Environment International 135 (2020) 105393

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Estimate ecotoxicity characterization factors for chemicals in life cycle assessment using machine learning models

Ping Hou[a,b], Olivier Jolliet[c], Ji Zhu[d], Ming Xu[a,e,*]

[a] *School for Environment and Sustainability, University of Michigan, Ann Arbor, MI, USA*
[b] *Michigan Institute for Computational Discovery & Engineering, University of Michigan, Ann Arbor, MI, USA*
[c] *Environmental Health Sciences, School of Public Heath, University of Michigan, Ann Arbor, MI, USA*
[d] *Department of Statistics, University of Michigan, Ann Arbor, MI, USA*
[e] *Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, MI, USA*

ARTICLE INFO

Handling Editor: Zhen Jason He

Keywords:
Life cycle assessment
Ecotoxicity
Hazardous concentration
Characterization factors
Machine learning
Quantitative structure-activity relationship (QSAR)

ABSTRACT

In life cycle assessment, characterization factors are used to convert the amount of the chemicals and other pollutants generated in a product's life cycle to the standard unit of an impact category, such as ecotoxicity. However, as a widely used impact assessment method, USEtox (version 2.11) only has ecotoxicity characterization factors for a small portion of chemicals due to the lack of laboratory experiment data. Here we develop machine learning models to estimate ecotoxicity hazardous concentrations 50% (HC_{50}) in USEtox to calculate characterization factors for chemicals based on their physical-chemical properties in EPA's CompTox Chemical Dashborad and the classification of their mode of action. The model is validated by ten randomly selected test sets that are not used for training. The results show that the random forest model has the best predictive performance. The average root mean squared error of the estimated HC_{50} on the test sets is 0.761. The average coefficient of determination (R^2) on the test set is 0.630, meaning 63% of the variability of HC_{50} in USEtox can be explained by the predicted HC_{50} from the random forest model. Our model outperforms a traditional quantitative structure-activity relationship (QSAR) model (ECOSAR) and linear regression models. We also provide estimates of missing ecotoxicity characterization factors for 552 chemicals in USEtox using the validated random forest model.

1. Introduction

Life cycle assessment (LCA) is a widely used analytical tool that examines the environmental impacts of a product along its whole life cycle (ISO, 2006; Rebitzer et al., 2004). A wide range of environmental impacts in product life cycles are associated with the usage and release of chemicals. For example, the life cycle of food generally includes production, harvesting, processing, packing, transport, marketing, consumption, and waste treatment and disposal. In every step of this life cycle, chemicals are used for food processing and preservation. The release of and exposure to those chemicals can impact ecosystem and human health. Quantifying the potential environmental impacts of chemicals is thus critical for LCA.

In LCA, the step of quantifying the impacts of chemicals and other pollutants is called life cycle impact assessment (LCIA). In this step, the amount of the chemicals and other pollutants generated in a product's life cycle are converted to the common unit of an impact category by characterization factors (ISO, 2006). Current practice of LCIA is significantly constrained by limited data on characterization factors of chemicals (Fantke et al., 2018a; Fantke et al., 2018b; Saouter et al., 2017a; Saouter et al., 2017b). For example, as the LCIA model endorsed by the Life Cycle Initiative – hosted by UN environment for toxicity impacts calculation, USEtox provides characterization factors for human and ecotoxicology impacts of chemicals (Frischknecht and Jolliet, 2016; Rosenbaum et al., 2008). Despite its wide applications in LCA, USEtox only offers characterization factors for approximately 3000 chemicals. As a comparison, the U.S. Environmental Protection Agency (EPA) has more than 85,000 chemicals listed under the Toxic Substances Control Act (Hinds and Weller, 2016). Even for the limited number of chemicals covered in USEtox, 19% and 67% of them miss ecotoxicity characterization factors and human toxicity characterization factors, respectively. Here, we mainly focus on estimating ecotoxicity characterization factors.

The lack of ecotoxicity and human toxicity characterization factors is essentially due to the lack of toxicity testing data, either from *in vivo* (i.e., within the living) animal tests or *in vitro* (i.e., within the glass)

[*] Corresponding author at: 440 Church St., Ann Arbor, MI 48109-1041, USA.
E-mail address: mingxu@umich.edu (M. Xu).

https://doi.org/10.1016/j.envint.2019.105393
Received 29 April 2019; Received in revised form 3 December 2019; Accepted 3 December 2019
Available online 18 December 2019

tests. Since laboratory tests are time-consuming and expensive, *in silico* toxicology (i.e., computational toxicology) are proposed to use computational methods to predict toxicity, prioritize chemicals, and guide drug design (Deeb and Goodarzi, 2012). *In silico* methods include structural alerts and rule-based models, read across (RA), dose-response and time-response models, pharmacokinetic and pharmacodynamic models, uncertainty factors models, and quantitative structure-activity relationship (QSAR) models. A detailed review of these methods can be found in (Raies and Bajic, 2016). Following the authors' guideline for choosing a method, RA and QSAR are the two most suitable methods for estimating characterization factors in LCIA based on the criteria that the endpoint (i.e., characterization factors) is quantitative and chemical properties data are available as continuous features (Fig. S1). RA predicts the unknown toxicity of a chemical using similar chemicals with known toxicity (Raies and Bajic, 2016). K nearest neighbor (KNN) in machine learning is an automatic algorithm to perform RA (Benfenati et al., 2019) and has been used to predict toxicity (Chavan et al., 2015).

QSAR refers to a broad area of inquiry on the relationship between chemical structures and biological activities of chemicals (Nantasenamat et al., 2009). It relates a set of predictor variables (e.g., number of carbon atoms in the molecule) to the response variable (e.g., bioaccumulation). The discovered relationship can then be used to predict the activities of new or untested chemicals. QSAR tools have been developed to predict aquatic ecotoxicity of chemicals, such as Ecological Structure Activity Relationships (ECOSAR) (Mayo-Bean et al., 2011), Kashinhou Tool for Ecotoxicity (KATE) (Furuhama et al., 2010), Toxicity Estimation Software Tool (TEST) (Martin, 2016), ADMET (absorption, distribution, metabolism, excretion, and toxicity) (Predictor, 2015), and Computer-Aided Discovery and REdesign for Aquatic Toxicity (CADRE-AT) (Kostal et al., 2015; Voutchkova et al., 2011; Voutchkova-Kostal et al., 2012). A recent review compares these tools in terms of acute aquatic toxicity experimental thresholds (LC_{50}) regarding OECD accepted freshwater fish species and any of 4 accepted time points (48 h, 72 h, 96 h, 120 h). Based on the root mean squared error (*RMSE*) on an external validation dataset, these tools are ranked from best to worst, i.e., ECOSAR (1.29) < TEST (1.32) < KATE (1.35) < ADMET (1.60). The *RMSE* cannot be calculated for CADRE-AT because it does not provide numeric LC_{50} estimates but only a regulatory category assignment, and it is currently not available to the public (Melnikov et al., 2016).

Traditional QSAR models, including ECOSAR, are mostly based on linear models (i.e., ordinary least squares regression), or advanced linear models (e.g., partial least square regression, principal component regression). Recent focus has been shifted towards more complex and nonlinear approaches, such as machine learning models (Tropsha, 2010). The applications of machine learning in environmental toxicology are still limited (Miller et al., 2018). A few studies develop machine learning models for predicting ecotoxicity of chemicals, such as decision tree (Singh et al., 2014) or discriminate analysis (Kostal et al., 2015; Voutchkova et al., 2011). Li et al (2017) compares six machine learning models to classify chemicals into categories based on 96 h LC_{50} to fish. Results show support vector machine and neural network give best results in acute toxicity estimation and they have a higher classification accuracy than the ECOSAR model. A few studies also successfully apply machine learning to predict ecotoxicity of chemicals (Saçan et al., 2015; Shoji, 2005; Tuulaikhuu et al., 2017). Most QSAR models predict toxicity regrading a certain species.

In life cycle assessment, ecotoxicity characterization factors are recommended to be based on effect data of at least three different species covering at least three different trophic levels to ensure a minimum variability of biological responses (Fantke et al., 2017). Therefore, hazardous concentration 50% (HC_{50}) aggregated by at least three different species are required to calculate effect factors (EF), which is an integral part of ecotoxicity characterization factors. EF expresses the potentially affected fraction of species due to concentration changes (Fantke et al., 2017). HC_{50} of many chemicals are currently unknown; thus, their characterization factors are missing as well.

QSAR also plays an important role in chemical risk assessment for screening toxicity of new chemicals, prioritizing chemicals for further assessment, and for chemical management and regularization (Mackay et al., 2003; Pradeep et al., 2016; Sangion and Gramatica, 2016; Sun et al., 2012). Nevertheless, the applications of machine learning models are suggested to be cautious for regularization purpose due to lack of transparency (ECHA, 2016). In chemical risk assessment, most studies use QSAR to classify chemicals into different levels of concern for screening and prioritization of chemicals for regulation. However, in life cycle assessment, numeric toxicity values are required to quantitatively evaluate the toxicity of chemicals along a product's life cycle.

Here we use machine learning models to predict hazardous concentration 50% (HC_{50}) of chemicals for calculating their ecotoxicity characterization factors in LCA.

There are increasing efforts to use machine learning models to estimate LCA data, including unit process data, characterization factors, and characterized LCA results. Most of the studies focus on estimating characterized LCA results for various products, such as chemicals (Song et al., 2017; Wernet et al., 2008), consumer products (Chen and Liau, 2001; Park et al., 2001; Park and Seo, 2003; Seo and Kim, 2006; Seo et al., 2005; Wisthoff et al., 2016), buildings (Azari et al., 2016), and others (Khoshnevisan et al., 2013; Nabavi-Pelesaraei et al., 2017; Ozbilen et al., 2013). Characterized LCA results are integrated results from unit process data and characterization factors. Some studies have estimated unit process data and their parameters (Chiang et al., 2011; Chiang and Roy, 2012; Piao et al., 2016; Yin et al., 2017). A few studies use machine learning models to estimate parameters for calculating characterization factors, such as removal rates (Sala et al., 2011), fate factors, and intake fractions (Birkved and Heijungs, 2011; Marvuglia et al., 2015a; Marvuglia et al., 2014; Marvuglia et al., 2015b). However, these parameters can be calculated by the multimedia fate and exposure model in USEtox. The lack of ecotoxicity characterization factors in USEtox is actually due to the lack of hazardous concentration 50% (HC_{50}). To the best of our knowledge, no study has been done using machine learning models to estimate HC_{50} of chemicals, which can then be used to calculate ecotoxicity characterization factors for LCIA.

In this study, we aim to estimate missing HC_{50} and ecotoxicity characterization factors for chemicals in USEtox entirely based on a data-driven approach. Relying on physical-chemical properties of chemicals and their classification of mode of action, we build machine learning models to estimate their hazardous concentration 50% (HC_{50}) values, which are later used to determine their respective ecotoxicity characterization factors. To evaluate the performance of our models, we compare their performance with those of traditional QSAR models, including ECOSAR model, ordinary least squares regression (OLS), partial least squares regression (PLS), and principal component regression (PCR).

2. Materials and methods

In USEtox, the ecotoxicity characterization factor of a chemical (CF_{eco} [$PAF{\cdot}m^3{\cdot}d{\cdot}kg^{-1}$]) is calculated by:

$$CF_{eco} = FF \times XF \times EF \tag{1}$$

whereFF [d] is the fate factor calculated by the multimedia transport and transformation model in USEtox, which characterizes the distribution of emitted contaminants among different compartments (e.g., urban air, agricultural soil, freshwater). XP [–] is the ecotoxicity exposure factor, calculated as the fraction of a chemical dissolved in freshwater. EF [$PAF * m^3{\cdot}kg^{-1}$] is the effect factor that relates ecosystem exposures and dissolved masses in the freshwater ecosystem to a measure of the potentially affected fraction (PAF, dimensionless) of exposed species.

As we mentioned above, missing CF_{eco} is generally due to missing HC_{50}, which are used to calculate EF:

$$\mathrm{EF} = \frac{0.5}{\mathrm{HC}_{50}} \tag{2}$$

where HC_{50} [kg·m^{-3}] is defined as the hazardous concentration of a chemical at which 50% of the freshwater species are exposed above their EC_{50}. The EC_{50} is the effective concentration at which 50% of a population displays an effect (e.g. mortality) in a laboratory test or a field test. The HC_{50} is calculated as the geometric mean of the EC_{50} of different species, for which the considered chemical was tested. Because LCA mainly deals with chronic effects, HC_{50} values are first derived from chronic EC_{50} in USEtox. Whenever chronic EC_{50} data are unavailable, best-estimate acute to chronic ratios are developed to extrapolate acute EC_{50} to chronic EC_{50} (Hauschild et al., 2008; Henderson et al., 2011; Rosenbaum et al., 2008).

2.1. Exploratory analysis and data filtering

USEtox version 2.11 contains 3077 organic chemicals and 27 inorganic metals. Here, we only consider the organic chemicals. Among them, 2499 out of 3077 organic chemicals have HC_{50} data, in which 283 HC_{50} values are derived from chronic EC_{50} and others are extrapolated from acute EC_{50} (Fantke et al., 2017). The physical-chemical characteristics of these chemicals are acquired from OPERA (Mansouri et al., 2018), which is available at EPA's CompTox Chemistry Dashboard (https://comptox.epa.gov). OPERA has physical-chemical characteristics data for 2307 chemicals among the 2499 chemicals. The characteristics include:

1. MW: molecular weight
2. AOH: atmospheric hydroxylation rate
3. BCF: bioconcentration factor
4. BioHL: biodegradation half-life
5. BP: boiling point
6. HL: henry's law constant
7. KM: fish biotransformation half-life
8. KOA: octanol/air partition coefficient
9. LogP (a.k.a., Kow): octanol-water partition coefficient
10. MP: melting point
11. KOC: soil adsorption coefficient
12. VP: vapor pressure
13. WS: water solubility

In addition, the mode of action (MoA) is recognized as an important determinant of chemical toxicity (Kienzler et al., 2017). Verhaar scheme (Verhaar et al., 1992) is developed to classify chemicals into five categories of modes of action based on the presence or absence of certain chemical structures and moieties. The five categories are class 1 (inert chemicals), class 2 (less inert chemicals), class 3 (reactive chemicals), class 4 (chemicals acting by a specific mechanism), and class 5 (unclassifiable chemicals). From class 1 to class 4, the toxicity of chemicals is generally increasing. We use a software program ToxTree (Patlewicz et al., 2008), which has a module of Verhaar scheme, to determine the classification of MoA of chemicals in USEtox. We treat them as a continuous variable (values = 1, 2, 3, 4, or 5) and add it as an additional variable besides the physical-chemical characteristics.

As a result, the collected data set has 2307 observations with one response variable (i.e., HC_{50}) and 14 predictor variables (i.e., 13 physical-chemical properties and the classification of MoA). We conduct an exploratory analysis to understand the basic characteristics and distributions of the HC_{50} values and physical-chemical characteristics of the corresponding chemicals. Data from different sources often have different levels of uncertainties. High uncertainty observations may distort the outcome and accuracy of a regression. Cook's distance measures the influence of each observation on the fitted response values and is used to remove outliers in the data (Cook, 1977; Kutner et al., 2005).

2.2. Machine learning models

In this study, we take a known set of HC_{50} of chemicals as output and their physical-chemical properties as input to train machine learning models. Since the response variable (i.e., HC_{50}) is continuous, this is a regression problem. We consider six commonly used regression machine learning models: k nearest neighbors (KNN), support vector machine (SVM), neural networks (NN), random forests (RF), adaptive boosting (AdaBoost), and gradient boosting machine (GBM).

KNN assumes that objects near each other are similar and finds the nearest neighbors by using a distance metric (e.g., Euclidean distance) based on input variables (Altman, 1992). KNN computes the average of the output of the k nearest neighbors as the estimation of the output of the query data point. When calculating the average, the data points in the neighbor can be weighted equally or by the inverse of their distance, in which case closer neighbors have a higher influence than further neighbors. K is a user-specified constant, the best choice of which depends upon the data.

SVM uses a nonlinear kernel function to map data into high-dimensional feature spaces and find a decision boundary (a.k.a., hyperplane) that minimizes the errors (Cortes and Vapnik, 1995). Radial basis function (RBF) and polynomial are accessible kernel functions. Gamma is the coefficient of the kernel function, with higher gamma, only points close to the hyperplane are considered in the calculation. To prevent overfitting, the penalty parameter of the error term (i.e., C) controls the tradeoff between smooth decision boundary and low error of the training points.

NN simulates the way biological nervous systems (e.g., brain) process information (Haykin and Network, 2004). A neural network is composed of "neurons" organized in multiple layers. The predictors (or inputs) form the first layer, and the responses (or outputs) form the last layer. Between the input layer and the output layer are one or more hidden layers interconnected with each other by hidden neurons. The value of each neuron in the hidden layers and output layer are calculated upon the outputs from neurons in the previous layer using an activation function. To minimize the errors between predicted output and the desired output, different optimization algorithms are available to train neural networks. Here, we consider only one hidden layer because a simple neural network model is parsimonious for easier interpretation and shorter training time.

Random forests (RF) generate an ensemble of decision trees (Svetnik et al., 2003). Each tree is a predictive model that uses a random subset of the training data and limits the number of variables used at each split. Random forests average many noisy but approximately unbiased trees, and hence reduce the prediction variance and obtain a more accurate and stable prediction.

Adaptive boosting (AdaBoost) is also an ensemble learning algorithm that creates a strong learner by iteratively adding weak learners (Schapire, 2013). AdaBoost is adaptive in that, with each iteration, a new weak learner is built accommodating fine-tuned weights that give priority to high error data in prior iterations. The result is a strong learner that has higher accuracy than the weak learners. AdaBoost commonly uses decision trees as the base learners.

Gradient boosting machine (GBM) is another ensemble learning algorithm that uses the boosting technique to combine several weak learners to form a strong learner (Friedman, 2001). GBM also uses decision trees as base learners. But different from AdaBoost, each subsequent tree in series is built on the errors identified by gradients instead of high weight data points.

2.3. Model development

To develop machine learning models with good performance on estimating HC_{50} of new chemicals, we split the data set into training and validation sets, and a test set. Five-fold cross-validation on training and validation sets are used to train and select the best parameters for each machine learning model, and the test set is then used to evaluate the predictive performance of the selected model. Specifically, we use coefficient of determination (R^2) and root mean squared error ($RMSE$) as model evaluation metrics. R^2measures how much of the variability of the HC_{50} values in USEtox can be explained by the predicted HC_{50} values. R^2 is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

where y_i is HC_{50} value of chemical i in USEtox, $\hat{y}_i$ is the predicated HC_{50} of chemical i, and $\bar{y}$ is the average of HC_{50} of chemicals in USEtox.

Since R^2 can be misleading sometimes, we also report root mean squared error ($RMSE$), which is a more helpful indicator of a model's usefulness (Alexander et al., 2015). $RMSE$ measures how much error there is between the predicted HC_{50} and the HC_{50} in USEtox. $RMSE$ is defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{4}$$

We use $RMSE$ as the criteria for model selection and also report R^2 at the same time. In the following model development process description, only $RMSE$ is mentioned as an example.

As shown in Fig. 1, we build our machine learning models in the following four steps:

(1) Data splitting. We randomly split the data into 70% for training and validation, and 30% for testing. Some machine learning models are sensitive to the scale of data, i.e., KNN, SVM, and NN. We normalize each input variable and response variable (i.e., HC_{50}) of the training set and validation set in the range of [-1, 1] via min-max normalization. Test set are also normalized based on the minimum and maximum values of the training set and validation set. On the other hand, tree-based methods (i.e., RF, AdaBoost, and GBM) do not require data preprocessing.

(2) Model selection using cross-validation. Each machine learning method has some parameters to tune. For the key parameters (Table 1), we use five-fold cross-validation (Fig. 2) on the training and validation sets to choose the best parameter combination. For each combination, the data are first randomly divided into five folds. Each time, four folds are used to train the model and the one fold left is used to validate the model and calculate the model performance. This process is repeated for five times with each fold being used as the validation set. The results are five $RMSE$ of the training sets and ten $RMSE$ of the validation sets, the average of which are respectively defined as the training $RMSE$ and the cross-validated $RMSE$. The parameters of the model with the smallest cross-validated$RMSE$ is then chosen as the best parameters. Other parameters are chosen based on trial and error (e.g., the kernel function of SVM and the optimizer of the NN) or commonly used default values (e.g., the minimum sample in a leaf of RF is set as 1 to allow the trees to fully grow).

(3) Model testing. The model is trained again with the whole training and validation sets and the best parameters identified from the previous step. The trained model is applied to the test set. The test $RMSE$ then evaluates the performance of the model. The lower the test $RMSE$, the better the model is at predicting the HC_{50} values for new chemicals.
In order to get an unbiased evaluation of the model performance, we repeat step 1 to 3 ten times with ten different splits of data to calculate the average test $RMSE$. Note that ten times repetition also generates ten best parameter combinations.

(4) Prediction. The best machine learning model with the tuned parameters is trained again using all available data (i.e., training and validation sets and test set). The trained model is then used to predict HC_{50} for chemicals of which HC_{50} are missing, and then calculate their CF_{eco} using FF and XF in USEtox.

2.4. Model performance comparison

We first compare the performance of our machine learning models with the performance of the ECOSAR model, which is proven to be the best model among several existing aquatic ecotoxicity QSAR tools (Melnikov et al., 2016). The predicted toxicity is either calculated by multiplying the baseline toxicity (Rand, 1995) by a toxicity reduction factor or by simple linear models with only one predictor variable (e.g., number of carbons in the molecular formula), depending on the chemical class. ECOSAR predicts EC_{50} for each chemical regarding different species. We aggregate them into HC_{50} following the guideline by USEtox (Fantke et al., 2017). The performance of ECOSAR is measured by comparing the HC_{50} in USEtox with the HC_{50} based on EC_{50} from ECOSAR. All chemicals in USEtox (after data filtering) have predictions in ECOSAR are used to evaluate the performance of the ECOSAR model.

We also compare our machine learning models with four linear

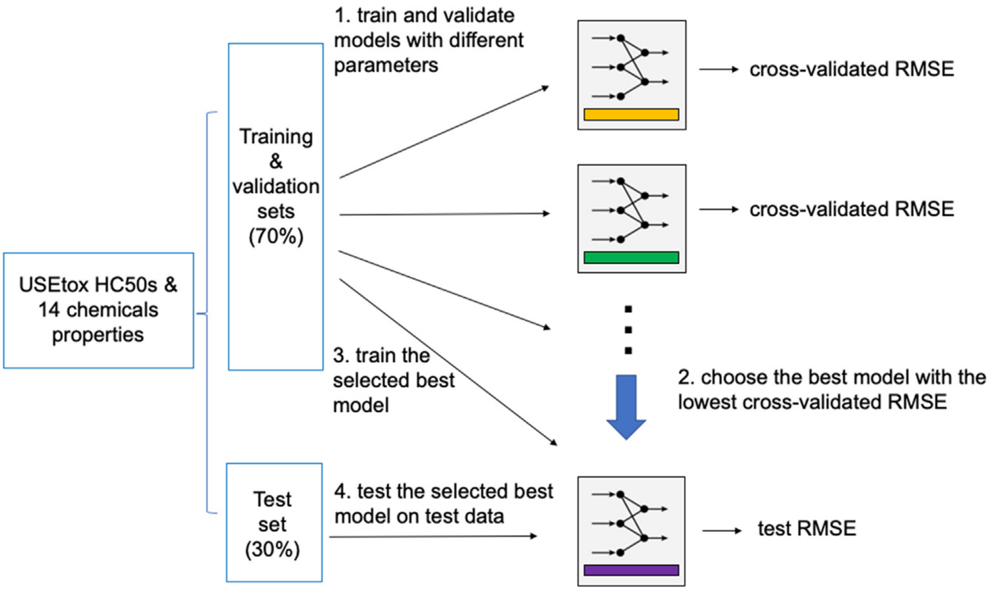

USEtox HC50s & 14 chemicals properties
Training & validation sets (70%)
Test set (30%)
1. train and validate models with different parameters
cross-validated RMSE
cross-validated RMSE
2. choose the best model with the lowest cross-validated RMSE
3. train the selected best model
4. test the selected best model on test data
test RMSE

Fig. 1. Steps for building machine learning models.

Table 1
Key parameters in the machine learning models tuned by cross validation.

Model	Key parameters tuned by cross validation	Options
K nearest neighbor (KNN)	n_neighbors (number of nearest neighbors)	5, 10, 15, 20, 25
	Weighting	Uniform, distance (i.e., Euclidean)
Support vector machine (SVM)	Gamma	0.001, 0.01, 0.1, 1
	C	0.01, 0.1, 1, 10
Neural network (NN)	n_neurons (number of hidden neurons)	5, 10, 15, 20, 25
	Optimizer	rmsprop, adam, sgd, adagrad, adadelta, adamax, nadam
Random forest (RF)	n_etimators (number of trees)	100, 500, 1000, 1500, 2000
	max_features (maximum number of features)	3, 4, 7, 14
Adaptive boosting (AdaBoost)	n_etimators (number of trees	100, 500, 1000, 1500, 2000
	Learning rate	0.001, 0.005, 0.01, 0.05, 0.1
Gradient boosting machine (GBM)	n_etimators (number of trees	100, 500, 1000, 1500, 2000
	Learning rate	0.001, 0.005, 0.01, 0.05, 0.1

regression models: including ordinary least squares (OLS) regression (with and without interaction terms), partial least squares (PLS) regression, and principle component regression (PCR) that we build based on the same data set. OLS chooses the parameters in a linear function by minimizing the sum of the squares of the difference between the observed and the predicted response variable by the linear function. The OLS model with interaction terms explores the interactions of any two variables, in order to capture their interdependence. PLS is a linear regression model by projecting the predictor variables and the response variable to a new space (Gomes et al., 2014). PCR first reduces the predictor variables using principal component analysis then uses the reduced variables in an OLS regression fit (Maitra and Yan, 2008). We use the same ten splits of data to evaluate the four models. We train the model using the training and validation set each time and test the model with the test set. For the PLS and PCR, we use 5-fold cross-validation on the training and validation set to choose the best number of components. This process is repeated on the ten splits of data to calculate the average test *RMSE* and test R^2 for each linear model.

2.5. Variable importance

Since not all the input variables are equally important in the model, we evaluate the relative importance of the variables and identify the important input variables for our machine learning models. Tree-based models (i.e., RF, AdaBoost, and GBM) can easily measure the relative importance of each feature on the prediction. At each split in each tree, the improvement on split-criterion is attributed to the splitting variable as its importance and is accumulated over all the trees in the model respectively for each variable. For regression problems, the split-criterion is the decrease of errors, which is calculated by the reduction in the sum of squared errors whenever a variable is chosen to split. Feature importance is then calculated as the decrease of errors weighted by the probability of reaching that node, which is the number of samples reaching the node divided by the total number of samples.

For other machine learning models (i.e., KNN, SVM, and NN), one way to test the importance of a variable is to shuffle or permutate the variable and see its impact on the model performance (Strobl et al., 2008). The procedure is first to get a benchmark test *RMSE* by training the model once and then predict multiple times while randomizing each variable in the test set. The difference of the benchmark test *RMSE* and the test *RMSE* after permuting the variable, meaning with and without the help of this variable, is used as an importance measure (i.e., permutation importance). If the test *RMSE* after randomizing a variable is lower than the benchmark test *RMSE*, it is an important variable. On the other hand, if nothing changes or the test *RMSE* is higher than the benchmark, it is less important in the model. In our study, we randomize 500 times and get an average test *RMSE* for each variable and compare it with the benchmark test *RMSE*.

The standardized coefficients of OLS models also imply the importance of variables. Standardized coefficients refer to how many standard deviations a dependent variable will change per standard deviation increase in the predictor variable. Similarly, the standard coefficients of PLS models have also been used in variable selection (Palermo et al., 2009). As for PCR models, the importance of variables can be calculated by their contributions to components multiplied by the standardized coefficients of the components.

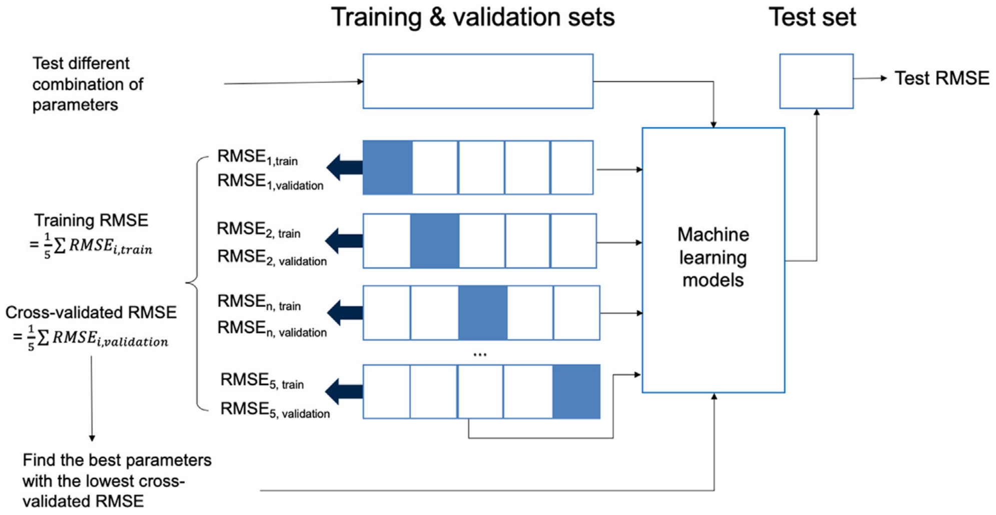


Fig. 2. Cross-validation.

Table 2
Performance of the random forests (RF) on ten different splits of data.

Splits of data	Five-fold cross validation				Best parameters	
	Training		Cross-validated		Number of estimators	Maximum features
	R^2	*RMSE*	R^2	*RMSE*		
1	0.948	0.278	0.607	0.763	2000	3
2	0.949	0.283	0.619	0.767	1000	4
3	0.946	0.286	0.621	0.757	500	4
4	0.945	0.288	0.606	0.773	1500	3
5	0.948	0.282	0.617	0.763	1000	4
6	0.947	0.286	0.624	0.762	500	4
7	0.950	0.279	0.629	0.756	2000	3
8	0.947	0.286	0.619	0.768	2000	3
9	0.949	0.286	0.628	0.769	500	3
10	0.947	0.284	0.605	0.771	1000	4
Average (standard deviation)	0.948 (0.001)	0.284 (0.003)	0.617 (0.009)	0.765 (0.006)	NA	NA

2.6. Uncertainty of the estimated HC_{50}

The outcome by the trained model is a point prediction with uncertainties that come from the model itself and noise in the input data. We evaluate the uncertainty of the estimated values by giving a confidence interval of the estimation. A confidence interval provides a range of model results and a probability that the model results will fall between the ranges when making predictions on new data. A robust way to determine confidence intervals for machine learning models is to use bootstrap, a common technique that can be used to derive empirical confidence intervals (DiCiccio and Efron, 1996). The basic idea is to resample the original data many times and use them to train the model respectively, and then we can have many predictions that can produce reasonable approximate confidence intervals of the predicted values. We train the model 100 times based on resampled training data and provide bootstrap confidence intervals of the predicted values for the chemicals without HC_{50} values in USEtox.

2.7. Application domain

An application domain is required in a QSAR study to express the scope and limitations of a model to specify the range of chemical properties for which the model is applicable (Netzeva et al., 2005). We use a distance-based method to define the application domain of our model. We first calculate a centroid of all the chemicals in the available data set based on their input variables. Then we calculate the distance to the centroid from each chemical. Finally, we identify a distance in which 90% of the chemicals are enclosed. This distance enclosed area is defined as the application domain of the model. When applying this model on a new chemical, one first calculates the distance of the new chemical to the centroid and determines whether it is in or outside the application domain. This also gives an estimation for the confidence of the predicted results.

3. Results

3.1. Exploratory analysis and data filtering

The exploratory analysis of the input variables shows that some of the 13 physical-chemical property variables are highly skewed (Fig. S2). Log-transforming the data helps alleviate such skewed distributions and help making the data patterns more interpretable. Fig. S3 shows the distribution of log-transformed physical-chemical property variables. After the log-transformation, most of the variables show approximately normal distribution. Therefore, highly skewed variables (i.e., MW, AOH, BCF, BioHL, HL, KM, VP, WS, LogP) are log-transformed before used to build the machine learning models and linear regression models. Note that the output variable HC_{50} is already log-transformed in USEtox.

We then remove the chemicals with Cook's distance larger than three times the mean Cook's distance. As a result, 2122 out of 2307 chemicals (92%) remain for building the machine learning models (Fig. S4). The off-diagonal plots in Fig. S5 show the correlation between any two variables after data filtering, including input and output variables. Fig. S6 lists the calculated Pearson correlation coefficient (ρ) beween any two variables. The Pearson correlation coefficients between HC_{50} and each variable in descending order of their absolute values are: WS ($\rho = 0.731$), LogP ($\rho = -0.660$), MP ($\rho = -0.658$), BCF ($\rho = -0.619$), MW ($\rho = -0.590$), KM ($\rho = -0.519$), BP ($\rho = -0.499$), KOA ($\rho = -0.498$), VP ($\rho = 0.381$), BioHL ($\rho = -0.225$), KOC ($\rho = -0.219$), AOH ($\rho = -0.108$), MoA ($\rho = -0.100$), and HL ($\rho = -0.012$).

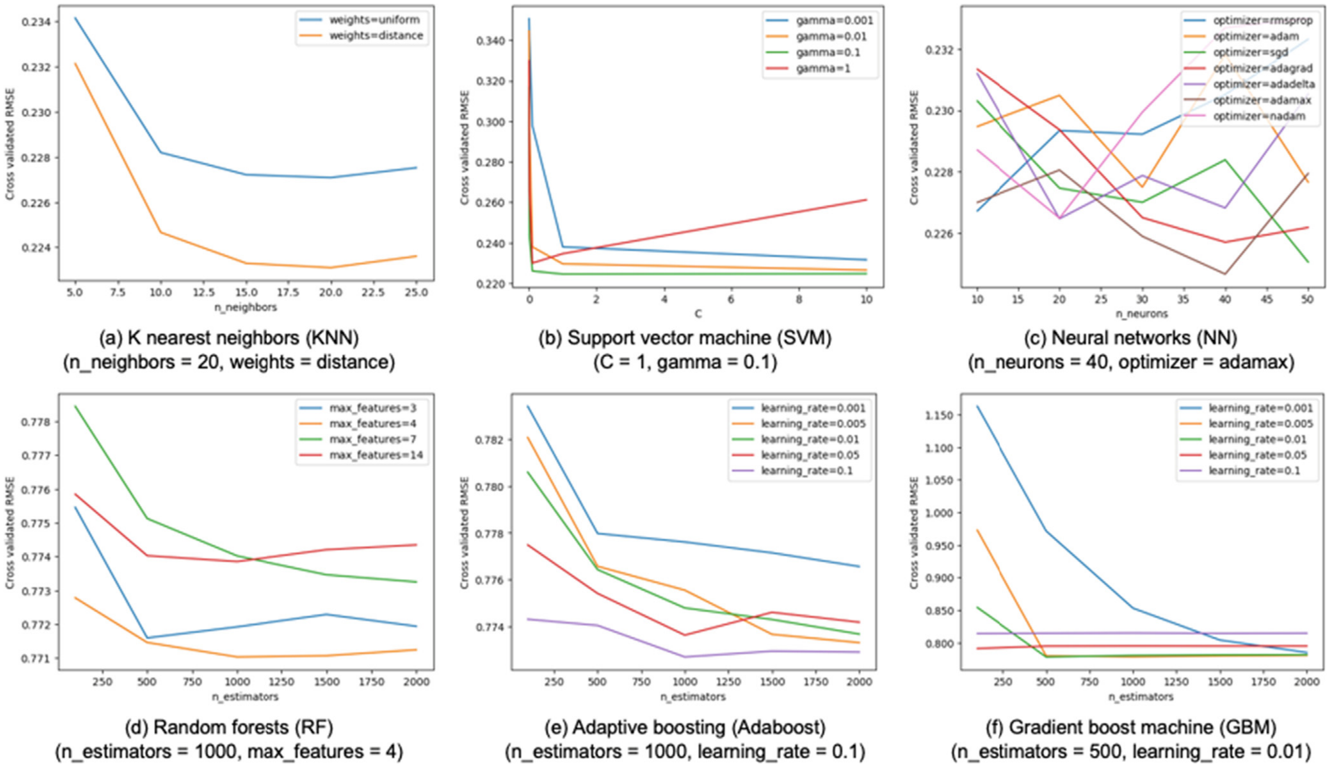
3.2. Performance of the machine learning models

We use cross-validation to choose the best parameters in each model. Take RF as an example. Table 2 shows the detailed results of the training, cross-validation, and best parameters selected based on each split of data. The training *RMSE* is unsurprisingly lower than the cross-validated *RMSE* because the model is trained to fit the training set. Cross-validated *RMSE*, evaluated on the data not used in training, is the criterion of selecting parameters. The best parameters selected based on different splits of data can be different, but the difference of cross-validated *RMSE* is trivial, showing by the small standard deviation. Table S1–S6 present the corresponding results for all the machine learning models.

Fig. 3. shows the model selection results for the 10th split of data. KNN is the most straightforward method among all the models. Weighting the k nearest neighbors based on their distance has a lower cross-validated *RMSE*, which makes sense that closer data points should have more influence on the prediction. The best number of nearest neighbors is 20, less than which does not have enough information for prediction, and more would introduce noise and increase errors. For SVM, our experiments show that the RBF kernel always performs better than the polynomial kernel (Table S7). Using RBF kernel, the best gamma is 0.1 and the best C is 1. For NN, our experiments show that relu performs better than other activation functions (Table S8). Using relu as the activation function, the best number of hidden neurons is 40 and the best optimizer is adamax. For RF, when the maximum number of features is 4, i.e., one-third of the total number of features, which is in line with the recommendation of the inventors for regression problems (Friedman et al., 2001). The best number of trees is 1000. For Adaboost, we use a decision tree of maximum depth is 9 and maximum features is 4 as the base learner. The cross-validated *RMSE* is lowest when the learning rate is 0.1 and the number of trees is to 1000. We use the same decision tree as the base learner for GBM. When the learning rate is 0.01 and the number of trees is 500, the cross-validated *RMSE* is the lowest.

Note that the cross-validation RMSE of KNN, SVM, and NN are much smaller than those of RF, Adaboost, and GBM. This is because each input and output variable are transformed in to the range of [−1, 1] before fitting the first three models since they are sensitive to the scale of the data.

After model selection, the machine learning models are trained again using the best parameters on the whole training and validation set, and then tested on the test set, which evaluates the model performance on "unknown" data. Table 3 shows the average performance of the machine learning models on ten different splits of data. According to the average test R^2 and test *RMSE*, RF has the best performance, followed by GBM, AdaBoost, KNN, SVM, and NN. RF, GBM, and AdaBoost are tree-based models. Trees in RF are independent of each other.



(a) K nearest neighbors (KNN)
(n_neighbors = 20, weights = distance)

(b) Support vector machine (SVM)
(C = 1, gamma = 0.1)

(c) Neural networks (NN)
(n_neurons = 40, optimizer = adamax)

(d) Random forests (RF)
(n_estimators = 1000, max_features = 4)

(e) Adaptive boosting (Adaboost)
(n_estimators = 1000, learning_rate = 0.1)

(f) Gradient boost machine (GBM)
(n_estimators = 500, learning_rate = 0.01)

Fig. 3. Model selection results by cross validation based on the 10th split of data.

The final prediction is the average of predictions of all trees in the forest. Different from RF, trees in AdaBoost and GBM are build based on the errors identified in previous trees. Therefore, the training *RMSE* of AdaBoost and GBM is lower than that of RF. Also shown by Fig. 4, the data points with extreme log HC_{50} values are not well predicted by RF, while AbaBoost and GBM are able to fix those data points. However, this does not help the majority of data points in between. The test *RMSE* of AdaBoost and GBM is not as good as that of RF. On average, RF has the best predictive performance.

KNN has excellent performance on the training set because it stores the entire training data to calculate the distance, but it performs not as good on the test set. SVM only uses a subset of training points in the decision function (i.e., support vectors), so the training *RMSE* is relatively large. NN also have large training *RMSE*, because we set the epochs (i.e., the number of times a model work through the training set) at 50 since after which the validation *RMSE* no longer decreases although training *RMSE* continue decreases as training goes on (Fig. S7). The performance on KNN, SVM, and NN on test set are not as good as tree-based models. This is because they in nature are not good at handling “mixed” types of data (Friedman et al., 2001) and require data to be scaled in the same range. For example, many NN applications involve images (each feature is a pixel) or speech signals (each feature is an amplitude sample), which all have the same kinds of values. In fact, the physical-chemical properties of chemicals are usually measured on very different scales (Fig. S3). On the other hand, tree-based models are more robust when dealing with such kind of data.

3.3. Performance of the ECOSAR model and the linear regression models

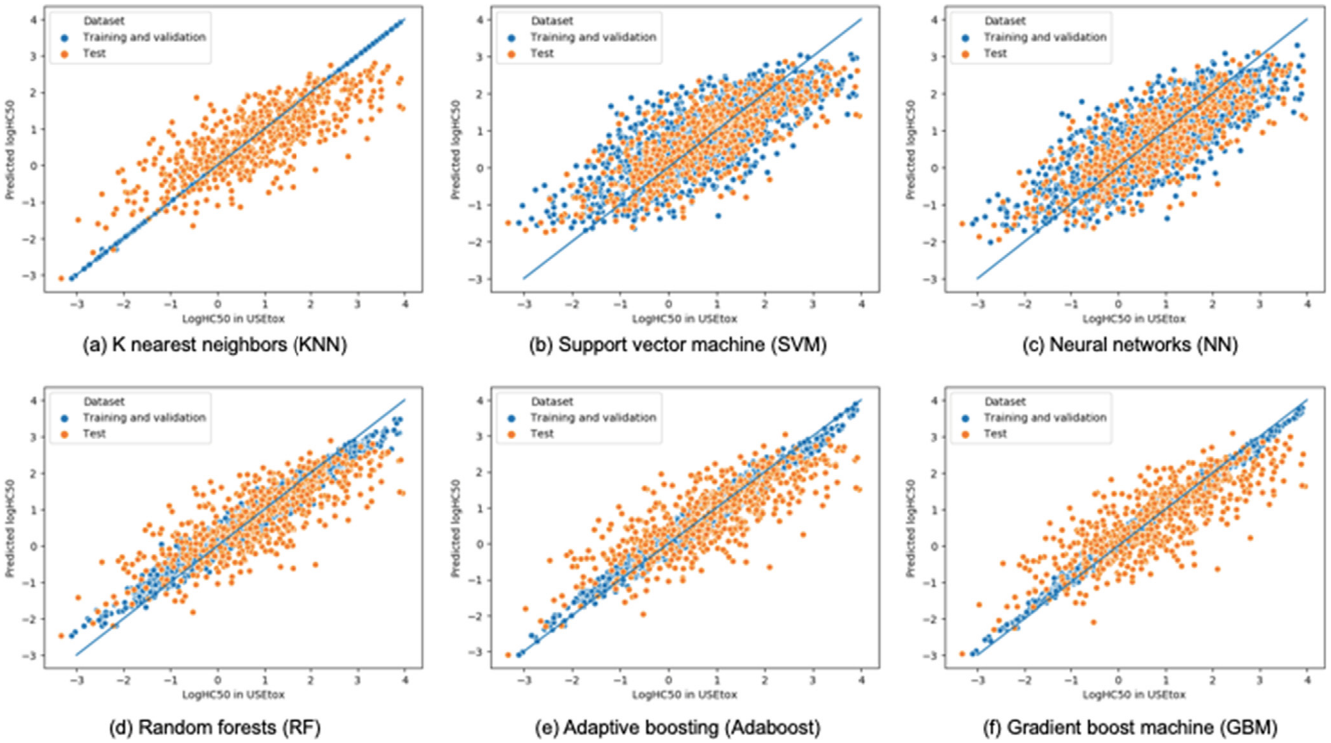
Table 4 shows the performance of the ECOSAR and the linear regression models. The ECOSAR model can be directly used to calculate the test *RMSE* and test R^2. For the filtered 2122 chemicals, ECOSAR predicts EC_{50} of at least three species for 2195 chemicals. The geometric mean of the EC_{50} (i.e., HC_{50}) is compared with HC_{50} in USEtox to evaluate the performance of ECOSAR model. Training is not needed for the ECOSAR model. The calculated test *RMSE* of the ECOSAR model is 1.398, which is close to 1.29 calculated by Melikov et al., (2016). The test R^2 is zero because of extreme outliers (Fig. S8).

The R^2 and *RMSE* of the linear regression models are calculated based on the same ten splits of data. The training *RMSE*of the OLS model with interactions is lower than the model without, but the test *RMSE*is higher than the model without. This means that the OLS model

Table 3
Performance of the machine learning models.

Models	Training**		Testing	
	R^2	*RMSE*	R^2	*RMSE*
K nearest neighbor (KNN)	0.999 (0.000)	0.028 (0.005)	0.623 (0.015)	0.767 (0.012)
Support vector machine (SVM)	0.654 (0.027)	0.728 (0.028)	0.611 (0.015)	0.780 (0.020)
Neural network (NN)	0.654 (0.018)	0.729 (0.016)	0.602 (0.016)	0.788 (0.021)
Random forest (RF)	0.948 (0.001)	0.281 (0.003)	0.630 (0.016)	0.761 (0.017)
Adaptive boosting (AdaBoost)	0.958 (0.010)	0.254 (0.026)	0.628 (0.017)	0.762 (0.018)
Gradient boosting machine (GBM)	0.985 (0.014)	0.141 (0.059)	0.629 (0.018)	0.761 (0.016)

*Numbers in parentheses are the standard deviation of the R^2 or *RMSE* on ten different splits of data.
**Training based on the selected best parameters and the whole training and validation set.



(a) K nearest neighbors (KNN)

(b) Support vector machine (SVM)

(c) Neural networks (NN)

(d) Random forests (RF)

(e) Adaptive boosting (Adaboost)

(f) Gradient boost machine (GBM)

Fig. 4. Scatter plots of log HC_{50} in USEtox and predicted log HC_{50} by machine learning models.

with interactions overfits the training data since the 14 variables with their interaction terms generates 105 variables, which could fit the noises in the training data, thus perform poorly on the test data. The test R^2 and test *RMSE* of the PLS model and the PCR model are the same as the OLS model without interactions. PLS and PCR are particularly useful when the number of variables is much larger than the number of observations (i.e., in high dimensional setting) (James et al., 2013). Here we have 14 variables and 2122 observations, therefore PLS and PCR do not show their advantage. Another reason is that PLS and PCR models are designed to resolve the multicollinearity issue by constructing latent independent variables underlying the collinear variables. Multicollinearity can make the coefficients in OLS models unstable and difficult to interpret. Variance inflation factor (VIF) indicates the extent to which multicollinearity is present in a regression analysis. The VIFs of the 14 variables are all in the acceptable range (a rule of thumb is when VIF is less than 10 (Hair et al., 2006)), which means no severe multicollinearity exists in the dataset (Table S9). As a result, PLS and PCR have the same performance as the OLS model. They all perform not as good as the machine learning models.

We conduct residue checks for the machine learning models and the OLS linear model to ensure the models are valid. The results show the residues are all approximately normally distributed; no systematic bias is detected (Fig. S9).

3.4. Variable importance

Our results show that machine learning models have better prediction performance than ECOSAR and linear regression models for estimating HC_{50} values of chemicals based on their physical-chemical properties and their classification of mode of action. However, most of the machine learning models (except KNN) are regarded as "black box". They are hard to interpret due to their complicated nonlinear functions (Muhlbacher et al., 2014; Rudin, 2019). Here we try to find reasonable explanations for the relative importance of variables in predicting HC_{50} values of chemicals. Among the machine learning models, RF has the best predictive performance. A great quality of RF and other tree-based models is that they can easily measure the relative importance of each feature on the prediction.

Fig. 5. shows the variable importance of the random forest model. The most important variables is WS. WS (water solubility) has the highest correlation coefficient ($\rho = 0.731$) with HC_{50}. Water solubility of a chemical influences its fate and transport in all environmental media and is especially relevant to exposure via aquatic pathways. Soluble chemicals are more available for traveling with water and for chemical and biological transformations (Bloom and de Serres, 1995). This might explain why higher WS is generally associated with higher ecotoxicity.

LogP, MP, BCF, and MW are relatively important variables. LogP

Table 4
Performance of the ECOSAR and the linear regression models.

Models	Training		Testing	
	R^2	*RMSE*	R^2	*RMSE*
ECOSAR	NA	NA	0 (NA)	1.398 (NA)
OLS without interaction	0.578 (0.004)	0.805 (0.008)	0.571 (0.010)	0.819 (0.018)
OLS with interaction	0.600 (0.019)	0.792 (0.023)	0.501 (0.066)	0.859 (0.045)
PLS	0.578 (0.004)	0.805 (0.007)	0.571 (0.010)	0.819 (0.020)
PCR	0.578 (0.004)	0.805 (0.007)	0.570 (0.010)	0.819 (0.020)

*Numbers in parentheses are the standard deviation of the test R^2 or test *RMSE* on ten different splits of data.

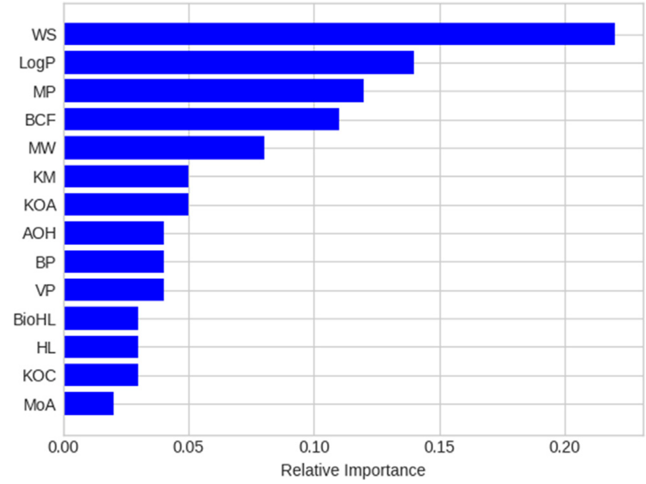


Fig. 5. Variable importance of the random forest model.

(octanol-water partitioning coefficient) is the ratio of a chemical's concentration in octanol (representing lipid "fat" in general) compared to water, measuring a chemical's affinity for the lipid portions of an organism's tissues. LogP has a relatively high negative correlation with HC_{50} ($\rho = -0.660$). Higher LogP indicates higher hydrophobicity, thus lower ectoxicity. MP (melting point) also has a relatively high correlation with HC_{50} ($\rho = -0.658$). MP and LogP have been together used to estimate aqueous solubility, which is a direct measure of the hydrophobicity of a chemical (Council, 2014). BCF (bioconcentration factor) can be expressed as the ratio of the concentration of a chemical in an organism to the concentration of the chemical in the surrounding environment, and, thus, it is relatively highly correlated with LogP ($\rho = 0.747$). MW (molecular weight) is negatively associated with WS ($\rho = -0.716$), which means the higher the MW, the lower the WS, and the lower the ecotoxicity. Other variables show relatively low importance in the RF model.

Table S10 lists the calculated variable importance in all machine learning and linear regression models. Fig. S10 shows the importance order of each variable in each model. The importance order of tree-based models is very similar to each other. Compared with other models, a major difference is that MoA is the least important variable in tree-based models, but is the first or second important variable in the other machine learning models and is relatively important in linear models. This is because MoA is a discrete variable. It has been observed that variables with a smaller number of unique values are less preferred in tree-based models (Altmann et al., 2010; Palermo et al., 2009). MoA is determined by the Verhaar scheme. From class 1 to class 4, the ecotoxicity of chemicals is generally increasing; thus it is an important variable in other models.

Another difference is that LogP is an important variable in tree-based models, but is the least important variable in linear models. This is because LogP is highly correlated with WS ($\rho = -0.82$), MP ($\rho = 0.79$), and BCF ($\rho = -0.75$), thus the independent information in LogP is little. However, in tree-baesd models, the variables for splitting trees are randomly selected; thus collinearity have little impact on the model (Sharma, 2011).

3.5. *Prediction of missing HC_{50} and CF_{eco} in USEtox*

in USEtox version 2.11., 578 chemicals do not have HC_{50} and CF_{eco} values. Among them, 552 chemicals have physical-chemical properties in CompTox. We estimate the HC_{50} and CF_{eco} for the 552 chemicals using the random forest model with 1000 trees and the maximum features for splitting trees is 4. Fig. S11 shows the physical-chemical property distributions of the data building the models and the data for prediction. In general, they share similar distributions. Appendix A provides the estimated HC_{50} with their 95% confidence interval and the relative distance to the application domain (less than one means in the domain, and larger than one means outside of the domain), together with their derived CF_{eco} values. These estimates do not replace the need for chemical-specific laboratory tests to obtain accurate HC_{50} values; but they serve as a useful reference when laboratory test data are not available.

4. Discussion

4.1. *Drawbacks and advantages of random forests and their application in LCA*

We estimate the missing HC_{50} in USEtox using a random forest model. Random forests have some drawbacks: (1) They do not have an explicit mathematical function; therefore, it is hard to interpret the reasons that some chemicals have high predicted ecotoxicity and others not. When applied in LCA studies, the derived characterization factors based on predicted HC_{50} values can first be used to calculate preliminary results, with which the chemicals with high impact in a product's life cycle can be identified. LCA practitioners can then look for more accurate data or conduct tests for the identified chemicals to improve the data quality of the LCA studies; (2) Random forests are in favor of numerical variables and categorical (and discrete) variables with more levels. Variable with less unique values (e.g., MoA) are shown less important in random forests; (3) Training a large number of deep trees use a lot of memory and can have high computational costs, but it can be easily paralleled since all trees are independent.

On the other hand, random forests have the following advantages: (1) Fast and less costly. Once trained, the prediction is pretty fast, so we can calculate characterization factors for a larger number of chemicals; (2) Good performance. As shown by the results, the random forest model outperforms other machine learning and linear models. The performance can be further improved with more data, either ecotoxicity data or relevant predictor variables; (3) Easy to tune. Random forests are not sensitive to parameters. Typically, two parameters need to be tuned, the number of trees and the number of features to be selected for splitting trees; (4) Wide applicability. Random forests can also be used to predict characterization factors for other environmental impacts in LCA given the existing data and relevant predictor variables.

4.2. *Limitations and future work*

In this work, we only use thirteen physical-chemical properties of chemicals in CompTox and an additional classification of mode of action variable. In addition to physical-chemical properties, other chemical databases contain a wide range of molecular descriptors, such as Dragon 7.0 (Mauri et al., 2006), Toxicity Estimation Software Tool (T.E.S.T.) (EPA, 2016), and VEGA (Benfenati et al., 2013). These additional data may provide more valuable information for predicting the ecotoxicity of chemicals that can be used in LCIA. For example, studies found that some variables (e.g., the energy gap between the highest occupied and lowest unoccupied frontier orbitals) not considered in our model are also effective at determining the acute aquatic ecotoxicity (Kostal et al., 2015; Voutchkova et al., 2011). We will consider these additional variables in our future research.

In this work, we focus on estimating effect factors which is the reason of missing ecotoxicity characterization factors in current USEtox (v2.11). In this version, fate factors and exposure factors are calculated by solving a set of mass balance equations characterizing the degradation and inter-compartment transfer processes. In the future development of USEtox, if more chemicals are to be included, random forests can also be used to estimate exposure factors and fate factors.

Our model is currently validated by independent test sets which are not used for training, but also come from USEtox. In future work, additional external validation is useful to further validate the model. We can validate by existing experimental data, e.g., EPA's Ecotox database (EPA, 2006), or by conducting laboratory or field tests.

We choose USEtox as our data source of HC_{50} because USEtox is the LCIA model endorsed by the Life Cycle Initiative – hosted by UN environment. Our data-driven approach takes known HC_{50} values of chemicals as inputs and estimate unknown HC_{50} values for other chemicals. HC_{50} in USEtox are aggregated from EC_{50} across species and test conditions, and the present extrapolated HC_{50} are subject to the same limitation as the HC_{50} used in the present version of USEtox, which did not take full advantage of all available chronic data. Once the latest recommendations for the Life Cycle Initiative on ecotoxicity effect factor (Owsianiak et al., 2019) have been fully implemented based on latest databases (e.g., Posthuma et al., 2019), there will be a need to apply the present approach. The future work can either directly extrapolate HC_{20} of EC_{10} or estimate chronic EC_{10} for each species based on consistent data sets and then aggregate these EC_{10} to HC_{20}, which may improve model performance.

In addition, ecotoxicity of chemicals in USEtox is currently based on freshwater toxicity without consideration of terrestrial or marine toxicity. Our results thus are also confined in the scope of freshwater toxicity. There are other ecotoxicity impact assessment models that covers terrestrial and marine toxicity besides freshwater toxicity. For example, ReCiPe2016 (Huijbregts et al., 2017) covers 18,590 chemicals and provides characterization factors with three perspectives (i.e., 20-year, 100-year, and infinite), Still, 71% of these chemicals do not have their 20-year characterization factors, and 27% without 100-year and infinite characterization factors. The model proposed in this paper can be applied with input variables calculated by CompTox and ToxTree.

It is important to use high quality data to train machine learning models because their performance is affected by outliers (Khamis et al., 2005). We filter the unreliable data by Cook’s distance. We tested the scenario that uses all available data in USEtox without data filtering to train the random forest model. Results show that the average test R^2 is 0.54 and the average test *RMSE* is 0.95. The model performs better after removing these chemicals. Detailed information of the removed chemicals is listed in Appendix A.

4.3. Implications for LCA

We envision three major implications for LCA research and practice. First, LCA practitioners can directly use the derived CF_{eco} values in LCA case studies. We provide estimates of HC_{50} values for 552 chemicals with confidence interval and application domain, together with their derived CF_{eco} values in Appendix A. These estimates do not replace the need for chemical-specific laboratory or field tests to obtain accurate HC_{50} values; instead, they serve as a useful reference when laboratory or field test data are not available.

Second, the HC_{50} values of chemicals that are not included in USEtox can be estimated using our method. First, one needs to collect or calculate the input variables of these chemicals for the model. CompTox provides physical-chemical properties for 875,000 chemicals. ToxTree (Patlewicz et al., 2008) can be used to assign the classification of MoA for chemicals. One then trains the model with the training data from USEtox (provided in Appendix A). Next, the missing HC_{50} can be predicted by the developed random forest model with 1000 trees and the maximum number of features is 4. We use Python Scikit-learn machine learning module to build our model. One can also use machine learning packages in other tools such R (e.g., randomForest) to train the model and predict missing $HC_{50.}$ Finally, one can calculate CF_{eco} using the estimated HC_{50} and other required inputs in the multimedia transport and transformation model in USEtox.

Third, one can develop random forest models to estimate characterization factors for other impact categories, e.g., human toxicity. Due to the limited availability of chronic data for estimating dose-response and disease incidences, 67.4% of the chemicals in USEtox do not have human toxicity characterization factors. Estimating human toxicity for chemicals represents an interesting future research direction. The first and most important task is to find variables relevant to human toxicity or other interested impacts. Then, one can develop random forests or other machine learning models following the same procedure as described in this paper.

Beyond LCA, although the application of machine learning models is suggested to be cautious due to lack of transparency in chemical risk assessment, they can still be used as a screening process to identify chemicals with high predicted ecotoxicity potentials.

5. Conclusions

Estimating ecotoxicity of chemicals is a difficult task because of the complex physical, chemical, and biological processes of how chemicals transform and interact in environmental media. The machine learning models, or computational models in general, explore these complex processes through the pattern revealed from observed data. Our results show that random forest model developed in this study performs better than ECOSAR, linear regression models, and other machine learning models. Although the model structure is not easy to interpret, the random forest model provides an efficient way to quickly predict the HC_{50} values and ecotoxicity characterization factors of chemicals for LCA and broader applications.

CRediT authorship contribution statement

Ping Hou: Conceptualization, Data curation, Formal analysis, Investigation, Visualization, Writing – original draft. **Olivier Jolliet:** Resources, Validation, Writing - review & editing. **Ji Zhu:** Methodology, Writing - review & editing. **Ming Xu:** Conceptualization, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank the anonymous reviewers for their insightful comments and suggestions.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2019.105393.

References

Alexander, D.L.J., Tropsha, A., Winkler, D.A., 2015. Beware of R-2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. J. Chem. Inf. Model. 55, 1316–1322.

Altman, N.S., 1992. An introduction to kernel and nearest-neighbor nonparametric regression. Am. Stat. 46, 175–185.

Altmann, A., Tolosi, L., Sander, O., Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. Bioinformatics 26, 1340–1347.

Azari, R., Garshasbi, S., Amini, P., Rashed-Ali, H., Mohammadi, Y., 2016. Multi-objective optimization of building envelope design for life cycle environmental performance. Energy Build. 126, 524–534.

Benfenati, E., Manganaro, A., Gini, G.C., 2013. VEGA-QSAR: AI Inside a Platform for Predictive Toxicology. PAI@ AI* IA.

Benfenati, E., Roncaglioni, A., Lombardo, A., Manganaro, A., 2019. Integrating QSAR, read-across, and screening tools: the VEGAHUB platform as an example. In: Hong, H. (Ed.), Advances in Computational Toxicology: Methodologies and Applications in Regulatory Science. Springer International Publishing Ag, Cham.

Birkved, M., Heijungs, R., 2011. Simplified fate modelling in respect to ecotoxicological and human toxicological characterisation of emissions of chemical compounds. Int. J. Life Cycle Assess. 16, 739–747.

Bloom, A.D., de Serres, F., 1995. Ecotoxicity and human health: a biological approach to environmental remediation edˆeds. CRC Press.

Chavan, S., Friedman, R., Nicholls, I.A., 2015. Acute toxicity-supported chronic toxicity prediction: a k-nearest neighbor coupled read-across strategy. Int. J. Mol. Sci. 16, 11659–11677.

Chen, J.L., Liau, C.-W., 2001. A simple life cycle assessment method for green product conceptual design. Second International Symposium on Environmentally Conscious Design and Inverse Manufacturing, 2001 Proceedings EcoDesign 2001. IEEE.

Chiang, T.-A., Che, Z., Wang, T.-T., 2011. A design for environment methodology for evaluation and improvement of derivative consumer electronic product development. J. Syst. Sci. Syst. Eng. 20, 260–274.

Chiang, T.-A., Roy, R., 2012. An intelligent benchmark-based design for environment system for derivative electronic product development. Comput. Ind. 63, 913–929.

Cook, R.D., 1977. Detection of influential observation in linear-regression. Technometrics 19, 15–18.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20, 273–297.

Council, N.R., 2014. A Framework to Guide Selection of Chemical alternatives edˆeds. National Academies Press.

Deeb, O., Goodarzi, M., 2012. In silico quantitative structure toxicity relationship of chemical compounds: some case studies. Current Drug Safety 7, 289–297.

DiCiccio, T.J., Efron, B., 1996. Bootstrap confidence intervals. Statistical Science 189–212.

ECHA, 2016. Practical Guide 5 How to Use and Report (Q) SARs. European Chemicals Agency.

EPA, U. ECOTOX database, 2006.

EPA, U.S. User's Guide for T.E.S.T. (version 4.2) (Toxicity Estimation Software Tool): A Program to Estimate Toxicity from Molecular Structure. 2016.

Fantke, P., Aurisano, N., Bare, J., Backhaus, T., Bulle, C., Chapman, P.M., De Zwart, D., Dwyer, R., Ernstoff, A., Golsteijn, L., Holmquist, H., Jolliet, O., McKone, T.E., Owsianiak, M., Peijnenburg, W., Posthuma, L., Roos, S., Saouter, E., Schowanek, D., van Straalen, N.M., Vijver, M.G., Hauschild, M., 2018a. Toward harmonizing ecotoxicity characterization in life cycle impact assessment. Environ. Toxicol. Chem. 37, 2955–2971.

Fantke, P., Aylward, L., Bare, J., Chiu, W.A., Dodson, R., Dwyer, R., Ernstoff, A., Howard, B., Jantunen, M., Jolliet, O., Judson, R., Kirchhubel, N., Li, D.S., Miller, A., Paoli, G., Price, P., Rhomberg, L., Shen, B., Shin, H.M., Teeguarden, J., Vallero, D., Wambaugh, J., Wetmore, B.A., Zaleski, R., McKone, T.E., 2018b. Advancements in life cycle human exposure and toxicity characterization. Environ. Health Perspect. 126.

Fantke, P.E., Bijster, M., Guignard, C., Hauschild, M., Huijbregts, M., Jolliet, O., Kounina, A., Magaud, V., Margni, M., McKone, T.E., Posthuma, L., Rosenbaum, R.K., van de Meent, D., van Zelm, R, 2017. USEtox® 2.0 user manual (Version 1). http://usetoxorg.

Friedman, J., Hastie, T., Tibshirani, R, 2001. The Elements of Statistical Learning edˆeds, Springer series in statistics New York.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29, 1189–1232.

Frischknecht, R., Jolliet, O., 2016. Global guidance for life cycle impact assessment indicators. UNEP/SETAC Life Cycle Initiative, Paris.

Furuhama, A., Toida, T., Nishikawa, N., Aoki, Y., Yoshioka, Y., Shiraishi, H., 2010. Development of an ecotoxicity QSAR model for the KAshinhou Tool for Ecotoxicity (KATE) system, March 2009 version. SAR QSAR Environ. Res. 21, 403–413.

Gomes, A.I., Pires, J.C.M., Figueiredo, S.A., Boaventura, R.A.R., 2014. Multiple linear and principal component regressions for modelling ecotoxicity bioassay response. Environ. Technol. 35, 945–955.

Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., Tatham, R.L., 2006. Multivariate data analysis 6th Edition. Pearson Prentice Hall New Jersey humans: Critique and reformulation. J. Abnormal Psychol. 87, 49–74.

Hauschild, M.Z., Huijbregts, M., Jolliet, O., MacLeod, M., Margni, M., van de Meent, D.V., Rosenbaum, R.K., McKone, T.E., 2008. Building a model based on scientific consensus for life cycle impact assessment of chemicals: The search for harmony and parsimony. Environ. Sci. Technol. 42, 7032–7037.

Haykin, S., Network, N., 2004. A comprehensive foundation. Neural Netw. 2, 41.

Henderson, A.D., Hauschild, M.Z., van de Meent, D., Huijbregts, M.A.J., Larsen, H.F., Margni, M., McKone, T.E., Payet, J., Rosenbaum, R.K., Jolliet, O., 2011. USEtox fate and ecotoxicity factors for comparative assessment of toxic emissions in life cycle analysis: sensitivity to key chemical properties. Int. J. Life Cycle Assess. 16, 701–709.

Hinds, R.d.C., Weller, J.L., 2016. Toxic Substances Control Act. Environmental Law Practice Guide, vol. 4.

Huijbregts, M.A., Steinmann, Z.J., Elshout, P.M., Stam, G., Verones, F., Vieira, M., Zijp, M., Hollander, A., van Zelm, R., 2017. ReCiPe2016: a harmonised life cycle impact assessment method at midpoint and endpoint level. Int. J. Life Cycle Assess. 22, 138–147.

ISO, 2006. 14040: Environmental Management-Life Cycle Assessment-Principles and Framework. ISO.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning edˆeds. Springer.

Khamis, A., Ismail, Z., Haron, K., Mohammed, A.T., 2005. The effects of outliers data on neural network performance. J. Appl. Sci. 5, 1394–1398.

Khoshnevisan, B., Rafiee, S., Omid, M., Mousazadeh, H., Sefeedpari, P., 2013. Prognostication of environmental indices in potato production using artificial neural networks. J. Cleaner Prod. 52, 402–409.

Kienzler, A., Barron, M.G., Belanger, S.E., Beasley, A., Embry, M.R., 2017. Mode of action (MOA) assignment classifications for ecotoxicology: an evaluation of approaches. Environ. Sci. Technol. 51, 10203–10211.

Kostal, J., Voutchkova-Kostal, A., Anastas, P.T., Zimmerman, J.B., 2015. Identifying and designing chemicals with minimal acute aquatic toxicity. PNAS 112, 6289–6294.

Kutner, M.H., Nachtsheim, C.J., Neter, J., Li, W., 2005. Applied Linear Statistical Models edˆeds. McGraw-Hill Irwin Boston.

Li, F.X., Fan, D.F., Wang, H., Yang, H.B., Li, W.H., Tang, Y., Liu, G.X., 2017. In silico prediction of pesticide aquatic toxicity with chemical category approaches. Toxicol. Res. 6, 831–842.

Mackay, D., Hubbarde, J., Webster, E., 2003. The role of QSARs and fate models in chemical hazard and risk assessment – Paper prepared for quantitative structure-activity relationships (QSAR) Proceedings of the QSAR 2002 Conference, Ottawa May 2002. Qsar & Combinatorial Science, vol. 22, pp. 106–112.

Maitra, S., Yan, J., 2008. Principle component analysis and partial least squares: two dimension reduction techniques for regression. Appl. Multivariate Statist. Models 79, 79–90.

Mansouri, K., Grulke, C., Judson, R., Williams, A., 2018. OPERA: a free and open source QSAR tool for predicting physicochemical properties and environmental fate endpoints. Abst. Papers Am. Chem. Soc. 255.

Martin, T., 2016. User's guide for TEST (version 4.2)(Toxicity Estimation Software Tool) A program to estimate toxicity from molecular structure. US EPA Office of Research and Development, Washington, DC. EPA/600/R-16/058 Google Scholar.

Marvuglia, A., Kanevski, M., Benetto, E., 2015a. Machine learning for toxicity characterization of organic chemical emissions using USEtox database: learning the structure of the input space. Environ. Int. 83, 72–85.

Marvuglia, A., Kanevski, M., Leuenberger, M., Benetto, E., 2014. Variables selection for ecotoxicity and human toxicity characterization using Gamma Test. In: International Conference on Computational Science and Its Applications. Springer.

Marvuglia, A., Leuenberger, M., Kanevski, M., Benetto, E., 2015b. Random forest for toxicity of chemical emissions: features selection and uncertainty quantification. J. Environ. Account. Manage. 3, 229–241.

Mauri, A., Consonni, V., Pavan, M., Todeschini, R., 2006. Dragon software: an easy approach to molecular descriptor calculations. Match 56, 237–248.

Mayo-Bean, K., Nabholz, J., Clements, R., Zeeman, M., Henry, T., Rodier, D., Moran, K., Meylan, B., Ranslow, P., 2011. Methodology document for the ECOlogical Structure-Activity Relationship Model (ECOSAR) class program: estimating toxicity of industrial chemicals to aquatic organisms using ECOSAR class program (Ver. 1.1). In: US Environmental Protection Agency, Office of Chemical Safety and Pollution Prevention, Office of Pollution Prevention and Toxics, Washington, DC.

Melnikov, F., Kostal, J., Voutchkova-Kostal, A., Zimmerman, J.B., Anastas, P.T., 2016. Assessment of predictive models for estimating the acute aquatic toxicity of organic chemicals. Green Chem. 18, 4432–4445.

Miller, T.H., Gallidabino, M.D., MacRae, J.I., Hogstrand, C., Bury, N.R., Barron, L.P., Snape, J.R., Owen, S.F., 2018. Machine learning for environmental toxicology: a call for integration and innovation. Environ. Sci. Technol. 52, 12953–12955.

Muhlbacher, T., Piringer, H., Gratzl, S., Sedlmair, M., Streit, M., 2014. Opening the black box: strategies for increased user involvement in existing algorithm implementations. IEEE Trans. Visual Comput. Graphics 20, 1643–1652.

Nabavi-Pelesaraei, A., Bayat, R., Hosseinzadeh-Bandbafha, H., Afrasyabi, H., Chau, K.-W., 2017. Modeling of energy consumption and environmental life cycle assessment for incineration and landfill systems of municipal solid waste management-A case study in Tehran Metropolis of Iran. J. Cleaner Prod. 148, 427–440.

Nantasenamat, C., Isarankura-Na-Ayudhya, C., Naenna, T., Prachayasittikul, V., 2009. A practical overview of quantitative structure-activity relationship. Excli. J. 8, 74–88.

Netzeva, T.I., Worth, A.P., Aldenberg, T., Benigni, R., Cronin, M.T., Gramatica, P., Jaworska, J.S., Kahn, S., Klopman, G., Marchant, C.A., 2005. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. ATLA 33, 155–173.

Owsianiak, M., Fantke, P., Posthuma, L., Saouter, E., Vijver, M., Backhaus, T., Schlekat, T., Hauschild, M., 2019. Chapter 7 Ecotoxicity. In: Frischknecht, R., Jolliet, O. (Eds.). Global guidance for life cycle impact assessment indicators – vol. 2.

Ozbilen, A., Aydin, M., Dincer, I., Rosen, M.A., 2013. Life cycle assessment of nuclear-based hydrogen production via a copper–chlorine cycle: a neural network approach. Int. J. Hydrogen Energy 38, 6314–6322.

Palermo, G., Piraino, P., Zucht, H.-D., 2009. Performance of PLS regression coefficients in selecting variables for each response of a multivariate PLS for omics-type data. Adv. Appl. Bioinform. Chem.: AABC 2, 57.

Park, J.-H., Seo, K.-K., Wallace, D., 2001. Approximate life cycle assessment of classified products using artificial neural network and statistical analysis in conceptual product design. In: 2001 Proceedings EcoDesign 2001: Second International Symposium on Environmentally Conscious Design and Inverse Manufacturing, IEEE.

Park, J.H., Seo, K.-K., 2003. Approximate life cycle assessment of product concepts using multiple regression analysis and artificial neural networks. J. Mech. Sci. Technol. 17, 1969–1976.

Patlewicz, G., Jeliazkova, N., Safford, R.J., Worth, A.P., Aleksiev, B., 2008. An evaluation of the implementation of the Cramer classification scheme in the Toxtree software. SAR QSAR Environ. Res. 19, 495–524.

Piao, W., Kim, C., Cho, S., Kim, H., Kim, M., Kim, Y., 2016. Development of a protocol to optimize electric power consumption and life cycle environmental impacts for operation of wastewater treatment plant. Environ. Sci. Pollut. Res. 23, 25451–25466.

Posthuma, L., van Gils, J., Zijp, M.C., van de Meent, D., de Zwart, D., 2019. Species sensitivity distributions for use in environmental protection, assessment, and management of aquatic ecosystems for 12 386 chemicals. Environ. Toxicol. Chem. 38, 905–917.

Pradeep, P., Povinelli, R.J., White, S., Merrill, S.J., 2016. An ensemble model of QSAR tools for regulatory risk assessment. J. Cheminf. 8, 9.

Predictor, A., 2015. Simulations plus. Inc, Lancaster, CA, USA, ver, vol. 7.

Raies, A.B., Bajic, V.B., 2016. In silico toxicology: computational methods for the prediction of chemical toxicity. Wiley Interdisciplinary Rev. – Comput. Mol. Sci. 6, 147–172.

Rand, G.M., 1995. Fundamentals of Aquatic Toxicology: Effects, Environmental Fate and

Risk Assessment edˆeds: CRC Press.

Rebitzer, G., Ekvall, T., Frischknecht, R., Hunkeler, D., Norris, G., Rydberg, T., Schmidt, W.P., Suh, S., Weidema, B.P., Pennington, D.W., 2004. Life cycle assessment Part 1: Framework, goal and scope definition, inventory analysis, and applications. Environ. Int. 30, 701–720.

Rosenbaum, R.K., Bachmann, T.M., Gold, L.S., Huijbregts, M.A., Jolliet, O., Juraske, R., Koehler, A., Larsen, H.F., MacLeod, M., Margni, M., 2008. USEtox—the UNEP-SETAC toxicity model: recommended characterisation factors for human toxicity and freshwater ecotoxicity in life cycle impact assessment. Int. J. Life Cycle Assess. 13, 532.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1, 206–215.

Sacan, M.T., Novic, M., Erturk, M.D., Minovski, N., 2015. Marine Algal Toxicity Models with Dunaliella tertiolecta: In Vivo and In Silico. Advances in Mathematical Chemistry and Applications, vol. 2, pp. 148–178.

Sala, S., Marinov, D., Pennington, D., 2011. Spatial differentiation of chemical removal rates from air in life cycle impact assessment. Int. J. Life Cycle Assess. 16, 748–760.

Sangion, A., Gramatica, P., 2016. Hazard of pharmaceuticals for aquatic environment: prioritization by structural approaches and prediction of ecotoxicity. Environ. Int. 95, 131–143.

Saouter, E., Aschberger, K., Fantke, P., Hauschild, M.Z., Bopp, S.K., Kienzler, A., Paini, A., Pant, R., Secchi, M., Sala, S., 2017a. Improving substance information in USEtox((R)), Part 1: Discussion on data and approaches for estimating freshwater ecotoxicity effect factors. Environ. Toxicol. Chem. 36, 3450–3462.

Saouter, E., Aschberger, K., Fantke, P., Hauschild, M.Z., Kienzler, A., Paini, A., Pant, R., Radovnikovic, A., Secchi, M., Sala, S., 2017b. Improving substance information in USEtox((R)), part 2: Data for estimating fate and ecosystem exposure factors. Environ. Toxicol. Chem. 36, 3463–3470.

Schapire, R.E., 2013. Explaining adaboost. Empirical inference. Springer.

Seo, K.-K., Kim, W.-K., 2006. Approximate life cycle assessment of product concepts using a hybrid genetic algorithm and neural network approach. Information Technology. Springer.

Seo, K.-K., Min, S.-H., Yoo, H.-W., 2005. Artificial neural network based life cycle assessment model for product concepts using product classification method. International Conference on Computational Science and Its Applications. Springer.

Sharma, D., 2011. Improving the art, craft and science of economic credit risk scorecards using random forests: Why credit scorers and economists should use random forests. Craft and Science of Economic Credit Risk Scorecards Using Random Forests: Why Credit Scorers and Economists Should Use Random Forests (June 9, 2011).

Shoji, R., 2005. The potential performance of artificial neural networks in QSTRs for predicting ecotoxicity of environmental pollutants. Curr. Comput. Aided Drug Des. 1, 65–72.

Singh, K.P., Gupta, S., Kumar, A., Mohan, D., 2014. Multispecies QSAR modeling for predicting the aquatic toxicity of diverse organic chemicals for regulatory toxicology. Chem. Res. Toxicol. 27, 741–753.

Song, R.S., Keller, A.A., Suh, S., 2017. Rapid life-cycle impact screening using artificial neural networks. Environ. Sci. Technol. 51, 10777–10785.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinf. 9, 307.

Sun, H.M., Xia, M.H., Austin, C.P., Huang, R.L., 2012. Paradigm shift in toxicity testing and modeling. AAPS J. 14, 473–480.

Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P., 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci. 43, 1947–1958.

Tropsha, A., 2010. Best practices for QSAR model development, validation, and exploitation. Mol. Inf. 29, 476–488.

Tuulaikhuu, B.A., Guasch, H., Garcia-Berthou, E., 2017. Examining predictors of chemical toxicity in freshwater fish using the random forest technique. Environ. Sci. Pollut. Res. 24, 10172–10181.

Verhaar, H.J., Van Leeuwen, C.J., Hermens, J.L., 1992. Classifying environmental pollutants. Chemosphere 25, 471–491.

Voutchkova, A.M., Kostal, J., Steinfeld, J.B., Emerson, J.W., Brooks, B.W., Anastas, P., Zimmerman, B., 2011. Towards rational molecular design: derivation of property guidelines for reduced acute aquatic toxicity. Green Chem. 13, 2373–2379.

Voutchkova-Kostal, A.M., Kostal, J., Connors, K.A., Brooks, B.W., Anastas, P.T., Zimmerman, J.B., 2012. Towards rational molecular design for reduced chronic aquatic toxicity. Green Chem. 14, 1001–1008.

Wernet, G., Hellweg, S., Fischer, U., Papadokonstantakis, S., Hungerbuhler, K., 2008. Molecular-structure-based models of chemical inventories using neural networks. Environ. Sci. Technol. 42, 6717–6722.

Wisthoff, A., Ferrero, V., Huynh, T., DuPont, B., 2016. Quantifying the Impact of Sustainable Product Design Decisions in the Early Design Phase Through Machine Learning. ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference: American Society of Mechanical Engineers.

Yin, L., Liao, Y., Zhou, L., Wang, Z., Ma, X., 2017. Life cycle assessment of coal-fired power plants and sensitivity analysis of CO2 emissions from power generation side. IOP Conference Series: Materials Science and Engineering: IOP Publishing.