Machine learning models based on molecular descriptors to predict toxicological and ecotoxicological characterization factors

Rémi Servien, Eric Latrille, Dominique Steyer Patureau, Arnaud Hélias

Machine learning models based on molecular descriptors to predict toxicological and ecotoxicological characterization factors

Rémi Servien[1,2], Eric Latrille[1,2], Dominique Patureau[1], Arnaud Hélias[3,4]

[1]INRAE, Univ. Montpellier, LBE, 102 Avenue des étangs, F-11000 Narbonne, France
[2]ChemHouse Research Group, Montpellier, France Univ Montpellier,
[3]ITAP, INRAE, Institut Agro, Montpellier, France
[4]ELSA, Research group for environmental LCSA and ELSA-Pact industrial chair, Montpellier, France
E-mail contact: remi.servien@inrae.fr

1. Introduction

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, Life Cycle Assessment (LCA) offers a structured, operational, and standardized methodological framework [1]. For human toxicity and freshwater ecotoxicity, USEtox® [2], was developed to produce a transparent and consensus characterization model. Unfortunately, to determine the characterization factors (CF) of a molecule, numerous physicochemical parameters (such as solubility, hydrophobicity, degradability) and detailed toxicological and ecotoxicological data must be provided. Obtaining these data is challenging, expensive and time-consuming. This remains an issue for number of existing and ever-increasing numbers of new chemicals, which leaves their impacts largely unknown. That is why more computational models are needed to complement experimental approaches, to decrease the experimental cost and to prioritize 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 [3] …) and are used to predict ecotoxicological data (LC50 for example). Recently, machine learning algorithms have been used to predict hazardous concentration 50% (HC50) based on 14 physicochemical characteristics [4] or on 691 more various variables [5]. But, the input variables of these models are not always only molecular descriptors that could be easily collected for any newly available compounds and/or the output variables are not directly the CFs but only prior parameters. To address this aim, we will test different methods (linear and non-linear) to build robust models that could directly predict lacking CFs (for ecotoxicological CF_{ET} and for human impact CF_{HT}) in continental freshwater, based on easy-to-obtain molecular descriptors.

2. Materials and methods

The different models were built and tested using the USEtox® database, namely the corrective release 2.12 [2], with the default landscape. The TyPol database [6] was used to collect 40 molecular descriptors on the different compounds. These easy-to-obtain molecular descriptors were constitutional (number of atoms …), geometric (Conolly molecular surface area), topological (connectivity indexes …) or quantum-chemical (Total energy …). A total of 274 compounds was in common between TyPol and USEtox®. These 274 compounds were used to build the model and assess their performances.

To predict the CFs using the molecular descriptors, we chose to compare one linear method (PLS) and two machine learning non-linear ones: the random forest and the support vector machines (SVM) [7]. We also tested clustering-then-predict approaches: a first clustering was performed on the whole TyPol database and then 3 models (PLS, RF and SVM) were derived for each cluster. By consequence, on each cluster we add 6 competing methods: 3 global and 3 cluster-then-predict ones. The different models were tested using a classical train-and-test approach by computing the absolute errors of the prediction on the test set.

3. Results and discussion

3.1. Models and prediction of the CF_{ET}

The performances of the different methods were assessed in the different clusters. The cluster-then-predict machine learning based methods were shown to have the best performances, highlighting the need of non-linear local models. The medians of the absolute errors were below one log which can be considered as an acceptable margin of error. One example of the different performances is provided in the Figure 1.

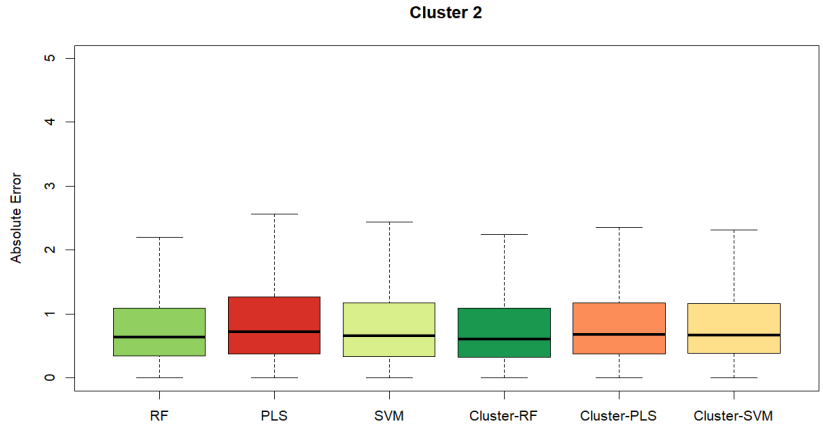

Figure 1: Log of the absolute errors of the different methods to predict CF_{ET}.

3.2. Models and prediction of the CF_{HT}

For the CF_{HT}, the same good performances were observed, mainly for machine learning global methods. One cluster was more difficult to predict than others, but overall performances were considered as good. Note that, as for CF_{ET}, the linear methods based on PLS are outperformed by the non-linear ones. More detailed results could be found in [7].

4. Conclusions

These results are very promising as the performances of the predictions of the machine learning approaches are below the level of uncertainty commonly assumed for these CF values (1 log) and as they are based on molecular descriptors that could be easily obtained for each compound without any ecotoxicity factor. 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. These predictive models were then used to complement the impact assessment of micropollutants release from wastewater treatment plants [8] where a lack of CFs was observed in a previous study [9]. It could be noticed that this machine learning predictive strategy could be applied to any other compartment and/or characterization factors, provided that a sufficiently large learning database already exists.

5. References

[1] Finkbeiner et al. 2006. The New International Standards for Life Cycle Assessment: ISO 14040 and ISO 14044. Int J Life Cycle Assess, 11(2):80–85. doi.
[2] USEtox®, 2020. USEtox® database system, https://USEtox®.org/model/download.
[3] DTU. 2015. Danish QSAR database. National Food Institute, Technical Univ of Denmark.
[4] Hou et al., 2020. Estimate ecotoxicity characterization factors for chemicals in life cycle assessment using machine learning models. Environ Int, 135, 105393. doi.
[5] Hou et al., 2020. Rapid Prediction of Chemical Ecotoxicity Through Genetic Algorithm Optimized Neural Network Models. ACS Sustainable Chem Eng, 8 (32), 12168-12176. doi.
[6] Servien et al. 2014. TyPol – A new methodology for organic compounds clustering based on their molecular characteristics and environmental behaviour, Chemosphere, 111, 613–622. doi.
[7] Servien et al. 2022. Machine learning models based on molecular descriptors to predict human and environmental toxicological factors in continental freshwater. Peer Community Journal, 2:e15. doi
[8] Servien et al. 2022. Machine learning to improve the impact assessment of micropollutants release from WWTP. Case Studies in Chemical and Environmental Engineering, 5, 100172. Open access link.
[9] Aemig et al. 2021. Impact assessment of a large panel of organic and inorganic micropollutants released by wastewater treatment plants at the scale of France. Water Res, 188: 116524, doi