Peer Community In Ecotoxicology & Environmental Chemistry

Predicting characterization factors of chemical substances from a set of molecular descriptors based on machine learning algorithms

Sandrine CHARLES *based on reviews by Sylvain Bart, Patrice Couture, Dominique Lamonica and 2 anonymous reviewers*

A recommendation of:

Open Access

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 *(2022), bioRxiv, 2021.07.20.453034, ver. 6 peer-reviewed and recommended by Peer Community in Ecotoxicology and Environmental Chemistry*

https://doi.org/10.1101/2021.07.20.453034

Published: 18 January 2022

Submitted: 21 July 2021, Recommended: 17 January 2022

Cite this recommendation as:
Sandrine CHARLES (2022) Predicting characterization factors of chemical substances from a set of molecular descriptors based on machine learning algorithms. *Peer Community In Ecotoxicology and Environmental Chemistry, 100001.* ***https://doi.org/10.24072/pci.ecotoxenvchem.100001***

Recommendation

Today, thousands of chemical substances are released into the environment because of human activities. It is thus crucial to identify all relevant chemicals that contribute to toxic effects on living organisms, also potentially disturbing the community functioning and the ecosystem services that flow from them. Once identified, chemical substances need to be associated with ecotoxicity factors. Nevertheless, getting such factors usually requires time-, resources- and animal-costly experiments that it should be possible to avoid. In this perspective, modelling approaches may be particularly helpful if they rely on easy-to-obtain information to be used as predictive variables.

Within this context, the paper of Servien et al. (2022) illustrates the use of machine learning algorithms to predict toxicity and ecotoxicity factors that were missing for a collection of compounds. Their modelling approach involve a collection of molecular descriptors as input variables. A total of 40 molecular descriptors were extracted from the TyPol database (Servien et al., 2014) as those describing the best how organic compounds behave within the environment. These molecular descriptors also have the advantage to be easily quantifiable for new chemical substances under evaluation. The performances of the proposed models were systematically checked and compared to the classical linear partial least square method, based on the calculation of the absolute error (namely, the difference between prediction and true value). This finally led to different best models (that is associated to the lowest median absolute error) according

to the classification of the 526 compounds comprised in the TyPol database in five clusters. These five clusters of different sizes gather chemical substances with different but specific molecular characteristics, also corresponding to different estimates of the characterization factors both in their median and within-variability.

In a final step, predictions of characterization factors were performed for 102 missing values in the USEtox® database (Rosenbaum et al., 2008) but also referenced in TyPol. This paper highlights that the molecular descriptors that explain the most the toxicity of the chemical substances in each cluster strongly differ. Nevertheless, these predictions, whatever the cluster, appear precise enough to be considered as relevant despite everything.

As a conclusion, this paper is a promising proof-of-concept in using machine learning modelling to go beyond some constraints around the toxicity evaluation of chemical substances, especially handling non-linearities and data-demanding calculations, in in an ever-changing world that is gradually depleting its resources without sufficient concern for the short-term risks to the environment and human health.

References

Rosenbaum RK, Bachmann TM, Gold LS, Huijbregts MAJ, Jolliet O, Juraske R, Koehler A, Larsen HF, MacLeod M, Margni M, McKone TE, Payet J, Schuhmacher M, van de Meent D, Hauschild MZ (2008) USEtox—the UNEP-SETAC toxicity model: recommended characterisation factors for human toxicity and freshwater ecotoxicity in life cycle impact assessment. The International Journal of Life Cycle Assessment, 13, 532. https://doi.org/10.1007/s11367-008-0038-4

Servien R, Latrille E, Patureau D, Hélias A (2022) Machine learning models based on molecular descriptors to predict human and environmental toxicological factors in continental freshwater. bioRxiv, 2021.07.20.453034, ver. 6 peer-reviewed and recommended by Peer Community in Ecotoxicology and Environmental Chemistry. https://doi.org/10.1101/2021.07.20.453034

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 behavior. Chemosphere, 111, 613–622. https://doi.org/10.1016/j.chemosphere.2014.05.020

Reviews

Evaluation round #2

DOI or URL of the preprint: *https://doi.org/10.1101/2021.07.20.453034*

Version of the preprint: 4

Author's Reply, None

Download author's reply

Decision by Sandrine CHARLES, 04 Jan 2022

Dear authors,

Thank you very much for your revised version of this manuscruipt that accounts for all suggestions given by the different reviewers. This finally led to an improved version that I am almost ready to recommend given still a very minor revision based on suggestions that I added directly on your point-by-point reply, here attached. Considering this should not take a lot of time and can be discussed or ignored. These are only suggestions for your consideration.

Best regards,

Sandrine Charles

Download recommender's annotations

Evaluation round #1

DOI or URL of the preprint: *10.1101/2021.07.20.453034*

Version of the preprint: 3

Author's Reply, 22 Nov 2021

Download author's replyDownload tracked changes file

Dear recommender,

We thank the reviewers very much for their constructive comments. Major revisions were done as required, and a detailed response to the reviewer comments, that carefully addresses, point-by-point, the issues raised in the comments, is provided attached. We hope that you will find the changes satisfactory and that this revised manuscript will be now considered for recommendation in PCI Ecotoxicology & Environmental Chemistry. Please note that the new version of the paper (and of the supplemental material) has been uploaded at BiorXiv. We are at your disposal if you need any further information. Thank you very much in advance for your attention.

Best regards,

Rémi Servien on behalf of co-authors.

Decision by Sandrine CHARLES, 22 Nov 2021

Dear authors,

First accept our apologize for the long duration of the review process regarding your paper. It took us a lot of time to find reviewers, the first two we got were not specialized enough into modelling to make us able to render a decision. We finally got three additional reviews that should help you in improving your manuscript in order to provide us with a revised version. Please provide this revision together with a point-by-point answer to reviewers' comments refering to the corresponding changes in your manuscript. Changes in your revised manuscript must be hightlighted to be clearly identifed compared to the previous version. If possible, please provide your revised manuscript and your answers on December the 21st 2021 at the latest.

Best regards,

S. Charles

Reviewed by Sylvain Bart, 23 Aug 2021

Servien et al presents a new method based on machine learning to predict ecotoxicological metrics for chemicals for which we don’t have these metrics. The approach is promising and complementary to the linear QSAR method which cannot deal with nonlinearity.

The graphical abstract is very informative and the introduction provides all the necessary information to understand the topic and the scientific gap addressed. All the methods and procedures are deeply described which is very appreciated for reader whom machine learning is not the primary expertise, like me.

In conclusion, the manuscript is well written, I don't see any major issue in the manuscript, and I would recommended it for publication in a peer reviewed journal

minor comment:

-I would suggest to carefully check all figure captions to ensure all necessary informations are given for the figures to be read by themselves. E.g. : Figure 4, Provide full name somewhere for RF, PLS etc.. ?

All the best

Reviewed by Patrice Couture, 27 Aug 2021

I would not provide an in-depth review of this manuscript, due to my very limited expertise in the area of the paper (I am an ecotoxicologist). This paper needs to be properly reviewed by experts in modeling. I only identified a few points that would need to be addressed to improve the clarity and the relevance of ecotoxicological terms like LC50 (see file attached).

I consider that the topic addressed in this paper is interesting and the approach proposed is promising. Overall, this work has the potential to provide very useful tools for environmental and human risk assessment of new chemicals that will reduce costs, time and use of live organisms.

Download the review

Reviewed by Dominique Lamonica, 11 Nov 2021

Download the review

Reviewed by anonymous reviewer, 21 Oct 2021

The paper frames itself in a line of research initiated by other researchers and pursued also by the same authors in previous works, i.e. the use of machine learning to predict human and environmental toxicity of chemicals (using the USETox database, but not only).

The application described in this paper is just another confirmation of the potential of this kind of approach.

The paper is rather well written, although it appears too concise in the description of the full path of modelling that was followed. In this sense, to facilitate the understanding of the model chain, I suggest inserting a clear flowchart or a figure like Fig. 1 in Hou et al. 2020 (Estimate ecotoxicity characterization factors for chemicals in life cycle assessment using machine learning models. Environment International, 135, 105393) or Fig. 1 in Marvuglia et al. 2013 (Machine learning for toxicity characterization of organic chemical emissions using USEtox database: learning the structure of the input space. Environment International 83: 72-85).

Besides these two articles, other exist on similar applications in the literature, that have not been cited in this manuscript. They authors might want to take a look at them to improve their state of the art:

- Marvuglia et al. 2014. Variables selection for ecotoxicity and human toxicity characterization using Gamma Test. In: B. Murgante et al. (Eds.): ICCSA 2014, Part III, LNCS 8581, pp. 640–652, 2014. Proceedings of the 14th International Conference on Computational Science and Applications (ICCSA 2014), University of Minho, Guimaraes, Portugal.

- Marvuglia et al. 2015. Random Forest for toxicity of chemical emissions: features selection and uncertainty quantification. Journal of Environmental Accounting and Management 3(3): 229-241;

- Song et al. 2017. Rapid Life-Cycle Impact Screening Using Artificial Neural Networks. Environ. Sci. Technol. 2017, 51, 10777−10785.

- Wu and Wang 2018. Machine Learning Based Toxicity Prediction: From Chemical Structural Description to Transcriptome Analysis. Int. J. Mol. Sci. 2018, 19, 2358; doi:10.3390/ijms19082358.

- Lysenko et al 2018. An integrative machine learning approach for prediction of toxicity-related drug safety. https://doi.org/10.26508/lsa.201800098.

- Song et al. 2021. Accelerating the pace of ecotoxicological assessment using artificial intelligence. Ambio. https://doi.org/10.1007/s13280-021-01598-8

At page 11, when the clustering protocol is described, it is not clear to me how the clustering is chosen. The authors mention that the whole algorithm is repeated 200 times. However, this is not a deterministic procedure and at each iteration a (slightly or not) different partitioning can come up. Therefore, a criterion of cluster quality is needed. For example, in hierarchical clustering, not always the cut height that determines how many clusters to choose, is clear. If I understand correctly, the error criterion that the authors use, pertain only to the evaluation of the forecasting capacity of the models to determine the two factors CFET and CFHT, but nothing is said on how to chose the best clustering partition. There are many cluster validity measures (see e.g. Vazirgiannis M. (2009) Clustering Validity. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_616).

At page 11, line 18, the term NA appears but it is not explained in the paper. It is only explained in the caption of table S2 in supporting information. I think it should also be explained in the text of the paper.

At page 21, line 2-3 read as follows: "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". This is true, but the important molecular descriptors (and the ranking of the descriptors overall) differ not only because we change from one cluster to another, but also because the best model changes from one cluster to the other. Therefore, how can we say that the important descriptors change only because of the cluster? To estimate how much of this change in ranking depends on the cluster and how much on the model used, the authors should provide the full ranking in each cluster for each model. Then one could calculate for example the change in raking position for each variable within the same cluster when passing from one model to the other.

In table 2, it is not clear how the descriptors are selected. Is it possible to add the % of variance of the output explained by each descriptor?

At page 24, the lines from 6 to 11 of the Conclusions are more fit for the introduction, rather than for the conclusions. I suggest moving this part there.

Suggested changes to the text:

- Page 3, line 11: begin the sentence with “therefore” rather than with “so”.

- Page 3, lines 23-24 from “To best” to “case-by-case basis”: this sounds like a repetition of something already mentioned above.

- Page 5, line 8: change “That’s why” with “That is why”.

- Page 6, line 28: change “that are” with “that is”.

- Page 9, line 26: add a comma after “performs well”.

- Page 10, line 18: correct “cluster-the-SVM” in “cluster-then-SVM”.

- Page 17, line 11: change “in each cluster” to “from one cluster to another”. The meaning changes, and I think my suggestion reflects better what you want to say.

- Page 17, line 13: begin the sentence with “therefore” rather than with “so”.

- Page 20, line 16: change “the more difficult” with “the most difficult”.

- Page 21, line 8: change “lonely” with “single”.

- Page 21, line 10: change “the more important” with “the most important”.

- Page 23, line 4: although also the cited paper (Lesnoff et al., 2020) uses the term “explicative”, I believe a more common term in statistics and machine learning is “explanatory”.

Reviewed by anonymous reviewer, 03 Nov 2021

The first impression reading the paper is that it contains some naïf considerations. The authors insist on the novelty of using non linear methods; those methods are in use since about 20 years, both in QSAR and many more modeling tasks. Using a non-linear method is the good practice today when simple linear methods fail.

So the novelty of the paper is not in choosing tools that are already accepted in QSAR; it can be in the idea of computing the characterization factors (CFs) using molecular descriptors instead of relying on the traditional LCA methods that depend on data (chemical, toxicological, etc.) not easily available for every chemical.

The authors compute 40 molecular descriptors (including some quantum chemical descriptors), selected since they appear relevant to describe the behavior of organic compounds in the environment. Then they apply both classifiers (using 3 modeling methods) and clustering, defining different local models for the 5 different clusters.

A point that should need more attention is the descriptor selection. In any modeling method (machine learning included) the features are important and a wider exploration of the features and their number is missing in the paper.

The combination of the classifiers with clustering is interesting in that the results can be more accepted by the users, which often like to consider also the compounds similar to the one under investigation.

As the authors report, USEtox® is commonly used; it provides in one single CF the chemical fate, the exposure, and the effect for each compound in a set of several thousands chemicals. Then the CF can be extended to other endpoints, both human and environmental (DALY and PDF). The observation that the computation of those final endpoints can be done in one model using directly the chemical information is the advantage of the proposed method over the traditional one.

In conclusion, even though the methods applied are quite common in QSAR, and the machine learning methods should be better applied, the paper proposes something new in the LCA domain.