From 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), the Netherlands
‡UvA Data Science Center, University of Amsterdam, the Netherlands
¶Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland, Australia
§Norwegian Institute for Water Research (NIVA), Oslo, Norway
∥Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, the Netherlands

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

Abstract

Most chemicals present in the human and environmental exposome are structurally unknown (i.e. \( \leq 1\% \)). The European Chemicals Agency (ECHA) and US Environmental Protection Agency (EPA) have listed approximately 800k chemicals that must be further investigated for their potential environmental and/or human health risk. A significant number of these chemicals have large enough global volumes of consumption (e.g. industrial and agrochemical) to reach the limits of detection of our analytical
chemistry methods and may be toxic. In this study we present a supervised classification model that directly connects the molecular descriptors of chemicals to their toxicity. As a proof of concept we used 907 experimentally defined LC50 values for acute fish toxicity. Our classification model explained ≈ 90% of variance in our data for the training set and ≈ 80% for the test set. Direct comparison of our classification model with the conventional strategy (i.e. QSAR regression models) resulted in a 5 fold decrease in the wrong chemical categorization for our model. This optimized model was employed to predict the toxicity categories of ≈ 32k chemicals (from the Norman SusDat). Finally, a comparison between the model based applicability domain (AD) vs 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 conventionally employed QSAR methods, makes this approach a viable tool for hazard identification and risk assessment of chemicals.

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. In fact theoretically for organic chemicals, there are at least $10^{60}$ possible structures for masses $\leq$ of 500 Da, further indicating the potential diversity of the human chemical space. The latest estimates of the environmentally relevant chemicals based on the chemical registries and production volumes are estimated to be between 350k and 800k. Most of these chemicals are structurally unknown with little to no knowledge about their fate nor toxicity. Moreover, experimental assessment of the fate and toxicity of such a large number of chemicals in not feasible, resulting in wider application of modeling approaches.

Prediction of the physiochemical properties and the biological activity (e.g. aquatic
toxicity) has been one of the main approaches to deal with the diversity in the chemical space. 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. These models are often built on very homogeneous training sets (i.e. similar chemical classes), hence the linearity assumption. In fact, recent efforts have been put into using more heterogeneous training sets as well as moving away from the linearity assumption. Independent from the 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). 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 covered 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.

For these predicted and measured activities (i.e. toxicity and/or properties) to be translated into chemical management actions, they are divided into different categories using thresholds based on expert knowledge. Examples for such categories are thresholds for persistence (P), bioaccumulation potential (B) and toxicity (T) defined under Registration, Evaluation, Authorization and Restriction of Chemicals (REACH). The chemicals that fall within specific categories are then furthered for more active monitoring and eventually for legislation. 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. 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 with respect to aquatic toxicity, where the QSAR based activity prediction step is skipped. Our approach directly converts the molecular descriptors into the chemical categories, avoiding the errors inherent in the activity prediction step. As a proof of concept, this strategy was tested with lethal concentration (LC50) values for 907 organic chemicals. We compared the results of our strategy with the conventional approach. Additionally, our modeling strategy was expanded to 32000 chemicals from Norman SusDat. 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. In the next step we assessed the prediction accuracy of the toxicity categories based on two different modeling strategies: conventional QSAR regression model vs direct classification. The regression model simulated the case where the toxicity is predicted based on descriptors via a QSAR model and then the chemical is put to a specific toxicity category. On the other hand, the classification model skipped the toxicity prediction step and directly classified the chemical of interest into a toxicity category. 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 toxicity categorization.
Figure 1: depicts all the overall workflow of the study from the raw data to the finally generated models.
**Dataset**

We employed two different datasets for our modeling\(^{20}\) and the model application\(^{35}\). The modeling dataset consisted of a compilation of the experimentally defined and curated LC50 values for 907 organic structures, Figure 2\(^{36}\). These chemicals included different chemical families, including pharmaceuticals, pesticides, conventional POPs, and industrial chemicals. The second dataset was an extract of around 32000 chemicals (31722 chemicals), including their predicted LC50 values from the Norman SusDat database\(^{36}\). This dataset included only the chemicals that were reported as within the applicability domain of the QSAR model developed by Aalizadeh et al.\(^{36}\) which was used for testing our model applicability, Figure 2\(^{36}\). This is the model employed by Norman Network for their risk assessment and chemical management. When checking the overlap between the two datasets, we observed around 100 common entries.

We calculated 2757 1D, 2D, 3D, and PubChem fingerprints for both datasets using PaDEL software package\(^{37}\) implemented via a python 3 wrapper called padelpy (https://github.com/ecrl/padelpy). Additionally, the name of the chemicals, their SMILES,\(^{38}\) and InChIKeys\(^{39}\) were retrieved from the PubChem database\(^{40}\) via pubchempy API (https://pubchempy.readthedocs.io/en/latest/).

In order to identify the unstable descriptors - caused by the lack of convergence during the structural optimization, we performed the descriptor calculations for the 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.\(^{41}\) After scaling, the variance of each descriptor in the 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 fish toxicity dataset are also stable for the Norman dataset. Therefore, the descriptors for this dataset was calculated only once. Additionally, the maximum of
each descriptor in the Norman dataset was compared to those from the training set. 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.

![Figure 2:](a) the distribution of the experimental LC50 values used for the model development and validation whereas b) shows the chemical space covered by the fish toxicity data (i.e. training and test sets) and the Norman dataset.)

**Modeling**

In this study, we developed three different models namely: 1) k-means clustering for the toxicity categorization; 2) QSAR regression model; and 3) a direct classification model. The details of each model strategy is provided below. All these models, once optimized, were used with the Norman dataset to further assess their applicability.
K-means Clustering for Toxicity Categorization

We generated four different toxicity categories (i.e. very low toxicity, low toxicity, moderate toxicity, and high toxicity), using the combination of the LC50 and monoisotopic mass of the chemicals. To build and validate our model, we employed a bootstrapping strategy with 500 iterations. Over each iteration, 90% of the fish toxicity data was used as the training set and 10% as the test set.

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). Also in this case, the fish toxicity dataset was divided into training set 90% and test set 10% of the full dataset. For the regression model, the model hyper-parameters 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. In this case we also used a threshold of 1% relative variable importance.

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 LC50 of the chemicals in the Norman dataset. In a second step, the predicted LC50 values were used to categorize the chemicals into the toxicity categories described above.

Descriptor Based Direct Classification Model

We developed, validated, and tested a classification model to convert the curated descriptors to the toxicity categories. For this model, we employed random forest classification, implemented via ScikitLearn.jl julia package.
For the direct classification, we split the fish toxicity dataset (i.e. curated the descriptors and toxicity categories) into training set 90% and test set 10% of the full dataset. The training set was used for the model development and optimization while the test set was for further evaluation of the dataset. 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, 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 fish toxicity dataset was used with the selected variables. In this case all variables were used for the final model building. This model was further tested with the test set, which was unknown to the model and was not used during the model development, optimization, and validation. Additionally, this model was used to categorize the Norman dataset into the 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. The leverage calculations were performed both on the full descriptor space as well as the descriptors relevant to specific models. This strategy enabled us to systematically assess which chemicals are well represented by the training set.

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 classification strategy to directly convert the molecular descriptors of chemicals to fish toxicity categories. This model was developed, validated, and tested via an experimentally defined dataset of LC50 values for 907 organic chemicals. The results of this strategy was directly compared to the conventional approach - first QSAR prediction and then categorization - both for the fish toxicity data and a dataset of ≈ 23000 chemicals from Norman SusDat.

Performance of K-means Clustering

The final k-means model resulted in a clustering accuracy of 97.5%. This model, then, was fed the full 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 four categories were used as labels in the classification model while the clustering model...
was used to convert the predicted toxicities based on the regression model to the toxicity categories.

Figure 3: shows the distribution of the four toxicity categories of the fish toxicity dataset via the best k-means model.

**Performance of Regression Model**

The residuals of the final and optimized 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, Figure 4. We observed up to two LC50 units overestimation of the toxicity for values \( \leq 2 \) while our model resulted in a slight underestimation of toxicity for LC50 values \( \geq 6 \), Figures 4 and S2. Finally, we used the optimized model to predict
the LC50 values for the Norman dataset. When comparing the results of our predictions to the predictions by Aalizadeh et al, a clear linear trend 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, S1. The most relevant variable was the Crippen logP value explaining around 35% variance the the final model. This logP was calculated based on 68 atomic contributions. On the other hand, the second variable was XLogP, implemented within PubChem. This logP calculation also uses the atomic contribution 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. Finally, the remaining relevant 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.

Performance of Classification Model

The optimized 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. 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 ones), Figure S4. Differently from the regression model, the most relevant variable only explained $\approx 1.5\%$ of variance 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.
The reported statistics and the selected variables in our classification model further indicated the applicability of our model for prediction of the 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 (i.e. regression model) and the direct classification strategy. The regression model resulted in predicted LC50 values that were converted to the toxicity categories, employing the k-mean model. On the other hand, the classification model directly predicted the toxicity categories. The predicted toxicity categories based on both methods were compared to the true categories coming from the measured LC50 values to evaluate the accuracy of each approach. The classification method resulted in around five times less misclassifications when compared
to the regression strategy. The classification model resulted in 35 cases of wrong toxicity
categorization while the regression approach yielded 152 cases of wrong toxicity categories,
Figure 5. The direct classification strategy showed a heterogeneous distribution of the mis-
categorized chemicals in the fish toxicity dataset. On the other hand the regression based
approach resulted in a larger density of wrong categorizations for moderate and low tox-
icity groups, Figure 5. There were 18 chemicals out of 35 that were wrongly categorized
by both approaches. An example of such chemicals was gamma-undecalactone (InChyKey:
PHXATPHONSXBIL-UHFFFAOYSA-N), which using both approaches was classified as a
moderately toxic chemical while based on its measured LC50 value (i.e. 2.09) it was a highly
toxic chemical. For this chemical the regression model predicted an LC50 value of 3.72,
resulting in a wrong toxicity category. As for the classification model, the combination of
logP and polarizability descriptors were the main drivers of the observed miscategorization.
Similar trends were observed for the 17 chemicals that were wrongly categorized by the clas-
sification approach (e.g. 3,7-dimethyloct-6-enal).

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

Overall, our classification strategy showed a better performance in identifying the toxicity
categories of the chemicals directly from the molecular descriptors, rather than passing via
a QSAR (i.e. regression model) approach. We also observed a higher level of consistency
between the categories generated by our models compared to other prediction method (i.e.
Aalizadeh model). These results further indicate the viability of the classification strategy as a means of chemical management.

Figure 5: depicts the correctly vs wrongly predicted toxicity categories based on a) the regression model (i.e. conventional QSAR model) and b) the classification strategy.

Applicability Domain

We also evaluated the impact of the applicability domain selection for the assessment of the model coverage of the chemicals space. To perform such assessment, we calculated the leverage for full descriptor space, regression model descriptors, and the classification model descriptors, Figure 6.

With the full descriptor space, only 585 entries were covered by the training set whereas using the regression model descriptors (i.e. the 9 most relevant ones) this resulted in around 31000 entries being covered by the training set. On the other hand, based on the descriptors of the 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. 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(diythylthiocarbamate) (InChyKey:
SPQHESGHZSSMQ-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 assessment (i.e. the leverage calculation) of a) the
training set (i.e. the full descriptor space), b) the regression model, and c) the classification model.
Conclusions

In this study, we have developed, optimized, and validated a direct classification strategy for prediction of the toxicity categories of organic chemicals. This model was directly compared to the conventional strategy consisting of QSAR prediction (regression based model) of the toxicity followed by toxicity categorization of the prediction. The direct classification model showed five folds higher accuracy in defining the toxicity category of 907 chemicals with experimentally defined LC50 values compared to the regression based models (i.e. the conventional approach). The results of our classification model showed its power in categorizing the chemicals based on their toxicity. This approach can be expanded to other hazard categories as well as to fate (e.g. mobility). Finally, it should be noted that such a strategy can overcome the continuity assumption of QSAR models, which directly impacts the size of the training set. In other words, the experimental datasets from different sources and for different chemical families can be grouped to generate larger training sets resulting in higher accuracy predictions.

Additionally, the training set vs model AD assessment was performed to examine the coverage of the chemical space of the training set. The data indicated that the training set AD is much more adequate for a systematic assessment than the model based ADs to avoid model extrapolation. Our results also suggested the need for larger training sets to generate models with higher levels of accuracy and a better coverage of the exposome chemical space.

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/](https://bitbucket.org/SSamanipour/toxcatpred-jl/) 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: 10.21942/uva.20089751, Norman SusDat: 10.21942/uva.20089787, and model output: 10.21942/uva.20089805).
padelpy: https://github.com/ecrl/padelpy
pubchempy API: https://pubchempy.readthedocs.io/en/latest/

Acknowledgement

SS is grateful the UvA Date 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).

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 website.

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


(11) 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). Environmental Science & Technology 2022,


(15) Mao, J.; Akhtar, J.; Zhang, X.; Sun, L.; Guan, S.; Li, X.; Chen, G.; Liu, J.; Jeon, H.-


(22) 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.


(26) 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.


(33) 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.


Review only.