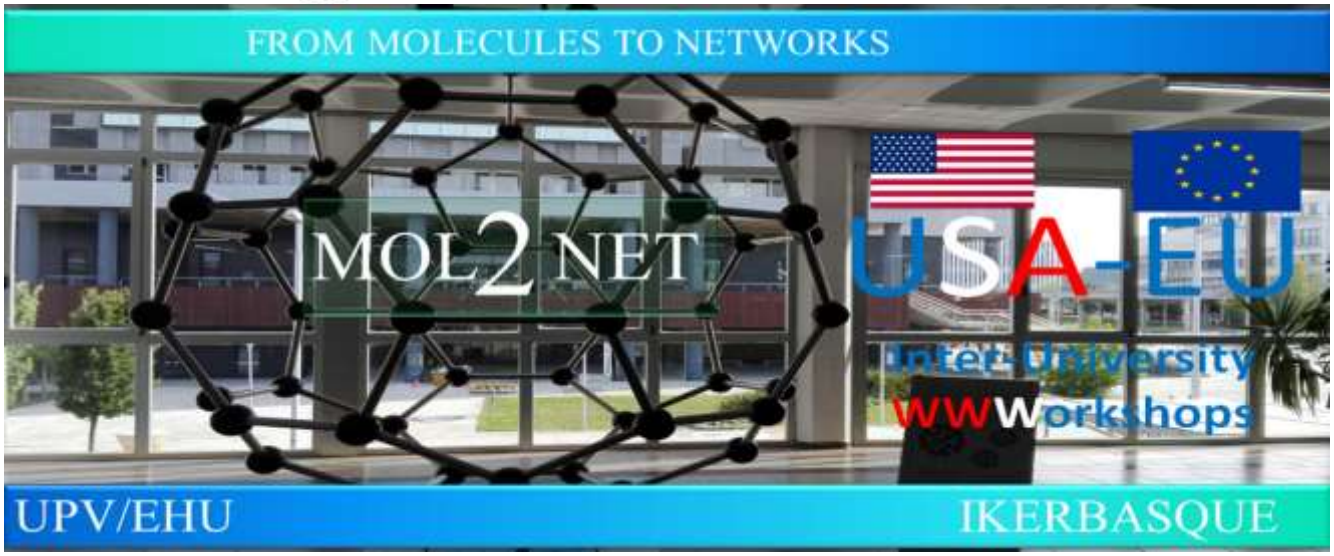
MDPI sciforum

MOL2NET, International Conference Series on Multidisciplinary Sciences



Machine learning-based prediction of toxicity of pesticide towards *Americamysis bahia*

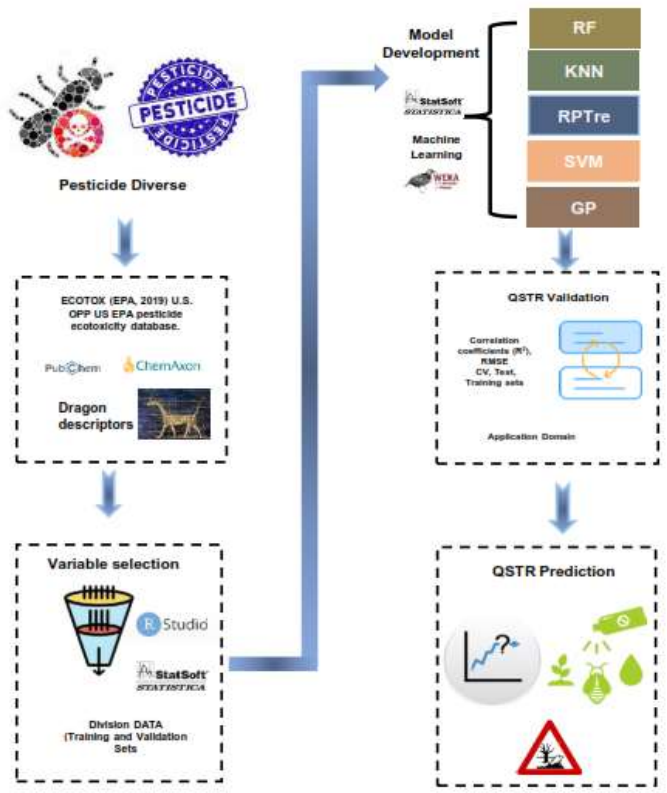
Manuel Mesías Nachimba-Mayanchi [a], *Karel Diéguez-Santana* [b]

[b] *Departamento Ciencias de la Vida, Universidad Estatal Amazónica, Paso Lateral Vía Napo, km 2.5, 160150, Puyo, Pastaza, Ecuador*

[b] *Department of Organic and Inorganic Chemistry, University of Basque Country UPV/EHU, 48940 Leioa, Spain*

.

Graphical Abstract	**Abstract.**
	Pesticides are toxic substances designed and widely applied throughout the world. However, their widespread use has received increasing attention from regulatory agencies due to the various acute and chronic effects they have on various organisms. In this study, QSTR (Quantitative Structure-Toxicity Relationship) models, based on nonlinear statistical techniques, have been established using five Machine Learning (ML) algorithms to predict the toxicity of pesticides on mysid shrimp (*Americamysis*



bahia). The optimal nonlinear model (Random Forest, $R^2 = 0.983$) was verified by internal (leave one cross-validation) and external validations. The validation results ($q_{int}^2 = 0.815$ and $q_{ext}^2 = 0.81$) were satisfactory in predicting acute toxicity in the saltwater crustacean (*A. bahia*) compared to other models reported in the literature. In addition, this model also predicted the toxicity of some pesticides without experimental data. With a p(LC50) value of 12.102, bromadiolone was the most toxic compound. It is the active compound of the product RASTOP BLOCKS (rodenticide), which is classified in the category "Extremely Hazardous".

Introduction

The increased use of chemicals in the world due to the development of science and technology has generated great concern about their potential toxicity to aquatic organisms [1]. Toxicity assessment of chemicals is necessary for all chemical industries before they are released into the market [2]. Traditionally, the toxicities of chemicals are obtained from animal tests. However, these toxicological experiments are not only ethically problematic, but also costly, time-consuming, and labor-intensive. A lethal concentration of 50% is used as a quantitative endpoint of toxicity [3].

Quantitative structure-activity relationship / toxicity (QSAR / QSTR) models are an important method for analyzing toxic mechanisms [4,5] and predicting the toxicity of organic chemicals, [6,7] even those that have not been synthesized. From databases of a given toxic property measured experimentally, and in a series of molecular descriptors, QSTR models are obtained that can be used to predict the toxicity of other chemical compounds that were not present in the training database. In silico tools would have a great economic benefit by reducing the experimental cost and speeding up the pesticide risk assessment process [8,9]. The topic introduces a new field in this sphere of knowledge, since it provides a useful tool for environmental risk assessment, through the prediction of the aquatic toxicity of organic compounds. In addition, it can be used for the estimation of toxicological and ecotoxicological properties of known products, thus achieving a rational use of resources in the selection of compounds that are less aggressive to humans and the environment. Many researchers have carried out QSAR studies to predict the toxicity of pesticides in aquatic systems [10,11].

The mysid shrimp (*Americamysis bahia*) is a saltwater crustacean that has been used as a toxicity test organism for endocrine disruptors and shown to be an optimal species for quantifying toxicological effects [12]. In addition, [13] considered that it may be useful in pesticide risk assessment for the European

Union Plant Protection Products Regulation. Measurements by [14] reflect that acute toxicity to estuarine crustaceans can occur even at low concentrations of different pesticides, reflecting the need to regulate the use of these substances and to consider toxicological analyses of sensitive species in the regulations. [14]. The objective of this work is to develop a nonlinear QSTR model for the acute toxicity of 289 pesticides in mysid shrimp (*A. bahia*).

Materials and Methods

Dataset collection

Compounds and experimental information were obtained from the ECOTOX database [15]. and the OPP pesticide ecotoxicity database (OPP, 2019). The variables considered in the database were: experimental medium (salt water), exposure time (96 h) and toxicity measure (median lethal concentration, LC50), which were converted to a negative logarithmic scale, -log LC50 or pLC50. Finally, 289 pesticides against *A. bahia* remained (102 herbicides, 96 insecticides, 59 fungicides, and 31 other pesticides).

Molecular descriptors

The structures of the compounds were generated with ChemAxon and 0-2D molecular descriptors were calculated with the DRAGON software [16]. After eliminating those descriptors that are equivalent to constants (or approximately constants) or whose pairwise correlation coefficients are above 0.90, 801 molecular descriptors were derived from the Dragon software for subsequent descriptor selection.

Weka QSTR-ML Implementations

The Weka -3.8.5 version was used to implement all of the learning algorithms [17]. Because of its extensive array of machine learning (ML) and data mining algorithms, this package is widely used in bioinformatics and chemoinformatics [18]. It is available for free at www.cs.waikato.ac.nz/ml. The REPTree method developed by Quinlan [19], and RF [20] were applied as representations of decision tree. Furthermore other techniques such as SVM [21], Gaussian Process [22], and k Nearest Neighbors (KNN) [23,24] were implemented. The statistical parameters obtained by the fitness of the models were the correlation coefficient (R^2) and Root Mean Square Error (RMSE) [6].

Results and Discussion

Results of Machine Learning Algorithms

To evaluate the performance of each ML model for the inhibition of *A. bahia* growth by the pesticides, the inhibition (LC50) of each compound predicted by the ML models was compared to that determined experimentally. Figure 1 presents an overview of the main statistical parameters calculated for each ML model. Based on the RMSE values estimated for each model, RF presents the best results (RMSE = 0.221) for the training and evaluation series. However, in the cross validation, SVM shows a slightly lower result ($RMSE_{SVM}$ = 0.559 vs $RMSE_{RF}$ = 0.595). Overall, the presented models demonstrate that these ML algorithms could be employed as a fast and useful tool to estimate the mean lethal concentration (LC_{50}) of pesticides. The 5 well-trained ML models were validated for the conditions of

the experimental results obtained from the literature. The results also demonstrated that the selected physicochemical properties of the pesticides could be used to interpret the aquatic toxicity of *A. bahia*.

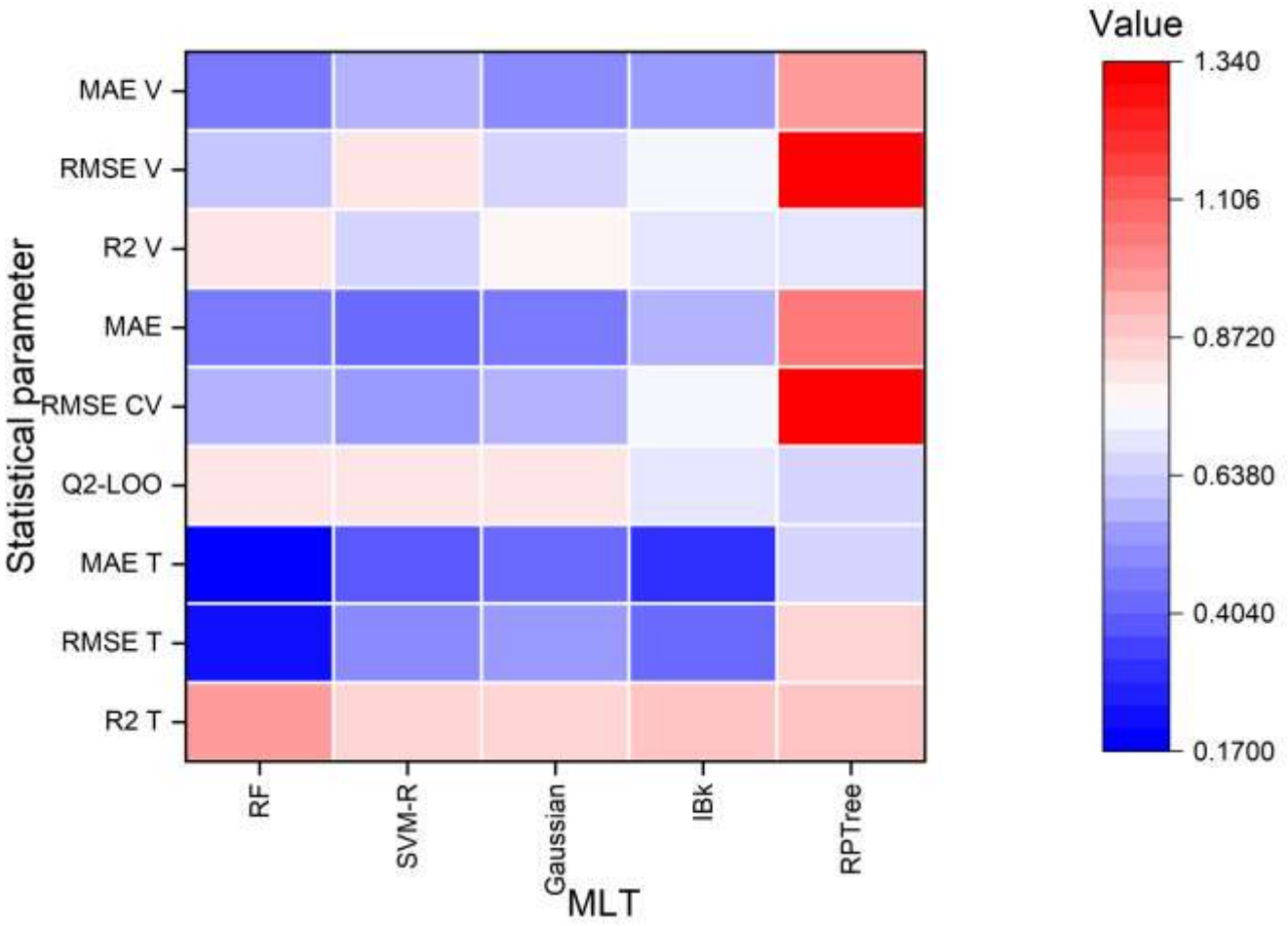


Figure 1. ML performance model results summary

Notes: R^2: Coefficients of determination, RMSE: root mean square error, and MAE: mean absolute error.

.

Conclusions

The QSTR models obtained present a high level of fitness, internal robustness, and external predictivity compared to models presented in other similar studies. Therefore, they can quantitatively determine the toxicity of pesticides in *Americamysis bahia*. The overall performance of the ML regression models was RF > IBK > RPTree SVM > GP, with RF presenting superiority among the models, with the highest values of R^2 = 0.812, and the lowest values of RMSE = 0.595 and MAE = 0.462 in cross-validation, presenting the same scenario in the training series and in the evaluation series. The ecotoxicological potential of 166 chemical compounds that are active ingredients in various pesticide products used worldwide was analyzed quantitatively. The toxicity calculation using the obtained model reveals that Bromadiolone is the most toxic compound, with a value of p(LC50) = 12.102; it is the active compound of the product RASTOP BLOCKS (rodenticide), which is primarily used on corn crops.

References

.

1. Stenzel, A.; Goss, K.-U.; Endo, S. Experimental Determination of Polyparameter Linear Free Energy Relationship (pp-LFER) Substance Descriptors for Pesticides and Other Contaminants: New Measurements and Recommendations. *Environmental Science & Technology* **2013**, *47*, 14204-14214, doi:10.1021/es404150e.
2. Yu, X. Prediction of chemical toxicity to Tetrahymena pyriformis with four-descriptor models. *Ecotoxicology and Environmental Safety* **2020**, *190*, 110146, doi:10.1016/j.ecoenv.2019.110146.
3. Chen, X.; Dang, L.; Yang, H.; Huang, X.; Yu, X. Machine learning-based prediction of toxicity of organic compounds towards fathead minnow. *RSC Advances* **2020**, *10*, 36174-36180, doi:10.1039/D0RA05906D.
4. Kar, S.; Gajewicz, A.; Puzyn, T.; Roy, K.; Leszczynski, J. Periodic table-based descriptors to encode cytotoxicity profile of metal oxide nanoparticles: A mechanistic QSTR approach. *Ecotoxicology and Environmental Safety* **2014**, *107*, 162-169, doi:10.1016/j.ecoenv.2014.05.026.
5. Pramanik, S.; Roy, K. Predictive modeling of chemical toxicity towards Pseudokirchneriella subcapitata using regression and classification based approaches. *Ecotoxicology and Environmental Safety* **2014**, *101*, 184-190, doi:10.1016/j.ecoenv.2013.12.030.
6. Dieguez-Santana, K.; Pham-The, H.; Villegas-Aguilar, P.J.; Le-Thi-Thu, H.; Castillo-Garit, J.A.; Casañola-Martin, G.M. Prediction of acute toxicity of phenol derivatives using multiple linear regression approach for Tetrahymena pyriformis contaminant identification in a median-size database. *Chemosphere* **2016**, *165*, 434-441, doi:10.1016/j.chemosphere.2016.09.041.
7. Heo, S.; Safder, U.; Yoo, C. Deep learning driven QSAR model for environmental toxicology: Effects of endocrine disrupting chemicals on human health. *Environmental Pollution* **2019**, *253*, 29-38, doi:10.1016/j.envpol.2019.06.081.
8. Pham-The, H.; Casañola-Martin, G.; Diéguez-Santana, K.; Nguyen-Hai, N.; Ngoc, N.T.; Vu-Duc, L.; Le-Thi-Thu, H. Quantitative structure–activity relationship analysis and virtual screening studies for identifying HDAC2 inhibitors from known HDAC bioactive chemical libraries. *SAR and QSAR in Environmental Research* **2017**, *28*, 199-220, doi:10.1080/1062936X.2017.1294198.
9. Diéguez-Santana, K.; Rivera-Borroto, O.M.; Puris, A.; Pham-The, H.; Le-Thi-Thu, H.; Rasulev, B.; Casañola-Martin, G.M. Beyond model interpretability using LDA and decision trees for α-amylase and α-glucosidase inhibitor classification studies. *Chemical biology drug design* **2019**, *94*, 1414-1421.
10. Coors, A.; Frische, T. Predicting the aquatic toxicity of commercial pesticide mixtures. *Environmental Sciences Europe* **2011**, *23*, 22, doi:10.1186/2190-4715-23-22.
11. Li, F.; Fan, D.; Wang, H.; Yang, H.; Li, W.; Tang, Y.; Liu, G. In silico prediction of pesticide aquatic toxicity with chemical category approaches. *Toxicol Res (Camb)* **2017**, *6*, 831-842, doi:10.1039/C7TX00144D.
12. Hirano, M.; Ishibashi, H.; Matsumura, N.; Nagao, Y.; Watanabe, N.; Watanabe, A.; Onikura, N.; Kishi, K.; Arizono, K. Acute Toxicity Responses of Two Crustaceans, Americamysis bahia and Daphnia magna, to Endocrine Disrupters. *Journal of Health Science* **2004**, *50*, 97-100, doi:10.1248/jhs.50.97.
13. Brock, T.C.M.; Van Wijngaarden, R.P.A. Acute toxicity tests with Daphnia magna, Americamysis bahia, Chironomus riparius and Gammarus pulex and implications of new EU requirements for the aquatic effect assessment of insecticides. *Environmental Science and Pollution Research* **2012**, *19*, 3610-3618, doi:10.1007/s11356-012-0930-0.
14. DeLorenzo, M.E.; Key, P.B.; Chung, K.W.; Sapozhnikova, Y.; Fulton, M.H. Comparative toxicity of pyrethroid insecticides to two estuarine crustacean species, Americamysis bahia and Palaemonetes pugio. *Environ. Toxicol* **2014**, *29*, 1099-1106, doi:10.1002/tox.21840.
15. OPP. U.S. OPP US EPA pesticide ecotoxicity database. **2019**.
16. Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. *DRAGON—Software for the Calculation of Molecular Descriptors.* , 2007.

17. Frank, E.; Hall, M.; Witten, I. *The WEKA workbench. "Data Mining: Practical Machine Learning Tools and Techniques"*, Fourth Edition ed.; Kaufmann, M., Ed.; 2016.
18. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* **2009**, *11*, 10-18.
19. Quinlan, J.R. Simplifying decision trees. *International Journal of Man-Machine Studies* **1987**, *27*, 221-234, doi:10.1016/S0020-7373(87)80053-6.
20. Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5-32, doi:10.1023/a:1010933404324.
21. Chang, C.-C.; Lin, C.-J. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems technology* **2011**, *2*, 1-27, doi:10.1145/1961189.1961199.
22. Colkesen, I.; Sahin, E.K.; Kavzoglu, T. Susceptibility mapping of shallow landslides using kernel-based Gaussian process, support vector machines and logistic regression. *Journal of African Earth Sciences* **2016**, *118*, 53-64, doi:10.1016/j.jafrearsci.2016.02.019.
23. Nam, N.-H.; Nga, D.-V.; Hai, D.T.; Dieguez-Santana, K.; Marrero-Poncee, Y.; Castillo-Garit, J.A.; Casanola-Martin, G.M.; Le-Thi-Thu, H. Learning from multiple classifier systems: Perspectives for improving decision making of QSAR models in medicinal chemistry. *Current topics in medicinal chemistry* **2017**, *17*, 3269-3288.
24. Dieguez-Santana, K.; M Rivera-Borroto, O.; Puris, A.; Le-Thi-Thu, H.; M Casanola-Martin, G. A Two QSAR Way for Antidiabetic Agents Targeting Using α-Amylase and α-Glucosidase Inhibitors: Model Parameters Settings in Artificial Intelligence Techniques. *Letters in Drug Design & Discovery* **2017**, *14*, 862-868, doi:10.2174/1570180814666161128121142.