RESEARCH ARTICLE

Grouping of complex substances using analytical chemistry data: A framework for quantitative evaluation and visualization

Melis Onel[1,2], Burcu Beykal[1,2], Kyle Ferguson[3], Weihsueh A. Chiu[3], Thomas J. McDonald[4], Lan Zhou[5], John S. House[6], Fred A. Wright[6,7], David A. Sheen[8], Ivan Rusyn[3☯*], Efstratios N. Pistikopoulos[1,2☯*]

1 Artie McFerrin Department of Chemical Engineering, Texas A&M University, College Station, TX, United States of America, **2** Texas A&M Energy Institute, Texas A&M University, College Station, TX, United States of America, **3** Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX, United States of America, **4** Department of Environmental and Occupational Health, Texas A&M University, College Station, TX, United States of America, **5** Department of Statistics, Texas A&M University, College Station, TX, United States of America, **6** Bioinformatics Research Center, North Carolina State University, Raleigh, NC, United States of America, **7** Departments of Statistics and Biological Sciences, North Carolina State University, Raleigh, NC, United States of America, **8** Chemical Sciences Division, National Institute of Standards and Technology, Gaithersburg, MD, United States of America

☯ These authors contributed equally to this work.
* irusyn@cvm.tamu.edu (IR); stratos@tamu.edu (ENP)

Abstract

A detailed characterization of the chemical composition of complex substances, such as products of petroleum refining and environmental mixtures, is greatly needed in exposure assessment and manufacturing. The inherent complexity and variability in the composition of complex substances obfuscate the choices for their detailed analytical characterization. Yet, in lieu of exact chemical composition of complex substances, evaluation of the degree of similarity is a sensible path toward decision-making in environmental health regulations. Grouping of similar complex substances is a challenge that can be addressed via advanced analytical methods and streamlined data analysis and visualization techniques. Here, we propose a framework with unsupervised and supervised analyses to optimally group complex substances based on their analytical features. We test two data sets of complex oil-derived substances. The first data set is from gas chromatography-mass spectrometry (GC-MS) analysis of 20 Standard Reference Materials representing crude oils and oil refining products. The second data set consists of 15 samples of various gas oils analyzed using three analytical techniques: GC-MS, GC×GC-flame ionization detection (FID), and ion mobility spectrometry-mass spectrometry (IM-MS). We use hierarchical clustering using Pearson correlation as a similarity metric for the unsupervised analysis and build classification models using the Random Forest algorithm for the supervised analysis. We present a quantitative comparative assessment of clustering results via Fowlkes–Mallows index, and classification results via model accuracies in predicting the group of an unknown complex substance. We demonstrate the effect of (i) different grouping methodologies, (ii) data set size, and (iii) dimensionality reduction on the grouping quality, and (iv) different analytical techniques on the characterization of the complex substances. While the complexity and

OPEN ACCESS

Citation: Onel M, Beykal B, Ferguson K, Chiu WA, McDonald TJ, Zhou L, et al. (2019) Grouping of complex substances using analytical chemistry data: A framework for quantitative evaluation and visualization. PLoS ONE 14(10): e0223517. https://doi.org/10.1371/journal.pone.0223517

Editor: Jie Zhang, Newcastle University, UNITED KINGDOM

Received: May 9, 2019

Accepted: September 23, 2019

Published: October 10, 2019

Copyright: This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the Creative Commons CC0 public domain dedication.

Data Availability Statement: The R Markdown documentation of unsupervised and supervised analysis workflow can be accessed through http://parametric.tamu.edu/research/Onel_etAl_2019_Rmarkdown.html The GC-MS, GCxGC-FID, and IM-MS data of Concawe, and GC-MS data of SRM samples can be downloaded from http://paroc.tamu.edu/Software/Onel_etAl_2019_data.zip.

Funding: This research was funded by U.S. National Institute of Health (NIH) grant P42 ES027704 (awardee: Dr. Ivan Rusyn), and Texas

A&M Energy Institute. The content of this publication does not necessarily represent the official views of the NIH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

variability in chemical composition are an inherent feature of complex substances, we demonstrate how the choices of the data analysis and visualization methods can impact the communication of their characteristics to delineate sufficient similarity.

1. Introduction

Products of petroleum refining are prototypical UVCB (Unknown or Variable composition, Complex reaction products and Biological materials) substances [1]. UVCBs are some of the most challenging substances for the industry and regulators, because there are few established frameworks for evaluating UVCBs under current chemical regulatory policy and ensuring that there is no underestimation of hazard to either workers or the general users of the end-products [2]. Indeed, the complexity of the chemical composition of petroleum substances [3, 4], and in particular their multi-constituent nature and variability in product composition based on the variability in crude oil stocks, poses unique challenges to the regulators and registrants of these substances [5].

Typically, individual UVCB substances are assigned into a product category based on the manufacturing processes, physical/chemical properties (including refining history and boiling point/carbon number ranges), and limited analytical chemical information (such as hydrocarbon classes) [1, 2]. However, such broad similarity parameters may not always be considered sufficient by the regulatory bodies, and new approaches to facilitate the grouping of UVCBs are needed [6]. Recent developments in high-resolution and multi-dimensional analytical techniques improve characterization of complex substances by providing greater resolution of their chemical composition [7, 8]. Despite these advances, full chemical characterization of complex substances, such as petroleum UVCB substances, is still largely unattainable [6]. This presents a challenge for defining "sufficient similarity" for a substance of interest in comparison to those substances that may have already been tested for their potential human and ecological effects [9, 10].

A variety of analytical methods can be used to rapidly profile chemical composition of environmental samples and UVCBs, and all of them produce complex high-dimensional data sets [6, 11]. Quantitative interpretation of high-dimensional data has been an active area of statistics and a number of algorithms have been applied to classify unknown samples, or to derive discriminating data features [8, 12]. For example, data integration, clustering and visualization techniques using ion mobility-mass spectrometry (IM-MS) data of a subset of UVCBs have been used to determine the group-specific similarities [13]. Comparative analyses have also been performed. For example, de Carvalho Rocha, Schantz [14] has utilized principal components analysis (MPCA), principal factors analysis (PARAFAC), and self-organizing maps (SOM) analysis to differentiate among various types of fuels via pattern recognition. Although SOMs produce visually appealing grouping maps (Fig 1), comparative assessment to determine the optimal grouping is a challenge [15, 16]. Additional pattern recognition analysis techniques [1, 17–19] have also been explored to interpret the patterns in complex data sets; however, the outcomes of these methods are largely qualitative in nature and rely on the subjective visual evaluation of the grouping outcomes rather than quantitative comparative metrics.

The work presented in this manuscript aims to bridge the gap between the quantitative evaluation and visual communication of the grouping analysis outcomes to find the optimal grouping of complex substances. In this study, we present a data-driven framework that

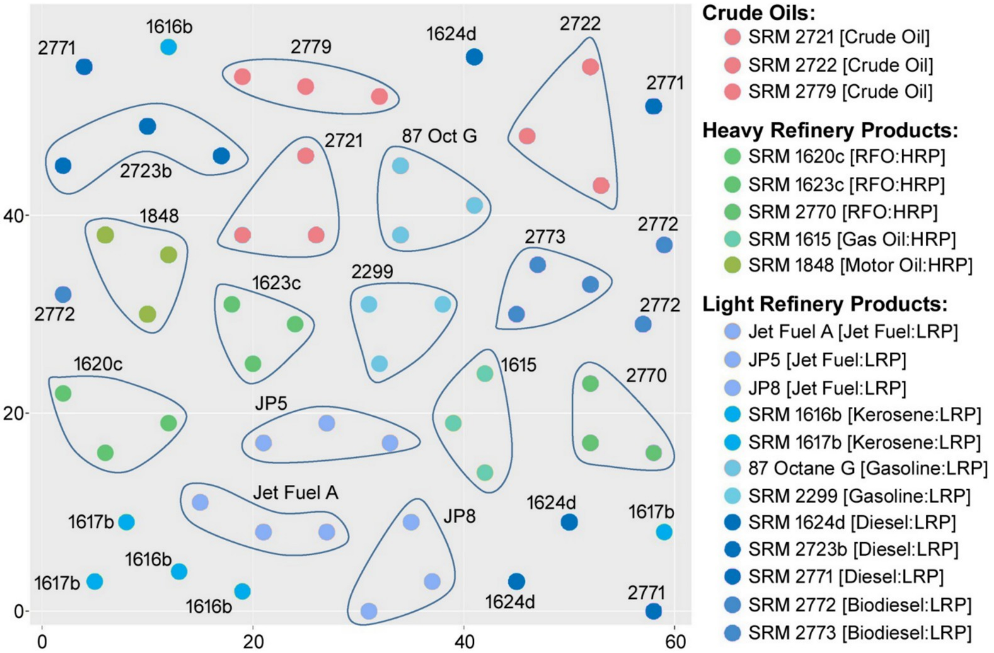


Fig 1. SOM recreated from de Carvalho Rocha, Schantz (14).

https://doi.org/10.1371/journal.pone.0223517.g001

includes two separate workflows for grouping of complex chemical substances. First, we present an unsupervised data analysis workflow based on hierarchical clustering, where the results are demonstrated through dendrograms, and the grouping quality is evaluated against existing manufacturing classes using the Fowlkes-Mallows (F-M) index. Second, we use supervised analysis. We have applied the Random Forest algorithm to train classification models to predict the manufacturing category information of these substances. The overall premise of the proposed framework is to provide (i) optimal grouping of complex substances, (ii) improved interpretation of the grouping results for decision-makers with the use of visualization techniques and identification of the most informative features, and (iii) comparative assessment of the grouping results by reporting quantitative metrics (i.e. the Fowlkes-Mallows index for clustering, and accuracy for classification analysis).

2. Materials & methods

2.1. Materials

In this study, we used two different sets of benchmark analytical chemistry data of: (i) 3 replicates of 20 Standard Reference Materials (SRM) (Table 1), and (ii) several petroleum UVCB substances, which were supplied by the European Petroleum Refiners Association AISBL, Concawe division (Brussels, Belgium) and referred to as "Petroleum UVCB samples" (Table 2). Specifically, SRMs are petroleum-related Certified Reference Materials and provided by the National Institute of Standards and Technology (NIST) [14]. In contrast, Petroleum UVCB samples were obtained from three separate refinement processes and categorized as straight run gas oils (SRGOs), other gas oils (OGOs), and vacuum and hydro-treated gas oils (VHGOs). Polycyclic-aromatic hydrocarbon (PAH), saturated hydrocarbon, and crude oil standards were provided by the Texas A&M Geochemical and Environmental Research Group (GERG) (College Station, TX).

Table 1. Standard Reference Materials (SRM) samples from de Carvalho Rocha, Schantz (14).

SRM ID	3-Class Grouping	9-Class Grouping	16-Class Grouping	Sample IDs
SRM 2722	Crude Oil	Crude Oil	Crude Oil (Heavy-Sweet)	petro203; petro204; petro205
SRM 2721	Crude Oil	Crude Oil	Crude Oil (Light-Sour)	petro274; petro275; petro276
SRM 2779	Crude Oil	Crude Oil	Gulf of Mexico Crude Oil	petro270; petro271; petro272
SRM 1615	Heavy Refinery Product	Gas Oil	Gas Oil	petro207; petro208; petro209
SRM 1848	Heavy Refinery Product	Motor Oil	Motor Oil Additive	petro218; petro219; petro220
SRM 2770	Heavy Refinery Product	RFO	S in Residual Fuel Oil	petro234; petro235; petro236
SRM 1623c	Heavy Refinery Product	RFO	S in Residual Fuel Oil	petro238; petro239; petro240
SRM 1620c	Heavy Refinery Product	RFO	S in Residual Fuel Oil	petro278; petro279; petro280
SRM 2773	Light Refinery Product	Biodiesel	Biodiesel (Animal-based)	petro230; petro231; petro232
SRM 2772	Light Refinery Product	Biodiesel	Biodiesel (Soy-based)	petro266; petro267; petro268
SRM 2723b	Light Refinery Product	Diesel	Low S Diesel	petro226; petro227; petro228
SRM 1624d	Light Refinery Product	Diesel	Sulfur in Diesel	petro214; petro215; petro216
SRM 2771	Light Refinery Product	Diesel	Zero S Diesel	petro222; petro223; petro224
Gasoline	Light Refinery Product	Gasoline	87 Octane Gasoline	petro258; petro259; petro260
SRM 2299	Light Refinery Product	Gasoline	S in gasoline	petro210; petro211; petro212
JP8	Light Refinery Product	Jet Fuel	Jet Fuel	petro246; petro247; petro248
JP5	Light Refinery Product	Jet Fuel	Jet Fuel	petro250; petro251; petro252
Jet Fuel A	Light Refinery Product	Jet Fuel	Jet Fuel	petro254; petro255; petro256
SRM 1617b	Light Refinery Product	Kerosene	S in Kerosene (High Level)	petro242; petro243; petro244
SRM 1616b	Light Refinery Product	Kerosene	S in Kerosene (Low Level)	petro262; petro263; petro264

*16-class grouping is based on designation by the National Institute of Standards and Technology (NIST), which was further grouped into 9 major classes. The 3-class grouping reflects the major refining distinctions among the SRMs.

https://doi.org/10.1371/journal.pone.0223517.t001

2.2. Chemical fingerprinting and experimental data processing

The analytical chemistry profile of SRMs was derived via Gas Chromatography-Mass Spectrometry (GC-MS) [14], whereas the chemical fingerprint of Petroleum UVCB substances was assessed with 3 different analytical chemistry techniques: (i) comprehensive two-dimensional gas chromatography with flame ionization detector (GC×GC-FID), (ii) GC-MS, and (iii) Ion Mobility Mass Spectrometry (IM-MS). The detailed experimental procedure is provided in Ferguson [20].

The GC-MS data from de Carvalho Rocha, Schantz (14) is a three-dimensional array, which consists of 23,248 elution times, and the 301 masses in the mass spectra for 60 Standard Reference Materials (triplicate runs of 20 samples). To reduce the computational complexity of the grouping analysis and the noise in the GC-MS data, we have selected 55 out of 301 m/z values (i.e. analytes) that correspond to Polycyclic Aromatic Hydrocarbons (PAHs) (S1 Table) and summed over the entire elution time dimension. This yields a two-dimensional (60 × 55) array, which is then used for grouping analysis.

2.3. Data analysis and visualization framework

We used two analysis workflows for grouping complex substances (Fig 2). In the unsupervised analysis, complex substances are grouped based on the similarity between the characteristics (i.e. analytical chemistry profiles) of the samples (complex substances) without prior knowledge of sample labels or categories. To evaluate the outcome of such grouping, we included a quantitative metric into the unsupervised analysis workflow to compare the outcome to a previously reported categorization of the samples (i.e. manufacturing classes). The details of the

Table 2. Petroleum UVCB samples.

Sample ID	Manufacturing class	CAS RN	CAS Name
CON07	OGO	64742-46-7	Distillates (petroleum), hydrotreated middle
CON09	OGO	64742-80-9	Distillates (petroleum), hydro-desulfurized middle
CON01	SRGO	64741-43-1	Gas oils (petroleum), straight-run
CON05	SRGO		
CON02	SRGO	68814-87-9	Distillates (petroleum), full-range straight-run middle
CON03	SRGO		
CON04	SRGO	68915-96-8	Distillates (petroleum), heavy straight-run
CON12	VHGO	64741-49-7	Condensates (petroleum), vacuum tower
CON13	VHGO	64741-58-8	Gas oils (petroleum), light vacuum
CON14	VHGO	64741-77-1	Distillates (petroleum), light hydrocracked
CON15	VHGO	64742-87-6	Gas oils (petroleum), hydrodesulfurized light vacuum
CON16	VHGO	68334-30-5	Fuels, diesel
CON17	VHGO	68476-30-2	Fuel oil, no. 2
CON18	VHGO	68476-31-3	Fuel oil, no. 4
CON20	VHGO	92045-24-4	Gas oils (petroleum), hydrotreated light vacuum

https://doi.org/10.1371/journal.pone.0223517.t002

proposed unsupervised analysis workflow are described in Section 2.3.2. In the supervised analysis, known categorizations/classes of the samples are used to build classification models, which can then be used to predict the class for an unknown substance. This idea is based on the read-across, where similar complex substances that are grouped together according to their physical/chemical properties may have similar effects [2]. The details of the proposed

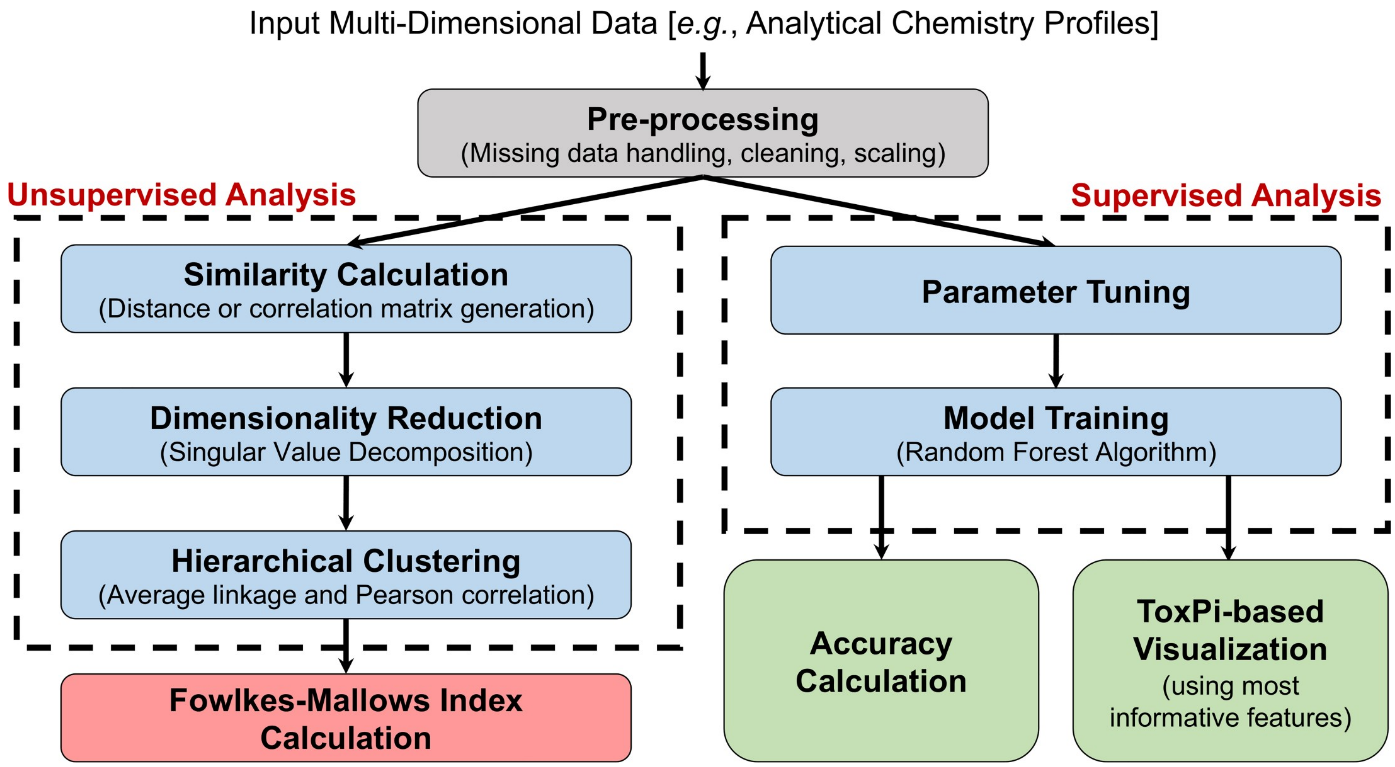

Input Multi-Dimensional Data [e.g., Analytical Chemistry Profiles]
Pre-processing
(Missing data handling, cleaning, scaling)
Unsupervised Analysis
Supervised Analysis
Similarity Calculation
(Distance or correlation matrix generation)
Dimensionality Reduction
(Singular Value Decomposition)
Hierarchical Clustering
(Average linkage and Pearson correlation)
Fowlkes-Mallows Index Calculation
Parameter Tuning
Model Training
(Random Forest Algorithm)
Accuracy Calculation
ToxPi-based Visualization
(using most informative features)

Fig 2. Data processing and visualization workflow.

https://doi.org/10.1371/journal.pone.0223517.g002

unsupervised analysis workflow are described in Section 2.3.3. Independent of which workflow (unsupervised or supervised analysis), the initial common step is data pre-processing, which is crucial to obtain robust and reliable grouping models (Section 2.3.1).

2.3.1. Data pre-processing. Data pre-processing steps include (i) data formatting, (ii) handling missing data, and (iii) data cleaning and scaling. Following these steps ensures data quality in order to build robust and reliable models. The application of these steps to each specific data set is provided below.

First the three-dimensional analytical chemistry data (the GC×GC-FID, GC-MS, and IM-MS data sets) needed to be unfolded into a 2-dimensional matrix. The GC-MS data of SRMs, after the experimental data processing step described in Section 2.2, is already two-dimensional. The unfolding was performed so that, for the final matrix, rows correspond to the complex substances, and columns are the analytical features (measurements), which are the concatenated values of carbon number and molecular class composition. This process yielded an array size of 15 × 310 for GC×GC-FID, 15 × 248 for GC-MS, and 15 × 403 for IM-MS data sets of the Petroleum UVCB samples.

Next, the two-dimensional analytical chemistry data sets were examined to detect any missing points. Although advanced missing data handling methods are sometimes used in complex data analysis [21], here the missing data points within analytical chemistry profiles indicate undetected chemical composition for a specific molecular class. Thus, we replaced the corresponding missing fields with zeros.

The data sets were further cleaned by removing the columns (carbon number–molecular class compositions) if they show negligible variation. Here, we removed columns with a standard deviation (SD) of 0, or $SD<0.05$ for the smaller Petroleum UVCB data set. This step reduced the number of features in GC×GC-FID from 310 to 192, GC-MS from 248 to 62, and IM-MS data from 403 to 68. This step did not eliminate any measurements from the 60 x 55 matrix of SRM samples.

The final step prior to data analysis is the scaling of the data sets. The clean two-dimensional arrays were scaled by using row-wise min-max scaling, where each row corresponds to a new sample and each column is a new analytical feature. Each row value was scaled by subtracting the minimum value of that row and then dividing it by the range of the corresponding row. Row-wise scaling was not performed on Petroleum UVCB data sets, because the data was already pre-processed within PetroOrg software [22] and row-wise scaled. Prior to the classification analysis, we also performed column-wise min-max scaling on the row-wise scaled data arrays. This additional scaling step is required to ensure that each measurement has approximately equal weight in training classification models.

2.3.2. Unsupervised analysis workflow. Unsupervised analysis examines the patterns of data to draw conclusions for the grouping structure of the samples without the reference categorization information. The two most prevalent unsupervised analysis techniques used in the literature are clustering analysis [23], as used in our workflow (Fig 2 left panel), and SOMs [24]. The detailed steps of this workflow are given below. The R Markdown documentation of this analysis for SRM samples is also provided in the Supplementary Material (S1 Text).

Under the unsupervised workflow, we performed cluster analyses using the data as described, and also after producing a reduced-rank data set, in order to judge the effect of using reduced-rank data (i.e., after "de-noising" the data). The original unscaled SRM and Petroleum UVCB data sets were linearly scaled and centered in a row-wise fashion (i.e. z-score normalization). The resultant data was then decomposed using singular value decomposition (SVD) [25, 26] in R to produce a reduced-rank data set corresponding to 85% of the variation in the original data.

Pearson correlations of both the original and reduced-rank analytical chemistry data were used as a similarity index for hierarchical cluster analysis of the samples using hclust in R with average linkage [27, 28].

A quantitative comparison of the clustering to the known substance categories (treated here as a "gold standard") was performed using the Fowlkes-Mallows (F-M) index [29]. The F-M index is traditionally used to compare two dendrograms but can also be used to compare a single dendrogram to a fixed categorization. We created two sets of hierarchical clustering dendrograms for both the Petroleum UVCB and SRM samples. First, an artificial dendrogram was generated by calculating the Euclidean distance between the indices (0/1) of a reference categorization. Second, the correlation matrices of the Petroleum UVCB and SRM samples were used to generate a dendrogram. Next, the dendrograms were both cut into the known number of manufacturing classes (i.e. 3 for Petroleum UVCB samples, and 3, 9, or 16 for SRM samples) to assess the number of the common complex substances in the obtained clusters. This number was then used to calculate the F-M index for the two groupings (i.e., comparing known categorization to the data-based grouping). The F-M index can be expressed as the geometric mean of precision and recall, two machine learning metrics that are widely used in data-driven modeling [23]. Expressed mathematically, we have

$$FM = \sqrt{\frac{TP}{TP+FP} \cdot \frac{TP}{TP+FN}}$$

where TP is True Positive, FP is False Positive, and FN is False Negative. TP indicates the number of complex substances that are grouped under category A in terms of manufacturing category and are also grouped under category A in terms of analytical chemistry profile. In contrast, FP and FN denote the number of complex substances that are grouped differently. The F-M index varies between 0 and 1, where 0 indicates the absence of any similarity, and 1 indicates 100% identity between reference categorization and clustering results. More details on the F-M index and other metrics for clustering comparison can be found in Wagner & Wagner [30]. The F-M index was calculated via the FM_index function of the dendextend package in R.

To test the statistical significance of the grouping results, we also calculated the distribution of the F-M index under the null hypothesis of no relation between two clustering dendrograms. This null distribution was generated by shuffling the group labels of samples using 1000 permutations, with an empirical *p*-value determined by the proportion of permuted F-M index values exceeding the observed. We used $\alpha = 0.05$ as a false-positive threshold. The null F-M index calculation with 1000 permutations of the group labels was performed via the Bk_permutations function of the dendextend package in R.

2.3.3. Supervised analysis workflow. Although unsupervised analysis can elucidate previously unknown structures in the data, supervised methods can identify the features most influential in classification. Moreover, in this context, supervised analysis may highlight substances that show comparatively poor similarity to the other members of the manufacturing category. Supervised learning algorithms are widely used in various engineering and sciences problems [31–40]. Here, we used the Random Forest decision-tree algorithm [41] to train models to predict manufacturing category from the features. The models were evaluated by their classification accuracy, and the results were visualized via ToxPi representation [42] for enhanced interpretation (Fig 2 right). The steps of the proposed supervised analysis workflow are provided below and applied to both Petroleum UVCB and SRM data sets. The documentation of the analysis through SRM samples was created using R Markdown and provided in the Supplementary Material (S1 Text).

In our implementation of the Random Forests, the number of analytical features was tuned via grid search using the trainControl function of the caret package in R, where each model training was performed using leave-one-out cross validation with 500 decision trees. The final Random Forest classifier was then built on the whole data set with 500 decision trees, where each tree was modeled by using the optimal number of analytical features. In addition, the ranking of the analytical features was obtained by calculating the mean decrease in classification accuracy among the 500 decision trees. This analysis was done via the randomForest function of the randomForest package.

To evaluate the classification model accuracy, an initial step was to extract the confusion matrix of the model, i.e. the number of true and falsely predicted samples for each class. Next, the classification accuracy was calculated, which is the percentage of true predicted number of samples from all classes with respect to the total number of samples.

In addition to the quantification of the classification models, we produced Toxicological Prioritization Index (ToxPi) profiles of complex substances by using the ranked analytical feature list from the classification analysis [42–44]. By integrating multiple data sources into an overall, weight-of-evidence score, and transforming them into clear visual rankings, ToxPi provides an effective way for visual communication of high-dimensional data sets. Here we integrated the top 10 most informative chromatographic features that were extracted during classification modeling step, to obtain the ToxPi visualization of complex substances.

3. Results & discussion

3.1. Quantifying the Unsupervised analysis

A recent study [14] has shown that GC-MS combined with unsupervised chemometric analysis can be used to differentiate among complex substances and mixtures. The authors have concluded that the SOM non-linear method proved to be effective in generating a separation model. However, the model is more difficult to interpret than the linear models such as MPCA and PARAFAC [45, 46]. The unified distance matrix of the SOM analysis of the 20 SRMs (Table 1) from de Carvalho Rocha, Schantz (14) is shown in Fig 1. The visualization of the data using SOM makes it apparent that the replicates of the same sample were clustered well (15 of 20 samples have all 3 replicates in close proximity to each other) (Fig 1). However, it is less obvious that the SOM analysis can discriminate among the broader categories of samples (3 classes: crude oils, heavy and light refinery products; 9 classes: crude oils, residual fuel oils, gas oil, motor oil, biodiesels, diesels, gasolines, kerosenes and jet fuels). Only jet fuels and gasoline samples of light refinery products were clustered close to each other (Fig 1).

To explore additional visualization methods, we used the data from de Carvalho Rocha, Schantz (14) to perform unsupervised clustering analysis of the samples (Fig 3). The results showed that all technical replicates of 20 substances were clustered tightly, which indicates high reproducibility of the analytical data from GC-MS analysis of these complex substances. However, when 3 or 9 broader manufacturing classes were considered, the samples were not clustered as closely as they were in the 16 manufacturing classes. For 9 class grouping, replicate samples of gas oils, biodiesels, and motor oils were grouped together in distinct clusters. In 3 class grouping results, only crude oil samples were grouped under one of the three clusters. Even though most of the light refinery products (one gasoline, three diesel, two jet fuels, and two kerosene samples) were clustered together in one of the three groups, one gasoline and two biodiesel samples fall into separate clusters (Fig 3). These analyses demonstrated that the analytical features derived from GC-MS were, by themselves, insufficient for justifying grouping of these complex substances into the manufacturing categories.

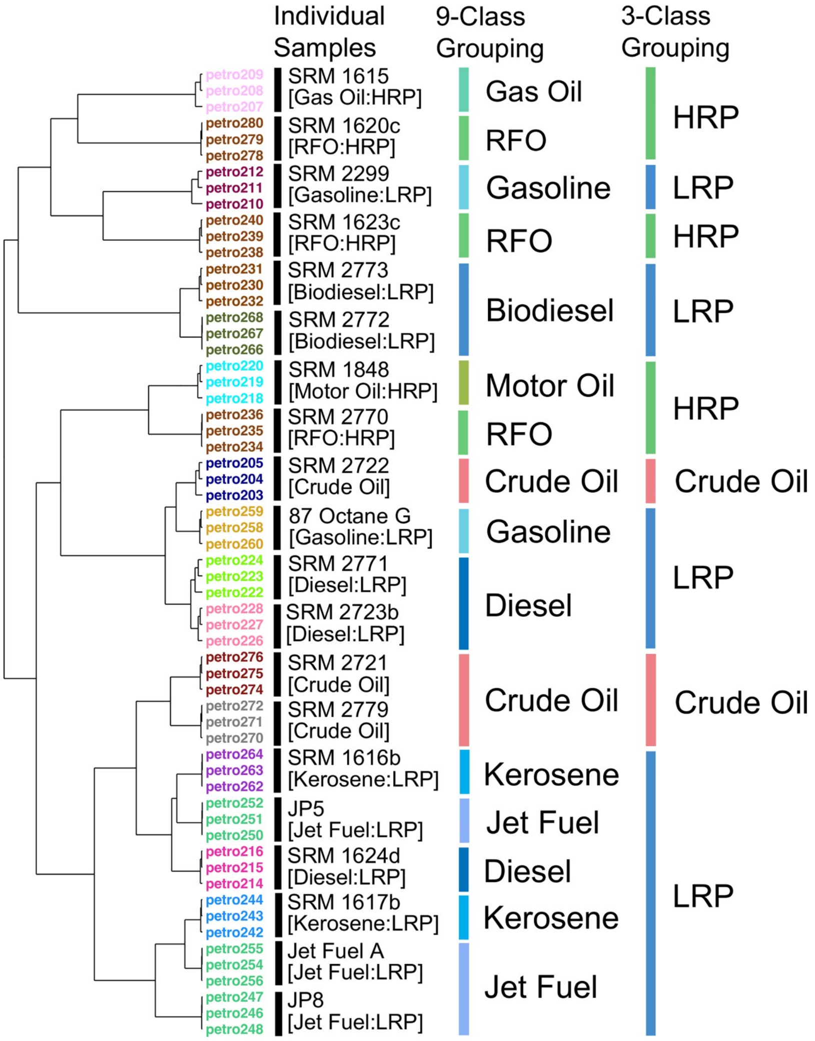
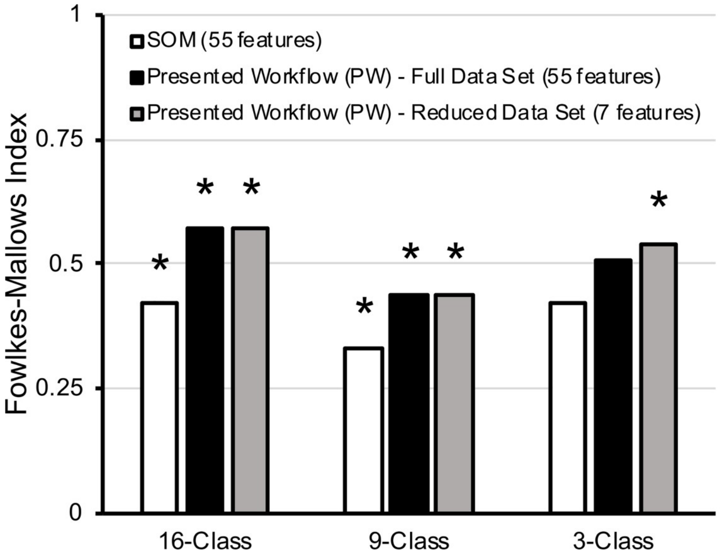


Fig 3. Dendrograms for the SRM samples clustering from the reduced data set into 3, 9 and 16 categories. LRP: Light Refinery Product, HRP: Heavy Refinery Product.

https://doi.org/10.1371/journal.pone.0223517.g003

Next, we aimed to quantitatively compare the outcomes of SOM and clustering analyses to the manufacturing class categories of these samples. We used the Fowlkes-Mallows (F-M) index to provide a quantitative metric for such comparisons [29]. Although there is no direct method to assess the grouping quality using the SOM analysis, we have extracted the x and y coordinates of each SRM sample on the SOM map as reported by de Carvalho Rocha, Schantz (14) (Fig 1) and used the Euclidean distance-based similarity matrix to obtain the F-M index for the SOM-based grouping analysis. The F-M index was also used to assess the effect of dimensionality reduction on the outcomes of clustering analyses.

Fig 4 displays the F-M indices for SOM-based analysis, as well as the presented unsupervised analysis workflow on full data set of 55 GC-MS features and a reduced set of 7 features after SVD. The *p*-values for the significance of the correspondence of the clustering compared to the known class assignment are also reported (Fig 4). The *p*-values obtained for 3-class grouping were higher than 0.05 for the SOM-based and original data set of SRM samples, implying these results were not statistically significant. Subtle differences differentiate these materials into 16 categories. When grouped under 3 categories, these differences presented



		F-M Index	p-value
16-Class	SOM	0.42	0.000
	PW - Full	0.57	0.000
	PW -Reduced	0.57	0.000
9-Class	SOM	0.33	0.000
	PW - Full	0.44	0.000
	PW -Reduced	0.44	0.000
3-Class	SOM	0.42	0.192
	PW - Full	0.51	0.257
	PW -Reduced	0.54	0.037

Fig 4. Fowlkes-Mallows index for the outcomes of clustering of SRM samples. * indicates that the results are statistically significant at the 0.05 level.

https://doi.org/10.1371/journal.pone.0223517.g004

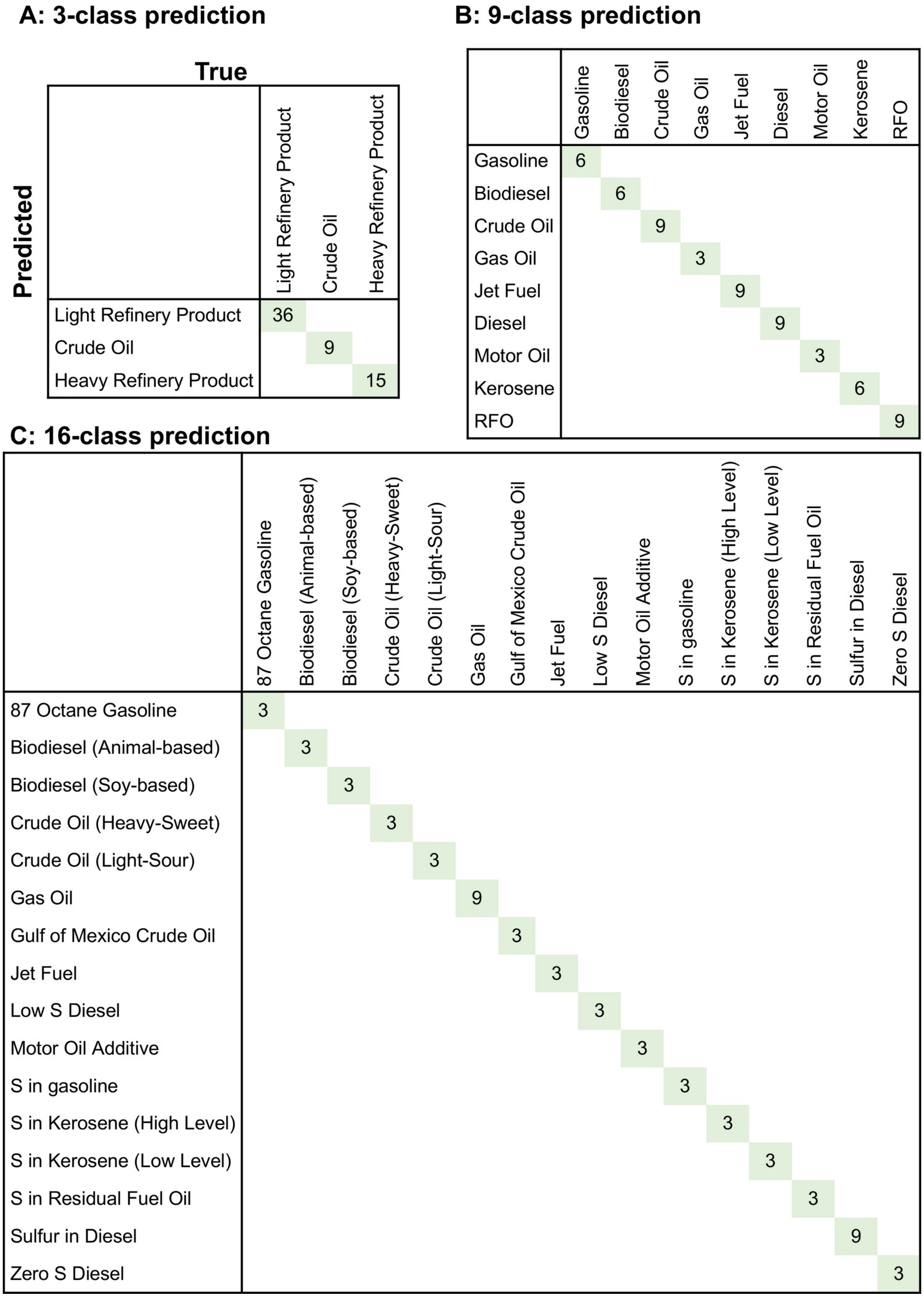
themselves as noise. Hence, the random permutation of these samples led to higher F-M indices by chance. In contrast, the *p*-values for 3-class grouping with the reduced data sets were lower than 0.05. This indicates evidence that dimensionality reduction eliminates redundant analytical features from the data sets (from 55 to 7) which further reduces the noise, leading to statistically significant results with a similar (and slightly improved) F-M index.

Based on 9 and 16-class groupings of SRMs, one can clearly observe that hierarchical clustering outperformed SOM analysis. The F-M index increased from 0.33 to 0.44, and 0.42 to 0.57 for 9 and 16-class groupings, respectively. Although the dimensionality reduction did not further increase the F-M index for 9 and 16 class groupings, it also did not hinder the grouping quality and provided equally good results with lower number features (7 out of 55).

3.2. Importance of substance sample size during supervised analysis

Here, we benefit from the read-across hypothesis that "complex substances that group similarly based on manufacturing may exhibit similar hazard profiles," and move from unsupervised to supervised analysis. To this end, we are building classification models using analytical chemistry profiles of samples. For each of the 20 SRM substances, GC-MS was run three times, which made the final GC-MS data set larger in terms of the number of samples. Thus, an interesting question that we can examine is that how many sample replicates would be adequate to develop data-driven models that can precisely differentiate class patterns.

Figs 5 and 6 demonstrate the confusion matrices obtained from the trained Random Forest classifiers. These matrices report known ("true") and predicted (through the trained Random Forest classifier) classes for each SRM sample. The results showed that we achieved 100% classification accuracy when we used all replicates provided in Table 1 (Table 3, Fig 5). The classification accuracy decreased to 65%, 35%, and 15% for 3, 9, and 16-class groupings when we only used 1 out 3 replicates (Table 3, Fig 6). The main reason for this fact is that the number of samples per group decreases as the number of classes increases. In particular, 14 out of 16 classes were represented with only a single sample during model training for the 16-class predictions (Fig 6C). Similarly, 5 out 9 classes were represented with only a single sample during model training for the 9-class predictions (Fig 6B). This decrease in the amount of information per



A: 3-class prediction

True

Predicted

	Light Refinery Product	Crude Oil	Heavy Refinery Product
Light Refinery Product	36		
Crude Oil		9	
Heavy Refinery Product			15

B: 9-class prediction

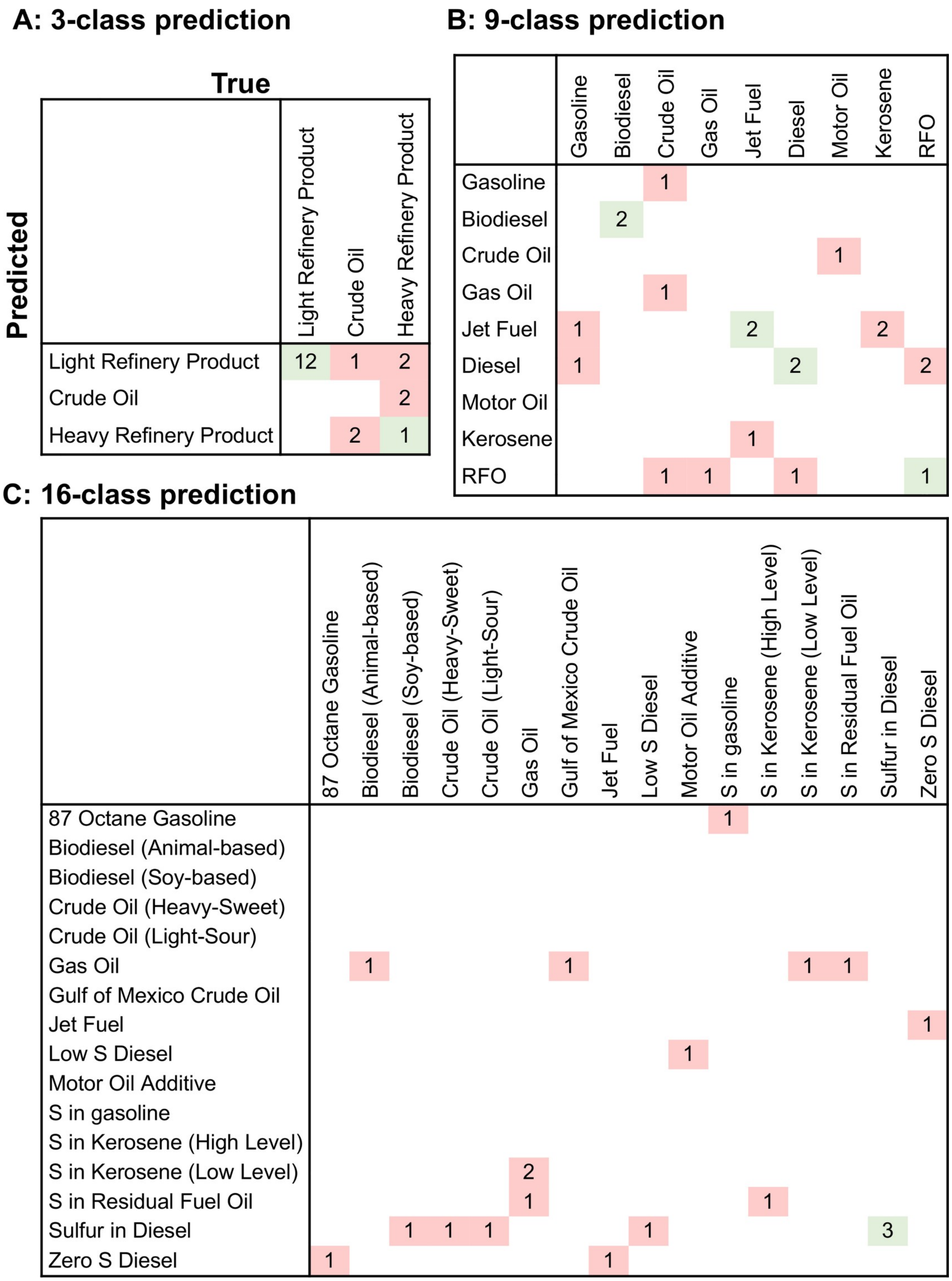
	Gasoline	Biodiesel	Crude Oil	Gas Oil	Jet Fuel	Diesel	Motor Oil	Kerosene	RFO
Gasoline	6								
Biodiesel		6							
Crude Oil			9						
Gas Oil				3					
Jet Fuel					9				
Diesel						9			
Motor Oil							3		
Kerosene								6	
RFO									9

C: 16-class prediction

	87 Octane Gasoline	Biodiesel (Animal-based)	Biodiesel (Soy-based)	Crude Oil (Heavy-Sweet)	Crude Oil (Light-Sour)	Gas Oil	Gulf of Mexico Crude Oil	Jet Fuel	Low S Diesel	Motor Oil Additive	S in gasoline	S in Kerosene (High Level)	S in Kerosene (Low Level)	S in Residual Fuel Oil	Sulfur in Diesel	Zero S Diesel
87 Octane Gasoline	3															
Biodiesel (Animal-based)		3														
Biodiesel (Soy-based)			3													
Crude Oil (Heavy-Sweet)				3												
Crude Oil (Light-Sour)					3											
Gas Oil						9										
Gulf of Mexico Crude Oil							3									
Jet Fuel								3								
Low S Diesel									3							
Motor Oil Additive										3						
S in gasoline											3					
S in Kerosene (High Level)												3				
S in Kerosene (Low Level)													3			
S in Residual Fuel Oil														3		
Sulfur in Diesel															9	
Zero S Diesel																3

Fig 5. Confusion matrices for SRM sample classification with 3 replicates. (A) 3-class, (B) 9-class, and (C) 16-class grouping.

https://doi.org/10.1371/journal.pone.0223517.g005


True
Predicted

A: 3-class prediction

	Light Refinery Product	Crude Oil	Heavy Refinery Product
Light Refinery Product	12	1	2
Crude Oil			2
Heavy Refinery Product		2	1

B: 9-class prediction

	Gasoline	Biodiesel	Crude Oil	Gas Oil	Jet Fuel	Diesel	Motor Oil	Kerosene	RFO
Gasoline			1						
Biodiesel		2							
Crude Oil							1		
Gas Oil			1						
Jet Fuel	1				2			2	
Diesel	1					2			2
Motor Oil									
Kerosene					1				
RFO			1	1		1			1

C: 16-class prediction

	87 Octane Gasoline	Biodiesel (Animal-based)	Biodiesel (Soy-based)	Crude Oil (Heavy-Sweet)	Crude Oil (Light-Sour)	Gas Oil	Gulf of Mexico Crude Oil	Jet Fuel	Low S Diesel	Motor Oil Additive	S in gasoline	S in Kerosene (High Level)	S in Kerosene (Low Level)	S in Residual Fuel Oil	Sulfur in Diesel	Zero S Diesel
87 Octane Gasoline											1					
Biodiesel (Animal-based)																
Biodiesel (Soy-based)																
Crude Oil (Heavy-Sweet)																
Crude Oil (Light-Sour)																
Gas Oil		1					1						1	1		
Gulf of Mexico Crude Oil																
Jet Fuel																1
Low S Diesel										1						
Motor Oil Additive																
S in gasoline																
S in Kerosene (High Level)																
S in Kerosene (Low Level)						2										
S in Residual Fuel Oil						1						1				
Sulfur in Diesel			1	1	1				1						3	
Zero S Diesel	1							1								

Fig 6. Confusion matrices for SRM sample classification with 1 replicate. (A) 3-class, (B) 9-class, and (C) 16-class grouping.

https://doi.org/10.1371/journal.pone.0223517.g006

class makes model learning significantly challenging (Table 3). Hence, we can conclude that single sample per class does not provide adequate information to capture the individual class characteristics. Moreover, the high-dimensional nature of the GC-MS data with 55 features further hindered the classification accuracy of SRM materials when using only one sample per

Table 3. Classification accuracy of SRM samples using sample replicates.

Prediction class type	Number of sample replicates used	Classification accuracy	Classification accuracy (permuted)	*p*-value
3-class	3	100%	44.8±7.0%	0.000
	1	65%	48.0±9.2%	0.023
9-class	3	100%	10.9±5.1%	0.000
	1	35%	6.6±6.9%	0.000
16-class	3	100%	6.5±4.1%	0.000
	1	15%	4.2±5.1%	0.019

https://doi.org/10.1371/journal.pone.0223517.t003

category. Yet, the prediction accuracies of the classifiers for each analysis were higher than those for random prediction, indicating they were statistically significant (Table 3). This was validated through *p*-value calculations by using the original and 1000 random permutation grouping results (Table 3). The confusion matrices generated from the average of 1000 permutations of SRM samples are provided in S1 Fig and S2 Fig for 3 and 1 replicates, respectively.

Similar trend was observed with Petroleum UVCB samples, where we built classification models using only 1 replicate of each sample (Table 4, S3 Fig). The results demonstrated that classification model accuracies for the Petroleum UVCB samples were not satisfactory, where the only statistically significant result was obtained from IM-MS data with 60% classification accuracy. Therefore, in order to build an accurate classification model, we need higher number of experimental replicates for each particular complex substance to capture and learn the non-linear characteristics of their chemical complexity. Although clustering can group the samples accurately independent of the sample size, given that measurements are significantly distinct from each other, sufficient data sample size is essential during the classification model building. Nonetheless, each experimental replicate leads to an additional cost and requires extra time, and resources. Thus, minimizing the number of sample replicates while achieving accurate predictive classifiers is of utmost importance. In this work, we observed that, given high quality analytical chemistry data, 3 replicates were sufficient to build accurate and robust classifiers. It is important to note that the sample size is critical during the model training phase, where the models benefit from higher number of samples. However, this is not the case for the testing phase where a single experiment is sufficient to predict its class information of an unknown complex substance.

3.3. Importance of substance number to class ratio during supervised analysis

Another important question that needs to be answered is the following: How precisely can we categorize a new, unknown, substance with a data-driven model which is trained with an analytical chemistry profile database of categorized substances with no prior labeled samples of the tested substance? Hypothetically, one can accurately classify an unknown substance when provided a classifier trained with an analytical chemistry profile database that includes a high number of substances per class. In other words, a new substance can be precisely labeled if the

Table 4. Classification accuracy of Petroleum UVCB samples.

Prediction class type	Analytical technique used	Classification accuracy	Classification accuracy (permuted)	*p*-value
3-class (VHGO, SRGO, OGO)	GC-MS	40.0%	39.9±13.5%	0.395
	GC×GC-FID	46.7%	39.4±13.9%	0.222
	IM-MS	60.0%	41.4±12.0%	0.047

https://doi.org/10.1371/journal.pone.0223517.t004

Table 5. 3-class classification accuracy of SRM substances excluding sample replicates.

Prediction class type	Classification accuracy	Classification accuracy (permuted)	*p*-value
3-class	75%	71.8±3.6%	0.092

https://doi.org/10.1371/journal.pone.0223517.t005

analytical chemistry profile database provides an accurate mean profile of particular classes. Here, we used 20 SRM substances to understand whether the number of substance-to-class ratio of the data set can enable accurate categorization of each SRM substance. To this end, we developed one Random Forest classification model per each of the 20 SRM substances and reported the overall classification accuracy. In particular, we excluded the analytical chemistry profile information from all 3 replicates of the selected SRM substance during model training, and then predicted the category of with the trained model.

For 3-class predictions, the classification accuracy was obtained as 75% (Table 5), where the confusion matrices for original and average of 1000 permutations of SRM substance group labels are provided in S4 Fig. However, the calculated *p*-value (Table 5) showed that the developed data-driven models were statistically insignificant, which we attribute to a low number of substance-to-class ratio. This ratio deteriorates as the number of categorizes increase for 20 SRM substances. As a confirmation of our observation, we also ran a random forest model in which we calculated an average feature profile per substance, collapsing the replicates in to one artificial feature vector. The resulting classification accuracy was very similar (in the range of 70%-75%, not shown).

S2 Table tabulates the number of substances and sample replicates per each class for 3-class, 9-class, and 16-class categorization. As can be seen from the S2 Table, removing all 3 replicates of a substance often corresponds to removing all samples of a class in several instances of a 9-class and 16-class analysis, thus hinders us to develop classification models for 2 (Category 4 and 7), and 14 categories (excluding only Category 6 and 15) for 9-class, and 16-class analysis, respectively. Therefore, we cannot report an overall classification accuracy for 9, and 16-class analysis. These results indicate that a high number for the substance-to-class ratio is crucial for accurate classification of an unknown substance with data-driven models that are trained without any previous samples of the tested substance. Therefore, we conclude that (i) continuous improvement of the analytical chemistry profile database used for model training with the addition of categorized substances per each class, and (ii) continuous update of the data-driven model are essential and necessary for accurate categorization of a new, unknown substance.

3.4. Facilitation of data interpretation via ToxPi representation

In addition to developing highly accurate classifier models to predict group/class information of an unknown complex substance using the sample replicates of categorized substances, we also reported the top 10 most informative features that distinctively identify the class patterns of SRM materials (Table 6). These informative features help us to facilitate the visual communication of the findings via ToxPi visualization as shown in Fig 7.

The ToxPi profiles of SRM samples successfully demonstrated the distinct nature of gas/motor oils, biodiesels, and crude oils (with the exception of SRM 2722) with respect to the rest of SRMs. Specifically for crude oils, all of the top 10 chromatographic features helped to identify crude oils among SRMs. Whereas for gas/motor oils, the profiles revealed the importance of C2-napthobenzothiophenes for further identification between them. Moreover, the ToxPi profiles showed that C3-phenanthrene/anthracenes, C2-naphthobenzothiophenes, and benzothiophene measurements were the characteristics of biodiesel samples and can differentiate them from the rest of the SRMs. We also observed the high similarity among a subgroup of

Table 6. Top 10 most informative GC-MS chromatographic features with respect to the classification accuracy of the petroleum substances. See S3–S5 Tables for the list of all chromatographic features and their respective ranks in each analysis.

	3-Class prediction		9-Class prediction		16-Class prediction	
GC-MS Chromatographic Feature	Rank*	Mean decrease in accuracy (%)#	Rank	Mean decrease in accuracy (%)	Rank	Mean decrease in accuracy (%)
C4-Naphthalenes	6	6.82	13	8.68	1	10.38
Naphthobenzothiophene	10	6.54	14	8.56	7	9.44
C2-Naphthobenzothiophenes	24	6.12	2	9.09	10	9.40
Benzothiophene	21	6.25	3	9.08	14	9.25
C3-Phenanthrene/anthracenes	5	6.90	6	8.86	33	8.55
C4-Phenanthrene/anthracenes	8	6.58	35	7.76	3	9.57
Benzo(b)fluoranthene	19	6.28	20	8.42	8	9.42
Dibenzothiophene	41	5.76	4	9.02	4	9.56
C1-Dibenzothiophenes	2	7.10	22	8.35	26	8.69
Benzo(k)fluoranthene	27	6.06	16	8.53	11	9.34

*Rank of the feature among 55 total for each classification analysis (3-, 9-, or 16-class prediction). Top 10 features with the overall highest rank in all three analyses were selected.

#Mean decrease in the accuracy of classification when this feature is removed from the analysis.

https://doi.org/10.1371/journal.pone.0223517.t006

light refinery products that includes jet fuels, kerosenes and gasolines, where the weight of benzothiophene remained to be the unique characteristic among all of them. This proves that the GC-MS data could not provide clear distinction among these substances. Finally, we noted the

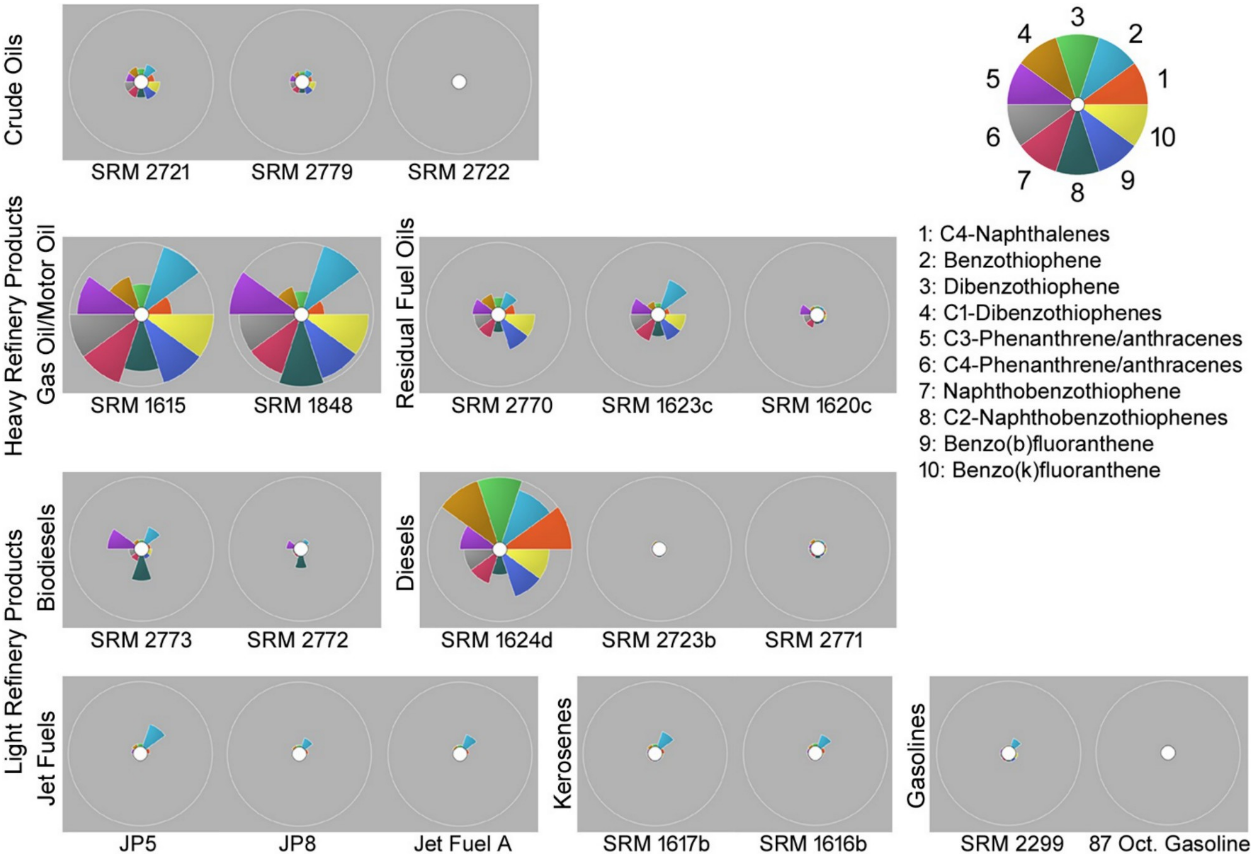


Fig 7. ToxPi visualization of SRM samples using top 10 most informative chromatographic features.

https://doi.org/10.1371/journal.pone.0223517.g007

major difference between diesel samples. Unlike the other two diesel samples, SRM 2723b and SRM 2771, most of the top 10 selected analytical features were significant for identifying SRM 1624d. In particular, dibenzothiophene, C1-dibenzothiophenes and C4-naphthalenes were the distinct measurements that differentiate SRM 1624d from the rest of the SRMs the most. Furthermore, the PCA of the extracted ToxPi scores helps us to depict the distinction between the complex substances by using the most informative analytical feature information (Fig 8).

3.5. Comparison of GC×GC-FID, GC-MS, and IM-MS techniques via Fowlkes-Mallows index

In addition to quantifying the grouping quality and class information, it is imperative to investigate the appropriate analytical chemistry technique that produces the optimal grouping for substances with complex chemistries. The majority of regulatory and standardized chemical compositional analysis protocols utilize GC-MS as the instrument of choice to fingerprint UVCB substances. Generally, a GC-MS instrument employs a capillary column, heated by an oven at a predetermined temperature gradient in order to separate compounds by boiling point and polarity. The eluting compounds are then ionized and analyzed by a detector. Since molecules of specific molecular classes maintain distinct mass ion fracture patterns, GC-MS is able to differentiate ion signals from multiple compounds. However, the column peak capacity of a GC-MS can become overloaded, causing a baseline hump termed as an unresolved complex mixture (UCM). In such cases, the column no longer has the resolving power to separate all the compounds within the sample, which is typically observed in petroleum substance analysis, since an individual petroleum substance contains more than 10,000 different chemical compounds. This may limit the number of molecules that can be effectively differentiated by the instrument and hinder a robust chemical fingerprint production.

However, in recent years, instrument resolution power and sensitivity has increased, allowing for more detailed characterization of complex substances. The incorporation of two gas chromatography columns with different selectivity (GC×GC-FID) increases the peak capacity of the instrument and allows for improved separation of molecules that form a UCM under GC-MS analysis. Moreover, ion mobility mass spectrometry (IM-MS) incorporates unique ionization methods, electron spray (ESI) or atmospheric photo ionization (APPI), along with separation techniques based on size, shape, and charge of the ionized molecule. This further increases the analytical sensitivity and enables improved chemical fingerprinting. Although these two techniques further enhance the ability to characterize complex substances like petroleum products, their application is still novel and not widely studied within the scientific, regulatory, or industrial communities [14, 47]. Despite the technological advances that are introduced by GC×GC-FID, and IM-MS techniques over GC-MS, there is no evidence examining any potential improvements on complex substance grouping. Therefore, we utilized the Fowlkes-Mallows index to provide comparative assessment between these three analytical chemistry techniques using the Petroleum UVCB samples.

Fig 9 demonstrates that GC×GC-FID and GC-MS yielded statistically insignificant F-M indices due to the limited sample size. IM-MS was the only one yielding statistically significant results, only after dimensionality reduction, which provided the most useful information to reveal the class differences among the Petroleum UVCB samples. Their corresponding clustering dendrograms are provided in Fig 10. Specifically, the F-M index of the grouping of Petroleum UVCB samples with 8 features generated via the IM-MS technique was 0.49. Although we could not draw specific conclusions among GC×GC-FID and GC-MS, we can report that IM-MS performed superior than the other two techniques in terms of capturing the chemical characteristics of complex substances.

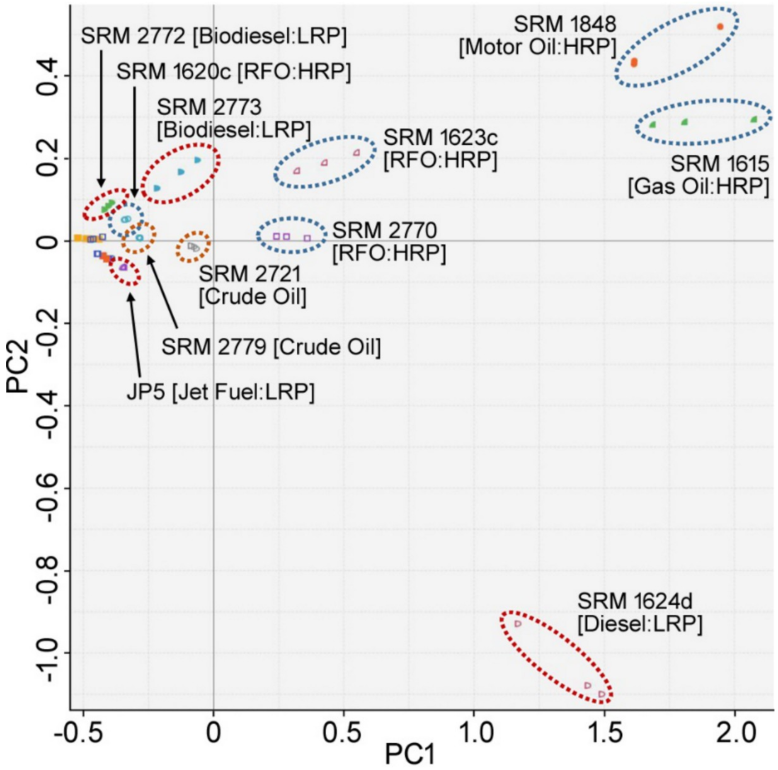
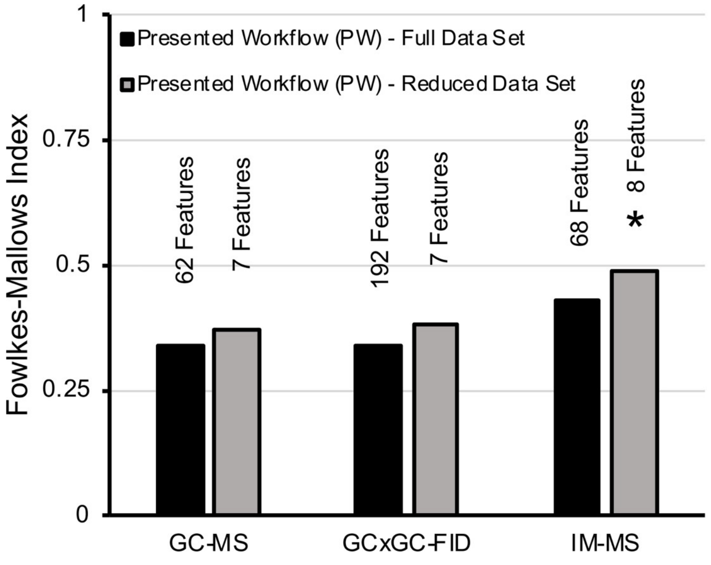


Fig 8. PCA of ToxPi scores.

https://doi.org/10.1371/journal.pone.0223517.g008

4. Conclusion

In this study, we established a data-driven framework for optimal grouping complex chemical substances based on their chemical characteristics, and provided quantitative and visual



		F-M Index	p-value
GC-MS	PW - Full	0.34	0.723
	PW - Reduced	0.37	0.176
GCxGC-FID	PW - Full	0.34	0.723
	PW - Reduced	0.38	0.169
IM-MS	PW - Full	0.43	0.166
	PW - Reduced	0.49	0.035

Fig 9. F-M index for the outcomes of clustering of Petroleum UVCB samples analyzed using 3 different techniques. * indicates that the results are statistically significant.

https://doi.org/10.1371/journal.pone.0223517.g009

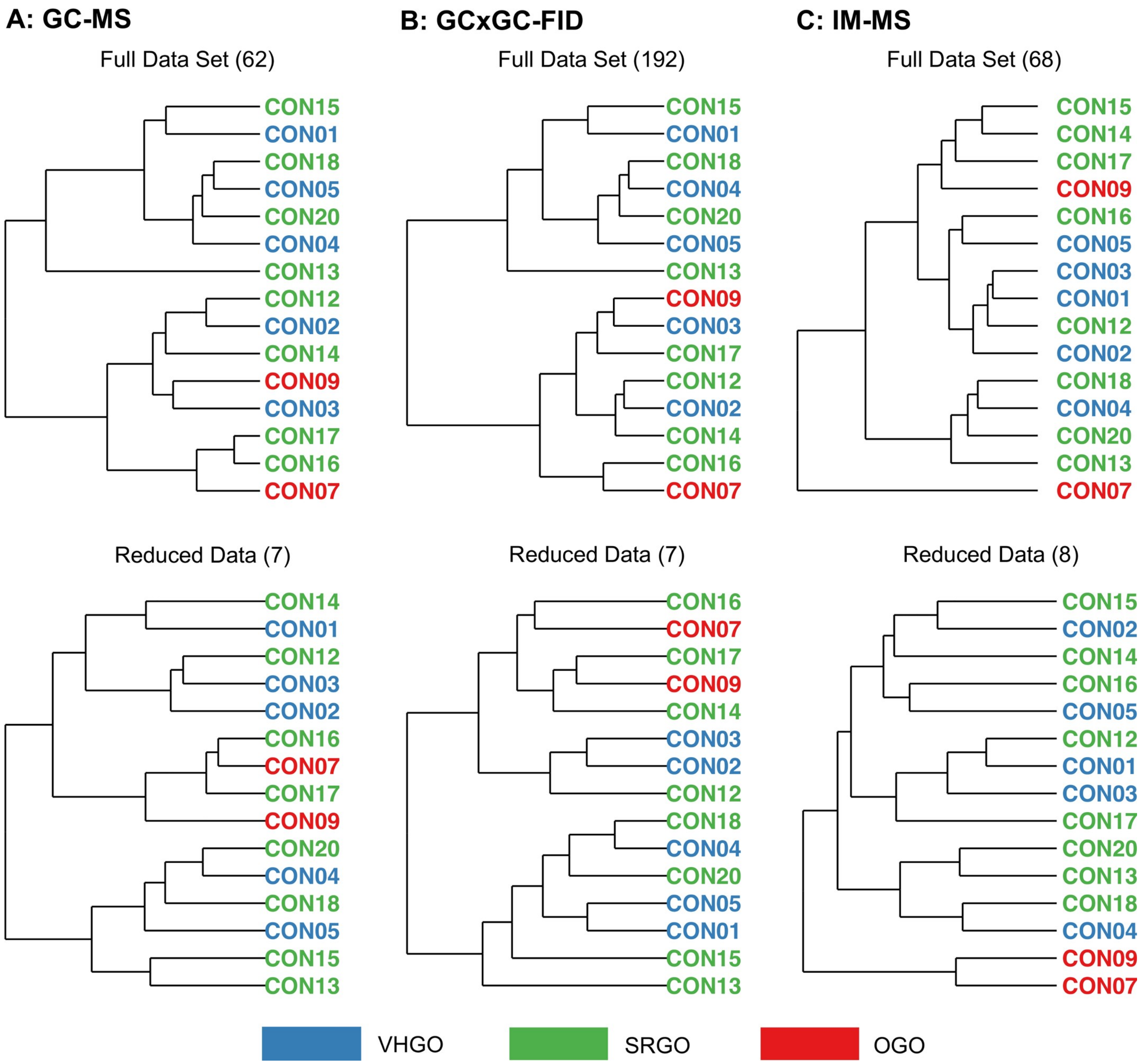


Fig 10. Dendrograms for Petroleum UVCB samples clustering from the reduced data set analyzed using 3 different techniques.

https://doi.org/10.1371/journal.pone.0223517.g010

evaluation to facilitate the interpretation of the complex chemical nature of substances/mixtures. The designed framework consists of two analysis workflows with two different perspectives. In unsupervised analysis workflow, we examined the grouping of the complex substances by using their chemical fingerprints derived from various analytical techniques, and quantitatively compared the grouping hierarchy to a reference categorization through F-M index. In contrast, in a supervised analysis workflow, we benefited from the "read-across" hypothesis, that is similar complex substances that are grouped together based on their chemical structure (i.e. manufacturing category) are prone to behave similarly in terms of environmental health risk assessment. Hence, we can train highly accurate classification models by using the available information on categorization of known complex substances. The generated models can then be used to predict the environmental health impact of future unknown complex substances. The common highlight of both workflows was on the quantitative metrics, which immensely facilitated the comparative assessment of different parameters, such as distinct analytical techniques, data set sizes, or different number of categorization of samples to elucidate the optimal grouping of complex substances. Additionally, we incorporated the ToxPi representation of complex substances with the most informative analytical features to further deliver insights from the developed data-driven classification models.

Our results have shown that in order to assess the statistical significance of grouping results, it is highly important to permute category labels of complex substances and to calculate *p*-value for the obtained results regardless of the selected workflow. In addition, the dimensionality reduction played a key role in reducing the noise in the extracted high-dimensional analytical chemistry profiles. Dimensionality reduction allowed similar or higher grouping quality with significantly reduced number of measurements. The selection of the most informative features further improved data interpretation significantly through advanced data visualization techniques, such as the ToxPi representation. This further facilitated the communication of complex substance characteristics with regulatory decision-makers.

Supporting information

S1 Text. Data source and R Markdown documentation.
(DOCX)

S1 Table. List of selected analytes from the GC-MS data of SRM samples for grouping analysis.
(DOCX)

S2 Table. Number of SRM substances and samples for 3-class, 9-class, and 16-class categorizations.
(DOCX)

S3 Table. List of all chromatographic features and their respective ranks in 3 class grouping analysis.
(DOCX)

S4 Table. List of all chromatographic features and their respective ranks in 9 class grouping analysis.
(DOCX)

S5 Table. List of all chromatographic features and their respective ranks in 16 class grouping analysis.
(DOCX)

S1 Fig. Average confusion matrices of 1000 permutations for SRM sample classification. (A) 3-class grouping, (B) 9-class grouping, and (C) 16-class grouping with 3 replicates.
(DOCX)

S2 Fig. Average confusion matrices of 1000 permutations for SRM sample classification. (A) 3-class grouping, (B) 9-class grouping, and (C) 16-class grouping with 1 replicate.
(DOCX)

S3 Fig. Original and average confusion matrices of 1000 permutations for Petroleum UVCB sample classification. (A) GC-MS, (B) GC×GC-FID, and (C) IM-MS data.
(DOCX)

S4 Fig. Original and average confusion matrices of 1000 permutations for SRM substance classification.
(DOCX)

Acknowledgments

The authors would like to acknowledge Dr. Terry L. Wade for his valuable contributions.-

Author Contributions

Conceptualization: Thomas J. McDonald, Fred A. Wright, Ivan Rusyn.

Data curation: Kyle Ferguson, Thomas J. McDonald, David A. Sheen, Ivan Rusyn.

Formal analysis: Melis Onel.

Funding acquisition: Ivan Rusyn, Efstratios N. Pistikopoulos.

Investigation: Melis Onel, Kyle Ferguson, Weihsueh A. Chiu, Lan Zhou, Fred A. Wright, Ivan Rusyn.

Methodology: Melis Onel, Lan Zhou, Fred A. Wright, Ivan Rusyn, Efstratios N. Pistikopoulos.

Software: Melis Onel, John S. House, Efstratios N. Pistikopoulos.

Supervision: Weihsueh A. Chiu, Thomas J. McDonald, Ivan Rusyn, Efstratios N. Pistikopoulos.

Validation: Ivan Rusyn.

Visualization: Melis Onel, Burcu Beykal, John S. House, Fred A. Wright, Ivan Rusyn.

Writing – original draft: Melis Onel, Burcu Beykal, Ivan Rusyn, Efstratios N. Pistikopoulos.

Writing – review & editing: John S. House, Fred A. Wright, Ivan Rusyn, Efstratios N. Pistikopoulos.

References

1. Clark CR, McKee RH, Freeman JJ, Swick D, Mahagaokar S, Pigram G, et al. A GHS-consistent approach to health hazard classification of petroleum substances, a class of UVCB substances. Regul Toxicol Pharmacol. 2013; 67(3):409–20. https://doi.org/10.1016/j.yrtph.2013.08.020 PMID: 24025648.
2. European Chemicals Agency. Read-Across Assessment Framework (RAAF)—considerations on multi-constituent substances and UVCBs. Helsinki, Finland: European Chemical Agency; 2017.
3. Redman AD, Parkerton TF. Guidance for improving comparability and relevance of oil toxicity tests. Mar Pollut Bull. 2015; 98(1–2):156–70. https://doi.org/10.1016/j.marpolbul.2015.06.053 PMID: 26162510.
4. Gestel CAMv. Mixture toxicity: linking approaches from ecological and human toxicology. Boca Raton: CRC Press; 2011. xxxviii, 280 p. p.
5. Dimitrov SD, Georgieva DG, Pavlov TS, Karakolev YH, Karamertzanis PG, Rasenberg M, et al. UVCB substances: methodology for structural description and application to fate and hazard assessment. Environ Toxicol Chem. 2015; 34(11):2450–62. https://doi.org/10.1002/etc.3100 PMID: 26053589.
6. CONCAWE. REACH–Analytical characterisation of petroleum UVCB substances. Brussels, Belgium: 2012 Contract No.: No. 7/12.
7. Bell M, Blais JM. "-Omics" workflow for paleolimnological and geological archives: A review. Sci Total Environ. 2019; 672:438–55. Epub 2019/04/10. https://doi.org/10.1016/j.scitotenv.2019.03.477 PMID: 30965259.

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. Epub 2014/06/20. https://doi.org/10.1002/mas.21438 PMID: 24942384.

9. Catlin NR, Collins BJ, Auerbach SS, Ferguson SS, Harnly JM, Gennings C, et al. How similar is similar enough? A sufficient similarity case study with Ginkgo biloba extract. Food Chem Toxicol. 2018; 118:328–39. Epub 2018/05/13. https://doi.org/10.1016/j.fct.2018.05.013 PMID: 29752982.

10. Grimm FA, Russell WK, Luo YS, Iwata Y, Chiu WA, Roy T, et al. Grouping of Petroleum Substances as Example UVCBs by Ion Mobility-Mass Spectrometry to Enable Chemical Composition-Based Read-Across. Environmental science & technology. 2017; 51(12):7197–207. https://doi.org/10.1021/acs.est.6b06413 PMID: 28502166.

11. Marshall AG, Rodgers RP. Petroleomics: chemistry of the underworld. Proc Natl Acad Sci U S A. 2008; 105(47):18090–5. https://doi.org/10.1073/pnas.0805069105 PMID: 18836082; PubMed Central PMCID: PMC2587575.

12. Rozett RW, Petersen EM. Methods of factor analysis of mass spectra. Anal Chem. 1975; 47(8):1301–8.

13. Grimm FA, Russell WK, Luo YS, Iwata Y, Chiu WA, Roy T, et al. Grouping of Petroleum Substances as Example UVCBs by Ion Mobility-Mass Spectrometry to Enable Chemical Composition-Based Read-Across. Environmental Science & Technology. 2017; 51(12):7197–207. https://doi.org/10.1021/acs.est.6b06413 WOS:000404087400062. PMID: 28502166

14. 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). 2017; 197:248–58. https://doi.org/10.1016/j.fuel.2017.02.025 PMID: 28603295; PubMed Central PMCID: PMC5464420.

15. Flexer A. Limitations of self-organizing maps for vector quantization and multidimensional scaling. Adv Neur In. 1997; 9:445–51. WOS:A1997BH93C00063.

16. Yin HJ. Connection between self-organizing maps and metric multidimensional scaling. Ieee Ijcnn. 2007:1025–30. https://doi.org/10.1109/Ijcnn.2007.4371099 WOS:000254291100179.

17. Rank J. Classification and risk assessment of chemicals: the case of DEHP in the light of REACH. The Journal of Transdisciplinary Environmental Studies. 2005; 4(3):1–15.

18. L'Yi S, Ko B, Shin D, Cho YJ, Lee J, Kim B, et al. XCluSim: a visual analytics tool for interactively comparing multiple clustering results of bioinformatics data. BMC Bioinformatics. 2015; 16 Suppl 11:S5. https://doi.org/10.1186/1471-2105-16-S11-S5 PMID: 26328893; PubMed Central PMCID: PMC4547151.

19. Filippova D, Gadani A, Kingsford C. Coral: an integrated suite of visualizations for comparing clusterings. BMC Bioinformatics. 2012; 13:276. https://doi.org/10.1186/1471-2105-13-276 PMID: 23102108; PubMed Central PMCID: PMC3576325.

20. Ferguson KC. Characterization of Complex Substances Used in Biological Profiling Through Determination of the Free Concentration Within In Vitro Assays. 2018.

21. Do KT, Wahl S, Raffler J, Molnos S, Laimighofer M, Adamski J, et al. Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. Metabolomics. 2018; 14(10):128. https://doi.org/10.1007/s11306-018-1420-2 PMID: 30830398; PubMed Central PMCID: PMC6153696.

22. PetroOrg Software Tallahassee, FL2014 [02/13/2019]. Available from: http://software.Petroorg.com.

23. Alpaydin E. Introduction to Machine Learning, 3rd Edition. Introduction to Machine Learning, 3rd Edition. 2014:1–613. WOS:000351537500022.

24. Kohonen T. The Self-Organizing Map. Proceedings of the Ieee. 1990; 78(9):1464–80. https://doi.org/10.1109/5.58325 WOS:A1990EC03500004.

25. Vanloan C. Computing the Cs and the Generalized Singular Value Decompositions. Numer Math. 1985; 46(4):479–91. https://doi.org/10.1007/Bf01389653 WOS:A1985ANR4700001.

26. Golub GH, Reinsch C. Singular Value Decomposition and Least Squares Solutions. Numer Math. 1970; 14(5):403–&. https://doi.org/10.1007/Bf02163027 WOS:A1970G000600001.

27. Chipman H, Tibshirani R. Hybrid hierarchical clustering with applications to microarray data. Biostatistics. 2006; 7(2):286–301. https://doi.org/10.1093/biostatistics/kxj007 WOS:000236436300009. PMID: 16301308

28. Xu R, Wunsch D. Survey of clustering algorithms. Ieee T Neural Networ. 2005; 16(3):645–78. https://doi.org/10.1109/Tnn.2005.845141 WOS:000228909900013. PMID: 15940994

29. Fowlkes EB, Mallows CL. A Method for Comparing 2 Hierarchical Clusterings. Journal of the American Statistical Association. 1983; 78(383):553–69. https://doi.org/10.2307/2288117 WOS:A1983RF90800005.

30. Wagner S, and Dorothea Wagner. Comparing clusterings: an overview. Karlsruhe: Universität Karlsruhe, Fakultät für Informatik. 2007.

31. Keasar C, McGuffin LJ, Wallner B, Chopra G, Adhikari B, Bhattacharya D, et al. An analysis and evaluation of the WeFold collaborative for protein structure prediction and its pipelines in CASP11 and CASP12. Sci Rep. 2018; 8(1):9939. https://doi.org/10.1038/s41598-018-26812-8 PMID: 29967418; PubMed Central PMCID: PMC6028396.

32. Kieslich CA, Tamamis P, Guzman YA, Onel M, Floudas CA. Highly Accurate Structure-Based Prediction of HIV-1 Coreceptor Usage Suggests Intermolecular Interactions Driving Tropism. PLoS One. 2016; 11(2):e0148974. https://doi.org/10.1371/journal.pone.0148974 PMID: 26859389; PubMed Central PMCID: PMC4747591.

33. Onel M, Beykal B, Wang MC, Grimm FA, Zhou L, Wright FA, et al. Optimal Chemical Grouping and Sorbent Material Design by Data Analysis, Modeling and Dimensionality Reduction Techniques. Comput-Aided Chem En. 2018; 43:421–6. https://doi.org/10.1016/B978-0-444-64235-6.50076–0 WOS:000441374200076.

34. Onel M, Kieslich CA, Guzman YA, Floudas CA, Pistikopoulos EN. Big Data Approach to Batch Process Monitoring: Simultaneous Fault Detection and Diagnosis Using Nonlinear Support Vector Machine-based Feature Selection. Comput Chem Eng. 2018; 115:46–63. https://doi.org/10.1016/j.compchemeng.2018.03.025 PMID: 30386002; PubMed Central PMCID: PMC6205516.

35. Onel M, Kieslich CA, Guzman YA, Pistikopoulos EN. Simultaneous Fault Detection and Identification in Continuous Processes via nonlinear Support Vector Machine based Feature Selection. Int Symp Process Syst Eng. 2018; 44:2077–82. https://doi.org/10.1016/B978-0-444-64241-7.50341-4 PMID: 30534633; PubMed Central PMCID: PMC6284809.

36. Onel M, Kieslich CA, Pistikopoulos EN. A nonlinear support vector machine-based feature selection approach for fault detection and diagnosis: Application to the Tennessee Eastman process. AIChE Journal. 2019.

37. Beykal B, Boukouvala F, Floudas CA, Pistikopoulos EN. Optimal design of energy systems using constrained grey-box multi-objective optimization. Comput Chem Eng. 2018; 116:488–502. https://doi.org/10.1016/j.compchemeng.2018.02.017 WOS:000448410000032. PMID: 30546167

38. Beykal B, Boukouvala F, Floudas CA, Sorek N, Zalavadia H, Gildin E. Global optimization of grey-box computational systems using surrogate functions and application to highly constrained oil-field operations. Comput Chem Eng. 2018; 114:99–110. https://doi.org/10.1016/j.compchemeng.2018.01.005 WOS:000439701100009.

39. Sorek N, Gildin E, Boukouvala F, Beykal B, Floudas CA. Dimensionality reduction for production optimization using polynomial approximations. Computat Geosci. 2017; 21(2):247–66. https://doi.org/10.1007/s10596-016-9610-3 WOS:000398928300005.

40. Mukherjee R, Onel M, Beykal B, Szafran AT, Stossi F, Mancini MA, et al. Development of the Texas A&M Superfund Research Program Computational Platform for Data Integration, Visualization, and Analysis. Computer Aided Chemical Engineering. 46: Elsevier; 2019. p. 967–72.

41. Breiman L. Random forests. Mach Learn. 2001; 45(1):5–32. https://doi.org/10.1023/A:1010933404324 WOS:000170489900001.

42. Marvel SW, To K, Grimm FA, Wright FA, Rusyn I, Reif DM. ToxPi Graphical User Interface 2.0: Dynamic exploration, visualization, and sharing of integrated data models. BMC Bioinformatics. 2018; 19(1):80. https://doi.org/10.1186/s12859-018-2089-2 PMID: 29506467; PubMed Central PMCID: PMC5838926.

43. Reif DM, Martin MT, Tan SW, Houck KA, Judson RS, Richard AM, et al. Endocrine profiling and prioritization of environmental chemicals using ToxCast data. Environ Health Perspect. 2010; 118(12):1714–20. https://doi.org/10.1289/ehp.1002180 PMID: 20826373; PubMed Central PMCID: PMC3002190.

44. Reif DM, Sypa M, Lock EF, Wright FA, Wilson A, Cathey T, et al. ToxPi GUI: an interactive visualization tool for transparent integration of data from diverse sources of evidence. Bioinformatics. 2013; 29(3):402–3. https://doi.org/10.1093/bioinformatics/bts686 PMID: 23202747; PubMed Central PMCID: PMC3988461.

45. Bro R. PARAFAC. Tutorial and applications. Chemometr Intell Lab. 1997; 38(2):149–71. https://doi.org/10.1016/S0169-7439(97)00032-4 WOS:A1997YH19600005.

46. Harshman RA, Lundy ME. Parafac—Parallel Factor-Analysis. Comput Stat Data An. 1994; 18(1):39–72. https://doi.org/10.1016/0167-9473(94)90132-5 WOS:A1994NY54800004.

47. Stout SA, Wang ZD. Chemical fingerprinting methods and factors affecting petroleum fingerprints in the environment. Standard Handbook Oil Spill Environmental Forensics: Fingerprinting and Source Identification, 2nd Edition. 2016:61–129. https://doi.org/10.1016/B978-0-12-809659-8.00003–6 WOS:000404774100003.