From Molecular Descriptors to Intrinsic Fish Toxicity of Chemicals: an Alternative Approach to Chemical Prioritization

Saer Samanipour,[*,†,‡,¶] Jake W. O'Brien,[¶] Malcolm J. Reid,[§] Kevin V. Thomas,[¶] and Antonia Praetorius[*,‖]

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

[‡] *UvA Data Science Center, University of Amsterdam, 1090 GD Amsterdam, the Netherlands*

[¶] *Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland, Brisbane Qld 4072, Australia*

[§] *Norwegian Institute for Water Research (NIVA), NO-0579 Oslo, Norway*

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

E-mail: s.samanipour@uva.nl; a.praetorius@uva.nl

Abstract

The European and US chemical agencies have listed approximately 800k chemicals where knowledge on potential risks to human health and the environment are lacking. Filling these data gaps experimentally is impossible so in-silico approaches and prediction are essential. Many existing models are however limited by assumptions (e.g. linearity and continuity) and small training sets. In this study we present a supervised direct classification model that connects molecular descriptors to toxicity. Categories can be either data-driven (using k-means clustering) or regulatory-defined. This was tested via 907 experimentally defined 96h LC50 values for acute fish toxicity. Our classification model explained $\approx 90\%$ of variance in our data for the training set and $\approx 80\%$ for the test set. This strategy gave a 5-fold decrease in the incorrect categorization compared to a QSAR regression model. Our model was subsequently employed to predict the toxicity categories of $\approx$ 32k chemicals. A comparison between the model-based applicability domain (AD) and the training set AD was performed, suggesting that the training set based AD is a more adequate way to avoid extrapolation when using such models. The better performance of our direct classification model compared to QSAR methods, makes this approach a viable tool for hazard and risk assessment of chemicals.

Synopsis

In this study an alternative machine learning-based strategy to conventional QSAR models is used for the toxicity categorization of chemicals using molecular descriptors and direct classification.

Introduction

The chemical space of the human exposome is ever expanding with a wider diversity of chemicals from both fate and toxicity points of view.[1–7] The latest estimates of the environmentally relevant chemicals based on the chemical registries and production volumes are estimated to be between 350k and 800k.[2,8] For most of these chemicals there is little to no knowledge about their environmental fate nor toxicity.[1–5,8,9] Since the experimental assessment of the fate and toxicity of such a large number of chemicals is not feasible, modeling approaches to predict hazard indicators play an increasingly important role in chemical prioritization and risk assessment.[10–13]

Prediction of the physicochemical properties and the biological activity (e.g. aquatic toxicity) has been one of the main approaches to deal with the structural diversity in the chemical space.[10–13] Most existing modeling strategies employ quantitative structure activity relationship (QSAR) models and rely on building linear and/or non-linear relationships between the structural descriptors and the modeled activity/property.[10,14–17] These models are often built on very homogeneous training sets (i.e. similar chemical classes), hence the linearity assumption.[17,18] In fact, recent efforts have been put into using more heterogeneous training sets as well as moving away from the linearity assumption.[13,14,18,19] Independent from the level of heterogeneity of the training dataset, QSAR models are very limited in the number of measured activities as well as the number of chemicals evaluated (e.g. around 1000 chemicals).[13,14,18,19] The main consequence of this limitation is the fact that the models are used in extrapolation mode when used for prediction. This implies that the new data points are not represented adequately by the chemicals within the training set, thus outside of the model applicability domain. The use of these models for extrapolation may potentially result in very large prediction errors.[13,19,20]

For these predicted and measured activities (i.e. toxicity and/or other properties) to be translated into chemical management actions, they are divided into different categories using thresholds based on expert knowledge.[1,3,21–24] Examples for such categories are environmental hazard categories defined by the Globally Harmonized System of Classification and Labelling of Chemicals (GHS) or thresholds for persistence (P), bioaccumulation potential (B) and toxicity (T) defined under the European Registration, Evaluation, Authorization and Restriction of Chemicals (REACH).[25] The chemicals that fall within specific categories are then furthered for more active monitoring and eventually for legislation.[24,26? –28] This process triggers wider experimental evaluation of chemicals within high priority categories, which may result in adjustment of the previously set thresholds, based on the new experimental evidence.[24,26,29] However, for this chemical management strategy to be effective, a more accurate and reliable chemical prioritization (i.e. chemical categorization) approach is warranted.

In this study we propose an alternative strategy for chemical prioritization on the example of acute aquatic toxicity, where the QSAR-based activity prediction step is skipped. Our direct classification model directly converts molecular descriptors into chemical categories, avoiding the errors inherent to the activity prediction step. As a proof of concept, this strategy was tested with experimentally determined 96h lethal concentration (LC50) values for fish, for 907 organic chemicals. We compared the results of our direct classification strategy with the conventional QSAR approach. Additionally, our modeling strategy was expanded to 32000 chemicals from Norman SusDat.[27] Finally, we performed a critical evaluation of applicability domains for all the models in this study.

Methods

Overall Workflow

The dataset used for our model development, validation, and testing consists of calculated descriptors, monoisotopic mass of each chemical, and experimentally determined LC50 values (96 hours) for acute fish toxicity (see details in Section Dataset). The LC50 values were divided into four categories namely: very low toxicity, low toxicity, moderate toxicity, and high toxicity via k-means clustering. This categorization followed the typical evidence-based effect modeling categorization.[30–32] Additionally, regulatory-defined toxicity categories were retrieved from the Globally Harmonized System of Classification and Labelling of Chemicals (GHS). We assessed the prediction accuracy of the two types of toxicity categories by employing two different modeling strategies: a conventional QSAR regression model vs direct classification (Figure 1. The QSAR regression model simulated the case where the acute fish toxicity (as LC50) is predicted based on molecular descriptors via a QSAR model and then the chemical is assigned a specific toxicity category in a separate step. On the other hand, the direct classification model skipped the LC50 prediction step and directly classified the chemical of interest into one of the initially defined toxicity categories. This comparison was performed for the full dataset (i.e. training set and test set) in order to assess the accuracy of each approach in acute fish toxicity categorization.

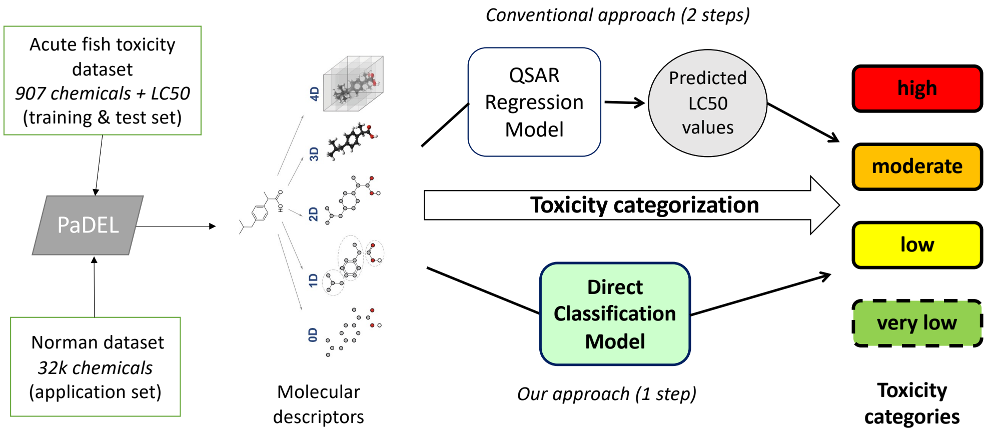


Figure 1: depicts the overall workflow of the study from the raw data to the finally generated models.

Datasets

We employed two different datasets for our model development[18] and the model application.[33] Our modeling dataset consisted of experimental acute fish toxicity values for 907 chemicals retrieved from three databases, namely OASIS, ECOTOX and EAT5 and provided by Cassotti et al.[18] The data consisted of the concentrations causing death in 50% of test fathead minnows (Pimephales promelas) over a test duration of 96 hours (LC50 96 hours). More details regarding the data curation is provided elsewhere.[18] We will refer to this dataset as "acute fish toxicity dataset" here after. The chemicals in this dataset covered different chemical families, including pharmaceuticals, pesticides, conventional persistent organic pollutants (POPs), and industrial chemicals. Throughout this article we refer to the 907 chemicals with measured toxicity and curated descriptors as full "acute fish toxicity dataset", the portion used for the model development/validation as training set, and the portion of the data used for additional model testing of the final model as test set.

The second dataset (hereafter referred to as "Norman dataset") was an extract of around 32000 chemicals (31722 chemicals), including their predicted 96h LC50 values for acute fish toxicity (Pimephales promelas) from the Norman SusDat database.[34] This dataset included only the chemicals that were reported as within the applicability domain of the QSAR model developed by Aalizadeh et al,[34] which was used for testing our model applicability, Figure 2. This is the model employed by Norman Network for their risk assessment and chemical management. When checking the overlap between the acute fish toxicity dataset and the Norman dataset, we observed around 100 common entries.

We calculated 2757 1D (i.e. constitutional/count descriptors), 2D (i.e. structural fragments), and 3D (i.e. graph invariants) molecular descriptors, and PubChem fingerprints for both datasets using PaDEL software package,[35] implemented via a python 3 wrapper called padelpy. Additionally, the name of the chemicals, their SMILES,[36] and InChiKeys[37] were retrieved from the PubChem database[38] via pubchempy API. In order to identify the unstable descriptors—caused by the lack of convergence during the structural optimization—we performed the descriptor calculations for the acute fish toxicity dataset in triplicates. The descriptors were scaled by the maximum of each descriptor in the training set to minimize the impact of the descriptor magnitude on the final models.[39] After scaling, the variance of each descriptor in the acute fish toxicity dataset was calculated and only the descriptors that had a variance below 0.1 were kept. We assumed that the stable descriptors for the acute fish toxicity dataset are also stable for the Norman dataset. Therefore, the descriptors for this dataset were calculated only once. Additionally, the maximum of each descriptor in the Norman dataset was compared to those from the training set (from the acute fish toxicity dataset). The descriptors that have this ratio larger than 100 were considered unstable and removed from both datasets, resulting in a total of 2036 final descriptors out of an initial 2780.

We also evaluated the coverage of the chemical spaces of the datasets by the means of Principal Component Analysis (PCA), Figure 2. The PCA is an unsupervised dimension reduction approach, which enabled us to assess the underlying trends in our datasets by combining several variables into a single principal component.[40] To perform PCA, we used the curated descriptors matrix and in total two principal components.

Toxicity Categories

To categorize the chemicals based on their acute fish toxicity, we employed two different strategies namely 1) applying k-means clustering to derive four categories from our acute fish toxicity dataset and 2) using predefined categories for acute aquatic hazard as defined in the GHS.[41]

K-means Clustering for Toxicity Categorization

The k-means strategy divided the chemicals into four categories consisting of high toxicity, moderate toxicity, low toxicity, and very low toxicity accounting for 96h LC50 values for fish toxicity and monoisotopic mass of the chemicals. The k-means clustering algorithm is an iterative clustering algorithm, where the distances between different measurements from

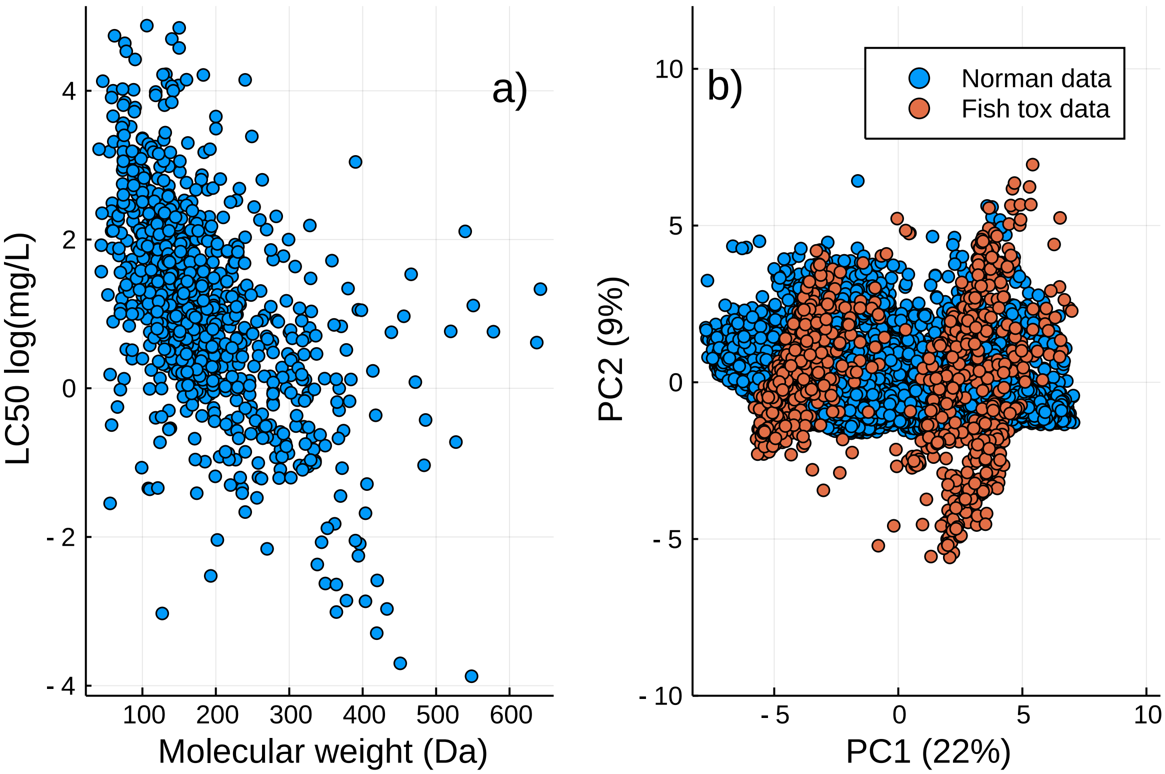


Figure 2: depicts a) the distribution of the experimental LC50 values used for the model development and validation whereas b) shows the chemical space via PCA covered by the acute fish toxicity data (i.e. training and test sets) and the Norman dataset, where the curated descriptors were used for the cluster analysis

.

a set of user defined centers (so called centroides) are used for clustering the data.[40] This algorithm has the advantage of incorporating more than one parameter, compared to expert manual judgment in the clustering. Additionally, this algorithm, given that it has randomly selected centroides in the first iteration, it requires further validation. Here we employed bootstrapping to assure the selected acute fish toxicity categories (i.e. clusters) are robust enough for predictive purposes. To do that, the fish toxicity data was randomly divided into 90% training set and 10% test set. The training set then was bootstrapped with replacement for 500 iterations, to guarantee that each model is built based on a unique dataset. The most commonly identified centroid over 500 iterations was selected as final model and for acute fish toxicity categorization. In the end, the final model was further tested using the test set. During the categorization, we provided the k-means algorithm with two variables namely 96h LC50 values and the monoisotopic masses, and the number of clusters of 4, following the category structures adapted by previous studies.[30]

GHS Categorization for Acute Aquatic Hazards

In addition to the k-mean clustering we also used the three categories for acute aquatic hazards of the GHS, which were hard set thresholds.[41]The three GHS-based categories for short-term (acute) aquatic hazard are based on thresholds derived from 96h LC50 values for acute fish toxicity: high toxicity (Category Acute 1: 96h LC50 for fish $\leq$ 1 mg/L), moderate toxicity (Category Acute 2: 1 mg/L $<$ LC50 $\leq$ 10 mg/L), and low toxicity (Category Acute 3: LC50 $>$ 10 mg/L LC50), Table 4.1.1 in Reference.[42]

Modeling

In this study, we developed two different models namely: a QSAR regression model and a direct classification model. The details of each model strategy is provided below. Both models, once optimized with the acute fish toxicity dataset, were used with the Norman dataset to further assess their applicability.

QSAR Regression Model

We developed, optimized, validated, and tested a random forest regression model using the curated descriptors (independent variables) and the experimentally defined LC50 values (dependent variable). Random forest is a decision tree based algorithm where several bootstrap data (i.e. training set) are given to several decision trees. This assures that the dataset given to each tree is unique.[40] Once the model is developed, the most common decision tree model outcome is considered as the random forest model prediction. The main advantage of the random forest modeling strategy is the ability to handle non-linearity and non-continuity in the data, which is highly relevant to toxicity prediction.[43] Here, the acute fish toxicity dataset was divided into training set (90% of the full dataset) and test set (10%). The training set was used for the model development and optimization while the test set was for further evaluation of the dataset. For the regression model, the model hyper-parameter optimization was performed with a two dimensional grid with the number of trees ranging from 100 - 1000 whereas the minimum number of points in each leaf varying from 1 - 21. The combination of 3 fold cross-validation and out-of-bag strategy enabled us to generate an optimized regression model while defining the importance of each variable. The variables that had relative levels of importance larger than 1% were considered as essential variables for the model. This strategy enabled us to quickly identify the most relevant variables to our model's accuracy.

The finally optimized regression model consisted of 600 trees, minimum 4 points in each leaf, and 8 variables. This regression model was employed to predict the 96h LC50 for fish toxicity of the chemicals in the Norman dataset. In a second step, the predicted LC50 values were used to categorize the chemicals into the two types of toxicity categories described above.

Descriptor-Based Direct Classification Model

We developed, validated, and tested a classification model to convert the curated descriptors to the acute fish toxicity categories. For this model, we employed random forest classification, implemented via ScikitLearn.jl julia package.[44]

For the direct classification, we split the acute fish toxicity dataset (i.e. curated the descriptors and toxicity categories) into training set (90% of the full dataset) and test set (10%). To optimize the main model hyper-parameters, the number of trees, and minimum number of points in each leaf, we generated a grid with 20 steps for each parameter ranging from 200 - 2000 and from 1 - 21 for the number of trees and minimum data points in leaf, respectively. For each model, we performed 3 folds of cross-validation to systematically assess the model accuracy. The model with the highest cross validation accuracy (i.e. 73%) was considered as the optimized classification model. This optimized classification model consisted of 1200 tress and minimum number of points in each leaf of 4. To avoid overfitting during the training process, when building the model, we set an out-of-bag cross-validation,[45] where only a randomly selected fraction (i.e. square root of the number of variables) of the variables were fed to individual trees. The combination of out-of-bag cross-validation and leaf purity was utilized to calculate the importance of individual variables on the final model. To select the relevant variables, we divided variance explained by each variable by the largest one and selected those that contributed more than 1% to the model, thus 230 out of 2036 variables.

To build the final model, the full acute fish toxicity dataset was used with the selected variables. In this case all the selected variables were used for the final model building. Additionally, this model was used to categorize the Norman dataset into the two types of acute fish toxicity categories directly based on the curated descriptors.

Applicability Domain

To assess whether a chemical is well represented by the model training set, we performed the applicability domain (AD) assessment. The AD assessment was done by calculating the leverage of each chemical compared to the training set.[34] The leverage was calculated using Eq.1, where X is the matrix of the training set (including the descriptors), x_i is the vector of descriptors for an individual chemical, and the h_{ii} is the calculated leverage. The leverage calculations are typically done only using the model variables, in other words only the descriptors used for the optimized model. In this study we performed both the full descriptor space (i.e. assuming the model using all the descriptors) and the model specific descriptors (i.e. conventional approach). This strategy enabled us to systematically assess which chemicals are well represented by the training set.

$$h_{ii} = x_i^T (X^T X)^{-1} x_i \tag{1}$$

Calculations

All calculations were performed using a personal computer (PC) with Intel Core i7 CPU and 16 GB of RAM operating Ubuntu 20.04.2 LTS. All the data processing and statistical analysis were performed using julia language 1.6.

Results and discussion

In this study, we developed a random forest-based direct classification model to convert the molecular descriptors of chemicals to predefined acute fish toxicity categories. This model was developed, validated, and tested via an experimentally defined dataset of 96h LC50 values for acute fix toxicity for 907 organic chemicals. The result of this strategy was directly compared to the conventional two-step approach—first QSAR-based property prediction and then toxicity categorization—both for the acute fish toxicity data and a dataset of $\approx$ 23000 chemicals from Norman SusDat.[33]

Toxicity Categorization

The final k-means model resulted in a clustering accuracy of 97.5%. This model, then, was fed the full acute fish toxicity dataset to define the toxicity category of each chemical in that dataset. The final model was saved as a binary file to be used for prediction (Figure 3). The k-means and GHS categories were used as labels in two separate runs of the direct classification model while the 96h LC50 values for acute fish toxicity predicted by the QSAR regression model were converted into the two types of acute toxicity categories in a second step.

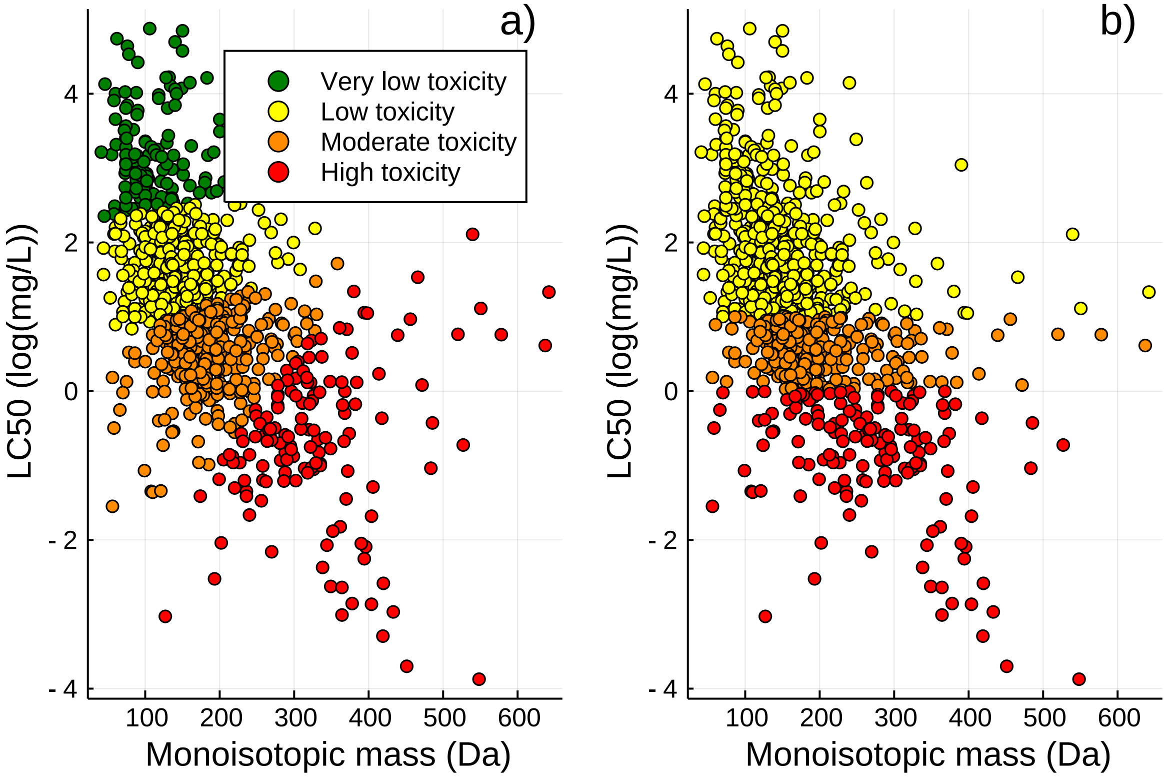


Figure 3: shows the distribution of the toxicity categories of the acute fish toxicity dataset via a) the best k-means clustering model and b) based on GHS categories.

When comparing the unsupervised k-means clustering-based categorization with the expert knowledge based categorization from the GHS, we see a high level of similarity in the thresholds (Figure 3). In fact the main differences were observed for chemicals with a molecular weight of $\geq$ 400 Da and LC50 values $\geq$ 1 mg/L (0 log(mg/L)). These chemicals in the k-means categorization were considered part of the high toxicity category while based on the GHS categories they were considered moderate to low toxicity. When calculating the similarity scores between the descriptors of those chemicals and the two categories, we consistently observed higher values for high toxicity category. This indicates that those chemicals may be structurally more similar to the high toxicity category rather than the moderate and/or low one. These similarities are better captured by the k-means model, given that it uses two variables (96h LC50 and monoisotopic mass) and Euclidean distances for the cluster creation.

Performance of QSAR Regression Model

The residuals of the final and optimized QSAR regression model were between -1 and 1 in LC50 units for $\approx$ 95% of the data (Figure S2). This model consisted of 600 trees and 8 variables, resulting in an R^2 of 0.86 for the training set and $\approx$ 0.7 for both median cross-validation and test set. The observed levels of accuracy was comparable to previously reported linear and non-linear QSAR models[17,34] (Figure 4). We observed up to 2.1 log(mg/L) overestimation of the LC50 for values $\leq$ -1 while our model resulted in a slight underestimation of toxicity for LC50 values $\geq$ 5 (Figures 4 and S2). Finally, we used the optimized model to predict the 96h acute fish toxicity LC50 values for the Norman dataset. When comparing the results of our predictions to the predictions by Aalizadeh et al,[34] a clear linear trend (i.e. Pearson correlation coefficient of 0.68) between the two predictions was observed, further indicating the validity of our model (Figure S3).

The optimized regression model included 8 variables from which two were related to the logP of the chemicals in the training set (Figure S1). The most relevant variable was the Crippen logP[46] value explaining around 35% variance of the final model. This logP was calculated based on 68 atomic contributions. On the other hand, the second variable was XLogP,[47] implemented within PubChem.[38,48] This logP calculation also uses the atomic contribution of 87 groups and additionally incorporates two correction factors, improving its accuracy and expanding its applicability. Another relevant variable for our regression model was the ZMIC1 descriptor which is a 2D descriptor indicating the level of symmetry in the structure.[35] Finally, the remaining relevant descriptors (i.e. excluding logP, XlogP and ZMIC1 descriptors) were related to molecular connectivity, polarizability, and hydrogen-bond donation, which all have shown to be relevant in explaining physico-chemical properties and toxicity of chemicals.[15,17,34]

Performance of Descriptor-Based Direct Classification Model

The optimized direct classification model resulted in a classification accuracy of 92% for the training set and around 80% for both the cross-validation and the test set, for the four k-means categories. The final model used 230 variables out of a total of 2036 curated descriptors. Similar to the regression model, most of the important variables were a combination of 2D descriptors and fingerprints (i.e. 3D) (Figure S4). These descriptors included the four logP calculations (e.g. CrippenlogP) as well as parameters related to polarizability and charge distribution. These parameters are all highly relevant to the mobility of the chemicals and their binding potential with the active sites.[15,18] Differently from the regression model, the most relevant variable only explained $\approx$ 1.5% of variance (vs 35% for the regression model) in the final model. Even though larger number of variables were included in the model, the total number of variables were less than 30% of the number of measurements resulting in a mathematically well-defined problem. Additionally, a larger number of variables enables a better assessment of the model applicability domain.

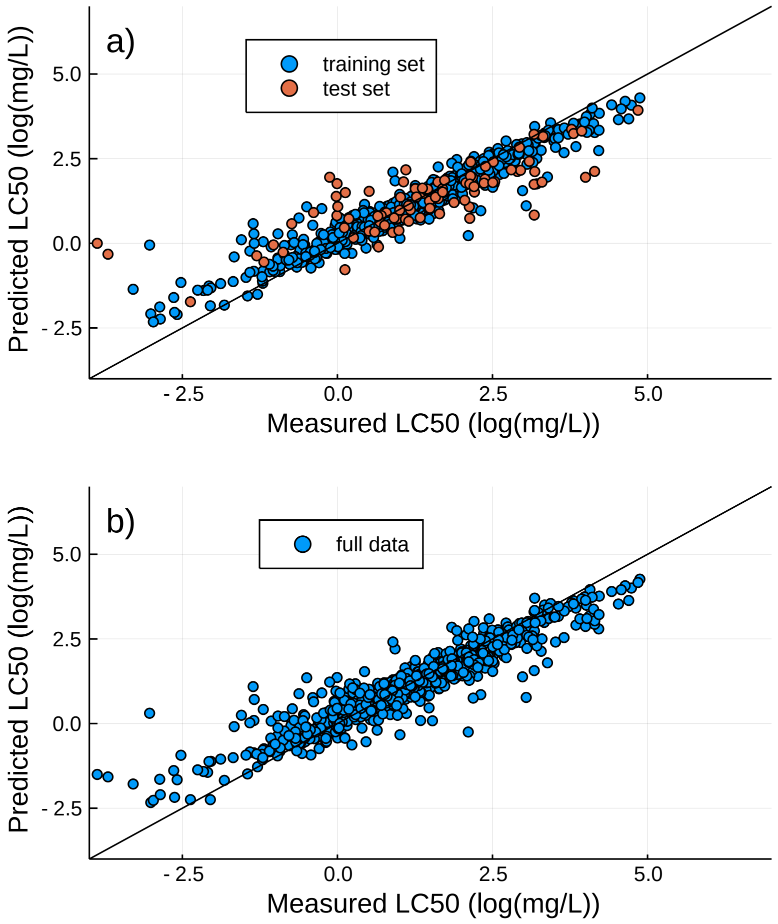


Figure 4: depicts the measured vs predicted 96h LC50 values (in log(mg/L)) for acute fish toxicity for a) the training and test set during the model optimization and b) the optimized model with the full acute fish toxicity dataset.

The direct classification model based on the three GHS categories, resulted in an accuracy of 94% for the training set and around 85% for the cross-validation and test set. This model, similar to the previous one, had 236 high importance variables that were included in the final model. The high importance variables (e.g. top 20) for both models were exactly the same as for the direct classification into the k-means categories with similar levels of variance explained.

The reported statistics and the selected variables in our classification models further indicated the applicability of our model for prediction of acute fish toxicity categories directly from the molecular descriptors.

Classification vs Regression

The fish toxicity data were used for predicting the toxicity categories via both the conventional QSAR regression model and the direct classification strategies. The QSAR regression model resulted in predicted LC50 values that were converted to the two types of acute fish toxicity categories in a subsequent step. On the other hand, the classification model directly predicted the toxicity categories. The predicted acute fish toxicity categories based on both methods were compared to the true categories coming from the measured 96h LC50 values for fish toxicity to evaluate the accuracy of each approach.

The direct classification method, for both cases, resulted in around four times fewer misclassifications when compared to the QSAR regression model. We observed 47 cases of misclassification for the k-means based categories while for GHS categories the misclassified cases were 41. This was in agreement with our expectations, given that the total number of classes in GHS categories were smaller, thus a lower probablity of wrong classification. For the QSAR regression model, we observed 178 cases of wrong classifications for k-means based categories whereas 163 incorrectly classified cases were observed for the GHS categories (Figure 5). The direct classification strategy showed a homogeneous distribution of the miscategorized chemicals in the acute fish toxicity dataset, for both the k-means and the GHS categories. For the k-means categorization the QSAR regression model resulted in large and homogenous distribution of wrong categorization while for the GHS approach we observed a high density of miscategorization for high and moderate toxicity groups, (Figure 5).

Around 85% of the miscategorized chemicals via direct classification overlapped with those wrongly categorized via the QSAR regression model, independently of the type of categories. For example a chemical that was consistently wrongly categorized by all the methods was 1-hydroxypyridine-2-thione (`InChyKey:YBBJKCMMCRQZMA-UHFFFAOYSA-N`) with measured LC50 of 0.95 μg/L (i.e. -3.02 log(mg/L)). This chemical was categorized as moderately toxic by both models while it is actually a high toxicity chemical. When looking at the structure of this chemical, it is clear that this chemical is not very well covered by our training set. In other words, there are not enough (at least 4) chemicals with a similar structure to this one in our training set. This further indicates that the addition of more diverse chemical structures to our training set will result in even more accurate prediction of the toxicity categories. Additionally, the replacement of the molecular descriptors with the topographical fingerprints,[49] given their stability, may further improve our prediction accuracy.

When comparing the distribution of the wrongly categorized chemicals, we observed a higher levels of homogeneity in the k-means categories compared to the GHS ones. This was consistent for both QSAR regression model and direct classification model. We also observed that for the GHS categories, both the QSAR regression-based and the direct classification model showed a high density of wrong categorization for chemicals at the border between high toxicity and moderate toxicity region. We interpret that this is mainly caused by larger number of categories and lower levels of rigidity in the k-means approach compared to hard set thresholds (i.e. GHS approach).

The predicted LC50 values using our optimized QSAR regression model followed by the k-means clustering categorization resulted in 81% consistent classification between the acute fish toxicity categories generated by the direct classification method (Figure S5). On the other hand, the predicted LC50 values using the model developed by Aalizadeh et al[34] resulted in only 37% consistent toxicity categories. This may be due to the fact that our QSAR regression and direct classification models both had the same training set as well as the fact that our QSAR regression model uses 8 descriptors while the model by Aalizadeh et al uses only 6 from which three are logP values.

Overall, our direct classification strategy showed a better performance in identifying the acute fish toxicity categories of the chemicals directly from the molecular descriptors, rather than passing via a QSAR regression model. We also observed a higher level of consistency between the categories generated by our models compared to another prediction method (i.e. Aalizadeh model). We interpret that the main reason behind the overall better performance of the direct classification approaches is first and foremost the fact the uncertainties associated with the QSAR regression models do not impact the categorization. Additionally, the inclusion of a larger number of descriptors in such models implies that higher levels of structural features are incorporated. In fact, the low level of variance explained by individual variables further confirms this hypothesis. Our direct classification model can be easily adapted to different types of pre-defined (acute fish toxicity) categories, as demonstrated here by classifying the chemicals following the categories for short-term (acute) aquatic hazard of the GHS. Overall, these results indicate the viability of the classification strategy as a means of chemical prioritization and management.

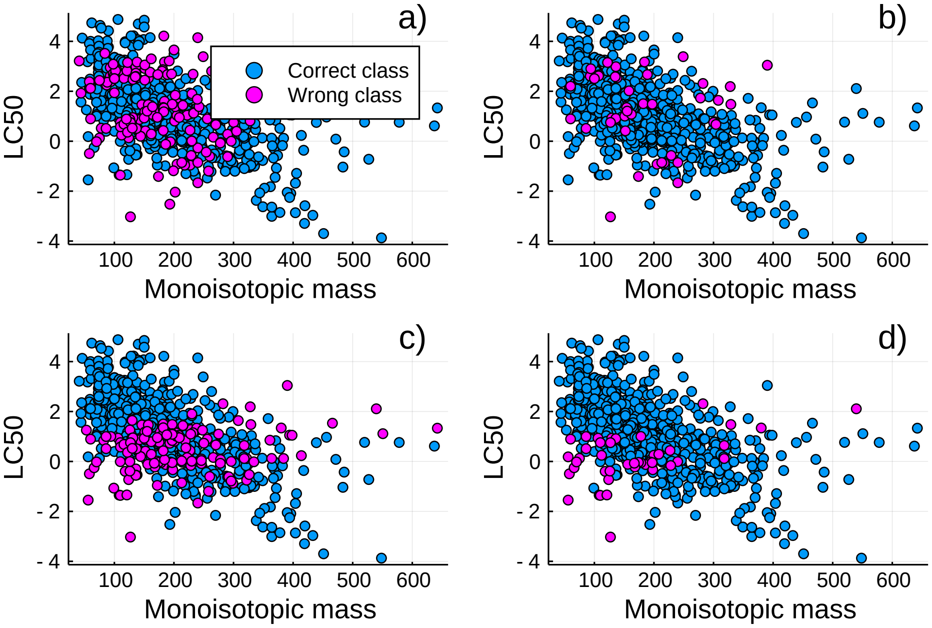


Figure 5: depicts the correctly vs wrongly predicted acute fish toxicity categories based on a) the QSAR regression model and k-means based categories; b) the direct classification strategy based on k-means categories; c) the QSAR regression model using the GHS categories categories; and d) the direct classification strategy with GHS categories.

Applicability Domain

We also evaluated the impact of the applicability domain AD selection for the assessment of the model coverage of the chemicals space. To perform such assessment, we calculated the leverage for full descriptor space, QSAR regression model descriptors, and the direct classification model descriptors. Figure 6 depicts the scores' plots for the training set and the Norman dataset and the associated applicability domains.

With the full descriptor space (i.e. the curated descriptors used for our model development), only 585 entries of Norman dataset were covered by the training set. Using the regression model descriptors (i.e. the 9 most relevant ones) resulted in around 31000 entries being covered by the training set. On the other hand, based on the descriptors of the direct classification model around 27000 entries were covered by the chemical space of the training set. The observed trend is in agreement with our expectations, given that the larger number of descriptors provides a better coverage of different structural characteristics of the chemicals. When looking at the covered chemical space by the training set (i.e. 96h LC50 for acute fish toxicity) and the chemicals within the AD of the training set (i.e. the full descriptor space) a good level of overlap is observed. This is not the case when looking at the model specific ADs, implying an extrapolation with a much larger level of prediction error. An example of such cases is carbonothioylbis(iminomethylene) bis(diethyldithiocarbamate) (InChyKey: SPQBHESGHZSSMQ-UHFFFAOYSA-N), which was covered by the regression model AD and was not covered by both the classification and the training set AD. In fact, this chemical was one of the most different chemicals compared to the chemicals in the Norman dataset (i.e. PC1 -11 and PC2 28). Therefore, it may be advisable to use the training set AD (i.e. the full descriptor space) to assess the training set coverage of the chemical space, rather than the individual model ADs.

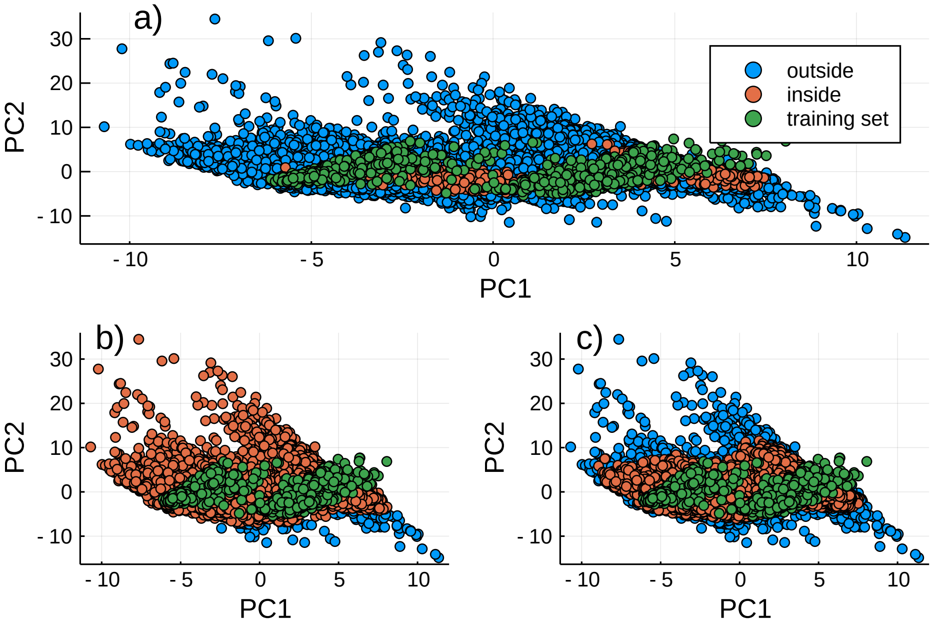


Figure 6: depicts the applicability domain (AD) assessment (i.e. the leverage calculation) of Norman dataset, based on a) the training set (i.e. the full molecular descriptor space), b) the QSAR regression model, and c) the direct classification model. The blue circles represent the chemicals that are outside of the AD while the orange circles are within the model applicability domain and the green circles are within the training set applicability domain.

Implications for Chemical Assessment

The results of our direct classification model showed its power in categorizing the chemicals in terms their acute fish toxicity based on their specific molecular descriptors. Our strategy can overcome the continuity assumption of QSAR models, which are conventionally used to fill experimental data gaps in chemical assessment of structurally similar compounds, directly impacting the size of the training set. In other words, with our direct classification approach the experimental datasets from different sources and for different chemical families can be grouped to generate larger training sets resulting in higher accuracy predictions. As demonstrated here with the direct classification of the chemicals in the Norman dataset into hazard categories defined by the GHS (based on acute fish toxicity), our approach can be adapted to different predefined categories as prescribed by various international regulations and/or classification or labeling systems. The direct classification approach can be expanded to other hazard categories (e.g. chronic toxicity) as well as to fate (e.g. mobility or persistence) and shows great potential for improving in-silico tools for chemical hazard and risk assessment.

Code Availability

The open access/source julia package for performing these calculations is available with MIT license using the link here `https://bitbucket.org/SSamanipour/toxcatpred-jl/src/main/`. Additionally, all the scripts for the model building is available in the same Bitbucket repository. Finally, the predictions of both models, and all three ADs are available for download and use via FigShare (Fish toxicity: `https://doi.org/10.21942/uva.20089751`, Norman SusDat: `https://doi.org/10.21942/uva.20089787`, and model output: `https://doi.org/10.21942/uva.20089805`).

padelpy: `https://github.com/ecrl/padelpy`

pubchempy API: `https://pubchempy.readthedocs.io/en/latest/`

ScikitLearn.jl: https://scikitlearnjl.readthedocs.io/en/latest/

Acknowledgement

SS is grateful to the UvA Data Science Center for financial support provided for this project. Moreover, SS is thankful to the members of CAST for the fruitful discussions. JO is the recipient of an NHMRC Emerging Leadership Fellowship (EL1 2009209). The Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland, gratefully acknowledges the financial support of the Queensland Health and Australian Research Council ARC Discovery Project (DP190102476). The authors thank Joanke van Dijk for help with identifying the regulatory hazard categories.

Supporting Information Available

The Supporting Information containing the details related to the samples, parameter settings, and the figures associated with the algorithms are available free of charge at ACS web site.

Author Information

Corresponding Author:
Saer Samanipour
Van 't hoff institute for molecular sciences (HIMS),
University of Amsterdam,
the Netherlands
Email: s.samanipour@uva.nl

ORCID

Saer Samanipour: 0000-0001-8270-6979

Jake W. O'Brien: 0000-0001-9336-9656

Malcolm J. Reid: 0000-0002-9988-4867

Kevin V. Thomas: 0000-0002-2155-100X

Antonia Praetorius: 0000-0003-0197-0116

References

(1) Muir, D. C. G.; Howard, P. H. Are There Other Persistent Organic Pollutants? A Challenge for Environmental Chemists. *Environ. Sci. Technol.* **2006**, *40*, 7157–7166.

(2) 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 & technology* **2020**, *54*, 2575–2584.

(3) Howard, P. H.; Muir, D. C. G. Identifying New Persistent and Bioaccumulative Organics Among Chemicals in Commerce. *Environ. Sci. Technol.* **2010**, *44*, 2277–2285.

(4) Howard, P. H.; Muir, D. C. G. Identifying New Persistent and Bioaccumulative Organics Among Chemicals in Commerce II: Pharmaceuticals. *Environmental Science & Technology* **2011**, *45*, 6938–6946.

(5) Howard, P. H.; Muir, D. C. G. Identifying New Persistent and Bioaccumulative Organics Among Chemicals in Commerce II: Pharmaceuticals. *Environ. Sci. Technol.* **2011**, *45*, 6938–6946.

(6) Escher, B. I.; Stapleton, H. M.; Schymanski, E. L. Tracking complex mixtures of chemicals in our changing environment. *Science* **2020**, *367*, 388–392.

(7) Vermeulen, R.; Schymanski, E. L.; Barabási, A.-L.; Miller, G. W. The exposome and health: Where chemistry meets biology. *Science* **2020**, *367*, 392–396.

(8) 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., et al. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *Journal of cheminformatics* **2017**, *9*, 1–27.

(9) 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 (UVCBs). *Environmen. Sci. Technol.* **2022**, *56*, 7448.

(10) Liu, H.; Papa, E.; Gramatica, P. QSAR Prediction of Estrogen Activity for a Large Set of Diverse Chemicals under the Guidance of OECD Principles. *Chem. Res. Toxicol.* **2006**, *19*, 1540–1548.

(11) Wang, T.; Yuan, X.-s.; Wu, M.-B.; Lin, J.-P.; Yang, L.-R. The advancement of multidimensional QSAR for novel drug discovery-where are we headed? *Expert opinion on drug discovery* **2017**, *12*, 769–784.

(12) Sigurnjak Bureš, M.; Cvetnić, M.; Miloloža, M.; Kučić Grgić, D.; Markić, M.; Kušić, H.; Bolanča, T.; Rogošić, M.; Ukić, Š. Modeling the toxicity of pollutants mixtures for risk assessment: a review. *Environmental Chemistry Letters* **2021**, *19*, 1629–1655.

(13) Mao, J.; Akhtar, J.; Zhang, X.; Sun, L.; Guan, S.; Li, X.; Chen, G.; Liu, J.; Jeon, K. M. S., Hyeon-Nae; No, K. T.; Wang, G. Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models. *Iscience* **2021**, *24*, 103052.

(14) Aalizadeh, R.; Ohe, P. C. v. d.; Thomaidis, N. S. Prediction of acute toxicity of emerging contaminants on the water flea Daphnia magna by Ant Colony Optimization–Support Vector Machine QSTR models. *Environ. Sci.: Processes Impacts* **2017**, *19*, 438–448.

(15) Aalizadeh, R.; Nika, M.-C.; Thomaidis, N. S. Development and application of retention

time prediction models in the suspect and non-target screening of emerging contaminants. *Journal of Hazardous Materials* **2019**, *363*, 277–285.

(16) Mansouri, K.; Grulke, C. M.; Judson, R. S.; Williams, A. J. OPERA models for predicting physicochemical properties and environmental fate endpoints. *J Cheminform* **2018**, *10*, 10.

(17) Ballabio, D.; Consonni, V. Classification tools in chemistry. Part 1: linear models. PLS-DA. *Analytical Methods* **2013**, *5*, 3790.

(18) 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.

(19) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; 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.

(20) Sheridan, R. P. Using random forest to model the domain applicability of another random forest model. *Journal of chemical information and modeling* **2013**, *53*, 2837–2850.

(21) Reppas-Chrysovitsinos, E.; Sobek, A.; MacLeod, M. Screening-level exposure-based prioritization to identify potential POPs, vPvBs and planetary boundary threats among Arctic contaminants. *Emerging Contaminants* **2017**, *3*, 85–94.

(22) Guo, J.; Sinclair, C. J.; Selby, K.; Boxall, A. B. Toxicological and ecotoxicological risk-based prioritization of pharmaceuticals in the natural environment. *Environmental toxicology and chemistry* **2016**, *35*, 1550–1559.

(23) Schulze, S.; Sättler, D.; Neumann, M.; Arp, H. P. H.; Reemtsma, T.; Berger, U. Using REACH registration data to rank the environmental emission potential of persistent and mobile organic chemicals. *Science of the Total Environment* **2018**, *625*, 1122–1128.

(24) Hale, S. E.; Arp, H. P. H.; Schliebner, I.; Neumann, M. Persistent, mobile and toxic (PMT) and very persistent and very mobile (vPvM) substances pose an equivalent level of concern to persistent, bioaccumulative and toxic (PBT) and very persistent and very bioaccumulative (vPvB) substances under REACH. *Environmental Sciences Europe* **2020**, *32*, 155.

(25) Williams, E. S.; Panko, J.; Paustenbach, D. J. The European Union's REACH regulation: a review of its history and requirements. *Critical reviews in toxicology* **2009**, *39*, 553–575.

(26) Kwiatkowski, C. F. et al. Scientific Basis for Managing PFAS as a Chemical Class. *Environ. Sci. Technol. Lett.* **2020**, *7*, 532–543.

(27) Dulio, V. et al. The NORMAN Association and the European Partnership for Chemicals Risk Assessment (PARC): let's cooperate! *Environmental Sciences Europe* **2020**, *32*, 100.

(28) Rüdel, H.; Körner, W.; Letzel, T.; Neumann, M.; Nödler, K.; Reemtsma, T. Persistent, mobile and toxic substances in the environment: a spotlight on current research and regulatory activities. *Environmental Sciences Europe* **2020**, *32*, 5.

(29) Dulio, V.; van Bavel, B.; Brorström-Lundén, E.; Harmsen, J.; Hollender, J.; Schlabach, M.; Slobodnik, J.; Thomas, K.; Koschorreck, J. Emerging pollutants in the EU: 10 years of NORMAN in support of environmental policies and regulations. *Environmental Sciences Europe* **2018**, *30*, 5.

(30) Moe, S. J.; Madsen, A. L.; Connors, K. A.; Rawlings, J. M.; Belanger, S. E.; Landis, W. G.; Wolf, R.; Lillicrap, A. D. Development of a hybrid Bayesian network model

for predicting acute fish toxicity using multiple lines of evidence. *Environmental Modelling & Software* **2020**, *126*, 104655.

(31) Linkov, I.; Massey, O.; Keisler, J.; Rusyn, I.; Hartung, T. From" weight of evidence" to quantitative data integration using multicriteria decision analysis and Bayesian methods. *Altex* **2015**, *32*, 3.

(32) Kjaerulff, U. B.; Madsen, A. L. Bayesian networks and influence diagrams. *Springer Science+ Business Media* **2008**, *200*, 114.

(33) Schymanski, E. Update on NORMAN-SusDat NORMAN-SLE (Suspect List Exchange). **2021**,

(34) Aalizadeh, R.; Peter, C.; Thomaidis, N. S. Prediction of acute toxicity of emerging contaminants on the water flea Daphnia magna by Ant Colony Optimization–Support Vector Machine QSTR models. *Environmental Science: Processes & Impacts* **2017**, *19*, 438–448.

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

(36) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **1988**, *28*, 31–36.

(37) Heller, S. R.; McNaught, A. D. The IUPAC international chemical identifier (InChI). *Chemistry International* **2009**, *31*, 7.

(38) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A., et al. PubChem substance and compound databases. *Nucleic acids research* **2016**, *44*, D1202–D1213.

(39) van den Berg, R. A.; Hoefsloot, H. C.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics* **2006**, *7*, 1–15.

(40) Hastie, T.; Tibshirani, R.; Friedman, J. H.; Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction*; Springer, 2009; Vol. 2; p 587.

(41) Miyagawa, M. Globally harmonized system of classification and labelling of chemicals (GHS) and its implementation in Japan. *Nihon Eiseigaku zasshi. Japanese Journal of Hygiene* **2010**, *65*, 5–13.

(42) Globally Harmonized System of Classification and Labelling of Chemicals (GHS Rev. 9, 2021) | UNECE. `https://unece.org/transport/standards/transport/dangerous-goods/ghs-rev9-2021`.

(43) Cassotti, M.; Ballabio, D.; Consonni, V.; Mauri, A.; Tetko, I. V.; Todeschini, R. Prediction of acute aquatic toxicity toward daphnia magna by using the ga-k nn method. *Alternatives to Laboratory Animals* **2014**, *42*, 31–41.

(44) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V., et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **2011**, *12*, 2825–2830.

(45) Cho, G.; Jung, K.; Hwang, H. Out-of-bag prediction error: A cross validation index for generalized structured component analysis. *Multivariate Behavioral Research* **2019**, *54*, 505–513.

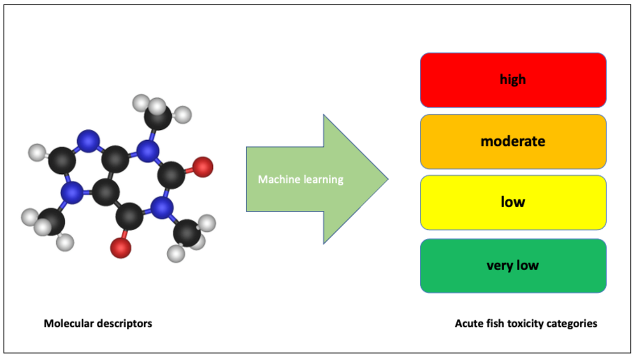
(46) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *Journal of chemical information and computer sciences* **1999**, *39*, 868–873.

(47) Cheng, T.; Zhao, Y.; Li, X.; Lin, F.; Xu, Y.; Zhang, X.; Li, Y.; Wang, R.; Lai, L. Computation of octanol- water partition coefficients by guiding an additive model with knowledge. *Journal of chemical information and modeling* **2007**, *47*, 2140–2148.

(48) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research* **2019**, *47*, D1102–D1109.

(49) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

TOC Art



Review only.