


Article

Integrative Chemical–Biological Grouping of Complex High Production Volume Substances from Lower Olefin Manufacturing Streams

Alexandra C. Cordova ^{1,2}, William D. Klaren ^{1,2,†}, Lucie C. Ford ^{1,2}, Fabian A. Grimm ^{1,2,‡}, Erin S. Baker ³, Yi-Hui Zhou ⁴ , Fred A. Wright ⁴ and Ivan Rusyn ^{1,2,*}

¹ Interdisciplinary Faculty of Toxicology, School of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, TX 77843, USA

² Department of Veterinary Physiology and Pharmacology, School of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, TX 77843, USA

³ Department of Chemistry, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

⁴ Departments of Statistics and Biological Sciences and Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27606, USA

* Correspondence: irusyn@tamu.edu; Tel.: +1-979-458-9866

† Current address: ToxStrategies, 31 College Pl, Asheville, NC 28801, USA.

‡ Current address: Clariant Produkte (Deutschland) GmbH, Arabellastraße 4a, 81925 Frankfurt am Main, Germany.

Abstract: Human cell-based test methods can be used to evaluate potential hazards of mixtures and products of petroleum refining (“unknown or variable composition, complex reaction products, or biological materials” substances, UVCBs). Analyses of bioactivity and detailed chemical characterization of petroleum UVCBs were used separately for grouping these substances; a combination of the approaches has not been undertaken. Therefore, we used a case example of representative high production volume categories of petroleum UVCBs, 25 lower olefin substances from low benzene naphtha and resin oils categories, to determine whether existing manufacturing-based category grouping can be supported. We collected two types of data: nontarget ion mobility spectrometry-mass spectrometry of both neat substances and their organic extracts and in vitro bioactivity of the organic extracts in five human cell types: umbilical vein endothelial cells and induced pluripotent stem cell-derived hepatocytes, endothelial cells, neurons, and cardiomyocytes. We found that while similarity in composition and bioactivity can be observed for some substances, existing categories are largely heterogeneous. Strong relationships between composition and bioactivity were observed, and individual constituents that determine these associations were identified. Overall, this study showed a promising approach that combines chemical composition and bioactivity data to better characterize the variability within manufacturing categories of petroleum UVCBs.

Keywords: UVCB; petroleum; regulatory risk assessment; read-across; ion mobility spectrometry



Citation: Cordova, A.C.; Klaren, W.D.; Ford, L.C.; Grimm, F.A.; Baker, E.S.; Zhou, Y.-H.; Wright, F.A.; Rusyn, I. Integrative Chemical–Biological Grouping of Complex High Production Volume Substances from Lower Olefin Manufacturing Streams. *Toxics* **2023**, *11*, 586. <https://doi.org/10.3390/toxics11070586>

Academic Editor: M. Moiz Mumtaz

Received: 27 May 2023

Revised: 24 June 2023

Accepted: 3 July 2023

Published: 5 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Regulatory agencies commonly categorize chemicals by the amount that is produced and/or imported into a particular jurisdiction; for example, substances whose aggregate quantities exceed some predefined amount per year are considered “high production volumes.” In the European Union, this would entail >1000 tons, and in the United States, the typical cutoff is >1 million pounds (~500 tons). Such substances receive heightened attention in terms of their hazard and risk evaluations and are typically subject to the most extensive testing requirements [1,2]. While most high-production-volume substances are mono-constituent chemicals, a large proportion are derivatives of petroleum refining that belong to a broad class of “unknown or variable composition, complex reaction products, or biological materials” substances (UVCBs). UVCBs from petroleum refining streams

pose unique challenges to regulators, both for registration and for human/ecological safety assessments [3–5]. These UVCBs are produced from crude oil, which is itself a highly complex and variable material; further, these substances are manufactured not to have an exact composition, but to meet technical specifications related to their use [6].

For regulatory registration and safety evaluation, petroleum UVCBs are grouped into categories of similar materials based on rather broad considerations about their composition and manufacturing methods [7]. It is assumed that substances manufactured to similar performance characteristics will have similar toxicological properties. These assumptions were the basis for the industry's voluntary data submissions on the mammalian toxicological hazards of petroleum UVCBs under the US EPA High Production Volume (HPV) Challenge Program in the early 2000s [8]. This program established broad categories of petroleum UVCBs based on physio-chemical properties and refining parameters, such as similar boiling ranges, process histories, or end-use types. However, regulators have been less than satisfied with this approach, especially in Europe, and have invited more informed justifications, such as detailed information on constituents, the extent of compositional variability, and assurances that the material that has been (or will be) used for any additional experiments is representative [9–14].

To overcome the challenges of grouping and read-across of petroleum UVCBs, two approaches have been recently proposed and tested. In the first approach, human cell-based in vitro studies have been conducted on a large number of substances and categories. These studies tested the hypothesis that in vitro biological activity signatures, both phenotypic and gene expression, can be used to support the grouping of UVCBs [15]. As many as 141 petroleum substances from 16 manufacturing categories [6] were tested in a compendium of 15 human cell types representing a variety of tissues [16]; of these, 6 cell types were also profiled for gene expression [17]. Petroleum substances were assayed in dilution series to derive point of departure (POD) estimates for bioactivity in each phenotype. While it was found that bioactivity was strongly correlated with the content of polycyclic aromatic compounds (PAC), the analysis also revealed substantial variability in bioactivity within each category. Some of these data were used in regulatory submissions to request waivers of animal testing requirements. However, the European Chemicals Agency (ECHA) did not accept the data as presented, in part because of the lack of detailed chemical compositional information [10].

Indeed, efforts to provide more detailed compositional characterization constitute a second approach to refining the current read-across of petroleum UVCBs. While there are many analytical methods that have been used to characterize the composition of these substances [7], they are largely insufficient for meeting regulatory requirements [14]. A number of novel ultra-high resolution and multi-dimensional mass spectrometry-based methods have been applied for the analysis of petroleum samples; however, most of these are yet to be adopted by industry or used in regulatory submissions [5]. Further, ultrahigh-resolution instruments and computational methods enabled the confident determination of molecular formulae for a large portion of these constituents in petroleum UVCBs [18,19]. The advantages of these novel techniques, such as ion mobility spectrometry-mass spectrometry (IMS-MS), as complements to traditional gas chromatography-mass spectrometry (GC-MS) have been demonstrated in a number of regulatory contexts—for grouping of crude oils [20] and petroleum UVCBs [21,22], for chemical speciation of oil weathering by-products [23,24], and for characterization of compositional variability of petroleum UVCBs [25].

While both bioactivity and detailed chemical analyses have been used separately to evaluate similarity in petroleum UVCBs, a combination of the approaches has not been undertaken. Inclusion of PAC and other physio-chemical properties together with cell-based bioactivity did show advantages to data interpretation [15–17]; therefore, an investigation of the utility of high-resolution analytical data is also warranted. Herein, a case example of representative high production volume categories of petroleum UVCBs, two lower olefin manufacturing streams, was used to determine whether the existing

grouping of the individual substances into these categories and further into “human health hazard” subcategories as defined under the US EPA HPV Challenge Program can be supported by the data from the new approach methodologies that included probing of both bioactivity and chemical composition. We tested 25 lower olefin substances belonging to the low benzene naphthas (LBN) and Resin Oils and Cyclodiene Dimer Concentrates (RO) categories. We collected two types of data: nontarget high-resolution IMS-MS analyses of each neat substance and their respective dimethyl sulfoxide (DMSO) extract, along with in vitro bioactivity of the DMSO extracts in five different cell types: human umbilical vein endothelial cells (HUVEC), as well as induced pluripotent stem cell (iPSC)-derived hepatocytes, endothelial cells, neurons, and cardiomyocytes. Using these data, we grouped substances and compared the groupings to those in the classes/sub-classes established by the HPV Challenge Program.

2. Experimental Section

2.1. Substances Used in This Study

All lower olefin substances used in this study (assigned number identifiers) and their respective streams are detailed in Table 1. In total, 25 neat substances (identified as 13 resin oils and 12 low benzene naphthas) were included in the analyses and were donated by member companies of the American Chemistry Council’s (ACC) Olefins Panel. Both the identity and origin of the individual substances were de-identified beyond each substance’s Chemical Abstract Service (CAS) number and manufacturing stream name. Select samples were categorized into different human health subcategories than originally proposed under the HPV Challenge Program based on the expert judgement of the authors and the information provided by the manufacturers. Table S1 details our reasoning for group assignments.

Table 1. Petroleum UVCBs from lower olefin categories that were tested in this study.

Sample ID *	Sponsored Stream *	CAS RN #	Human Health Hazard Subcategory #
Low Benzene Naphthas			
83757 83806	Pyrolysis C7	68527-23-1 68478-10-1	Group I: High Toluene Streams
83946	Pyrolysis C7-C8	68527-23-1 68919-15-3	
84070 84003	Hydrotreated C7-C8		
84075	Hydrotreated C7+	64742-48-9	Group II: Mixed Aromatics Streams
84068	Hydrotreated C8-C10	68512-78-7 64742-48-9	
83979 84024	Hydrotreated C7-C12	64742-48-9 68516-20-1	
83931	Solvent Naphtha	68512-78-7	
83984 83683	Pyrolysis C7-C12	68516-20-1 64742-83-2 68746-45-9	Group III: Pyrolysis C7-C12
83758	C9+ from o-xylene	68333-88-0 68553-14-0	Group V: C9+ from o-xylene unit
84082	Solvent Naphtha	68512-78-7	Not Defined Properly

Table 1. Cont.

Sample ID *	Sponsored Stream *	CAS RN #	Human Health Hazard Subcategory #
Resin Oils and Cyclodiene Dimer Concentrates			
83981	DCPD, High Purity	77-73-6	Group I: DCPD High Purity & Related Streams
84023	High DCPD Resin Oil	68477-54-3 68477-40-7	
83955	Distillates (petroleum), steam cracked. C8-C12; High DCPD Resin Oil	68477-54-3	
83956	Resins Distillates (petroleum), cracked stripped steam cracked. C10-C12	68477-40-7	
84543	HYDROCARBONS, C5-RICH, DICYCLOPENTADINE Resin; DCPD Concentrate Distillates (Petroleum) steam cracked C5-C12 fraction	68527-24-2	
83949	Low DCPD Resin Oil	68477-54-3 68516-20-1	Group II: Low DCPD Resin Oil & Resin Former
83980	Low DCPD Resin Oil	68477-54-3	
84012		68516-20-1	
84074		68516-20-1	
83618	Dicyclopentadiene Resin Grade (3a,4,7,7a-tetrahydro-4,7-methano-1H-indene/Alkenes, C9-11, C10-rich)	2647-00-4	Group III: MCPD Dimer
83985	Resin Feed (Distillates (petroleum), steam-cracked, C8-12 fraction/C9 mixture rich in indene and vinyltoluene/Complex mixture of (mainly aromatic) C9–C10 hydrocarbons); Dicyclopentadiene Resin	68477-54-3	Not Defined Properly
83879	Resin Distillates, steam cracked. C8-C12 (Extract residue (coal), light oil, alk, acid ext, indene fraction	68477-54-3	
83998	Resin—Distillate cracked, ethylene manufacturing by-product, C9-C10		

* Sample IDs and sponsored stream information as provided by the ACC Olefins Panel. # CAS RN and human health hazard subcategory assignments were made by the authors by matching the stream names, as provided by the sponsor, to the information in the US EPA Screening-Level Hazard Characterization documents for Low Benzene Naphthas, Resin Oils, and Cyclodiene Dimer Concentrates categories [26,27].

Samples were stored at -80°C until analyzed or otherwise processed. From each substance, an organic extract was prepared using DMSO and cyclohexane, a method that preferentially extracts PAC from petroleum-containing samples, according to the standard American Society for Testing and Materials (ASTM) IP 346 method [28]. Briefly, 4 g of each substance was dissolved in 10 mL cyclohexane. The cyclohexane fraction was then extracted twice with 10 mL pre-equilibrated 10:1 DMSO/cyclohexane. The two subsequent DMSO fractions were collected in a 20 mL glass vial and stored at -80°C until used in the experiments. It is important to note that throughout this study, the substances are referred to by a five-digit ID (e.g., 84070) prefixed by either “N” representing a “neat” substance, or “E” representing its DMSO extract.

2.2. IMS-MS Analysis of Neat Substances and DMSO Extracts

All substances were analyzed using an ion mobility spectrometry (IMS) instrument coupled to a quadrupole time-of-flight (QTOF) mass spectrometer (MS) (model G6560A, Agilent Technologies, Santa Clara, CA, USA). Neat and extracted samples were prepared for IMS-MS analysis as follows. A glass syringe was first used to add 100 μL of each sample to a glass vial. Substances were then diluted 3 \times by adding 200 μL of 50:50 acetonitrile/toluene buffer and vortexing. The glass syringe was rinsed in triplicate with acetone, hexane, and methanol between the preparation of each sample. All samples were analyzed using an atmospheric pressure photoionization (APPI) source in positive ion mode and were injected at a flow rate of 50 $\mu\text{L}/\text{min}$. The appropriate tune mix was used to calibrate the instrument prior to sample runs, and samples were collected for 1.5 min each. Washes with acetone and methanol were conducted at least three times between samples. Other instrument parameters were consistent with prior studies examining similar substances using an APPI ion source in positive mode [29].

Upon acquisition, IMS-MS raw data files for neat substances and corresponding extracts were first calibrated in IMS-MS Browser B.08.00 software (Agilent Technologies, Santa Clara, CA, USA) using the tune mix file obtained prior to the sample run. The tune mix file was verified to have mass accuracies within ± 5 ppm m/z for each calibrant peak. Calibrated files for neat substances and extracts were then processed using Agilent Mass Profiler software to obtain two separate sets of detected compounds, or “features”, and their abundances in each sample. A library of compounds was then used to match identities to detected features based on m/z and collisional cross section ($^{DT}CCS_{N_2}$) values for each compound [30]. $^{DT}CCS_{N_2}$ values are a quantitative representation of the size and shape of individual features, derived from the drift time (DT) of each feature [31–33]. $^{DT}CCS_{N_2}$ is unique to each detected species and can be used to identify targeted species within a nontarget dataset [18,32]. Datasets for neat substances and extracts, including library-matched anchor features, were then exported from MassProfiler for chemical characterization. Raw IMS-MS data files for neat samples and extracts can be found in Tables S2 and S3, respectively.

Chemical characterization was conducted following a modified workflow detailed previously [18]. In brief, datasets were first processed to only include features at an abundance ≥ 5000 in at least one sample to minimize unnecessary amplification of noise. Filtered data matrices can be found in Tables S4 and S5 for neat samples and extracts, respectively. Anchor features were then manually verified using the $^{DT}CCS_{N_2}$ library to ensure m/z fell within a range of ± 5 ppm and $\pm mDa$ and $^{DT}CCS_{N_2}$ values fell within a range of $\pm 1\%$. Kendrick mass defect (KMD) was then calculated in the context of CH_2 functional units to enable feature organization in homologous series and molecular formula identification of hydrocarbon species. The series were then validated using KMD-H homologous series and $^{DT}CCS_{N_2}$ values [18]. Once a maximum number of features were characterized with confidence, double bond equivalence (DBE) for individual features was determined based on assigned molecular formulas as follows [34,35]:

$$DBE = \#C + 1 - (\#H/2) + (\#N/2) \quad (1)$$

Feature abundances that appear in terms of % Total Abundance throughout this publication were calculated by normalization to the sum of abundances of all filtered features (Abundance > 5000). Data matrices with molecular formulas and DBE assignments can be found in Tables S6 and S7 for neat samples and extracts.

2.3. In Vitro Bioactivity Experiments

In total, five organotypic human cell types were used to conduct bioactivity experiments. Organotypic cell types derived from induced pluripotent stem cells (iPSC) were acquired from FUJIFILM-Cellular Dynamics (Madison, WI, USA) and included cardiomyocytes (Cat #R1007, Lot 1299716), endothelial cells (Cat #R1022, Lot 1833921), hepatocytes (Cat #R1027, Lot 7000716), and neurons (Cat #R1013, Lot 1227535). In addition, primary human umbilical vein endothelial cells (HUVEC; Cat #C2519A, Lots 0000433795 and 0000460587, Lonza, Basel, Switzerland). These cell types were selected based on previous studies with petroleum UVCBs that showed them to be most informative for grouping [15–17]. All cells were cultured and prepared for treatment based on modified manufacturer protocols (Cellular Dynamics and Lonza) as detailed elsewhere [15,36–42].

All in vitro experiments were conducted by first preparing a chemical stock plate containing extracts of each substance and all controls (except assay-specific positive controls) in 100% DMSO in a 384-well plate. The compounds in the chemical stock plate were then serially diluted in appropriate cell-specific culture media into working plates at $5\times$ or $2\times$ the desired extract concentration for testing in each cell-specific assay plate. Working plates contained extracts with 2% or 1% DMSO for further dilution to 0.5% or 0.25% (for neurons) DMSO in all assay plates. Thus, in the assay plates, each cell type was exposed to the extracts across five final concentrations: 500 $\mu g/mL$, 50 $\mu g/mL$, 5 $\mu g/mL$, 0.5 $\mu g/mL$, and 0.05 $\mu g/mL$ for neurons (in 0.25% DMSO), or 1000 $\mu g/mL$, 100 $\mu g/mL$,

10 µg/mL, 1 µg/mL, and 0.1 µg/mL for all other cell types (in 0.5% DMSO). Cell-specific exposure times, controls, phenotypes, and endpoints measured are detailed in Tables S8 and S9. The “method blank” vehicle control [16] was DMSO that was carried through the IP 346 extraction procedure without the inclusion of a petroleum substance.

The experimental design consisted of running a singleton of all the test substance extracts on a single 384-well plate (using only the inner 308 wells) with full concentration response. The inter- and intra-plate controls were included to ensure that the concentration responses observed were not artefacts of the experimental design. Inter-plate controls consisted of running each plate twice; this allowed for a duplicate to be obtained of all substance extracts but also ensured reproducibility between plates. Intra-plate controls were added to ensure that the single values were consistent within a plate. Two olefin substance extracts were selected at random to be present a second time on each plate in a full concentration response representation.

Raw data generated during the in vitro assays was normalized to method blank vehicle control values. The normalized values represent a percent response to the method blank. The normalization was performed for all raw values, including positive/negative controls, using the formula:

$$\text{Normalized Value} = \left(\frac{\text{Raw Value}}{\text{Average of Method Blank Wells}} \right) \times 100 \quad (2)$$

To ensure the integrity of the data, several aspects were assessed for each endpoint (data not shown). First, vehicle effects were determined by comparing method blank vehicle, DMSO, and media wells to ensure no effect of the vehicle. The positive cytotoxic control, tetraoctyl ammonium bromide, was also evaluated on all cells. Second, cell type and assay specific positive controls were examined for concentration response with a nonlinear line fit (Hill function) to ensure that the cells were performing as expected from previous publications elsewhere [15,37–42]. Third, inter-plate replicate controls were plotted as a scatterplot, with one replicated as the x-value and the other replicated as the y-value. Pearson’s *r* and Spearman’s *ρ* correlations were calculated, along with *p*-values of significance, and experiments were deemed reproducible if correlations were significant and >0.8. Lastly, intra-plate replicates were plotted as concentration responses with a nonlinear fit (Hill function) to determine if outliers were present.

Upon quality control evaluation, concentration-response data for each endpoint were analyzed to obtain corresponding PODs. Concentration-response data were first normalized to the average of all vehicle treatments (100%). For most of the cell types and phenotypes, a POD was defined as the point where a logistically fitted line departed 10% from the mean of the vehicle control values (EC₁₀). Previous investigations have used this POD [39]. Cell- and phenotype-specific PODs are shown in Table S8.

Biological PODs were then analyzed using the Toxicological Prioritization Index (ToxPi) software to generate ToxPi scores [43,44]. First, individual ToxPis were generated for each cell type, with each slice representing a phenotype and equally weighted depending on the number of phenotypes tested per cell type (Table S8). The contribution of each POD element to the ToxPi scores was scaled from lowest bioactivity (ToxPi element = 0) to highest bioactivity (ToxPi element = 1) using the formula:

$$\text{ToxPi Value} = 1 - \frac{\log_{10}(\text{POD}) - \log_{10}(\text{POD}_{\min})}{\log_{10}(\text{POD}_{\max}) - \log_{10}(\text{POD}_{\min})} \quad (3)$$

Total ToxPi scores for each cell type were then represented in a separate analysis as individual slices to generate an overall ToxPi depicting all cell types. All substances were included as “available chemicals” in the software settings, and each cell type tested was displayed as an individual pie slice. The distribution for each slice was log-scaled and equally weighted in its contribution to the overall ToxPi.

2.4. Clustering of Substances Using IMS-MS and Bioactivity Data

Grouping of LBN and RO categories as well as human health subcategories for both biological and chemical data was conducted using unsupervised hierarchical clustering via *hclustfunc* in *heatmaply* and *gplots* packages in *RStudio*.

2.5. Predicting Bioactivity Based on IMS-MS Chemical Profiles

For prediction of the bioactivity from the individual chemical features in neat or extracted samples, an extension of the penalized ridge regression approach as developed in [45] was used. Briefly, the approach performs multivariate ridge regression for the multivariate linear model $Y = XB + \text{error}$, where Y ($n \times m$) and X ($n \times p$) are scaled bioactivity and feature matrices with dimensions shown, and B ($p \times n$) is a coefficient matrix. Here, n is the sample size of substances, m is the number of bioactivity measurements, and p is the number of features used in the predictions. Briefly, one can envision the bioactivity data as a multi-dimensional readout Y with n rows, where each row included data for one endpoint, cell-specific overall ToxPi scores, or an overall ToxPi score incorporating all cell types together. The matrix had 25 columns for each chemical, classified by their category. A chemical predictor matrix for neat substances X had 225 rows (features comprising >1% of at least one sample) and 25 columns (one per sample). Similarly, a separate chemical predictor matrix for DMSO extracts had 212 rows (features > 1%) and 25 columns. Prior to fitting, all data columns were centered and scaled to unit variance for comparability and to ensure no predictor dominated simply due to scale differences.

The fitted model is truly multivariate because a single tuning penalty λ is applied, with $\hat{B} = (X^T X + \lambda I)^{-1} (X^T Y)$ (which is the ridge regression approach) and final prediction $\hat{Y} = X \hat{B}$. λ was evaluated on a grid such that $\log_{10}(\lambda)$ varied uniformly from -1.0 to 6.0 in increments of 0.1 . Evaluations were performed using leave-one-out cross validation, i.e., prediction for elements of Y from the i^{th} sample used coefficients obtained after removing the i^{th} sample, to avoid overfitting. The selection of the tuning parameter was performed to give the minimum mean squared prediction error. Final predictions were returned to the original Y scale by multiplying each column by the original standard deviation and adding the original mean. The entire procedure was then run again to predict features by reversing the assignment of X and Y matrices.

As a measure of model fit for each bioactivity feature, the Pearson correlation r between the observed bioactivity values and the values predicted in cross-validation was used. Standard cross-validation principles [46] rely on the fact that the test sample (which is singular under leave-one-out cross-validation) is held out for model training, and thus each test set prediction is often treated as independent of the training set. However, a subtle internal dependency can arise due to the scaling of X and Y , which is performed once. In addition, our final prediction tuning parameter was selected once, outside of the cross-validation loop. Thus, as a conservative measure without requiring complicated double cross-validation loops, p -values for the predicted-observed r using a permutation procedure were computed. A total of 1000 permutations of the sample indices in Y and X were performed, with the mean and standard deviations of the (null) r values used to compute a statistic $z = (r - E(r))/SD(r)$, which was compared to a standard normal distribution in a two-sided test. The resulting p -values for each bioactivity feature were then corrected for multiple comparisons by computing the Benjamini–Hochberg q -value [47] using the *R* v4.1 *p.adjust* package.

3. Results and Discussion

The overall experimental workflow is shown in Figure 1. Both neat substances (two manufacturing categories, 25 substances in total, Table 1) and their respective DMSO extracts were analyzed using nontarget IMS-MS. DMSO extracts of each test substance were used for in vitro assays across four induced pluripotent stem cell-derived cell types (cardiomyocytes, endothelial cells, hepatocytes, and neurons) and human umbilical vein endothelial cells (HUVEC).

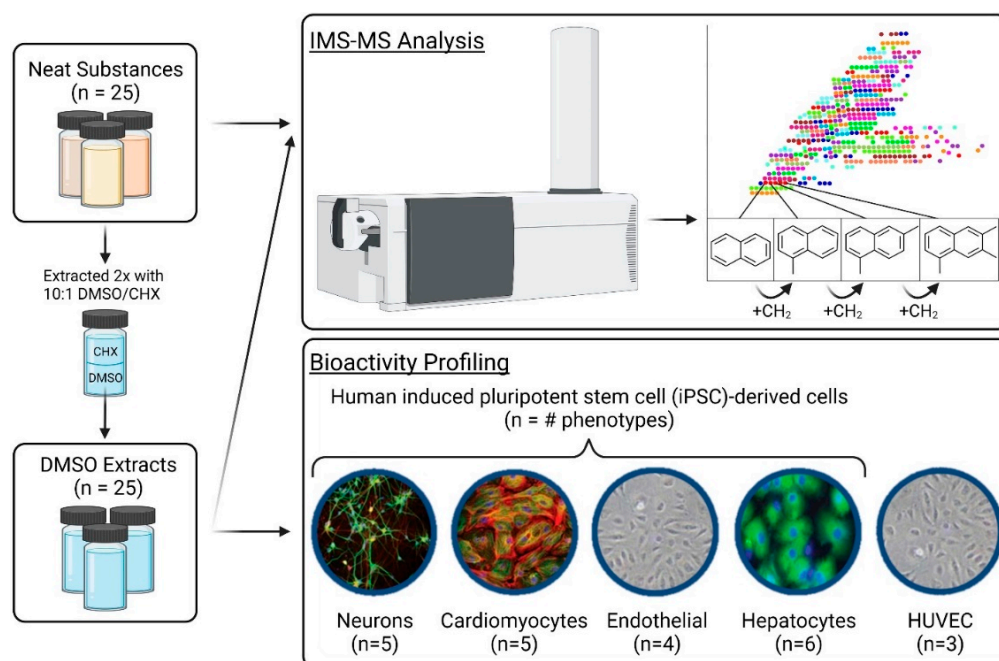


Figure 1. Schematic of the overall experimental design depicting the chemical analytical and bioactivity profiling of neat substances and their respective DMSO extracts ($n = 25$).

3.1. Compositional Characterization and Similarity between Test Substances

Regulatory guidelines require compositional characterization and assessment of the variability between substances to (1) determine the applicability domain of a category, (2) confirm membership in that category, and (3) establish a basis for read-across of toxicological properties [14]. To fulfil these criteria for the substances tested herein, the chemical profiles obtained with IMS-MS nontarget analysis were first analyzed separately within the LBN and RO categories (Figures 2 and 3). Figure 2A shows the profiles of the substances originally identified as belonging to the LBN category, both in terms of the raw abundance of various constituents and as a percentage of the total abundance within each sample. A complete list of sponsored streams is available in the US EPA Screening-Level Hazard Characterization for LBN (see access links to the documents in Table S10) [26]. According to US EPA [26], the LBN category comprises 12 unique chemical identifiers and 9 production streams; in this study, substances were available for experiments that represented 10 identifiers and 8 production streams.

First, raw abundance profiles showed the complexity of the substances and the variability in their composition within and across human health subcategories (Table 1). The substances belonging to subcategory I, high toluene streams, were the least complex of the LBN substances tested in terms of the overall raw abundance of the constituents. This was expected, because substances belonging to this subcategory should be composed of C7–C8 range constituents, while the LBN category as a whole consists “*primarily of C7 to C12 aromatic and cycloaliphatic hydrocarbons*” [26]. Similar observations were made when the data were expressed in percent abundance, although abundance normalization demonstrates a more homogeneous LBN category than raw abundances (Figure 2A, bottom). The composition of DMSO extracts had little overlap with the corresponding neat products, both in terms of raw and normalized abundance. Figure 2B shows hierarchical clustering of the samples using analytical data. It is evident that while some substances from the same human health subcategory cluster together, others are not sufficiently similar using the chemical compositional profiles from IMS-MS analyses.

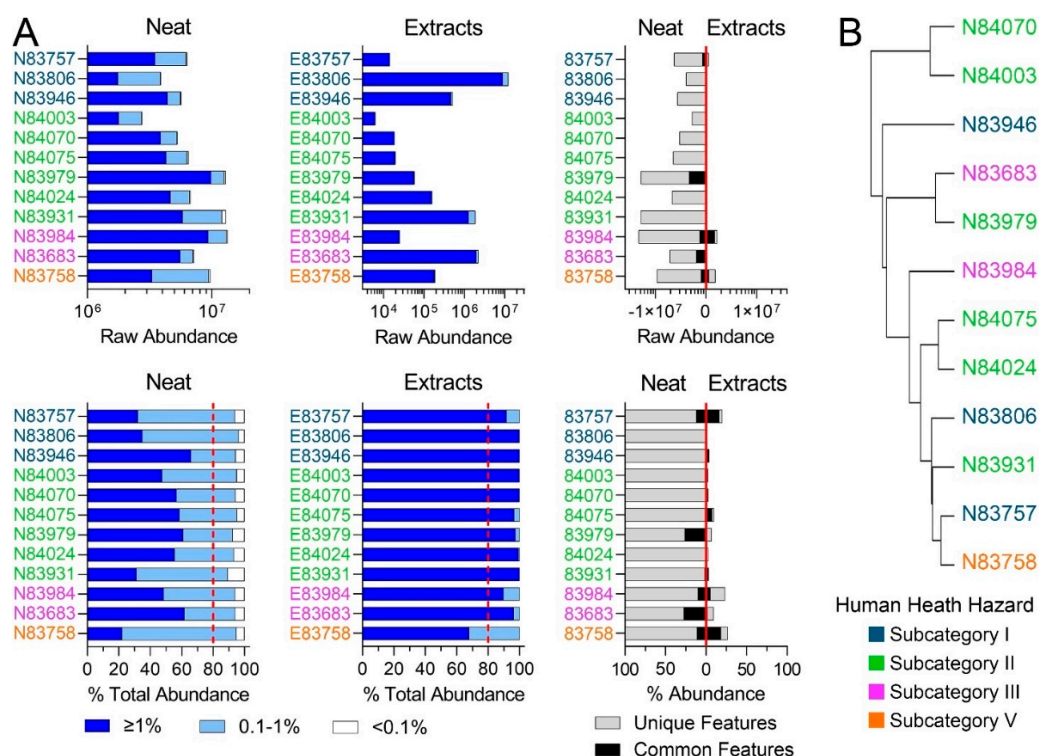


Figure 2. (A) Feature abundances for Low Benzene Naphthas category. The top row depicts the raw abundances of features detected for neat products, product extracts, and the abundance of features characterized by the same molecular formula in neat and corresponding extract substances. The bottom row depicts the same features normalized to the total abundance of features per substance. Dark blue bars denote features present at $\geq 1\%$, light blue bars denote features present at 0.1–1%, and white bars denote features present at $<0.1\%$ abundance. Dotted red lines refer to ECHA's 80% minimum threshold [14] for UVCB characterization. The third plot in each row shows features present in both the neat samples and DMSO extracts (black bars) and features unique to each (grey bars). (B) Hierarchical clustering portraying the chemical similarity of LBN neat substances based on IMS-MS profiles. Substances closer together have the most similar chemical profiles. Colors indicate pre-assigned health hazard groups.

Similar observations were made for RO substances (Figure 3A). A complete list of sponsored streams is available in the US EPA Screening-Level Hazard Characterization for RO (see access links to the documents in Table S10) [27]. The US EPA specified that this category includes 11 unique chemical identifiers in 9 production streams; herein, we tested substances from 6 identifiers and 5 production streams. Three RO substances that were available for this study could not be defined into one of the existing subcategories. Raw abundance profiles again demonstrated the variation in chemical composition among substances. Subcategory I exhibited the most variation, while subcategory II exhibited the most similar substance profiles. This was supported by hierarchical clustering (Figure 3B). Corresponding DMSO extracts showed the variation between substances and, although to a greater extent than for LBN samples, still captured a very small fraction of the corresponding neat substances (Figure 3A).

Second, the most recent regulatory guidance on substance chemical characterization [14] details the extent of information needed for UVCBs, including constituent identities and concentrations. Compounds present at $\geq 1\%$ abundance must comprise at least 80% of the sample to warrant more extensive characterization of molecular structures for hazard evaluation. For cases where the 80% threshold is not met, it is “not technically possible or impractical” to identify the individual constituents, and “structural similarity must be demonstrated by other means.” “Other means” may include pre-existing information on start-

ing materials and manufacturing processes or fingerprinting analysis; however, analytical methods must enable “the provision of information on a sufficient proportion of constituents . . . [to cover] >95% of the constituents of a substance” [14]. Thus, for analyses in Figures 2 and 3, constituents were classified as comprising $\geq 1\%$, $0.1\text{--}1\%$, and $<0.1\%$ of a sample for all LBN and RO neat products and extracts. For both categories, features of $\geq 1\%$ abundance in the neat substances did not meet the ECHA’s 80% threshold, meaning that the use of “other means” to characterize the composition of the neat substances may be justified. However, for toxicity testing, it is equally important to characterize the DMSO extracts used to expose the substances. Features of $\geq 1\%$ abundance in the extracts constituted >80% of each substance, meaning that constituents of concern at concentrations below 0.1% may also need to be identified using additional analytical techniques. Without analytical reference standards to confirm the structural identities of these low-concentration constituents, analyses herein were restricted to putative molecular formulae. Ultra-high-resolution techniques or structure-based modeling approaches may be better suited to confirm the identities of these constituents. Still, the number of species present in each sample $<0.1\%$ is vast, and structural identification of all constituents of concern would be a daunting task.

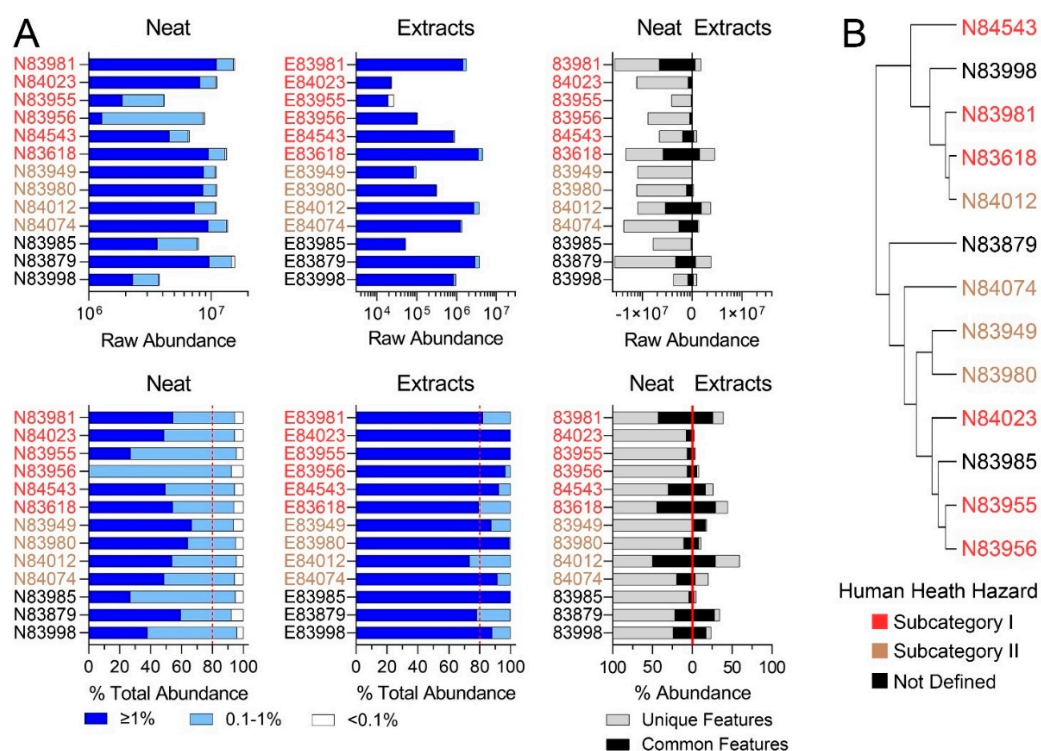


Figure 3. (A) Feature abundances for the Resin Oils category. The top row depicts the raw abundances of features detected for neat products, product extracts, and the abundance of features characterized by the same molecular formula in neat and corresponding extract substances. The bottom row depicts the same features normalized to the total abundance of features per substance. Dark blue bars denote features present at $>1\%$, light blue bars denote features present at $0.1\text{--}1\%$, and white bars denote features present at $<0.1\%$ abundance. Dotted red lines refer to ECHA’s 80% minimum threshold [14] for UVCB characterization. The third plot in each row shows features present in both the neat samples and DMSO extracts (black bars) and features unique to each (grey bars). (B) Hierarchical clustering portraying the chemical similarity of RO neat substances based on IMS-MS profiles. Substances closer together have the most similar chemical profiles. Colors indicate pre-assigned health hazard groups.

Third, the composition of the constituents obtained using nontarget IMS-MS was compared to the typical constituents reported in REACH Category Identity Profiles [48,49], information that is derived using traditional analytical methods (Tables 2 and 3). These data are typically reported for a limited number of the most abundant constituents, which are

known to vary among registered substances. The reported LBN constituent list [48] includes 45 compounds with CAS numbers mapping to 18 unique molecular formulas (Table 2). Figure 4 shows the raw and relative abundances for IMS-MS-observed constituents that matched these formulas. As expected, their abundance varied among substances within each category. We compared the reported typical ranges with those from IMS-MS analyses (Table 2). Even though the data was obtained using different analytical methods and on different samples, we reason that by normalizing abundances as a percent of the total sample, it is possible to perform meaningful comparisons. Overall, IMS-MS data were well within the typical range for all constituents, with the maximum observed concentration for any single constituent being toluene at 6.7%. Still, because the IMS-MS approach provides higher resolution and more individual constituents are detected, the relative abundances were lower than those typically reported using other techniques. Seven molecular formulae spanning 10 CAS numbers were below the limit of detection.

Table 2. Typical (as defined in [48]) versus observed (this study) constituents for substances in the Low Benzene Naphthas category.

Constituent	CAS RN	Formula	Typical Concentration (%)	Typical Concentration Range (%)	Observed (IMS-MS) Range (%)
Toluene	108-88-3	C ₇ H ₈	~30	0–≤50	0–0.08
Ethylbenzene	100-41-4	C ₈ H ₁₀	~20	0–≤45	0–0.2
Xylenes	1330-20-7	C ₈ H ₁₀	~15	0–≤30	0–0.2
m-Xylene	108-38-3	C ₈ H ₁₀	~7	0–≤15	0–0.2
p-Xylene	106-42-3	C ₈ H ₁₀	~5	0–≤10	0–0.2
o-Xylene	95-47-6	C ₈ H ₁₀	~2.5	0–≤5	0–0.2
Ethyltoluene	25550-14-5	C ₉ H ₁₂	~20	0–≤45	0.01–0.21
1,2,4-Trimethylbenzene	95-63-6	C ₉ H ₁₂	~12	0–≤21	0.01–0.21
Propylbenzene	103-65-1	C ₉ H ₁₂	~8	0–≤15	0.01–0.21
1,2,3-Trimethylbenzene	526-73-8	C ₉ H ₁₂	~6	0–≤12	0.01–0.21
Isopropylbenzene	98-82-8	C ₉ H ₁₂	~1.5	0–≤9	0.01–0.21
3-Ethyltoluene	620-14-4	C ₉ H ₁₂	~3	0–≤5	0.01–0.21
4-Ethyltoluene	622-96-8	C ₉ H ₁₂	~1	0–≤2	0.01–0.21
1,3,5-Trimethylbenzene	108-67-8	C ₉ H ₁₂	~1	0–≤2	0.01–0.21
Indene	95-13-6	C ₉ H ₈	~15	0–≤40	0.22–2.6
Methylstyrene	1319-73-9	C ₉ H ₁₀	~5	0–≤36	0.05–1.3
Indane	496-11-7	C ₉ H ₁₀	~7	0–≤13	0.05–1.3
2,3,3a,4,7,7a-Hexahydro-4,7-methano-1H-indene	19398-83-5	C ₁₀ H ₁₄	~22	0–≤30	0–0.49
Dihydrodicyclopentadiene	4488-57-7	C ₁₀ H ₁₄	~15	0–≤25	0–0.5
1,2,3,5-Tetramethylbenzene	527-53-7	C ₁₀ H ₁₄	~8	0–≤16	0–0.5
1,2,4,5-Tetramethylbenzene	95-93-2	C ₁₀ H ₁₄	~6	0–≤11	0–0.5
1,2-Dimethyl-4-ethylbenzene	934-80-5	C ₁₀ H ₁₄	~5	0–≤11	0–0.5
1,3-Dimethyl-4-ethylbenzene	874-41-9	C ₁₀ H ₁₄	~2	0–≤4	0–0.5
1,4-Dimethyl-2-ethylbenzene	1758-88-9	C ₁₀ H ₁₄	~2	0–≤3	0–0.5
2-Methyl-2-butene	513-35-9	C ₅ H ₁₀	~7	0–≤14	0–1.32
Cyclopentane	287-92-3	C ₅ H ₁₀	~6	0–≤11	0–1.32
Trans-2-pentene	646-04-8	C ₅ H ₁₀	~5	0–≤10	0–1.32
Cis-2-pentene	627-20-3	C ₅ H ₁₀	~2	0–≤3	0–1.32

Table 2. Cont.

Constituent	CAS RN	Formula	Typical Concentration (%)	Typical Concentration Range (%)	Observed (IMS-MS) Range (%)
Naphthalene	91-20-3	C ₁₀ H ₈	~6	0–≤12	0.41–6.7
Tetralin	119-64-2	C ₁₀ H ₁₂	~3	0–≤6	0.05–1.4
Dicyclopentadiene	77-73-6	C ₁₀ H ₁₂	~15	0–≤2	0.05–1.4
Styrene	100-42-5	C ₈ H ₈	~2	0–≤5	0.01–0.26
1-Methylnaphthalene	90-12-0	C ₁₁ H ₁₀	~1	0–≤2	0.7–2.7
Isomer of Methylindene	N/A	N/A	~13	0–≤36	n.d.
C-10 Aromatic	N/A	N/A	~13	0–≤36	n.d.
N-pentane	109-66-0	C ₅ H ₁₂	~16	0–≤31	n.d.
Isopentane	78-78-4	C ₅ H ₁₂	~9	0–≤17	n.d.
Methylcyclohexane	108-87-2	C ₇ H ₁₄	~14	0–≤27	n.d.
Ethylcyclopentane	1640-89-7	C ₇ H ₁₄	~12	0–≤23	n.d.
Cis-1,2-dimethylcyclopentane	1192-18-3	C ₇ H ₁₄	~2	0–≤3	n.d.
Cyclopentene	142-29-0	C ₅ H ₈	~7	0–≤14	n.d.
N-heptane	142-82-5	C ₇ H ₁₆	~7	0–≤14	n.d.
Tetrahydrodicyclopentadiene	6004-38-2	C ₁₀ H ₁₆	~5	0–≤10	n.d.
N-octane	111-65-9	C ₈ H ₁₈	~4	0–≤7	n.d.
Benzene	71-43-2	C ₆ H ₆	~0	<0.1	n.d.

Table 3. Typical (as defined in [49]) versus observed (this study) constituents for substances in the Resin Oils category.

Constituent	CAS RN	Formula	Typical Concentration (%)	Typical Concentration Range (%)	Observed (IMS-MS) Range (%)
DCPD	77-73-6	C ₁₀ H ₁₂	~40	0–≤80	0.02–0.55
Vinyltoluene	25013-15-4	C ₉ H ₁₀	~30	0–≤60	0.03–0.37
4-Methylstyrene	622-97-9	C ₉ H ₁₀	~20	0–≤40	0.03–0.37
Indan	496-11-7	C ₉ H ₁₀	~7.5	0–≤25	0.03–0.37
2-Phenylpropene	98-83-9	C ₉ H ₁₀	~5	0–≤20	0.03–0.37
3-Methylstyrene	100-80-1	C ₉ H ₁₀	~10	0–≤20	0.03–0.37
2-Methylstyrene	611-15-4	C ₉ H ₁₀	~7.5	0–≤15	0.03–0.37
Cyclopentane	287-92-3	C ₅ H ₁₀	~25	0–≤50	0–0
2-Methylbut-2-ene	513-35-9	C ₅ H ₁₀	~5	0–≤10	0–0
Ethyltoluene	25550-14-5	C ₉ H ₁₂	~20	0–≤40	0–0.19
Trimethylbenzenes (TMB)	25551-13-7	C ₉ H ₁₂	~20	0–≤40	0–0.19
Isopropylbenzene	98-82-8	C ₉ H ₁₂	~15	0–≤30	0–0.19
1,2,4-Trimethylbenzene	95-63-6	C ₉ H ₁₂	~7.5	0–≤15	0–0.19
m-Ethyltoluene	620-14-4	C ₉ H ₁₂	~5	0–≤13	0–0.19
1,3,5-Trimethylbenzene	108-67-8	C ₉ H ₁₂	~5	0–≤10	0–0.19
Propylbenzene	103-65-1	C ₉ H ₁₂	~5	0–≤10	0–0.19
4,7-Methano-1H-indene, 2,3,3a,4,7,7a-hexahydro-	19398-83-5	C ₁₀ H ₁₄	~10	0–≤20	0–0.2
Dihydrodicyclopentadiene	4488-57-7	C ₁₀ H ₁₄	~5	0–≤12	0–0.2
1,2,4,5-Tetramethylbenzene	95-93-2	C ₁₀ H ₁₄	~5	0–≤10	0–0.2

Table 3. Cont.

Constituent	CAS RN	Formula	Typical Concentration (%)	Typical Concentration Range (%)	Observed (IMS-MS) Range (%)
Xylenes	1330-20-7	C ₈ H ₁₀	~10	0–≤20	0–0.11
Ethylbenzene	100-41-4	C ₈ H ₁₀	~5	0–≤15	0–0.11
Naphthalene	91-20-3	C ₁₀ H ₈	~20	0–≤40	0.13–3.14
Methylnaphthalene	90-12-0	C ₁₁ H ₁₀	~5	0–≤15	0.08–2.04
Methyldicyclopentadiene	25321-13-5	C ₁₁ H ₁₄	~10	0–≤21	0.02–0.44
Toluene	108-88-3	C ₇ H ₈	~10	0–≤20	0–0.03
Styrene	100-42-5	C ₈ H ₈	~12.5	0–≤25	0–0.04
Indene	95-13-6	C ₉ H ₈	~35	0–≤80	0.16–1.67
4-Ethyl-3-octene	53966-51-1	C ₁₀ H ₂₀	~40	0–<80	n.d.
Methylindenes	29036-25-7	C ₁₀ H ₁₀	~10	0–≤70	n.d.
1,2-Dihydronaphthalene	447-53-0	C ₁₀ H ₁₀	~12.5	0–≤25	n.d.
2,3,6-Trimethyl-4-octene	63830-65-9	C ₁₁ H ₂₂	~20	0–≤50	n.d.
1,3-Pentadiene	504-60-9	C ₅ H ₈	~16	0–≤51	n.d.
Cyclopentene	142-29-0	C ₅ H ₈	~15	0–≤25	n.d.
(3Z)-Penta-1,3-diene	1574-41-0	C ₅ H ₈	~10	0–≤20	n.d.
(E)-3-Dodecene	7239-23-8	C ₁₂ H ₂₄	~5	0–≤10	n.d.
Benzene	71-43-2	C ₆ H ₆	~1.0	0–≤3	n.d.
Phenol	108-95-2	C ₆ H ₆ O	~0	0–≤7	n.d.
n-Hexane	110-54-3	C ₆ H ₁₄	~0	0–≤0.2	n.d.

Based on IMS-MS data, LBN substances exhibited similar relative abundances of the reported constituents, and little variation was observed between human health subcategories [26]. Subcategory I substances are typically distinguished by high toluene content, although other, higher m/z compounds at a higher abundance than toluene (C₇H₈) for substances 83757, 83806, and 83946 were detected in this study. Subcategory II substances are expected to contain toluene, ethylbenzene (C₈H₁₀), and xylenes (C₈H₁₀, all isomers included); these were all detected by IMS-MS in relatively high amounts (though not as high as C₉–C₁₀ compounds), although ethylbenzene and xylene isomers could not be distinguished without analytical reference standards. Typical components for subcategory III include toluene, xylene isomers, styrene (C₈H₈), and naphthalene (C₁₀H₈). Naphthalene was the constituent detected by IMS-MS in the highest abundance for both samples belonging to subcategory III (2.4–2.7%), while the other component chemicals were detected at a lesser abundance (~0.05%). Subcategory V has no reported specific constituents apart from being described as “C₉+ from *o*-xylene unit”; sample 83758 fit this description, and C₉H₈, C₉H₁₀, C₁₀H₈, C₁₀H₁₀, C₁₀H₁₂, and C₁₈H₂₀ were all listed constituents of highest abundance [26].

The reported RO constituent list [49] included constituents with 38 CAS numbers that mapped to 20 unique molecular formulae (Table 3); these are expected to comprise between 0% and 80% of any RO substance. Constituents matching twelve of these unique molecular formulae were detected by IMS-MS in RO substances tested herein, ranging in abundance from 0% to 3.14% (naphthalene; Figure 4, Table 3). Eight molecular formulae representing 11 CAS numbers were not detected by IMS-MS. Unlike human health subcategories for LBN, RO subcategories are distinguished mostly by varying levels of dicyclopentadiene (DCPD, C₁₀H₁₂). As expected, DCPD was one of the highest detected constituents using IMS-MS across RO substances, both in subcategories I (high DCPD) and II (low DCPD). Samples for substances representing subcategory III, for which methylcyclopentadiene dimer (MCPD; C₁₂H₁₆) is an additional supporting chemical [27], were not available for

this study. More detailed analyses for Figure 4 can be found in Tables S11 and S12 for LBN and RO, respectively.

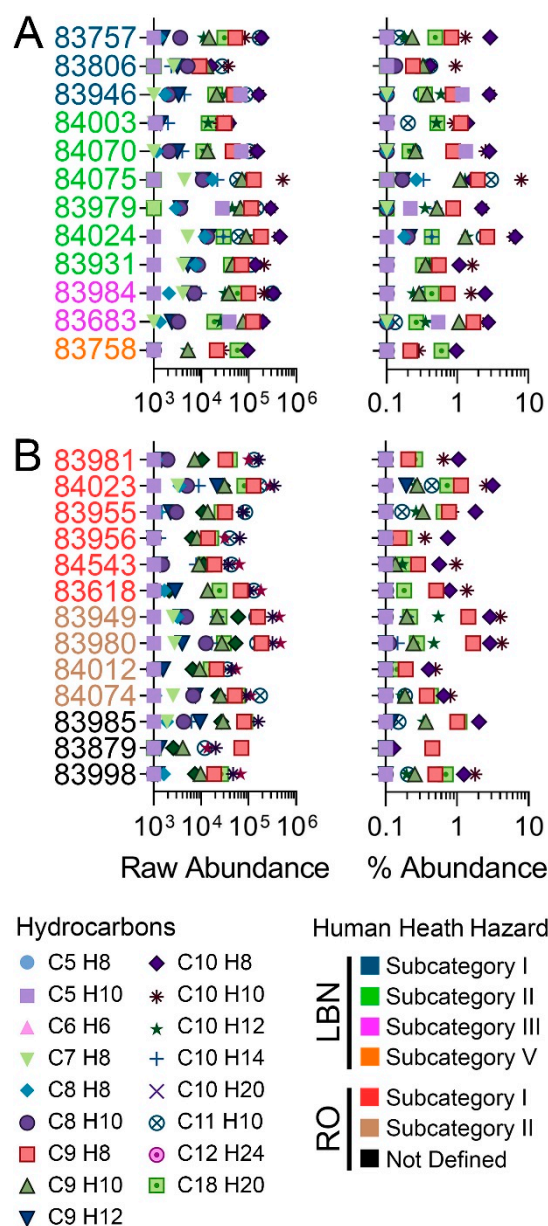


Figure 4. Total feature raw (left) and percent (right) abundance of unique molecular formulas representing typical constituents for LBN (A) and RO (B) categories, see color legend. Constituents were selected based on the substance profiles [48,49]. More detailed analysis can be found in Tables S11 and S12.

3.2. Bioactivity Profiling

Characterizing the composition of UVCBs is critical to establishing structural similarity and the applicability domain of a category [14]. Still, the inherent variability between substances presents uncertainty that may be addressed through the evaluation of the bioactivity of the individual substances. Herein, bioactivity profiling, i.e., testing of the concentration-response effects of the DMSO extracts of the petroleum UVCBs on various human cells and endpoints, was conducted. This analysis aimed to determine whether (i) similarity in bioactivity would be observed within each category and (ii) similar bioactivity profiles would be concordant with chemical similarity from IMS-MS data. The ToxPi approach for integrating bioactivity data across different phenotypes and cell types [15,16,50] is a

common method for visualization and ranking of substances. Here, the data from 20 phenotypes across 5 cell types (Table S8) were integrated by constructing substance-specific ToxPi to represent bioactivity, where one pie slice equates to the overall ToxPi score derived for each cell type. ToxPi profiles were assembled within each tested category, LBN and RO, whereby the bioactivity is relative within that category. Greater bioactivity (i.e., lower POD) is represented by a larger ToxPi score and a bigger pie slice. Unsupervised hierarchical clustering was then used to assess the similarity between the bioactivity profiles of different substances within each category (Figure 5). Overall, RO substances (Figure 5B) exhibited greater bioactivity than LBN substances (Figure 5A). This finding corroborates previous reports, which showed greater bioactivity of higher carbon-range vacuum and hydrotreated gas oils as compared to lower carbon-range straight-run gas oils [15]. Similar to the observations with chemical composition (Figures 2 and 3), there was some, albeit not complete, similarity in bioactivity profiles within each human health subcategory.

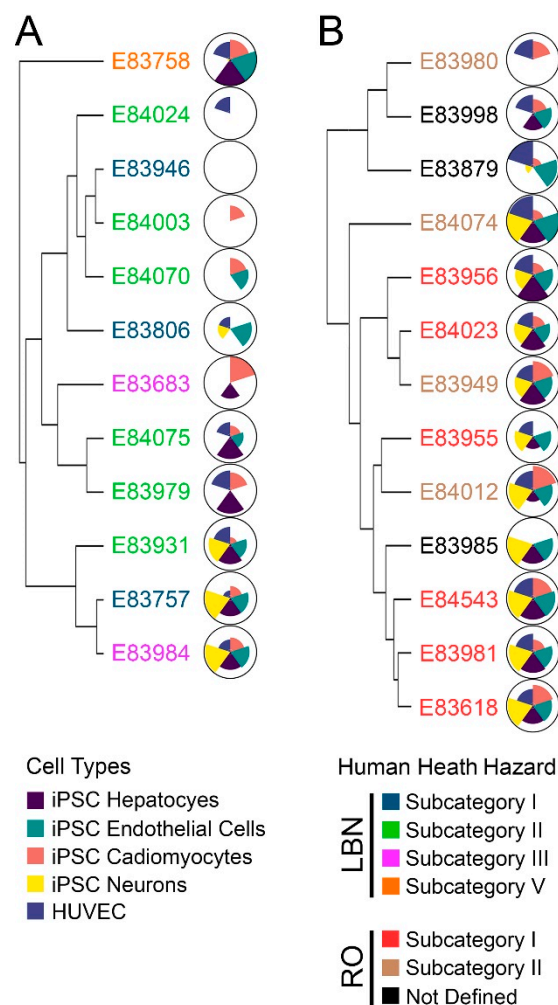


Figure 5. (A) Hierarchical clustering based on bioactivity profiles for LBN DMSO extracts. The name of each substance is colored by the prescribed health hazard group. Corresponding ToxPi diagrams depict overall substance toxicity; each slice represents one cell type, including all assessed phenotypes. Cell types tested include iPSC-derived hepatocytes (purple), endothelial cells (green), cardiomyocytes (pink), neurons (yellow), and HUVEC (blue). Larger pie slices indicate greater toxicity for that substance and cell type. (B) Hierarchical clustering based on bioactivity profiles for RO DMSO extracts. Substance names are again colored by prescribed health hazard groups, and respective ToxPi charts show an overall greater toxicity of RO substances as compared to LBN substances.

In the LBN category, subcategory II exhibited the greatest bioactivity similarity. Some of the substances had lower bioactivity (E84024, E84003, and E84070) as compared to others (E84075, E83979, and E83931). This result is concordant with the data on chemical composition; samples 84003 and 84070 are two of the least complex LBN substances tested, and they also exhibited few effects in vitro. Similarly, samples 84075 and 83979 were of comparable chemical complexity and elicited similar bioactivity profiles. iPSC-derived neurons and hepatocytes were the most affected cell types across the LBN substances.

In the RO category, subcategory I exhibited the most similarity in bioactivity profiles; five out of the six samples assigned to subcategory I demonstrated bioactivity in all cell types tested. The sixth sample (E83955) was bioactive in four out of five cell types, albeit to a lesser extent. These results were also generally concordant with the chemical composition data in Figures 2 and 3; subcategory I substances 83956 and 84023 were closely related, while 84543 and 83981 also exhibited compositional concordance. Bioactivity was observed more consistently across all cell types for RO substances; still, iPSC-derived endothelial cells, neurons, and hepatocytes were the cell types for which bioactivity was observed most often.

3.3. Comparison of Bioactivity and Chemical Composition

Human health evaluations for petroleum UVCBs are typically based on substance grouping using physio-chemical properties and manufacturing processes, followed by an assessment of possible hazards by several constituents. The bioactivity profiling described above (Figure 5) grouped substances based on similarity in bioactivity, but the grouping was not fully concordant with existing HPV categories; therefore, “substance similarity” was examined using both chemical profiles and bioactivity. Specifically, the objective of this study was to assess chemical and in vitro data together to determine whether chemical composition may align with trends in bioactivity (Figures 6 and 7). First, the overall chemical composition clustering of samples in the LBN category (Figure 2) was split into four sub-groups based on clustering (Figure 6A). To visualize the hydrocarbon composition of each substance, the carbon number range was plotted versus double bond equivalence (DBE) and abundance (Figure 6B). This typical data presentation for petroleum UVCBs allows a visual assessment of the complexity of each sample, as well as the range of hydrocarbon types that are present. Aromaticity, measured by DBE, varied from a minimum of 1 (low aromaticity, likely olefin or alkane species) to a maximum of 30+ (highly aromatic species). Overall, the chemical profiles of most LBN samples were within the expected C7–C12 range; however, many samples contained an appreciable number of constituents in the C40 range that are aromatic. Generally, samples with a higher carbon number range exhibited greater bioactivity across all cell types tested. The first subgroup (N84070, N84003, and N83946) included the least bioactive substances of all tested samples; based on their compositional signatures, these samples were clustered based on the high abundance of C40+ constituents. The second subgroup (N83683 and N83979) exhibited a high abundance of <C20 constituents, as well as some that were >C40 (although these were not as highly abundant as in the first subgroup). Between the two substances in this subgroup, the most bioactivity was contributed by iPSC-derived cardiomyocytes and hepatocytes. The third subgroup (N83984, N84075, and N84024) displayed the most chemical similarity between N84075 and N84024, although these substances only had hepatocyte bioactivity in common. N83984 had a chemical abundance distributed over a wider carbon range (up to C40) and exhibited greater bioactivity in all cell types. Finally, the last subgroup (N83806, N83931, N83757, and N83758) presented the greatest chemical variability and carbon number range when compared to the other LBN samples. The three substances (N83931, N83757, and N83758) with the largest carbon number range and high levels of aromatic species exhibited some of the highest bioactivity, which was especially notable in iPSC-derived neurons, endothelial cells, and hepatocytes.

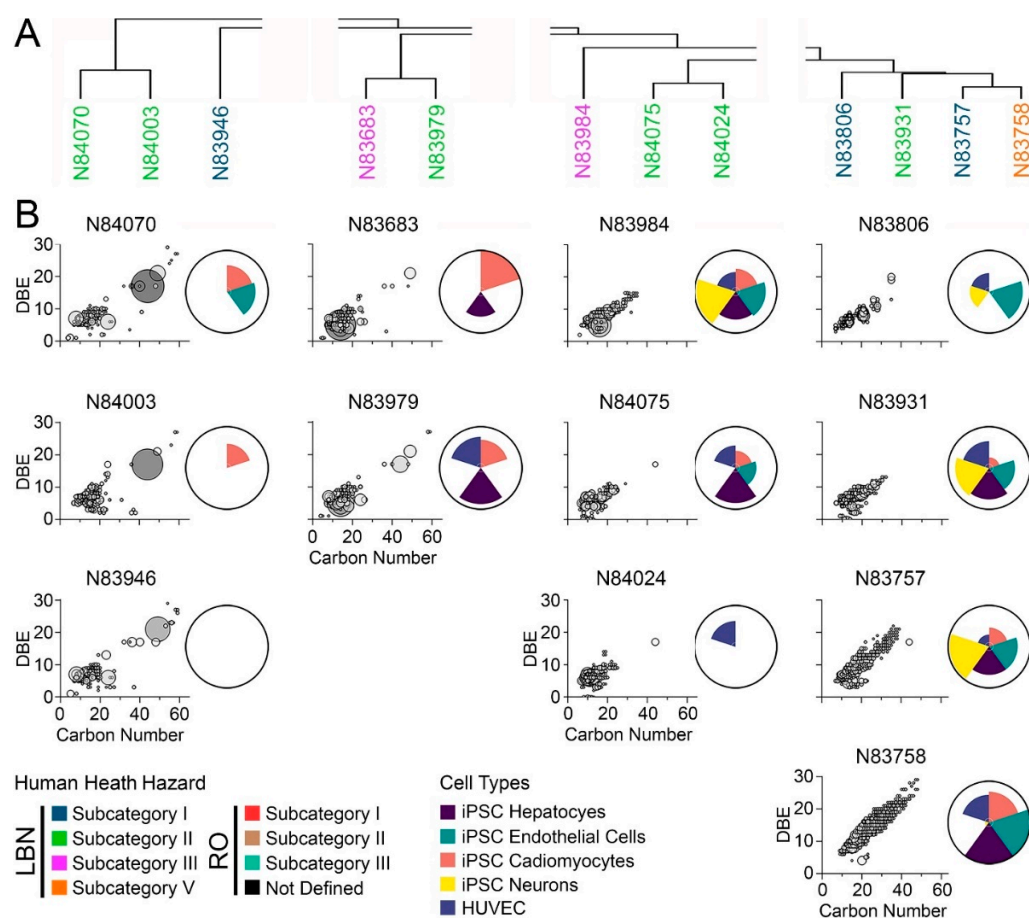


Figure 6. (A) Hierarchical clustering based on chemical profiles of neat LBN substances. Substance names are colored by prescribed health hazard groups. (B) IMS-MS chemical profiles depicted as carbon number versus double bond equivalence. Larger bubble sizes and a darker grey color depict more abundant features. Adjacent ToxPi charts show overall bioactivity across iPSC-derived hepatocytes (purple), endothelial cells (green), cardiomyocytes (pink), and neurons (yellow), as well as HUVEC (blue).

The same analyses were conducted for RO substances (Figure 7). Three subcategories were examined (Figure 7A). Most of the samples had the greatest number of constituents (Figure 7B) in the C7–C20 range; however, all substances had constituents in the C20–C40 range, and one substance (N83956) extended to C50+. Like LBN substances, greater bioactivity across all cell types tested was generally associated with a larger carbon number range. All RO substances exhibited a larger carbon number range than LBN substances (except for N83757 and N83758). Substances in the first RO subgroup (N84543, N83998, N83981, N83618, and N84012) exhibited bioactivity in all cell types tested except sample N83998, which was not bioactive in iPSC-derived neurons. Three of these substances belong to human health subcategory I (N84543, N83981, and N83618). Of the four substances belonging to the second subgroup (N83879, N84074, N83949, and N83980), three were in human health subcategory II and had generally comparable chemical profiles; still, N83980 exhibited bioactivity only in iPSC-derived cardiomyocytes and HUVEC, whereas N84074 and N83949 exhibited bioactivity in all cell types tested. Similar conclusions could be drawn for the third subgroup (N84023, N83985, N83955, and N83956); these substances are members of RO human health subcategory I and showed considerable overlap in chemical composition. Despite the difference in carbon number ranges between N84023 and N83956, their bioactivity profiles shared a closer resemblance to each other.

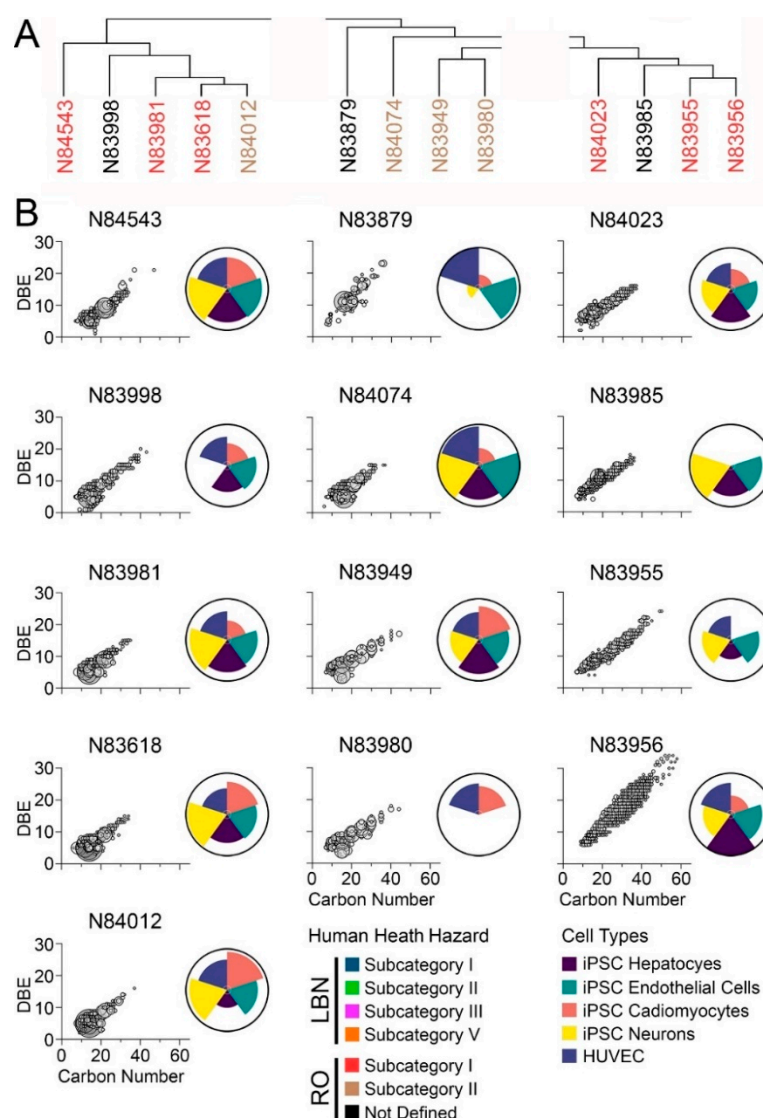


Figure 7. (A) Hierarchical clustering based on chemical profiles of neat RO substances. Substance names are colored by prescribed health hazard groups. (B) IMS-MS chemical profiles depicted as carbon number versus double bond equivalence. Larger bubble sizes and a darker grey color depict more abundant features. Adjacent ToxPi charts show overall bioactivity across iPSC-derived hepatocytes (purple), endothelial cells (green), cardiomyocytes (pink), and neurons (yellow), as well as HUVEC (blue).

3.4. Determining What Chemical Constituents May Be Associated with Bioactivity

Data on PAC content with 3+ aromatic rings is conventional analytical chemistry-based information that is used to judge the potential health hazards of petroleum UVCBs; higher PAC content is assumed to have higher bioactivity [51,52]. However, regulatory bodies such as ECHA are typically hesitant to rely on these data alone in hazard evaluation, reasoning that PAC content may not necessarily represent the entire bioactive fraction [10]. It was also argued by ECHA that such a broad characterization does not provide enough information to justify the application of read-across [10]. Indeed, considerable heterogeneity in both chemical composition [25] and bioactivity [15–17] of substances within current petroleum UVCB categories, based on the physio-chemical properties and manufacturing process, has been previously observed; therefore, the findings presented in Figures 2–7 for LBN and RO categories are not unexpected. While such heterogeneity in both overall chemical composition and bioactivity cannot be used directly to justify similarity between substances

in each category, determination of whether there may be statistically significant associations among specific chemical constituents and bioactivity phenotypes has not been previously attempted for petroleum UVCBs.

Therefore, machine learning was used to predict overall and cell type-specific bioactivity from the IMS-MS chemical profiles for the tested substances (Figure 8). This approach has previously been used to provide a refined analysis of bioactive components in case studies of other complex substances [45] and mixtures [53]. Even though neither chemical composition, nor bioactivity data separately replicated existing categories/sub-categories of the tested substances, the overall bioactivity of each sample was found to be strongly associated (multiple testing-corrected q -value <0.1) with the chemical profiles of both neat and DMSO-extracted samples (Figure 8A and B, top). Interestingly, the data from iPSC-derived neurons and endothelial cells was also strongly associated with the chemical profiles of the neat substances (Figure 8A, middle and bottom), but not of the DMSO extracts (Figure 8B, middle and bottom). Next, it was determined what constituents in the neat samples were most influential in this multivariate prediction analysis (Figure 8C). Of the seven constituents that were significantly associated with bioactivity, all were high-molecular-weight PAC belonging to homologous series with pyrene, fluorene, or naphthalene. Only one constituent could not be identified with high confidence using a workflow for IMS-MS data analysis of petroleum substances [18]. Table S13 shows a list of potential names that could be assigned to the seven hydrocarbon features driving bioactivity and their corresponding hazard classifications. Further, the relative abundance of these constituents in each tested sample (Figure 8D) was compared. It was found that there was an overall higher abundance of these constituents in RO substances as compared to LBN substances, supporting the previous observation that RO substances were generally more bioactive.

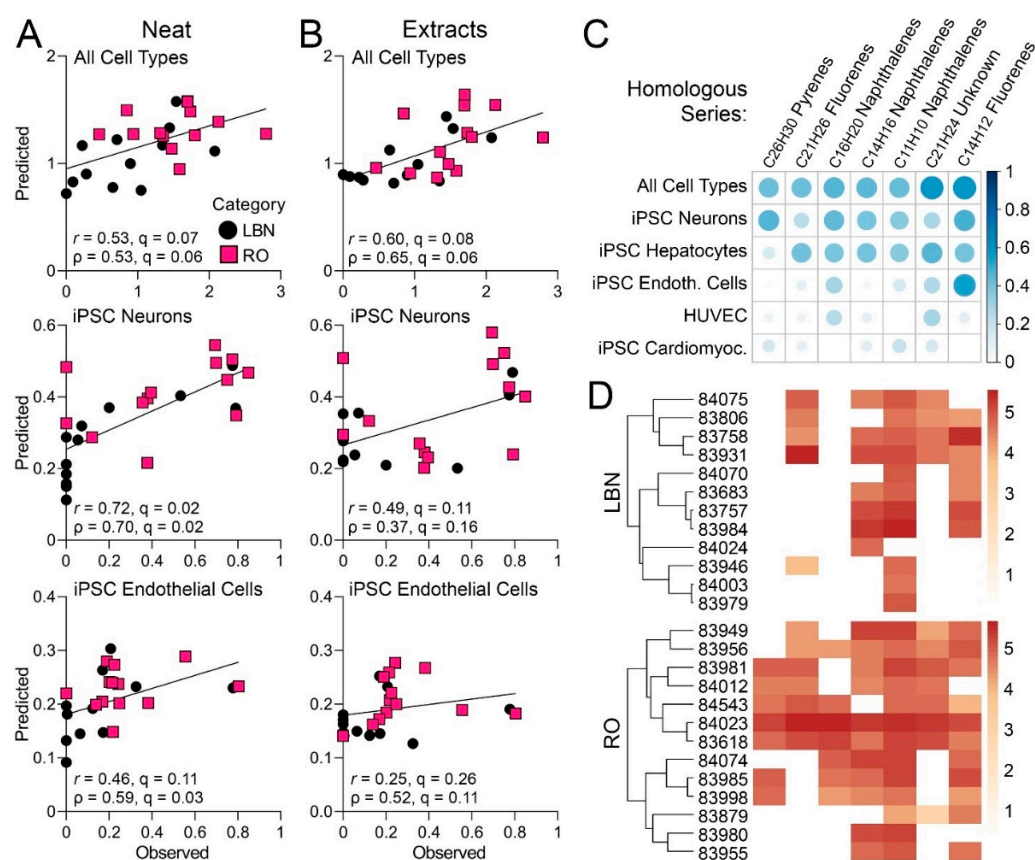


Figure 8. (A,B) Each scatterplot shows bioactivity (y -axis) for the overall ToxPi score (top), iPSC-derived endothelial cells (middle), and iPSC-derived neurons (bottom) as predicted from the chemical

profiles of neat (A) and corresponding extracts (B). Observed bioactivity is shown on the x-axis. Bioactivity prediction was conducted using the penalized regression approach described in Methods. The predicted values were obtained by leave-one-out cross validation, where the prediction model was developed with each sample left out of analysis, and the model applied to the features of the held-out sample. The most informative validations were chosen with the highest prediction r (Pearson coefficient) and the lowest q (false discovery rate value). (C) Correlation plot depicting the hydrocarbon compounds from neat samples that were most significantly predictive of the overall ToxPi score based on cross-validation analyses. Bubble size represents the Pearson correlation between feature abundance and ToxPi score overall as well as for individual cell types. Positive correlations are shown in blue, whereas negative correlations are shown in red. (D) Heatmap depicting the relative abundance of each feature in each sample tested. A darker color indicates higher abundance.

4. Discussion

This study is novel because it used new analytical and toxicological approaches to examine both the chemical composition and biological effects of complex petroleum UVCBs. Samples were from two HPV categories, and this study aimed to determine the extent of chemical and bioactivity similarity among substances that have been previously assigned to these categories using physio-chemical properties and manufacturing process information. The main questions of this study were four-fold: (1) To what extent can petroleum UVCBs be characterized using novel analytical methods such as IMS-MS to meet the most recent ECHA advice on substance characterization for read-across [14]? (2) How much *chemical* variability is to be expected within and between existing LBN and RO manufacturing categories? (3) How much *biological* variability is to be expected within and between existing LBN and RO manufacturing categories? Additionally, (4) What constituents are potential drivers of bioactivity in complex petroleum UVCBs?

First, it was found that in the DMSO extracts (but not in the neat substances) in the RO and LBN manufacturing categories, the sum of constituents present in amounts $\geq 1\%$ of the overall substance was above the 80% ECHA threshold [14]. This means that additional analyses need to be performed to further identify the constituents of concern below 1%; for this, higher resolution analytical instruments such as Orbitrap and Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometry (MS) may be more suitable [5,19]. In addition, subsequent application of targeted chromatographic approaches would also be needed to confirm the structural identities of identified constituents of interest [5,54–56].

Second, broad chemical concordance was observed for substances belonging to the same category; however, considerable variability was observed between substances in the same category and even subcategory. This was likely a result of inherent substance variability or reaction byproduct impurities from manufacturing processes. While compositional variability is to be expected, recent advice from ECHA calls for the characterization of such variability. Not only is there a need to provide compositional characterization of the substances identified by different CAS RN but grouped into a category, but also characterization of the variability of the same product across manufacturing batches and refineries [14]. The analysis of at least five independent (i.e., production batch) samples from all registrants of a substance is the most recent threshold proposed by ECHA [14]. To establish this, novel analytical techniques such as IMS-MS, Orbitrap-MS, and FT-ICR-MS are most appropriate [5]. A recent study showed 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 [25].

Third, similarity in bioactivity was observed within the overall LBN and RO categories; however, less concordance was evident within previously proposed HPV human health subcategories. This observation is similar to that from a larger study of other petroleum UVCBs, where 141 substances spanning 6 product categories were tested in 15 human organotypic cell types to investigate substance similarity using both bioactivity signa-

tures [16] and transcriptomic profiles [17]. These studies showed that the bioactivity and transcriptomic data correlate strongly with the PAC content of each substance and can be used to rank overall categories in a way similar to that using other hazard data (typically from animal and genotoxicity studies); however, they cannot be used to substantiate existing groupings. These data are still highly informative, as a combination of bioactivity and transcriptomic data could be integrated to make decisions as to the selection of class-representative worst-case petroleum UVCBs for subsequent evaluation in vivo [57].

Fourth, this study is also informative in terms of the hazard evaluation of petroleum UVCBs. Due to the chemical complexity of petroleum UVCBs, there is no harmonized methodology for their risk assessment; both whole mixture and constituent-based approaches can be used [58–60]. The constituent-based approach is most commonly used for petroleum UVCBs [61,62]; however, the approaches to the selection of the chemical constituents of interest are yet to be standardized [4,63,64]. Furthermore, petroleum UVCBs are typically tested as the whole substance (in vivo) or as a DMSO extract (in vitro), rather than as individual constituents or groups of constituents [65]. The results presented herein are consistent with the historical observations that the potential hazards of petroleum UVCBs are largely determined by their PAC 3–7 ring content [66–68] and previous observations that PAC content is the strongest “driver” of in vitro bioactivity [16,17,69]. In addition, this study also provides specific details on what constituents, rather than PAC 3–7 overall, are most strongly associated with in vitro bioactivity. Such an approach, assessing relationships between high-dimensional chemical profiles and multi-dimensional bioactivity phenotypes, is informative for defining constituents of interest for component-based risk assessment of petroleum UVCBs. This is especially beneficial in scenarios such as environmental disasters, where exposure assessment and hazard evaluation are time sensitive [50,53].

This study is not without limitations. The availability of samples, a common challenge in studies of large-volume produced substances, limited our ability to characterize the intra-category and sub-category variability. Even though we tested 25 samples that were representative of two manufacturing categories and multiple sub-categories within them, the desired replication was lacking. Prior studies showed that a single sample per category may not provide adequate information to capture the individual category characteristics [70]. Updated ECHA advice also addressed this limitation, specifying that constituent concentrations in “*at least five independent samples of the substance . . . from different production batches . . . as produced by all the registrants*” must be included to characterize the variability [14]. However, obtaining samples for the analysis of petroleum UVCBs is a well-known challenge that cannot be easily addressed because samples need to be provided by the individual manufacturers and cannot be commercially procured from standard chemical suppliers. Some studies have begun to address compositional variability within production batches [25]; still, additional investigation is warranted to examine variability in bioactivity within production batches as well.

Our study used one analytical approach to characterize the chemical composition of tested substances; however, products of petroleum refining are highly complex, and both separation, ionization, and detection methods may affect the molecules that are identifiable using each technique [5,19,71]. Therefore, the analytical results presented herein should be interpreted with caution. For example, we reason that while they may be used for the purpose of relative comparisons among substances and categories, they should not be used to infer the exact chemical composition or absolute concentrations of the individual constituents.

In addition, DMSO extraction, a widely used method to enable testing complex petroleum substances [51,72], captures only a fraction of the neat substance. This is a concern for regulators, who maintain that solvent extraction may restrict the bioactive fraction to only constituents that are soluble in biocompatible solvents such as DMSO [10]. Recent developments in the field have therefore adapted alternative dosing techniques as potential solutions to enable more high-throughput in vitro testing [73–78], and future

studies of petroleum UVCBs may utilize these alternative approaches for delivering the substances in small-volume in vitro methods.

Another well-recognized challenge of using in vitro bioactivity for hazard-based evaluations of chemicals is the translation of in vitro results to apical in vivo phenotypes. The complex composition of UVCBs makes it difficult to conduct traditional in vitro-to-in vivo extrapolation from bioactive concentrations to human exposures [79]. It is still debated as to whether bioactivity should be used only for screening and prioritization [80], for grouping and read-across [81], or to establish health-protective points of departure for screening-level assessments [82]. The use of in vitro bioactivity data in regulatory decision-making is rapidly evolving, and regulators currently indicate that the results of cell-based studies should be confirmed with additional assays, including studies in animals [83].

5. Conclusions

Overall, this study demonstrates the benefits of simultaneous assessment of both chemical composition and bioactivity when evaluating the potential hazard properties of petroleum UVCBs. We found that based on the samples analyzed herein, existing categories, based largely on the manufacturing considerations and intended future uses of these products, may be considered heterogeneous in terms of their composition and bioactivity. While additional work is needed to evaluate a larger compendium of substances, including different manufacturing batches of the same substance and testing alternative in vitro delivery methods for these “difficult to test” substances, we conclude that an approach that combines chemical composition and bioactivity data is sensible. These complementary data streams provide information that will enable a more comprehensive and confident characterization of similarities, differences, and variability between and within manufacturing categories of petroleum UVCBs.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/toxics11070586/s1>, Table S1: LBN & RO US EPA Human Health Hazard Subcategory Assignments; Table S2: Raw IMS-MS Data Matrix (Before Processing) for Neat Samples; Table S3: Raw IMS-MS Data Matrix (Before Processing) for DMSO Extracts; Table S4: Filtered IMS-MS Data Matrix (After Processing) for Neat Samples; Table S5: Filtered IMS-MS Data Matrix (After Processing) for DMSO Extracts; Table S6: Filtered IMS-MS Data Matrix with Molecular Formula Assignments: Neat Samples; Table S7: Filtered IMS-MS Data Matrix with Molecular Formula Assignments: DMSO Extracts; Table S8: Cell-Specific Phenotypes and Endpoints Measured; Table S9: Positive Controls for All Cell Types Tested; Table S10: Links to US EPA HPV Documents for LBN and RO Categories; Table S11: LBN: Detailed Analyses of Expected vs. Observed Constituent Abundances; Table S12: RO: Detailed Analyses of Expected vs. Observed Constituent Abundances; Table S13: Potential Identities for Features Driving Bioactivity.

Author Contributions: Conceptualization, W.D.K., F.A.G., A.C.C. and I.R.; Methodology, W.D.K., A.C.C., L.C.F., F.A.G., E.S.B., Y.-H.Z., F.A.W. and I.R.; Validation, A.C.C., W.D.K., L.C.F. and I.R.; Formal Analysis, A.C.C., W.D.K., F.A.G., Y.-H.Z., F.A.W. and E.S.B.; Writing—Original Draft Preparation, A.C.C.; Writing—Review and Editing, A.C.C., L.C.F., W.D.K., F.A.G., Y.-H.Z., F.A.W., E.S.B. and I.R.; Visualization, A.C.C. and I.R.; Supervision, I.R.; Funding Acquisition, I.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded, in part, by a contract with the Foundation for Chemistry Research and Initiatives, a 501(c)(3) tax-exempt organization established by the American Chemistry Council (Washington, DC). IR, ACC, ESB, FAW, WDK, and FAG were partially supported by the National Institute of Environmental Health Sciences grants P42 ES027704 and T32 ES026568 (WDK, ACC and LCF). The views expressed in this manuscript do not reflect those of the funding agencies. The use of specific commercial products in this work does not constitute endorsement by the funding agencies.

Institutional Review Board Statement: This study is not subject to institutional review.

Informed Consent Statement: Not applicable.

Data Availability Statement: All pertinent data are included in Supplementary Materials.

Acknowledgments: We are grateful to the anonymous reviewers for their valuable feedback. The graphical abstract and Figure 1 were produced with BioRender. Chemical structures were obtained from ChemSpider.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

UVCB, unknown, variable, complex reaction byproducts, or biological materials; HPV, high production volume; POD, point of departure; IMS-MS, ion mobility spectrometry-mass spectrometry; GC-MS, gas chromatography; PAC, polycyclic aromatic compounds; LBN, low benzene naphthas; RO, resin oils and cyclodiene dimer concentrates; DMSO, Dulbecco's Modified Eagle Medium; HUVEC, human umbilical vein endothelial cells; iPSC, induced pluripotent stem cell; ACC, American Chemistry Council; CAS, Chemical Abstract Service; ASTM, American Society for Testing and Materials; QTOF, quadrupole time of flight; APPI, atmospheric pressure photoionization; CCS, collisional cross section; DT, drift time; KMD, Kendrick mass defect; DBE, double bond equivalence; ToxPi, Toxicological Prioritization Index; ECHA, European Chemicals Agency; REACH, Registration, Evaluation, and Authorization of Chemicals; DCPD, dicyclopentadiene; MCPD, methylcyclopentadiene dimer; FT-ICR, Fourier transform ion cyclotron resonance; MS, mass spectrometry.

References

- Bishop, P.L.; Manuppello, J.R.; Willett, C.E.; Sandler, J.T. Animal use and lessons learned in the U.S. High production volume chemicals challenge program. *Environ. Health Perspect.* **2012**, *120*, 1631–1639. [[CrossRef](#)] [[PubMed](#)]
- Rogers, M.D. Risk analysis under uncertainty, the precautionary principle, and the new eu chemicals strategy. *Regul. Toxicol. Pharmacol.* **2003**, *37*, 370–381. [[CrossRef](#)] [[PubMed](#)]
- Lai, A.; Clark, A.M.; Escher, B.I.; Fernandez, M.; McEwen, L.R.; Tian, Z.; Wang, Z.; Schymanski, E.L. The next frontier of environmental unknowns: Substances of unknown or variable composition, complex reaction products, or biological materials (uvcb). *Environ. Sci. Technol.* **2022**, *56*, 7448–7466. [[CrossRef](#)]
- Salvito, D.; Fernandez, M.; Arey, J.S.; Lyon, D.Y.; Lawson, N.; Deglin, S.; MacLeod, M. The path to uvcb ecological risk assessment: Grappling with substance characterization. *Environ. Toxicol. Chem.* **2022**, *41*, 2649–2657. [[CrossRef](#)] [[PubMed](#)]
- Roman-Hubers, A.T.; Cordova, A.C.; Barrow, M.P.; Rusyn, I. Analytical chemistry solutions to hazard evaluation of petroleum refining products. *Regul. Toxicol. Pharmacol.* **2023**, *137*, 105310. [[CrossRef](#)]
- CONCAWE. *Hazard Classification and Labelling of Petroleum Substances in the European Economic Area—2020*; CONCAWE: Brussels, Belgium, 2020.
- CONCAWE. *Guidance to Registrants on Methods for Characterisation of Petroleum Uvcb Substances for Reach Registration Purposes*; CONCAWE: Brussels, Belgium, 2020.
- McKee, R.H.; 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.* **2014**, *33*, 4S–16S. [[CrossRef](#)]
- ECHA. *Read-Across Assessment Framework (Raaf)—Considerations on Multi-Constituent Substances and Uvcb*; European Chemical Agency: Helsinki, Finland, 2017.
- ECHA. *Testing Proposal Decision on Substance ec 295-332-8 “Extracts (Petroleum), Deasphalted Vacuum Residue Solvent”*; European Chemicals Agency: Helsinki, Finland, 2020.
- ECHA. *Guidance for Identification and Naming of Substances under Reach and Clp*; European Chemical Agency: Helsinki, Finland, 2017; Volume 2.1.
- ECHA. *Testing Proposal Decision on Substance ec 265-182-8 “Gas Oils (Petroleum), Hydrodesulfurized”*; European Chemicals Agency: Helsinki, Finland, 2021.
- ECHA. *Testing Proposals Decision on Substance ec 265-110-5 “Extracts (Petroleum), Residual Oil Solvent”*; European Chemicals Agency: Helsinki, Finland, 2020.
- ECHA. *Advice on Using Read-Across for Uvcb Substances—Obligations Arising from Commission Regulation 2021/979, Amending Reach Annexes*; European Chemicals Agency: Helsinki, Finland, 2022.
- Grimm, F.A.; Iwata, Y.; Sirenko, O.; Chappell, G.A.; Wright, F.A.; Reif, D.M.; Braisted, J.; Gerhold, D.L.; Yeakley, J.M.; Shepard, P.; et al. A chemical-biological similarity-based grouping of complex substances as a prototype approach for evaluating chemical alternatives. *Green Chem.* **2016**, *18*, 4407–4419. [[CrossRef](#)]
- House, J.S.; Grimm, F.A.; Klaren, W.D.; Dalzell, A.; Kuchi, S.; Zhang, S.D.; Lenz, K.; Boogaard, P.J.; Ketelslegers, H.B.; Gant, T.W.; et al. Grouping of uvcb substances with new approach methodologies (nams) data. *ALTEX* **2021**, *38*, 123–137. [[CrossRef](#)]
- House, J.S.; Grimm, F.A.; Klaren, W.D.; Dalzell, A.; Kuchi, S.; Zhang, S.D.; Lenz, K.; Boogaard, P.J.; Ketelslegers, H.B.; Gant, T.W.; et al. Grouping of uvcb substances with dose-response transcriptomics data from human cell-based assays. *ALTEX* **2022**, *39*, 388–404. [[CrossRef](#)]

18. Roman-Hubers, A.T.; Cordova, A.C.; Aly, N.A.; McDonald, T.J.; Lloyd, D.T.; Wright, F.A.; Baker, E.S.; Chiu, W.A.; Rusyn, I. Data processing workflow to identify structurally related compounds in petroleum substances using ion mobility spectrometry-mass spectrometry. *Energy Fuels* **2021**, *35*, 10529–10539. [\[CrossRef\]](#)
19. Palacio Lozano, D.C.; Thomas, M.J.; Jones, H.E.; Barrow, M.P. Petroleomics: Tools, challenges, and developments. *Annu. Rev. Anal. Chem.* **2020**, *13*, 405–430. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Roman-Hubers, A.T.; McDonald, T.J.; Baker, E.S.; Chiu, W.A.; Rusyn, I. A comparative analysis of analytical techniques for rapid oil spill identification. *Environ. Toxicol. Chem.* **2021**, *40*, 1034–1049. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Grimm, F.A.; Russell, W.K.; Luo, Y.S.; Iwata, Y.; Chiu, W.A.; Roy, T.; Boogaard, P.J.; Ketelslegers, H.B.; Rusyn, I. Grouping of petroleum substances as example uvcb by ion mobility-mass spectrometry to enable chemical composition-based read-across. *Environ. Sci. Technol.* **2017**, *51*, 7197–7207. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Onel, M.; Beykal, B.; Wang, M.; Grimm, F.A.; Zhou, L.; Wright, F.A.; Phillips, T.D.; Rusyn, I.; Pistikopoulos, E.N. Optimal chemical grouping and sorbent material design by data analysis, modeling and dimensionality reduction techniques. *ESCAPE* **2018**, *43*, 421–426.
23. Aeppli, C.; Mitchell, D.A.; Keyes, P.; Beirne, E.C.; McFarlin, K.M.; Roman-Hubers, A.T.; Rusyn, I.; Prince, R.C.; Zhao, L.; Parkerton, T.F.; et al. Oil irradiation experiments document changes in oil properties, molecular composition, and dispersant effectiveness associated with oil photo-oxidation. *Environ. Sci. Technol.* **2022**, *56*, 7789–7799. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Roman-Hubers, A.T.; Aeppli, C.; Dodds, J.N.; Baker, E.S.; McFarlin, K.M.; Letinski, D.J.; Zhao, L.; Mitchell, D.A.; Parkerton, T.F.; Prince, R.C.; et al. Temporal chemical composition changes in water below a crude oil slick irradiated with natural sunlight. *Mar. Pollut. Bull.* **2022**, *185*, 114360. [\[CrossRef\]](#)
25. Roman-Hubers, A.T.; Cordova, A.C.; Rohde, A.M.; Chiu, W.A.; McDonald, T.J.; Wright, F.A.; Dodds, J.N.; Baker, E.S.; Rusyn, I. Characterization of compositional variability in petroleum substances. *Fuel* **2022**, *317*, 123547. [\[CrossRef\]](#)
26. US EPA. *Screening-Level Hazard Characterization: Low Benzene Naphthas Category*; US EPA: Washington, DC, USA, 2010.
27. US EPA. *Screening-Level Hazard Characterization: Resin Oils and Cyclodiene Dimer Concentrates Category*; US EPA: Washington, DC, USA, 2010.
28. ASTM International. *Standard Test Method for Determining Carcinogenic Potential of Virgin Base Oils in Metalworking Fluids*; ASTM International: West Conshohocken, PA, USA, 2014.
29. Zheng, X.; Dupuis, K.T.; Aly, N.A.; Zhou, Y.; Smith, F.B.; Tang, K.; Smith, R.D.; Baker, E.S. Utilizing ion mobility spectrometry and mass spectrometry for the analysis of polycyclic aromatic hydrocarbons, polychlorinated biphenyls, polybrominated diphenyl ethers and their metabolites. *Anal. Chim. Acta* **2018**, *1037*, 265–273. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Baker, E.S. Collision Cross Section Database. Available online: <https://brcwebportal.cos.ncsu.edu/baker/> (accessed on 15 December 2020).
31. Ahmed, A.; Cho, Y.J.; No, M.H.; Koh, J.; Tomczyk, N.; Giles, K.; Yoo, J.S.; Kim, S. Application of the mason-schamp equation and ion mobility mass spectrometry to identify structurally related compounds in crude oil. *Anal. Chem.* **2011**, *83*, 77–83. [\[CrossRef\]](#)
32. Ponthus, J.; Riches, E. Evaluating the multiple benefits offered by ion mobility-mass spectrometry in oil and petroleum analysis. *Int. J. Ion Mobil. Spec.* **2013**, *16*, 95–103. [\[CrossRef\]](#)
33. Dodds, J.N.; Baker, E.S. Ion mobility spectrometry: Fundamental concepts, instrumentation, applications, and the road ahead. *J. Am. Soc. Mass. Spectrom.* **2019**, *30*, 2185–2195. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Luo, R.J.; Schrader, W. Development of a non-targeted method to study petroleum polyaromatic hydrocarbons in soil by ultrahigh resolution mass spectrometry using multiple ionization methods. *Polycycl. Aromat. Comp.* **2022**, *42*, 643–658. [\[CrossRef\]](#)
35. Korsten, H. Characterization of hydrocarbon systems by dbe concept. *AIChE J.* **1997**, *43*, 1559–1568. [\[CrossRef\]](#)
36. Grimm, F.A.; Iwata, Y.; Sirenko, O.; Bittner, M.; Rusyn, I. High-content assay multiplexing for toxicity screening in induced pluripotent stem cell-derived cardiomyocytes and hepatocytes. *Assay Drug Dev. Technol.* **2015**, *13*, 529–546. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Iwata, Y.; Klaren, W.D.; Lebakken, C.S.; Grimm, F.A.; Rusyn, I. High-content assay multiplexing for vascular toxicity screening in induced pluripotent stem cell-derived endothelial cells and human umbilical vein endothelial cells. *Assay Drug Dev. Technol.* **2017**, *15*, 267–279. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Sirenko, O.; Crittenden, C.; Callamaras, N.; Hesley, J.; Chen, Y.W.; Funes, C.; Rusyn, I.; Anson, B.; Cromwell, E.F. Multiparameter in vitro assessment of compound effects on cardiomyocyte physiology using ipsc cells. *J. Biomol. Screen* **2013**, *18*, 39–53. [\[CrossRef\]](#)
39. Sirenko, O.; Cromwell, E.F.; Crittenden, C.; Wignall, J.A.; Wright, F.A.; Rusyn, I. Assessment of beating parameters in human induced pluripotent stem cells enables quantitative in vitro screening for cardiotoxicity. *Toxicol. Appl. Pharmacol.* **2013**, *273*, 500–507. [\[CrossRef\]](#)
40. Sirenko, O.; Grimm, F.A.; Ryan, K.R.; Iwata, Y.; Chiu, W.A.; Parham, F.; Wignall, J.A.; Anson, B.; Cromwell, E.F.; Behl, M.; et al. In vitro cardiotoxicity assessment of environmental chemicals using an organotypic human induced pluripotent stem cell-derived model. *Toxicol. Appl. Pharmacol.* **2017**, *322*, 60–74. [\[CrossRef\]](#)
41. Sirenko, O.; Hesley, J.; Rusyn, I.; Cromwell, E.F. High-content assays for hepatotoxicity using induced pluripotent stem cell-derived cells. *Assay Drug Dev. Technol.* **2014**, *12*, 43–54. [\[CrossRef\]](#)
42. Sirenko, O.; Hesley, J.; Rusyn, I.; Cromwell, E.F. High-content high-throughput assays for characterizing the viability and morphology of human ipsc-derived neuronal cultures. *Assay Drug Dev. Technol.* **2014**, *12*, 536–547. [\[CrossRef\]](#)
43. Reif, D.M.; Sypa, M.; Lock, E.F.; Wright, F.A.; Wilson, A.; Cathey, T.; Judson, R.R.; Rusyn, I. Toxpi gui: An interactive visualization tool for transparent integration of data from diverse sources of evidence. *Bioinformatics* **2013**, *29*, 402–403. [\[CrossRef\]](#) [\[PubMed\]](#)

44. Marvel, S.W.; To, K.; Grimm, F.A.; Wright, F.A.; Rusyn, I.; Reif, D.M. Toxpi graphical user interface 2.0: Dynamic exploration, visualization, and sharing of integrated data models. *BMC Bioinf.* **2018**, *19*, 80. [\[CrossRef\]](#)
45. Luo, Y.S.; Chen, Z.; Blanchette, A.D.; Zhou, Y.H.; Wright, F.A.; Baker, E.S.; Chiu, W.A.; Rusyn, I. Relationships between constituents of energy drinks and beating parameters in human induced pluripotent stem cell (ipsc)-derived cardiomyocytes. *Food Chem. Toxicol.* **2021**, *149*, 111979. [\[CrossRef\]](#)
46. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer Series in Statistics; Springer: Berlin, Germany, 2016; p. 767.
47. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate—A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **1995**, *57*, 289–300. [\[CrossRef\]](#)
48. Lower Olefins and Aromatics REACH Consortium. *Category j Identity Profile—Low Benzene Naphthas*; Lower Olefins and Aromatics REACH Consortium: Brussels, Belgium, 2021; p. 4.
49. Lower Olefins and Aromatics REACH Consortium. *Category l Identity Profile—Resin Oils and Cyclic Dienes*; Lower Olefins and Aromatics REACH Consortium: Brussels, Belgium, 2021; p. 5.
50. Chen, Z.; Lloyd, D.; Zhou, Y.H.; Chiu, W.A.; Wright, F.A.; Rusyn, I. Risk characterization of environmental samples using in vitro bioactivity and polycyclic aromatic hydrocarbon concentrations data. *Toxicol. Sci.* **2021**, *179*, 108–120. [\[CrossRef\]](#)
51. Mackerer, C.R.; Griffis, L.C.; Grabowski, J.S., Jr.; Reitman, F.A. Petroleum mineral oil refining and evaluation of cancer hazard. *Appl. Occup. Environ. Hyg.* **2003**, *18*, 890–901. [\[CrossRef\]](#) [\[PubMed\]](#)
52. Goyak, K.O.; Kung, M.H.; Chen, M.; Aldous, K.K.; Freeman, J.J. Development of a screening tool to prioritize testing for the carcinogenic hazard of residual aromatic extracts and related petroleum streams. *Toxicol. Lett.* **2016**, *264*, 99–105. [\[CrossRef\]](#)
53. Chen, Z.; Jang, S.; Kaihatu, J.M.; Zhou, Y.H.; Wright, F.A.; Chiu, W.A.; Rusyn, I. Potential human health hazard of post-hurricane harvey sediments in galveston bay and houston ship channel: A case study of using in vitro bioactivity data to inform risk management decisions. *Int. J. Environ. Res. Public Health* **2021**, *18*, 13378. [\[CrossRef\]](#)
54. Wise, S.A.; Rodgers, R.P.; Reddy, C.M.; Nelson, R.K.; Kujawinski, E.B.; Wade, T.L.; Campiglia, A.D.; Liu, Z. Advances in chemical analysis of oil spills since the deepwater horizon disaster. *Crit. Rev. Anal. Chem.* **2022**, 1–60. [\[CrossRef\]](#)
55. Stout, S.A.; Wang, Z. *Standard Handbook oil Spill Environmental Forensics: Fingerprinting and Source Identification*, 2nd ed.; Academic Press: Cambridge, MA, USA, 2016.
56. Chainet, F.; Ponthus, J.; Lienemann, C.P.; Courtiade, M.; Donard, O.F. Combining fourier transform-ion cyclotron resonance/mass spectrometry analysis and kendrick plots for silicon speciation and molecular characterization in petroleum products at trace levels. *Anal. Chem.* **2012**, *84*, 3998–4005. [\[CrossRef\]](#)
57. Tsai, H.D.; House, J.S.; Wright, F.A.; Chiu, W.A.; Rusyn, I. A tiered testing strategy based on in vitro phenotypic and transcriptomic data for selecting representative petroleum UVCBs for toxicity evaluation in vivo. *Toxicol. Sci.* **2023**, *193*, 219–233. [\[CrossRef\]](#) [\[PubMed\]](#)
58. OECD. *Considerations for Assessing the Risks of Combined Exposure to Multiple Chemicals*, Series on Testing and Assessment No. 296; Environment, Health and Safety Division, Ed.; Organisation for Economic Co-Operation and Development: Paris, France, 2018.
59. U. S. Environmental Protection Agency. *A Framework for a Computational Toxicology Research Program in Ord*; U. S. Environmental Protection Agency: Washington, DC, USA, 2003.
60. Efsa Scientific Committee; More, S.J.; Bampidis, V.; Benford, D.; Bragard, C.; Hernandez-Jerez, A.; Bennekou, S.H.; Halldorsson, T.I.; Koutsoumanis, K.P.; Lambre, C.; et al. Guidance document on scientific criteria for grouping chemicals into assessment groups for human risk assessment of combined exposure to multiple chemicals. *EFSA J.* **2021**, *19*, e07033.
61. Verhaar, H.J.; Morroni, J.R.; Reardon, K.F.; Hays, S.M.; Gaver, D.P., Jr.; Carpenter, R.L.; Yang, R.S. A proposed approach to study the toxicology of complex mixtures of petroleum products: The integrated use of qsar, lumping analysis and pbpk/pd modeling. *Environ. Health Perspect.* **1997**, *105* (Suppl. S1), 179–195. [\[PubMed\]](#)
62. Bierkens, J.; Geerts, L. Environmental hazard and risk characterisation of petroleum substances: A guided "walking tour" of petroleum hydrocarbons. *Environ. Int.* **2014**, *66*, 182–193. [\[CrossRef\]](#)
63. Yordanova, D.G.; Patterson, T.J.; North, C.M.; Camenzuli, L.; Chapkanov, A.S.; Pavlov, T.S.; Mekenyan, O.G. Selection of representative constituents for unknown, variable, complex, or biological origin substance assessment based on hierarchical clustering. *Environ. Toxicol. Chem.* **2021**, *40*, 3205–3218. [\[CrossRef\]](#)
64. Redman, A.D.; Parkerton, T.F.; Leon Paumen, M.; Butler, J.D.; Letinski, D.J.; den Haan, K. A re-evaluation of petrotox for predicting acute and chronic toxicity of petroleum substances. *Environ. Toxicol. Chem.* **2017**, *36*, 2245–2252. [\[CrossRef\]](#)
65. McKee, R.H.; Adenuga, M.D.; Carrillo, J.C. Characterization of the toxicological hazards of hydrocarbon solvents. *Crit. Rev. Toxicol.* **2015**, *45*, 273–365. [\[CrossRef\]](#)
66. Gray, T.M.; Simpson, B.J.; Nicolich, M.J.; Murray, F.J.; Verstuyft, A.W.; Roth, R.N.; McKee, R.H. Assessing the mammalian toxicity of high-boiling petroleum substances under the rubric of the hpv program. *Regul. Toxicol. Pharmacol.* **2013**, *67*, S4–S9. [\[CrossRef\]](#)
67. Murray, F.J.; Roth, R.N.; Nicolich, M.J.; Gray, T.M.; Simpson, B.J. The relationship between developmental toxicity and aromatic-ring class profile of high-boiling petroleum substances. *Regul. Toxicol. Pharmacol.* **2013**, *67*, S46–S59. [\[CrossRef\]](#)
68. Nicolich, M.J.; Simpson, B.J.; Murray, F.J.; Roth, R.N.; Gray, T.M. The development of statistical models to determine the relationship between aromatic-ring class profile and repeat-dose and developmental toxicities of high-boiling petroleum substances. *Regul. Toxicol. Pharmacol.* **2013**, *67*, S10–S29. [\[CrossRef\]](#)

69. Kamelia, L.; de Haan, L.; Ketelslegers, H.B.; Rietjens, I.; Boogaard, P.J. In vitro prenatal developmental toxicity induced by some petroleum substances is mediated by their 3- to 7-ring pah constituent with a potential role for the aryl hydrocarbon receptor (AhR). *Toxicol. Lett.* **2019**, *315*, 64–76. [[CrossRef](#)] [[PubMed](#)]
70. Onel, M.; Beykal, B.; Ferguson, K.; Chiu, W.A.; McDonald, T.J.; Zhou, L.; House, J.S.; Wright, F.A.; Sheen, D.A.; Rusyn, I.; et al. Grouping of complex substances using analytical chemistry data: A framework for quantitative evaluation and visualization. *PLoS ONE* **2019**, *14*, e0223517. [[CrossRef](#)] [[PubMed](#)]
71. Marshall, A.G.; Rodgers, R.P. Petroleomics: Chemistry of the underworld. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 18090–18095. [[CrossRef](#)] [[PubMed](#)]
72. CONCAWE. *The Use of the Dimethyl Sulphoxide (DMSO) Extract by the ip 346 Method as an Indicator of the Carcinogenicity of Lubricant Base Oils and Distillate Aromatic Extracts*; CONCAWE: Brussels, Belgium, 1994.
73. Johann, S.; Gossen, M.; Behnisch, P.A.; Hollert, H.; Seiler, T.B. Combining different in vitro bioassays to evaluate genotoxicity of water-accommodated fractions from petroleum products. *Toxics* **2020**, *8*, 45. [[CrossRef](#)]
74. Cordova, A.C.; Ford, L.C.; Valdiviezo, A.; Roman-Hubers, A.T.; McDonald, T.J.; Chiu, W.A.; Rusyn, I. Dosing methods to enable cell-based in vitro testing of complex substances: A case study with a pah mixture. *Toxics* **2022**, *11*, 19. [[CrossRef](#)]
75. Hammershoj, R.; Birch, H.; Sjöholm, K.K.; Mayer, P. Accelerated passive dosing of hydrophobic complex mixtures—controlling the level and composition in aquatic tests. *Environ. Sci. Technol.* **2020**, *54*, 4974–4983. [[CrossRef](#)]
76. Trac, L.N.; Sjöholm, K.K.; Birch, H.; Mayer, P. Passive dosing of petroleum and essential oil uvcb-whole mixture toxicity testing at controlled exposure. *Environ. Sci. Technol.* **2021**, *55*, 6150–6159. [[CrossRef](#)]
77. Stibany, F.; Schmidt, S.N.; Schaffer, A.; Mayer, P. Aquatic toxicity testing of liquid hydrophobic chemicals—Passive dosing exactly at the saturation limit. *Chemosphere* **2017**, *167*, 551–558. [[CrossRef](#)]
78. Smith, K.E.; Oostingh, G.J.; Mayer, P. Passive dosing for producing defined and constant exposure of hydrophobic organic compounds during in vitro toxicity tests. *Chem. Res. Toxicol.* **2010**, *23*, 55–65. [[CrossRef](#)]
79. Wambaugh, J.F.; Hughes, M.F.; Ring, C.L.; MacMillan, D.K.; Ford, J.; Fennell, T.R.; Black, S.R.; Snyder, R.W.; Sipes, N.S.; Wetmore, B.A.; et al. Evaluating in vitro-in vivo extrapolation of toxicokinetics. *Toxicol. Sci.* **2018**, *163*, 152–169. [[CrossRef](#)]
80. Kleinstreuer, N.C.; Yang, J.; Berg, E.L.; Knudsen, T.B.; Richard, A.M.; Martin, M.T.; Reif, D.M.; Judson, R.S.; Polokoff, M.; Dix, D.J.; et al. Phenotypic screening of the toxcast chemical library to classify toxic and therapeutic mechanisms. *Nat. Biotechnol.* **2014**, *32*, 583–591. [[CrossRef](#)] [[PubMed](#)]
81. Anklam, E.; Bahl, M.I.; Ball, R.; Beger, R.D.; Cohen, J.; Fitzpatrick, S.; Girard, P.; Halamoda-Kenzaoui, B.; Hinton, D.; Hirose, A.; et al. Emerging technologies and their impact on regulatory science. *Exp. Biol. Med.* **2022**, *247*, 1–75. [[CrossRef](#)] [[PubMed](#)]
82. Paul Friedman, K.; Gagne, M.; Loo, L.H.; Karamertzanis, P.; Netzeva, T.; Sobanski, T.; Franzosa, J.A.; Richard, A.M.; Lougee, R.R.; Gissi, A.; et al. Utility of in vitro bioactivity as a lower bound estimate of in vivo adverse effect levels and in risk-based prioritization. *Toxicol. Sci.* **2020**, *173*, 202–225. [[CrossRef](#)] [[PubMed](#)]
83. U.S. EPA. *Availability of New Approach Methodologies (NAMS) in the Endocrine Disruptor Screening Program (EDSP)*; Office of Chemical Safety and Pollution Prevention, Office of Research and Development: Washington, DC, USA, 2022.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.