HHS Public Access

Author manuscript

Fuel (Lond). Author manuscript; available in PMC 2023 June 01.

Published in final edited form as:

Fuel (Lond). 2022 June 01; 317: . doi:10.1016/j.fuel.2022.123547.

Characterization of Compositional Variability in Petroleum Substances

Alina T. Roman-Hubers[1], **Alexandra C. Cordova**[1], **Arlean M. Rohde**[1], **Weihsueh A. Chiu**[1], **Thomas J. McDonald**[2], **Fred A. Wright**[3], **James N. Dodds**[4], **Erin S. Baker**[4], **Ivan Rusyn**[1,*]

[1]Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, Texas 77843, United States;

[2]Departments of Environmental and Occupational Health, Texas A&M University, College Station, Texas 77843, United States;

[3]Departments of Statistics and Biological Sciences, Raleigh, North Carolina 27695, United States

[4]Department of Chemistry, North Carolina State University, Raleigh, North Carolina 27695, United States

Abstract

In the process of registration of substances of Unknown or Variable Composition, Complex Reaction Products or Biological Materials (UVCBs), information sufficient to enable substance identification must be provided. Substance identification for UVCBs formed through petroleum refining is particularly challenging due to their chemical complexity, as well as variability in refining process conditions and composition of the feedstocks. This study aimed to characterize compositional variability of petroleum UVCBs both within and across product categories. We utilized ion mobility spectrometry (IMS)-MS as a technique to evaluate detailed chemical composition of independent production cycle-derived samples of 6 petroleum products from 3 manufacturing categories (heavy aromatic, hydrotreated light paraffinic, and hydrotreated heavy paraffinic). Atmospheric pressure photoionization and drift tube IMS-MS were used to identify

*Corresponding author: Ivan Rusyn, Texas A&M University, 4458 TAMU, College Station, TX 77843-4458, USA. Telephone: +1-979-458-9866; irusyn@tamu.edu.

Alina Roman-Hubers: Conceptualization, Methodology, Validation, Formal analysis, Data curation, Visualization, Writing- Original draft preparation

Alexandra C. Cordova: Methodology, Validation, Formal analysis, Data curation.

Arlean M. Rohde: Conceptualization, Visualization, Investigation.

Weihsueh A. Chiu: Software, Validation, Writing - Review & Editing

Thomas J. McDonald: Methodology, Resources, Writing- Review and Editing

Fred A. Wright: Software, Writing - Review & Editing

James N. Dodds: Methodology, Validation, Writing - Review & Editing

Erin S. Baker: Conceptualization, Methodology, Resources, Writing - Review & Editing

Ivan Rusyn: Conceptualization, Resources, Writing - Review & Editing, Visualization, Supervision, Project administration, Funding acquisition

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

structurally related compounds and quantified between- and within-product variability. In addition, we determined both individual molecules and hydrocarbon blocks that were most variable in samples from different production cycles. We found that detailed chemical compositional data on petroleum UVCBs obtained from IMS-MS can provide the information necessary for hazard and risk characterization in terms of quantifying the variability of the products in a manufacturing category, as well as in subsequent production cycles of the same product.

1. Introduction

Crude oil refining involves complex physical and chemical processes such as distillation, cracking, isomerization, reforming, alkylation and hydrodesulphurization, ultimately yielding petroleum products of certain performance characteristics that are subsequently used for a variety of applications [1, 2]. Because of the chemical complexity and variability of the oil feedstocks, as well as differences in the refining process conditions within and across manufacturing sites, it is expected that the types and quantities of hydrocarbons and other constituents present in downstream products may vary both within and between manufacturers, even for the same refining processes and products [3]. This inherent compositional complexity and variability of petroleum substances, products that fall into the class known as substances of unknown, variable composition, complex reaction products, or biological materials (UVCBs), presents unique challenges for their registration and evaluation [4, 5]. Current naming conventions and grouping of petroleum UVCBs into manufacturing categories is based on the information on their general composition such as carbon chain length and boiling point ranges, other physicochemical properties, performance characteristics, and proposed use(s) [2, 4]. While Chemical Abstract Service (CAS) and European Inventory of Existing Commercial Chemical Substances (EINECS) identifications have been assigned to petroleum products, they are further grouped into broad manufacturing categories for registration and regulatory evaluation [6]. The existing nomenclature for petroleum UVCBs, either for the individual product identifiers or for broad manufacturing categories, is deemed generally sufficient for the purpose of naming and identification of these products [2, 4, 5].

Once identified, petroleum substances must be registered following the laws and regulations of the jurisdiction where they are to be manufactured or used. The European Union REACH (Registration, Evaluation, Authorization and Restriction of Chemicals) regulation [7] specifies human health and the environment hazard data requirements that must be met before authorization is given for their use. Most petroleum UVCBs have considerable data gaps that need to be addressed in the process of registration either through additional testing or read-across to another substance, or other members in a group of substances, that have the requisite information for registration purpose [8]. Recent proposals for grouping of “similar” petroleum UVCBs into manufacturing categories, in part based on the “*broadly similar [chemical] composition*,” have been questioned by the regulatory bodies as suitable for read-across, such as the European Chemicals Agency (ECHA). Concerns were raised about the strength of the justification for the proposed grouping and read-across [9]. Recently, the European Commission has amended Annex XI of REACH clarifying that for the application of read-across/grouping, “*structural similarity for UVCB substances shall be established*

on the basis of similarities in the structures of the constituents (…) and variability in the concentration of these" [10]. Indeed, the chemical variability in petroleum products is one well-appreciated concern, because petroleum substances "*are UVCBs and are manufactured to specifications based on performance characteristics rather than chemical composition, analysis of the same substance manufactured in the same location at different times could show a considerable variation in composition*" [11].

To address the challenge of quantifying the variability of the petroleum UVCBs, a number of analytical approaches have been used to characterize their composition ranging from physicochemical analyses to detailed mass spectrometry (MS)-based methods [3, 11]. Despite recent advances in petroleomics including novel high-resolution (HR) MS methods [12], the ability to thoroughly assess the chemical composition of petroleum UVCBs, including the analysis of isomeric species, in the context of REACH has yet to be shown due to the similarity of the hydrocarbon components (*i.e.*, presence of isomeric species). In addition, the incomplete understanding of variability (*i.e.*, among samples from independent production cycles) represents a key barrier to the application of read-across between products in the same category, as any quantification of substance-to-substance similarity must be informed by within-substance variability. Accordingly, we set out to quantify both within- and between-product variability for representative petroleum UVCBs. We used both gas chromatography (GC)-MS and ion mobility spectrometry (IMS)-MS techniques, because recent studies have demonstrated the utility of IMS-MS for determining the composition of petroleum UVCBs [13–17]. We identified structurally related hydrocarbon and heteroatom compounds, examined hydrocarbon blocks, and characterized within-product variability.

2. Methods

2.1. Samples of petroleum products.

A total of six refined petroleum products (Table 1) were used in this study. For evaluation, we selected three products from three broad manufacturing categories of "Solvent naphtha, heavy aromatic products" (marked as AR) and "Petroleum distillates, hydrotreated [light or heavy] paraffinic" (marked as BO). Sample selection was meant to be representative of a wide range of expected chemical complexity of petroleum UVCBs. For each product, samples were obtained from 2–3 independent production cycles at the same refinery (samples were collected 2–3 months apart), resulting in a total of 16 samples (Supplemental Table 1). For GC-MS analyses, samples were weighed and dissolved in dichloromethane (CAS no. 75-09-2, catalog no. 34856; Sigma-Aldrich, St. Louis, MO) to a final concentration of 1 mg/mL. For the IMS-MS analyses, 1 mg of each sample was first dissolved in 9 mL of 1:1 (v/v) mixture of toluene (CAS no. 108-88-3, catalog no. 34866; Sigma-Aldrich) and methanol (CAS no. 67-56-1, catalog no. 34860; Sigma-Aldrich). Next, 25 μL of the solution was mixed with 300 μL of the same mixture of toluene and methanol and injected directly.

2.2. GC-MS instrumental analysis and data processing.

The modified United States Environmental Protection Agency (US EPA) method 8270 was carried out in full scan analysis mode using an Agilent 7890 GC (Agilent Technologies, Santa Clara, CA) interfaced with a Hewlett-Packard (HP) 5976 MS. Additionally, a HP-5ms Ultra Inert Column (30 m × 0.25 μm × 0.25 mm; catalog no. G3900–63001; Agilent Technologies) was used to chromatographically separate the petroleum hydrocarbons. Instrumental operating conditions were as follows: mass range 40 to 500m/z, splitless injector, injection volume of 2 μL, column flow 1 mL/min, helium carrier gas. Initial temperature of the injection port was held at 250°C. The oven was initially set to 50°C with a hold time of 4 min; then, the oven was programmed at a rate of 6°C/min until it reached the final holding temperature of 300°C with a final hold time of 20 min. Individual full-scan total ion chromatograms for each sample were processed using ChemStation Data Analysis Software (Agilent). Raw data consisted of 10,127 scans at 1 atomic mass unit (amu) bins from 40 to 500 amu. Data for each amu bin across all scans was averaged and the final data matrix consisted of an average abundance value for each amu bin for one sample. The data for 16 samples were combined into a two-dimensional data matrix of mass range versus average fragment ion intensities. See Supplemental Table 2 for the resulting GC-MS data matrix and Supplemental Figure 1 for the GC-MC chromatograms for each sample.

2.3. IMS-MS instrumental analysis and data processing.

For the IMS-MS analyses, we utilized an Agilent Technologies G6560A platform coupling drift tube IMS (resolving power (RP) ≈ 60) and a quadrupole time-of-flight (QTOF) mass spectrometer (RP ≈ 25 000). In all experiments, the drift tube was filled with nitrogen gas and the samples were ionized with an atmospheric pressure photoionization (APPI) source (model G1917C; Agilent Technologies). The instrument was calibrated prior to running samples according to the Agilent protocol for 50–1,700 *m/z* range, using the atmospheric pressure chemical ionization (APCI)-L Low Concentration tuning mix solution (part #G1969–85010, Agilent Technologies). The petroleum samples (200 μL) were then infused directly at a flow rate of 50 μL/min, and analysis included three technical replicates for each sample. Instrumental and source parameters were as follows: APPI positive ion mode, sample analysis time 1.5 min; source parameters: gas temperature 325 °C, vaporizer 350 °C, drying gas 10 L/min, nebulizer 30 psi, VCap 3000, fragment 400 V, 110 RF Vpp 750. The following acquisition parameters were defined for each instrumental run: mass range 50 to 1700 *m/z*, frame rate 1 frames/s, IM transient rate 18 transients/frame, maximum drift time 60 ms, time-of-flight transient rate 600 transients/IM transient, trap fill time 20 000 μs and trap release time 300 μs. QTOF parameters were as follows: firmware Ver 18.723, Rough Vac 2.71 torr, Quad Vac 3.68×10^{-5} torr, TOF Vac 3.47×10^{-7} torr, drift tube pressure 3.940 torr, trap funnel pressure 3.790 torr, chamber voltage 5.96 μA, and capillary voltage 0.076 μA. Data were obtained using the MassHunter Acquisition software (Agilent; ver. 08.00). Each sample was analyzed in triplicate in three independent experimental batches using the instrument and setting as detailed above. Two of these experiments were conducted at Texas A&M University on separate days (about 1 month apart) by different individuals. One of the experiments was conducted at North Carolina State University. These replication studies were using the same model IMS-MS instrument and experimental conditions, but were conducted by three different individuals.

IMS-MS raw data files from each instrumental run were processed using MassHunter Browser Acquisition Data software (Agilent Technologies, ver. 08.00) to derive nitrogen gas-filled drift tube collisional cross section ($^{DT}CCS_{N2}$) values for all detected features [18]. In this manuscript, a feature is defined as a potential molecule's isotopic envelope and by having both MS and IMS dimensions, all isotopes must occur at the same IMS drift time. Next, data files for all samples and their respective technical replicates (16 samples × 3 technical replicates = 48 files) were uploaded to Agilent MassProfiler software (Agilent Technologies, B.08.00) for feature alignment based on drift time (±5.0%) and mass (±15ppm+5mDa). Finally, aligned raw data matrices for each experimental batch (Supplemental Table 3) were filtered to select features with abundance >5,000 in two out of three technical replicates for each sample. These filtered data (Supplemental Table 4) include information on the constituents present in high abundance that would be of most relevance with regards to hazard evaluation of petroleum UVCBs [19]. The filtering parameters were selected based on the general consideration of the presence of ^{13}C isotopic partner for individual features and previous data analyses [17] that showed erosion in confidence for molecular formulae assignments for the features of low abundance; however, alternative thresholds may be selected using the datasets provided in Supplemental Table 3.

After alignment and filtering as detailed above, the data (Supplemental Table 4), including technical replicates (n=3), was used for feature identification using an IMS-MS data processing workflow detailed elsewhere [17]. Briefly, each feature was cross-referenced to a $^{DT}CCS_{N2}$ standard library containing a number of hydrocarbon standards [20]. Features were deemed matching to a molecule in the database at a $^{DT}CCS_{N2}$ tolerance of ±1% and an *m/z* tolerance of ±5 ppm and ±2 mDa. Then, the Kendrick Mass Defect (KMD) was calculated for each feature using base units of CH_2 (14.01565) and H (1.00783) to identify features that fall into homologous series (KMD-CH_2 of ±1.00 parts per thousand, ppt). Next, the elemental composition was assigned to each feature if it was in homologous series using the $^{DT}CCS_{N2}$ library matched features as reference points, as well as based on the KMD-H analyses as detailed in [17] (Supplemental Table 5). Carbon chain length and double bond equivalency (DBE) of each feature were calculated from the elemental composition [21]. Based on the elemental formula and other properties, each feature was assigned a carbon chain length and hydrocarbon class.

2.4. Data analysis.

We reasoned that quantifying overall similarity among samples would be instructive to illustrate the informativeness of the data. Thus, the GC-MS and IMS-MS data matrices of all samples (Supplemental Tables 2–4) were used for hierarchical clustering [22] based on Spearman correlation and average linkage using the *hclust* package in R Studio (ver. R-4.1.0). The correlation among samples was then visualized in a dendrogram. In order to assess the similarity between clusters, the Fowlkes-Mallows (FM) index [23] was used to evaluate the concordance of experimental data-derived clustering to that of the pre-defined manufacturing categories, products, independent production cycles, and technical replicates of each sample. The technical replicates and independent production cycles were considered as separate instances of the same substance, and the FM index was compared between the pre-defined categories and the hierarchical clustering having a number of clusters equal

to the number of manufacturing categories, using the *cuttree* command on the clustered tree. FM index values can range from 0 (no correspondence) to 1 (perfect correspondence). Principal components analysis (PCA) was carried out to evaluate similarity between the products and samples using the *prcomp* and *ggplot* packages in R Studio (4.1.0) and based on characterized features, carbon chain length, hydrocarbon class and heteroatom species. For analysis at the individual ion level (*i.e.*, full dataset of 55,466 features), the differences in abundance of each feature in samples from the independent production cycles were quantified as the maximum of the absolute value (if at least 3 samples were available) of the fold change difference when comparing across all pairs of samples from independent production cycles. For each feature, a p-value for variability was assessed by a one-way analysis of variance, using production cycle as a factor. Correction for multiple comparisons across features was performed using the Benjamini-Hochberg *q*-value computed in R using *p.adjust* (Supplemental Table 6). For the analysis at the level of carbon chain length, hydrocarbon class and heteroatom profile, the variability in chemical composition among samples of independent production cycles was evaluated based on the relative abundance of the molecules in each aggregated set of features using two-way ANOVA with Sidak's multiple comparison test (GraphPad Prism 9.0, San Diego, CA) followed by the Bonferroni correction [24].

3. Results

This study evaluated 16 samples of 6 oil refining-derived products that fall into three broad manufacturing categories of petroleum UVCBs (Table 1). We began by analyzing samples using conventional GC-MS technique. Figure 1A shows superimposed GC-MS full-scan chromatograms for representative samples of each product (see Supplemental Figure 1 for the similar chromatograms of each sample). These data clearly demonstrate the difference among samples from diverse manufacturing categories and the "cuts" of hydrocarbons varied considerably among samples as evidenced by the retention time differences. The "solvent naphtha (petroleum), heavy aromatic" products AR150 and AR200 were readily separated by GC-MS. The "distillates (petroleum), hydrotreated" light (BO60), or heavy (BO100, BO220 and BO600) products were more complex as they yielded characteristic unresolved complex mixture (UCM) "humps" on the GC-MS chromatograms. To visualize the similarity among products in the GC-MS data, we used the data matrix of averaged intensities for each of the amu bins (from 40 to 500 amu) to conduct unsupervised hierarchical clustering analysis. Figure 1B shows that samples from independent production cycles of the same product clustered together, except for one BO220 sample (production cycle 1). Moreover, solvent naphtha samples and the hydrotreated paraffinic distillate samples also formed distinct groups commensurate with their manufacturing category and CAS# groupings (Table 1). Samples of product BO60 from independent production cycles were most dissimilar to each other, yet they still clustered into their own group. Based on GC-MS data, the concordance in the clustering of the samples, as compared to pre-determined manufacturing category assignments for each sample, was modest (FM index of 0.49).

Samples were next analyzed using the IMS-MS platform. Figure 2 shows representative two-dimensional nested spectra for representative samples of each product where they are

plotted by *m/z* (x-axis) and drift time (y-axis, parameter used to calculate $^{DT}CCS_{N2}$) with feature abundance represented by color intensity. Supplemental Figure 2 shows IMS-MS nested spectra for each sample analyzed. These plots illustrate the differences in both complexity (total number of features) and changes in *m/z* and structural sizes (IMS) of the individual constituents in the samples. For example, samples of solvent naphtha (petroleum) heavy aromatic products AR150 and AR200 contained compounds in the mass range of 50–400 *m/z*. The hydrotreated distillates light (BO60) product contained lower *m/z* range species as compared to the hydrotreated distillates heavy (BO100, BO220 and BO600) products that spanned a mass range of up to 700 *m/z*.

Unsupervised hierarchical clustering utilizing *m/z*, $^{DT}CCS_{N2}$ and feature abundance data from the IMS-MS analyses was then used to compare composition similarity among the samples. For these analyses, both the full data matrix (Supplemental Table 3) and filtered (i.e., most abundant features) data was assessed (Supplemental Table 4). Figure 3 illustrates the full and filtered abundance dendrograms where the technical replicates were averaged for the analyses. When the full IMS-MS data was used (Supplemental Table 3A), samples from independent production cycles of the same manufacturing category clustered together resulting in three main clusters. For example, samples of BO60 product clustered closer to the AR150 and AR200 products and not with the other products (BO100, BO220 and BO600). Additionally, the concordance in sample clustering for pre-determined manufacturing category assignments was excellent (FM index of 1.0). When the IMS-MS datasets were filtered for only highest abundance features (Supplemental Table 4A), similar clustering was observed, and the FM index for this analysis was also 1.0. Furthermore, when technical replicate samples were included in these analyses, similar results were obtained (Supplemental Figure 3). Specifically, technical replicates of each sample clustered together, and then with the samples from different production cycles for each product, and finally with other products within a manufacturing category.

Our next assessment was to determine how well petroleum UVCBs group using IMS-MS data obtained in independent experiments by distinct operators and in a different laboratory. For these studies, the samples were analyzed at Texas A&M on the same instrument but by a different operator and then at NC State University by another operator and instrument, but the same model of IMS-MS platform (G6560) and an identical experimental protocol. In all cases, samples were prepared independently from the neat stocks of each product (see Methods) before each instrumental analysis. Abundance-filtered data (Supplemental Table 4), where technical replicates were averaged, were used for the following comparisons. Irrespective of the laboratory or operator, strong correlation between samples from independent production cycles was evident as products clustered within their manufacturing category and CAS# (Figure 4 and Supplemental Table 7). The FM index values for clustering were 1.0 for two experiments (Figures 4A–B) and 0.86 for the third one (Figure 4C). These results indicate high reproducibility of the IMS-MS technique for the analysis of similarities in samples of complex-composition petroleum UVCBs.

While clustering using multidimensional data from untargeted IMS-MS is useful to establish the overall similarity of the samples for the purpose of substance identification [16], this information may not be adequate for product registration because it does not provide

sufficient detail on the chemical composition of each sample. To address this challenge, we identified structurally related compounds [17] in analyzed petroleum products to obtain molecular formula assignments to the high abundance features. Because each of three independent IMS-MS experiments (Figure 4) yielded similar clustering of the samples, filtered IMS-MS data from one of the experiments was used herein (Supplemental Table 4A). Molecular formulas for each feature in the 16 samples are provided in Supplemental Table 5. Similar to our previous findings of analysis of refined products or crude oils [17], we were able to assign molecular formulas to 93% of the high abundance features across all samples.

Because the composition of petroleum UVCBs is typically presented using the hydrocarbon block method which groups closely related compounds by their carbon chain length and hydrocarbon class [25, 26], we used the assigned molecular formulas and other information from the KMD analysis (*i.e.*, homologous series and double bond equivalence) to aggregate the data into hydrocarbon blocks (Supplemental Table 8). We also determined whether any of the identified molecules were heteroatoms (Supplemental Table 9). With these data, we performed PCA to visualize the similarities between samples of independent production cycles, as well as differences among products across manufacturing categories (Figure 5). When all high abundance features with assigned molecular formulas (n=1,417, Figure5A) or carbon chain length (Figure 5B) were used for the PCA, four groups were discernable in the first two principal components. Group 1 and Group 2 distinguished between two heavy aromatic products (AR150 and AR200), while Group 3 separated the light (BO60) and Group 4 contained the heavy (BO100, BO220 and BO600) hydrotreated paraffinic distillate products. Interestingly, the latter group appeared more homogenous even though it contained samples from three different products. Additionally, samples from independent production cycles were closely aligned to each other. The molecular formula-level data showed tighter grouping between samples of the same product, while the data on hydrocarbon blocks (Figure 5C) or heteroatom profiles (Figure 5D) allowed fewer distinctions among product groups, there was wider separation between samples from production cycles.

To further evaluate the variability in petroleum UVCBs, we analyzed the relative abundance of the molecules in hydrocarbon blocks or heteroatoms between samples from independent production cycles of the same product (Supplemental Tables 8–9). Figure 6 shows an example of this analysis for product BO220. Figures 6A–C show the relative abundance of each hydrocarbon block, as well as total abundance for each carbon chain length and hydrocarbon class. It is evident that while the overall ranges in carbon chain length and hydrocarbon classes were largely concordant, the abundances of the constituents in each hydrocarbon block varied. Significant differences were observed in most highly abundant hydrocarbon blocks (Figures 6D–E). In addition, the relative proportion of O_1-containing heteroatoms was also significantly different between production cycles (Figure 6F).

Similar analyses were performed for each product and the quantitation of the variability is presented in Figure 7. For the carbon chain length data (Figure 7A), product BO220 was most variable in terms of the number and range of molecules that were significantly different between production cycles. Even though product BO600 was equally complex in terms of the overall range of hydrocarbons, only C_{35}-containing molecules varied significantly

between production cycles. Products AR150 and AR220 showed variability in about one-third of the hydrocarbon blocks. Product BO60 showed no variability, and product BO100 showed variability in only a few blocks; however, there were only two samples available for those products and therefore limited variability should be interpreted with caution. Similar findings were observed with the hydrocarbon class data (Figure 7B). Most production cycle-associated variability was found in mono-, di- and tri-aromatic compounds. For the heteroatom data (Figure 7C), only products AR150 and BO220 showed variability with O_1-containing molecules, these were the most abundant and variable heteroatoms.

Even though the analysis of within-product variability based on the hydrocarbon block method has broad utility, such data are deemed insufficient in terms of satisfying the regulatory need for "*detailed chemical characterization*" of petroleum UVCBs for product registration purposes. Therefore, we also used the data on high abundance features to characterize the variability between samples of each product that were derived from independent production cycle. For this, we examined (i) the degree to which the individual features varied between samples from different production cycles (*i.e.*, p-values for each molecule), and (ii) the average relative abundance of each feature in a product across production cycles (Supplemental Table 10). Figure 8 shows the results of this analysis for each product. The figure shows that the feature abundance threshold set during data processing was concordant with the REACH Regulation threshold of 0.1% abundance for constituents of concern in complex substances [27]. It is evident that there are hundreds of constituents in each examined product that are present in quantities above 0.1%. These included expected abundant amounts (5–10%) of naphthalene and related mono- and di-aromatic hydrocarbons (Supplemental Table 11). However, few constituents were significantly different between samples (using a cutoff based on the Bonferroni-corrected p-value which was a false discovery rate of 5% corrected for the total number of features) of the same product from independent production cycles. We found no constituents that are both significantly different and reasonably abundant in products BO60, BO100 and BO600, even though these products spanned the degree of complexity of the entire dataset in terms of the number of high abundance features. Product AR200 had the largest number of constituents identified as variable and abundant. To the contrary of the results with a hydrocarbon block method data analysis (Figures 6–7), product BO220 only had 3 constituents above the variability and abundance thresholds, even though the total number of constituents with suggestive significance was large. Product AR150 had only one constituent above the thresholds. Table 2 lists the constituents for each product that were identified as above the thresholds in Figure 8. A number of the constituents identified in these analyses are currently listed by ECHA as Annex III substances, which are substances predicted to likely present health or environmental hazard [28]. One constituent in product AR200, anthracene, a feature whose identity was identified using IMS-MS data from a chemical standard, is identified by ECHA as a substance of very high concern.

4. Discussion

The analytical chemistry challenges in petroleomics are many [29, 30], and the potential solutions range from well-established physicochemical and analytical methods [3, 11] to novel HRMS techniques [12, 31–34]. However, there appears to be a growing chasm

between the research-driven advances in HRMS for petroleomics, and the needs of the practitioners in the industry and regulatory agencies. "*Sufficient*" characterization of highly complex petroleum UVCBs as products allowed into commerce and trade at various economic areas, such as the European Union where REACH defines data requirements [7], is a pressing regulatory need. As recently as 10 years ago, it was noted in a report by a major trade association of the petroleum refiners in Europe that conventional MS-derived "*data obtained by direct analysis of a petroleum UVCB substance, in which all constituents are ionized and fragmented simultaneously, would be too complex to allow meaningful interpretation*" [3]. Indeed, regulatory submissions of petroleum UVCBs do not typically use conventional MS-based data, or more contemporary HRMS data, but rather include information that defines chemical composition broadly, for example into hydrocarbon blocks [25, 26]. The typical substance identity information provided to the regulators such as ECHA consists of the manufacturing process description, various physicochemical data (boiling point and carbon chain length ranges, etc.), relative proportions of constituents in major hydrocarbon classes (saturates, aromatics, resins, asphaltenes, etc.), and relative content of various polycyclic aromatic compounds (by the number of aromatic rings). Invariably, regulators express dissatisfaction that the individual chemical constituents, their structural features, and quantitative metrics for the intrinsic variability of the products that are being registered are not attainable using the analytical methods on which the industry is relying heavily. For example, in a recent decision from ECHA on a testing proposal for grouping of substances in the "Residual aromatic extracts" manufacturing category, the agency concluded that chemical similarity between products has not been established for the purpose of registration (*i.e.*, read-across), because "*no qualitative or quantitative comparative assessment of the compositions of the different category members*" has been presented [9]. This representative ECHA decision further noted that because of the intrinsic compositional variability of petroleum UVCBs, detailed information in support of the "*chemical similarity*" argument would need to include (i) detailed data on the composition of the test sample(s) (both individual constituents and "major hydrocarbon classes"), as well as (ii) data on intrinsic chemical variability among products in a category [9], these requirements were recently added to Annex XI of REACH [10].

A number of recently developed multidimensional HRMS techniques, including IMS-MS, when applied to the analysis of petroleum samples, demonstrated excellent molecular resolution and the ability to characterize high molecular weight hydrocarbons, including isomeric species [16, 30, 32, 35–41]. In addition, a number of chemometric methods have been proposed in conjunction with HRMS data on oil and petroleum products for fingerprinting and source identification [12, 41–43], as well as the capability of the molecular and structural identity of chemical constituents and hydrocarbon blocks [17, 44, 45]. Therefore, we reasoned that the opportunity exists to demonstrate the value of multidimensional HRMS petroleomics as a solution to current challenges in chemical characterization of petroleum UVCBs for regulatory decision-making purpose. To this effect, this study aimed to demonstrate how one of the multidimensional HRMS petroleomics techniques, IMS-MS, can be used to address the regulatory challenges by (i) providing qualitative and quantitative information on the composition of representative complex petroleum substances, and (ii) using this information to characterize the variability

of the constituents in the substances manufactured in different production cycles or those grouped into the same broad category. Our choice of APPI ionization in positive mode with IMS-MS as an analytical technique was informed by prior studies demonstrating improved resolution of isomeric aromatic species in petroleum samples [17, 46]. Specifically, we aimed to take advantage of the IMS-MS technique-derived data on the differences in drift time among various hydrocarbons of the same atomic composition (*i.e.*, isomeric species), rather than focus on increasing the resolution in the *m/z* dimension, a common goal in petroleomics studies afforded by ultrahigh resolution Fourier transform ion cyclotron resonance (FTICR) MS [47] and other HRMS techniques [41]. Recent studies demonstrated the utility of IMS-MS for determining the chemical composition of petroleum substances and crude oils [13–16] and we proposed a chemometric method for deducing the chemical compositional information for both refined products and crude oils that uses $^{DT}CCS_{N2}$ information to increase confidence in the evaluation of the chemical composition of the features in homologous series [17]. Overall, we hypothesized that high resolution untargeted IMS-MS analysis, in conjunction with a petroleomics data processing workflow and chemometric evaluation, would enable detailed characterization of the most abundant ionizable molecules in petroleum UVCBs, providing quantitative data on substance-to-substance variation that will inform overall hazard assessment. To test this hypothesis, we evaluated both a range of petroleum products, and samples from independent production cycles of the same product.

Overall, we highlight four major advances afforded by this study. First, we demonstrate how IMS-MS data can be used to evaluate broad similarity among substances while also identifying the degree of variability within a class or between production batches of the same substance. By comparing and contrasting the IMS-MS data to that from GC-MS, we confirm advantages in both resolving power, and coverage of the high molecular weight compounds. GC-MS is used widely to characterize the composition of various fuels and to classify and group the fuels [43]. In addition, GC×GC-flame ionization detection (FID) technique [26, 48–50] is also commonly used for petroleum analyses to derive “hydrocarbon blocks” for substance identification purposes [25, 51]. It was previously shown that the multidimensional data from these techniques can be used for fingerprinting of oils or grouping petroleum UVCBs, but that IMS-MS data typically affords greater classification and fingerprinting accuracy [16, 42]. In this study, we found a similar pattern, with IMS-MS data superior to that from GC-MS for grouping and classification of the samples.

Second, this study goes farther than grouping and classification as we were able to assign confident molecular formulas to most (on average 93% across all samples) of the high abundance features from IMS-MS data. To achieve this, we selected only the highest quality abundant features at the expense of focusing on a relatively small fraction (~2%) of all detected features. While the process of dimensionality reduction may seem counter to the desire to provide as detailed chemical characterization of the UVCBs as possible, the following considerations support our approach: (i) the confidence in molecular formula assignments for the features beyond those with highest abundance erodes rapidly [17], (ii) even though IMS-MS is able to resolve tens of thousands of features in petroleum UVCBs, numerous molecules are still undetected either due to ion suppression or instrument sensitivity [32, 52], and (iii) if these chemical composition data are to be used for regulatory

decisions, it is acknowledged that the priority shall be given to the highest abundance constituents in complex substances. For example, according to Articles 7(2) and 33 of REACH Regulation [27], the abundance threshold of 0.1% (w/w) is set (for the purposes of either notification of substances in articles, or communication of information on substances in articles) for constituents that are classified as substances of very high concern. This implies that the focus on the highest abundance features when analyzing detailed chemical composition of petroleum UVCBs would be responsive to REACH Regulation requirements, because other molecules in each sample are likely present at amounts far below the 0.1% threshold.

Third, a very important consideration for the use of an analytical method for regulatory decision-making is its accessibility and reproducibility. Both GC-MS and GC×GC-FID are used to generate data for regulatory submissions because these methods have been standardized [53, 54]. In this regard, commercialization of the drift tube IMS-MS made these instruments available in a standard configuration leading to a growing number of publications demonstrating their use for petroleomics [16, 17, 32]. In addition, studies of reproducibility of IMS-MS-derived experimental parameters such as standardized drift tube, nitrogen CCS values ($^{DT}CCS_{N2}$) were conducted using hundreds of molecules across multiple laboratories and illustrated the potential of this technique for providing confident molecular identifiers for a broad range of discovery-based analyses [18, 20]. This study, while not a formal cross-laboratory standardization analysis, does demonstrate that samples can be confidently compared across operators in the same laboratory and across laboratories. Therefore, this technique and approach have promise for wider application as they are based on a commercially-available instrument and also a fairly rapid analysis based on gas phase separations and direct injection that does not require extensive sample preparation.

Finally, because of the ability to deduce molecular identifications for hundreds of molecules in complex petroleum UVCBs, a number of existing challenges with chemical characterization of petroleum UVCBs for hazard assessment are potentially resolved. Specifically, it is possible to identify constituents and determine their abundance for consideration as potential substances of concern. Because the hydrocarbon block method [51] is widely used for the characterization of human health and environmental hazards of petroleum UVCBs, the IMS-MS data with molecular identifiers can be used to construct data matrices similar to those generated in GC×GC-FID, but where the identity of the constituents in each block are known. In addition, the variability between independent production cycles and among samples in the same product category can be quantitatively characterized; if it is found that samples are significantly variable, it is now possible to determine whether such variability may impact potential hazardous properties of the entire substance and reduce uncertainty in grouping.

One limitation of this study, similar to other analytical studies of petroleum UVCBs, is that the complete chemical characterization of petroleum UVCBs is unattainable. The extent of the molecular resolution depends on the type of ionization and detection methods and instruments, as well as sample processing and other factors [55]. For example, the APPI ionization used herein, albeit a preferred method for characterization of nonpolar petroleum fractions [56, 57], is not applicable to the analysis of paraffins. Still, our method is suitable

for evaluation of polycyclic aromatic compounds, which include polycyclic aromatic hydrocarbons and heteroatoms, substances that have been associated with carcinogenic activity [58, 59]. We also note that other HRMS methods can be used for characterization of chemical composition of petroleum UVCBs [12, 60]. In this regard, by coupling HRMS with additional separation techniques, such as GC-APCI [61] or ion mobility [15, 62], additional characterization of isomers can be achieved. It is important to distinguish and characterize structural isomers in petroleum UVCBs to understand potential variability in the manufacturing process chemistry and the effects of different oil feed stocks [63].

5. Conclusion

This study evaluated samples of 6 petroleum products (heavy aromatic, hydrotreated light paraffinic, and hydrotreated heavy paraffinic) from 2–3 production cycles using GC-MS and APPI-IMS-MS. The resulting data were used for classification and grouping using several unsupervised algorithms as either untargeted data, or after structurally related compounds in each sample were identified with confidence using multidimensional data analysis workflow. Between- and within-substance variability was quantified and the types of hydrocarbon blocks, and individual molecules, that were variable in samples of different production cycles were identified. Sample analysis was conducted in different laboratories to examine reproducibility of the grouping and classifications. Overall, these data show that IMS-MS can be used to provide chemical compositional data on petroleum UVCBs, information that is needed to characterize the variability in substances from different production cycles. Such chemical characterization can be used to support hazard evaluations and address the regulatory need for qualitative and quantitative comparative assessment of the chemical composition of petroleum UVCBs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This study was supported, in part, by grants from the National Institutes of Health (P30 ES029067 and P42 ES027704) and the National Academies Gulf Research Program (2000008942). A.T. Roman-Hubers and A. C. Cordova were supported, in part, by a training grant from the National Institutes of Health (T32 ES0226568). The views expressed in this manuscript are those of the authors and not of the funding agencies or their academic institutions. The authors also wish to acknowledge gracious assistance from Mr. Michael Smith for donating the samples used in these studies.

References

[1]. Kaiser MJ, A review of refinery complexity applications, Pet Sci, 14 (2017) 167–194.

[2]. Salvito D, Fernandez M, Jenner K, Lyon DY, de Knecht J, Mayer P, MacLeod M, Eisenreich K, Leonards P, Cesnaitis R, Leon-Paumen M, Embry M, Deglin SE, Improving the Environmental Risk Assessment of Substances of Unknown or Variable Composition, Complex Reaction Products, or Biological Materials, Environ Toxicol Chem, 39 (2020) 2097–2108. [PubMed: 32780492]

[3]. CONCAWE, REACH – Analytical characterisation of petroleum UVCB substances, in, Brussels, Belgium, 2012.

[4]. 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, Regulatory toxicology and pharmacology : RTP, 67 (2013) 409–420. [PubMed: 24025648]

[5]. ECHA, Guidance for identification and naming of substances under REACH and CLP, in, European Chemical Agency, Helsinki, Finland, 2017.

[6]. CONCAWE, Hazard Classification and Labelling of Petroleum substances in the European Economic Area – 2020, in, Brussles, Belgium, 2020.

[7]. Williams ES, Panko J, Paustenbach DJ, The European Union's REACH regulation: a review of its history and requirements, Critical reviews in toxicology, 39 (2009) 553–575. [PubMed: 19650717]

[8]. CONCAWE, REACH roadmap for Petroleum Substances, in, Brussels, Belgium, 2019.

[9]. ECHA, Testing Proposal Decision on Substance EC 295-332-8 "Extracts (petroleum), deasphalted vacuum residue solvent", in, European Chemicals Agency, Helsinki, Finland, 2020.

[10]. European Commission, COMMISSION REGULATION (EU) 2021/979 of 17 June 2021 amending Annexes VII to XI to Regulation (EC) No 1907/2006 of the European Parliament and of the Council concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), in, Official Journal of the European Union, Brussels, Belgium, 2021.

[11]. CONCAWE, Concawe Substance Identification Group Analytical Program Report (Abridged Version), in, Brussels, Belgium, 2019.

[12]. Palacio Lozano DC, Thomas MJ, Jones HE, Barrow MP, Petroleomics: Tools, Challenges, and Developments, Annu Rev Anal Chem (Palo Alto Calif), 13 (2020) 405–430. [PubMed: 32197051]

[13]. 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, 88 (2016) 12152–12160. [PubMed: 28193022]

[14]. Farenc M, Corilo YE, Lalli PM, Riches E, Rodgers RP, Afonso C, Giusti P, Comparison of Atmospheric Pressure Ionization for the Analysis of Heavy Petroleum Fractions with Ion Mobility-Mass Spectrometry, Energ Fuel, 30 (2016) 8896–8903.

[15]. Ruger CP, Le Maitre J, Maillard J, Riches E, Palmer M, Afonso C, Giusti P, Exploring Complex Mixtures by Cyclic Ion Mobility High-Resolution Mass Spectrometry: Application Toward Petroleum, Anal Chem, 93 (2021) 5872–5881. [PubMed: 33784070]

[16]. Roman-Hubers AT, McDonald TJ, Baker ES, Chiu WA, Rusyn I, A comparative analysis of analytical techniques for rapid oil spill identification, Environ Toxicol Chem, 40 (2021) 1034–1049. [PubMed: 33315271]

[17]. Roman-Hubers AT, Cordova AC, Aly NA, McDonald TJ, Lloyd DT, Wright FA, Baker ES, Chiu WA, Rusyn I, Data Processing Workflow to Identify Structurally Related Compounds in Petroleum Substances Using Ion Mobility Spectrometry-Mass Spectrometry, Energ Fuel, 35 (2021) 10529–10539.

[18]. Stow SM, Causon TJ, Zheng X, Kurulugama RT, Mairinger T, May JC, Rennie EE, Baker ES, Smith RD, McLean JA, Hann S, Fjeldsted JC, An Interlaboratory Evaluation of Drift Tube Ion Mobility-Mass Spectrometry Collision Cross Section Measurements, Anal Chem, 89 (2017) 9048–9055. [PubMed: 28763190]

[19]. McKee RH, Medeiros AM, Daughtrey WC, A proposed methodology for setting occupational exposure limits for hydrocarbon solvents, J Occup Environ Hyg, 2 (2005) 524–542. [PubMed: 16174635]

[20]. Baker ES, Collision Cross Section database, in, 2021.

[21]. McLafferty FW, F. T, Interpretation of Mass Spectra, 4th ed., University Science Books, Mill Valley, Californa, 1993.

[22]. Everitt B, Cluster analysis, Qual Quant, 14 (1980) 75–100.

[23]. Fowlkes EB, Mallows CL, A Method for Comparing Two Hierarchical Clusterings, J Am Stat Assoc, 78 (1983) 553–569.

[24]. Dunn OJ, Multiple comparison among means, J Am Stat Assoc, 52–64.

[25]. CONCAWE, Environmental Risk Assessment of Petroleum Substances: The Hydrocarbon Block Method, in, Brussels, Belgium, 1996.

[26]. CONCAWE, Investigating the HCBM – GCxGC relationship: an elution model to interpret GCxGC retention times of petroleum substances, in, Brussels, Belgium, 2019.

[27]. ECHA, Guidance on requirements for substances in articles, in, European Chemicals Agency, Helsinki, Finland, 2017.

[28]. ECHA, Preparation of an inventory of substances suspected to meet REACH Annex III criteria, in, European Chemicals Agency, Helsinki, Finland, 2016.

[29]. Marshall AG, Rodgers RP, Petroleomics: chemistry of the underworld, Proc Natl Acad Sci U S A, 105 (2008) 18090–18095. [PubMed: 18836082]

[30]. Marshall AG, Rodgers RP, Petroleomics: the next grand challenge for chemical analysis, Acc Chem Res, 37 (2004) 53–59. [PubMed: 14730994]

[31]. 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, 34 (2015) 248–263. [PubMed: 24942384]

[32]. 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, 7 (2015) 4450–4463.

[33]. Terra LA, Filgueiras PR, Tose LV, Romao W, de Souza DD, de Castro EV, de Oliveira MS, Dias JC, Poppi RJ, Petroleomics by electrospray ionization FT-ICR mass spectrometry coupled to partial least squares with variable selection methods: prediction of the total acid number of crude oils, Analyst, 139 (2014) 4908–4916. [PubMed: 25068148]

[34]. Islam A, Cho Y, Ahmed A, Kim S, Data Interpretation Methods for Petroleomics, Mass Spectrometry Letters, 3 (2012) 63–67.

[35]. Hsu CS, Hendrickson CL, Rodgers RP, McKenna AM, Marshall AG, Petroleomics: advanced molecular probe for petroleum heavy ends, J Mass Spectrom, 46 (2011) 337–343. [PubMed: 21438082]

[36]. Guillemant J, Albrieux F, Lacoue-Negre M, Pereira de Oliveira L, Joly JF, Duponchel L, Chemometric Exploration of APPI(+)-FT-ICR MS Data Sets for a Comprehensive Study of Aromatic Sulfur Compounds in Gas Oils, Anal Chem, 91 (2019) 11785–11793. [PubMed: 31441637]

[37]. Palacio Lozano DC, Gavard R, Arenas-Diaz JP, Thomas MJ, Stranz DD, Mejia-Ospino E, Guzman A, Spencer SEF, Rossell D, Barrow MP, Pushing the analytical limits: new insights into complex mixtures using mass spectra segments of constant ultrahigh resolving power, Chem Sci, 10 (2019) 6966–6978. [PubMed: 31588263]

[38]. Grimm FA, Russell WK, Luo YS, Iwata Y, Chiu WA, Roy T, Boogaard PJ, Ketelslegers HB, Rusyn I, Grouping of Petroleum Substances as Example UVCBs by Ion Mobility-Mass Spectrometry to Enable Chemical Composition-Based Read-Across, Environmental science & technology, 51 (2017) 7197–7207. [PubMed: 28502166]

[39]. Fernandez-Lima FA, Becker C, McKenna AM, Rodgers RP, Marshall AG, Russell DH, Petroleum crude oil characterization by IMS-MS and FTICR MS, Analytical chemistry, 81 (2009) 9941–9947. [PubMed: 19904990]

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

[41]. Niyonsaba E, Manheim JM, Yerabolu R, Kenttamaa HI, Recent Advances in Petroleum Analysis by Mass Spectrometry, Anal Chem, 91 (2019) 156–177. [PubMed: 30428670]

[42]. Onel M, Beykal B, Ferguson K, Chiu WA, McDonald TJ, Zhou L, House JS, Wright FA, Sheen DA, Rusyn I, Pistikopoulos EN, Grouping of complex substances using analytical chemistry data: A framework for quantitative evaluation and visualization, PloS one, 14 (2019) e0223517. [PubMed: 31600275]

[43]. de Carvalho Rocha WF, Schantz MM, Sheen DA, Chu PM, Lippa KA, Unsupervised classification of petroleum Certified Reference Materials and other fuels by chemometric

analysis of gas chromatography-mass spectrometry data, Fuel (Lond), 197 (2017) 248–258. [PubMed: 28603295]

[44]. Gabelica V, Shvartsburg AA, Afonso C, Barran P, Benesch JLP, Bleiholder C, Bowers MT, Bilbao A, Bush MF, Campbell JL, Campuzano IDG, Causon T, Clowers BH, Creaser CS, De Pauw E, Far J, Fernandez-Lima F, Fjeldsted JC, Giles K, Groessl M, Hogan CJ Jr., Hann S, Kim HI, Kurulugama RT, May JC, McLean JA, Pagel K, Richardson K, Ridgeway ME, Rosu F, Sobott F, Thalassinos K, Valentine SJ, Wyttenbach T, Recommendations for reporting ion mobility Mass Spectrometry measurements, Mass Spectrom Rev, 38 (2019) 291–320. [PubMed: 30707468]

[45]. Koch BP, Dittmar T, Witt M, Kattner G, Fundamentals of molecular formula assignment to ultrahigh resolution mass data of natural organic matter, Anal Chem, 79 (2007) 1758–1763. [PubMed: 17297983]

[46]. Borsdorf H, Nazarov EG, Miller RA, Atmospheric-pressure ionization studies and field dependence of ion mobilities of isomeric hydrocarbons using a miniature differential mobility spectrometer, Anal Chim Acta, 575 (2006) 76–88. [PubMed: 17723575]

[47]. Marshall AG, Blakney GT, Beu SC, Hendrickson CL, McKenna AM, Purcell JM, Rodgers RP, Xian F, Petroleomics: a test bed for ultra-high-resolution Fourier transform ion cyclotron resonance mass spectrometry, Eur J Mass Spectrom (Chichester), 16 (2010) 367–371. [PubMed: 20530823]

[48]. Frysinger GS, Gaines RB, Xu L, Reddy CM, Resolving the unresolved complex mixture in petroleum-contaminated sediments, Environmental science & technology, 37 (2003) 1653–1662. [PubMed: 12731850]

[49]. Gaines RB, Frysinger GS, Hendrick-Smith MS, Stuart JD, Oil spill source identification by comprehensive two-dimensional gas chromatography, Environmental science & technology, 33 (1999) 2106–2112.

[50]. Van De Weghe H, Vanermen G, Gemoets J, Lookman R, Bertels D, Application of comprehensive two-dimensional gas chromatography for the assessment of oil contaminated soils, J Chromatogr A, 1137 (2006) 91–100. [PubMed: 17055525]

[51]. Bierkens J, Geerts L, Environmental hazard and risk characterisation of petroleum substances: a guided “walking tour” of petroleum hydrocarbons, Environ Int, 66 (2014) 182–193. [PubMed: 24607926]

[52]. Hawkes JA, D’Andrilli J, Agar JN, Barrow MP, Berg SM, Catalán N, Chen H, Chu RK, Cole RB, Dittmar T, Gavard R, Gleixner G, Hatcher PG, He C, Hess NJ, Hutchins RHS, Ijaz A, Jones HE, Kew W, Khaksari M, Palacio Lozano DC, Lv J, Mazzoleni LR, Noriega-Ortega BE, Osterholz H, Radoman N, Remucal CK, Schmitt ND, Schum SK, Shi Q, Simon C, Singer G, Sleighter RL, Stubbins A, Thomas MJ, Tolic N, Zhang S, Zito P, Podgorski DC, An international laboratory comparison of dissolved organic matter composition by high resolution mass spectrometry: Are we getting the same answer?, Limnol Oceanogr Methods, 18 (2020) 235–258.

[53]. US EPA, Method 8270E (SW-846): Semivolatile Organic Compounds by Gas Chromatography/ Mass Spectrometry (GC/MS), in, US Environmental Protection Agency, Washington, DC, 2014.

[54]. ASTM International, UOP Method 990–11: Organic Analysis of Distillate by Comprehensive Two-Dimensional Gas Chromatography with Flame Ionization Detection, in, ASTM INternational, West Conshohocken, PA, 2011.

[55]. Palacio Lozano DC, Chacon-Patiño ML, Gomez-Escudero A, Barrow MP, Chapter 32 | Mass Spectrometry in the Petroleum Industry,” in Fuels and Lubricants Handbook: Technology, Properties, Performance, and Testing, MNL37–2ND-EB Fuels and Lubricants Handbook: Technology, Properties, Performance, and Testing, 2 (2019).

[56]. Kauppila TJ, Kuuranne T, Meurer EC, Eberlin MN, Kotiaho T, Kostiainen R, Atmospheric pressure photoionization mass spectrometry. Ionization mechanism and the effect of solvent on the ionization of naphthalenes, Anal Chem, 74 (2002) 5470–5479. [PubMed: 12433075]

[57]. Purcell JM, Hendrickson CL, Rodgers RP, Marshall AG, Atmospheric pressure photoionization proton transfer for complex organic mixtures investigated by fourier transform ion cyclotron resonance mass spectrometry, J Am Soc Mass Spectrom, 18 (2007) 1682–1689. [PubMed: 17689097]

[58]. Ayala-Cabrera JF, Galceran MT, Moyano E, Santos FJ, Chloride-attachment atmospheric pressure photoionisation for the determination of short-chain chlorinated paraffins by gas

chromatography-high-resolution mass spectrometry, Anal Chim Acta, 1172 (2021) 338673. [PubMed: 34119025]

[59]. McKee RH, White R, The mammalian toxicological hazards of petroleum-derived substances: an overview of the petroleum industry response to the high production volume challenge program, Int J Toxicol, 33 (2014) 4S–16S. [PubMed: 24351873]

[60]. Rodgers RP, McKenna AM, Petroleum analysis, Anal Chem, 83 (2011) 4665–4687. [PubMed: 21528862]

[61]. Barrow MP, Peru KM, Headley JV, An added dimension: GC atmospheric pressure chemical ionization FTICR MS and the Athabasca oil sands, Anal Chem, 86 (2014) 8281–8288. [PubMed: 25036898]

[62]. Maillard JF, Le Maitre J, Ruger CP, Ridgeway M, Thompson CJ, Paupy B, Hubert-Roux M, Park M, Afonso C, Giusti P, Structural analysis of petroporphyrins from asphaltene by trapped ion mobility coupled with Fourier transform ion cyclotron resonance mass spectrometry, Analyst, 146 (2021) 4161–4171. [PubMed: 34047731]

[63]. Lalli PM, Corilo YE, Rowland SJ, Marshall AG, Rodgers RP, Isomeric Separation and Structural Characterization of Acids in Petroleum by Ion Mobility Mass Spectrometry, Energ Fuel, 29 (2015) 3626–3633.

Highlights

- Registration of petroleum products requires detailed compositional information
- Traditional analytical methods are insufficient for such detailed characterization
- IMS-MS can provide detailed chemical compositional data on petroleum products
- This study used IMS-MS to characterize compositional variability in petroleum products
- Detailed compositional characterization from IMS-MS can support hazard evaluations

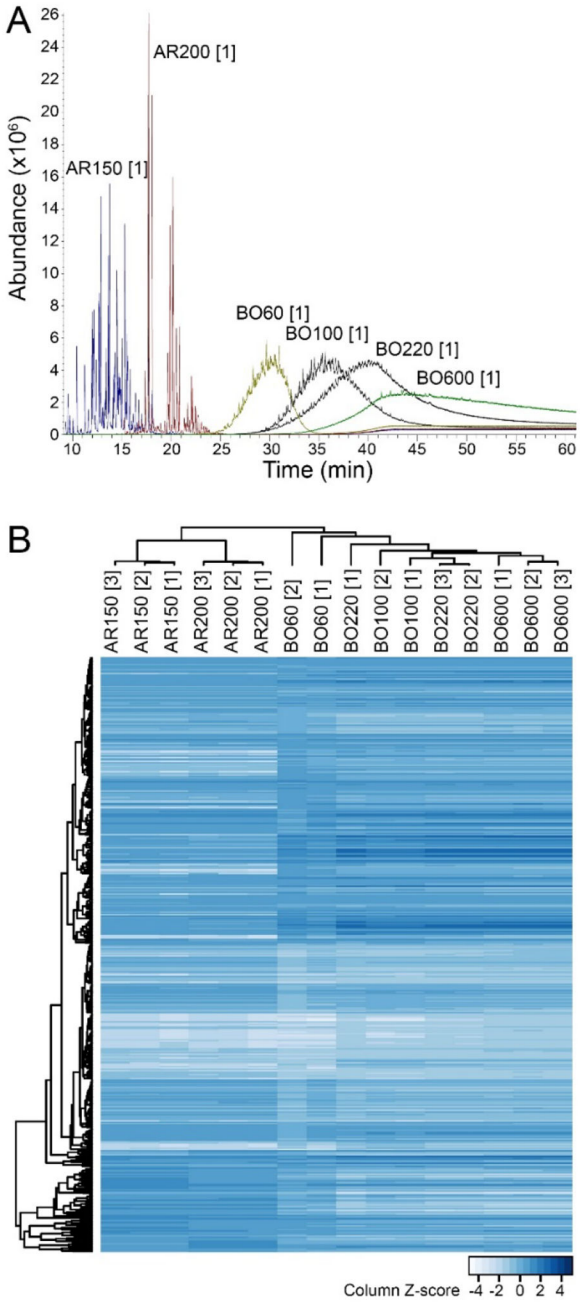

Figure 1. GC-MS full scan analysis of petroleum UVCB products included in this study.

(A) Superimposed GC-MS total ion chromatograms (time vs. abundance) for representative samples (see Table 1 for sample annotations). Individual chromatograms for each sample are shown in Supplemental Figure 1. **(B)** Hierarchical clustering analysis of the average abundance of the detected compound ion fragments in a mass range of 40–500 amu in 10,127 scans (see Supplemental Table 2 for the raw data). Both samples (columns) and features (rows) were clustered (Spearman correlation, average linkage method). Feature abundance was z-scaled for each sample with lower abundance features indicated by light blue and higher abundance features indicated by dark blue colors.

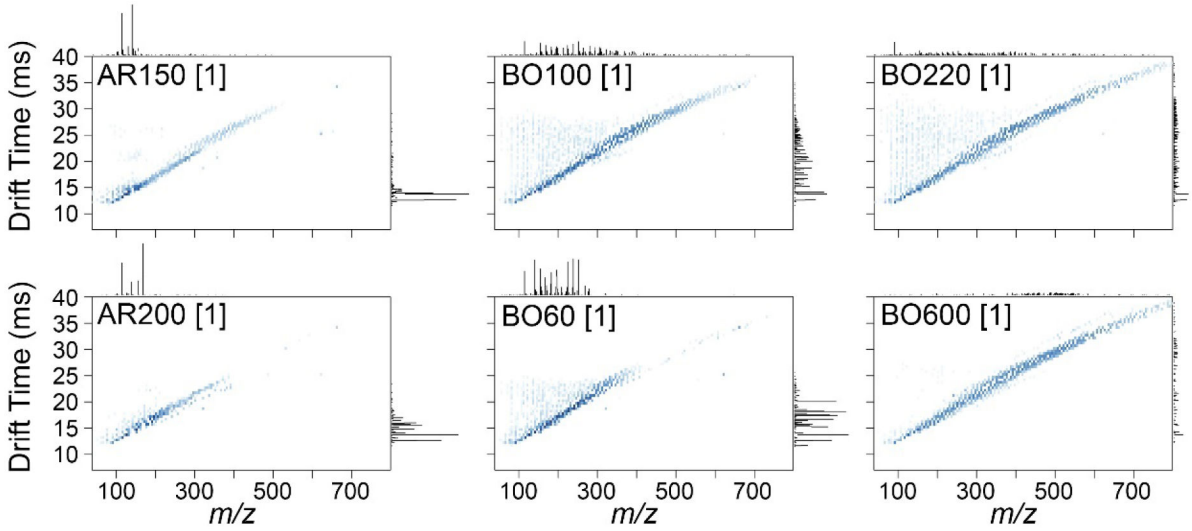

Figure 2. Representative nested IMS-MS spectra (APPI+ ion mode) for petroleum UVCB products included in this study.

Representative samples (see Table 1 for sample annotations) are shown, data for other samples are shown in Supplemental Figure 2. Individual features are shown as dots in the 2D scatterplot where x-axes are *m/z*, y-axes are drift time, and feature intensities are indicated by the color intensity. The density histograms of the features are shown at the top (for *m/z*) or on the right (for drift time) of each plot.

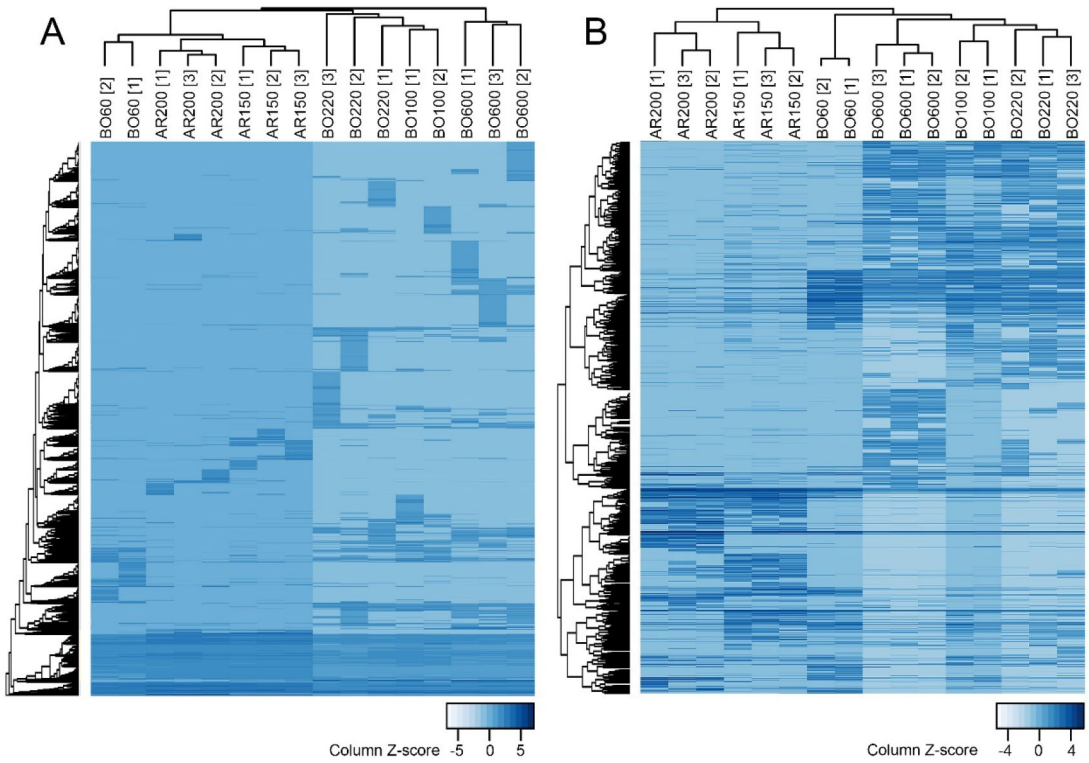

Figure 3. Unsupervised hierarchical clustering of petroleum UVCB products using IMS-MS data.

Shown are heatmaps (illustrating relative feature abundance) that were products of hierarchical clustering analysis (Spearman correlation, average linkage method) for 16 samples (see Table 1 for sample annotations) analyzed in one of the experimental runs. Technical replicates of each sample were averaged for each feature. **(A)** Full dataset (Supplemental Table 3A; 55,466 features). **(B)** Filtered dataset (Supplemental Table 4A; 1,530 features).

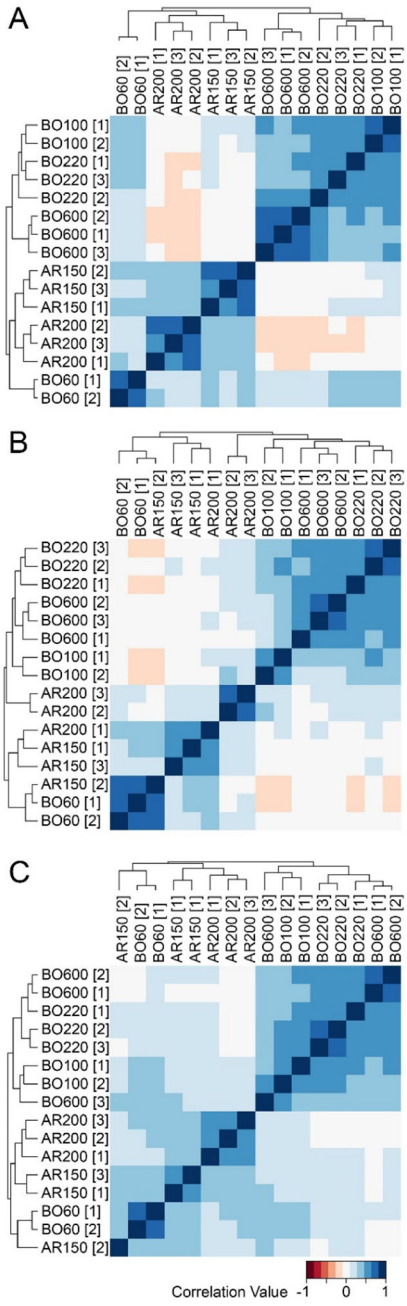

Figure 4. Inter- and intra-laboratory reproducibility of grouping petroleum UVCB products using untargeted IMS-MS analyses conducted in independent experiments.

The samples were analyzed using an identical experimental protocol either at Texas A&M on the same instrument but by a different operator (**A** and **B**) or at NC State University by another operator and instrument, but the same model of IMS-MS platform (**C**). Correlation values are listed in Supplemental Table 7 and shown using a color gradient as indicated in the legend at the bottom of the figure.

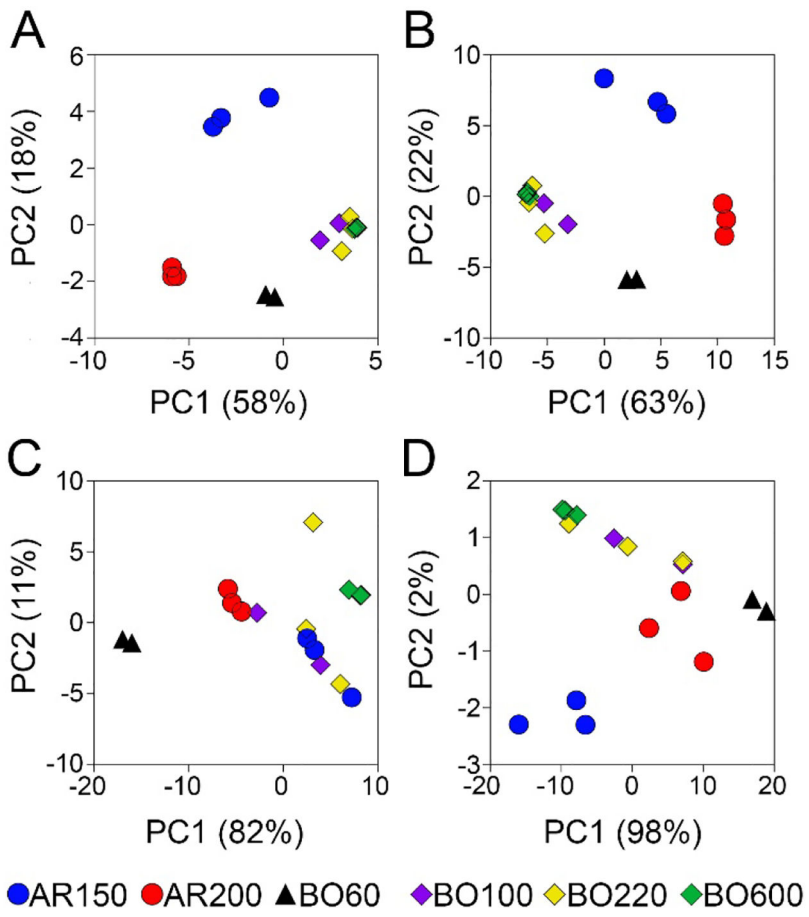

Figure 5. The Principal Component Analysis grouping of petroleum UVCB products.

(A) Grouping based on the relative abundance of all features with assigned molecular formulas (Supplemental Table 5). **(B)** Grouping based on the carbon chain length distribution (Supplemental Table 8). **(C)** Grouping based on the hydrocarbon class (Supplemental Table 8). **(D)** Grouping based on the heteroatom profile (Supplemental Table 9). Colors represent individual samples of the same product as indicated in the legend at the bottom of the figure.

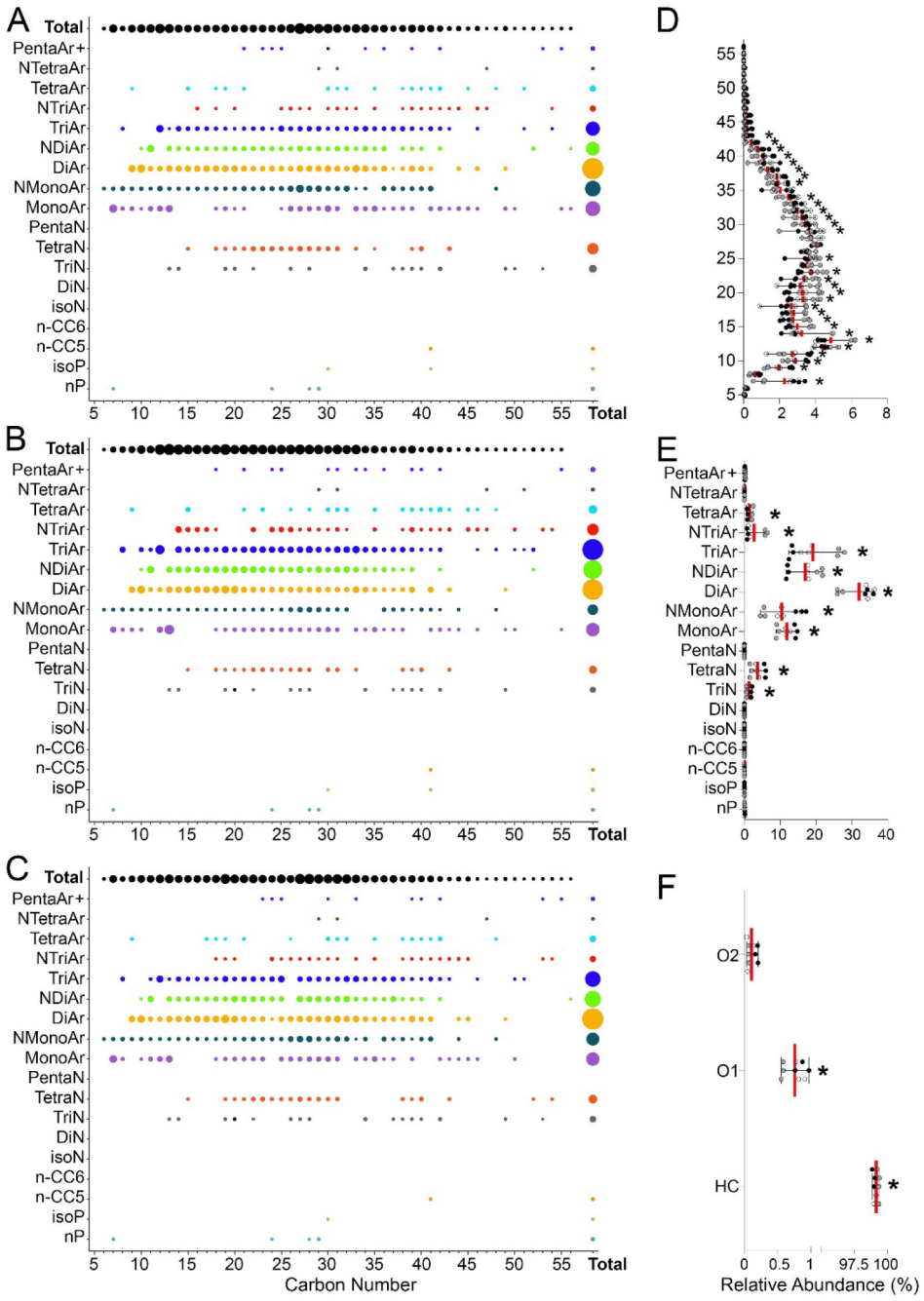

Figure 6. Hydrocarbon block matrix for samples from independent manufacturing cycles of product BO220.

(A–C) Dot plots representing the relative abundance (each sample is scaled to 100%) of the constituents in different hydrocarbon blocks (hydrocarbon class vs carbon chain length) in three independent samples (see Supplemental Tables 8–9 for data on each product). **(D–F)** Relative abundance distribution for the carbon chain length **(D)**, hydrocarbon class **(E)** and heteroatom content **(F)** where symbols represent individual technical replicates (same color) of the samples from independent manufacturing cycles (shades of gray). Red vertical lines are mean and whiskers are min-max range. Asterisks (*) denote blocks with statistically

significant (p_{adj}-value <0.05, Supplemental Table 10) variability among samples of product BO220 from independent manufacturing cycles.

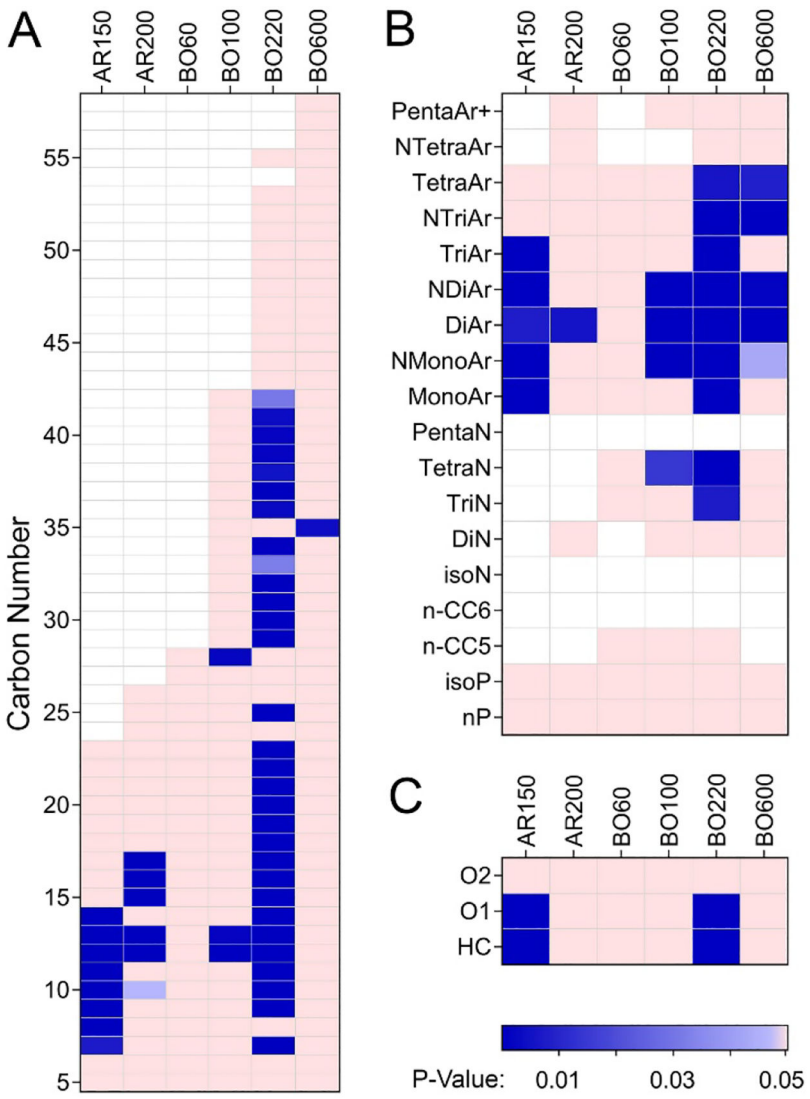

Figure 7. Variability in hydrocarbon blocks (A–B) and heteroatom content (C) for independent manufacturing cycles of petroleum UVCB products.

Heatmaps show whether relative abundance of the constituents in different hydrocarbon blocks or heteroatom classes were significantly variable (p_{adj}<0.05, see Supplemental Table 9) among samples from independent manufacturing cycles. Colors represent significance (see legend at the bottom of the figure, white indicates that there were no constituents in that hydrocarbon block).

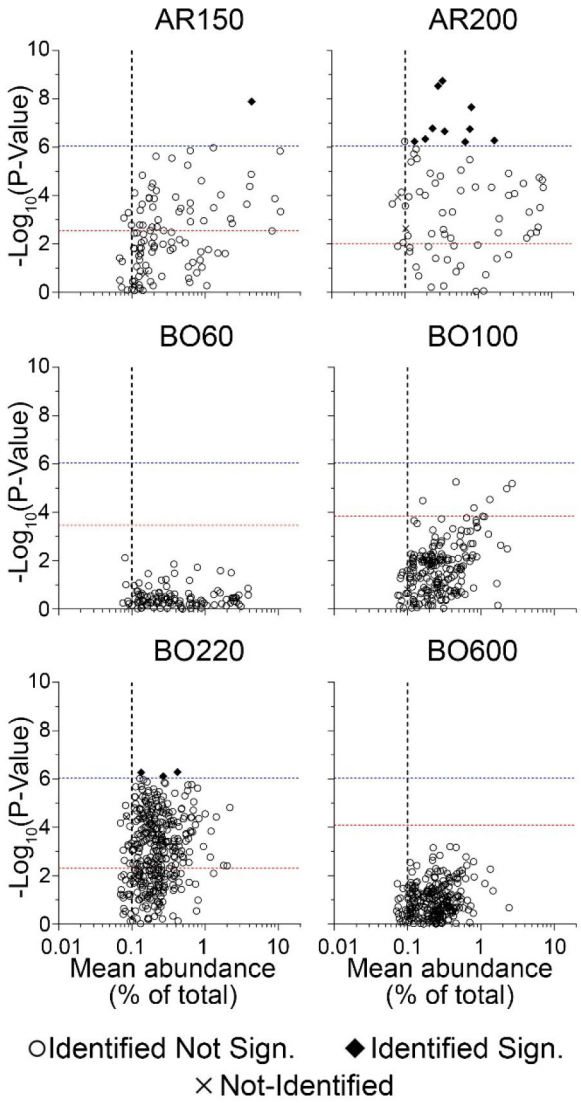

Figure 8. Identification of the individual features that are both abundant and significantly variable among samples from independent manufacturing cycles of each petroleum UVCB product.

The scatted plots show features that were present in each product based on their relative abundance (x-axis) and significance in variability (y-axis, p-values were converted to $-\text{Log}_{10}$ values). Vertical dotted lines indicate the 0.1% relative abundance threshold. Horizontal lines indicate product-specific (red dotted line corresponding to the p-value at false discovery rate of 5%) and global (across all samples, $-\text{Log}_{10}$(p-value) = 6.05, blue dotted lines) thresholds for multiple-corrected significance values. Black diamonds indicate features that were exceeding both global variability significance and abundance thresholds (see Table 2 for the complete list). Open circles (features with molecular formulae assigned) and "x" symbols (no molecular formulae assigned) indicate features that were not significant based on the global variability significance threshold.

Table 1.

Petroleum refining products used in this study. Samples of the same product (identified by a sample ID) are numbered consecutively based on their date of collection. See Supplemental Table 1 for additional information.

Sample ID	CAS #	Name	Substance Definition
AR150 [1] AR150 [2] AR150 [3]	64742-94-5	Solvent naphtha (petroleum), heavy aromatic	A complex combination of hydrocarbons obtained from distillation of aromatic streams. It consists predominantly of aromatic hydrocarbons having carbon numbers predominantly in the range of C_9 through C_{16} and boiling in the range of approximately 165°C to 290°C (330°F to 554°F).
AR200 [1] AR200 [2] AR200 [3]			
BO60 [1] BO60 [2]	64742-55-8	Distillates (petroleum), hydrotreated light paraffinic	A complex combination of hydrocarbons obtained by treating a petroleum fraction with hydrogen in the presence of a catalyst. It consists of hydrocarbons having carbon numbers predominantly in the range of C_{15} through C_{30} and produces a finished oil with a viscosity of less than 100 SUS at 100°F (19cSt at 40°C). It contains a relatively large proportion of saturated hydrocarbons.
BO100 [1] BO100 [2]	64742-54-7	Distillates (petroleum), hydrotreated heavy paraffinic	A complex combination of hydrocarbons obtained by treating a petroleum fraction with hydrogen in the presence of a catalyst. It consists of hydrocarbons having carbon numbers predominantly in the range of C_{20} through C_{50} and produces a finished oil of at least 100 SUS at 100°F (19cSt at 40°C). It contains a relatively large proportion of saturated hydrocarbons.
BO220 [1] BO220 [2] BO220 [3]			
BO600 [1] BO600 [2] BO600 [3]			

Table 2.

A list of features that exceeded the thresholds for both abundance of 0.1% and significance (multiple testing-corrected p-value) in three tested products. See Figure 8 for additional details.

Product Name	Feature ID*	Relative abundance, % total (mean±SD)	Fold Difference**	$-Log_{10}$ (p-value)***	Inferred formula#	Hydrocarbon class	Putative feature identity†	REACH indication of concern¶
AR150	8	4.3±2.6	3.01	7.89	$C_{12}H_8$	TriAr	Acenaphthylene	Annex III substances
AR200	37	1.6 ± 1.4	10.6	6.28	$C_{15}H_{11}$	TriAr	Methyphenantrene or methylanthracene	Annex III substances
	121	0.80 ± 0.22	26.5	7.66	$C_{17}H_{14}$	NTriAr	Cyclopenteno-phenanthrene	-
	26	0.77 ± 0.39	2.97	6.75	$C_{14}H_{10}$	TriAr	Anthracene	PBT, SVHC
	90	0.66 ± 0.14	1.51	6.22	$C_{12}H_6$	MonoAr	Triethynylbenzene	-
	77	0.35± 0.074	7.94	6.66	$C_{16}H_{18}$#	DiAr	Diphenylbutane	-
	341	0.32 ± 0.10	10.6	8.75	$C_{16}H_{16}$#	TriAr	Propylfluorene	-
	340	0.28 ± 0.080	5.03	8.53	$C_{15}H_{14}$#	TriAr	Ethylfluorene	-
	73	0.24 ± 0.63	11.6	6.78	$C_{17}H_{16}$	TriAr	Trimethylphenanthrene	Annex III substances
	475	0.19 ± 0.049	12	6.34	$C_{17}H_{18}$#	NDiAr	Benzyl-tetrahydronaphthalene, or ethyl-methyl-dihydroanthracene	-
	501	0.13 ± 0.033	1.56	6.23	$C_{13}H_{12}$#	DiAr	Methyl-phenylbenzene	Annex III substance
BO220	334	0.42 ± 0.095	1.57	6.29	$C_{17}H_{22}$	DiAr	Heptylnaphthalene	-
	32	0.27 ± 0.086	1.65	6.12	$C_{31}H_{50}$	DiAr	Henicosanylnaphthalene	-
	213	0.13 ± 0.0012	12.3	6.27	$C_{14}H_8$#	DiAr	Diethynylnaphthalene	-

*, See Supplemental Table 5A for additional information on each feature.

**, The maximum of the absolute value (if at least 3 samples were available) of the fold difference when comparing across all pairs of samples from independent production cycles.

***, The minimum p-value (converted to a $-Log_{10}$) from unequal variance t-test for comparing the differences in abundance of a feature between samples from independent production cycles of a product.

#, Table 5A lists these features in their radical form (−H)

†, Putative identification based on the data analysis workflow as detailed in Methods, or based on a match to a library standard (*i.e.*, anthracene).

¶The indications of concern for each putatively identified molecule based on the information in ECHA database (https://echa.europa.eu/). PBT, persistent, bioaccumulative, or toxic; SVHC, substance of very high concern; -, no information was included in the database as of 10/2021.