UC Santa Barbara

UC Santa Barbara Previously Published Works

Title

Accelerating the pace of ecotoxicological assessment using artificial intelligence.

Permalink

https://escholarship.org/uc/item/04q7f238

Journal

Ambio, 51(3)

ISSN

0044-7447

Authors

Song, Runsheng
Li, Dingsheng
Chang, Alexander
et al.

Publication Date

2022-03-01

DOI

10.1007/s13280-021-01598-8

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Ambio 2022, 51:598–610
https://doi.org/10.1007/s13280-021-01598-8

KUNGL. VETENSKAPS-AKADEMIEN
THE ROYAL SWEDISH ACADEMY OF SCIENCES

Check for updates

RESEARCH ARTICLE

Accelerating the pace of ecotoxicological assessment using artificial intelligence

Runsheng Song, Dingsheng Li, Alexander Chang, Mengya Tao, Yuwei Qin, Arturo A. Keller, Sangwon Suh

Received: 3 December 2020 / Revised: 12 April 2021 / Accepted: 29 June 2021 / Published online: 24 August 2021

Abstract Species Sensitivity Distribution (SSD) is a key metric for understanding the potential ecotoxicological impacts of chemicals. However, SSDs have been developed to estimate for only handful of chemicals due to the scarcity of experimental toxicity data. Here we present a novel approach to expand the chemical coverage of SSDs using Artificial Neural Network (ANN). We collected over 2000 experimental toxicity data in Lethal Concentration 50 (LC50) for 8 aquatic species and trained an ANN model for each of the 8 aquatic species based on molecular structure. The R^2 values of resulting ANN models range from 0.54 to 0.75 (median R^2 = 0.69). We applied the predicted LC50 values to fit SSD curves using bootstrapping method, generating SSDs for 8424 chemicals in the ToX21 database. The dataset is expected to serve as a screening-level reference SSD database for understanding potential ecotoxicological impacts of chemicals.

Keywords Chemical toxicity · Environmental toxicity · Life cycle assessment · Machine learning · QSAR

INTRODUCTION

Climate change, habitat losses and the exposure to various man-made chemicals are major threats to global biodiversity (Hartley 2002; Vörösmarty 2010; Malaj 2014). According to the Red List of Threatened Species by the International Union for Conservation of Nature (IUCN), 1256 out of the total 8455 threats are associated with pollution, of which 251 are due solely to the pesticide and herbicide (The IUCN Red List of Threatened Species).

Our understanding of chemical's toxicity footprints on the ecosystem, however, is hampered by the sheer number and diversity of the chemicals used by the society, their wide variation in sensitivity across species, and the high costs—and therefore the scarcity—of experimental toxicity data (Bressler et al. 2006; Holmstrup 2010; Martin 2019). The number of unique chemicals have been produced or used in the European Union (EU) countries in excess of one tonne per year reached 15,000 in 2018 and is growing in the past years (ECHA Publishes Official Statistics for the Last REACH Registration Deadline). Different species may exhibit dramatically different sensitivity to the same chemical. Pyrethroid, for example, is extremely toxic to insects, but it is well tolerated by most mammals (Wolansky and Harrill 2008).

The Species Sensitivity Distribution (SSD) is an approach that allows estimating the potential ecosystem impacts of a chemical considering the variation in the sensitivity of species to toxicants. SSD uses the statistical distribution of toxicity data points (Lethal Concentration, or LC50, for example) across multiple species as a proxy measure for the ecotoxicological impact of single stressor to the entire community (forum, E.-U. E. 1998; Posthuma et al. 2001). SSDs, combined with an assessment factor, are often used in risk assessment to estimate the Predicted No Effect Concentration (PNEC) (Raimondo et al. 2008; Ping et al. 2011). In environmental risk assessment, PNEC is often regarded as the safe concentration for chemical under which the entire aquatic ecosystem is unlikely be adversely affected (Calow and Forbes 2003; Cunningham et al. 2009). Furthermore, SSD can be also used in Life Cycle Assessment (LCA) and Life Cycle Impact Assessment (LCIA), where the Hazard Concentration at which half of

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13280-021-01598-8.

© The Author(s) 2021
www.kva.se/en

the species are adversely affected, or HC50 value, is often used to derive the ecotoxicity Characterization Factors (CFs) of chemicals (Rosenbaum 2008; Henderson 2011).

The challenge is that experimental toxicity data are scarce, and developing an SSD of a chemical requires multiple toxicity data points across multiple species (Garner et al. 2015). The recommended minimum sample size ranges from 8 to 15 (Newman et al. 2009; Lowry 2012). The ECOTOX database, one of largest databases for experimental toxicity values, contains about 500 organic chemicals with experimental toxicity data for aquatic species, and only about 80 aquatic species have been tested on more than 5 organic chemicals. In USETOX, which is one of the characterization models widely used in LCA, only about 2000 CFs, which can be calculated using SSD, were derived using experimental toxicity data that exist for organic chemicals (Rosenbaum 2008). The scarcity of experimental toxicity data is the primary barrier for developing SSDs and for understanding the ecotoxicological impact of chemicals (Andersen and Krewski 2009).

One of the approaches to overcome the scarcity of experimental data is the use of Quantitative Structure–Activity Relationship (QSAR) models. QSAR models estimate a chemical's bioactivity or toxicity using the structure of the chemical in the absence of experimental data (Cherkasov 2014). QSARs often use linear regression or logistic regression models (Worth and Cronin 2003; Chen et al. 2012). Mayer and colleagues for example, predicted chronic toxicity of chemicals to multiple fish species using linear regression model and acute toxicity test data (Mayer et al. 1994). Raevsky and colleagues estimated the LC50 values of chemicals to Guppy, Fathead Minnow and Rainbow Trout using chemical similarity approach (Raevsky et al. 2008). These QSARs, however, are designed to be applied to targeted groups of chemicals such as those with nonpolar Mode of Action (MOA) (Raevsky et al. 2008), and, when applied to other groups of chemicals, fail to provide reliable predictions (Cherkasov 2014).

Recent progresses in machine learning techniques, however, opens an entirely new avenue of opportunities for developing predictive models (Haupt et al. 2008). Artificial Neural Network (ANN), for example, has been successfully applied to predict rate constants and reaction rates of chemicals in atmosphere (Allison 2016) and extreme weather (Liu, et al. 2016), and QSARs using simpler neural networks have also been used to estimate acute toxicity of chemicals to few aquatic species using inputs in varies formats. For example, Devillers developed QSAR model to estimate the acute toxicity of pesticide for *Lepomis macrochirus* (Devillers 2001). Martin and colleagues provided a new model in Neural Networks to estimate the LC50 (96 h) for *Fathead Minnow*, and achieved satisfying performance (Martin and Young 2001). However, because of the development of SSDs requires the ecotoxicity data in comparable experimental conditions applied across various taxa, therefore, the existing QSARs from different studies cannot be assembled together to generate SSDs, and currently there isn't an established method to estimate SSDs with machine learning techniques. Researchers typically combine existing QSARs for multiple species, which may employ disparate machine learning models, limiting the interpretability of the results when used simultaneously to estimate SSDs. Furthermore, the existing QSARs were developed with various sizes of training dataset, which are often smaller than few hundreds of chemicals (Burden and Winkler 1999; Devillers 2001; Devillers 2001; Kaiser 2003).

In this study, we present a novel approach to develop SSDs for organic chemicals using assembled machine learning techniques, taking only molecular structure information as the input. We collected over 2000 experimental ecotoxicity data points in LC50, produced under comparable experimental conditions for 8 aquatic species. Using these data and molecular descriptors, we developed ANN models to estimate the ecotoxicity of chemicals in LC50. A total of 8 ANN models were trained on experimental toxicity data for each of 8 aquatic species: *Pimephales Promelas, Daphnia Magna, Oryzias Latipes, Oncorhynchus Mykiss, Lepomis Macrochirus, Cyprinodon Variegatus, Americamysis Bahia* and other water fleas. The performances of the predictive SSDs were evaluated on existing SSDs built by experimental data. The uncertainties of the ANN models as well as the predictive SSDs were analyzed. Finally, we applied our model and estimated the SSDs for over 8000 organic chemicals in the Toxicology Testing in the 21st Century (ToX21) database and characterized their SSDs as well as the HC5 values. The performances of log-normal, Gamma and Weibull distributions to fit SSD were also evaluated.

MATERIALS AND METHODS

Ecotoxicity dataset collection

We collected 2521 experimental ecotoxicity data for non-ionizable organic chemicals on 8 aquatic species: *Pimephales Promelas, Daphnia Magna, Oryzias Latipes, Oncorhynchus Mykiss, Lepomis Macrochirus, Cyprinodon Variegatus, Americamysis Bahia* and other water fleas, from major public databases, including ECOTOX, eChem, EFSA and HSDB (Todeschini and Consonni 2008; eChemPortal-Home; ECOTOX|MED|US EPA; ESFA; Hazardous Substances Data Bank (HSDB)). Data from peer-reviewed literatures was also added as supplementary data source to

© The Author(s) 2021

develop the neural network models in this study (Russom et al. 1997; Devillers 2001; Martin and Young 2001; Raevsky et al. 2008; Results of ecotoxicity tests data conducted by Ministry of the Environment in 2014; Austin et al. 2015; Toropov 2017). The number of experimental data collected for each species can be found in Fig. S1 in supplementary information. To ensure data quality of the ecotoxicity dataset we collected from this study, the critical experimental conditions, such as the testing duration, chemical purity and *pH* values were strictly controlled. We filtered the datasets and used only the LC50 data with 96 h of duration for all species except for water fleas, for which 48 h' data was used due to the concerns of dataset size. Chemical purity must be higher than 85%. And the *pH* value must be in the range of 5 to 9. Experimental data that not meet these requirements was discarded. For chemical with multiple experimental values, the geometric mean was used in the final dataset. To utilize some of the discarded data, and to increase the diversity of the species taxa, experimental values that met our data selection procedure for other water fleas in ECOTOX database was combined and treated as an individual species in this study. Within this category, there are 20 chemicals for *Ceriodaphnia Dubia,* 13 chemicals for *Daphnia Pulex and* 63 chemicals for Mix Water Flea (not specified). Additional information, such as the CAS number, SMILEs, molecular weight and the chemical names were also collected, for referencing purpose. The unit of the LC50 values were converted to log10(LC50) in μmol/L. The final dataset is available in the supplementary information. mol/L. The final dataset is available in the supplementary information.

Two-step molecular descriptor selection

The original molecular structural descriptors were calculated using Python packages rdkit and Mordred (rdkit: The official sources for the RDKit library. 2017; Moriwaki et al. 2018). The descriptor calculators can produce over 2000 descriptors for a single chemical, including basic physicochemical properties and autocorrelation descriptors. Large amount of descriptors could lead to overfitting problem (Kaiser 2003; Cherkasov 2014). Two-step feature selection procedures: filter-based plus tree-based feature selection, were used in this study to extract more meaningful descriptors (Guyon and Elisseeff 2003; Saeys et al. 2007).

Filter-based feature selection removes descriptors that have low variance, as well as the descriptors have high mutual correlations with others (Stojić et al. 2010). Tree-based feature selection method ranks the importance of each descriptor by their contribution to the prediction results in a decision tree model (Sugumaran et al. 2007) (Broderius and Kahl 1985). In this study, during the filter-based feature selection, descriptors with variance lower than 10 were discarded. Then, the correlations between every leftover descriptor were calculated and the second descriptor was discarded if a descriptor pair has correlation higher than 0.6. A decision tree regressor in Python package Sklearn was used as the basis for the tree-based feature selection on the remaining descriptors (Pedregosa 2011). The descriptors that contribute to the toxicity endpoint three times higher than the mean contribution were selected as the final descriptors in this study. As a result, the final descriptors are same for every chemical for one species, but are different between species (different ANN models). In this study, we used 8 to 15 structural descriptors for developing our models. The most frequently utilized molecular descriptor was SLogP (Wildman–Crippen LogP), which appeared in all models. Xp-2dv (2-ordered Chi path weighted by valence electrons) and PEOE_VSA6 (MOE Charge VSA Descriptor 6) were used in more than 3 models. The full list of descriptors used to develop each model in can be found in Table S6 of the supplementary information.

The development of neural networks models and their applicable domain

ANNs were used as the modeling basis of the QSARs in this study. The ANNs were developed using Tensorflow and Keras in Python 2.7 (Chollet 2015; Abadi et al. 2016). The hyper-parameters of ANNs that were optimized through fivefold cross-validation in this study, including the number of hidden layer(s), the number of hidden neuron(s) in each layer, the regularization factor and the type of activation function. These hyper-parameters were optimized by minimizing the mean square error (MSE) of the ANN models while holding others constant. The final models were built using the hyper-parameters that generated the lowest MSE during cross-validation. The final model performances were reported on 20 chemicals that were randomly selected and left out during model development for each species. The ANNs were built on the rest of data.

ANNs have better performance on inputs that are like the training data. We used Euclidean distance from the input descriptors to the centroid of our training data as the metric to evaluate the Applicable Domain (AD) in this study. The Euclidean distance is calculated as:

$$d_n = \sqrt{\sum (X_i - C_i)^2} \quad (1)$$

where d_n is the distance of chemical n to the centroid of training data C; X_i and C_i are the ith molecular descriptors of the input chemical and the training data. The centroid of the training data was calculated as the mean value of the molecular descriptors of all chemicals in the training data.

© The Author(s) 2021
www.kva.se/en

Whether an input chemical falls inside the model AD was determined by comparing a threshold value K with the distance d_n. For each ANNs, we first selected an initial K and then grouped the chemicals in the validation dataset by their distance to the centroid of the training data comparing with the K value (smaller or larger). The differences of the MSEs between these two groups were calculated. We then gradually increased the K value. The MSE differences changed accordingly since the chemicals within each group are different. We selected the K value that has the largest MSE difference to be the final threshold for model AD. The performance of this AD estimation was reported on the chemicals in the testing dataset.

The development of SSDs and their uncertainties

SSD is a statistical distribution that illustrate the variation in the response of species to the exposure of chemicals. The development of SSD begins with the generation of individual toxicity value of chemicals to species. In this study, we used LC50 values of chemicals to aquatic species. The LC50s are ranked from low to high, or the most sensitive to the least sensitive species. On the SSD graph, as shown on Fig. 2, the x-axis is the concentration of chemical, and the y-axis stands for the percentage of species affected. For each data point, the location on y-axis is the Median Rank position of it. Which is calculated using the ppoint function in R, and reproduced in Python (ppoints function|R Documentation).

Therefore, the LC50 values are used to estimate the Cumulative Distribution Function (CDF) of a selected distribution. Most of the SSDs were fitted using normal or log-normal distributions (Wheeler et al. 2002; Aldenberg and Rorije 2013). Other statistical method including log-logistic distribution and Burr Type III method are also exist but have not been widely used (Aldenberg and Slob 1993; Wheeler et al. 2002). In this study, we used log-normal distribution as the basic distribution to fit SSDs, which is justified by the OVL analysis. The CDF of log-normal distribution is presented in Eq. (2):

$$F_x(x) = \Phi\left(\frac{(\ln x) - \mu}{\sigma}\right) \tag{2}$$

where Φ is the CDF for a standard normal distribution $N(0, 1)$, shown in Eq. (3), and μ and σ are the mean and standard deviation.

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \tag{3}$$

In this study, the decision of using log-normal distribution to fit SSD was made through running Overlapping Coefficient Analysis (OVL) testing on the screening results of ToX21 database. OVL is a measurement for the similarity of distributions, which compare the percentage of overlapping of the Probability Density Function (PDF) (Qin and Suh 2017). Equation (4) shows the mathematic representation of OVL for distributions $f_a(x)$ and $f_b(x)$:

$$\Delta(f_a(x), f_b(x)) = \int \min\{f_a(x), f_b(x)\} \mathrm{d}x \tag{4}$$

For each chemical in ToX21 library, the actual distribution of the LC50 values on 8 species were compared with the empirical distributions that are fitted using the mean and standard deviation values on log-normal, Weibull and Gamma distributions. The area of overlapping was calculated.

Bootstrapping approach was used to estimate the uncertainty of SSD due to the limited amount of data points (MacKinnon et al. 2004). During each iteration of bootstrapping, eight data points were resampled using the fitted distribution curve and the newly sampled data points were used to construct new distribution curve. This process was repeated for 1000 times, generating the upper and lower bounds of SSD for each chemical. The uncertainty of the QSAR predictions were also considered in the SSDs. Depending on whether the chemical fell inside or outside a model AD, different MSEs were attached to the QSAR predicted values. Therefore, the upper and lower bounds of SSDs can be reported.

Database screening

The chemical list in the ToX21 project is used as the candidates to be screened against the models developed in this study (US EPA 2015a). ToX21project aims to develop better toxicity assessment techniques in high-throughput robotic screening system. To date, about 10 000 chemicals have been tested under the project, and the screening results help to identify chemicals for further investigation (US EPA 2015a). We removed inorganic chemicals, ionized chemicals and chemicals that can't find SMILEs within this list. As a result, 8424 chemicals are left and developed predictive SSDs using the models in this study. Among these chemicals, 1239 chemicals fell into the ADs for more than 4 (out of 8) ANN models. We considered these predictive SSDs are trustful and discarded the rest of predictive SSDs.

HC5 values for these (1239) chemicals were derived from the predictive SSDs. Among them, 218 chemicals were registered in the ECHA database, therefore we were able to find the production bands for them (Registered

© The Author(s) 2021

substances-ECHA). To consider ecotoxicity and production volume at the same time when comparing chemicals, we considered them when evaluating the threatening of the candidate chemicals. The threatening is calculated as described in Eq. (5). The screening results for all chemicals, as well as their production band can be found in the supporting information.

$$T = \frac{P}{HC5} \tag{5}$$

where T stands for the threatening (tonne·L/year·umol), which is a comparative score; P (tonne/year) is the annual production band reported in ECHA database; $HC5$ (umol/L) is the hazardous concentration read from the predictive SSD.

RESULTS

ANN model performances and applicable domain

Figure 1 shows the performance of the ANN models. Circles in blue color are the training data points and triangles in red are randomly selected testing data. If predicted values match perfectly with experimental values, all the data points would be perfectly aligned on the line, $x = y$. The R^2 of these ANN models ranged from 0.54 to 0.75 (mean 0.67, medium 0.69) on the testing data. More information about the hyper-parameters of these models, such as the number of hidden layers and neurons, is summarized in Table 1. Other details about the model structure, including the activation functions and regularization factors during training can be found in the supplementary information (Table S1). According to the results on the testing chemicals, the models for *Daphnia Magna* and *Oncorhynchus Mykiss* showed the highest R^2 on testing data (0.75), followed by the *Lepomis Macrochirus* (0.72) and *Pimephales Promelas* (0.71) models, while the *Oryzias Latipes* model showed the lowest R^2 on the testing data (0.54).

To evaluate the model prediction confidence, we employed the concept of Applicable Domain (AD) to characterize the prediction accuracies of the ANN models and serves as a proxy to estimate whether a chemical is appropriate for the QSARs. The results of AD analysis are presented in Table S3 in the supplementary information. Among the ANN models that we developed, *Oncorhynchus Mykiss* and the *Lepomis Macrochirus* models have the narrowest ADs. For these two models, the mean square error (MSEs) of the testing data inside of the ADs was 6%, while those outside of AD were 15% and 22%, respectively. For the *Pimephales Promelas* model, however, the average MSEs inside and outside of AD were 8% to 220%, respectively, indicating limited utility of the model outside of AD.

Predictive species sensitivity distributions and evaluations

Using our ANN models, we estimated the LC50 values for 8424 chemicals from the ToX21 database for each of the 8 aquatic species. We also estimated the prediction errors of the ANN models, as well as the inherent error of SSDs due to the limited number of data points. These SSDs can be found in the supporting information. Given the large number of chemicals in our results, we randomly selected a few chemicals to compare our predictive SSDs with the SSDs derived from experimental data. Elaborated here is one of them, DCMU (*3-(3,4-dichlorophenyl)-1,1-dimethylurea*), an algaecide.

The predictive SSD for DCMU is shown in red line in Fig. 2. The figure also shows the uncertainty range of the ANN-derived SSD in gray. This uncertainty range was calculated using the prediction error of each ANN model, which was determined by whether this chemical fell into the AD of each model or not. For a comparison, we collected experimental data for the same species, and we located experimental LC50 values for the same list of species other than *Oryzias Latipes*, which were unavailable in the literature and databases available to us. Using these experimental values, we constructed an SSD as shown by the green line in Fig. 2. According to the SSD derived from experimental values, the HC5 of DCMU is about 1.82 mmol/L, whereas the HC5 from the ANN-based SSD ranged from 2.51 to 3.24 mmol/L. Both experimental SSD and the predictive SSD show that *Pimephales Promelas* has the best tolerance to DCMU in water, with an experimental LC50 of 61.7 mmol/L and a predicted LC50 of 75.9 mmol/L. Figure 2 indicates that the predicted SSD tends to show lower toxicity for this chemical at lower concentration (i.e., < 0.5 Log μmol/L), and higher toxicity with higher concentration (i.e., > 1.5 Log μmol/L), which will be discussed in the next section.

Another 10 organic chemicals were randomly selected from the ECOTOX databases to evaluate the SSDs derived from our ANN models. We collected experimental LC50 data of these chemicals on other species than the aforementioned 8 species in order to avoid any overlap with the training data we used to develop our ANN models. Given the inherent uncertainty in SSDs due to the limited number of data points, we used the bootstrapping technique to visualize the potential range of SSDs. The mean, lower,

© The Author(s) 2021
www.kva.se/en

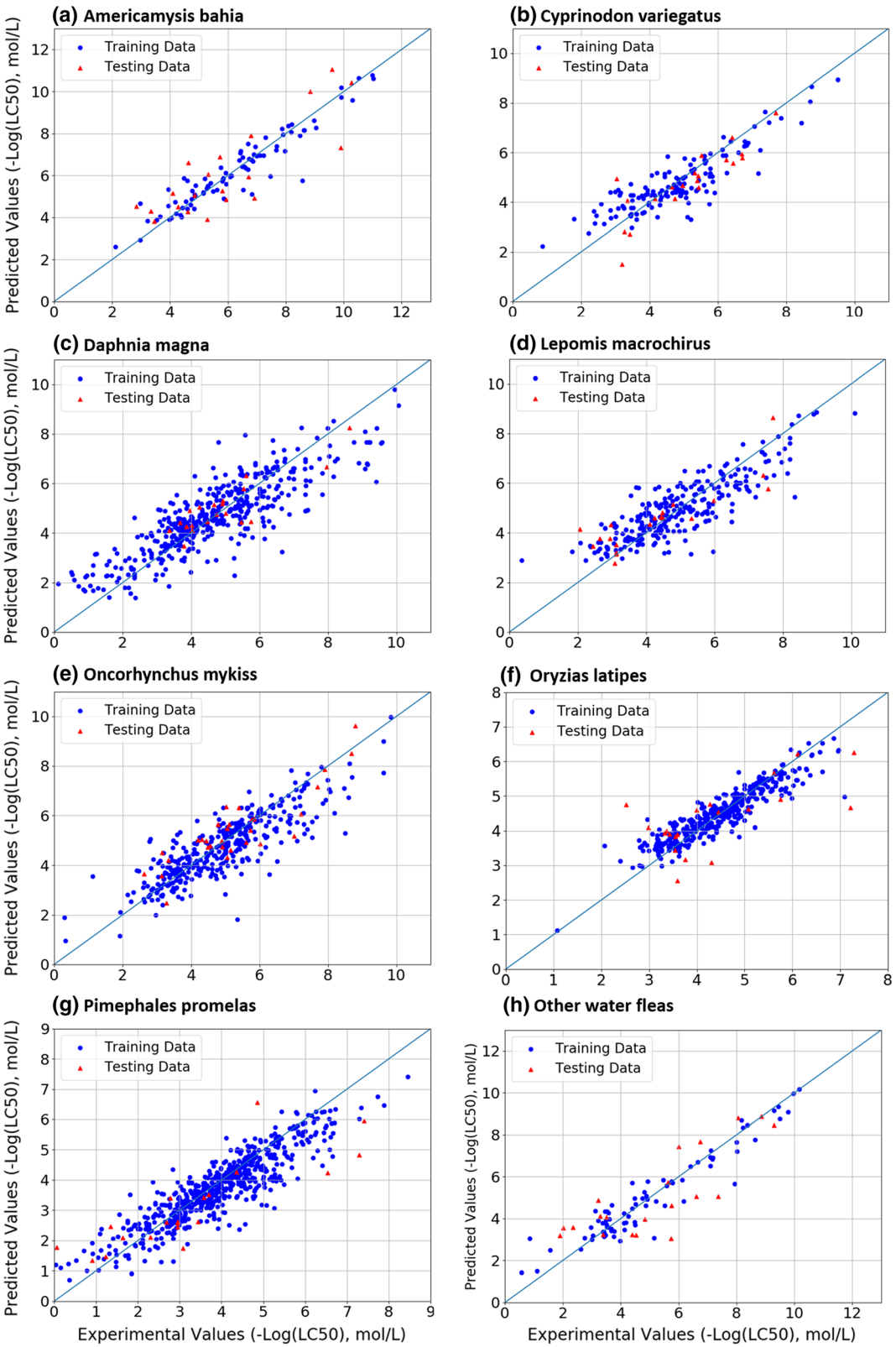

Fig. 1 The performances of all models in this study on the training data (blue circles) and testing data (red triangles). The horizontal axis is the experimental values, and the vertical axis is the predicted values. The model structures were tuned using cross-validation technique

© The Author(s) 2021

Table 1 The performance of the ANN models on the testing data for the 8 aquatic species in R^2. The number of hidden layers and hidden neurons for each ANN model

	*PP	*DM	*OL	*OM	*LM	*CV	*AB	*OWF
Model performance (R^2) on testing data	0.71	0.75	0.54	0.75	0.72	0.66	0.67	0.63
Number of hidden layer	2	1	2	2	2	2	1	2
Number of hidden neuron in each layer	32 × 16	16	64 × 32	64 × 32	32 × 16	16 × 8	16	16 × 8

*Spcies acronyms: *Americamysis Bahia* (A.B.); *Daphnia Magna* (D.M.); *Lepomis Macrochirus* (L.M.); *Oncorhynchus Mykiss* (O.M.); *Cyprinodon Variegatus* (C.V.); *Oryzias Latipes* (O.L.); *Pimephales Promelas* (P.P.) and Other Water Fleas (O.W.F.)

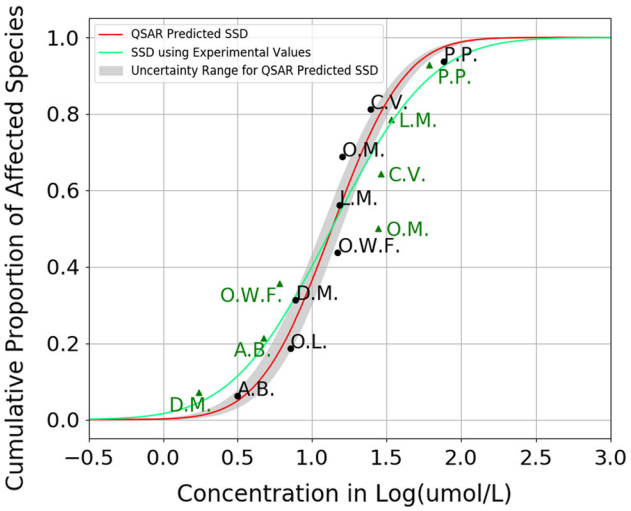

Fig. 2 The SSD of DCMU (solid red line) constructed using the ANN-based LC50 values (black points), along with the uncertainty of ANN predictions (gray area), based on the model AD estimation for *Americamysis Bahia* (A.B.), *Daphnia Magna* (D.M.), *Lepomis Macrochirus* (L.M.), *Oncorhynchus Mykiss* (O.M.), *Cyprinodon Variegatus* (C.V.), *Oryzias Latipes* (O.L.), *Pimephales Promelas* (P.P.), and Other Water Fleas (O.W.F.). The SSD in green was constructed using experimental LC50 values found for 7 species)

and upper bounds of HC50 (hazardous concentration for 50% of the species) values on both predictive and experimental SSD curves are presented in Table 2. The Overlapping Coefficient (OVL) score in Table 2 shows the percentage of overlapping of the area of the predictive distribution and the experimental distribution. The detailed model prediction data for each of the chemicals, as well as the experimental LC50 values can be found in Table S4, and in the supplementary information. The predictive SSD, experimental SSD along with their overlapping area for chemical *chlorpyrifos* (*2921-88-2*) are presented in Fig. S2 as an example.

Table 2 shows that the predicted HC50 values generated by the ANN models are generally in line with the experimental SSDs. The OVL results show that 8 out 10 chemicals have OVL score higher than 70%, which means that 70% of the area in the predictive SSD overlap with the SSD generated by the experimental data. Among them, the predictive SSD for the chemical *diazinon* (333-41-5) shares the largest overlapping area with the experimental SSD (96.8%), followed by the chemical *chlorpyrifos* (2921-88-2) by 92.0% overlapping area. The predictive SSD shows

Table 2 The HC50 values of 10 chemicals in the ECOTOX database, along with the mean HC50 values for both ANN-based SSD and the experimental SSD, as well as the percentage of overlapping of the distributions based on the predictive and experimental SSDs

Chemical CAS	Chemical name	HC50 mean (lower, upper bounds) in log (μmol/L)		OVL score (%)
		Predicted	Experimental	
50-29-3	*Clofenotane*	− 0.45 (− 1.5, 0.62)	− 0.85 (− 1.43, − 0.26)	70.6
87-86-5	*Pentachlorophenol*	0.32 (0.04, 0.62)	0.23 (− 0.11, 0.54)	89.6
58-89-9	*Lindane*	1.29 (0.26, 2.22)	0.87 (0.36, 1.4)	65.8
60207-90-1	*Propiconazole*	0.64 (0.08, 1.25)	0.88 (0.5, 1.25)	75.9
138261-41-3	*Lmidacloprid*	2.1 (1.4, 2.8)	1.65 (0.66, 2.7)	77.6
115-29-7	*Endosulfan*	− 0.46 (− 1.09, 0.23)	− 0.99 (− 2.12, 0.1)	72.0
2921-88-2	*Chlorpyrifos*	− 0.03 (− 0.63, 0.66)	0.01 (− 0.76, 0.84)	92.0
206-44-0	*Fluoranthene*	0.9 (0.22, 1.58)	0.23 (− 0.04, 0.54)	50.3
62-53-3	*Aniline*	2.48 (2.21, 2.76)	2.71 (2.04, 3.42)	55.2
333-41-5	*Diazinon*	0.1 (− 0.72, 0.91)	0.04 (− 0.81, 0.87)	96.8

© The Author(s) 2021
www.kva.se/en

the lowest OVL score is the one for chemical *fluoranthene* (206-44-0) with OVL score 50.3% and followed by the SSD for chemical *aniline* (62-53-3) with OVL score 55.2%.

We used the 97.5% percentile and the 2.5% percentile as the upper and lower bounds, respectively, of the 1000 time bootstrapping when fitting LC50 values to SSDs. Mean values of predicted HC50 for all 10 chemicals are within the upper and lower bounds of experimental counterparts, regardless of the species and number of data points. Figure 3 shows the mean SSD curves for chemical *chlorpyrifos* (2921-88-2), as well as the upper and lower bounds according to 1,000 times of bootstrapping (in light colors) for both experimental (red) and predictive (blue) SSDs. The range of experimental and predictive SSD are mostly overlapped according to Fig. 3. The HC50 values of *chlorpyrifos* based on predictive SSD range from 0.23 to 4.57 µmol/L, and the experimental HC50 values range from 0.17 to 6.92 µmol/L. On both curves, fishes tend to be more sensitive to the exposure of *chlorpyrifos*. The species have the highest tolerance on the experimental SSD is *Sialis Lutaria* (Insects/Spiders) with LC50 61.66 umol/L, and on the predictive SSD is other water fleas with LC50 436.52 umol/L.

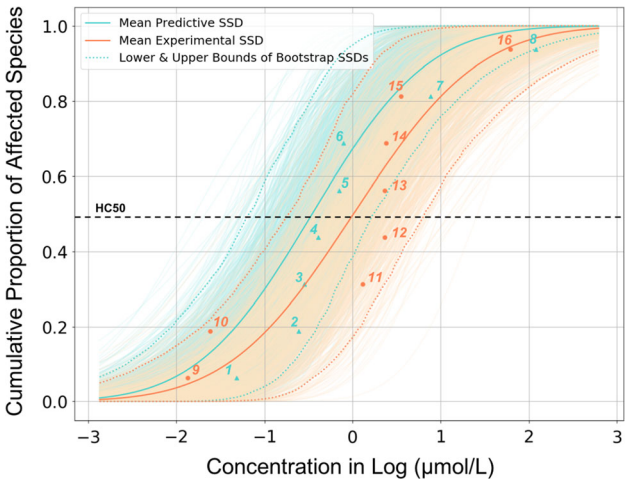

Fig. 3 The mean (solid blue line), upper (97.5%), and lower (2.5%) bounds (dash blue lines) of the predictive SSD, and the mean (solid red line), upper (97.5%), and lower (2.5%) bounds (dash red lines) of the experimental SSD for *chlorpyrifos*. Each data point and numbers on the curves represent a species for corresponding data group (predictive, blue, or experimental, red). 1: *Americamysis Bahia* (Crustaceans, shrimp); 2: *Cyprinodon Cariegatus* (Fish); 3: *Daphnia Magna* (Crustaceans, water flea); 4: *Lepomis Macrochirus* (Fish); 5: *Pimephales Promelas* (Fish); 6: *Oncorhynchus Mykiss* (Fish); 7: *Oryzias Latipes* (Fish); 8: Other water fleas (Crustaceans, water flea); 9: *Pungitius Pungitius* (Fish); 10: *Gasterosteus Aculeatus* (Fish); 11: *Neocaridina Denticulate* (Crustaceans, shrimp); 12: *Lctalurus Punctatus* (Fish); 13: *Aplexa Hypnorum* (Molluscs); 14: *Carassius Auratus* (Fish); 15: *Zilchiopsis Collastinensis* (Crustaceans); 16: *Sialis Lutaria* (Insects/Spiders)

Screening-level chemical ecotoxicity analysis

We applied our models to the organic chemicals in the ToX21 dataset. As a result, SSDs for 8424 organic chemicals were generated, among which 1240 fell into the AD for at least four ANN models (out of eight). Their predicted LC50 values, predictive HC5 and SSDs can be found in the supplementary information.

We mapped these chemicals with the ones registered as high-volume chemicals in European Chemicals Agency (ECHA) database (Registered substances-ECHA). We identified top 10 chemicals, of which the products of production volume and toxicity (inverse of HC5) are the highest (Table 3). If no additional data are available, these chemicals deserve attention given their high volume of usage and the high ecotoxicity, according to our screening-level analysis results (see IS for the full screening results). Among them 4,4′-diphenylmethane diisocyanate (101-68-8, MDI) shows the highest volume X toxicity value. MDI is widely used in the manufacture of polyurethane. MDI makes up about 60% of the global production of diisocyanate in 2000 (Randall and Lee 2002), and the U.S. demand for pure MDI was about 200 million pounds in 2008 (US EPA 2015b). Under certain circumstances, MDI can be released from adhesive and sealants in a format that isn't completed reacted, therefore cause potential occupational exposure (US EPA 2015b).

OVL testing

SSDs can be fitted by different statistical distributions. We used the coefficient of overlapping (OVL) method to compare the performance of different statistical distributions: log-normal, Weibull and Gamma, when fitting SSD curves. As the results show, the average OVL score of log-normal distribution was 0.82. More than 93% of the 8424 SSDs have OVL score higher than 0.60 on log-normal distribution. The comparison between log-normal, Weibull and Gamma distributions is presented in Fig. S3. The average OVL scores for Weibull and Gamma distributions were 0.71 and 0.67, respectively. Log-normal distribution was the one that has the highest average OVL score among all distributions we tested. The resulting standard log-normal SSD function shows the average logmean (µ) and average GSD (geometric standard deviation, σ) of 3.21 and 2.58, respectively, for the 8424 SSDs.

© The Author(s) 2021

Table 3 The top chemicals with the highest threatening among the registered chemicals in the ECHA database

Chemical name	Chemical CAS	HC5 umol/L	Number of chemicals in model AD	Production band in ECHA (thousand tonnes year^{-1})
4,4′-Diphenylmethane diisocyanate	101-68-8	0.19	4	100–1000
2-Ethylhexyl acrylate	103-11-7	3.1	4	100–1000
2-Ethylhexyl nitrate	27247-96-7	5.6	5	100–1000
Anthraquinone	84-65-1	0.13	4	1–10
tert-Butylperoxy 2-ethylhexyl carbonate	34443-12-4	0.19	4	1–10
Dodecanoic acid	143-07-7	2.6	4	10–100
2-Methyl-4′′′-(methylthio)-2-morpholinopropiophenone	71868-10-5	0.29	5	1–10
Methyl dodecanoate	111-82-0	3.3	4	10–100
6H-Dibenzo[c,e][1,2]oxaphosphinine 6-oxide	35948-25-5	0.37	5	1–10
1,3-Benzenedicarboxylic acid	121-91-5	57.1	4	100–1000

DISCUSSION

To our knowledge, our study is the first that consolidated aquatic ecotoxicity data from multiple data sources, and used them for large-scale SSD development using ANN. The resulting dataset, which is, to our best knowledge, the largest of its kind, is made freely available through our website. The predictive SSD, can be used for screening analysis to estimate the safety concentration of chemicals in aquatic ecosystem. Our results can also be used for as the reference ecotoxicity data in LCIA when better quality data are lacking, which is a acknowledged problem in LCA (Reap et al. 2008).

The performance of QSAR models developed in this study was promising and the results were comparable tothe existing QSARs in literature (Buccafusco et al. 1981; Devillers 2001; Toropov 2017). Our study demonstrates that advanced machine learning models can be used to improve the performance of QSAR models.

We collected extensive chemical toxicity dataset from reputable databases including ECOTOX and eChem. The datasets collected include over 2000 data points with comparable experimental conditions for reliable QSAR modeling. The QSAR models developed in this study using ANN can achieve $R^2 > 0.7$, when the training dataset is larger than 300 data points for a single species. With lesser data points, the ANN model must be reduced to simpler structure, which compromises its ability to estimate chemical toxicity. For future studies, even larger training dataset would be desirable to minimize the chance of missing out uncommon chemical structure–toxicity relationships.

We used the QSAR models to create predictive SSDs in this study. Our predictive SSDs showed a good performance compared with the SSDs created by experimental values. In Fig. 2, the predictive SSD showed lower toxicity at low concentration, and higher toxicity at high concentration. This is mainly due to the fact that the experimental SSD was created with less toxicity data points, compared with our predictive SSD. It is expected that the accuracy of SSD increases with more toxicity data points, and more diverse taxa (Posthuma et al. 2001, 2019). It is notable that the predictive SSD showed closer HC50 values compared with the experimental SSDs.

We recommend that the predictive model to be used as a supplementary to experimental data. Our models cannot replace SSDs derived from experimental toxicity data, as it has prediction uncertainties, and focusing on single stressor in this study. Furthermore, since ANN is a “black box” model, it should not be used to interpret the mechanism of ecotoxicity with molecular structure (Stojić et al. 2010). Given the current scarcity of experimental data and the high cost of developing them, however, we believe that our results demonstrate the potential for machine learning techniques to be used as a proxy for SSDs when better information is lacking. Furthermore, the rapidly growing number of chemicals in the lab and in the marketplace makes it challenging for experimental data alone to meet the needs for understanding the potential ecotoxicological impact of chemicals. We believe that our results can serve as a screening tool in the absence of experimental data to prioritize the candidates for further analysis. We view machine learning techniques not as a replacement of but as a complementary tool for experimental studies.

© The Author(s) 2021
www.kva.se/en

Experimental toxicity studies are also crucial for improving the quality of machine learning models to follow. High species sensitivity or low HC5 values in our SSD database should constitute a reason for in-depth testing, although predicted low species sensitivity or high HC5 values alone should not be taken as a proof that the chemical is safe.

We believe that the complementarity between predictive modeling and experimental studies can be further improved by standardizing the conditions for toxicity experiments and reporting. First of all, we cannot emphasize enough the importance of standard data exchange protocol on experimental conditions which is critical to accommodate machine readability of experimental data. Due to the poor documentation and the lack of standard data exchange protocol, extracting data on experimental conditions from existing literature and databases required painstaking effort. Second, consistency in experimental methods is crucial. We could not utilize many valuable experimental data points because one or more experimental conditions were not identical to the rest of the dataset. The variation in experimental conditions in e.g., duration of exposure, temperature, and chemical purity, significantly degraded the value of experimental toxicity data. A wider adoption of standard protocols for documenting and sharing toxicity testing results is urgently needed to tap into and maximize the value of experimental toxicity data for predictive modeling. While there are existing standards and guidelines including the OECD Test Guidelines, the Good Laboratory Practice (GLP) principles, and the Catalogue of Standard Toxicity Tests for Ecological Risk Assessment (Epa 1994), a universal applicable testing guideline is still lacking.

Machine learning techniques for ecotoxicological applications are still in a nascent stage, and there is considerable room for improvement in our study. Experimental data of better quality and quantity will improve the performances of the ANNs. Our models do not properly represent the toxicological impacts under multi-stressor conditions, because the experimental data used for training our model are all based on single chemical species. In fact, mixtures of chemicals are scarcely tested for ecotoxicity, and the development of protocols for mixture testing and reporting is in its infancy. In reality, however, ecosystem species are exposed to multiple chemicals at any given time. Although there are some studies that evaluate the concentration addition effect of chemical mixture (Hermens et al. 1984; Broderius and Kahl 1985; Wolf et al. 1988; Niederlehner et al. 1998), given that the number of possible combinations of chemical mixtures in both composition and proportion is extremely large, experimental data alone won't be able to meet the growing needs of data. Additional data and research are needed to adequately address the ecotoxicological impacts of multiple stressors, especially in the context of using SSDs.

Acknowledgements This publication was developed under Assistance Agreement No. 83557901 awarded by the US Environmental Protection Agency to the University of California, Santa Barbara. It has not been formally reviewed by EPA. The views expressed in this document are solely those of the authors and do not necessarily reflect those of the Agency. EPA does not endorse any products or commercial services mentioned in this publication.

Author contributions RS, AK, and SS designed the study. RS, MT, DL, and AC collected the ecotoxicity data. RS trained the machine learning models. YQ conducted the OVL testing in this study. RS and SS wrote this manuscript. All authors have reviewed the manuscript. The works in this manuscript done by RS and MT were conducted during their time at UC Santa Barbara.

Code availability All code in this study were developed in Python 2.7 (Anaconda2) on a machine with Ubuntu 16.04 LTS system. The code and the pretrained ANN models are available in GitHub: https://github.com/RunshengSong/QSAR_SSD_Toolbox. Instruction on how to reproduce the results in this study is also provided in the Github repository.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

REFERENCES

Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, and A. Davis, et al. 2016. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. *ArXiv160304467 Cs*.

Aldenberg, T., and E. Rorije. 2013. Species Sensitivity Distribution estimation from uncertain (QSAR-based) effects data. *Alternatives to Laboratory Animals* 41: 19–31.

Aldenberg, T., and W. Slob. 1993. Confidence limits for hazardous concentrations based on logistically distributed NOEC toxicity data. *Ecotoxicology and Environmental Safety* 25: 48–63.

Allison, T.C. 2016. Application of an artificial neural network to the prediction of OH radical reaction rate constants for evaluating global warming potential. *The Journal of Physical Chemistry B* 120: 1854–1863.

Andersen, M.E., and D. Krewski. 2009. Toxicity testing in the 21st century: Bringing the vision to life. *Toxicological Sciences* 107: 324–330.

Austin, T., M. Denoyelle, A. Chaudry, S. Stradling, and C. Eadsforth. 2015. European Chemicals Agency dossier submissions as an experimental data source: Refinement of a fish toxicity model for

predicting acute LC50 values. *Environmental Toxicology and Chemistry* 34: 369–378.
Bressler, D.W., J.B. Stribling, M.J. Paul, and M.B. Hicks. 2006. Stressor tolerance values for benthic macroinvertebrates in Mississippi. *Hydrobiologia* 573: 155–172.
Broderius, S., and M. Kahl. 1985. Acute toxicity of organic chemical mixtures to the fathead minnow. *Aquatic Toxicology* 6: 307–322.
Buccafusco, R.J., S.J. Ells, and G.A. LeBlanc. 1981. Acute toxicity of priority pollutants to bluegill (*Lepomis macrochirus*). *Bulletin of Environment Contamination and Toxicology* 26: 446–452.
Burden, F.R., and D.A. Winkler. 1999. Robust QSAR models using bayesian regularized neural networks. *Journal of Medicinal Chemistry* 42: 3183–3187.
Calow, P., and V.E. Forbes. 2003. *Peer reviewed: Does ecotoxicology inform ecological risk assessment?* Washington: ACS Publications.
Chen, B., R.P. Sheridan, V. Hornak, and J.H. Voigt. 2012. Comparison of random forest and Pipeline Pilot Naive Bayes in prospective QSAR predictions. *Journal of Chemical Information and Modeling* 52: 792–803.
Cherkasov, A., E.N. Muratov, D. Fourches, A. Varnek, I.I. Baskin, M. Cronin, J. Dearden, and P. Gramatica, et al. 2014. QSAR modeling: Where have you been? Where are you going to? *Journal of Medicinal Chemistry* 57: 4977–5010.
Chollet, F. *Keras*. GitHub, 2015.
Cunningham, V.L., S.P. Binks, and M.J. Olson. 2009. Human health risk assessment from the presence of human pharmaceuticals in the aquatic environment. *Regulatory Toxicology and Pharmacology* 53: 39–45.
De Wolf, W., J.H. Canton, J.W. Deneer, R.C.C. Wegman, and J.L.M. Hermens. 1988. Quantitative structure-activity relationships and mixture-toxicity studies of alcohols and chlorohydrocarbons: Reproducibility of effects on growth and reproduction of Daphnia magna. *Aquatic Toxicology* 12: 39–49.
Devillers, J. 2001. A general QSAR model for predicting the acute toxicity of pesticides to lepomis macrochirus. *SAR and QSAR in Environmental Research* 11: 397–417.
ECHA Publishes Official Statistics for the Last REACH Registration Deadline. https://www.chemsafetypro.com/Topics/EU/ECHA_Publishes_2018_REACH_Registration_Statistics.html.
eChemPortal-Home. https://www.echemportal.org/echemportal/propertysearch/page.action;jsessionid=D34DADB24143BE5071985CCDC085AA77?pageID=0.
ECOTOX | MED | US EPA. https://cfpub.epa.gov/ecotox/ecotox_home.cfm.
ESFA. https://dwh.efsa.europa.eu/bi/asp/Main.aspx.
forum, E.-U. E. protection agency R. assessment. 1998. *Guidelines for ecological risk assessment*. US Environmental protection agency.
Garner, K.L., S. Suh, H.S. Lenihan, and A.A. Keller. 2015. Species sensitivity distributions for engineered nanomaterials. *Environmental Science and Technology* 49: 5753–5759.
Guyon, I., and A. Elisseeff. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3: 1157–1182.
Hartley, M.J. 2002. Rationale and methods for conserving biodiversity in plantation forests. *Forest Ecology and Management* 155: 81–95.
Haupt, S.E., A. Pasini, and C. Marzban. 2008. *Artificial intelligence methods in the environmental sciences*. New York: Springer.
Hazardous Substances Data Bank (HSDB). https://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?HSDB.
Henderson, A.D., M.Z. Hauschild, D. de Meent, M.A. Huijbregts, H.F. Larsen, M. Margni, T.E. McKone, and J. Payet, et al. 2011. USEtox fate and ecotoxicity factors for comparative assessment of toxic emissions in life cycle analysis: Sensitivity to key chemical properties. *The International Journal of Life Cycle Assessment.* 16: 701–709.
Hermens, J., H. Canton, N. Steyger, and R. Wegman. 1984. Joint effects of a mixture of 14 chemicals on mortality and inhibition of reproduction of Daphnia magna. *Aquatic Toxicology* 5: 315–322.
Holmstrup, M., A.M. Bindesbøl, G.J. Oostingh, A. Duschl, V. Scheil, H.R. Köhler, S. Loureiro, and A.M. Soares, et al. 2010. Interactions between effects of environmental chemicals and natural stressors: A review. *Science Total and Environment* 408: 3746–3762.
Kaiser, K.L.E. 2003. The use of neural networks in QSARs for acute aquatic toxicological endpoints. *Journal of Molecular Structure* 622: 85–95.
Liu, Y., E. Racah, J. Correa, A. Khosrowshahi, D. Lavers, K. Kunkel, M. Wehner, and W. Collins, et al. 2016. Application of deep convolutional neural networks for detecting extreme weather in climate datasets. In *Int'l Conf. on Advances in Big Data Analytics | ABDA'16 |*, 81–88. http://worldcomp-proceedings.com/proc/p2016/ABD6152.pdf.
Lowry, G.V., B.P. Espinasse, A.R. Badireddy, C.J. Richardson, B.C. Reinsch, L.D. Bryant, A.J. Bone, and A. Deonarine, et al. 2012. Long-term transformation and fate of manufactured Ag nanoparticles in a simulated large scale freshwater emergent wetland. *Environmental Science & Technology* 46: 7027–7036.
Martin, T.M., and D.M. Young. 2001. Prediction of the acute toxicity (96-h LC50) of organic compounds to the fathead minnow (Pimephales promelas) using a group contribution method. *Chemical Research in Toxicology* 14: 1378–1385.
MacKinnon, D.P., C.M. Lockwood, and J. Williams. 2004. Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research* 39: 99–128.
Malaj, E., C. Peter, M. Grote, R. Kühne, C.P. Mondy, P. Usseglio-Polatera, W. Brack, and R.B. Schäfer, et al. 2014. Organic chemicals jeopardize the health of freshwater ecosystems on the continental scale. *Proceedings of the National Academy of Sciences* 111: 9549–9554.
Martin, O.V., J. Adams, A. Beasley, S. Belanger, R.L. Breton, T. Brock, V.A. Buonsante, and M. Galay Burgos, et al. 2019. Improving environmental risk assessments of chemicals: Steps towards evidence-based ecotoxicology. *Environment International* 128: 210–217.
Mayer, F.L., G.F. Krause, M.R. Ellersieck, G. Lee, and D.R. Buckler. 1994. Predicting chronic lethality of chemicals to fishes from acute toxicity test data: Concepts and linear regression analysis. *Environmental Toxicology and Chemistry* 13: 671–678.
Moriwaki, H., Y.-S. Tian, N. Kawashita, and T. Takagi. 2018. Mordred: A molecular descriptor calculator. *Journal of Cheminformatics* 10: 4.
Newman, M.C., D.R. Ownby, L.C. Mézin, D.C. Powell, T.R. Christensen, S.B. Lerberg, and B.A. Anderson. 2009. Applying species-sensitivity distributions in ecological risk assessment: Assumptions of distribution type and sufficient numbers of species. *Environmental Toxicology and Chemistry* 19: 508–515.
Niederlehner, B.R., J. Cairns, and E.P. Smith. 1998. Modeling acute and chronic toxicity of nonpolar narcotic chemicals and mixtures to Ceriodaphnia dubia. *Ecotoxicology and Environmental Safety* 39: 136–146.
Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, and P. Prettenhofer, et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12: 2825–2830.
Ping, Q., Y. Wang, and J. Wang. 2011. Aquatic predicted no-effect-concentration derivation for perfluorooctane sulfonic acid. *Environmental Toxicology and Chemistry* 30: 836–842.

© The Author(s) 2021
www.kva.se/en

ppoints function | R Documentation. https://www.rdocumentation.org/packages/stats/versions/3.5.2/topics/ppoints.
Posthuma, L., G.W. Suter II., and T.P. Traas. 2001. *Species sensitivity distributions in ecotoxicology*. Boca Raton: CRC Press.
Posthuma, L., J. van Gils, M.C. Zijp, D. van de Meent, and D. de Zwart. 2019. Species sensitivity distributions for use in environmental protection, assessment, and management of aquatic ecosystems for 12 386 chemicals. *Environmental Toxicology and Chemistry* 38: 905–917.
Qin, Y., and S. Suh. 2017. What distribution function do life cycle inventories follow? *International Journal of Life Cycle Assessment* 22: 1138–1145.
Raevsky, O.A., V.Y. Grigorev, E.E. Weber, and J.C. Dearden. 2008. Classification and quantification of the toxicity of chemicals to Guppy, Fathead Minnow and Rainbow Trout: Part 1 nonpolar narcosis mode of action. *QSAR Combinatorial Sciecne* 27: 1274–1281.
Raimondo, S., D.N. Vivian, C. Delos, and M.G. Barron. 2008. Protectiveness of species sensitivity distribution hazard concentrations for acute toxicity used in endangered species risk assessment. *Environmental Toxicology and Chemistry* 27: 2599–2607.
Randall, D., and S. Lee. 2002. *The polyurethanes book*. New York: Wiley.
rdkit: The official sources for the RDKit library. RDKit, 2017.
Reap, J., F. Roman, S. Duncan, and B. Bras. 2008. A survey of unresolved problems in life cycle assessment. *International Journal of Life Cycle Assessment* 13: 374.
Registered substances—ECHA. https://echa.europa.eu/information-on-chemicals/registered-substances.
Results of eco-toxicity tests data conducted by Ministry of the Environment in Japan. 2014.
Rosenbaum, R.K., T.M. Bachmann, L.S. Gold, M.A. Huijbregts, O. Jolliet, R. Juraske, A. Koehler, and H.F. Larsen, et al. 2008. USEtox—the UNEP-SETAC toxicity model: Recommended characterisation factors for human toxicity and freshwater ecotoxicity in life cycle impact assessment. *The International Journal of Life Cycle Assessment.* 13: 532–546.
Russom, C.L., S.P. Bradbury, S.J. Broderius, D.E. Hammermeister, and R.A. Drummond. 1997. Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow (Pimephales promelas). *Environmental Toxicology and Chemistry* 16: 948–967.
Saeys, Y., I. Inza, and P. Larrañaga. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23: 2507–2517.
Stojić, N., S. Erić, and I. Kuzmanovski. 2010. Prediction of toxicity and data exploratory analysis of estrogen-active endocrine disruptors using counter-propagation artificial neural networks. *Journal of Molecular Graphics and Modelling* 29: 450–460.
Sugumaran, V., V. Muralidharan, and K.I. Ramachandran. 2007. Feature selection using Decision Tree and classification through Proximal Support Vector Machine for fault diagnostics of roller bearing. *Mechanical Systems and Signal Processing* 21: 930–942.
The IUCN Red List of Threatened Species. http://www.iucnredlist.org/.
Todeschini, R., and V. Consonni. 2008. *Handbook of molecular descriptors*. New York: Wiley.
Toropov, A.A., A.P. Toropova, M. Marzo, J.L. Dorne, N. Georgiadis, and E. Benfenati. 2017. QSAR models for predicting acute toxicity of pesticides in rainbow trout using the CORAL software and EFSA's OpenFoodTox database. *Environmental Toxicology and Pharmacology* 53: 158–163.
US EPA. 1994. Catalogue of standard toxicity tests for ecological risk assessment. 2: 4.
US EPA, O. 2015a. Toxicology Testing in the 21st Century (Tox21). *US EPA* https://www.epa.gov/chemical-research/toxicology-testing-21st-century-tox21.
US EPA, O. 2015b. Methylene Diphenyl Diisocyanate (MDI) And Related Compounds. *US EPA* https://www.epa.gov/assessing-and-managing-chemicals-under-tsca/methylene-diphenyl-diisocyanate-mdi-and-related.
Vörösmarty, C.J., P.B. McIntyre, M.O. Gessner, D. Dudgeon, A. Prusevich, P. Green, S. Glidden, and S.E. Bunn, et al. 2010. Global threats to human water security and river biodiversity. *Nature* 467: 555–561.
Wheeler, J.R., E.P.M. Grist, K.M.Y. Leung, D. Morritt, and M. Crane. 2002. Species sensitivity distributions: Data and model choice. *Marine Pollution Bulletin* 45: 192–202.
Wolansky, M.J., and J.A. Harrill. 2008. Neurobehavioral toxicology of pyrethroid insecticides in adult animals: A critical review. *Neurotoxicology and Teratology* 30: 55–78.
Worth, A.P., and M.T. Cronin. 2003. The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects. *Journal of Molecular Structure* 622: 97–111.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

AUTHOR BIOGRAPHIES

Runsheng Song is a doctoral candidate at Bren School of Environmental Science & Management, University of California, Santa Barbara. His research focuses on applying machine learning on life cycle assessment, and environmental impact assessment.
Address: Bren School of Environmental Science and Management, University of California, Santa Barbara, Santa Barbara, CA 98121, USA.
e-mail: runsheng@ucsb.edu; srs90324@gmail.com

Dingsheng Li is an assistant professor at the School of Community Health Sciences at the University of Nevada, Reno. His research focuses human exposure to chemicals and resulting health risks with various models.
Address: University of Nevada, Reno, 1664 N Virginia St, Reno, NV 89557, USA.
e-mail: dingshengl@unr.edu

Alexander Chang is a graduate student at the Emory Rollins School of Public Health, Emory University.
Address: Emory Rollins School of Public Health, 1518 Clifton Rd, Atlanta, GA 30322, USA.
e-mail: theamazingchang@gmail.com

Mengya Tao is a doctoral candidate at Bren School of Environmental Science & Management, University of California, Santa Barbara. Her research focuses on fate and transport model in chemical exposure.
Address: Bren School of Environmental Science and Management, University of California, Santa Barbara, Santa Barbara, CA 98121, USA.
e-mail: mengya120@gmail.com

Yuwei Qin is a doctoral candidate at Bren School of Environmental Science & Management, University of California, Santa Barbara. Her research focuses on uncertainties in life cycle assessment.
Address: Bren School of Environmental Science and Management, University of California, Santa Barbara, Santa Barbara, CA 98121, USA.
e-mail: rainsmile333@gmail.com

© The Author(s) 2021
www.kva.se/en

Arturo A. Keller is a professor at Bren School of Environmental Science & Management, University of California, Santa Barbara. His research focuses on the sustainable use of chemicals and materials in our modern society.
Address: Bren School of Environmental Science and Management, University of California, Santa Barbara, Santa Barbara, CA 98121, USA.
e-mail: keller@bren.ucsb.edu

Sangwon Suh (✉) is a professor at Bren School of Environmental Science & Management, University of California, Santa Barbara. His research focuses on the sustainability of the human-nature complexity through the understanding of materials and energy exchanges between them.
Address: Bren School of Environmental Science and Management, University of California, Santa Barbara, Santa Barbara, CA 98121, USA.
e-mail: suh@bren.ucsb.edu

© The Author(s) 2021
www.kva.se/en