sample is the heaviest considering all gas oils in the database (Table 1) and contains very aromatic and very alkylated compounds such as NBT or with DBE > 15. Moreover, EBGO sample has the lowest proportion of light compounds such as BT, as seen in Part 1. Then, it is logically negatively correlated to these quite “light” variables regarding the very heavy composition of EBGO. This is also demonstrated over the analysis of the Q-Residual plot (Figure S6 in Supporting Information) showing that EBGO is the only sample with contributions for very heavy variables (V>600, that is DBEs superior to 13). In comparison, Q-Residual contributions of all other samples are related to variables with DBE lower than 13.

In this study, PCA was revealed to be a good and exhaustive exploratory method. Particularly, some gas oils groups have been clearly identified over PC1-PC2 score plot depending on process used for their production. These groupings have been validated by using external samples (mixed blends) that were well projected according to their composition. As a consequence, this method could be applied to unknown sample in order to identify the process used that cannot be directly identified considering only FT-ICR MS complex data.

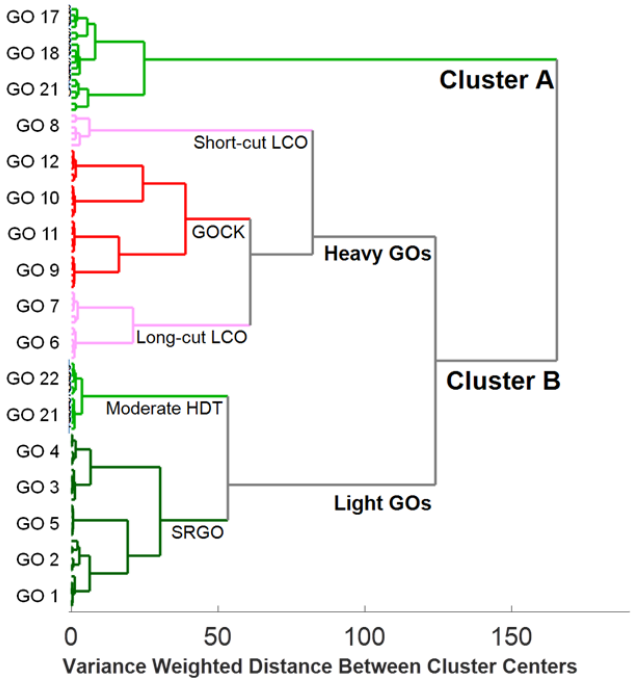


Figure 5: Dendrogram obtained with HCA Ward's method. Heavy gas oils cluster refers to heavy character of gas oil such as GOCK and LCO samples, while light gas oils cluster refers to quite light samples such as moderate HDT and SRGO samples

HCA

APPI(+)-FT-ICR MS data from 17 of 23 samples (MIX, FBGO and EBGO samples not considered) was assessed using Hierarchical Cluster Analysis (HCA) with Ward’s method distance measurement. As usual, the MS data transformation pipeline (Figure 1) has been applied prior this new chemometric exploration. In other words, the 102 pseudo-spectra have been directly used in HCA. Dendrogram shown in Figure 5 resumed the results of HCA clustering. Two major clusters observed with HCA are related to the sulfur content of samples. Deep HDT samples in cluster A (i.e samples GO 17, 18 and 21) are separated from all other samples in cluster B because of their very low sulfur content (Table 1). Considering cluster B, several sub-clusters are obtained depending on the process origin of the sample (LCO, GOCK, SRGO...). Samples with similar characteristics are grouped into several sub-clusters as seen in the dendrogram (Heavy gas oils: LCO with GOCK and light gas oils: moderate HDT with SRGO). It is also interesting to focus on the localization of sample 8 in the dendrogram. This sample has been obtained by LCO process but exhibits low contents in sulfur and nitrogen. Its boiling temperature range is much lower than other LCO samples (166-304 °C) due to a short distillation cut. Actually, according to boiling temperature range values, this sample could be considered as a kerosene sample. This separation from other LCO samples has demonstrated the HCA ability to perform an exhaustive analysis of all FT-ICR MS data by separating samples according to hydrotreatment level, as well as highlighting unique sample among classes. To sum up, HCA has confirmed PCA classification and has highlighted short-cut LCO that was already put forward in PCA.

CONCLUSION

Combining both high resolution mass spectrometry and chemometric methods could be one of the most interesting ways to extract unique variables among all data generated by FT-ICR MS. This methodology has been applied here to study differences or similarities among several processed gas oils samples focusing on aromatic sulfur compounds. First, the analysis of APPI(+)-FT-ICR MS data has put forward differences in composition, especially in terms of relative abundances of several heteroatomic sulfur families within gas oils obtained from different processes. These differences have been successfully correlated to the macroscopic properties of the samples such as sulfur and nitrogen contents or boiling temperature ranges. By applying chemometric methods such as PCA and HCA on a preprocessed MS data set, complementary levels of detail have been observed. Classification has been obtained within each origin process and validated by mixed blends samples. Distinction between moderate and deep hydrotreatment has been achieved even though they have very similar mass spectra. Closer look at loadings obtained from these methods has allowed us to identify key compounds such as C3-DBT that drove most of the separation within classes. The combination of design of experiment, advanced MS data processing and unique data rearrangement into a 1D matrix allowed a robust, repeatable and coherent gas oil dataset analysis. To go further, cutting-edge techniques such as ion mobility could then be very useful to get information about isomers of this compound, as well as providing structural characterization of these refractory compounds.

ASSOCIATED CONTENT

Supporting Information

Data processing workflow used, PC1-PC2 score plot from PCA considering only LCO samples, C3-DBT relative intensity in samples, PC3-PC4 and PC5-PC6 score plots, PC3, PC4, PC5 and PC6 loadings plot, Q-Residual contributions of different gas oils samples (PDF)

AUTHOR INFORMATION

Corresponding Author

*marion.lacoue-negre@ifpen.fr

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank warmly Yury Tsybin, Konstantin Nagornov and Anton Kozhinov (Spectroswiss) for providing software tools and help regarding data processing.

REFERENCES

(1) U.S Energy & Information Administration. Percentages of Total Imported Crude Oil by API Gravity. https://www.eia.gov/dnav/pet/pet_move_ipct_k_m.htm.

(2) Billon, A.; Morel, F.; Morrison, M. E.; Peries, J. P. Les procédés IFP HYVAHL(r) et SOLVAHL(r) de conversion de résidus. *Rev. Inst. Fr. Pét.* **1994**, *49*, 495–507.

(3) Colyar, J. J.; Kressmann, S.; Boyer, C.; Schweitzer, J. M.; Viguie, J. C. Improvements of Ebullated-Bed Technology for Upgrading Heavy Oils. *Oil & Gas Science and Technology - Rev. IFP* **2000**, *55*, 397–406.

(4) Tissot, B. P.; Welte, D. H. Composition of Crude Oils. In *Petroleum Formation and Occurrence*, 2., nd ed. 1984. Softcover reprint of the original 2nd ed. 1984; Tissot, B.P., Welte, D.H., Eds.; Springer Berlin: Berlin, **2013**; pp 375–414.

(5) Wauquier, J.-P. *Raffinage du pétrole, Pétrole brut, Produits pétroliers, Schémas de fabrication*; Editions Technip, **1994**.

(6) Stanislaus, A.; Marafi, A.; Rana, M. S. Recent advances in the science and technology of ultra low sulfur diesel (ULSD) production. *Catalysis Today* **2010**, *153*, 1–68.

(7) Nagai, M.; Masunaga, T.; Hanaoka, N. Hydrodenitrogenation of carbazole on a molybdenum/alumina catalyst. Effects of sulfiding and sulfur compounds. *Energy Fuels* **1988**, 645–651.

(8) Valencia, D.; García-Cruz, I.; Uc, V. H.; Ramírez-Verduzco, L. F.; Aburto, J. Refractory Character of 4,6-Dialkyldibenzothiophenes: Structural and Electronic Instabilities Reign Deep Hydrodesulfurization. *Chemistry Select* **2018**, *3*, 8849–8856.

(9) Hughey, C. A.; Rodgers, R. P.; Marshall, A. G.; Qian, K.; Robbins, W. K. Identification of acidic NSO compounds in crude oils of different geochemical origins by negative ion electrospray Fourier transform ion cyclotron resonance mass spectrometry. *Organic Geochemistry* **2002**, *33*, 743–759.

(10) Klein, G. C.; Rodgers, R. P.; Marshall, A. G. Identification of hydrotreatment-resistant heteroatomic species in a crude oil distillation cut by electrospray ionization FT-ICR mass spectrometry. *Fuel* **2006**, *85*, 2071–2080.

(11) Marshall, A. G. Fourier transform ion cyclotron resonance detection: principles and experimental configurations. *International Journal of Mass Spectrometry* **2002**, *215*, 59–75.

(12) Ijames, C.F; Wilkins, C.L. Fourier Transform Ion Cyclotron Mass Spectrometry using pseudo-random noise excitation. *Chemical Physics Letters* **1984**, *108,* 58–62.

(13) Cho, Y.; Ahmed, A.; Islam, A.; Kim, S. Developments in FT-ICR MS instrumentation, ionization techniques, and data interpretation methods for petroleomics. *Mass spectrometry reviews* **2015**, *34*, 248–263.

(14) Gutiérrez Sama, S.; Farenc, M.; Barrère-Mangote, C.; Lobinski, R.; Afonso, C.; Bouyssière, B.; Giusti, P. Molecular Fingerprints and Speciation of Crude Oils and Heavy Fractions Revealed by Molecular and Elemental Mass Spectrometry: Keystone between Petroleomics, Metallopetroleomics, and Petrointeractomics. *Energy Fuels* **2018**, *32*, 4593–4605.

(15) Marshall, A. G.; Rodgers, R. P. Petroleomics: The next grand challenge for chemical analysis. *Accounts of chemical research* **2004**, *37*, 53–59.

(16) Marshall, A. G.; Rodgers, R. P. Petroleomics: chemistry of the underworld. *Proceedings of the National Academy of Sciences of the United States of America* **2008**, *105*, 18090–18095.

(17) Muller, H.; Adam, F. M.; Panda, S. K.; Al-Jawad, H. H.; Al-Hajji, A. A. Evaluation of quantitative sulfur speciation in gas oils by Fourier transform ion cyclotron resonance mass spectrometry: Validation by comprehensive two-dimensional gas chromatography. *Journal of the American Society for Mass Spectrometry* **2012**, *23*, 806–815.

(18) Purcell, J. M.; Hendrickson, C. L.; Rodgers, R. P.; Marshall, A. G. Atmospheric pressure photoionization fourier transform ion cyclotron resonance mass spectrometry for complex mixture analysis. *Analytical chemistry* **2006**, *78*, 5906–5912.

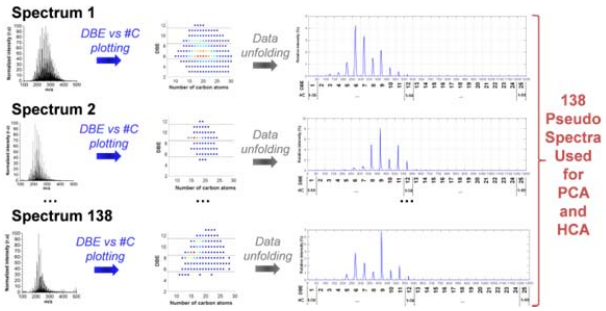
(19) Purcell, J. M.; Juyal, P.; Kim, D.-G.; Rodgers, R. P.; Hendrickson, C. L.; Marshall, A. G. Sulfur Speciation in Petroleum: Atmospheric Pressure Photoionization or Chemical Derivatization and Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Energy Fuels* **2007**, *21*, 2869–2874.

(20) Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **1933**, *24*, 417–441.

(21) Law, J.; Jolliffe, I. T. Principal Component Analysis. *The Statistician* **1987**, *36*, 37–52.

(22) Corilo, Y. E.; Podgorski, D. C.; McKenna, A. M.; Lemkau, K. L.; Reddy, C. M.; Marshall, A. G.; Rodgers, R. P. Oil spill source identification by principal component analysis of electrospray ionization Fourier transform ion cyclotron resonance mass spectra. *Analytical chemistry* **2013**, *85*, 9064–9069.

(23) Witt, M.; Timm, W. Determination of Simulated Crude Oil Mixtures from the North Sea Using Atmospheric Pressure Photoionization Coupled to Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Energy Fuels* **2015**, *30*, 3707–3713.
(24) Hur, M.; Yeo, I.; Park, E.; Kim, Y. H.; Yoo, J.; Kim, E.; No, M.-h.; Koh, J.; Kim, S. Combination of statistical methods and Fourier transform ion cyclotron resonance mass spectrometry for more comprehensive, molecular-level interpretations of petroleum samples. *Analytical chemistry* **2010**, *82*, 211–218.
(25) Mazur, D. M.; Harir, M.; Schmitt-Kopplin, P.; Polyakova, O. V.; Lebedev, A. T. High field FT-ICR mass spectrometry for molecular characterization of snow board from Moscow regions. *The Science of the total environment* **2016**, *557-558*, 12–19.
(26) Yeo, I.-J.; Lee, J.-W.; Kim, S.-H. Application of Clustering Methods for Interpretation of Petroleum Spectra from Negative-Mode ESI FT-ICR MS. *Bulletin of the Korean Chemical Society* **2010**, *31*, 3151–3155.
(27) Guillemant, J.; Albrieux, F.; Pereira de Oliveira, L. C.; Lacoue-Nègre, M.; Duponchel, L.; Joly, J. F. Insights from nitrogen compounds in gas oils highlighted by high-resolution Fourier transform mass spectrometry. Under review. *Anal. Chem.,* **2019**.
(28) Herrera, L. C.; Grossert, J. S.; Ramaley, L. Quantitative aspects of and ionization mechanisms in positive-ion atmospheric pressure chemical ionization mass spectrometry. *Journal of the American Society for Mass Spectrometry* **2008**, *19*, 1926–1941.
(29) Huba, A. K.; Huba, K.; Gardinali, P. R. Understanding the atmospheric pressure ionization of petroleum components: The effects of size, structure, and presence of heteroatoms. *The Science of the total environment* **2016**, *568*, 1018–1025.
(30) Kilgour, D. P. A.; Nagornov, K. O.; Kozhinov, A. N.; Zhurov, K. O.; Tsybin, Y. O. Producing absorption mode Fourier transform ion cyclotron resonance mass spectra with non-quadratic phase correction functions. *Rapid communications in mass spectrometry* **2015**, *29*, 1087–1093.
(31) Zhurov, K. O.; Kozhinov, A. N.; Fornelli, L.; Tsybin, Y. O. Distinguishing analyte from noise components in mass spectra of complex samples: Where to cut the noise? *Analytical chemistry* **2014**, *86*, 3308–3316.
(32) Islam, A.; Cho, Y.-J.; Ahmed, A.; Kim, S.-H. Data Interpretation Methods for Petroleomics. *Mass Spectrometry Letters* **2012**, *3*, 63–67.
(33) Kozhinov, A. N.; Zhurov, K. O.; Tsybin, Y. O. Iterative method for mass spectra recalibration via empirical estimation of the mass calibration function for Fourier transform mass spectrometry-based petroleomics. *Analytical chemistry* **2013**, *85*, 6437–6445.
(34) Purcell, J. M.; Rodgers, R. P.; Hendrickson, C. L.; Marshall, A. G. Speciation of nitrogen containing aromatics by atmospheric pressure photoionization or electrospray ionization fourier transform ion cyclotron resonance mass spectrometry. *Journal of the American Society for Mass Spectrometry* **2007**, *18*, 1265–1273.
(35) Morales-Valencia, E. M.; Castillo-Araiza, C. O.; Giraldo, S. A.; Baldovino-Medrano, V. G. Kinetic Assessment of the Simultaneous Hydrodesulfurization of Dibenzothiophene and the Hydrogenation of Diverse Polyaromatic Structures. *ACS Catalysis.* **2018**, *8*, 3926–3942.
(36) Zeuthen, P.; Knudsen, K. G.; Whitehurst, D. D. Organic nitrogen compounds in gas oil blends, their hydrotreated products and the importance to hydrotreatment. *Catalysis Today* **2001**, *65*, 307–314.
(37) van Looij, F.; van der Laan, P.; Stork, W.H.J.; DiCamillo, D.J.; Swain, J. Key parameters in deep hydrodesulfurization of diesel fuel. *Applied Catalysis A: General* **1998**, *170*, 1–12.
(38) Eide, I.; Zahlsen, K. A Novel Method for Chemical Fingerprinting of Oil and Petroleum Products Based on Electrospray Mass Spectrometry and Chemometrics. *Energy Fuels* **2005**, *19*, 964–967.
(39) Macaud, M.; Milenkovic, A.; Schulz, E.; Lemaire, M.; Vrinat, M. Hydrodesulfurization of Alkyldibenzothiophenes: Evidence of Highly Unreactive Aromatic Sulfur Compounds. *Journal of Catalysis* **2000**, *193*, 255–263.


Spectrum 1
DBE vs #C plotting
Data unfolding
Spectrum 2
DBE vs #C plotting
Data unfolding
Spectrum 138
DBE vs #C plotting
Data unfolding
138 Pseudo Spectra Used for PCA and HCA