Prioritization of unknown features based on predicted toxicity categories

Viktoriia Turkina,[*,†] Jelle T. Gringhuis,[†] Sanne Boot,[†] Annemieke Petrignani,[†] Garry Corthals,[†] Antonia Praetorius,[‡] Jake W. O'Brien,[¶,†] and Saer Samanipour[*,†,§,‖]

[†] *Van 't Hoff Institute for Molecular Sciences (HIMS), University of Amsterdam, 1090 GD, Amsterdam, the Netherlands*

[‡] *Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, 1090 GE, Amsterdam, the Netherlands*

[¶] *Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland, 20 Cornwall Street, Woolloongabba, QLD, 4102, Australia*

[§] *UvA Data Science Center, University of Amsterdam, Amsterdam*

[‖] *Queensland Alliance for Environmental Health Sciences (QAEHS), 20 Cornwall Street, Woolloongabba, QLD, 4102, Australia*

E-mail: v.turkina@uva.nl; s.samanipour@uva.nl

Abstract

Complex environmental samples contain a diverse array of known and unknown constituents. While Liquid Chromatography coupled with High-Resolution Mass Spectrometry (LC-HRMS) Non-Targeted Analysis (NTA) has emerged as an essential tool for the comprehensive study of such samples, the identification of individual constituents remains a significant challenge, primarily due to the vast number of detected features in each sample. To address this, prioritization strategies are frequently employed to narrow the focus to the most relevant features for further analysis.

In this study, we developed a novel prioritization strategy that directly links fragmentation and chromatographic data to aquatic toxicity categories, bypassing the need for individual compound identification. Given that features are not always well-characterized through fragmentation, we created two models: 1) a Random Forest Classification (RFC) model, which classifies fish toxicity categories based on MS1, retention, and fragmentation data—expressed as cumulative neutral losses (CNLs)—when fragmentation information is available, and 2) a Kernel Density Estimation (KDE) model that relies solely on retention time and MS1 data when fragmentation is absent. Both models demonstrated accuracy comparable to structure-based prediction methods. We further tested the models on a pesticide mixture in a tea extract measured by LC-HRMS, where the CNLs-based RFC model achieved 0.76 accuracy and the KDE model reached 0.61, showcasing their robust performance in real-world applications.

Synopsis

This study presents a novel prioritization strategy for LC-HRMS NTA, which directly links chromatographic and fragmentation data to the activity of unknown features, bypassing the need for their identification

Introduction

The increasing number and variety of produced chemicals hinders the investigation and assessment of their exposure levels and impacts on human and environmental health.[1–3] Only specific groups of chemicals are routinely monitored as a regulatory measure, therefore, limiting the exposure assessment by excluding a diverse mixture of chemicals.[4,5] In order to investigate chemical exposure realistically, accounting for the diversity of structures and their properties in samples necessitates the application of unbiased approaches.[5–7]

A major portion of contaminants of emerging concern (CECs) is covered by semi-polar and polar organic compounds.[8,9] High-resolution tandem mass spectrometry coupled with reverse phase liquid chromatography (LC-HRMS) is a well-known technique to analyze such chemicals.[9,10] To limit the bias towards a small set of standardized analytes, the LC-HRMS nontargeted (NTA) approach has been successfully developed and employed in environmental studies.[11–16] NTA does not require prior knowledge of the chemicals present in a sample and enables the analysis of all detectable compounds. However, the complex nature of environmental samples results in highly convoluted data, requiring careful and robust preprocessing.[17–20] Generally the number of detected features in NTA—data points constructed by retention time, chromatographic peak intensity, and m/z of precursor ions—often counts in the hundreds or thousands, though typically fewer than 5% of these can be identified.[19,21–23]

Confident identification of detected features requires considerable computational and research resources.[17,23,24] Therefore, prioritization is used as strategy to focus the available resources on the most relevant features for a given study's objectives.[25] The prioritized species of interest undergo more detailed and thorough investigation. There are two ways of prioritization: online and offline. Online prioritization refers to a variety of parameters to acquire the data, e.g. to perform data-dependent acquisition,[25] consequently, limiting the acquired data according to predefined parameters.[21] On the other hand, offline prioritization strategies employ post hoc analysis of MS1 and MS2 (both data-dependent (DDA) and data-independent (DIA) acquisition) data, including intensity-based prioritization, statistical analysis, structural, and quantitative structure-activity relationship (QSAR) evaluation.[6] The online approach results in a cleaner dataset for further processing with the risk of losing valuable information about highly relevant features. Thus, the offline approach is more advantageous for environmental studies and can also be used for retrospective studies.[26]

Environmental studies primarily focus on the properties and potential activities of detected chemicals, including their ecological toxicity.[27,28] As a result, prioritization strategies often aim to highlight features that may pose significant (eco)toxicological risks.[29] For example, Peets et al. developed the MS2Tox R package, which uses a regression model to predict toxicity (-[LOG(mM)]) for organisms like fish, water fleas, and algae.[30,31] Similarly, Rahu et al.. developed an algorithm to predict endocrine-disrupting activity using binary or multi-label classification,[32] while Arturi and Hollender created MLinvitroTox, a tool to classify chemicals as toxic or non-toxic based on nearly 400 target-specific and over 100 cytotoxic endpoints.[33]

All these studies rely on predicting molecular formulas and fingerprints from MS2 spectra using tools like CSI:FingerID/SIRIUS.[34] Therefore, the accuracy of these models is closely tied to the precision of the predicted molecular fingerprints. Incorrectly calculated fingerprints can significantly distort the predictions, and this accuracy can be compromised by various experimental factors, such as incomplete fragment detection, false positive fragments, or missing characteristic fragments due to instrumental noise. Bypassing the molecular fingerprint prediction and instead predicting chemical activity directly from chromatographic and MS data potentially reduces this uncertainty and enhances the interpretability of results.

Aquatic toxicity has been shown to be related to hydrophobicity of the chemicals, their molecular weights and structural alerts.[35–37] Each detected feature indirectly includes structural information: chromatographic retention times and mass spectral (MS) information, like accurate mass (MS1) and fragmentation spectra (MS2). Therefore, this information can be used to assess the toxicity of the chemicals without any knowledge of molecular structure, formula or fingerprint.

In this study, we developed a prioritization strategy based on acute aquatic toxicity using chromatographic and fragmentation information of unknown features. Instead of identifying each feature, we directly described their activity without any knowledge of molecular structure. We constructed two models: a Random Forest Classification (RFC) model to assign toxicity categories based on fragmentation data, and a Kernel Density Estimation (KDE) model to determine the probability of features belonging to specific toxicity classes.

These algorithms allow us to characterize unknown features at both the mass spectral and chromatographic levels. The performance of the constructed models was evaluated both on spectra from databases with measured toxicity values and on tea extracts spiked with 253 pesticide standards.

Methods

Since fragmentation information in NTA is often incomplete, we developed a probabilistic algorithm to predict toxicity categories that relies solely on chromatographic data, specifically retention time and MS1 spectra. This algorithm uses KDE in both the retention and mass domains to identify regions of chromatograms that are more likely to correspond to higher or lower toxicity, effectively mapping toxicity. Additionally in the presence of reliable fragmentation data, the CNLs-based RFC model can be applied to categorize toxicity based on detected fragments.

Overall workflow

To develop the models we employed a semi-supervised approach. First, a structure-based classification model was built using an experimentally generated acute fish toxicity dataset and molecular fingerprints. This model was then used to expand toxicity knowledge by predicting the toxicity category for chemicals from large, existing spectral and environmentally relevant databases. The CompTox dataset[38] was used to develop a KDE model, while for the CNLs-based model, HRMS spectra were collected from sources such as MassBank, MoNa, GNPS, and NIST.[39–42] Finally, the results of the developed models were compared and validated using an experimentally acquired pesticide mixture in a tea extract (Figure 1).

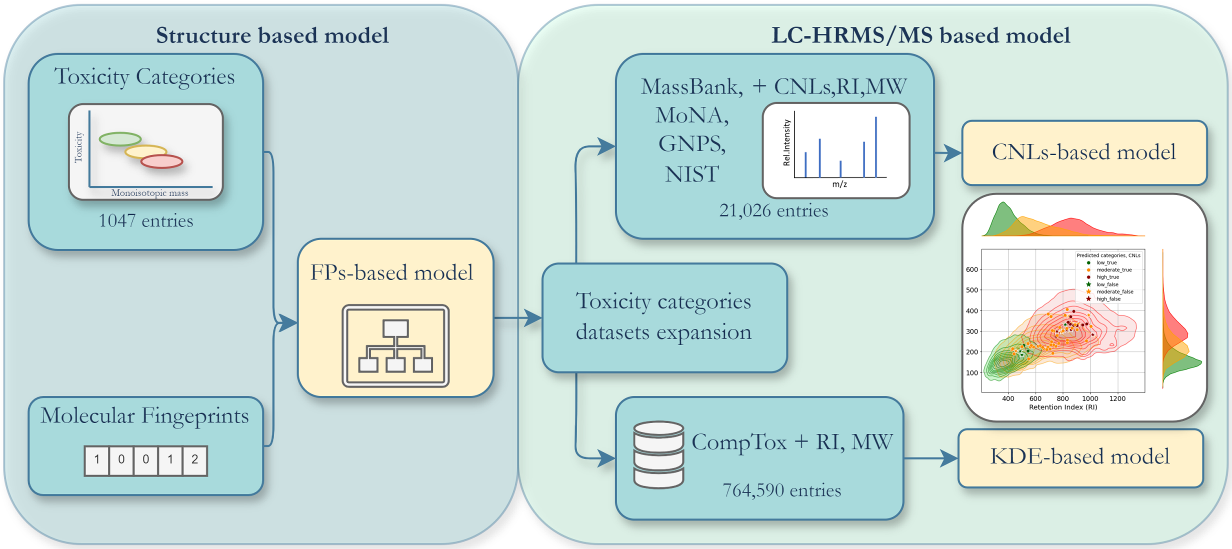


Figure 1: Overall workflow. A structure-based model was developed using a fish toxicity dataset and molecular fingerprints (FPs). This FPs-based model was then applied to the CompTox and Spectral databases, incorporating additional information such as molecular weight (MW) and retention indices (RI). A KDE-based model was constructed based on this data. Furthermore, fragmentation information (i.e., characteristic neutral losses, CNLs) was integrated to develop a CNLs-based model.

Datasets

For the model development, validation, and testing we employed four different datasets namely: a set of experimental acute fish toxicity values - referred to as the fish toxicity data set, the CompTox dataset, a Cumulative Neutral Loss (CNL) dataset, and experimentally acquired LC-HRMS chromatogram of standard pesticide mixture in tea extracts - referred to as pesticide mixture dataset.

Fish toxicity dataset

The fish toxicity dataset used in this study was compiled from two sources: Cassotti et al.[43] and Schür et al.[44] These sources are themselves compilations of numerous empirical studies, relying on data collected from a wide range of experiments. The dataset includes concentration values, expressed in LOG(mg/L), representing the dose that causes 50% mortality in fathead minnows (Pimephales promelas) over a 96-hour exposure period (LC50). In total 1047 chemicals in the dataset come from various chemical families, including pharmaceuticals, pesticides, conventional persistent organic pollutants, and industrial chemicals.[45] Due to the inherent variability in *in vivo* experiments, where results can vary even under identical conditions, the dataset includes multiple repeated measurements for each chemical. The median value from these repetitions was used as the reference LC50 value for each compound.

The CompTox dataset

The CompTox Chemicals Dashboard, developed by the U.S. Environmental Protection Agency (EPA), is a vast database that has been curated over 15 years through both manual and automated efforts as part of the EPA's DSSTox project. It contains detailed information on 764,590 chemicals (as of March 2024), including identifiers such as CAS numbers and SMILES, synonyms, and key properties like solubility, logKow, and molecular weight. Additionally, it provides data on physicochemical characteristics, environmental fate, exposure, usage, and available toxicity results from both *in vivo* and *in vitro* studies. For our purposes, the CompTox dataset was filtered to exclude salts, metalloids, and compounds containing elements not typically found in fish toxicity datasets or not analyzable via reverse phased liquid chromatography (RPLC), leaving only chemicals composed of H, B, C, N, O, F, Si, P, S, Cl, Br, and I.

CNLs dataset

The CNLs were used in place of fragments to reduce the dimensionality of the dataset while retaining the structural information embedded in the spectra. The CNLs dataset was generated from entries in the MassBank EU, MoNA, GNPS, and NIST20 databases. These entries were obtained using electrospray ionization (positive mode) with a mass resolution of ≥ 5000, proton adducts, and included data from various mass analyzers (e.g., Q-TOF and Orbitrap) with different collision energies. To ensure data quality, only spectra with at least three recorded fragments were included. When multiple spectra were available for a single compound, they were merged using a ±0.01 Da mass window, yielding 21026 unique spectra. Rare fragments, present in only a single instance across multiple entries, were discarded to focus on more representative data.

For each compound, the CNLs were calculated by subtracting the fragment m/z values from the precursor ion mass. These CNLs were then converted into bit vectors representing masses from 0 to 1000 Da with a step size of 0.01 Da (± 5 mDa mass tolerance). In this representation, a value of 1 indicated the presence of a CNL in the spectrum, while a value of 0 indicated its absence. To further refine the dataset, CNLs that occurred in fewer than 100 spectra were removed, resulting in a final dataset of 2440 common CNLs, with mass values ranging from 1.03 to 382.19 Da.

In addition to the CNL bit vectors, the monoisotopic mass and predicted retention index were added as continuous features. To calculate monoisotopic mass, the mass of proton, 1.008 Da, was subtracted from the measured m/z value of the proton adduct. To predict retention indices the previously reported by van Herwerden et al. model based on molecular fingerprints and retention indices calculated on the cocamide scale was employed.[47] This combination of features allowed us to incorporate key structural and chromatographic information into the model while minimizing the number of variables. To interpret the importance of individual CNLs, a search was conducted in the CompTox Chemicals Dashboard to identify relevant substructures. The prominent substructures frequently identified in the search results were used to inform model interpretation.

Pesticide mixture dataset

The prioritization models were further tested on experimentally acquired datasets of varying levels of tea extract matrix spiked with 253 pesticide standards. The samples were analysed by the previously reported RPLC-ESI-Q-TOF method.[48] The three standard mixtures of pesticides from Neochema were diluted with filtered water/ethanol solution (50%:50%, v/v, Blank) or filtered tea extract that was further diluted 1:10 or 1:100 with Blank to final analyte concentrations of 100 μg/L. Each sample was analysed twice. A detailed description of the samples, acquisition method, and data processing is provided in Supporting Information, section S1.

To obtain features and fragmentation information of detected pesticides, suspect screening was performed by applying the Universal Library Search Algorithm (ULSA).[49] A suspect list of the pesticides was compiled from the MS2 spectra found in the MassBank EU, MoNA, and NIST20 databases with a resolution higher than 5000 for each of these chemicals. To minimize detection of false positive fragments a mass tolerance of 0.010 Da and a minimum MS1 intensity of 2000 were applied for matching acquired fragments with suspects. Additionally, to increase the confidence of correctly detected fragments, a retention time tolerance of 0.5 min was employed for comparison of MS1 and fragments profiles.

To convert retention times in retention indices based on the cocamide series a linear regression model was built between retention time of 18 detected pesticides in all six chromatograms and their predicted retention indices (Figure S1. The model was applied to the rest of the detected pesticides (Equation S1). Finally, the developed models were applied to the detected suspects to assign toxicity categories. Since 96h LC50 mortality of fathead minnows were unknown for detected pesticides, the predicted toxicity categories based on the structure presented via molecular fingerprints were used as benchmark toxicity categories.

Modeling

Assigning toxicity categories

A previous study demonstrated the advantage of using clustering algorithms as an alternative to expert-based toxicity categorization, which typically relies on fixed LC50 or EC50 thresholds.[45] These expert-defined thresholds often fail to account for the uncertainty in LC50 measurements or the structural similarities between chemicals. The k-means clustering algorithm, an unsupervised iterative method, clusters data based on the distances between measurements and a set of user-defined centroids. This approach takes advantage of structural similarities between chemicals when forming clusters.

In this study, we applied k-means clustering to categorize chemicals from a fish toxicity dataset, using scaled LC50 values [LOG(mg/L)], monoisotopic mass, and six elemental mass defects (EMDs): CO, CCl, CN, CS, CF, and CH as input features.[50] Since LC50 values primarily determine toxicity while structural information captures relationships between chemical structures, we assigned a weight of 1/7 to each of the monoisotopic mass and the six EMDs. This ensured that LC50 values and structural information contributed equally.

The algorithm converged after 12 iterations with an inertia of 323, forming three distinct clusters (Figure S2). These clusters were then labeled as low, moderate, and high toxicity, based on their positions along the y-axis (LC50 values). The final clustering allowed clear differentiation between low toxicity (n=238), high toxicity (n=202), and moderate toxicity (n=467) chemicals, providing a robust categorization informed by both toxicity data and structural characteristics.

Classification models

For modeling, a Random Forest Classifier (RFC) algorithm was implemented in Julia with scikit-learn.[51] Random Forest is a supervised algorithm that constructs several unique decision trees from bootstrap data. After model development, the major voting of predictions resulting from the individual trees is used to produce the final RFC model prediction.[52]

Fingerprint-based model: To train the fingerprint-based (FPs-based) model, the fish toxicity dataset (n = 1047) was split into a training set (n = 854), test set (n = 95), and global test set (n = 98) via stratified sampling (Figure S2). The input for the classification model was the optimizied fingerprint[53] and the output was toxicity categories. The molecular fingerprints were calculated based on six different non-hashed molecular fingerprints, namely, Atom Pair 2D Count (AP2DC),[54] Electrotopological state (E-state),[55] Klekotha-Roth Count (KRC),[56] Molecular Access Systems (MACCS),[57] PubChem,[58] and Substructure Keys Count

(SSC) FPs.[59] 340 bits out of combined 7073 were selected based on relevance to acute fish toxicity prediction.[53] The model was optimized on a training set, validated with a test set and finally tested with a global test set. Hyperparameter optimisation for the toxicity random forest classification models consisted of a number of estimators varying between 100 and 400 with a step of 100, minimum samples leaf between 2 and 10 with a step of 2, a maximum number of features of sqrt, log2, a maximum depth between 10 and 20 with a step of 2. Cross-validation of the hyperparameter optimisation was 3-fold. This FPs-based model then was utilized to predict the toxicity categories for CompTox and CNLs datasets.

CNLs-based model: To train the CNLs-based model the CNLs dataset was randomly split into a training set (n = 18835), test set (n = 2093), and global test set (n = 98). The input for the classification model was the CNL values converted to a bit vector, monoisotopic mass, and predicted retention indices. The output of the CNL-based model was predicted probability of assignment to toxicity categories. The probability was further converted into predicted category based on the defined threshold. The threshold was calculated based on the difference in probabilities between all three categories using test set only. Hyperparameter optimisation was performed in the same manner as described in section ”Fingerprint-based model”. Furthermore, to account for a class imbalance the weighted version of RFC was employed.

KDE based model

The KDE algorithm[60] was used for toxicity categories mapping on chromatographic dimensions using monoisotopic mass, predicted retention indices, and toxicity categories. Given fitted KDE, we calculated the probability for each chemical to belong to one of the predetermined toxicity categories using the formula;

$$\hat{f}(\mathbf{x};\mathbf{H}) = \frac{1}{n|\mathbf{H}|^{1/2}} \sum_{i=1}^{n} K\left(\mathbf{H}^{-1/2}(\mathbf{x}-\mathbf{X}_i)\right) \tag{1}$$

where $\hat{f}(\mathbf{x})$ represents the estimated density at point x, n is the number of data points, d is the number of dimensions, H is the $d \times d$ bandwidth matrix, K is the multivariate kernel function, $|\mathbf{H}|$ is the determinant of the bandwidth matrix H, and $\mathbf{H}^{-1/2}$ is the matrix square root of the inverse of H. Because the classification data is unbalanced, the number of samples in each class was used as weights. This effectively cancels out the $1/n$ in the formula and allows for comparison between the kernel densities.

Bandwidth selection was carried out using Silverman's rule of thumb;

$$sigma_j = \left(\frac{4\hat{\sigma}_j^5}{3n_k}\right)^{\frac{1}{5}} \tag{2}$$

where σ_j denotes the standard deviation of the j-th feature in class C_k. This method ensures a smooth estimation of the density function, aiding in the effective prediction of toxicity classifications.

To evaluate the performance of the KDE based model, densities for all three toxicity classes were calculated for the fish toxicity dataset as only available experimentally measured values. The toxicity category was assigned based on the highest calculated probability (density).

Applicability domain calculations

The Applicability Domain (AD) defines the parameter space within which a predictive model is expected to make reliable predictions. In this study, we utilized the leverage approach to define and evaluate the AD of our predictive model.[61] The AD was assessed by calculating the leverage for each sample using the equation;

$$h_i = x_i^\top (X^\top X)^{-1} x_i \tag{3}$$

where x_i is the vector of predictor variables for the i-th sample, and X is the matrix of predictor variables for all training samples. The leverage values indicate the influence of each sample on the model's predictions.

The warning leverage threshold h^* is determined to identify samples that have high leverage and may lead to unreliable predictions. The threshold is typically set to three times the average leverage in the training set;

$$h^* = 3 \times h_{\text{avg}} \tag{4}$$

Thus, predictions for samples with leverage values greater than these thresholds are considered less reliable and are flagged as outside the AD. To ensure these do not affect the kernel density estimation, the weights for each sample were set at $1/h$.

Performance assessment

For the evaluation of the performance of the multiclass classification model, several metrics were computed, including precision, recall, F1 score, and accuracy. The classification metrics provide an overview of the model's performance for each class. The calculations for these metrics are as follows:

Precision: The ratio of correctly predicted positive observations to the total predicted positives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

Recall (Sensitivity): The ratio of correctly predicted positive observations to all observations in the actual class.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

F1 Score: The weighted average of Precision and Recall.

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

Accuracy: The ratio of correctly predicted observations to the total observations

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{8}$$

To further validate the classification models, the confusion matrix was computed, which provided a detailed breakdown of the model's predictions compared to the actual class labels. Because of the class imbalance, the confusion matrix was normalized to represent fractions, making it easier to interpret the proportion of correct and incorrect predictions for each class.

Calculations

All calculations were performed on a personal computer (PC) with an Intel Core i7-1260P central processing unit and 32 GB of RAM operating Windows 10 Education version 22H2. All of the data processing and statistical analyses were performed using Julia language version 1.6.7. The computational algorithms were computed using the ScikitLearn.jl version 0.7 library. The fish toxicity data set, the filtered CompTox data set, and the pesticide mixture dataset are publicaly available at `https://doi.org/10.5281/zenodo.14188167`. Additionally, a script to perform the calculations can be accessed at at`https://bitbucket.org/viktoriiaturkina/toxmap_prioritization.jl/src/main`.

Results and Discussion

In this study we developed two classification algorithms, machine learning and probabilistic, to predict fish toxicity categories based only on the chromatographic and fragmentation without identified structural information. Additionally, to assign the toxicity categories we trained the classification model based on the molecular FPs. For the probabilistic algorithm, the CompTox database was mapped on RPLC-HRMS space and KDE was fitted to each toxicity category assigned based on the known chemical structures (FPs). For the machine learning algorithm, the RFC model was trained on precursor ion mass (proton adduct), retention indices, and fragmentation information in the form of CNLs bit vector.

FPs-based model

The fingerprint-based model developed in this study was an RFC model trained on 900 structures described with optimized molecular fingerprint[53] and their experimental LC50 [LOG(mg/L)] values.[43–45] Optimized hyperparameters of the model were determined as 300 estimators, a minimum of 10 samples split, a maximum depth of 20, and sqrt as the number of features for split. To assess the performance of the model confusion matrix, accuracy, precision, recall, and f1-score for each toxicity category were calculated (Table 1).

The accuracy of the optimized model was 0.86 for the training set, 0.75 for the test set and 0.70 as a mean accuracy of 3-fold cross-validation, which was comparable to previously reported QSAR models for toxicity predictions.[62,63] Performance assessment of the prediction results for chemicals from the test set showed that the model consistently performed well for low and moderate toxicity categories with an F1-Score 0.80 and 0.73 correspondingly. However, the F1-Score for the high toxicity category was lower compared to the other two categories, 0.54. Based on the confusion matrix and leverage analysis (Figure S4), the lower recall for the high-toxicity class in the test set was primarily due to the misclassification of compounds with relatively low leverage values into the moderate-toxicity class. Compounds with lower leverage generally have lower monoisotopic mass and show greater similarity to compounds in the moderate and low-toxicity categories, that influences the accuracy of high-toxicity category predictions. Moreover, the chemicals assigned to the high-toxicity category exhibited greater variability in the structural representation with the smallest number of chemicals included, 145 for the training set and 13 for the test set.

To ensure a broader applicability domain for predictive purposes, the final fingerprint-based classification model was trained using a combination of the training and test sets. A separate global test set, containing measured toxicity values and available LC-HRMS data, was reserved for the final model assessment. The validated model was then used to predict toxicity categories for the CompTox and CNLs datasets, expanding the dataset and supporting the development of a probabilistic approach and the training of a CNLs-based model.

Table 1: The performance assessment of the fingerprint-based classification model.

	Class	Precision	Recall	F1-Score	Support	Accuracy
Training set	Low	0.90	0.83	0.87	263	0.86
	Moderate	0.83	0.94	0.88	446	
	High	0.92	0.66	0.77	145	
Test set	Low	0.72	0.75	0.73	24	0.75
	Moderate	0.81	0.79	0.80	58	
	High	0.54	0.54	0.54	13	
Combined dataset	Low	0.88	0.83	0.85	287	0.85
	Moderate	0.83	0.93	0.87	504	
	High	0.88	0.65	0.75	158	

KDE-based model

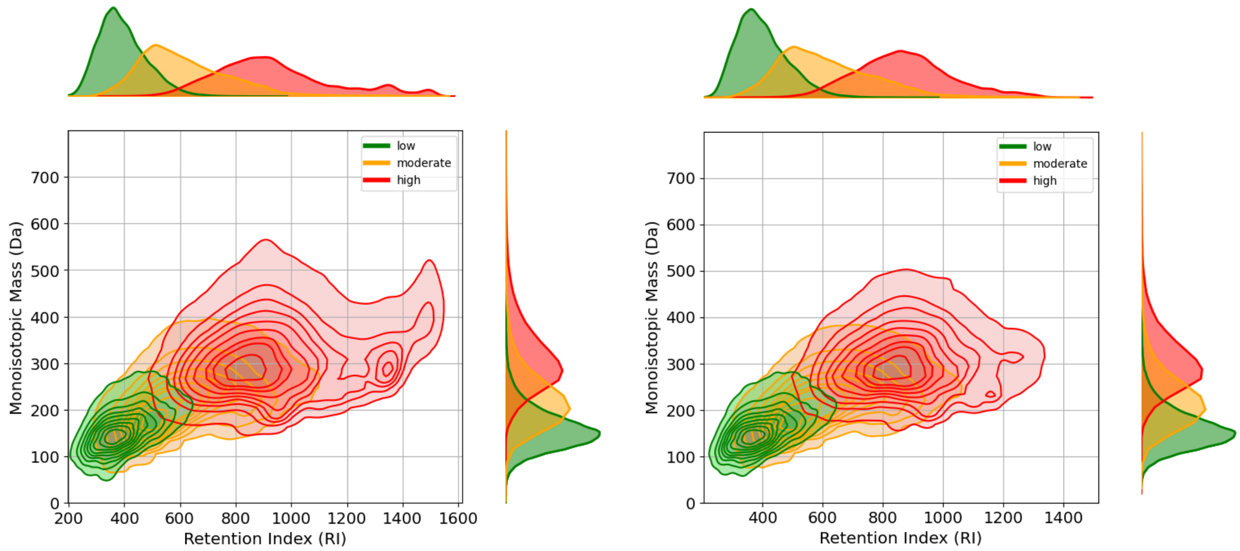
To create a KDE algorithm the toxicity categories and retention indices were predicted for each chemical from the CompTox dataset based on their structures. This information was then used to map toxicity regions within the RPLC-MS domain, linking toxicity categories to features in the chromatogram.

Out of the 764,590 chemicals of the filtered Comptox database, 83,969 chemicals (10.9%) were classified as low toxicity, 458,302 chemicals (59.9%) as moderate toxicity, and 222,319 chemicals (29.0%) as high toxicity class. However, only 20.8% of the CompTox dataset is well represented by the training set, meaning the calculated leverage for these chemicals is less than the leverage threshold, 3 times of the mean value of leverage in the training set, 1.02. To decrease the uncertainty of KDE we used calculated leverage values as weights to account for the less certain prediction of the fingerprint-based model. Therefore, chemicals that were well represented in the training set, gave more weights for the density estimation and more accurately represented the results. The weights used were $1/h$ with h being the leverage of a compound (Figure 2a).

The CompTox dataset contains a wide variety of structures, with the range of monoisotopic mass between 12.0 and 5,628.0 Da, and the predicted retention indices from 178.6 to 1,530.5 RI value. (Figure S3) A partion of this space cannot be analysed with RPLC, either due to insufficient retention (e.g., highly hydrophilic chemicals, often with low masses, i.e., $<$ 70 Da) or excessive retention (e.g., highly hydrophobic chemicals with high mass and high retention indices). Several recent studies investigated algorithms for assigning chemicals to be analyzable with RPLC or other types of chromatography.[46,64–66] We applied the machine learning algorithm developed by van Herwerden et al. to the CompTox dataset to filter out chemicals that will not be analysed and therefore detected using RPLC. Of the 671,395 chemicals predicted to fall within the RPLC space, 78,087 (11.6%) were classified as low toxicity, 431,089 (64.2%) as moderate toxicity, and 162,219 (24.2%) as high toxicity category. Thus, mainly compounds predicted as highly toxic fell outside of RPLC space (Figure 2b).

Overall accuracy of the algorithm was 0.69. Similar results were observed for overall precision, recall and f1-score. The metrics of predictions for moderate and high toxicity categories showed comparable performance with precision 0.75-0.76, recall 0.6-0.64, and f1-score 0.67-0.70. Results for low-toxicity group with 0.58 for precision and 0.84 for recall, differed from the results for the rest of the categories. Overall, the algorithm showed consistently good performance among all categories.

Only 1.3% of chemicals from low-toxicity group were classified as highly toxic chemicals and 9.4% of high-toxicity chemicals were classified as low-toxicity chemicals. This indicated that low toxicity compounds generally do not occupy the region of high toxicity class compounds. The larger error occurs in misclassification between moderate and low toxicity groups (28%) and between high and moderate toxicity groups (30%) (Figure S5). The developed algorithm based on KDE was as accurate as the FPs-based model, even though no actual structural information was provided to it.



(a) **(b)**

Figure 2: KDE joint plots of filtered on RPLC space Comptox dataset (a), and applied to KDE weights based on calculated leverage (b)

CNLs-based model

For CNLs-based model construction, we predicted toxicity categories and retention indices for the CNLs dataset. Out of 20,928 chemicals in the dataset 3,152 chemicals were predicted as low-toxicity chemicals, 12,662 as moderate, and 5,114 as high toxicity. The number of chemicals assigned to different groups shows an imbalance in the dataset; therefore to train the model, the weighted version of RFC was employed.

The optimized the CNLs-based model showed an overall accuracy of 0.76 for training set and 0.72 for test set. Therefore, the final model showed a high level of accuracy, comparable to fingerprint-based classification model. Moreover, precision, recall, and F1-score for training and test set show consistent result for all three toxicity groups (Table S2), indicating excellent performance of the model regardless of the class imbalance.

This model consisted of 600 trees, a minimum of two samples in each leaf, a maximum depth of 20, and ”sqrt” as a maximum number of features. A difference in probability threshold of 0.08 was applied to the classification results: if the difference in predicted probabilities across all three toxicity categories was less than 0.08 (or 8%), the chemical was assigned to the moderate toxicity category. This is because, in the KDE model, the moderate toxicity category is the most uncertain and tends to overlap with both the low and high toxicity categories.

The model explained 85% of the variance using 592 variables, including CNLs with masses ranging from 2.01 Da to 354.19 Da. Based on impurity decrease across trees, the largest contributions to the model came from retention index (RI) and accurate mass, cumulatively accounting for 36%. These were followed by contributions from CNLs with masses such as 17.03 Da (1.7%), likely corresponding to NH_3, 45.01 Da (CH_3NO), 27.01 Da (HCN), and 63.03 Da (HNO_2 or CH_2SO), each contributing around 1%. While these structures are plausible, alternative small molecules corresponding to these masses cannot be disregarded. Higher mass CNLs had an additional contribution, including 294.11 Da, potentially representing multiple aromatic rings and/or benzyloxy substructures, 216.13 Da, possibly linked to aromatic heterocycles like indole or quinoline, and 174.11 Da, which may correspond to a thiocarboxylic ester. This indicates that the model primarily relies on MS1 data, supplemented by structural information that linked to the potentially toxicity-relevant substructures.

Comparison of the KDE and CNLs-based models

Based on the confusion matrices for training and test set the comparable performance to KDE algorithm was observed. The main misclassifications occur for a moderate toxicity group, while low toxicity barely ever classified as high toxicity group (1%) and vice versa (0%) (Figure S6). However, the test and training set for the KDE model and CNLs based model were different therefore one-to-one comparison cannot be performed using these datasets. Thus, to further evaluate and compare prediction power between all three generated models we used a global test set. These 98 entries had both the CNL bit vector and fish toxicity experimentally determined values and were completely unknown to any of the models.

When comparing KDE and CNLs-based models with the FPs-model, the confusion matrices indicate that the FPs-based model produced comparable results in predicting low and high toxicity categories. This is a significant outcome, given that the KDE and CNLs-based models do not account for the structural representation of the compounds and are directly linked to chromatographic and mass spectrometry data. However, FPs-based model outperformed the other models in predicting the moderate toxicity group, achieving an accuracy of 0.88 compared to 0.47 for the KDE model and 0.43 for the CNL model (Figure S6). This can be explained by the substantial overlap in the RPLC-MS space assigned to the moderate toxicity category, which has a mass range of 110.11 to 336.30 and retention indices between 285.28 and 1139.23, with the regions for low and high toxicity categories.

To further demonstrate the applicability of the developed approach, we applied it to an experimentally acquired dataset of a standard pesticide mixture at varying concentrations in a tea extract. Between 141 and 155 pesticides were detected out of 253 in the standard mixture across all three levels of the tea matrix (no tea, 100-fold dilution, and 10-fold dilution). The minimum number of 141 pesticides was detected in the 100-fold diluted tea, while 155 were detected in the 10-fold diluted tea. No clear trend was observed between the number of detected pesticides and the matrix dilution. Additionally, the number of detected and matched fragments was compared across the three matrices. The median number of fragments per pesticide was 12 to 13, with the 25th and 75th percentiles ranging from 7 to 21-23. Overall, no significant influence of the matrix level was observed on the number of detected pesticides or the number of corresponding fragments. Since matrix effects and other parameters can influence not only the number but also the quality of detected fragments, the CNLs-based model was applied independently to each chromatogram (Table S3, Figure S7).

To keep prediction results comparable between the models based on KDE and CNLs, we retained only the features corresponding to the standard pesticide mixture detected in all samples. As part of the ULSA workflow, Suspect Screening was applied to identify corresponding features at both the chromatographic and fragmentation levels. In total 84 pesticides were detected in all three different matrices, 18 of which were used to establish a correlation between retention time and retention indices, yielding an $R^2 = 0.91$. The remaining 66 pesticides were used for further testing. Based on molecular structure, the FPs-based model classified the 66 pesticides as follows: 3 assigned to the low toxicity category, 48 to moderate toxicity, and 15 to high toxicity.

When applied to the detected pesticides, the CNLs-based model achieved an average accuracy of 0.76-0.80, while the KDE model had a lower accuracy of 0.61. Additionally, the accuracy of the CNLs model was not significantly influenced by different matrix levels, as variations in accuracy across repeated measurements were within 3%, similar to the variation between different matrices. This suggests that the CNLs model maintains stability and reliability across diverse experimental conditions, making it particularly useful for complex environmental matrices. We examined chemicals that were misclassified by either the KDE-based or CNLs-based models, meaning they were assigned to a different toxicity category compared to the benchmark (FPs-based) model. For instance, Metsulfuron-methyl (INCHIKEY: RSMUVYRMZCOLBH-UHFFFAOYSA-N) was predicted to belong to the high toxicity category by the KDE-based model, while both the CNLs-based and FPs-based models classified it as moderate toxicity. While there are no measured LC50 values available for fathead minnows, experimental LC50 values for Bluegill Sunfish and Rainbow Trout reported in the Pesticide Ecotoxicity Database from the EPA indicate LC50 values greater than 2 [LOG(mg/L)] for a 96-hour exposure. Based on classification criteria, chemicals with LC50 values above 2 [LOG(mg/L)] and a monoisotopic mass of 381.07 Da typically fall into the moderate or low toxicity categories for fathead minnows. The CNLs and FPs models, therefore, offered a more accurate classification in this case, reflecting the expected behavior of Metsulfuron-methyl in aquatic species. On the other hand, Thiabendazole (INCHIKEY:

WJCNZQLZVWNLKY-UHFFFAOYSA-N) was assigned to the moderate toxicity category by the CNLs-based model, while the KDE- and FPs-based models classified it as low toxicity. Again, no data were available for fathead minnows, but the LC50 values for Bluegill Sunfish and Rainbow Trout, as recorded by the EPA, range between -0.25 and 1.8 [LOG(mg/L)]. With an accurate mass of 201.04 Da, this chemical is typically assigned to either moderate or low toxicity categories. The discrepancy in classification highlights the challenges of classifying compounds that fall near the boundary between toxicity classes. In this case, the CNLs model's assignment of moderate toxicity may be more reflective of the compound's potential toxicity, given the reported LC50 values and structural characteristics.

Both models performed well on experimental and global test datasets. Overall, the CNLs-based model outperformed the KDE-based model in the pesticide dataset, demonstrating superior predictive power and more consistent classification.

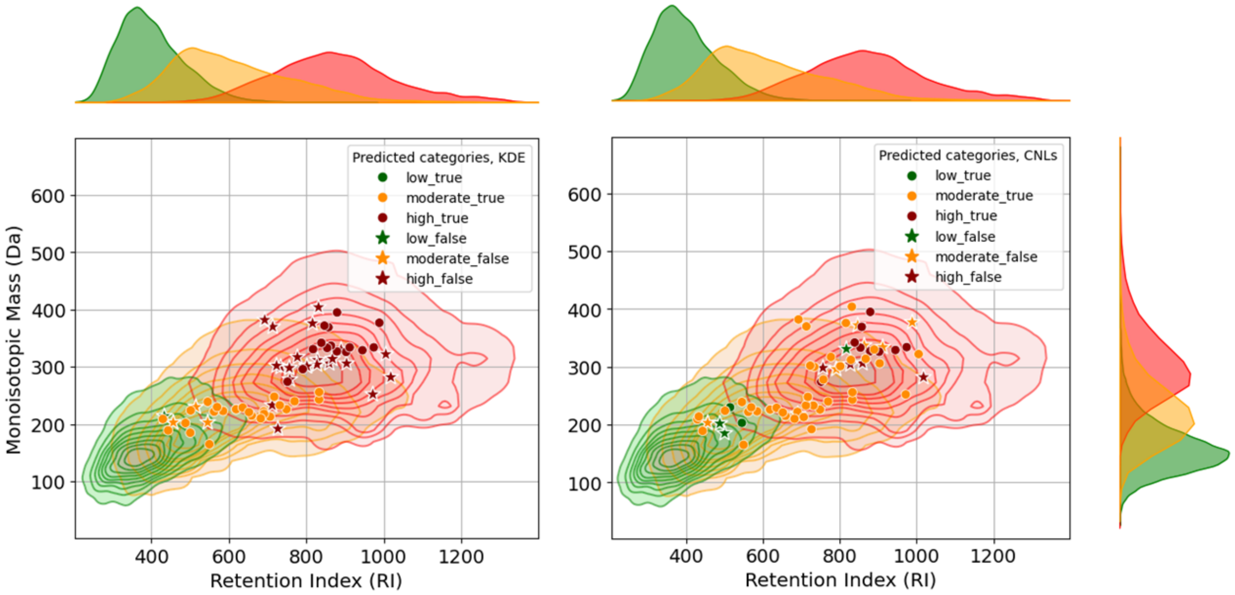


Figure 3: KDE map of RPLC space with the predicted toxicity categories for the detected pesticides based on KDE model (a), and CNLs based model (b).

Several recent studies have focused on linking high-resolution accurate mass tandem mass spectrometry (MS2) data to EC50 and LC50 values. These studies primarily aim to predict LC50 or EC50 values, which are then converted into toxicity categories.[30–33] Typically, this process involves converting the spectra into estimated molecular formulas and predicted probabilistic molecular fingerprints using tools like SIRIUS. The generated probabilistic fingerprints are then used for LC50 or EC50 prediction. As a result, the performance of these models depends on the accuracy of the molecular formula and fingerprint predictions. In contrast, our approach bypasses the need for molecular structure prediction by directly linking mass spectral signals to toxicity categories. Instead of using a regression model to predict continuous EC50 or LC50 values, we apply a classification algorithm that assigns compounds directly into toxicity categories. By relying solely on chromatographic and fragmentation data, our CNL- and KDE-based algorithms present a unique approach that differs from the structure-dependent models used in previous studies, making direct comparisons challenging.

Implications

In this study, we showed a novel feature prioritization approach based on predicting toxicity categories, by employing two algorithms: a probabilistic KDE-based model and a machine learning-based CNLs model. Due to the complex nature of environmental samples and the highly convoluted data, fragmentation information is not consistently available for all features. To mitigate the influence of the MS2 signal availability we developed KDE-based algorithm, which relies solely on MS1 and retention of the features. On the other hand the combination of retention and MS1 and fragmentation patterns or CNLs enhances the prediction accuracy. Our approach connects chromatographic and fragmentation data directly to toxicity categories, bypassing the need for compound identification, a common requirement in the currently available prioritization strategies. Moreover, our models demonstrated accuracy comparable to those that depend on structural information (FPs-based model), enabling effective toxicity predictions even for compounds with unknown structures. This makes our models not only valuable for potential prioritization of unknown features in current NTA studies but also suitable for enhancing retrospective analyses of historical datasets.

One of the current limitations of the developed approach is the limited number of measured toxicity values, retention indices, and high-resolution mass spectra, which reduces the model's applicability domain. Furthermore, the model is currently developed based on data from a single trophic level and acute toxicity. More specific mode of action models can give more insight into hazard scores and risk assessment. Another challenge arises from the fact that the model was trained on clean spectra obtained from publicly available spectral databases. In real LC-HRMS measurements, co-eluted compounds, instrumental artifacts, and background noise can introduce false positive fragments, potentially reducing prediction accuracy. As a result, robust and reliable data preprocessing is critical especially for DIA experiments to ensure the accuracy of the model.

Acknowledgement

The authors are thankful to the members of Environmental Modeling & Computational Mass Spectrometry (www.emcms.info). S.S., A.P., and V.T. are thankful to the UvA Data Science Center and ChemistryNL TKI for the financial support (projects Edified and SCOPE). The authors thank Denice van Herwerden for her help in setting up the calculations of elemental mass defects. J.W.O. is the recipient of a National Health and Medical Research Council (NHMRC) Investigator Grant (EL12009209) funded by the Australian Government. K.V.T, and S.S. also acknowledge funding support from the Australian National Health and Medical Research Council (NHMRC; APP1185347). The Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland (UQ), gratefully acknowledges the financial support of Queensland Health. This research was funded in whole or in part by NHMRC Emerging Leadership Fellowship EL1 2009209. For the purposes of open access, the author has applied a Creative Commons Attribution (CC BY) license to any Author

Accepted Manuscript version arising from this submission.

Notes

The script to perform the calculations is available at `https://doi.org/10.5281/zenodo.14188167`. The datasets used for the development of the algorithms are available at `https://bitbucket.org/viktoriiaturkina/toxmap_prioritization.jl/src/main`

Supporting Information Available

The Supporting Information is available at `XXX`.

References

(1) Landrigan, P. J. et al. The Lancet Commission on pollution and health. *The Lancet* **2018**, *391*, 462–512.

(2) Arp, H. P. H.; Aurich, D.; Schymanski, E. L.; Sims, K.; Hale, S. E. Avoiding the Next Silent Spring: Our Chemical Past, Present, and Future. *Environmental Science and Technology* **2023**, *57*, 6355–6359.

(3) Wang, Z.; Walker, G. W.; Muir, D. C.; Nagatani-Yoshida, K. Toward a Global Understanding of Chemical Pollution: A First Comprehensive Analysis of National and Regional Chemical Inventories. *Environmental Science and Technology* **2020**, *54*, 2575–2584.

(4) Rudén, C.; Hansson, S. O. Registration, Evaluation, and Authorization of Chemicals (REACH) Is but the First Step–How Far Will It Take Us? Six Further Steps to Improve the European Chemicals Legislation. *Environmental Health Perspectives* **2010**, *118*, 6.

(5) Hernández, F. et al. The role of analytical chemistry in exposure science: Focus on the aquatic environment. *Chemosphere* **2019**, *222*, 564–583.

(6) Ciccarelli, D.; Samanipour, S.; Rapp-Wright, H.; Bieber, S.; Letzel, T.; O'Brien, J. W.; Marczylo, T.; Gant, T. W.; Vineis, P.; Barron, L. P. Bridging knowledge gaps in human chemical exposure via drinking water with non-target screening. *Critical Reviews in Environmental Science and Technology* **2024**,

(7) Samanipour, S.; Barron, L. P.; van Herwerden, D.; Praetorius, A.; Thomas, K. V.; O'Brien, J. W. Exploring the Chemical Space of the Exposome: How Far Have We Gone? *JACS Au* **2024**, *4*, 2412–2425.

(8) Gosetti, F.; Mazzucco, E.; Gennaro, M. C.; Marengo, E. Contaminants in water: non-target UHPLC/MS analysis. *Environ. Chem. Lett.* **2016**, *14*, 51–65.

(9) Richardson, S. D.; Manasfi, T. Water Analysis: Emerging Contaminants and Current Issues. *Analytical Chemistry* **2024**, *96*, 8184–8219.

(10) Krauss, M.; Singer, H.; Hollender, J. LC-high resolution MS in environmental analysis: From target screening to the identification of unknowns. *Analytical and Bioanalytical Chemistry* **2010**, *397*, 943–951.

(11) Sobus, J. R.; Wambaugh, J. F.; Isaacs, K. K.; Williams, A. J.; McEachran, A. D.; Richard, A. M.; Grulke, C. M.; Ulrich, E. M.; Rager, J. E.; Strynar, M. J.; Newton, S. R. Integrating tools for non-targeted analysis research and chemical safety evaluations at the US EPA. *Journal of Exposure Science & Environmental Epidemiology 2017 28:5* **2017**, *28*, 411–426.

(12) Overdahl, K. E.; Sutton, R.; Sun, J.; Destefano, N. J.; Getzinger, G. J.; Ferguson, P. L. Assessment of emerging polar organic pollutants linked to contaminant pathways within an urban estuary using non-targeted analysis. *Environmental Science: Processes & Impacts* **2021**, *23*, 429–445.

(13) Creusot, N.; Huba, K.; Borel, C.; Ferrari, B. J.; Chèvre, N.; Hollender, J. Identification of polar organic chemicals in the aquatic foodweb: Combining high-resolution mass spectrometry and trend analysis. *Environment International* **2024**, *183*, 108403.

(14) Kutarna, S.; Tang, S.; Hu, X.; Peng, H. Enhanced Nontarget Screening Algorithm Reveals Highly Abundant Chlorinated Azo Dye Compounds in House Dust. *Environmental Science and Technology* **2021**, *55*, 4729–4739.

(15) Szabo, D.; Fischer, S.; Mathew, A. P.; Kruve, A. Prioritization, Identification, and Quantification of Emerging Contaminants in Recycled Textiles Using Non-Targeted and Suspect Screening Workflows by LC-ESI-HRMS. *Analytical Chemistry* **2024**, *96*, 14150–14159.

(16) Braun, G.; Herberth, G.; Krauss, M.; König, M.; Wojtysiak, N.; Zenclussen, A. C.; Escher, B. I. Neurotoxic mixture effects of chemicals extracted from blood of pregnant women. *Science* **2024**, *386*, 301–309.

(17) Petrie, B.; Barden, R.; Kasprzyk-Hordern, B. A review on emerging contaminants in wastewaters and the environment: Current knowledge, understudied areas and recommendations for future monitoring. *Water Research* **2015**, *72*, 3–27.

(18) Hohrenk, L. L.; Itzel, F.; Baetz, N.; Tuerk, J.; Vosough, M.; Schmidt, T. C. Comparison of Software Tools for Liquid Chromatography-High-Resolution Mass Spectrometry Data Processing in Nontarget Screening of Environmental Samples. *Analytical Chemistry* **2020**, *92*, 1898–1907.

(19) Hollender, J.; Schymanski, E. L.; Singer, H. P.; Ferguson, P. L. Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? *Environ. Sci. Technol.* **2017**, *51*, 11505–11512.

(20) Lennon, S. et al. Harmonized quality assurance/quality control provisions to assess completeness and robustness of MS1 data preprocessing for LC-HRMS-based suspect

screening and non-targeted analysis. *TrAC Trends in Analytical Chemistry* **2024**, *174*, 117674.

(21) Hulleman, T.; Turkina, V.; O'Brien, J. W.; Chojnacka, A.; Thomas, K. V.; Samanipour, S. Critical Assessment of the Chemical Space Covered by LC-HRMS Non-Targeted Analysis. *Environmental Science and Technology* **2023**, *57*, 14101–14112.

(22) González-Gaya, B.; Lopez-Herguedas, N.; Bilbao, D.; Mijangos, L.; Iker, A. M.; Etxebarria, N.; Irazola, M.; Prieto, A.; Olivares, M.; Zuloaga, O. Suspect and non-target screening: the last frontier in environmental analysis. *Analytical Methods* **2021**, *13*, 1876–1904.

(23) Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J. Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environ. Sci. Technol.* **2014**, *48*, 2097–2098.

(24) Kind, T.; Fiehn, O. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanalytical Reviews 2010 2:1* **2010**, *2*, 23–60.

(25) Szabo, D.; Falconer, T. M.; Fisher, C. M.; Heise, T.; Phillips, A. L.; Vas, G.; Williams, A. J.; Kruve, A. Online and Offline Prioritization of Chemicals of Interest in Suspect Screening and Non-targeted Screening with High-Resolution Mass Spectrometry. *Analytical Chemistry* **2024**, *96*, 3707–3716.

(26) Rilievo, G.; Boscolo, S.; Pettenuzzo, S.; Matozzo, V.; Fabrello, J.; Roverso, M.; Bogialli, S. From a validated targeted method to a retrospective UHPLC-HRMS non-targeted analysis unveiling COVID-19-related contaminants in clams. Have we bias in marine model organisms for ecotoxicological studies? *Chemosphere* **2024**, *364*, 142994.

(27) Silva, M.; Capps, S.; London, J. Community-Engaged Research and the Use of Open Access ToxVal/ToxRef In Vivo Databases and New Approach Methodologies (NAM)

to Address Human Health Risks From Environmental Contaminants. *Birth Defects Research* **2024**, *116*, e2395.

(28) Vermeulen, R.; Schymanski, E. L.; Barabási, A. L.; Miller, G. W. The exposome and health: where chemistry meets biology. *Science (New York, N.Y.)* **2020**, *367*, 392.

(29) Meekel, N.; Vughs, D.; Béen, F.; Brunner, A. M. Online Prioritization of Toxic Compounds in Water Samples through Intelligent HRMS Data Acquisition. *Analytical Chemistry* **2021**, *93*, 5071–5080.

(30) Peets, P.; Wang, W. C.; Macleod, M.; Breitholtz, M.; Martin, J. W.; Kruve, A. MS2Tox Machine Learning Tool for Predicting the Ecotoxicity of Unidentified Chemicals in Water by Nontarget LC-HRMS. *Environmental Science and Technology* **2022**, *56*, 15508–15517.

(31) Peets, P.; Rian, M. B.; Martin, J. W.; Kruve, A. Evaluation of Nontargeted Mass Spectral Data Acquisition Strategies for Water Analysis and Toxicity-Based Feature Prioritization by MS2Tox. *Environmental Science and Technology* **2024**, *58*, 17406–17418.

(32) Rahu, I.; Kull, M.; Kruve, A. Predicting the Activity of Unidentified Chemicals in Complementary Bioassays from the HRMS Data to Pinpoint Potential Endocrine Disruptors. *Journal of Chemical Information and Modeling* **2024**, *64*, 3093–3104.

(33) Arturi, K.; Hollender, J. Machine Learning-Based Hazard-Driven Prioritization of Features in Nontarget Screening of Environmental High-Resolution Mass Spectrometry Data. *Environmental Science and Technology* **2023**, *57*, 18067–18079.

(34) Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.; Dorrestein, P. C.; Rousu, J.; Böcker, S. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods 2019 16:4* **2019**, *16*, 299–302.

(35) Zushi, Y.; Hanari, N.; Nabi, D.; Lin, B. L. Mixture Touch: A Web Platform for the Evaluation of Complex Chemical Mixtures. *ACS Omega* **2020**, *5*, 8121–8126.

(36) Zhang, R.; Guo, H.; Hua, Y.; Cui, X.; Shi, Y.; Li, X. Modeling and insights into the structural basis of chemical acute aquatic toxicity. *Ecotoxicology and Environmental Safety* **2022**, *242*, 113940.

(37) Yang, H.; Lou, C.; Li, W.; Liu, G.; Tang, Y. Computational Approaches to Identify Structural Alerts and Their Applications in Environmental Toxicology and Drug Discovery. *Chemical Research in Toxicology* **2020**, *33*, 1312–1322.

(38) Williams, A. J.; Grulke, C. M.; Edwards, J.; McEachran, A. D.; Mansouri, K.; Baker, N. C.; Patlewicz, G.; Shah, I.; Wambaugh, J. F.; Judson, R. S.; Richard, A. M. The CompTox Chemistry Dashboard: A community data resource for environmental chemistry. *Journal of Cheminformatics* **2017**, *9*, 1–27.

(39) MassBank — MassBank Europe Mass Spectral DataBase. https://massbank.eu/MassBank/.

(40) MassBank of North America. https://mona.fiehnlab.ucdavis.edu/downloads.

(41) Reference Spectral Libraries - GNPS Documentation. https://ccms-ucsd.github.io/GNPSDocumentation/downloadlibraries/.

(42) NIST20: Updates to the NIST Tandem and Electron Ionization Spectral Libraries — NIST. https://www.nist.gov/programs-projects/nist20-updates-nist-tandem-and-electron-ionization-spectral-libraries.

(43) Cassotti, M.; Ballabio, D.; Todeschini, R.; Consonni, V. A similarity-based QSAR model for predicting acute toxicity towards the fathead minnow (Pimephales promelas). *SAR and QSAR in Environmental Research* **2015**, *26*, 217–243.

(44) Schür, C.; Gasser, L.; Perez-Cruz, F.; Schirmer, K.; Baity-Jesi, M. A benchmark dataset for machine learning in ecotoxicology. *Scientific Data 2023 10:1* **2023**, *10*, 1–20.

(45) Samanipour, S.; O'Brien, J. W.; Reid, M. J.; Thomas, K. V.; Praetorius, A. From Molecular Descriptors to Intrinsic Fish Toxicity of Chemicals: An Alternative Approach to Chemical Prioritization. *Environmental Science and Technology* **2023**, *57*, 17950–17958.

(46) van Herwerden, D.; Nikolopoulos, A.; Barron, L. P.; O'Brien, J. W.; Pirok, B. W.; Thomas, K. V.; Samanipour, S. Exploring the chemical subspace of RPLC: A data driven approach. *Analytica Chimica Acta* **2024**, *1317*, 342869.

(47) Aalizadeh, R.; Nikolopoulou, V.; Thomaidis, N. S. Development of Liquid Chromatographic Retention Index Based on Cocamide Diethanolamine Homologous Series (C(n)-DEA). *Analytical Chemistry* **2022**, *94*, 15987–15996.

(48) van Herwerden, D.; O'Brien, J. W.; Lege, S.; Pirok, B. W. J.; Thomas, K. V.; Samanipour, S. Cumulative Neutral Loss Model for Fragment Deconvolution in Electrospray Ionization High-Resolution Mass Spectrometry Data. *Anal. Chem.* **2023**,

(49) Samanipour, S.; Reid, M. J.; Bæk, K.; Thomas, K. V. Combining a Deconvolution and a Universal Library Search Algorithm for the Nontarget Analysis of Data-Independent Acquisition Mode Liquid Chromatography-High-Resolution Mass Spectrometry Results. *Environmental Science and Technology* **2018**, *52*, 4694–4701.

(50) van Herwerden, D.; O'Brien, J. W.; Choi, P. M.; Thomas, K. V.; Schoenmakers, P. J.; Samanipour, S. Naive Bayes classification model for isotopologue detection in LC-HRMS data. *Chemometrics and Intelligent Laboratory Systems* **2022**, *223*, 104515.

(51) Pedregosa FABIANPEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA,

VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(52) Svetnik, V.; Liaw, A.; Tong, C.; Christopher Culberson, J.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 1947–1958.

(53) Turkina, V.; Messih, M. R.; Kant, E.; Gringhuis, J.; Petrignani, A.; Corthals, G.; O'Brien, J. W.; Samanipour, S. Molecular Fingerprints Optimization for Enhanced Predictive Modeling. **2024**,

(54) Smith, D. H.; Carhart, R. E.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *Journal of Chemical Information and Computer Sciences* **1985**, *25*, 64–73.

(55) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *Journal of Chemical Information and Computer Sciences* **1995**, *35*, 1039–1045.

(56) Klekota, J.; Roth, F. P. Chemical substructures that enrich for biological activity. *Bioinformatics (Oxford, England)* **2008**, *24*, 2518–2525.

(57) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *Journal of chemical information and computer sciences* **2002**, *42*, 1273–1280.

(58) PubChem Substructure Fingerprint V1.3. 2009; `https://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf`.

(59) Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry* **2011**, *32*, 1466–1474.

(60) Węglarczyk, S. Kernel density estimation and its application. **2018**,

(61) Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR & Combinatorial Science* **2007**, *26*, 694–701.

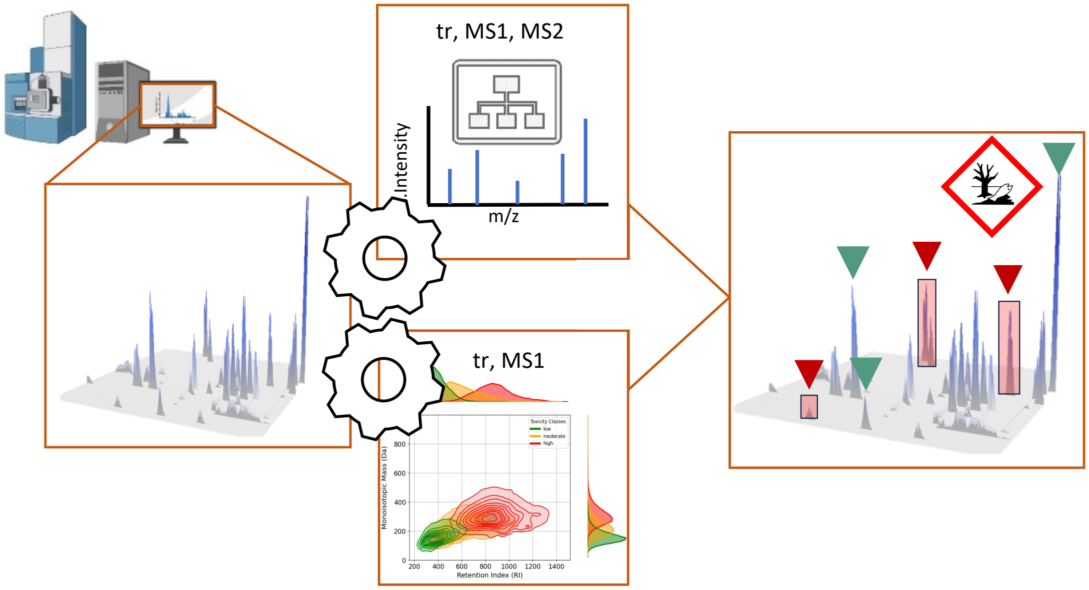
(62) Yu, X. Global classification models for predicting acute toxicity of chemicals towards Daphnia magna. *Environmental Research* **2023**, *238*, 117239.

(63) Sheffield, T. Y.; Judson, R. S. Ensemble QSAR Modeling to Predict Multispecies Fish Toxicity Lethal Concentrations and Points of Departure. *Environmental Science and Technology* **2019**, *53*, 12793–12802.

(64) Black, G. et al. Exploring chemical space in non-targeted analysis: a proposed ChemSpace tool. *Analytical and bioanalytical chemistry* **2022**, *415*, 35.

(65) Lowe, C. N.; Isaacs, K. K.; McEachran, A.; Grulke, C. M.; Sobus, J. R.; Ulrich, E. M.; Richard, A.; Chao, A.; Wambaugh, J.; Williams, A. J. Predicting compound amenability with liquid chromatography-mass spectrometry to improve non-targeted analysis. *Analytical and bioanalytical chemistry* **2021**, *413*, 7495.

(66) Alygizakis, N.; Konstantakos, V.; Bouziotopoulos, G.; Kormentzas, E.; Slobodnik, J.; Thomaidis, N. S. A Multi-Label Classifier for Predicting the Most Appropriate Instrumental Method for the Analysis of Contaminants of Emerging Concern. *Metabolites* **2022**, *12*, 199.



TOC for Review.