Machine learning models based on molecular descriptors to predict human and environmental toxicological factors in continental freshwater

Rémi Servien, Eric Latrille, Dominique Patureau, Arnaud Hélias

Machine learning models based on molecular descriptors to predict human and environmental toxicological factors in continental freshwater

Rémi Servien[a,b,*], Eric Latrille[a,b], Dominique Patureau[a], Arnaud Hélias[c,d]

[a]INRAE, Univ. Montpellier, LBE, 102 Avenue des étangs, F-11000 Narbonne, France

[b]ChemHouse Research Group, Montpellier, France

[c]ITAP, Univ Montpellier, INRAE, Institut Agro, Montpellier, France

[d]ELSA, Research group for environmental life cycle sustainability assessment and ELSA-Pact industrial chair, Montpellier, France

*corresponding author : remi.servien@inrae.fr

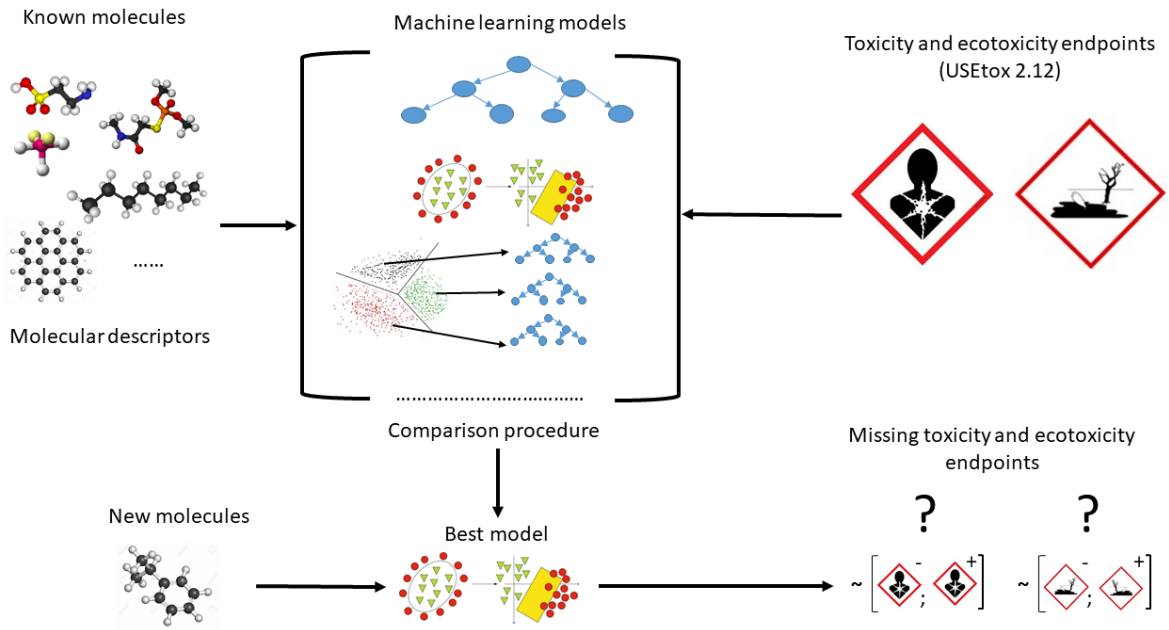
Highlights:

- Characterization factors (for human health and ecotoxicological impacts) were predicted using molecular descriptors.
- Several linear or non-linear machine learning methods were compared.
- The non-linear methods tend to outperform the linear ones using a train and test procedure. Cluster-then-predict approaches often show the best performances, highlighting their usefulness.
- This methodology was then used to derive characterization factors that were missing for more than a hundred chemicals in USEtox®.

Abstract: It is a real challenge for life cycle assessment practitioners to identify all relevant substances contributing to the ecotoxicity. Once this identification has been made, the lack of

corresponding ecotoxicity factors can make the results partial and difficult to interpret. So, it is a real and important challenge to provide ecotoxicity factors for a wide range of compounds. Nevertheless, obtaining such factors using experiments is tedious, time-consuming, and made at a high cost. A modelling method that could predict these factors from easy-to-obtain information on each chemical would be of great value. Here, we present such a method, based on machine learning algorithms, that used molecular descriptors to predict two specific endpoints in continental freshwater for ecotoxicological and human impacts. The different tested machine learning algorithms show good performances on a learning database and the non-linear methods tend to outperform the linear ones. The cluster-then-predict approaches usually show the best performances which suggests that these predicted models must be derived for somewhat similar compounds. Then, predictions were derived from the validated model for compounds with missing toxicity/ecotoxicity factors.

Graphical abstract:



Keywords: machine learning, Life Cycle Assessment, characterisation factors, toxicity, ecotoxicity, continental freshwater.

1. Introduction

Recent legislations such as the Registration, Evaluation, Authorization and restriction of Chemicals (REACH) regulation in the EU requires that manufacturers of substances and formulators register to provide eco/toxicological data for substances with volume higher than one metric ton per year. As an example, the U.S. Environmental Protection Agency (EPA) has more than 85,000 chemicals listed under the Toxic Substances Control Act (Hinds and Weller, 2016). So, robust (eco)-toxicological data are quickly needed to make informed decisions on how to regulate new chemicals. These data must also be coupled with environmental exposures and sources data, to better understand the impact on the environment.

To address the cause-effect relationships between the flow of molecules emitted by human activities and the consequences for ecosystems and humans, LCA offers a structured, operational, and standardized (Finkbeiner et al., 2006) methodological framework. Two main steps are at the core of this approach:

- Quantification of the masses of substances emitted into the environment through the Life Cycle Inventory (LCI). While it is possible to rely on databases that facilitate this inventory work for the background of the system under study, this task must nevertheless be carried out on a case-by-case basis to represent all the specificities of the foreground elements. To best describe human activities, their specificities must be represented on a case-by-case basis. This is the task of the LCA practitioner.
- Calculation of the impacts on ecosystems and human health of these emitted masses. Due to the complexity of environmental mechanisms, it is not possible to (re)model impact pathways on a case-by-case basis. Therefore, LCA uses characterization factors (CF) to assess the potential impacts of a compound. Concretely, if two

compounds are emitted with the same mass, the one with the higher CFs will have the higher impact. Then, CFs are multiplied by the emitted masses of each compound to determine the impacts. CFs are not recalculated for each study but provided within a Life Cycle Impact Assessment (LCIA) method.

For a given impact, the LCIA method designer refers to the knowledge of the scientific community to model the mechanisms involved. For human toxicity and freshwater ecotoxicity, USEtox® (Rosenbaum et al., 2008), was developed by life cycle initiative under the United Nations Environmental Programme (UNEP) and the Society for Environmental Toxicology and Chemistry (SETAC) (Henderson et al. 2011) to produce a transparent and consensus characterization model. USEtox® is also used for the European Product Environmental Footprint (PEF) (Saouter et al., 2020). This model gathers in one single characterization factor the chemical fate, the exposure, and the effect for each of the several thousands of organic and inorganic compounds. Then, the USEtox® model results can be extended to determine endpoint effects expressed as total (i.e. cancer and non-cancer) disability-adjusted life years (DALY) for human health impacts and potentially disappeared fraction of species (PDF) for ecotoxicological impacts. The PDF represents an increase in the fraction of species potentially disappearing as a consequence of emission in a compartment while the DALY represents an increase in adversely affected life years. These endpoints are now consensual at an international level (Verones et al., 2017).

If the structure of the USEtox® multimedia model is always the same, to determine the CF of a molecule, numerous physico-chemical parameters (such as solubility, hydrophobicity, degradability) and detailed toxicological and ecotoxicological data must be provided. For example, EC50 values (i.e. the effective concentration at which 50% of a population died) for at least three species from three different trophic levels are required for the ecotoxocological effect factor.

Over the past few decades, thousands of tests (in laboratory and field) have been carried out to evaluate the potential hazard effects of chemicals (He et al., 2017). Usually, toxicity testing has relied on *in vivo* animal models, which is extremely costly and time-consuming (Xia et al., 2008). In recent years, under societal pressures, there has been a significant paradigm shift in toxicity testing of chemicals from traditional *in vivo* tests to less expensive and higher throughput *in vitro* methods (National Research Council, 2007). However, it is still extremely hard to test the number of existing and ever-increasing numbers of new chemicals, which leaves their impacts largely unknown. That's why more computational models are needed to complement experimental approaches to decrease the experimental cost and determine the prioritization for those chemicals which may need further *in vivo* studies. Such models already exist, like QSAR models that are mostly linear models based on the chemical structure of compounds (Danish QSAR database (DTU, 2015), ECOSAR (Mayo-Bean et al., 2011), VEGA (Benfenati et al., 2013)) and are used to predict ecotoxicological data (LC50) needed for REACH for example. Recently, machine learning algorithms have been used to predict hazardous concentration 50% (HC50) based on 14 physico-chemical characteristics (Hou et al., 2020a) or on 691 more various variables (Hou et al., 2020b). In the case of USEtox®, despite its wide use in LCA, it only offers characterization factors for approximately 3000 chemicals and even for this limited number of compounds, 19% of ecotoxicity CFs and 67% of human toxicity CFs are missing.

The objective of this article is thus to propose a new way of calculating CFs using machine learning approaches to solve the problem of nonlinearity that could affect a linear QSAR method. This makes it possible, when the CFs are not determined due to lack of time or lack of data, to propose values based solely on easily identifiable molecular descriptors. Here, the main differences with the above-cited methods are twofold: first, our input variables are only molecular descriptors that could be easily collected for any newly available compounds; second, our output variables are directly the CFs that are closer to the endpoints (DALY and PDF) than the HC50 or the LC50 (i.e. the acute aquatic toxicity experimental threshold). These two specific endpoints will be studied in the present paper through the emission of compounds

in continental freshwater and will be named CF_{ET} for ecotoxicological impacts and CF_{HT} for human ones. To address this aim, we will test different methods (linear and non-linear) and assess their performances, to build a robust model that could predict CFs that are currently lacking.

2. Materials & Methods

2.1. USEtox® database

The last version of the USEtox® database was downloaded, namely the corrective release 2.12 (USEtox®, 2020). The whole USEtox® 2.12 database contains 3076 compounds.

2.2. TyPol database

We recently developed TyPol (Typology of Pollutants), a classification method based on statistical analyses combining several environmental parameters (i.e., sorption coefficient, degradation half-life, Henry constant) and an ecotoxicological parameter (bioconcentration factor BCF), and structural molecular descriptors (i.e., number of atoms in the molecule, molecular surface, dipole moment, energy of orbitals). Molecular descriptors are calculated using an *in silico* approach (combining Austin Model1 and Dragon software). In the present paper, we only extract and use the molecular descriptors from the TyPol database, as this information could be easily collected for any new compound. The 40 descriptors included in the TyPol database have been selected based on a literature review on QSAR equations used to predict the main environmental processes as degradation, sorption, volatilization. These 40 descriptors were the ones most frequently used in the equations, meaning describing the best the behaviour of organic compounds in the environment. By consequence, even if no environmental parameters are directly incorporated as input in our model, some information that are directly linked to them are included in the 40 molecular descriptors. These descriptors

1 are constitutional, geometric, topological, and quantum-chemical descriptors (see Table 1); 36
2 described the 2D-structure of the compound while the other four are linked to its 3D-structure.
3 An important advantage of the unique use of molecular descriptors is that they are easily and
4 quickly computable for not yet synthesized compounds. For more details, we refer the
5 interested reader to Servien et al. (2014) where the choice of the 40 molecular descriptors is
6 described in details. Now, TyPol gathers 526 compounds, including pesticides,
7 pharmaceuticals and their transformation products (Benoit et al. 2017, Traoré et al. 2018).
8
9 **Table 1** – List of the 40 molecular descriptors in TyPol

Category	Molecular descriptors		
Constitutional	Number of atoms	Number of non-H atoms	Number of hydrogen atoms
	Number of hydrogen atoms	Number of carbon atoms	Number of nitrogen atoms
	Number of oxygen atoms	Number of phosphorus atoms	Number of sulfur atoms
	Number of fluorine atoms	Number of chlorine atoms	Number of halogen atoms
	Number of bonds	Number of non-H bonds	Number of double bonds
	Number of triple bonds	Number of multiple bonds	Number of rotatable bonds
	Number of aromatic bonds	Sum of conventional bond order	Number of rings
	Number of circuits	Molecular weight	
Geometric	Connolly molecular surface area		

Topological	Connectivity index of order 0	Connectivity index of order 1	Connectivity index of order 2
	Connectivity index of order 3	Connectivity index of order 4	Connectivity index of order 5
	Valence connectivity index of order 0	Valence connectivity index of order 1	Valence connectivity index of order 2
	Valence connectivity index of order 3	Valence connectivity index of order 4	Valence connectivity index of order 5
Quantum-chemical	Polarizability	Electric dipole moment	HOMO energy
	LUMO energy	Total energy	

2.3. Machine learning methods

To predict the CFs using the molecular descriptors we use three modelling methods combined. The first method is a linear well-known prediction method namely the Partial Least Squares (PLS) (Wold, 1985). It finds the multidimensional directions in the observable variable (molecular descriptor) space that explains the maximum multidimensional variance direction in the predicted variable (CF) space. That provides a linear regression model based on the observable variables to predict the predicted variable. We also choose to compare two machine learning methods adapted to non-linear problems: the random forest (Breiman 2001) and the support vector machines (SVM) (Drucker et al. 1996). Random forests are a machine learning method, for classification or, in our case, regression, that operate by constructing a multitude of decision trees that uses a random subset of the training data and limits the number of variables used at each split and outputting the mean prediction (regression) of the individual

trees. SVM constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space in which the problem is linearly separable.

These choices allow us to compare several ideas. The PLS is a simple linear method that will not exhibit good performances if the underlying relationship is not linear. The SVM and RF methods are well-known non-linear machine learning algorithms that used to show good results in this kind of problem (Hou et al., 2020a).

All the models were computed in the freeware R (R core team, 2019). The PLS has been computed using the package mixOmics (Rohart et al., 2017), the random forests using the package randomForest (Liaw et al., 2002), and the SVM using the package e1071 (Meyer et al., 2019). These 3 modelling methods have some parameters that needed to be fixed: the number of latent components for the PLS (fixed using the tune.pls function), the number of variables randomly sampled as candidates at each split for the random forests (selected using the tune.randomForest function) and, for the SVM, the gamma parameter of the radial kernel and the cost of constraints violation (using the tune.svm function). All these different tune functions are based on cross-validation (i.e. a training/test procedure to find the best value for the parameters) using default function values.

2.4. Clustering-based model

A recent popular way to make predictions is to use a cluster-then-predict approach. That is, clustering is used for pre-classification which is to arrange a given collection of input patterns into natural meaningful clusters. Then, the clustering results are used to construct a predictor in each cluster. The main idea of the cluster-then-predict approach is that if the clustering performs well the prediction will be easier by modelling only similar compounds. If a new compound with no CF_{ET} and/or CF_{HT} is investigated, the clustering can easily be applied to it before the prediction model itself. The cluster-then-predict approach has already been applied

with success in various domains such as sentiment prediction (Sony et al., 2015), finance (Tsai et al., 2014), chemometrics (Minh Maï Le et al., 2018). So we decided to use the clustering given by the TyPol application (more details in Servien et al., 2014) based on the whole database and the molecular descriptors. Note that the TyPol clustering has already been shown relevant on various occasion: in combination with mass spectrometry to categorize tebuconazole products in soil (Storck et al., 2016), to explore the potential environmental behaviour of putative chlordecone transformation products (Benoit et al., 2017) or to classify pesticides with similar environmental behaviours (Traore et al., 2018). So, the clustering procedure of TyPol was applied on the whole database of 526 compounds using the 40 molecular descriptors. This approach provides us a global clustering based on all the available information contained in the TyPol database. It is based on PLS, hierarchical clustering and an optimal choice of the number of clusters and is detailed in Servien et al. (2014). The obtained clustering is given in Supplementary Figure S1 and relies on 5 different clusters.

Based on this clustering, we then define three other competing methods. For these methods, a different model (with different parameters) will be derived for the compounds in each cluster. Consequently, six different models will be calibrated and tested for each CF prediction: global PLS, global SVM, global random forest, cluster-then-PLS, cluster-the-SVM and cluster-then-Random Forest.

2.5. Comparison procedure

To assess the performances of the different models we will use the following procedure:

1. Split each cluster between a training set (85% of the dataset) and a test set (15%). The test set is not used for any step of the procedure (such as the imputation of the missing data, the calibration of the parameters …).
2. Imputation of the NA values (less than 1%) in the descriptor matrix using the NIPALS algorithm (Wold, 1985).

3. Tune the parameters and train the specific models by performing cross-validation on the training set. We have 3 global models to train (PLS, random forest, and SVM) and the cluster-then-test models (PLS, random forest and SVM for each cluster).
4. Test the different models on the test set. Compute the absolute error.
5. Back to step 1.

The Typol clustering focused on the common compounds is plotted in Supplementary Figure S2 and we see that, for cluster 5, the 3 global models are the only ones available as we can't define a cluster-then-test model due to a lack of data (only one compound remaining). The whole algorithm is repeated 200 times. All the performances are compared in terms of absolute error. The absolute error is the absolute difference between the prediction and the true value. It has been shown to be the most natural and unambiguous measure of error (Willmott et Matsuura, 2005) and is chosen to be easily comparable to the assumed error on the experimental CFs (2-3 logs, see Rosenbaum, 2008). For each cluster, we chose the model with the lowest median absolute error.

Then, the best model is calibrated and computed on the whole cluster. Finally, it is applied to the compounds, according to their clusters, with a CF_{ET} (or a CF_{HT}) equals to NA to provide a prediction. For the compounds in cluster 5, this best model cannot be a cluster-then-predict one and, by consequence, is a global one. To assess the robustness of our prediction we derive a 95% prediction interval for each prediction. The type of model and its corresponding parameters are fixed during this process, according to the best model of the cluster. For example, if the best model of cluster 1 was the random forest approach, random forest models are used with the parameters optimized during the previous step. Then, we perform a leave-one-out bootstrap on the dataset that was used to compute the model (the whole dataset if the model is global, only the data lying in the dedicated cluster if that is a cluster-then-predict model) and a new model is computed on this leave-one-out sample. A prediction is carried for

each leave-one-out model and the 2.5% and 97.5% quantile of these predictions are computed and considered as the prediction interval (Hou et al., 2020a).

The five descriptors contributing the most to the prediction are then derived for each chosen model to assess the differences between models and to interpret their relevance. For a random forest model, these descriptors are calculated using variable permutations (Breiman, 2001), for the SVM they are the descriptors with the higher coefficients in absolute value.

3. Results

3.1. Descriptive analysis of the intersection of the TyPol and the USEtox® databases

As the objective of this proof-of-concept study is to predict USEtox® CF_{ET} and CF_{HT} using the molecular descriptors contained in TyPol, we could only use the compounds that are present in both databases. This results in 274 compounds that are detailed in Table S1 in supplementary material and the range of their CF_{ET} and CF_{HT} values are summarized in the boxplots in Figures 1 and 2. Note that for the 274 common compounds there are 15 NA values for the CF_{ET} and 102 for the CF_{HT}.

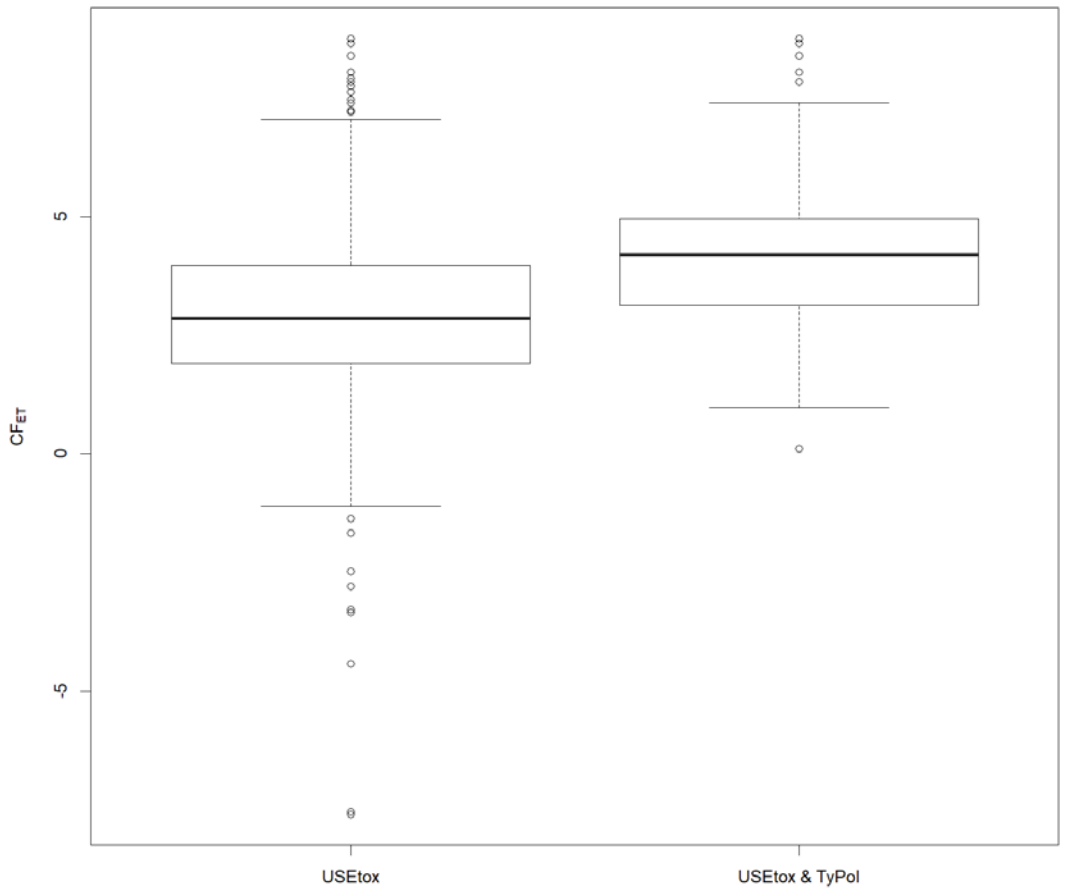


Figure 1- Boxplots of the CF_{ET} for the USEtox® database and the common molecules between the USEtox® and the TyPol databases. This CF_{ET} is equal to the $\log_{10}(PDF.m^3.d.kg^{-1})$.

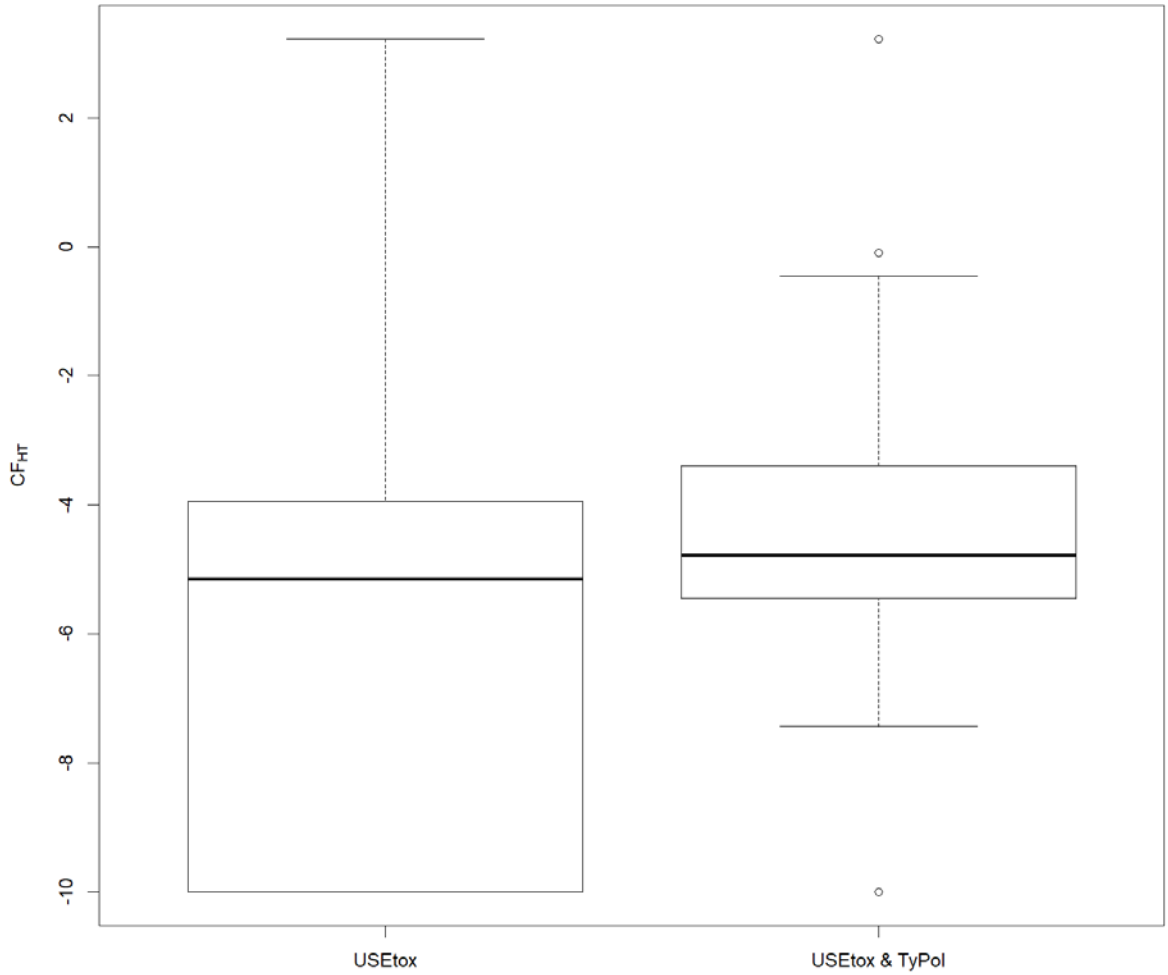


Figure 2 Boxplots of the CF_{HT} for the USEtox® database and the common molecules between the USEtox® and the TyPol databases. This CF_{HT} is equal to $\log_{10}((DALY+\varepsilon).kg^{-1})$. The ε is needed as some values of the DALY are exactly equal to zero. ε has been chosen equal to 1e-10 to be below the minimum of the USEtox® database (5e-9).

We could see on these two figures that the common compounds present higher CF_{ET} and CF_{HT} values than the one of the complete USEtox® database: it focuses on the more dangerous compounds as their boxplots are above the USEtox® counterparts.

The Typol clustering focused on the common compounds is plotted in Supplementary Figure S2 and the boxplots of each molecular descriptor per cluster are given in Supplementary Figure S3 with different indicators in Table S2. We could see that they are clustered in 5 groups with different sizes (respectively 33 compounds in the first black cluster, 122 compounds in the second red cluster, 91 compounds in the third green cluster, 27 compounds in the fourth blue cluster, and one compound in the fifth brown cluster). Cluster 1 grouped compounds with a high number of aromatic bonds, double bonds, rotatable bonds, and multiple bonds. Cluster 2 is an intermediate one between clusters 1 and 3, with less extreme values. Cluster 3 is made of compounds with the lowest molecular mass. Cluster 4 gathered compounds presenting a high number of halogens, rings, and circuits. The unique compound in the fifth cluster is erythromycin (highest molecular mass and number of H and C, lowest number of rings) and, obviously, no cluster-then-predict model could be built for this cluster

As a first analysis of the clustering given by TyPol, we could see in Figure 3 below the boxplots of the CF_{ET} and CF_{HT} within the 5 clusters.

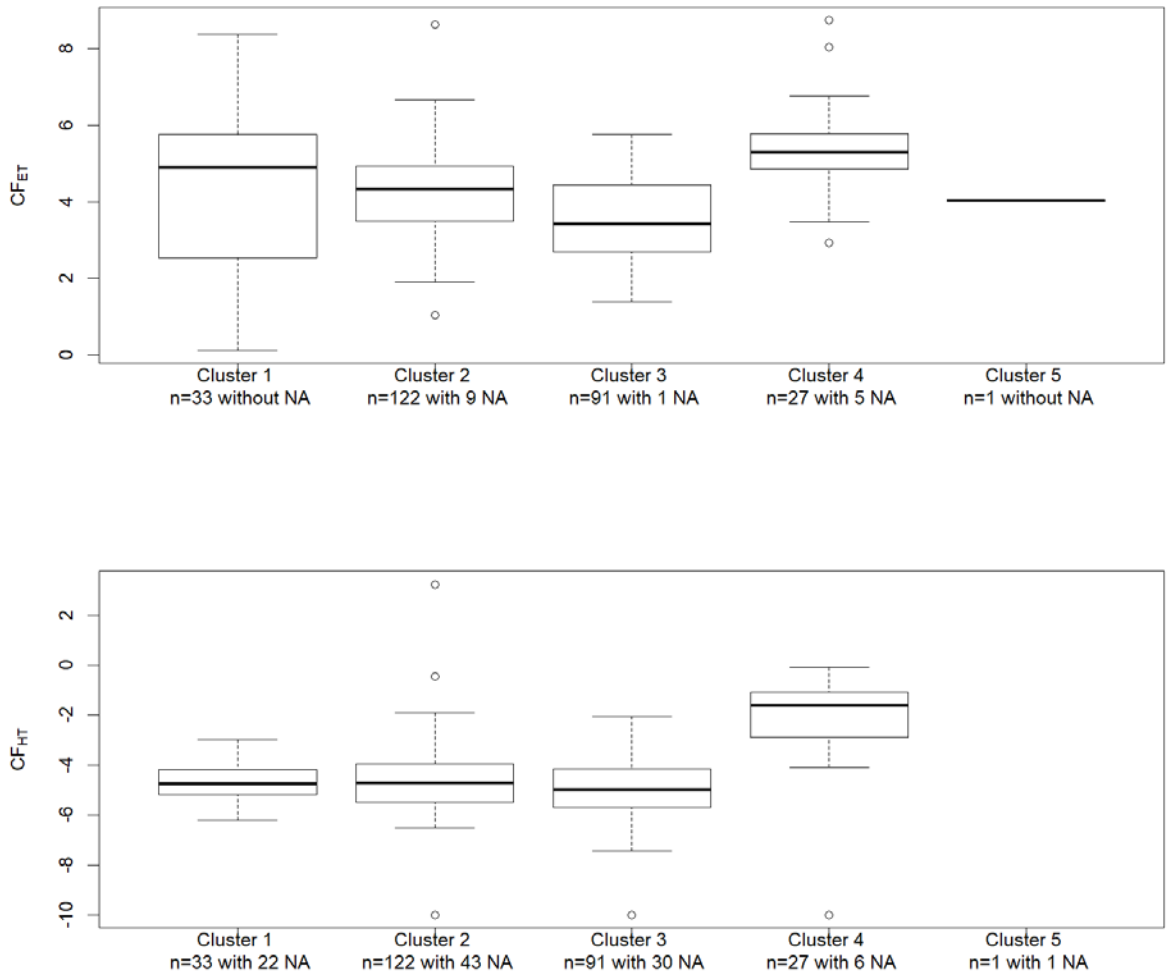


Figure 3- Boxplot by cluster for the CF_{ET} and CF_{HT} values. Note that the unique compound of Cluster 5 has no CF_{HT} value. The size of the clusters and the numbers of NA are gathered in the legend.

The predictions will be made difficult for the CF_{ET} of cluster 1 as it covers a wide range whereas it includes a relatively small number of compounds. On the contrary, cluster 3 covers a small range with no extreme values and includes a high number of compounds, for this cluster the cluster-then-predict approach could produce interesting results.

3.2. Models and prediction of the CF_{ET}

3.2.1. Performances of the machine learning methods

The methodology described in the previous section was applied to our dataset and gave the results gathered in Figure S4 for the global results and in Figure 4 for the results detailed on each cluster.

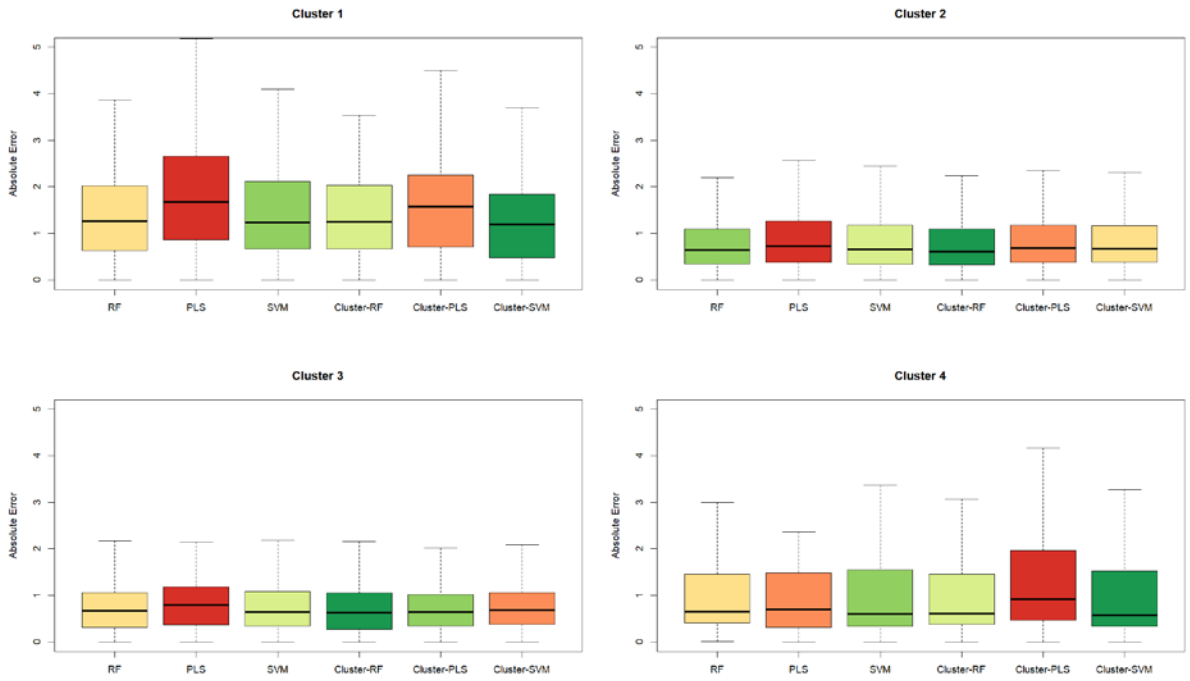


Figure 4 - Performances of the different methods in terms of the log of the absolute error of the CF_{ET} with respect to the different clusters. In each cluster, the models are coloured from green (best) to red (worst) according to their median of the absolute error.

The performances are not similar in each cluster. For example, performances of all methods for cluster 1 are very poor (median absolute error above 1) whereas performances for cluster 4 seem good despite its smallest size (median absolute error around 0.6). So, a future prediction of an unknown compound which lies in cluster 1 will be less reliable than in other clusters. Note that we could not test this in the next section as no NA value is present in this cluster 1.

The cluster-then-predict methods seem more appropriate in each cluster. The cluster-then-RF approach has the best performances (with a global median absolute error equals to 0.64 and the best performances on clusters 2 and 3), even if there is not a big difference between the different methods. The cluster-then-SVM is also the best method for the two clusters 1 and 4. The linear methods (PLS and cluster-then-PLS) have higher absolute errors but are competitive. The individual predictions of the best method in each cluster are reported in Figure S5.

3.2.2. Prediction with the best model

Then we apply the best model in each cluster: a cluster-then-predict approach using SVM for clusters 1 and 4 and using random forest for clusters 2 and 3. To compare the different models in each cluster and give an idea of what are the important molecular descriptors we provide the five most important molecular descriptors for each cluster in the following table.

Table 1- The five most important molecular descriptors for each best model for each cluster. The most important descriptors are in the first line of the table.

Cluster 1: cluster-then-SVM model	Cluster 2: cluster-then-RF model	Cluster 3: cluster-then-RF model	Cluster 4: cluster-then-SVM model
HOMO energy	Number of Chlorine atoms	Number of triple bonds	Number of double bonds
Molecular surface area	Number of halogen atoms	Molecular mass	Number of Nitrogen atoms
Number of Sulfur atoms	Number of Oxygen atoms	Number of Phosphorus atoms	HOMO energy
Connectivity index chi-5	Molecular mass	Number of Oxygen atoms	Number of triple bonds

Connectivity index chi-3	Number of bonds	Number of halogen atoms	Electric dipole moment

We could see in this Table that the important molecular descriptors strongly differ from one cluster to another, highlighting the usefulness of the cluster-then-predict approaches.

Then the models were used to predict the missing CF_{ET} of the common compounds between USEtox® and TyPol databases. These values are by consequence new estimations of the CF_{ET} for compounds on which we have no information. The prediction intervals are relatively small: less than 0.5 $\log_{10}$ in a log scale which highlights the robustness of the estimation. They are given in Table S3. No NA value was present in cluster 1 with no prediction for this cluster. For cluster 2 gathering molecules with intermediate molecular mass, 9 CF_{ET} values were predicted for various kinds of compounds. One value concerns the antibiotic sulfamethazine and its value is quite near to the one of sulfamethoxazole and sulfadiazine of the same sulphonamide antibiotic family constituted of the sulphonamide group ($-S(=O)_2-NR_2R_3$). Cluster 3 grouped compounds with the lowest molecular mass and the lowest median CF_{ET} like ibuprofen, phthalates, cresol constituted of monoaromatic ring substituted with methyl, carboxylic groups. The CF_{ET} prediction for acetylsalicylic acid seemed coherent with the value of the nearest compounds (herbicides mecoprop) of this group. Cluster 4 gathered compounds with the highest median CF_{ET} and that presented a high number of rings halogenated or not, like PAH and hormones. The 5 CF_{ET} predicted concerned 4 PAHs and 1 hormone. By comparison to the 2 other PAHs present in this cluster, the 4 predicted CF_{ET} are quite similar and higher. Concerning the prediction for the hormone, the CF_{ET} is intermediate between the CF_{ET} of the 3 other hormones in the cluster. It seems that all these 5 predicted values are very closed, falling near the median value of this cluster.

3.3. Models and prediction of the CF_{HT}

3.3.1. Performances of the methods

Let us recall that we have more NA values for the CF_{HT} (102) than for the CF_{ET} (15). The performances of the methods are illustrated in the following figure.

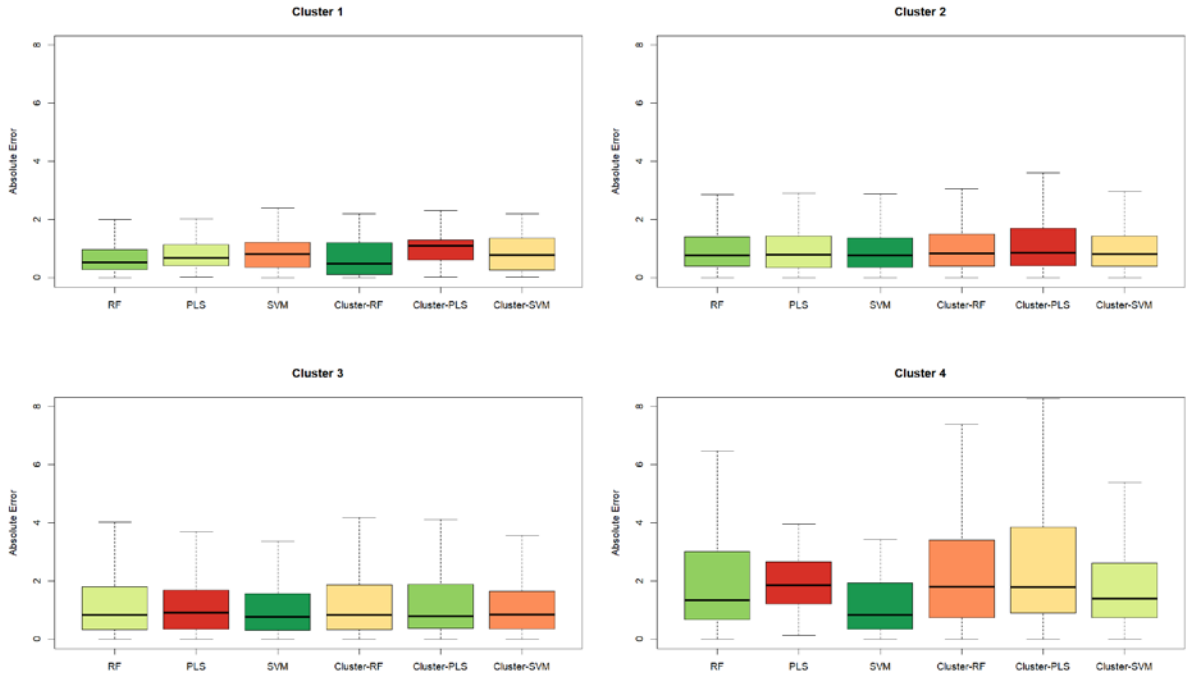


Figure 5- Performances of the different methods in terms of the log of the absolute error of the CF_{HT} with respect to the different clusters. In each cluster, the models are coloured from green (best) to red (worst) according to their median of the absolute error.

We observe that, despite its small size (11 compounds), the CF_{HT} of the first cluster are well predicted (with the best performance for the cluster-then-RF approach). It could be explained by the small range of the CF_{HT} values of this cluster, as illustrated on the boxplot in Figure 3. The performances of all the methods are comparable on clusters 2 and 3 where the best method is the SVM. Cluster 4 seems to be the more difficult to predict: all the methods have their worst results on this cluster and, if the SVM has an acceptable median absolute error of

0.82, all the medians of the other methods are above 1.3. Global performances of the different methods are given in Supplementary Figure S6. Note that, as for CF_{ET}, the linear methods based on PLS are outperformed by the other ones.

3.3.2. Prediction with the best model

The global SVM model was then calibrated and computed on the whole dataset. It was then used to predict the compound of clusters 2, 3, 4, and 5. Let us recall that there is a lonely molecule in cluster 5 and, as it has a NA value for its CF_{HT}, the best global model (SVM) is used. For cluster 1, a cluster-then-RF model is computed. The more important descriptors of these two models are gathered in the following table.

Table 2- Five most important molecular descriptors for each best model for each cluster. The most important descriptors are in the first line of the table.

Cluster 1 : cluster-then-RF model	Cluster 2, 3, 4 and 5: SVM model
Number of Fluorine atoms	Number of halogen atoms
Connectivity index chi-5	Electric dipole moment
Connectivity index chi-1	Number of double bonds
Number of circuits	Number of Chloride atoms
Number of rings	Number of Oxygen atoms

Then, this model was used to predict the CF_{HT} value for the 102 common compounds without a CF_{HT} value. These predictions are reported in Supplementary Table S4. As for the CF_{ET}, the small width of the prediction interval (less than a $\log_{10}$ in a log scale) highlights the robustness of the approach even with a relatively small number like estimations made for compounds that lie in cluster 1. In this cluster 1, CF_{HT} for a phthalate (DEHP) is already known, but the one for

diisodecyl and diisononyl phthalate was predicted with value in the same range. The 3 cyclines (tetracycline, aureomycin, and oxytetracycline) present in cluster 1, presented also similar predicted CF_{HT}. This was also the case for triclosan and triclocarban in cluster 2. Similar predicted and known CF_{HT} were found for four herbicides from the substituted urea family (linuron, diuron, monolinuron, isoproturon) in cluster 3. Cluster 4 gathered a small number of molecules but with the highest median CF_{HT}, the predicted CF_{HT} of the organochlorine insecticide isodrin was similar to another congener of the same family, aldrin.

4. Discussion

It is a real and important challenge to provide characterization factors for a wide range of compounds. Obviously, it is expected that these new calculated factors have an acceptable margin of error. As reported in UNEP/SETAC (2019), it is commonly assumed that the uncertainty of the characterization factors can vary by approximately 2-3 orders of log-magnitude (Rosenbaum et al. 2008) or significantly higher (up to 7 orders) if all sources of uncertainty are considered (Douziech et al. 2019). Using our methodology, we can exhibit a median absolute error of 0.62 log for the prediction of the CF_{ET} and 0.75 log for the prediction of the CF_{HT}. These results are very promising as they are below the level of uncertainty commonly assumed and as they are based on molecular descriptors that could be easily obtained for each compound without ecotoxicity factor. Based on this fact we could already provide 15 new CF_{ET} and 102 new CF_{HT} for the common molecules between USEtox® and TyPol without a previous value.

The idea of predicting ecotoxicity characterization factors for chemicals using machine learning algorithms has already been used (Hou et al., 2020a and 2020b). But, here, our findings go further. Indeed, we show that we could directly obtain accurate estimations of endpoint values from easy-to-obtain molecular descriptors. This will open the door to the fast characterization of each new unknown compound that appears, including transformation

products. We also show that the cluster-then-predict approach can give better performances than the usual ones. This local approach confirms that local models could be an efficient prediction method when heterogeneity of data generates nonlinear relations between the response and the explicative variables (Lesnoff et al., 2020).

Across the clusters and models, there is a general trend that the non-linear models tend to outperform the linear ones. This suggests that a linear model is not fully adequate to capture the complexity of the relationship between the molecular descriptors and the CFs. However, the use of linear model for e.g. a QSAR is likely due to the ease of interpreting its coefficients, while interpretation is much more challenging for machine learning approaches such as random forest or SVM. Thus, the advantages or drawbacks of linear/non-linear approaches must be balanced according to the final goal of each study. Here, as the main goal is to calculate the most accurate CFs, non-linear models seem more suited. We must also mention that a new emerging field is developing tools needed to help make black-box models (e.g. random forest) more interpretable (Bénard et al., 2021).

The difficult interpretability of the machine learning models used in this study can thus be viewed as a limitation. On another side, even if we already had an acceptable number of compounds on our training datasets, the model accuracies would benefit of the inclusion of new compounds. These compounds could be carefully chosen to improve the models where there is a clear need (i.e. where the performances of the models are not good enough), for example in the cluster 1 for CF_{ET} or in the cluster 4 for CF_{HT}.

One of the interests of USEtox® and its three-step structure (fate - exposure - effect) is that it can be adapted to some specific contexts (a more accurate and spatialised fate model, a different exposure...) while keeping the steps that are not modified. However, these adaptations of USEtox® are not widely used and are reserved for advanced users. Our approach does not allow this, with a direct one-step estimation of CFs. It was designed to provide default CF values for molecules where information is missing. We have chosen to directly predict the CF by simplicity, as the first tests revealed that doing three models (for the three steps) and then calculating the CFs produced less accurate results. It would however be

an interesting perspective to estimate only some of the stages by these learning approaches and to combine them with stages modelled in a classical way in USEtox®

5. Conclusion

In a recent study, Aemig et al. (2021) studied the potential impacts on Human health and aquatic environment of the release of 286 micropollutants (organic and inorganic) at the scale of France. One of their conclusion was that, due to a lack of characterization factors, these impacts could be assessed only for 1/3 of these molecules. This paper fills this gap by providing a new modelling method to derive characterization factors from easily obtainable molecular descriptors. The results presented here show that models that can handle non-linearity and that could be adapted to a small number of compounds (using the cluster-then-predict approaches) are the best suited. By consequence, the missing characterization factors, as well as those of new molecules, could now be quickly estimated with an overall good precision. More generally, one of the key factors in the evaluation of toxicity and ecotoxicity in LCA lies in the construction of the characterization factors: a task requiring a large amount of data and a consequent investment of time. The use of machine learning allows us to go beyond these constraints. This makes it possible to obtain characterization factor values in a fast and simple way, which can be used as long as conventionally established CFs are not available.

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.

Acknowledgments

The authors are grateful to Pierre Benoit, Laure Mamy, and Virginie Rossard for their work on TyPol.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Supplementary materials

Supplementary material associated with this article can be found, in the online version.

References

Aemig, Q., Hélias, A., Patureau, D., 2021. Impact assessment of a large panel of organic and inorganic micropollutants released by wastewater treatment plants at the scale of France, Water Research, 188, 116524, https://doi.org/10.1016/j.watres.2020.116524.

Bénard, C., Biau, G., da Veiga, S., Scornet, E, 2021. Interpretable random forests via rule extraction. In International Conference on Artificial Intelligence and Statistics, vol. 130 of Proceedings of Machine Learning Research, 937–945 (PMLR, 2021).

Benfenati, E., Manganaro, A., Gini, G.C., 2013. VEGA-QSAR: AI Inside a Platform for Predictive Toxicology. CEUR Workshop Proceedings, 21-28.

Benoit, P., Mamy, L., Servien, R., Li, Z., Latrille, E., Rossard, V., Bessac, F., Patureau, D., Martin-Laurent, F., 2017. Categorizing chlordecone potential degradation products to explore

their environmental fate, Science of the Total Environnement, 574, 781–795. https://doi.org/10.1016/j.scitotenv.2016.09.094.

Breiman, L., 2001. Random Forests, Machine Learning, 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Cortes, C., Vapnik, V., 1995. Support-vector networks, Machine Learning, 20 (3), 273–297. https://doi.org/10.1007/BF00994018.

Douziech, M., Oldenkamp, R., van Zelm, R., King, H., Hendriks, A.J., Ficheux, A.-S., Huijbregts, M.A.J., 2019. Confronting variability with uncertainty in the ecotoxicological impact assessment of down-the-drain products, Environment International, 126, 37-45, https://doi.org/10.1016/j.envint.2019.01.080.

Drucker, H., Burges, C.C., Kaufman, L., Smola, A.J., Vapnik, V., 1997, Support Vector Regression Machines, Advances in Neural Information Processing Systems 9, NIPS, 155–161, MIT Press. https://dl.acm.org/doi/10.5555/2998981.2999003.

DTU, 2015. Danish QSAR database. Danish QSAR group, National Food Institute, Technical University of Denmark.

Finkbeiner, M., Inaba, A., Tan, R., Christiansen, K., Klüppel, H.-J., 2006. The New International Standards for Life Cycle Assessment: ISO 14040 and ISO 14044. The International Journal of Life Cycle Assessment, 11 (2), 80–85. https://doi.org/10.1065/lca2006.02.002.

He, J., Tang, Z., Zhao, Y., Fan, M., Dyer, S. D., Belanger, S. E., Wu, F., 2017. The combined QSAR-ICE models: practical application in ecological risk assessment and water quality

criteria, Envronnemental Science & Technology, 51, 8877. https://doi.org/10.1021/acs.est.7b02736.

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, The International Journal of Life Cycle Assessment, 16, pp. 701-709 https://doi.org/10.1007/s11367-011-0294-6.

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

Hou, P., Jolliet, O., Zhu, J., Xu, M., 2020a. Estimate ecotoxicity characterization factors for chemicals in life cycle assessment using machine learning models. Environment International, 135, 105393. https://doi.org/10.1016/j.envint.2019.105393.

Hou, P., Zhao, B., Jolliet, O., Zhu, J., Wang, P., Xu, M., 2020b. Rapid Prediction of Chemical Ecotoxicity Through Genetic Algorithm Optimized Neural Network Models, ACS Sustainable Chemistry & Engineering, 8 (32), 12168-12176. https://dx.doi.org/10.1021/acssuschemeng.0c03660.

Lesnoff, M., Metz, M., Roger, JM., 2020. Comparison of locally weighted PLS strategies for regression and discrimination on agronomic NIR data, Journal of Chemometrics, 34(5), e3209, https://doi.org/10.1002/cem.3209.

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest, R News, 2(3), 18-22. http://CRAN.R-project.org/doc/Rnews/.

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.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2019. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-2. https://CRAN.R-project.org/package=e1071.

Minh Maï Le, L., Kégl, B., Gramfort, A., Marini, C., Nguyen, D., Cherti, M., Tfaili, S., Tfayli, A., Baillet-Guffroy, A., Prognon, P., Chaminade, P., Caudron, E., 2018. Optimization of classification and regression analysis of four monoclonal antibodies from Raman spectra using collaborative machine learning approach, Talanta, 184, 260-265, https://doi.org/10.1016/j.talanta.2018.02.109.

National Research Council, 2007. Toxicity Testing in the 21st Century: A Vision and a Strategy; National Academies Press, https://doi.org/10.17226/11970.

R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org/index.html.

Rohart, F., Gautier, B., Singh, A., Le Cao, K.-A., 2017. mixOmics: An R package for omics feature selection and multiple data integration, PLoS computational biology, 13(11), e1005752. https://doi.org/10.1371/journal.pcbi.1005752.

Rosenbaum, R.K., Margni, M., Jolliet, O., 2007. A flexible matrix algebra framework for the multimedia multipathway modelling of emission to impacts, Environment International, 33(5),624-634. https://doi.org/10.1016/j.envint.2007.01.004.

Rosenbaum, R.K., Bachmann, T. M., Gold, L. S., Huijbregts, M.A.J., Jolliet, O., Juraske, R., Koehler, A., Larsen, H.F., MacLeod, M., Margni, M., McKone, T.E., Payet, J., Schuhmacher, M., van de Meent, D., Hauschild, M.Z., 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 (7), 532–546. https://doi.org/10.1007/s11367-008-0038-4.

Saouter, E., Biganzoli, F., Ceriani, L., Versteeg, D., Crenna, E., Zampori, L., Sala, S., Pant, R., 2020. Environmental Footprint: Update of Life Cycle Impact Assessment Methods – Ecotoxicity freshwater, human toxicity cancer, and non-cancer, Publications Office of the European Union, Luxembourg, https://doi.org/10.2760/300987.

Servien, R., Mamy, L., Li, Z., Rossard, V., Latrille, E., Bessac, F., Patureau, D., Benoit, P., 2014. TyPol - a new methodology for organic compounds clustering based on their molecular characteristics and environmental behaviour, Chemosphere, 111, 613–622. https://doi.org/10.1016/j.chemosphere.2014.05.020.

Soni, R., Mathai, K.J., 2016. An Innovative 'Cluster-then-Predict' Approach for Improved Sentiment Prediction. In: Choudhary R., Mandal J., Auluck N., Nagarajaram H. (eds) Advanced Computing and Communication Technologies. Advances in Intelligent Systems and Computing, vol 452. Springer, Singapore. https://doi.org/10.1007/978-981-10-1023-1_13.

Storck, V., Lucini, L., Mamy, L., Ferrari, F., Papadopoulou, E.S., Nikolaki, S., Karas, P.A., Servien, R., Karpouzas, D.G., Trevisan, M., Benoit, P., and Martin-Laurent, F, 2016.

Identification and characterization of tebuconazole transformation products in soil by combining suspect screening and molecular typology, Environmental Pollution, 208 B, 537-545. https://doi.org/10.1016/j.envpol.2015.10.027.

Traore, H., Crouzet, O., Mamy, L., Sireyjol, C., Rossard, V., Servien, R., Latrille, E., Martin-Laurent, F., Patureau, D., Benoit, P., 2018. Clustering pesticides according to their molecular properties, fate and effects by considering additional ecotoxicological parameters in the TyPol method, Environmental Science and Pollution Research, 25(5), 4728-4738. https://doi.org/10.1007/s11356-017-0758-8.

Tsai, C.-F., 2014. Combining cluster analysis with classifier ensembles to predict financial distress, Information Fusion, 16, 46-58. https://doi.org/10.1016/j.inffus.2011.12.001.

UNEP-SETAC, 2019. Global Guidance for Life Cycle ImpactAssessment Indicators: Volume 2. https://www.lifecycleinitiative.org/training-resources/global-guidance-for-life-cycle-impact-assessment-indicators-volume-2/ (accessed Nov 22, 2020).

USEtox® 2020: USEtox® database system, https://USEtox®.org/model/download.

Verones, F., Bare, J., Bulle, C., Frischknecht, R., Hauschild, M., Hellweg, S., Henderson, A., Jolliet, O., Laurent, A., Liao, X., et al., 2017. LCIA Framework and Cross-Cutting Issues Guidance within the UNEP-SETAC Life Cycle Initiative, Journal of Cleaner Production, 161, 957–967. https://doi.org/10.1016/j.jclepro.2017.05.206.

Willmott, C., Matsuura, K., 2005. Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance, Climate Research, 30, 79. https://doi.org/10.3354/cr030079.

Wold, H., 1985. Partial least squares, In Kotz, Samuel; Johnson, Norman L. (eds.), Encyclopedia of statistical sciences, vol 6, New York, Wiley.

Xia, M., Huang, R., Witt, K.L., Southall, N., Fostel, J., Cho, M.-H., Jadhav, A., Smith, C.S., Inglese, J., Portier, C.J., Tice, R.R., Austin, C.P., 2008. Compound cytotoxicity profiling using quantitative high-throughput screening, Environmental Health Perspectives*, 116* (3), 284–291, https://doi.org/10.1289/ehp.10727.