Using machine learning to fill data gaps in the toxicity characterization of chemicals for life cycle assessment

Tianran Ding[1], Gustavo Larrea-Gallegos[1], Antonino Marvuglia[1], Thomas Schaubroeck[1]

1. Introduction

With the increasing number of marketed chemicals, understanding their effect on different dimensions of sustainability has become a priority. Since chemicals can be found along the different stages of the supply chains of most products and services, it is important to understand their sustainability from an holistic perspective. For this purpose, Life-Cycle Assessment (LCA) has been commonly selected as the suitable tool to quantify the impacts associated with the life-cycle of a product or chemical. The mapping between environmental flows and the affectation to the environment is done through the use of coefficients known as characterization factors (CF). While CFs are developed to address different types of environmental impacts, in this study we focused on toxicity-related impacts.

In the literature, different methods were proposed to develop toxicity-related CFs for numerous chemicals, such as USEtox and Environmental Footprint (EF) (Saouter et al., 2018). Nevertheless, not all chemicals are characterised in these methods. For instance, USEtox provides CFs for around 3000 chemicals and a recent update of this method (i.e., EF version 3) expanded the chemical coverage to around 6000. However, the provided CFs are still not enough to cover the whole range of chemicals, especially the ones newly developed. In these methods, calculating new CFs requires collecting various data concerning chemical properties to calculate factors related to fate, exposure, and effect. However, some of these data are not always available since they traditionally come from experimental tests that are cost- and time-consuming, confidential or non-transparent, and could face legislation restrictions on in-vivo tests on animals.

To address this, recent literature has adopted the use of data-driven approaches such as machine learning (ML) to predict parameters or metrics that are later fed into the USEtox method, such as fate and intake factors (Marvuglia et al., 2015), species sensitivity distribution (SSD) (Song et al., 2022), hazard concentration 50% (HC_{50}) (Hou et al., 2020a, 2020b; Li et al., 2024). These works rely on the use of USEtox database to train their models, but they have not been updated or improved using novel toxicity datasets, such as the one provided by the new version of the EF 3 method (Saouter et al., 2018).

The aim of this study is to fill these gaps by exploring different ML algorithms to produce off-the-shelf models of toxicity prediction. For this, the models are trained to predict hazard concentration 20% (HC_{20}) of chemicals using only SMILE representation as input, so it can be used later in the pipeline of the novel EF 3 methodology. The novelty of our

[1] Luxembourg Institute of Science and Technology, tianran.ding@list.lu

study relies on the use of a new dataset that contains almost two times more data entries if compared to the formerly used USEtox dataset.

2. Materials and methods

Based on the different types of approaches found in the literature, we have identified two main methodological pathways that are commonly explored (see Figure 1). In this study we selected the first pathway that looks towards predicting HC_{20}. The predicted HC_{20} are then used to calculate the effect factor (EfF) that measures the potentially affected fraction (PAF) of exposed masses in the freshwater ecosystem in the EF method.

HC_{20} is predicted using two algorithms: the eXtreme Gradient Boosting (XGboost) and Gaussian processes. Molecular descriptors were obtained from SMILE labels using an open source cheminformatics library (i.e., RDKit) and HC_{20} were collected from the EF dataset. The final dataset contains 5540 observations with one predicted variable (i.e., HC_{20}) and 256 characteristics as initial predictors.

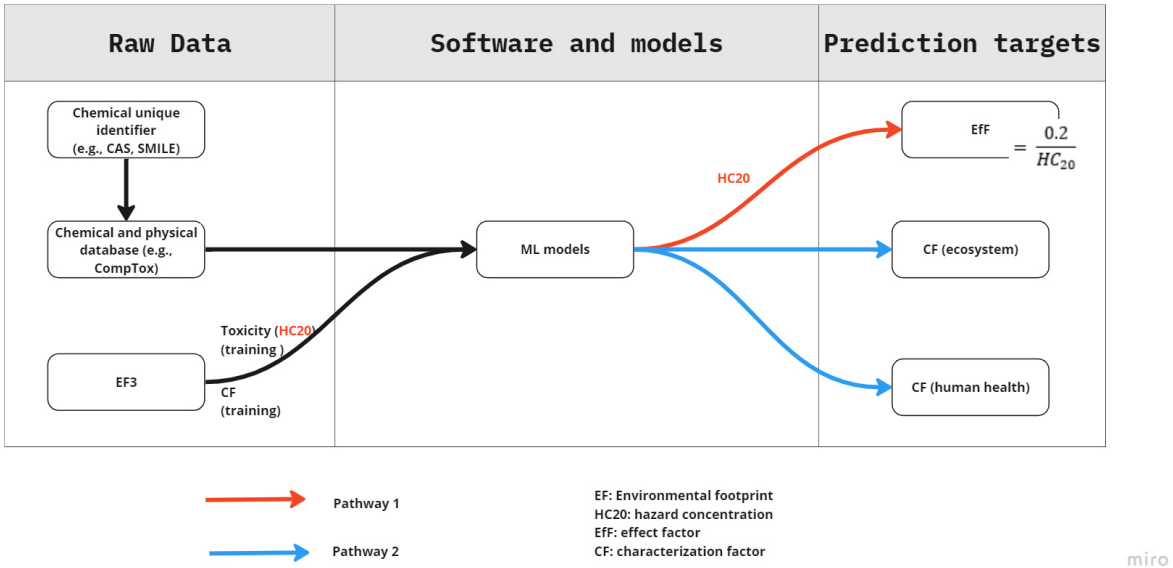

Figure 1. Illustrations of two methodological pathways common in literature. This work follows pathway 1.

The data were split into training (70%) and testing (30%) sets. Three-fold cross-validation was used to choose the best hyperparameter combination from a grid search. The best model was then used to predict the test set using the coefficient of determination (R^2) as indicator of performance of the model.

3. Results

In our preliminary results, the XGBoost model yielded an R^2 of 0.46 (see Figure 2), which sets a new benchmark for this training dataset. With respect to the Gaussian Process, a similar R^2 of 0.47, suggests that this model has a similar capacity of capturing the underlying complexity of the dataset.

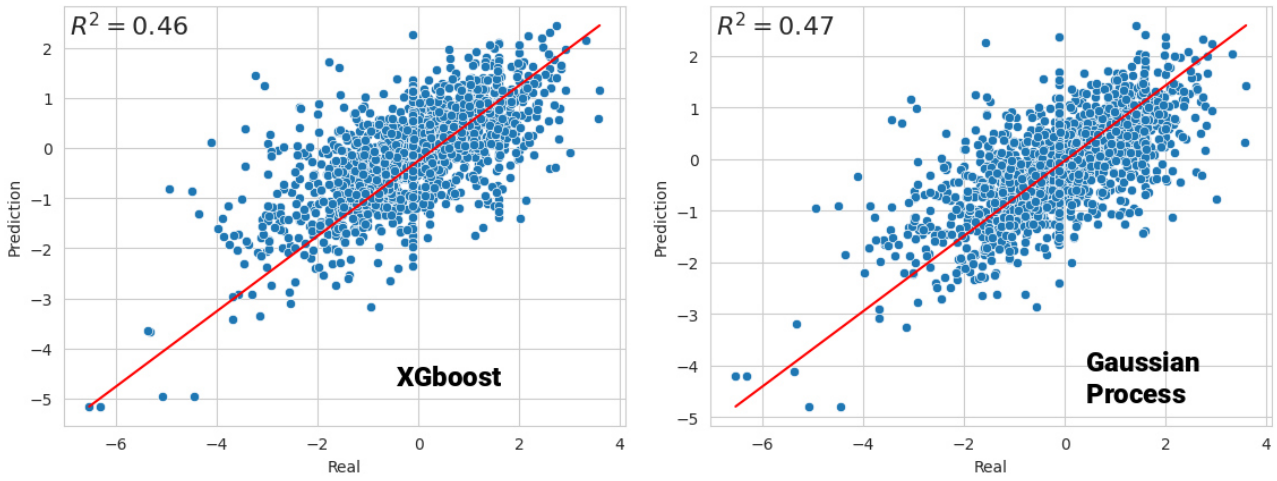

Figure 2. Performance of predictions obtained from the XGBoost and Gaussian Process model using coefficient of determination as indicator.

4. Conclusions

This study embarked on a required quest to harness ML in bridging data gaps inherent in the toxicity characterization of chemicals, a pressing need amid the burgeoning array of chemicals in the market. Through the adept application of ML algorithms, namely XGBoost and Gaussian processes, this research is working on predicting chemical toxicity more accurately, leveraging on the latest EF data and methodologies.

While preliminary, our results indicate that a further improvement of the algorithms is required. This may suggest that further research should be oriented to the adoption of more sophisticated algorithms such as deep learning. Finally, further steps in this study will include the exploration of the methodological pathway 2, which would imply the delivery of a completely off-the-shelf tool to be used in LCA studies.

5. Acknowledgements

This study is part of the CALIMERO project, funded by the European Union with the grant number: 101060546.

6. References

Hou, P., Jolliet, O., Zhu, J., Xu, M., 2020a. Estimate ecotoxicity characterization factors for chemicals in life cycle assessment using machine learning models. Environ. Int. 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 Sustain. Chem. Eng. 8, 12168–12176. https://doi.org/10.1021/acssuschemeng.0c03660

Li, D., Qin, J., Hong, J., 2024. Toward a comprehensive life cycle aquatic ecotoxicity assessment via machine learning: Application to coal power generation in China. J. Clean. Prod. 445, 141373. https://doi.org/10.1016/j.jclepro.2024.141373

Marvuglia, A., Kanevski, M., Benetto, E., 2015. Machine learning for toxicity characterization of organic chemical emissions using USEtox database: Learning the structure of the input space. Environ. Int. 83, 72–85. https://doi.org/10.1016/j.envint.2015.05.011

Saouter, E., Biganzoli, F., Ceriani, L., Versteeg, D., Crenna, E., Zampori, l, Sala, S., Pant, R., 2018. Environmental Footprint: Update of Life Cycle Impact Assessment Methods – Ecotoxicity freshwater, human toxicity cancer, and noncancer.

Song, R., Li, D., Chang, A., Tao, M., Qin, Y., Keller, A.A., Suh, S., 2022. Accelerating the pace of ecotoxicological assessment using artificial intelligence. Ambio 51, 598–610. https://doi.org/10.1007/s13280-021-01598-8