HHS Public Access

Author manuscript

Environ Sci Technol. Author manuscript; available in PMC 2018 June 20.

Published in final edited form as:

Environ Sci Technol. 2017 June 20; 51(12): 7197–7207. doi:10.1021/acs.est.6b06413.

Grouping of Petroleum Substances as Example UVCBs by Ion Mobility-Mass Spectrometry to Enable Chemical Composition-Based Read-Across

Fabian A. Grimm[1], **William K. Russell**[2], **Yu-Syuan Luo**[1], **Yasuhiro Iwata**[1], **Weihsueh A. Chiu**[1], **Tim Roy**[3], **Peter J. Boogaard**[4], **Hans B. Ketelslegers**[5], and **Ivan Rusyn**[1,*]

[1]Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX [2]Laboratory for Biological Mass Spectrometry, Department of Chemistry, Texas A&M University, College Station, TX [3]Department of Natural Science, University of South Carolina, Beaufort, SC [4]Shell International BV, The Hague, The Netherlands [5]European Petroleum Refiners Association, Concawe Division, Brussels, Belgium

Abstract

Substances of Unknown or Variable composition, Complex reaction products, and Biological materials (UVCBs), including many refined petroleum products, present a major challenge in regulatory submissions under the EU Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) and US High Production Volume regulatory regimes. The inherent complexity of these substances, as well as variability in composition obfuscates detailed chemical characterization of each individual substance and their grouping for human and environmental health evaluation through read-across. In this study, we applied ion mobility mass spectrometry in conjunction with cheminformatics-based data integration and visualization to derive substance-specific signatures based on the distribution and abundance of various heteroatom classes. We used petroleum substances from four petroleum substance manufacturing streams and evaluated their chemical composition similarity based on high-dimensional substance-specific quantitative parameters including *m/z* distribution, drift time, carbon number range, and associated double bond equivalents and hydrogen-to-carbon ratios. Data integration and visualization revealed group-specific similarities for petroleum substances. Observed differences within a product group were indicative of batch- or manufacturer-dependent variation. We demonstrate how high-resolution analytical chemistry approaches can be used effectively to support categorization of UVCBs based on their heteroatom composition and how such data can be used in regulatory decision-making.

*Corresponding author: Ivan Rusyn, MD, PhD, Department of Veterinary Integrative Biosciences, 4458 TAMU, Texas A&M University, College Station, TX 77843, (979) 458-9866, irusyn@tamu.edu.

Supporting information

Supporting figures showing petroleum product group-specific similarities in the top 10 most frequent heteroatom classes based on absolute counts and relative abundances; replicability of IM-MS analysis; chemical fingerprinting, cluster analysis, and principal components analysis of petroleum substances based on absolute counts of their heteroatom class distribution.

Other authors declare no relevant conflicts of interest.

Introduction

Health and environmental assessments of the UVCB (Unknown or Variable composition, Complex reaction products or Biological materials) substances are some of the most challenging areas in regulatory toxicology.[1] UVCBs are identified on global chemical inventories with unique identifiers (EC or CAS numbers) and names; however, these are not always adequately specific to permit unambiguous identification.[2] Detailed chemical characterization of UVCBs is a challenge because of the complexity of their composition, as well as variability in composition between batches or manufacturers.[3, 4] Petroleum substances are one example of UVCBs; they are both very chemically complex and have variable composition which may be due to manufacturing processes, intended uses, or sources of crude oil. The range of petroleum substance types is wide, they are most commonly used as transportation fuels, in energy/heat-raising applications, and in manufacturing of other chemicals and products.

Due to the very high production volumes, almost all petroleum substances were registered in the 2010 REACH deadline; they represent a quarter of the top 20 most frequently registered substances in the European Union.[5] Most of the REACH registration submissions for petroleum substances incorporated some form of read-across argument to address existing data gaps.[6] Read-across requires justification to establish “sufficient similarity” between data-poor and data-rich substances, which traditionally rests on the similarities in chemical structure, metabolites, or biological effects. Substance identification is a critical initial step in evaluating similarity. Because of chemical complexity and variability, identification of petroleum substances is based on their manufacturing process and performance characteristics.[4] Specifically, ECHA guidance states that characteristics like “the name, carbon-chain length range, boiling point, viscosity, cut-off values and other physical properties are generally more helpful than compositional information.” At the same time, analytical data (UV/Vis, infra-red, nuclear magnetic resonance or mass spectrometry) is also considered useful for substance identification in REACH, especially if it allows for establishing a “chromatographic fingerprint” of a complex substance.[4]

Mass spectrometry-enabled analyses of complex petroleum substances[7–11] are especially attractive as they aim to provide comprehensive characterization of their organic composition and yield data that may be used to address not only the complexity, but also inherent natural and process-related variability in end-products. Recent developments in high-resolution mass spectrometry[8] and novel separation techniques[7, 12] enable ever greater ability to confront the identification of petroleum substances. For example, comprehensive two-dimensional chromatography (GC×GC) which exploits orthogonal separation[13] has been coupled with time-of-flight mass spectrometry (TOF MS) to provide an effective basis for compound identification.[14] GC×GC-based mass spectrometry methods yield substance composition information consisting of a wide range of carbon number and many distinct chemical structural classes (e.g., linear and branched alkanes, cycloalkanes, parent and alkyl mono-, di-, and poly-aromatics). High-resolution mass spectrometry enables identification of non-carbon atoms (referred to as heteroatom classes, or molecules that also contain N, O, and S), double bond equivalents (DBE, the number of rings plus double bonds involving carbon), and carbon number.[10] Fourier transform ion cyclotron resonance mass spectrometry

(FT-ICR MS), TOF MS and Orbitrap MS have been also coupled to different high-resolution separation techniques for more precise structural and compositional characterization of the compound classes.[7, 8, 15, 16]

While much progress has been made in developing novel analytical methods for the analysis of complex petroleum-derived UVCBs, less attention is dedicated to the use of these data in regulatory decision-making. Cheminformatics-based techniques that aim to define "blocks" of compounds in each petroleum substance[17–19] have taken advantage of the advances in analytical chemistry-based techniques. For example, chromatographically-defined aromatic ring class (ARC) profiles and other physico-chemical properties of petroleum substances have been used for modeling health and environmental hazards and risks[19–21], but they have not been widely used in read-across. Regulatory opinions on read-across submissions of petroleum substances frequently cite considerable uncertainties on these approaches. These uncertainties may be alleviated, at least in part, with novel data integration techniques analyzing multiple sources of relevant high content data, as presented previously in the context of petroleum UVCBs[22], including data from advanced analytical methods. The current paper focuses on the latter, in which we aimed to use ion mobility TOF MS to characterize heteroatom profiles of a number of petroleum substances from several distinct product categories and evaluate the use of these data to support grouping and read-across of these substances.

Materials and Methods

The overall study design is summarized in Figure 1.

Materials

LC/MS grade methanol Optima™ (CAS 67-56-1), toluene Optima™ (CAS 108-88-3) and formic acid (CAS 64-18-6) were purchased from ThermoFisher (Grand Island, NY). Naphthalene (CAS 91-20-3, 99% purity), phenanthrene (CAS 85-01-8, 98% purity), 1,2-benzanthracene (CAS 56-55-3, 99% purity), benzo[ghi]perylene (CAS 191-24-2, 98% purity), benzo[a]pyrene (CAS 50-32-8, 98% purity), and coronene (CAS 191-07-1, 98% purity) were purchased from Sigma Aldrich (St. Louis, MO). Samples of petroleum substances from four different manufacturing streams, straight run gas oils (SRGO, n=5), other gas oils (OGO, n=2), vacuum & hydrotreated gas oils (VHGO, n=8), and heavy fuel oils (HFO, n=3) were graciously provided by Concawe (Brussels, Belgium). SRGO are produced by atmospheric distillation of crude oil and typically contains hydrocarbons in the range from C_9 to C_{25} and has a boiling point range from ~150 to 400 ˚C. VHGO are produced by distillation under reduced pressure (light vacuum) and subsequently treated with hydrogen in the presence of a catalyst to remove hetero compounds and reduce aromatic compounds to cycloalkanes. VHGO typically comprise of hydrocarbons in the range from C_{13} to C_{30} and has a boiling point range of ~230 to 450˚C. OGO are produced from crude oil distillates by a series of subsequent treatments to remove hetero and aromatic compounds and typically have hydrocarbon ranges from C_{10} to C_{27} and a boiling point range from ~150 to 400˚C. HFO are residual refinery distillates from thermal and cracking

units with typical hydrocarbon ranges from C_{20} to C_{50} and boiling point ranges from ~350 to 650˚C.

Ion mobility-mass spectrometry

In accordance with published standard procedures,[8, 23] all petroleum substances were diluted prior to IM-MS analysis to a final concentration of 1 mg/ml using 1:1 (v/v) methanol:toluene containing 0.5% (v/v) formic acid to increase protonation efficiency and [M+H]+ ion generation from basic compounds for IM-MS analysis in positive ion mode. Toluene as a solvent is typical for the analysis of petroleum Cho Y, Ahmed A, Islam A, Kim S. Developments in FT-ICR MS instrumentation, ionization techniques, and data interpretation methods for petroleomics. Mass Spectrom Rev. 2015. 34:248–63; Hsu CS, Hendrickson CL, Rodgers RP, McKenna AM, Marshall AG. Petroleomics: advanced molecular probe for petroleum heavy ends. J Mass Spectrom. 2011. 46:337–43 No partitioning of the solvent or precipitation of sample upon dilution with the MeOH: toluene mixture was observed. Diluted samples were directly infused into a Waters Synapt G2-S Q-TOF MS (Waters Corporation, Milford, MA).

The instrument assembly contained an electrospray ionization source, equipped with a fused capillary electrospray tip prepared in house; voltage of the tip was set to 2.5kV. The sampling cone was set at 40V, the extraction cone was set at 4V, the source temperature was maintained at 120°C, the cone gas flow was kept at 40 l/h. The Nano Flow Gas pressure was set at 2 Bar, and a purge gas at a flow of 75 l/h was used for all samples. Following initial ionization, molecules were separated according to their volume/charge ratio by IM and the drift times were recorded. The IM wave velocity was set to 800 m/s and the wave height set to 40V. Mass analysis was achieved using a Time-of-Flight high-resolution mass analyzer. All samples were acquired for 2 minutes and each measurement was repeated two times. Data were acquired using MassLynx software (Waters, Milford, MA).

Reproducibility Assessment

The reproducibility of IM-MS data was assessed using correlation analysis of identifiable features in duplicate sample runs. IM-MS data were acquired as specified above and spectra were integrated for feature identification using DriftScope 2.0 software (Waters). Peak detection parameters were as follows: minimum drift time peak width: 1.0 bins; drift time range: auto; retention time range: auto; MS resolution: 5000; minimum intensity threshold: 500. Unique features in duplicate runs were then matched based on matching high-resolution m/z values and associated drift times using an assumed precision up to two-decimals. Acquired abundances for duplicate determinations of unique features were analyzed using Pearson and Spearman correlation analysis in Prism 5 (Graphpad, San Diego, CA).

Data processing

Waters Synapt IM-MS outputs (.raw format) were imported into PetroOrg software (http://software.petroorg.com; Omics, Tallahassee, FL) for quantitative feature identification. Peaks were identified across an *m/z* range between 0 and 2500 and were defined by a minimum of 11 counts/peak using a variable abundance threshold in positive ion mode. Raw data were then recalibrated according to the developer's recommendations using the most abundant

heteroatom class, the N1 class (close shell), as a reference. Assignment settings for individual feature identification included fixed carbon (0–100) and hydrogen (4–200) atom ranges, as well as heteroatom number limitations for nitrogen (0–2), oxygen (0–2), and sulfur (0–2) containing molecules. The maximum error and noise threshold were kept at their respective default values, 5.0 ppm and 0. Calibration-specific settings included a maximum error of 50 ppm and an abundance threshold of 0.05. ESI was selected as the ionization type.

Quantitative descriptors derived for each petroleum substance were exported as Microsoft Excel files and included a list of identifiable features based on their unique combination of recalibrated m/z ratio, drift time, and relative abundance. Derived descriptors for each molecule included the heteroatom class assignment, theoretical high-resolution mass, elemental composition, carbon chain length, double bond equivalents (DBE), and hydrogen/carbon (H/C) and nitrogen/carbon (N/C) ratios.

Aromatic Ring Class (ARC) analysis

Weight percentages of the polycyclic aromatic compounds in all petroleum substances were determined as detailed previously.[24, 25] Briefly, 4 g of each substance was dissolved in 10 mL of cyclohexane and subsequently extracted with 10 mL DMSO twice. Extracts were combined and diluted with two volumes of 4% (w/v) sodium chloride prior to a two-step extraction with 20 and 10 mL of cyclohexane. The organic fractions were combined and washed twice with 5 ml water. Following filtration using anhydrous sodium sulfate the cyclohexane was evaporated to near dryness at 40°C. Remaining solvent was then evaporated for 30 min at 80°C. The amount of each extract was then determined using the weight difference of the empty flask and following solvent evaporation. The extract was subsequently dissolved in cyclohexane yielding a final concentration of 50 mg/ml and stored at 4°C. ARC content analysis was then performed by gas chromatography-coupled mass selective detection (GC/MSD). Sample separation was achieved on a (30m; 0.25 mm; 0.25 mm) Zebron-5HT capillary column (Phenomenex, Torrance, CA) using a starting temperature of 70°C (30 sec), followed by a temperature gradient of 5°C/min until the final temperature of 300°C was reached and held for 35 minutes. Likewise, detector and injection port temperatures were kept at 300°C. Helium was used as the carrier gas at a rate of 30 cm/sec. Quantitative integration of the GC/MSD chromatograms was achieved using standards of naphthalene, phenanthrene, 1,2-benzanthracene, benzo[a]pyrene, benzo[ghi]perylene, and coronene. The resulting ARC profiles consisted of weight percentages by ring number.

In silico data processing for chemical data-integrative grouping of petroleum substances

In order to group petroleum substances according to global similarities in their chemical composition, we utilized a variety of computational approaches for data processing and visualization. For qualitative comparison of the chemical complexity of individual substances, we generated intensity plots showing either carbon chain length vs DBE or H/C ratio (van Krevelen plots) for the most abundant heteroatom class (N_1), directly in PetroOrg.

For quantitative data-integrative groupings, we initially extracted data matrices containing information on the sample-specific heteroatom class distribution, based on absolute counts

of individual features or their relative abundance. Sample-associated ARC content data were processed in the same manner. Using these data sets, we conducted an unbiased cluster analysis using the gplots software package in R studio (version 3.1.1). Cluster analysis provided a dendrogram summarizing the compositional correlation between petroleum substances, as well as heatmap visualization of the absolute or relative abundance of heteroatom classes and ARC content contained in specific samples. Compositional similarities among petroleum substances of the various manufacturing streams was also evaluated by principal components analysis using the scatterplot3D and plotrix software packages in R studio.

Results

Global compositional analysis of petroleum substances using IM-MS

IM-MS heatmaps, *i.e.* plots based on the chemical constituent's *m/z* (mass divided by charge number, x-axis), drift time (time for each ion to traverse within a homogeneous electric field in the ion mobility spectrometer, y-axis), and abundance (color) provide a visual fingerprint of the sample-specific feature distribution[26] that can serve as a qualitative indicator of the overall chemical complexity of petroleum substances (Figure 2). IM-MS heatmaps were obtained for petroleum substances from four manufacturing streams, other gas oils (OGO), straight run gas oils (SRGO), vacuum & hydrotreated gas oils (VHGO), and heavy fuel oils (HFO). These indicate strong qualitative and quantitative compositional similarities within a substance class, and differences between classes. OGOs represent the least complex, in terms of heteroatom content, example of the four manufacturing streams examined herein. This is evident from the overall lower feature abundance, as well as the absence of feature regions in the IM-MS heatmaps as compared to other petroleum substances. SRGOs and VHGOs featured a high degree of overall compositional resemblance, covering both similar *m/z* and drift time ranges and feature abundance. HFOs were the most complex substances tested in this study, exhibiting significantly higher feature abundance than any other petroleum substance class, as well as significantly broader feature distribution. This complexity was qualitatively and quantitatively well-preserved across all three tested HFO substances.

To generate chemical signatures and conduct global compositional fingerprinting of petroleum substances, IM-MS data were processed for feature identification using PetroOrg software (Figure 3A). Quantitative outputs of these data included unique feature identifiers (e.g., *m/z*, drift time, predicted molecular weight, etc.) in addition to variable characteristics including feature abundance, carbon chain length, H/C ratio, and DBEs. A cumulative summary of the distribution breakdown, *i.e.* absolute and relative abundances of all sample-specific heteroatom classes, was also obtained. Within assigned ranges for carbon (0–100), hydrogen (4–200), nitrogen (0–2), oxygen (0–2), and sulfur (0–2) atoms, the number of identifiable peaks varied widely between the three gas oil classes, with OGOs offering the lowest number of identifiable peaks (213–257), SRGO (269–1774) and VHGO (244–1724) covering similar ranges, and HFOs (4163–7580) exhibiting the highest complexity (Figure 3B). Identifiable features across all petroleum extracts constituted a total of 82 unique heteroatom classes (Supplemental Table 1). As expected, there were identifiable trend lines

in IM-MS data of the substances analyzed herein (e.g. Paglia and Astarita, Nature Protocols 12, 797–813, 2017). However, based on the spatial distribution of heteroatom classes across the IM-MS spectra, m/z and drift time of the identifiable features were closely correlated for the identifiable heteroatom classes and there does not appear to be unique separation of the trend lines. We found that, for the same substance shown in Figure 3A (see image below) and the region used for quantitation (black dots in Figure 3A),

The top 10 most abundant heteroatom classes across all samples, based either on absolute counts or relative abundance, *i.e.* following sample-specific normalization of peak counts, are shown in Figures 3C and 3D. There was good overall concordance between both approaches. The N_1 heteroatom class was the predominant one constituting approximately 34% of the bulk of identifiable features and being detectable in all petroleum substance extracts. Likewise, the N_1S_1 and N_2S_1 heteroatom classes were quantifiable in all tested samples, contributing 13% and 6% based on relative abundance of all identifiable features, respectively.

Comparison of top 10 heteroatom class distribution based on the absolute number of the features revealed strong quantitative similarities among individual samples within each petroleum substance manufacturing category (Supplemental Material, Figure S1). OGOs appeared as the least complex samples with less than 25 counts per group. A distinguishing feature for OGOs was the absence of features belonging to the $N_1{}^{13}C_1$ heteroatom class. SRGOs and VHGOs exhibited higher complexity as compared to OGOs, with the N_1 and N_1S_1 representing the most abundant heteroatom classes, while comparably low levels of N_2 and N_1O_2 containing compounds were also common characteristics. HFOs exhibited on average more than ten-fold higher counts per heteroatom class compared to gas oils. The N_1 class was again predominant with between 800 and 1000 counts per sample. The O_1, N_1O_2 and N_2O_1 classes were the least abundant ones across all HFOs, but still provided a large number of counts (>100) per class.

While a count-based comparison was primarily reflective of quantitative similarities and differences among the samples, to elucidate qualitative characteristics of the substances we also compared the distribution of top 10 heteroatom classes based on their relative abundance (Supplemental Material, Figure S2). The relative abundance does not take overall quantitative differences into account, as it is a result of a normalization of the counts for the individual heteroatom features to the total identifiable ion content in each sample. However, the relative feature distribution is pronounced, allowing improved qualitative comparison between global chemical sample compositions. In fact, OGOs were quite distinct with respect to overabundance of the O_1 heteroatom class, whereas samples of the other two gas oil categories contained comparably lower amounts of the O_1 heteroatom class molecules. By contrast, N_1 and N_1S_1 classes were most abundant in SRGOs and VHGOs with other groups being present at significantly lower amounts. HFOs, despite also featuring the N_1 class as their most abundant chemical contributor, displayed a higher abundance of the less frequent heteroatom classes, i.e. N_2O_1, N_1S_2, as compared to the gas oils.

Reproducibility Assessment of the IM-MS data

Qualitative analysis of replicate IM-MS spectra for representative OGO (CON-07), SRGO (CON-02), VHGO (CON-12), and HFO (A083/13) samples indicates virtually indistinguishable drift time and *m/z* patterns, including feature abundance distributions as compared to original sample runs (Supplemental Material, Figure S3).

In order to provide a quantitative basis for estimating sample replicability, unique features were determined using standardized peak identification settings in DriftScope 2.0 software. Identification was based on matching high-resolution *m/z* values and associated drift times and revealed a wide range of compositional complexity of the substances among in four manufacturing streams tested. While *m/z* values and drift times are invariable identifiers of the individual constituents in complex substance matrices, their relative abundance is a variable characteristic that provides a meaningful descriptor for quantitative replicability analysis. Correlation analysis based on replicate abundances for unique features revealed strong linear correlation for all four petroleum samples: Pearson's r correlation coefficients were 0.99 for duplicates of CON-02, CON-05, and CON-12 substances and 0.82 for a more complex A083/13 substance. Similar results were found using Spearman's ρ values (0.96 for CON-02, CON-05, CON-12 and 0.72 for A083/13). All correlations were statistically significant ($p<0.0001$).

Semi-quantitative compositional similarity assessment

To evaluate the feasibility of using the most frequent and abundant heteroatom class, the N_1 class, in semi-quantitative compositional similarity analysis, we prepared heatmap abundance visualization plots for chemical property correlations, i.e. carbon chain length distribution versus DBE (Figure 4A) and carbon chain length distribution versus H/C ratio (Figure 4B). These plots show that OGOs are the least complex substances tested, with a narrow carbon chain length distribution within this heteroatom class (approx. 15–23). SRGOs and VHGOs exhibited a wider range of carbon-containing molecules (approx. 11–40), as well as wider ranges for both DBEs (approx. 4–17) and H/C ratio (approx. 0.6–2) as compared to OGOs (approx. 6–10 and 1–1.7). By contrast, HFOs contain molecules with 10 to 80 carbons, while having a DBE range of approx. 4–60 and H/C ratios ranging from 0.3 to higher than 2. Carbon number ranges for heteroatom classes of petroleum substances are thus concordant with their expected distillation profiles.[27] The density plots indicated overall compositional similarities within each petroleum substance category, but they were also reflective of the fact that samples within a given category exhibit compositional variation, albeit to a much lower extent than differences among the samples of different manufacturing class.

ARC content analysis

Characterization of the wt.% concentration of each of seven ring classes of aromatic compounds in complex petroleum substances is a common strategy for estimating their hazard properties.[20, 28, 29] ARC profiles[28] of the substances tested in our study comprised of sample-specific weight percentages for aromatic compounds containing 1 through 7 rings.[30] Total weight percentages of aromatic compounds were obtained across all petroleum substances as detailed in Methods, they ranged between 2.1 and 9.9% (4.7–6.5% for

SRGOs, 2.3–6.6% for OGOs, 2.1–9.9% for VHGOs, and 2.3–5.4% for HFOs). It should be noted that HFOs can vary substantially in their composition and 2.3–5.4% aromatics content is not unusual for HFOs, even though it can be significantly greater depending on the source of the sample and performance specifications for a refinery stream. The relative distribution among ARC revealed comparably low concentrations of single aromatic ring containing molecules (0–9%), whereas compounds containing two (9–83%) or three (14–80%) aromatic rings comprised the most abundant groups based on the total ARC content. 4-ring containing compounds were less prevalent in SRGOs (0–13%), OGOs (0%), and VHGOs (0–19%), but still constituted 22% and 36% of the ARC content in two of HFOs. Molecules containing five or more aromatic rings were only sporadically observable in gas oils, whereas they were still prevalent (0–21%) in HFOs.

Chemical data-integrative fingerprinting of petroleum substances

Integration and visualization of IM-MS data sets for chemical fingerprinting and groupings of petroleum substances was achieved using (i) heatmap-based cluster analysis (Figure 5 and Supplemental Material, Figure S4) and (ii) principal components analysis (Figure 6 and Supplemental Material, Figure S5) using absolute count- and relative abundance-based heteroatom class distributions, as well as aromatic ring class mass percentages as underlying data matrices.

A cluster analysis based on the relative abundance of all 82 identified heteroatom classes provided a visualization of the complexity of the heteroatom distribution for each sample and allowed the identification of unique compositional patterns (Figure 5, Supplemental Table 1). As described previously for the abundance of the top 10 most frequent heteroatom classes, OGOs were characterized by the highly abundant O_1 class, but also due to the consistent presence of the $O_1{}^{13}C_1$, O_1S_1, O_2S_1, $N_2O_1S_1$, and $N_2O_1S_2$, as well as the absence of the $N_1{}^{13}C_1$ heteroatom classes. SRGOs did not exhibit uniquely identifying features and largely resembled the heteroatom class distribution obtained from VHGOs. The most abundant heteroatom classes in SRGOs and VHGOs, N_1, O_1, $N_1{}^{13}C_1$, N_1S_1, and N_2S_1, were also detectable at similar amounts in HFOs. The higher complexity of HFOs, which represented the least refined product class in this study, was reflected by an overall higher number of detectable heteroatom classes. The N_1S_2 class was among the unique identifying features of HFOs, albeit there were a larger number of low abundance classes that were only detectable in HFOs. The associated dendrogram showed clusters consisting of HFOs (A092/13, A087/13, A083/13) and OGOs (CON-07 and CON-09), with the HFOs separating from gas oil samples. SRGOs and VHGOs were not well separated. Principal components analysis also demonstrated sample and manufacturing stream-specific compound groupings with clear separation between HFOs and the gas oils. OGOs separated from SRGOs and VHGOs with three principal components, whereas the latter two groups were highly overlapping.

When absolute counts, rather than relative abundances, were used as the data input for chemical fingerprinting, a more pronounced separation between HFOs and gas oils was evident (Supplemental Material, Figure S4). Still, the overall trends observable in the cluster analysis remained identical between two approaches, *i.e.* HFOs and OGOs each cluster

together with HFOs clearly separating from all gas oils, whereas SRGOs and VHGOs remain indistinguishable from each other. Principal components analysis based on absolute counts shows strong overlap and a narrow distribution of OGOs, SRGO, and VHGOs, with HFOs clearly distinct but broadly dispersed (Supplemental Material, Figure S5). Conversely, the less complex ARC profile data matrix did not provide uniquely identifying, group-specific patterns (Figure 5A) or result in distinct separation of HFOs from gas oils as reflected in visualization using clustering (Figure 5B), a technique which groups samples based on their overall correlation, and principal components analysis (Figure 6), a dimensionality reduction technique that visualizes the relative position of samples in high-dimensional space.

Discussion

Most chemicals in commerce do not meet all of the REACH data requirements with regard to mammalian toxicity and data gaps are most often addressed through substance similarity-based read-across. Grouping and read-across are encouraged in regulatory submissions for both mono-constituents and UVCBs.[31–33] However, the inherent compositional complexity and variability of many UVCBs creates unique challenges for their grouping and read-across. Regulatory guidelines allow for inclusion of multidimensional data to support read-across submissions; these can include data on biological activity (toxicological and mechanistic) and bioavailability (including metabolite profiling).[31, 32] We recently demonstrated that comprehensive bioactivity assessment using organotypic *in vitro* models and high-content screening technologies provides an effective means towards grouping and read-across of petroleum substances in a way reflective of their manufacturing process.[22, 34]

Still, while biological data are one way to define groupings of substances, demonstration of chemical similarity within a group remains a strong requirement, it is an even greater challenge for UVCB substances.[32] Conventional analytical techniques, primarily GC-MS have been used to provide high-dimensional substance composition information, primarily of the hydrocarbon constituents of petroleum substances, which typically form well over 90% of their total mass.[7, 35] Advancements in IM-MS-based technologies have improved sample separation efficiency by incorporating shape and size-based dimensionality, and are amenable for high (TOF-MS) and ultra-high (FT-MS) mass analyses of petroleum substances.[16, 36] Collectively, these approaches are defined under the term petroleomics; they are useful for detailed chemical characterization to meet regulatory requirements towards grouping of petroleum substances for read across.[7, 10, 15, 37]

ESI-coupled IM-MS is particularly well-suited for the analysis of polar molecules and thus is a promising technique for compositional fingerprinting of petroleum substances based on their heteroatom constituents, *i.e.* content and distribution of oxygen-, nitrogen-, and sulfur-containing molecules.[10, 37] While ESI-coupled IM-MS has been shown to be useable for characterization of very complex substances such as crude oil and refined products, this study demonstrates its utility also for grouping of petroleum substances. Specifically, we show how chemical signatures characteristic for petroleum manufacturing classes can be used for data-integrative groupings of petroleum substances based on their hetero-atom content. Even though heteroatom-containing molecules comprise only a relatively minor

fraction (<10% based on mass percentage[10]) of crude oil and refined petroleum products, heteroatom-based fingerprints can be used to support groupings of petroleum substances in support of regulatory decision-making by read-across.

We demonstrate global chemical similarity in heteroatom class distribution and abundance of petroleum substances belonging to one of four manufacturing streams, including three types of gas oils (OGOs, SRGOs, VHGOs) and HFOs. Importantly, while we observed good separation between the gas oils and HFOs, qualitative and quantitative differences among samples of a given product group provide a measure for manufacturer- and batch-dependent compositional variation. IM-MS heatmaps reveal certain chemical-specific regions that are consistent with feature distributions and determined compositions, mostly nitrogen and oxygen-containing compounds, in previously published IM-MS data of crude oil and gasoline.[15] These characteristics reflect advancements in the removal of sulfur-containing compounds in modern refining.[38] Detailed feature identification still remains technically challenging due to the complexity of petroleum substances and resulting ion suppression; these limitations preclude the use of such data for read-across of the individual components of different UVCBs. However, we show that heteroatom signatures derived from IM-MS data enable effective communication of the degree of chemical similarity among UVCBs for scientific information-driven grouping of complex substances, information that should improve confidence in read-across.

Compositional information based on heteroatom feature identification reveals sample complexities that are largely concordant with previous characterizations based on GC-MS analyses.[27] For example, observed heteroatom-based carbon number ranges fall within, but did not entirely cover expected ranges for gas oils and the more complex HFOs. HFOs were clearly segregated from all gas oil samples due to their increased qualitative and quantitative complexity, primarily due to the presence of higher molecular weight, *i.e.* less volatile, constituents resulting in increased carbon number-, DBE-, and H/C ratio ranges. However, while traditional ARC content analysis reveals more pronounced differences among the three HFO samples, consistent with a lower degree of refinement and more pronounced product diversity as compared to gas oils, the heteroatom signatures indicated strong similarities and thus did not reflect the variation in the hydrocarbon composition. OGOs were characterized as the samples bearing the lowest complexity, and were therefore distinguishable from their related gas oils within the SRGO or VHGO categories. SRGOs and VHGOs were indistinguishable using the current methodology, which could be reflective of their overall high degree of chemical similarity and could provide a rationale for clustering them in a single category. Altogether, chemical signature based-comparisons were highly concordant with previously observed trends based on bioactivity profiles.[22] Specifically, the case study of about two dozen petroleum substances from the same product groups used herein revealed distinct groups of petroleum substances similar to the manufacturing process-based categories based on the concentration-response bioactivity profiles of cell function and toxicity in human induced pluripotent stem cell-derived cardiomyocytes and hepatocytes. Both studies indicated some overlap among SRGOs and VHGOs, but also a clear distinction between these gas oils and OGOs as well as HFOs.

Another important consideration of our findings is the relationship between substance chemical composition and possible adverse effects. For example, ARC content of petroleum substances tested herein has been related to their dermal carcinogenic potential.[24, 25, 30] Thus, it may be not surprising that ARC while content may be less useful for chemical fingerprinting and empirical groupings of petroleum substances as compared to heteroatom-class distribution. This raises a question of whether heteroatom content of petroleum substances, in addition to being useful for classification and grouping, may be indicative of their potential toxicity. Comprehensive correlation analysis between heteroatom content and biological activity of petroleum substances will require more extensive data sets and the utility of these data for toxicity prediction cannot be evaluated as only a limited number of substances was examined.

There are a number of other limitations to the approach presented in this work. Electrospray ionization is most efficient for polar molecules. Because many heteroatom (N, O, S)-containing components of petroleum substances are highly polar, ESI is specific and especially efficient in generating their gas-phase ions which make up 10% or less of the crude oil content and can be partially unstable.[10] Hydrocarbons on the other hand are, except for some relatively polar molecules, not effectively ionized by ESI but constitute on average over 90% of crude oil components. Thus, our methodology, while already useful for grouping of complex petroleum substances, is not sufficiently comprehensive with respect to the hydrocarbon complexity of these substances. There are several potential solutions to address this limitation. First, additional chemical and physico-chemical characterization of the same substances with other analytical techniques can be utilized to achieve more holistic approach as no single technology is capable of providing a comprehensive chemical fingerprint of complex petroleum samples. For example, two-dimensional[39] and conventional gas chromatography methods,[40] or vacuum ultraviolet photoionization mass spectrometry,[11] can derive additional features suitable for “fingerprinting” of the individual compounds in UVCBs. Second, alternative ionization methods, such as field desorption/ionization[41], atmospheric pressure chemical ionization, and atmospheric pressure photoionization,[42] may provide access to many of the remaining petroleum components including cycloalkanes, polycyclic aromatic hydrocarbons, and even to benzo- and dibenzothiophenes, and furans that are present at very low amounts. Third, future studies should be targeted at adapting and optimizing the herein presented approach towards chemical fingerprinting of more acidic components in negative ion mode.[15]

However, even in the absence of hydrocarbon-based signatures and the more acidic heteroatom-containing molecules, the concordance among observed heteroatom-class composition in positive ion mode, previously determined associated bioactivities[22, 43], and manufacturing process-associated sample clustering indicates the potential of these data to be applicable as a surrogate data layer that can potentially suffice to justify chemical similarity within a regulatory category. An added benefit is that IM-MS data can be acquired considerably faster as compared to, for example, GC-MS (minutes vs hours), allowing both higher sample throughput and a timely analysis for samples.

Overall, this work charts a path to an analytical chemistry-based method for grouping chemically similar UVCB substances which is a major step to increase confidence in read-

across. We successfully derived heteroatom class distribution and abundance profiles for diverse petroleum substances representing several distinct manufacturing streams. These data were used to demonstrate manufacturing-group specific similarities, as well as batch- and manufacturer-dependent variation, using vizualizations that communicate global compositional fingerprinting of complex petroleum substances. Importantly, there is concordance between groupings based on chemical signatures and the manufacturing process-based categories, as well as sample bioactivity.[22]

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was funded, in part, by grants from the United States Environmental Protection Agency (STAR RD83516602) and the National Institutes of Health (P42 ES005948). F. Grimm was the recipient of the 2015 Society of Toxicology-Colgate Palmolive postdoctoral fellowship in *in vitro* toxicology. P. J. Boogaard is employed by Shell International BV, a manufacturer of many of the substances tested herein. H. B. Ketelslegers is employed Concawe, a division of the European Petroleum Refiners Association, a trade group that provided all substances used in these analyses. T. Roy conducts contract research for a variety of commercial clients, including Concawe.

References

1. Rasmussen K, Pettauer D, Vollmer G, Davis J. Compilation of EINECS: Descriptions and definitions used for UVCB substances: Complex reaction products, plant products, (post-reacted) naturally occurring substances, micro-organisms, petroleum products, soaps and detergents, and metallic compounds. Tox Env Chem. 1999; 69(3–4):403–416.
2. Clark CR, McKee RH, Freeman JJ, Swick D, Mahagaokar S, Pigram G, Roberts LG, Smulders CJ, Beatty PW. A GHS-consistent approach to health hazard classification of petroleum substances, a class of UVCB substances. Regul Toxicol Pharmacol. 2013; 67(3):409–20. [PubMed: 24025648]
3. International Energy Agency. Energy Statistics Manual. Paris, France: 2005.
4. European Chemicals Agency. Guidance for identification and naming of substances under REACH and CLP. Helsinki, Finland: 2016.
5. ECHA. Registration Statistics. Vol. 2015. Helsinki, Finland: Jan. 2015
6. Boogaard PJ, Banton MI, Dalbey W, Hedelin AS, Riley AJ, Rushton EK, Vaissiere M, Minsavage GD. A consistent and transparent approach for calculation of Derived No-Effect Levels (DNELs) for petroleum substances. Regul Toxicol Pharmacol. 2012; 62(1):85–98. [PubMed: 22178770]
7. Fernandez-Lima FA, Becker C, McKenna AM, Rodgers RP, Marshall AG, Russell DH. Petroleum crude oil characterization by IMS-MS and FTICR MS. Anal Chem. 2009; 81(24):9941–7. [PubMed: 19904990]
8. Cho Y, Ahmed A, Islam A, Kim S. Developments in FT-ICR MS instrumentation, ionization techniques, and data interpretation methods for petroleomics. Mass Spectrom Rev. 2015; 34(2):248–63. [PubMed: 24942384]
9. ASTM International. UOP990-11. ASTM International; West Conshohocken, PA: 2011. Organic Analysis of Distillate by Comprehensive Two-Dimensional Gas Chromatography with Flame Ionization Detection.
10. Marshall AG, Rodgers RP. Petroleomics: chemistry of the underworld. Proc Natl Acad Sci U S A. 2008; 105(47):18090–5. [PubMed: 18836082]
11. Worton DR, Zhang H, Isaacman-VanWertz G, Chan AW, Wilson KR, Goldstein AH. Comprehensive Chemical Characterization of Hydrocarbons in NIST Standard Reference Material 2779 Gulf of Mexico Crude Oil. Environ Sci Technol. 2015; 49(22):13130–8. [PubMed: 26460682]

12. Ibrahim YM, Baker ES, Danielson WF 3rd, Norheim RV, Prior DC, Anderson GA, Belov ME, Smith RD. Development of a New Ion Mobility (Quadrupole) Time-of-Flight Mass Spectrometer. Int J Mass Spectrom. 2015; 377:655–662. [PubMed: 26185483]
13. Murray JA. Qualitative and quantitative approaches in comprehensive two-dimensional gas chromatography. J Chromatogr A. 2012; 1261:58–68. [PubMed: 22647189]
14. Manzano C, Hoh E, Simonich SL. Quantification of complex polycyclic aromatic hydrocarbon mixtures in standard reference materials using comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry. J Chromatogr A. 2013; 1307:172–9. [PubMed: 23932031]
15. Santos JM, Galaverna RD, Pudenzi MA, Schmidt EM, Sanders NL, Kurulugama RT, Mordehai A, Stafford GC, Wisniewski A, Eberlin MN. Petroleomics by ion mobility mass spectrometry: resolution and characterization of contaminants and additives in crude oils and petrofuels. Anal Methods. 2015; 7(11):4450–4463.
16. Ibrahim YM, Garimella SV, Prost SA, Wojcik R, Norheim RV, Baker ES, Rusyn I, Smith RD. Development of an Ion Mobility Spectrometry-Orbitrap Mass Spectrometer Platform. Anal Chem. 2016; 88(24):12152–12160. [PubMed: 28193022]
17. Dimitrov SD, Georgieva DG, Pavlov TS, Karakolev YH, Karamertzanis PG, Rasenberg M, Mekenyan OG. UVCB substances: methodology for structural description and application to fate and hazard assessment. Environ Toxicol Chem. 2015; 34(11):2450–62. [PubMed: 26053589]
18. Quann RJ. Modeling the chemistry of complex petroleum mixtures. Environ Health Perspect. 1998; 106(Suppl 6):1441–8. [PubMed: 9860903]
19. Redman AD, Parkerton TF, Comber MH, Paumen ML, Eadsforth CV, Dmytrasz B, King D, Warren CS, den Haan K, Djemel N. PETRORISK: a risk assessment framework for petroleum substances. Integr Environ Assess Manag. 2014; 10(3):437–48. [PubMed: 24687890]
20. Gray TM, Simpson BJ, Nicolich MJ, Murray FJ, Verstuyft AW, Roth RN, McKee RH. Assessing the mammalian toxicity of high-boiling petroleum substances under the rubric of the HPV program. Regul Toxicol Pharmacol. 2013; 67(2 Suppl):S4–9. [PubMed: 23247262]
21. Bierkens J, Geerts L. Environmental hazard and risk characterisation of petroleum substances: a guided "walking tour" of petroleum hydrocarbons. Environ Int. 2014; 66:182–93. [PubMed: 24607926]
22. Grimm FA, Iwata Y, Sirenko O, Chappell GA, Wright FA, Reif DM, Braisted J, Gerhold DL, Yeakley JM, Shepard P, Seligmann B, Crittenden C, Roy T, Boogaard PJ, Ketelslegers HB, Rohde AM, Rusyn I. A chemical–biological similarity-based grouping of complex substances as a prototype approach for evaluating chemical alternatives. Green Chem. 2016; 18:4407–4419. [PubMed: 28035192]
23. Hsu CS, Hendrickson CL, Rodgers RP, McKenna AM, Marshall AG. Petroleomics: advanced molecular probe for petroleum heavy ends. J Mass Spectrom. 2011; 46(4):337–43. [PubMed: 21438082]
24. Roy TA, Johnson SW, Blackburn GR, Mackerer CR. Correlation of mutagenic and dermal carcinogenic activities of mineral oils with polycyclic aromatic compound content. Fundam Appl Toxicol. 1988; 10(3):466–76. [PubMed: 3286347]
25. McKee RH, Nicolich M, Roy T, White R, Daughtrey WC. Use of a statistical model to predict the potential for repeated dose and developmental toxicity of dermally administered crude oil and relation to reproductive toxicity. Int J Toxicol. 2014; 33(1 Suppl):17S–27S. [PubMed: 24179028]
26. Metz TO, Baker ES, Schymanski EL, Renslow RS, Thomas DG, Causon TJ, Webb IK, Hann S, Smith RD, Teeguarden JG. Integrating ion mobility spectrometry into mass spectrometry-based exposome measurements: what can it add and how far can it go? Bioanalysis. 2017; 9(1):81–98. [PubMed: 27921453]
27. CONCAWE. REACH – Analytical characterisation of petroleum UVCB substances. Brussels, Belgium: 2012.
28. Roth RN, Simpson BJ, Nicolich MJ, Murray FJ, Gray TM. The relationship between repeat-dose toxicity and aromatic-ring class profile of high-boiling petroleum substances. Regul Toxicol Pharmacol. 2013; 67(2 Suppl):S30–45. [PubMed: 23751816]

29. Murray FJ, Roth RN, Nicolich MJ, Gray TM, Simpson BJ. The relationship between developmental toxicity and aromatic-ring class profile of high-boiling petroleum substances. Regul Toxicol Pharmacol. 2013; 67(2 Suppl):S46–59. [PubMed: 23680405]

30. Roy TA, Blackburn GR, Mackerer CR. Evaluation of Analytical End-Points to Predict Carcinogenic Potency of Mineral-Oils. Polycycl Aromat Comp. 1994; 5(1–4):279–287.

31. OECD. Guidance document on grouping of chemicals. Environment Directorate, Organisation for Economic Co-operation and Development (OECD); Paris, France: 2014.

32. ECHA. How to report read-across and categories. European Chemical Agency; Helsinki, Finland: 2012.

33. Berggren E, Amcoff P, Benigni R, Blackburn K, Carney E, Cronin M, Deluyker H, Gautier F, Judson RS, Kass GE, Keller D, Knight D, Lilienblum W, Mahony C, Rusyn I, Schultz T, Schwarz M, Schuurmann G, White A, Burton J, Lostia AM, Munn S, Worth A. Chemical Safety Assessment Using Read-Across: Assessing the Use of Novel Testing Methods to Strengthen the Evidence Base for Decision Making. Environ Health Perspect. 2015; 123(12):1232–40. [PubMed: 25956009]

34. Grimm FA, Iwata Y, Sirenko O, Bittner M, Rusyn I. High-Content Assay Multiplexing for Toxicity Screening in Induced Pluripotent Stem Cell-Derived Cardiomyocytes and Hepatocytes. Assay Drug Dev Technol. 2015; 13(9):529–46. [PubMed: 26539751]

35. Reddy CM, Quinn JG. GC-MS analysis of total petroleum hydrocarbons and polycyclic aromatic hydrocarbons in seawater samples after the North Cape oil spill. Marine Poll Bull. 1999; 38(2): 126–135.

36. Ponthus J, Riches E. Evaluating the multiple benefits offered by ion mobility-mass spectrometry in oil and petroleum analysis. Int J Ion Mobil Spec. 2013; 16(2):95–103.

37. Zhan DL, Fenn JB. Electrospray mass spectrometry of fossil fuels. Int J Mass Spectrom. 2000; 194(2–3):197–208.

38. CONCAWE. Sulphur dioxide emissions from oil refineries and combustion of oil products in western europe and hungary. Brussels, Belgium: 1998.

39. Gros J, Reddy CM, Aeppli C, Nelson RK, Carmichael CA, Arey JS. Resolving biodegradation patterns of persistent saturated hydrocarbons in weathered oil samples from the Deepwater Horizon disaster. Environ Sci Technol. 2014; 48(3):1628–37. [PubMed: 24447243]

40. Wang Z, Yang C, Yang Z, Hollebone B, Brown CE, Landriault M, Sun J, Mudge SM, Kelly-Hooper F, Dixon DG. Fingerprinting of petroleum hydrocarbons (PHC) and other biogenic organic compounds (BOC) in oil-contaminated and background soil samples. J Environ Monit. 2012; 14(9):2367–81. [PubMed: 22796730]

41. Schaub TM, Hendrickson CL, Quinn JP, Rodgers RP, Marshall AG. Instrumentation and method for ultrahigh resolution field desorption ionization fourier transform ion cyclotron resonance mass spectrometry of nonpolar species. Anal Chem. 2005; 77(5):1317–24. [PubMed: 15732913]

42. Purcell JM, Hendrickson CL, Rodgers RP, Marshall AG. Atmospheric pressure photoionization fourier transform ion cyclotron resonance mass spectrometry for complex mixture analysis. Anal Chem. 2006; 78(16):5906–12. [PubMed: 16906739]

43. Tsitou P, Heneweer M, Boogaard PJ. Toxicogenomics in vitro as an alternative tool for safety evaluation of petroleum substances and PAHs with regard to prenatal developmental toxicity. Toxicol In Vitro. 2015; 29(2):299–307. [PubMed: 25481525]

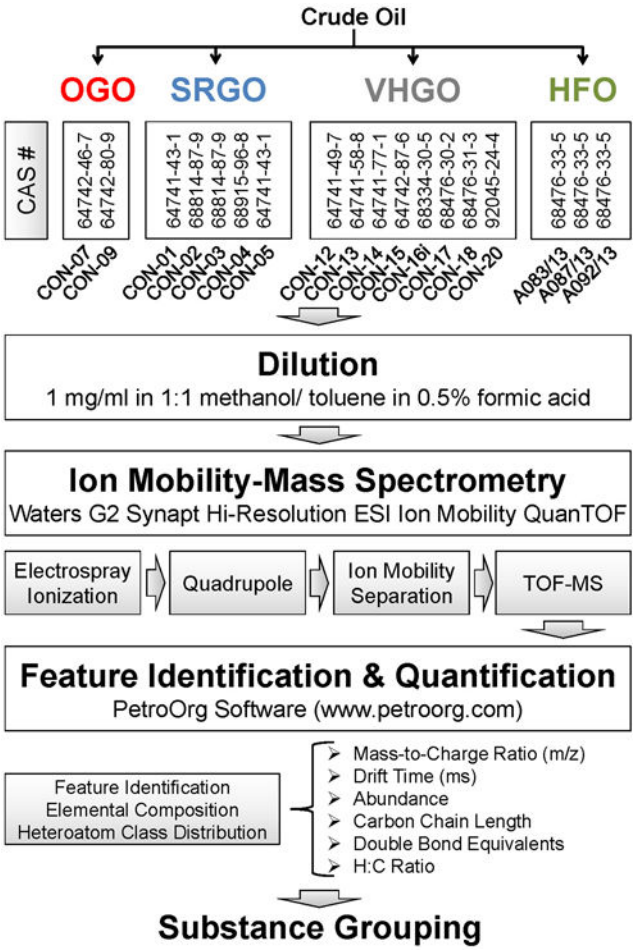

Figure 1.

Workflow for IM-MS based quantitative chemical grouping of petroleum substances from Other Gas Oils (OGO), Straight Run Gas Oils (SRGO), Vacuum & Hydrotreated Gas Oils (VHGO), and Heavy Fuel Oils (HFO) manufacturing streams.

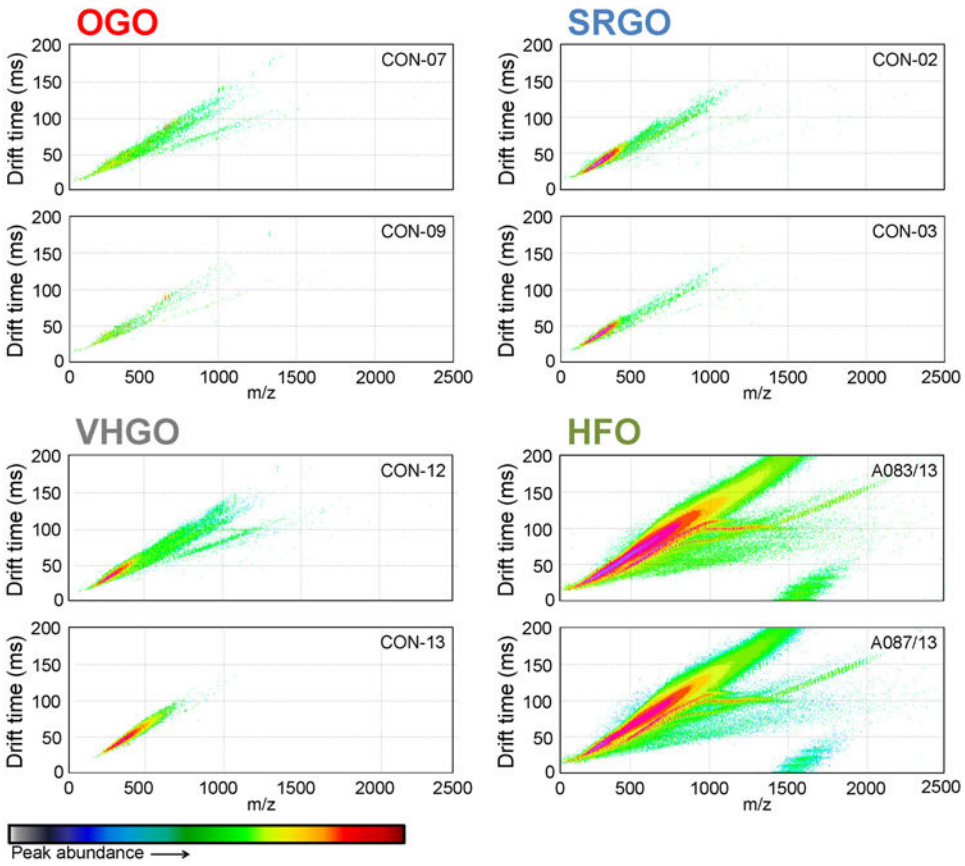

Figure 2.

IM-MS spectra for representative petroleum substances. The figure depicts IM-MS spectra (m/z vs drift time, feature abundance is highlighted by color intensity) for each two representative of OGO, SRGO, VHGO, and HFO.

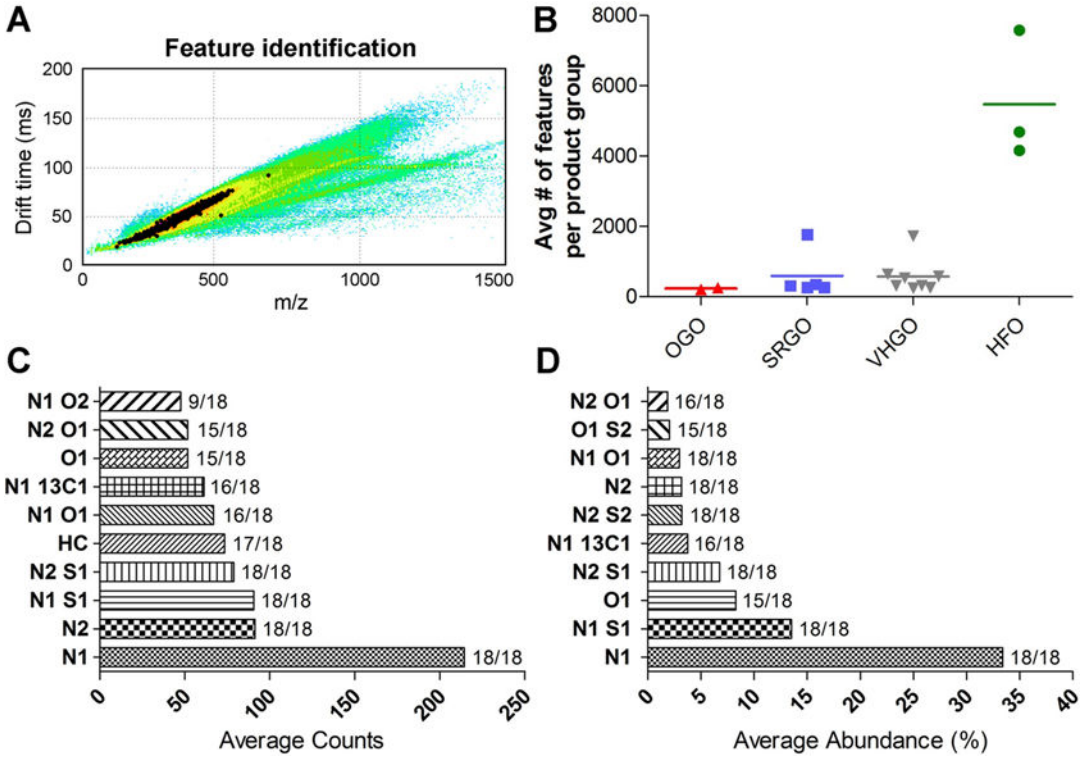

Figure 3.

IM-MS data processing and global feature identification. Computational integration of IM-MS data sets in PetroOrg software (CON-20 shown as a representative example) enables identification of the unique features (highlighted in black) and quantitation of heteroatom class distribution for each sample **(A)**. Plots in **(B)** depict the total number of features identified in each sample grouped into petroleum manufacturing streams. Averages for the top 10 most abundant heteroatom classes based on raw counts and relative abundance are shown in **(C)** and **(D)**, respectively. The frequency of feature detection in the various petroleum substances is indicated.

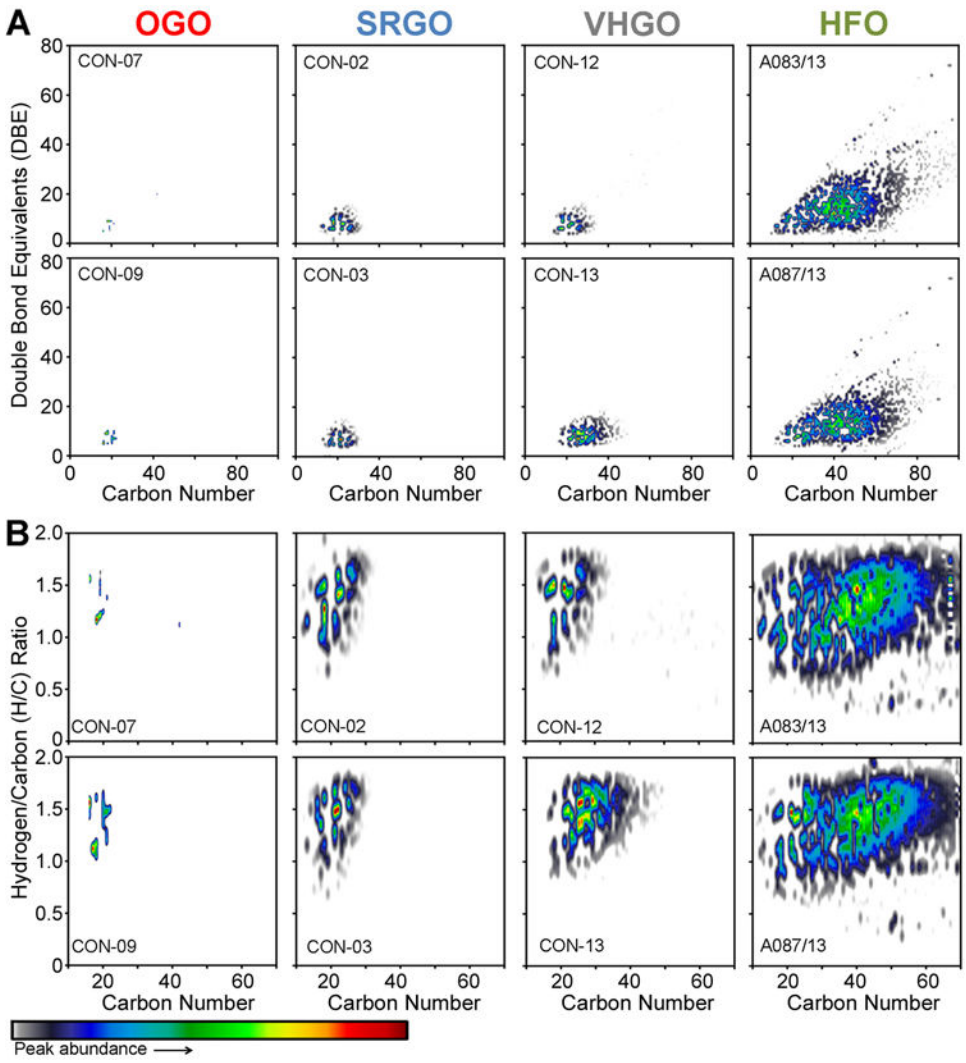

Figure 4.

Chemical composition-based grouping of petroleum substances. Plots visualize unique feature distributions for the N1 heteroatom class, i.e. carbon chain length vs double bond equivalents (DBE) **(A)** and carbon chain length vs hydrogen-to-carbon (H/C) ratio **(B)**. Two representative samples for each of four manufacturing streams (OGO, SRGO, VHGO, and HFO) are shown. The feature density is reflected through increasing color intensity.

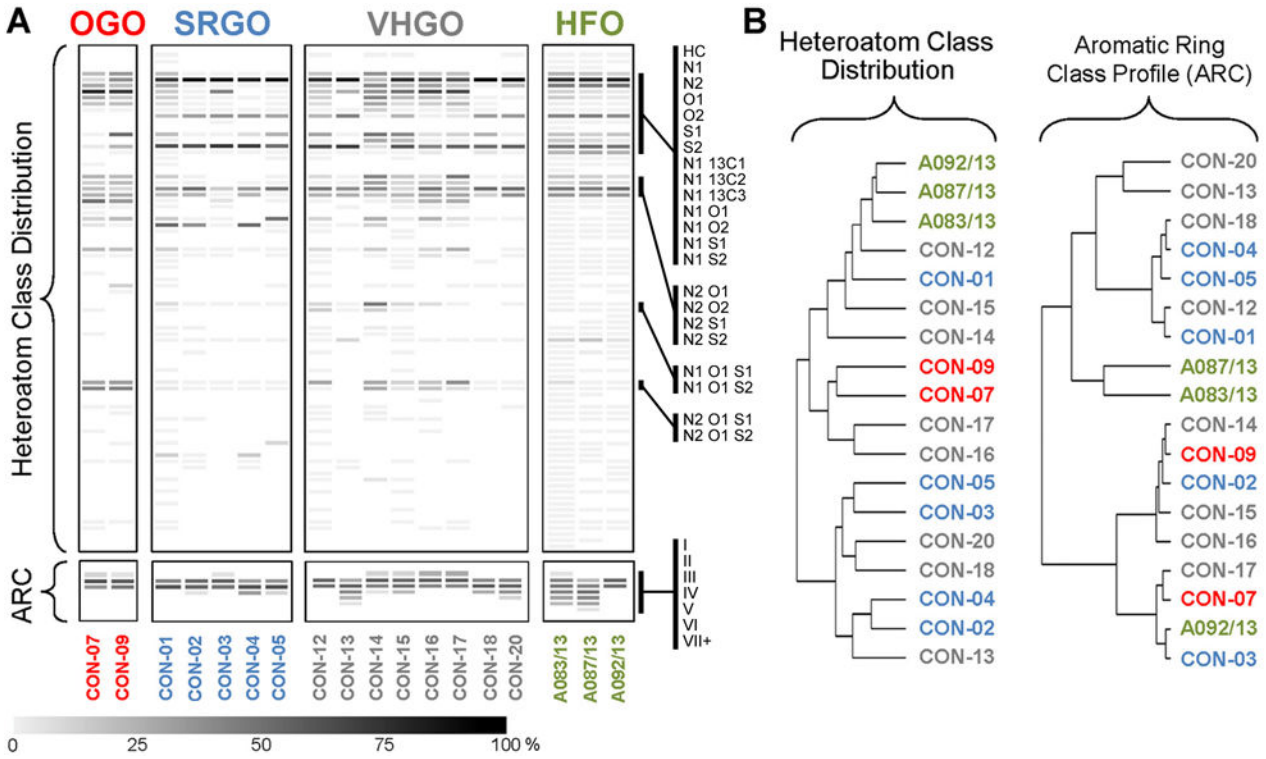

Figure 5.

Chemical fingerprinting and cluster analysis of petroleum substances. Heatmap **(A)** and cluster analysis **(B)** were performed based on the relative frequency of heteroatom class distribution or aromatic ring class (ARC) content. Representative informative features are identified as a sidebar in panel A. Complete data table with all heteroatom classes is included as Supplemental Table 1.

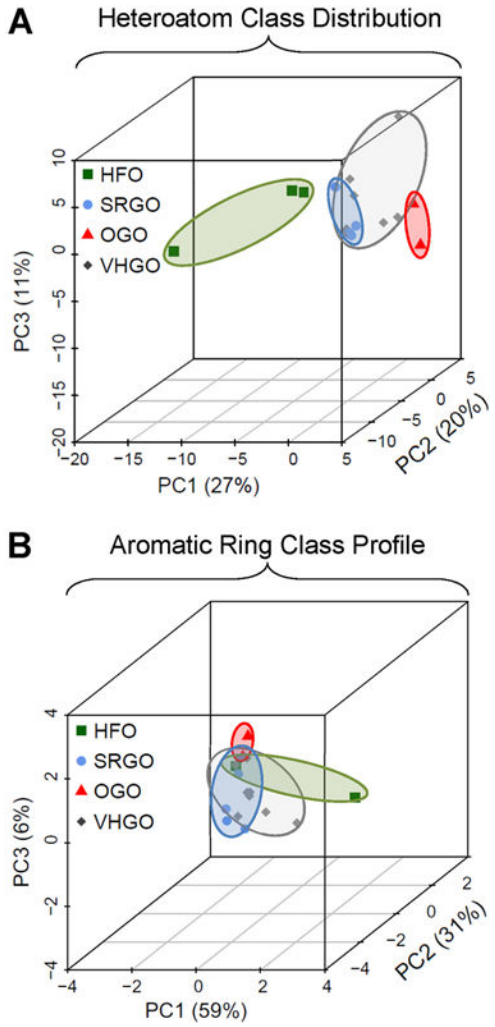

Figure 6.

Principal components-based grouping of petroleum substances. Principal components analysis was based on the relative heteroatom class distribution (**A**) and aromatic ring class profile (**B**) of individual petroleum substances.