Water Research 174 (2020) 115583

Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres

Application of chemometric methods and QSAR models to support pesticide risk assessment starting from ecotoxicological datasets

Francesco Galimberti [a, *], Angelo Moretto [a, b], Ester Papa [c, **]

[a] ICPS, International Centre for Pesticides and Health Risk Prevention, ASST Fatebenefratelli-Sacco, Milan, Italy
[b] Department of Biomedical and Clinical Sciences, Università degli Studi di Milano, Milan, Italy
[c] QSAR Research Unit in Environmental Chemistry and Ecotoxicology, University of Insubria, Varese, Italy

ARTICLE INFO

Article history:
Received 25 October 2019
Received in revised form
10 January 2020
Accepted 1 February 2020
Available online 6 February 2020

Keywords:
Pesticide
QSAR
Ecotoxicology
Endpoint

ABSTRACT

The EFSA '*Guidance on tiered risk assessment for edge-of-field surface waters*' underscores the importance of in silico models to support the pesticide risk assessment. The aim of this work was to use in silico models starting from an available, structured and harmonized pesticide dataset that was developed for different purposes, in order to stimulate the use of QSAR models for risk assessment. The present work focuses on the development of a set of in silico models, developed to predict the aquatic toxicity of heterogeneous pesticides with incomplete/unknown toxic behavior in the water compartment. The generated models have good fitting performances (R^2: 0.75–0.99), they are internally robust (Q^2loo: 0.66–0.98) and can handle up to 30% of perturbation of the training set (Q^2 lmo: 0.64–0.98). The absence of chance correlation was guaranteed by low values of R^2 calculated on scrambled responses (R^2 Y_{scr}: 0.11–0.38). Different statistical parameters were used to quantify the external predictivity of the models (CCC_{ext}: 0.73–0.91, Q^2 ext-Fn: 0.53–0.96).

The results indicate that all the best models are predictive when applied to chemicals not involved in the models development. In addition, all models have similar accuracy both in fitting and in prediction and this represents a good degree of generalization.

These models may be useful to support the risk assessment procedure when experimental data for key species are missing or to create prioritization lists for the general a priori assessment of the potential toxicity of existing and new pesticides which fall in the applicability domain.

1. Introduction

Agriculture occupies a very important place in the European Union (EU) as an economic activity and as a source of food for population. In the last 50 years, it has gone from a national and intensive agriculture to a common and sustainable agriculture [Villaverde et al., 2019]. Pesticides, have a massive role in the agricultural framework, due to the fact that they have always been widely used in agriculture to prevent or control pests, diseases, weeds, and other plant pathogens to reduce or eliminate yield losses and maintain high quality of agricultural products. They are intentionally used to cause adverse effects on target organisms; but adverse effects in non-target organisms may arise as well, as a consequence of the exposure to these products. Currently, the integrated management of pest, diseases and weeds, seeks to maximize crops production and, at the same time, to reduce the impact of pesticides on living beings and on the environment by making better use of them at socially acceptable and economically viable levels [Villaverde et al., 2019]. For this reason, pesticides are developed through very strict, complex and time-consuming regulation processes [Reg.EC 1107/2009] to function with reasonable certainty (reduce the uncertainty) and minimal impact on human health and the environment. In addition, Directive 2009/128/EC [European Commission] promotes the sustainable use of pesticides through the integrated pest management. Pesticide registration is a scientifically-based, legal, and also administrative process, where the potential to cause adverse effects on human health and the environment associated with the use of pesticide products, is assessed by conducting several tests [Reg.EU 283/2013]. Effects in any non-target species may translate into ecosystem unbalance and food-chain disruption that ultimately may affect human health

* Corresponding author. Via G.B. Grassi 74, ICPS, Pad. 17, 20157, Milan, Italy.
** Corresponding author. University of Insubria, Department of Theoretical and Applied Sciences (DiSTA), via J.H. Dunant 3, 21100, Varese, Italy.
E-mail addresses: francesco.galimberti@icps.it (F. Galimberti), ester.papa@uninsubria.it (E. Papa).

https://doi.org/10.1016/j.watres.2020.115583
0043-1354/

and edible species [Damalas et al. 2011]. As mentioned, regulatory bodies rely on extensive *in vivo* and *in vitro* testing to support regulatory decisions on human health and environmental risks. It is a known fact that *in vivo* tests often require a large number of laboratory animal studies which can consume significant amounts of resources in terms of budget and time for testing and evaluation. Alternatives to animal testing [Reg.EU 283/2013] were proposed to overcome some of the drawbacks associated with animal experiments and avoid the unethical procedures. *In silico* methods, based on quantitative structure-activity relationships (QSAR) are worldwide recognized alternatives to animal testing as well as cells and tissue cultures, alternative organisms, lower vertebrates, invertebrates and microorganisms [Ranganatha et al.][Doke et al.][Scholz et al.] [Villaverde et al., 2017]. QSAR can potentially result in significant cost savings during the pesticide safety assessment process, having a great economic impact for the agrochemical sector. The reduction in the necessary laboratory or field evaluations will decrease time and cost, leading to faster commercialization of formulations and thus promoting the productivity and competitiveness of European agriculture. In silico tools are especially important when experimental studies are not adequate because of ethical reasons or because experimental studies are either too complex or not viable [Villaverde et al., 2017] [Zhu et al.].

QSAR models are based on a variety of mathematical approaches to predict activities and properties of untested chemicals on the basis of their molecular structure. QSAR methods have a long history of use both for the design pre-synthesis and the regulatory assessment of pharmaceuticals, pesticides, and other chemicals [TWG on Pesticide 2012]. Therefore, considering the limitations of current testing approaches, the growing public attention for ethical issues related to *in-vivo* tests and breedings, and the rapid development and the advantages of computational predictive methods, companies and Regulatory Agencies have started in the last decade to support the use of QSARs to enhance the efficiency of hazard and risk assessment processes [ECHA] [76/768/EEC]. In particular, in 2007 the European REACH regulation (Regulation Evaluation Authorization of Chemicals) [Reg. EC 1907/2006] promoted the regulatory use of in silico (i.e. models, grouping and read across procedures) and *in vitro* alternatives to animal testing. Since then, specific Guidance Documents and other tools were made available by the Organization for Economic Cooperation and Development (OECD), the European Chemical Agency (ECHA) [ECHA] [OECD 2015] and the EU commission JRC [Triebe et al.], to increase the transparent use of these methods.

In line with the current regulatory approaches mentioned above, the aim of the present work was, to use data collected within a project commissioned by the European Food Safety Authority (EFSA) [Galimberti et al.] to estimate pesticide ecotoxicity values on the basis of their chemical structure. In particular, data measured for aquatic organisms representative of different levels of biological complexity were used for the creation of *ad hoc* QSAR models. These approaches, which are proposed as alternative methods in the Guidance on tiered risk assessment for edge-of-field surface waters [EFSA J. 2013], should serve as quantitative tools to predict the Effect Concentrations (EC_x) of other substances of interest having no experimental toxicity data available. Ecotoxicological studies are designed to identify the adverse effects produced by a substance to selected species and to characterize the dose-response relationship for the adverse effects. The aim of the present work is to maximize the toxicological information available to describe the whole set of 70 pesticides within different taxonomic groups by using chemometric approaches.

2. Material and methods

2.1. Dataset and data control

Within the project "Comparison of NOEC values to EC_{10}/EC_{20} values, including confidence intervals, in aquatic and terrestrial ecotoxicological risk assessment" a collection of data from the ecotoxicological section of 70 pesticide approval dossiers was performed [Galimberti et al.]. Ecotoxicological data of studies from the pesticide approval dossiers have to strictly follow internationally agreed test guidelines, such as OECD or US-OPPTS. These guidelines allow a high standardization of the study performance and consequently the results. Moreover, in the aforementioned collection of data [Galimberti et al.] ecotoxicological studies have been peer reviewed from National Experts in the pesticide assessment procedure and these data have been processed through a statistical modelling analysis performed, in 2015, by the Wageningen University. The pesticides were mainly distributed through the herbicide, fungicide and insecticide pesticide functional classes (respectively: 32, 44 and 19% and 5% of combinations of functional classes). Only studies concerning chronic toxicity were taken into account and in particular on the following taxa: algae, aquatic invertebrates, aquatic plants, birds, earthworms, fish, mammals, non-target plants, soil arthropods and terrestrial arthropods. Different effects were investigated with numerous biologic parameters. Among all the effects, development, growth and reproduction of the different tested organisms were the most frequent in the pool of selected studies. All the data were managed and stored into an MS Access database. The database consisted of 952 single entries collected for different pesticide active substances tested in different taxa, species, effects and biologic parameters (like Cell count, length, weight, young produced.) in addition to information related to time of exposure as well as the number of tested doses, and the dose max. Each substance was characterized by the CAS Registry Number, the molecular formula, and the molecular weight. The single entries were EC_{10}, EC_{20} and EC_{50} values (with their upper and lower confidential limits) and the NOEC value (calculated and reported from the pesticide study tests). A subset of the dataset (only studied aquatic organisms) is reported in supplementary material. To create a strong dataset in order to build robust QSAR models, data (as reported in the supplementary materials) were curated according to the following steps:

- Presence of mixtures and stereoisomeric compounds in the studied samples (pesticides). In general, mixtures cannot be run through QSAR models, nor can synergistic or antagonistic effects of chemicals in mixtures be accounted for because models typically use single, discrete chemical structures as input. Regarding stereoisomeric pesticides, the user should be aware that QSAR predictions generated by using representative 2D structure do not accurately reflect the true 3D conformation of the active ingredients, i.e. they miss information on stereochemistry [TWG on Pesticide 2012].
- Grouping data for data consistency. According to the *'Guidance on tiered risk assessment for plant protection products for aquatic organisms'* [EFSA PPR 2013] and on the basics of statistical homogeneity, data on substances were aggregated by the same taxa, species, effects and biologic parameters on a first attempt of analysis.
- Outliers. An outlier is an observation point that is distant from other observations; it may be due to variability in the measurement or it may indicate experimental error which might be sometimes excluded from the data set. In QSAR, compounds that have unexpected biological activity and are unable to fit in a QSAR model are known as outliers. These are valuable in

defining the limitations under which compounds act by a common molecular mechanism modelled by one or more descriptors, and also in defining the experimental limitations of the biological test data. Thus, particular attention should be payed to the outliers and the reason for their peculiarity be sought [Rajeshwar et al.].

2.2. Molecular descriptors

The SMILE strings available for the 70 active substances were imported into the PaDEL descriptor software [Yap] for the calculation of the molecular descriptors. The SMILE Notation, an acronym which stands for Simplified Molecular Input Line Entry System, i.e. a string notation used to describe the nature and topology of molecular structures, were derived according to the PubChem Project [Pubchem], and where available, these notations were compared with the ones reported in the approval dossiers of the active substances to cross-validate the information acquired. The SMILE Notation was retrieved because is the starting point to calculate the molecular descriptors for the selected active substances. Molecular descriptors [Todeschini et al.], are the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule, into a useful number. Version 2.21 of PaDEL was used; all the 1D, 2D descriptors were calculated and also the PubChem fingerprints. The number of calculated descriptors (1875) with PaDEL software [Yap] was large, and included molecular descriptors and fingerprints in order to have the possibility to represent different features of the chemical structure in different ways. However, a lot of them resulted to be inter-correlated and redundant, giving very similar structural information. For this reason a pre-filtering procedure was applied to the dataset to exclude semi-constant descriptors (no information brought by the variable itself) and to exclude descriptors too inter-correlated (same information brought by many variables). Constant variables were excluded if greater than 80% and inter-correlated variables were excluded if correlated for a percentage greater than 90%. After this procedure, the number of remaining variables was 525. This data curation procedure was performed with QSARINS Software [QSARINS][Gramatica et al., 2013] [Gramatica et al., 2014].

2.3. QSAR modelling

The dataset was prepared to be used as the input for the QSAR model generation i.e. all the responses were converted into μmol/L and transformed into Log (EC_x). Multiple linear regression (MLR) QSAR models by OLS (Ordinary Least Squares) method and Genetic Algorithm for variable selection were generated and validated in the software QSARINS.

Principal Component Analysis (PCA) [Wold S. et al., 1987] was then performed on the curated molecular descriptors, in order to explore the distribution of the data in the chemical and experimental space, and to highlight possible outliers and/or particular clusters/patterns.

The next step was the development of the QSAR models according to the OECD principles for the development of QSARs for regulatory use [OECD 2004][OECD 2007].

As mentioned above the software QSARINS was used to generate MLR-OLS models. In a first step, exhaustive selection was performed by exploring the fitting of all the possible combinations of up to 2 variables included in the models. In a second step, a selection procedure based on a Genetic Algorithm (GA), was used to select the best population of models. The statistical quality of the models was determined by quantification of the coefficient of determination R^2, which represents the fitting ability of the models, and of the Q^2leave-one-out and Q^2leave-more-out (Q^2loo and Q^2lmo, respectively) which represent the internal robustness of the models [Wehrens et al.].

In particular, the GA evolution performed in QSARINS optimizes the Q^2leave-one-out parameter.

In addition, the QUIK rule (set to 0.03−0.05 value in this study) was applied to control the risk of chance correlation, and ensure that the total correlation among the descriptors selected in each model is not higher than their correlation with the modelled response. Furthermore the scrambling of the responses (i.e. Y scramble) was performed to identify and exclude models possibly obtained by chance [Rücker et al.][Gramatica et al., 2004]: low R^2 and Q^2loo values recalculated for each model applied on scrambled responses are expected in the absence of chance correlation.

The best models, according to fitting and robustness were then checked for their ability to predict the endpoint of interest for external molecules i.e. not considered during the development/calibration of the models. To perform the external validations the data available for each response were divided (split) into training and prediction sets before to run the GA. These sets were used to fit the models and to check their external predictivity, respectively. The splitting was performed by ranking the available data in ascending order of the response and putting one every five chemicals in the prediction set (i.e. 20% was set as prediction, while the remaining 80% was kept in the training-set).

The external predictivity was quantified using multiple external validation parameters, which are calculated by QSARINS such as Q^2 ext-F1 [Shi et al.], Q^2 ext-F2 [Schüürmann et al.], Q^2 ext-F3 [Consonni et al.], CCC_{ext} [Chirico et al.] [Chirico et al., 2012] [Lin], as well as the Root Mean Squared of Errors (RMSE). This last parameter summarizes the overall error of the model in training, cross validation and external sets (i.e. $RMSE_{tr}$, $RMSE_{cv}$, and $RMSE_{ext}$ respectively). QSARINS also provides the Mean Absolute Error (MAE) which is a measure of difference between two continuous variables (predicted and measured variables), a common measure to forecast errors.

Therefore, summarizing, the best models were selected as the most externally predictive out of GA based populations, at different levels of complexity and ranked according to decreasing robustness (Q^2loo, Q^2lmo) and fitting (R^2) [Cherkasov et al.].

Finally, the analysis of the applicability domain (AD) of the models allowed for the identification of influential and/or problematic compounds (response and structural outliers). The leverage approach was used to quantify the structural space of the models, which depends on the modelling descriptors and helps in the identification of chemicals which are influent in the selection of the modelling descriptors or are structurally dissimilar from the training set compounds. The Williams plot (hat values vs standardized residuals for each chemical) was used to assess the presence of both response outliers (i.e. compounds with cross-validated standardized residuals greater than 2.5 standard deviation units), and structural outliers (i.e. compounds with leverage value (h) higher than h^* ($(3\ p + 1)/n$), where p is the number of variables of the model and n is the number of compounds in the training set) [Sangion et al.]. The Williams plot is indeed a graph representing the Standard residuals vs HAT i/i. The Hat value of leverage is used for domain applicability assessment. Hat values represent the “distance” of the molecules to the model structural space.

3. Results and discussion

3.1. Data setup

The combination of data available to describe pesticide active substance of pesticide, taxa, species, effects and biologic parameters resulted in a dataset of 952 single entries for aquatic and terrestrial organisms. In the first screening of the data two

substances were excluded from the dataset: Dodine and Pyrethrins. These pesticides are mixtures of different chemical components and as such they can't be modelled by the here proposed QSAR approach and were excluded from the dataset [ECHA 2008]. The number of the single combinations, with the deletion of Dodine and Pyrethrins records was reduced to 929. Duplicate values were excluded as well as aquatic organisms were kept, and only valid combinations of taxa-effect-parameter-species were kept for a total number of 125 useful combinations for the calculation of the models.

The following step taken in the data setup was to explore the dataset of descriptors in order to find outliers that might bring to uncertainties of results of the QSAR models. A Principal Component Analysis (PCA) was performed on the set 525 selected variables.

Fig. 1 shows the distribution of the 70 active ingredients in the new PC1 vs PC2 space (about 25% of the total data variance). This graph is useful to analyze the behavior of the samples (pesticides) in the different Components and their similarity on the basis of the structural information used as input for the analysis.

In Fig. 2 the scree plot shows the number of principal components versus the corresponding eigenvalue. Eigenvalues represent the variance explained by each PC in decreasing order with largest value associated to PC1.

The scree plot is useful to determine how many principal components are necessary to cover a percentage of variance of about 80%. Usually the ideal pattern is a steep curve, followed by a bend and then a straight line, behavior which is highlighted with the circle in the figure. It means that in the first ten principal components most of the information content is kept.

Fig. 3 is a matrix plot, i.e. a multiple plot generated from different combinations of the first 10 PCs.

In particular the outliers are highlighted in light-blue in Fig. 3. These pesticides (Amitrole, Chlorothalonil, Dazomet, Emamectin and Lufenuron) appear structurally different from the other chemicals in the matrix plot, and need to be carefully monitored in the further analysis and model generation.

3.2. QSAR modelling

In a first step models were developed on the basis of split datasets in order to provide external validations for the best combinations of variables selected for each endpoints from the respective GA populations. All the possible combinations of the selected descriptors were firstly explored up to two. The selected Fitness function was the Q^2loo function. The QUIK rule was set to 0.030 (i.e the model variables (X) to response (Y) correlation must be at least 3% higher than within the X block). Then the Genetic Algorithm was run up to four descriptors [Haupt et al.], using the following settings: 2000 iterations (gen. per size), mutation rate of 65% and a population size of 500 models. The selection of the best model within the final GA-based population took into account the principle of the "Occam's Razor", i.e. law of parsimony, which states that among competing hypothesis, the one with the fewest assumptions should be selected. This law is usually applicable to QSAR because large number of descriptors can make the interpretation and explanation of the models more difficult, and often a small number of molecular descriptors outperforms significantly more complex combinations which may lead to overfitting [Cherkasov et al.].

Table 1 shows the statistical parameters calculated for the training and the prediction sets of the split models.

All the best split models reported in Table 1 have good fitting performances (R^2: 0.75–0.99) and therefore they are able to estimate with good approximation the experimental data used for the model development. These models are internally robust since they can handle up to 30% of perturbation with little/no change in Q^2 loo and Q^2 lmo values ((Q^2loo: 0.66–0.98; Q^2 lmo: 0.64–0.98). The low values of R^2_{Yscr} (0.11–0.38) indicates that the models are not affected by chance correlation. The statistical parameters calculated to quantify the external predictivity of the models were all satisfying according to thresholds reported in the literature [Tropsha]. The range of Q^2 ext-Fn: 0.53–0.96 suggests that the models are reasonably predictive when applied to chemicals that were not

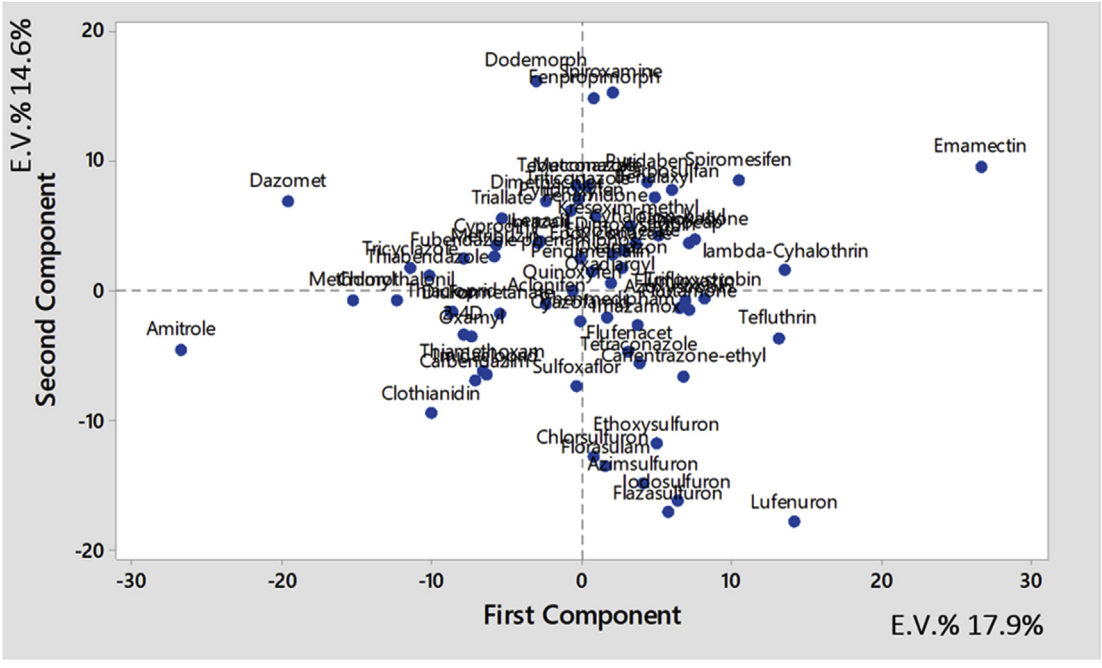


Fig. 1. Score Plot of the selected pesticides in the new XY space: PC1 (E.V.% 17.9%) vs PC2 (E.V.% 14.6%).

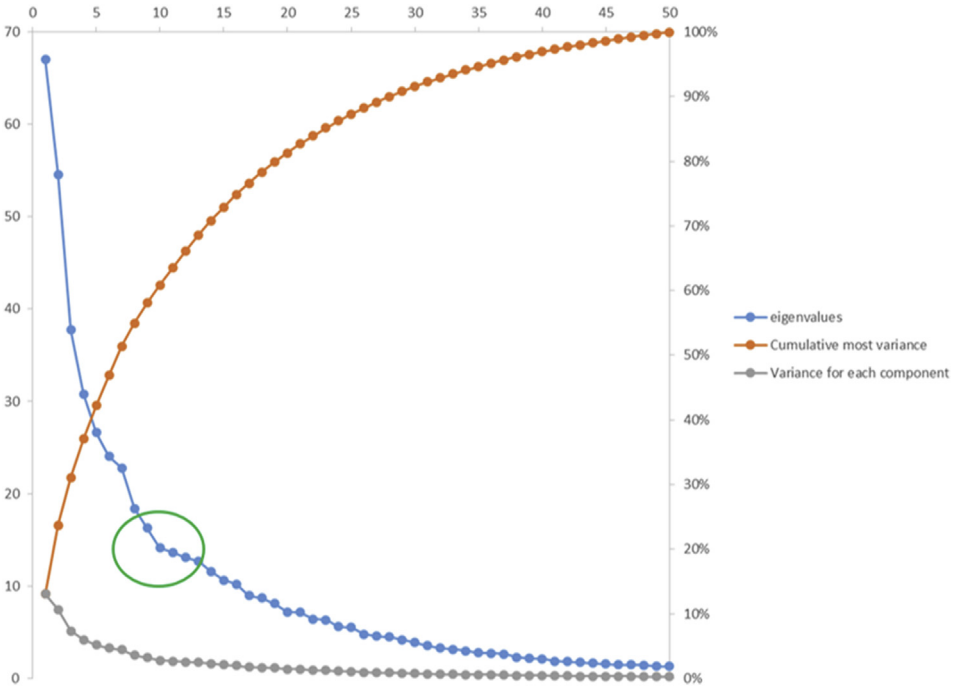


Fig. 2. Scree Plot displays the number of the principal components versus its corresponding eigenvalue. Cumulative variance and variance of each components are showed too.

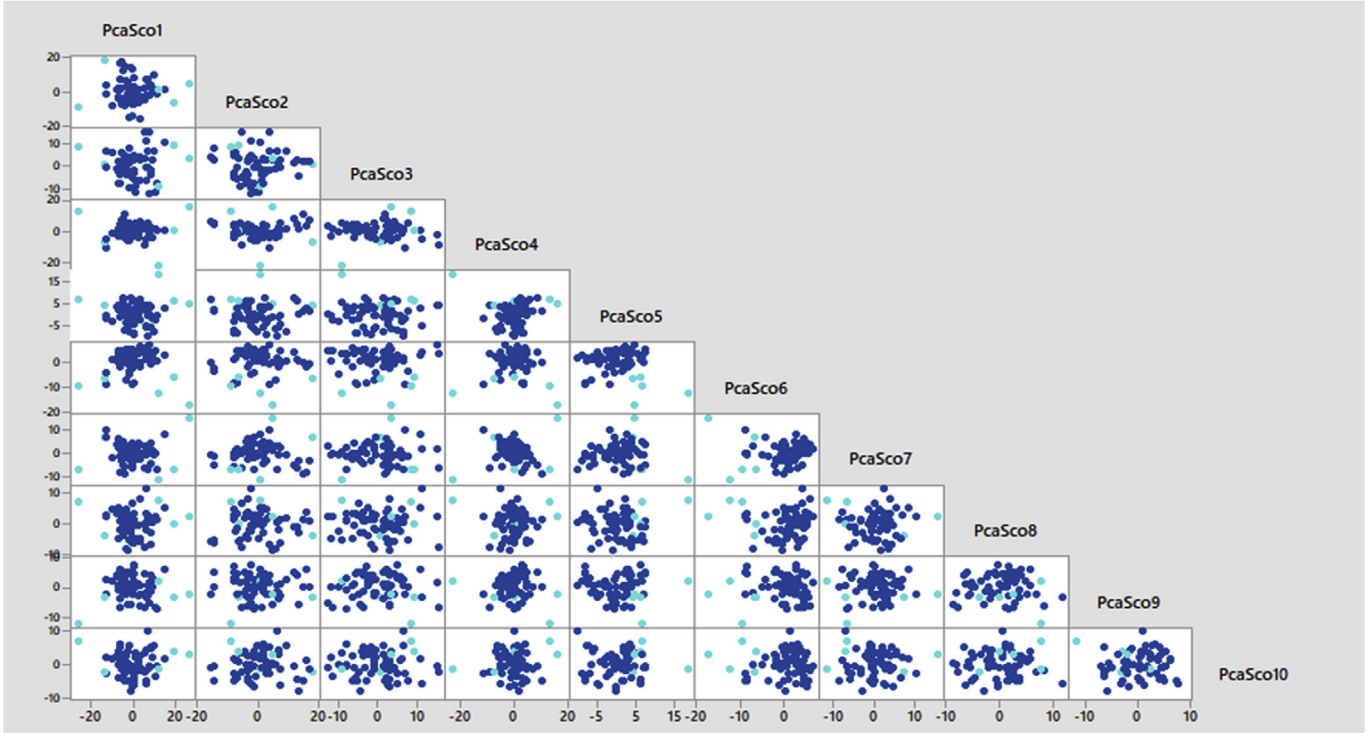


Fig. 3. Matrix Plot: in this graphic are presented all the combination of the scores of the PCA of the first 10 PCs.

involved in the development of the models.

In addition the $RMSE_{ext}$ values of each model on external chemicals were very close to the $RSME_{tr}$ on training chemicals. This confirms that the models had similar accuracy both in fitting and in prediction and thus a good degree of generalization.

Table 2 reports the molecular descriptors combination selected by the GA for each dataset and their relative importance, based on the standardized coefficients, and their definition.

In addition, Table 2 lists the descriptors selected in the models and provides a general description of these variables. Most of them are calculated from 2D structural information encoded into SMILES (supplementary materials section) on the basis of connectivities

Table 1
The combinations represent the selected aquatic taxa/effects/biological parameters/species combinations; the splitting scheme shows which criteria was used to split the pesticides within the training set and the prediction set and how many pesticides were included in both categories: full model means that all the pesticides with a response coming from the pesticide study tests were kept in the training set, instead the column Order By Response refers to a 20% splitting generated after ascending ordering the response the remaining 80% were moved to the training set.

Combination				Splitting			Fitting	Robustness		Chance correlation	External Validation	Accuracy				
code	Taxa tested	Effect Measured	Parameter tested	Specie	Scheme	Ntr	Next	R^2	Q^2loo	Q^2lmo	R^2_{YSCR}	CCC_{ext}	Q^2ext Fn	$RMSE_{tr}$	$RMSE_{ext}$	MAE_{tr}	MAE_{ext}
AL1	Algae	Growth	Cell Count	*Selenastrum capricornutum*	OrderByResponse	22	6	0.83	0.70	0.66	0.19	0.87	0.71 −0.77	0.52	0.67	0.38	0.55
					Full Model	28	none	0.82	0.73	0.70	0.15	none	none	0.55	0.6637 ($RMSE_{cv}$)	0.43	0.5183 (MAE_{cv})
AI4	Aquatic Invertebrates	Reproduction	Young produced	*Daphnia magna*	OrderByResponse	23	6	0.77	0.66	0.64	0.14	0.73	0.53 −0.73	0.77	0.83	0.66	0.57
					Full Model	29	none	0.75	0.66	0.64	0.11	none	none	0.77	0.9064 ($RMSE_{cv}$)	0.63	0.7405 (MAE_{cv})
AP5	Aquatic Plants	Growth	Frond Number	*Lemna gibba L.*	OrderByResponse	10	3	0.96	0.91	0.89	0.22	0.91	0.74 −0.82	0.32	0.71	0.26	0.63
					Full Model	13	none	0.94	0.90	0.89	0.17	none	none	0.39	0.5165 ($RMSE_{cv}$)	0.33	0.4295 (MAE_{cv})
FS6	Fish	Growth	Length	*Pimephales promelas*	OrderByResponse	9	3	0.99	0.98	0.98	0.38	0.79	0.57 −0.85	0.12	0.62	0.10	0.49
					Full Model	12	none	0.96	0.94	0.93	0.27	none	none	0.29	0.3674 ($RMSE_{cv}$)	0.21	0.2894 (MAE_{cv})
FS7	Fish	Growth	Weight	*Oncorhynchus mykiss*	OrderByResponse	9	3	0.97	0.90	0.90	0.38	0.90	0.85 −0.96	0.25	0.26	0.22	0.21
					Full Model	12	none	0.96	0.92	0.91	0.27	none	none	0.25	0.3549 ($RMSE_{cv}$)	0.22	0.3267 (MAE_{cv})

within the molecules and atomic properties such as electronegativities and polarizabilities. We want to highlight that these descriptors do not encode for specific fragments, but they capture global properties of the molecular structure such as spatial autocorrelations (2D autocorrelation), molecular heterogeneity (BCUTS), and electronic accessibility (E-State), which may be quantified by complex calculations. Differently, Pubchem Fragments listed in Table 2 encode for specific patterns in the structures, therefore the presence/absence of these specific chemical features (e.g. functional groups or molecular portions) in a model can be easily directly associated to variations in the studied effect.

A more in depth comment of these descriptors is provided in the following sections. However, regarding the interpretation of these molecular descriptors it is necessary to bear in mind that the here proposed models consist of multivariate combinations of descriptors selected through a statistically driven procedure (i.e. GA-based selection). Therefore, none of the descriptors can independently explain the behavior of the modelled endpoint i. e only the combination of descriptors allows the accurate modelling of the studied responses.

The plot of experimental versus predicted toxicity values for both the full and the split models are reported, as well as an example of Williams plot (i.e. representative for the split model), which describes the model's AD.

Therefore, having verified the predictive ability of the best variables selected by the GA for each modelled endpoint, and considering the limited amount of chemicals included in each training set, the best equations were newly calibrated using all the available experimental information (Full model).

3.3. QSAR model for algae

3.3.1. Selenastrum capricornutum toxicity test AL1- endpoint based on growth (cell count)

The full model for EC_{50} in *Selenastrum capricornutum* was calibrated on 28 pesticides. The endpoint, expressed as Log (EC_{50}), ranged from −2.47 to 2.45 μmol/L (i.e. from 1.19 to 71,600 μg/L). The equation of the full QSAR model, and relative plots (Fig. 4) for AL1 are given below:

$$LOG(EC_{50})_{AL1} = 0.45 + 1.55\mathbf{MATS8e} - 1.28\mathbf{PubchemFP645} - 4.78\mathbf{MATS4e} - 1.33\mathbf{PubchemFP346} \quad (1)$$

$N = 28 \quad R^2 = 0.82 \quad Q^2_{loo} = 0.73 \quad Q^2_{LMO} = 0.70 \quad R^2_{Yscr} = 0.15$
$RMSE_{tr} = 0.55 \quad RMSE_{cv} = 0.66$
$MAE_{tr} = 0.42 \quad MAE_{cv} = 0.52$

The most important variables selected in this model are two Moran autocorrelation coefficients [Todeschini et al.] [Todeschini et al., 2000] MATS4e (std. Coefficient −0.51) and MATS8e (std. Coefficient 0.42). These variables, encode for structural information related to the degree of autocorrelation between numerical values of a property (i.e. electronegativity) at a specific topological distance (i.e. distance 4 and 8, respectively). The other variables selected in the model are Pubchem Fingerprints. These descriptors, as specified above, are Boolean values reflecting the presence or not of a chemical characteristic in a chemical structure. In particular, PubchemFP645 and 346 [SMARTS Theory] (std. Coefficient −0.45 and −0.52) refer to O=C−N−C−C and C (~C) (~H) (~O) fragments, respectively. In this dataset, the presence of both fragments in the molecular structure, in addition to negative values or values close to zero of the autocorrelation descriptors increases the toxic

Table 2
Selected molecular descriptors for each model.

Combinations	Descriptors	Std coefficient	Range		Definition
		(full model)	min	max	
AL1	MATS8e	0.42	−0.67	1.33	Moran autocorrelation - lag 8/weighted by Sanderson electronegativities (2D)
AL1	PubchemFP645	−0.45	0.00	1.00	O=C−N−C−C; Simple SMARTS patterns - These bits test for the presence of simple SMARTS patterns, regardless of count, but where bond orders are specific and bond aromaticity matches both single and double bonds.
AL1	MATS4e	−0.51	−0.75	0.35	Moran autocorrelation - lag 4/weighted by Sanderson electronegativities (2D)
AL1	PubchemFP346	−0.52	0.00	1.00	C (~C) (~H) (~O); Simple atom nearest neighbors - These bits test for the presence of atom nearest neighbor patterns, regardless of bond order (denoted by "~") or count, but where bond aromaticity (denoted by ":") is significant.
AI4	BCUTp-1h	0.92	6.34	13.43	nlow highest polarizability weighted BCUTS (eigenvalue-based descriptors)
AI4	MIC0	−0.66	9.90	40.94	Modified information content index (neighborhood symmetry of 0-order)
AI4	PubchemFP12	−0.80	0.00	1.00	≥ 16 C; Hierarchic Element Counts - These bits test for the presence or count of individual chemical atoms represented by their atomic symbol.
AP5	BCUTc-1l	0.84	−0.41	−0.24	nhigh lowest partial charge weighted BCUTS (eigenvalue-based descriptors)
AP5	maxwHBa	0.56	0.00	5.31	Electro topological State Atom Type Descriptor: Maximum E-States for weak Hydrogen Bond acceptors
FS6	SpMax5_Bhi	−0.63	1.23	3.74	Burden Modified Eigenvalues Descriptor: Largest absolute eigenvalue of Burden modified matrix - n 5/weighted by relative first ionization potential
FS6	minaaCH	−0.40	0.00	2.29	Electro topological State Atom Type Descriptor: Minimum atom-type E-State::CH:
FS6	PubchemFP613	0.67	0.00	1.00	C−N−C−C−C; Simple SMARTS patterns - These bits test for the presence of simple SMARTS patterns, regardless of count, but where bond orders are specific and bond aromaticity matches both single and double bonds.
FS7	GATS6e	0.48	0.00	2.19	Geary autocorrelation - lag 6/weighted by Sanderson electronegativities
FS7	SRW5	0.43	0.00	4.26	Walk Count Descriptor: Self-returning walk count of order 5 (ln (1+x)
FS7	PubchemFP179	−0.86	0.00	1.00	≥ 1 saturated or aromatic carbon-only ring size 6; Rings in a canonic Extended Smallest Set of Smallest Rings (ESSSR) ring set - These bits test for the presence or count of the described chemical ring system.

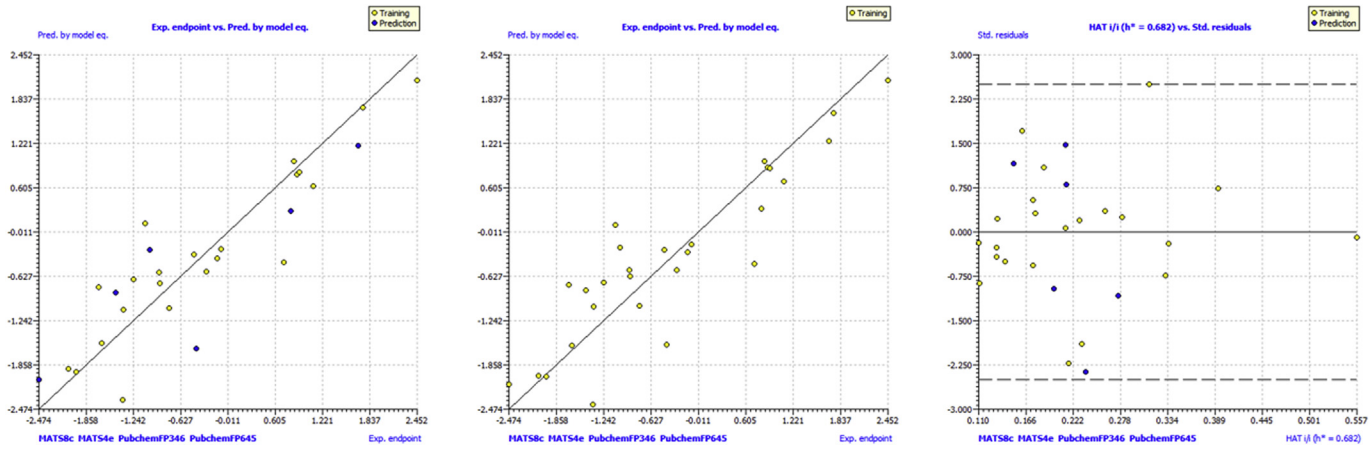


Fig. 4. The plot of experimental versus predicted endpoints for AL1 for the split model (left), for the full model (middle) and Williams plot for AL1 (right).

potency of the chemicals (more negative Log (EC_{50})).

3.4. QSAR model for aquatic invertebrates

3.4.1. Daphnia magna toxicity test AI4- endpoint based on reproduction (young produced)

The Log (EC_{50}) for the 29 pesticides with available data in *Daphnia magna* ranged from −4.37 to 3.93 μmol/L. The equation of the full QSAR model, and related plots (Fig. 5) for AI4 are given below:

$$LOG(EC_{50})_{AI4} = -11.09 + 1.42\mathbf{BCUTp-1h} - 2.61\mathbf{PubchemFP12} - 0.16\mathbf{MIC0} \quad (2)$$

$$N = 29 \quad R^2 = 0.75 \quad Q^2_{loo} = 0.66 \quad Q^2_{LMO} = 0.64 \quad R^2_{Yscr} = 0.11$$
$$RMSE_{tr} = 0.77 \quad RMSE_{cv} = 0.90 \quad MAE_{tr} = 0.63 \quad MAE_{cv} = 0.74$$

The BCUTp-1h descriptor [Pearlman et al.][Burden et al.][Burden et al., 1997][Kang et al.] (std. Coefficient 0.92) is the most important descriptor in the model. This variable takes into account both connectivity and atomic properties relevant to intermolecular interactions. B-CUT descriptors are calculated from a matrix representation of the molecular graph where diagonal elements encode for atomic properties such as, in this case, polarizability. In the proposed equation BCUTp-1h is positively correlated with the response, therefore the most toxic compounds in the present dataset are characterized by small values of the B-CUT descriptor.

The PubchemFP12 fragment [Pubchem fingerprints] (std. Coefficient −0.80) comes from the Hierarchic Element Counts and brings information related to molecular dimension and atom diversity (i.e. number of C atoms equal to or greater than 16). Since this descriptor is inversely related to the endpoint, the absence of the aforementioned pattern (i.e. PubchemFP12 = 0) is associated with low toxicity values (i.e. large Log (EC_{50})). Finally the MIC0 descriptor [Todeschini et al., 2000] (std. Coefficient −0.66) is an index of neighborhood symmetry (Modified information content index - neighborhood symmetry of 0-order).

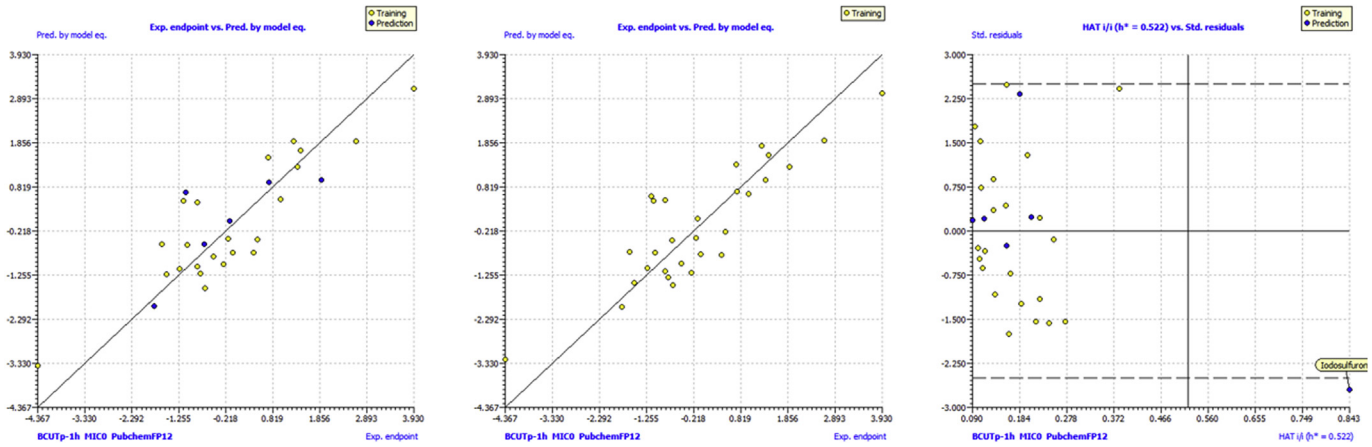


Fig. 5. The plot of experimental versus predicted endpoints for Al4 for the split model (left), for the full model (middle), and Williams plot for Al4 (right).

3.5. QSAR model for aquatic plants

3.5.1. Lemna gibba toxicity test AP5 - endpoint based on growth (frond number)

This dataset is very limited and composed only of 13 pesticides with Log (EC_{50}) values ranging from −2.86 to 2.78 μmol/L. The equation of the full QSAR model, and related plots (Fig. 6) for AP5 are given below:

$$LOG(EC_{50})_{AP5} = 6.14 + 0.84\mathbf{BCUTc1l} - 2.26\mathbf{maxwHBa} \quad (3)$$

$$N = 13 \quad R^2 = 0.94 \quad Q^2_{loo} = 0.90 \quad Q^2_{LMO} = 0.89 \quad R^2_{Yscr} = 0.17$$
$$RMSE_{tr} = 0.39 \quad RMSE_{cv} = 0.52 \quad MAE_{tr} = 0.33 \quad MAE_{cv} = 0.43$$

Also in this model a descriptor from the burden group is selected as the most relevant, i.e. BCUTc-1l (std. Coefficient 0.84). As mentioned above modified burden descriptors encode for the distribution of a property (partial charge in this case) in the molecule [Todeschini et al., 2000]. In this dataset the value of this descriptor BCUTc-1l increases (i.e. becomes more negative) with chemical dimension and is directly correlated with the response (i.e. more toxic chemicals have large negative BCUTc-1 values). In addition, the descriptor maxwHBa (Maximum E-States for weak Hydrogen Bond acceptors - std. Coefficient 0.56) is directly correlated to the studied endpoint.

3.6. QSAR models for fish

3.6.1. Pimephales promelas test FS6- endpoint based on growth (length)

The dataset available to model EC50 in *Pimephales promelas* was also very small, with toxicity values, expressed as Log (EC_{50}), between −2.91 and 2.34 μmol/L. The equation of the full QSAR model, and related plots (Fig. 7) for FS6 are given below:

$$LOG(EC_{50})_{FS6} = 7.49 + 2.43\mathbf{PubchemFP}613 - 0.69\mathbf{minaaCH} - 2.67\mathbf{SpMax}5\_\mathbf{Bhi} \quad (4)$$

$$N = 12 \quad R^2 = 0.96 \quad Q^2_{loo} = 0.94 \quad Q^2_{LMO} = 0.93 \quad R^2_{Yscr} = 0.27$$
$$RMSE_{tr} = 0.29 \quad RMSE_{cv} = 0.37 \quad MAE_{tr} = 0.21 \quad MAE_{cv} = 0.29$$

As for the QSAR model developed for *Lemna gibba* it should be noted that this model is based on a very limited number of data. The limited external predictivity of the model reported in Table 1 is due to the exclusion of essential structural information from the training set. This information is important to help the calibration of the coefficients of the model for an accurate prediction of compounds Epoxiconazole and Carbosulfan which behave as strong outliers in the split model. However the full model is robust when checked for cross validation.

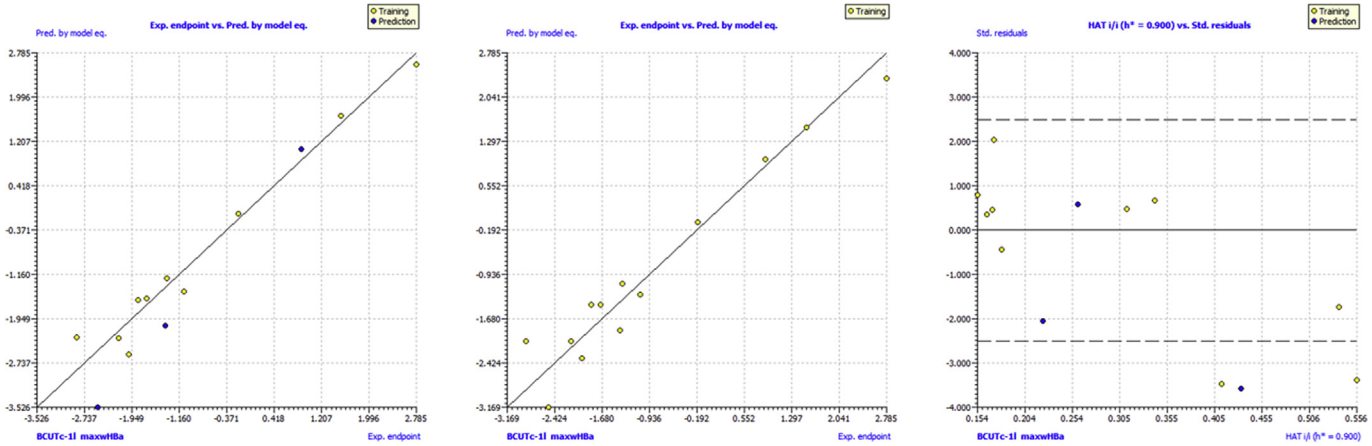


Fig. 6. The plot of experimental versus predicted endpoints for AP5 for the split model (left), for the full model (middle), and Williams plot for Al5 (right).

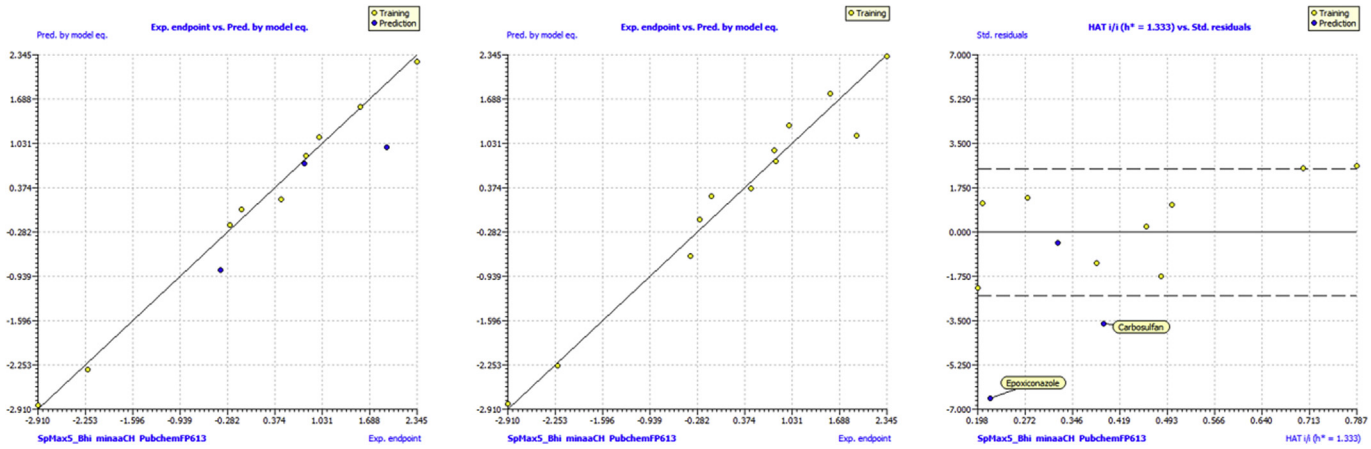


Fig. 7. The plot of experimental versus predicted endpoints for FS6 for the split model (left), for the full model (middle), and Williams plot for FS6 (right).

The most important descriptors in the model are the Fingerprint PubchemFP613 and the descriptor minaaCH from the electro-topological state indices. PubchemFP613 checks for the presence of the C—N—C—C—C SMARTS pattern. The most toxic among the chemicals investigated for toxicity to Pimephales i.e. Pyridaben and lambda-Cyalothrin miss this fragment in their molecular structure; the descriptor minaaCH (i.e. Minimum atom-type E-State aromatic—CH—aromatic) encodes information related to the connectivity of a specific atom (topological environment) and the electronic interactions due to all other atoms in the molecule. The topological relationship is based on the graph distance to each other atom. Atom-type E-state indices are calculated by summing the E-state values (e.g. electronic information) of all atoms of the same atom type in the molecule [Todeschini et al., 2009]. These descriptors help to capture information associated to electronic accessibility of an atom therefore the probability of interaction with another molecule. In this model the selected E-state descriptor is inversely correlated to the studied endpoint. Finally the SpMax5_Bhi belongs from the eigenvalue-based descriptor family [Todeschini et al., 2009] and takes into account molecular heterogeneity based on atoms electronegativities weighted by ionization potential. In the studied dataset large chemicals, have large values of this descriptor which is inversely related with the studied endpoint (i.e. large SpMax5Bhi values corresponds to small Log (EC_{50}) values).

3.6.2. *Oncorhynchus mykiss toxicity test FS7- endpoint based on growth (weight)*

The Log (EC_{50}) expressed in μmol/L measured for the 12 pesticides for *Oncorhynchus mykiss* ranged from −2.25 to 2.73. The best results were obtained using three variables (shown in equation (5)). The full QSAR model, and relative plots (Fig. 8) for FS7 are given below:

$$LOG(EC_{50})_{FS7} = -0.56 + 1.68\mathbf{GATS6e} + 0.464\mathbf{SRW5} - 2.54\mathbf{PubchemFP179} \quad (5)$$

$$N = 12 \quad R^2 = 0.96 \quad Q^2_{loo} = 0.92 \quad Q^2_{LMO} = 0.91 \quad R^2_{Yscr} = 0.27$$
$$RMSE_{tr} = 0.25 \quad RMSE_{cv} = 0.35 \quad MAE_{tr} = 0.22 \quad MAE_{cv} = 0.33$$

The GATS6e Descriptor Geary autocorrelation - lag 6/weighted by Sanderson electro-negativities is part of the Autocorrelation descriptors. The Geary coefficient is a distance-type function varying from zero to infinity. Brings information related to chemical complexity/heterogeneity and is positively related to the studied endpoint [Todeschini et al., 2009]. The PubchemFP179 Fragment is part of a class of descriptors which tests the presence of a specific chemical ring system. In the studied dataset, this descriptor helps to identify chemicals that present at least one or more saturated or aromatic carbon-only ring of size 6, which tend to be more toxic (i.e. lower toxicity values) than chemicals characterized by the presence of other ring systems (e.g. heterocycles or heteroaromatic).

The SRW5 is part of the Walk Count Descriptors and it is defined as the Self-returning walk count of order 5 (ln (1+x). This is a simple molecular descriptor based on counting defined elements of a compound [Todeschini et al., 2009] and it brings information related to molecular dimension and the presence of rings.

It should be noted that the complexity of Eq. (5) (i.e. number of variables) is rather high considering the dimension of the training set. Therefore, in order to be in agreement with the parsimony principle, best solution would be, in a future refinement, to find a combination of lower complexity and comparable accuracy as Equation (5) or, as an alternative to increase the number of samples for the training/prediction sets.

3.7. *Evaluation of the applicability domain of the proposed QSARs*

After assessing the robustness and predictivity of all models, it is interesting to analyze, for each taxa, the percentage of predicted data which fall inside the applicability domain (AD) of the respective QSAR. Results of this analysis are shown in Table 3. All the ADs of the developed models showed a good coverage of the dataset demonstrating that the predictions generated by these QSARs are reliable for several pesticides. This is interesting considering the limited amount of data originally available for some of the endpoints (i.e. 12-13 data points). The number of pesticides falling inside the ADs of all the models is 51 (75% of all selected pesticides), i.e. considering only the interpolated, reliable predictions.

The distribution of molecular weight were investigated too, without any noticeable meaning.

3.8. *Investigation of the multi-specie toxicity profile of the studied pesticides*

Principal Component Analysis (PCA) was used to investigate the toxicity profile of the studied pesticides according to their experimental or estimated toxicity in different taxa. The PCA was firstly performed on the 51 pesticides which had all interpolated

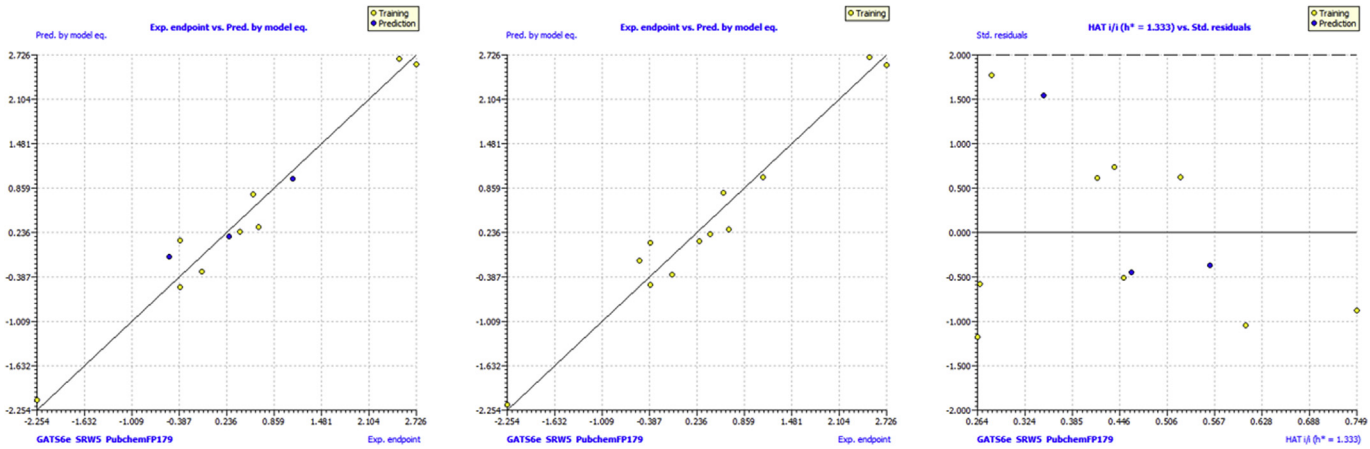


Fig. 8. The plot of experimental versus predicted endpoints for FS7 for the split model (left), for the full model (middle), and Williams plot for FS7 (right).

Table 3
Percentage of pesticides, over a total of 68, falling outside the Applicability Domain of each determined QSAR model and considering the whole set of models together. In the last column all the pesticides falling outside the applicability domain of each determined model are reported. Highlighted with an asterisk (*), the pesticides previously identified as outliers of the multivariate analysis (Amitrole, Chlorothalonil, Dazomet, Emamectin and Lufenuron).

Taxa (code – training set)	Species	Outside AD	Pesticides Outside AD
Algae (AL1 - 28)	*Selenastrum capricornutum*	2.9% (2/68)	2-4D, Methomyl
Aq. Invertebrates (AI4 - 30)	*Daphnia magna*	4.4% (3/68)	Amitrole*, Iodosulfuron, Lufenuron*
Aquatic Plants (AP5 - 14)	*Lemna gibba*	8.8% (6/68)	Dodemorph, Flumioxazin, Imazalil, Oxadiargyl, Oxamyl, Spiroxamine
Fish (FS6 - 13)	*Pimephales promelas*	4.4% (3/68)	Amitrole*, Chlorothalonil*, Methomyl
Fish (FS7 - 13)	*Oncorhynchus mykiss*	11.8% (8/68)	Amitrole*, Chlorothalonil*, Clothianidin, Epoxiconazole, Imidacloprid, Lambda-Cyhalothrin, Methomyl, Tefluthrin
Aquatic Organisms	*All the cited above*	25% (17/68)	All the above cited pesticides

predictions, using only predicted values. The biplot (i.e. plots of the loadings and of the scores) is presented in Fig. 9 whereas Fig. 10 shows the Loading plots for the aforementioned PCA.

In all the loading plots, each loading aims at the direction of maximum Log (EC_{50}), i.e. high distance from the center of the axis corresponds to low toxicity for the displayed taxa. On the left side of Fig. 9 (i.e. PC1 vs PC2 biplot) there are the active ingredients which tend to be toxic to all the species. These compounds should be prioritized as the most hazardous for the aquatic environment represented in this study and are listed Table 4.

PC2 discriminates between autotrophic (negative) and heterotrophic (positive) organisms: algae and aquatic plants stand completely separated from the rest of the loadings in PCs where PC2 is present, so this distinction is quite clear and this PC can be used as descriptor for toxicity of Aquatic Algae (AL1) and Aquatic Plants (AP5) vs. Invertebrates (AI4) and Fish (FS6 and FS7). For example in the PC1vsPC2 graph, herbicides (in green), which are the most notable pesticides for aquatic plants and algae, are displayed on the opposite of AL1 and AP5 loadings.

These patterns were identified also in the PCA based on the integration of experimental and predicted toxicity values which supports the reliability of the here proposed QSARs.

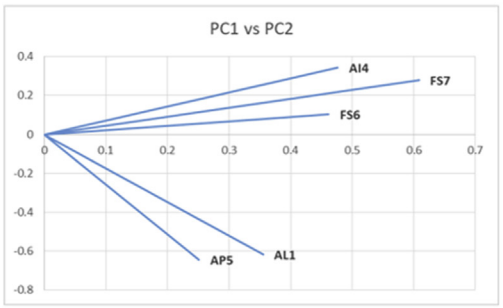
In addition we attempted to categorize the 51 studied substances on the basis of toxicity thresholds for hazard characterization as proposed by US EPA Aquatic Risk Assessment [EPA]. In particular, three levels of aquatic toxicity concern for acute and chronic toxicity are used in this approach, which are shown in Table 5. The PCA biplot with scores labelled according to the proposed categorization is shown in Fig. 11.

Even if NOEC or EC_{10} values would be more suitable for the aforementioned categorization, we performed our analysis on the basis of EC_{50} data (experimental and predicted) available in the present work which were categorized according to the pattern proposed by the US EPA reported in Table 5. High-concern pesticides were identified as those which counted 4 or more than 4 EC_{50} values exceeding the EPA thresholds on the basis of 5 species; medium-concern pesticides were identified as those which counted 2 or 3 EC_{50} values exceeding the EPA thresholds; low-concern pesticides were identified as those which counted less than 2 EC_{50} values exceeding the EPA thresholds Table 6.

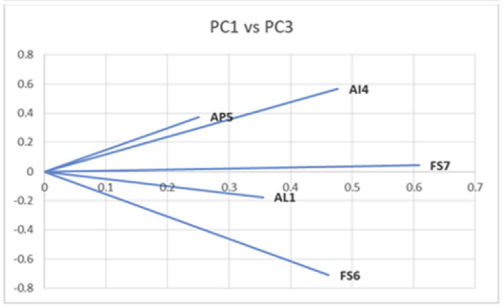
While only three of the potentially most hazardous pesticides belongs to the insecticides and nematocides classes, more than 15 belong to the fungicide and herbicides classes. The use of these molecules should be monitored as well as their fate in the environment because of the potential hazard they may pose to non-target aquatic organisms.

3.9. Further considerations

Regulation (EC) No. 1107/2009 encourages the use and development of non-experimental tests to anticipate the possible health and environmental risks of pesticides [Villaverde et al., 2017], but the current regulatory pesticide toxicity testing and assessment approaches, in the framework of environmental risk assessment, remain to a large extent based on a checklist of *in vivo* tests, conducted in accordance with standardized test guidelines or protocols such as OECD Test Guidelines. While this approach has evolved over the past half century, it is unlikely to efficiently meet legislative



Components	Variance	Cumulative E.V. (%)
1	34.5	34.5
2	27.0	61.5
3	18.7	80.2
4	10.9	91.2
5	8.8	100.0



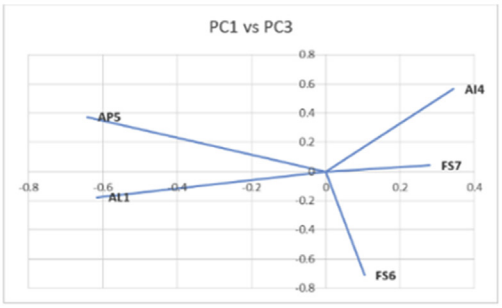


Fig. 9. Biplot of the PCA.

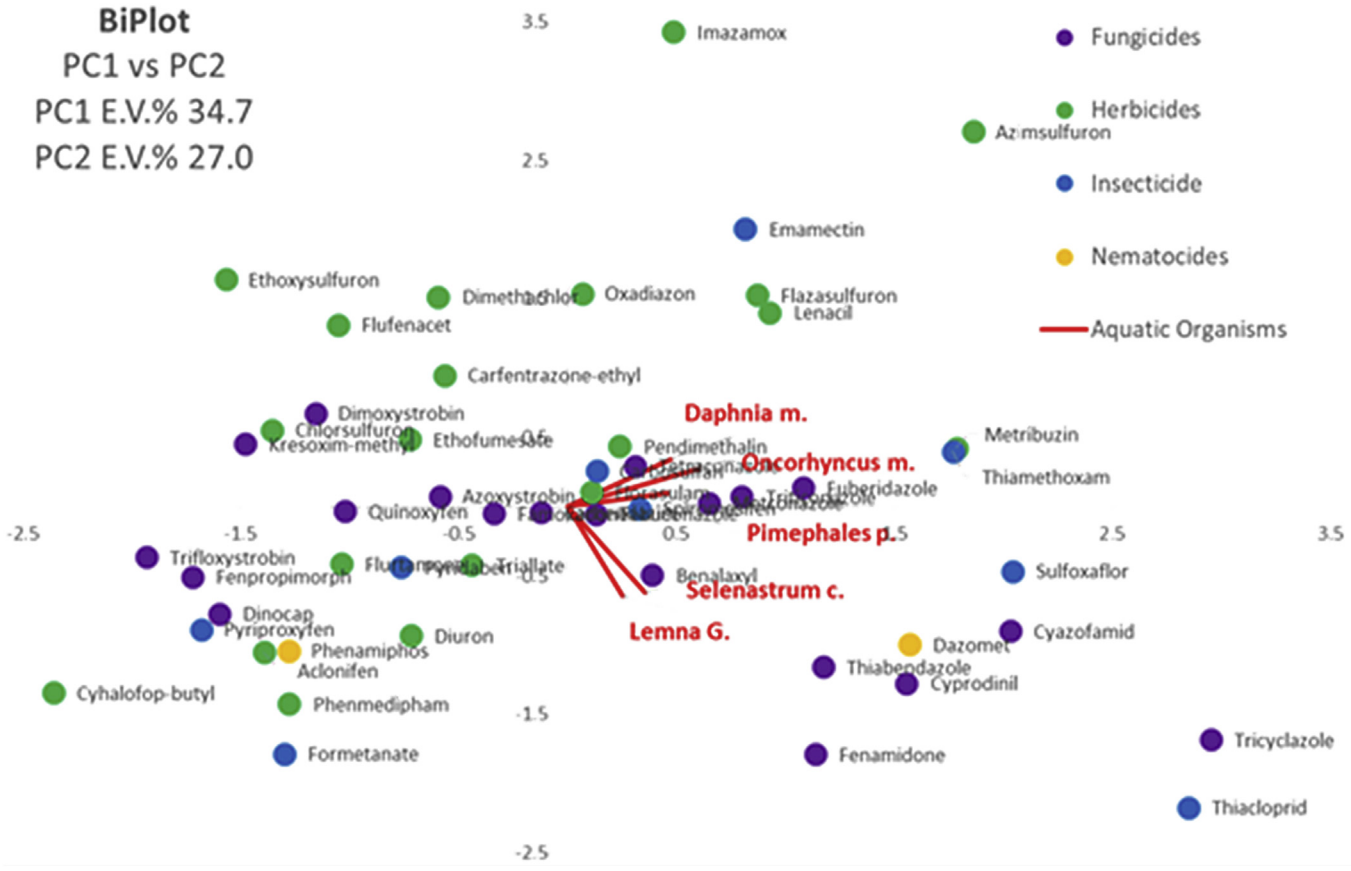


Fig. 10. Loading plots of the PCA

mandates that require increased numbers of chemical assessments to be undertaken without a concomitant increase in the use of animals and resources. New approaches are necessary to close the gap between the number of chemicals in use and the number assessed to date. Modern computational methodologies for toxicological testing and chemical risk assessment are currently a topic of great interest amongst researchers and the regulatory community, because of their potential for predicting chemical toxicity and reducing animal testing [Benfenati et al., 2017]. As mentioned above, the EU Regulation on pesticides still provides, as

Table 4
Most hazardous pesticides for aquatic compartment among the dataset according to PC1 ranking of EC_{50} μg/L].

Pesticide Category	Most hazardous pesticides for aquatic organism
Fungicides	Trifloxystrobin; Fenpropimorph; Kresoxim-methyl; Dinocap; Dimoxystrobin; Quinoxyfen
Herbicides	Cyhalofop-butyl; Ethoxysulfuron; Aclonifen; Phenmedipham; Chlorsulfuron; Flurtamone; Flufenacet
Insecticides	Pyriproxyfen; Formetanate
Nematocides	Phenamiphos

Table 5
EPA aquatic toxicity concern.

	Low Concern	Moderate Concern	High Concern
Acute	>100 mg/L	1–100 mg/L	<1 mg/L
Chronic	>10 mg/L	0.1–10 mg/L	<0.1 mg/L

Table 6
Classes concern toxicity based on count of high-concern pesticides according to EPA scheme.

	Green	Orange	Red
Count of High-Concern over the 5 aquatic species	<2	[2–3]	≥ 4

mandatory data requirements, that ecotoxicological studies have to be performed to test aquatic toxicity of substances. The EFSA Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface water opens a window on in silico tests, but it also aims the attention at a major concern, i.e. the danger of underestimating the real toxicity or hazard of the given substances. As a consequence, it is, currently, difficult that non-testing methods replace *in vivo* studies in the risk assessment as they are not legally accepted as data requirements. What lacks is a global repository of harmonized, structured and certified data in order to reduce uncertainty which is actually the big problem of in silico strategies [Villaverde et al., 2019], even if, for example, EFSA has recently launched calls for tenders for the creation of datasets that may be used for the development of pesticide-oriented computational tools. In the US, the regulation on pesticides, relies on the Integrated Approaches to Testing and Assessment (IATA): IATA have the potential to integrate existing data on pesticides with the results of alternative methods (e.g., biochemical/cellular assays, QSAR) leading to the refinement, reduction, and/or replacement of conventional test requirements [TWG on Pesticide 2012]. Nowadays the European Union science hub is going to integrate the IATA approach to the EU context and this might open a window to new perspectives and innovative scenarios. [Kolesnyk], [Jaworska], [Villaverde et al., 2019].

4. Conclusions

Ecotoxicological data of aquatic organisms gathered from 70 active substances' approval dossiers were collected into a storage MS Access database. In particular pesticide Effect Concentrations

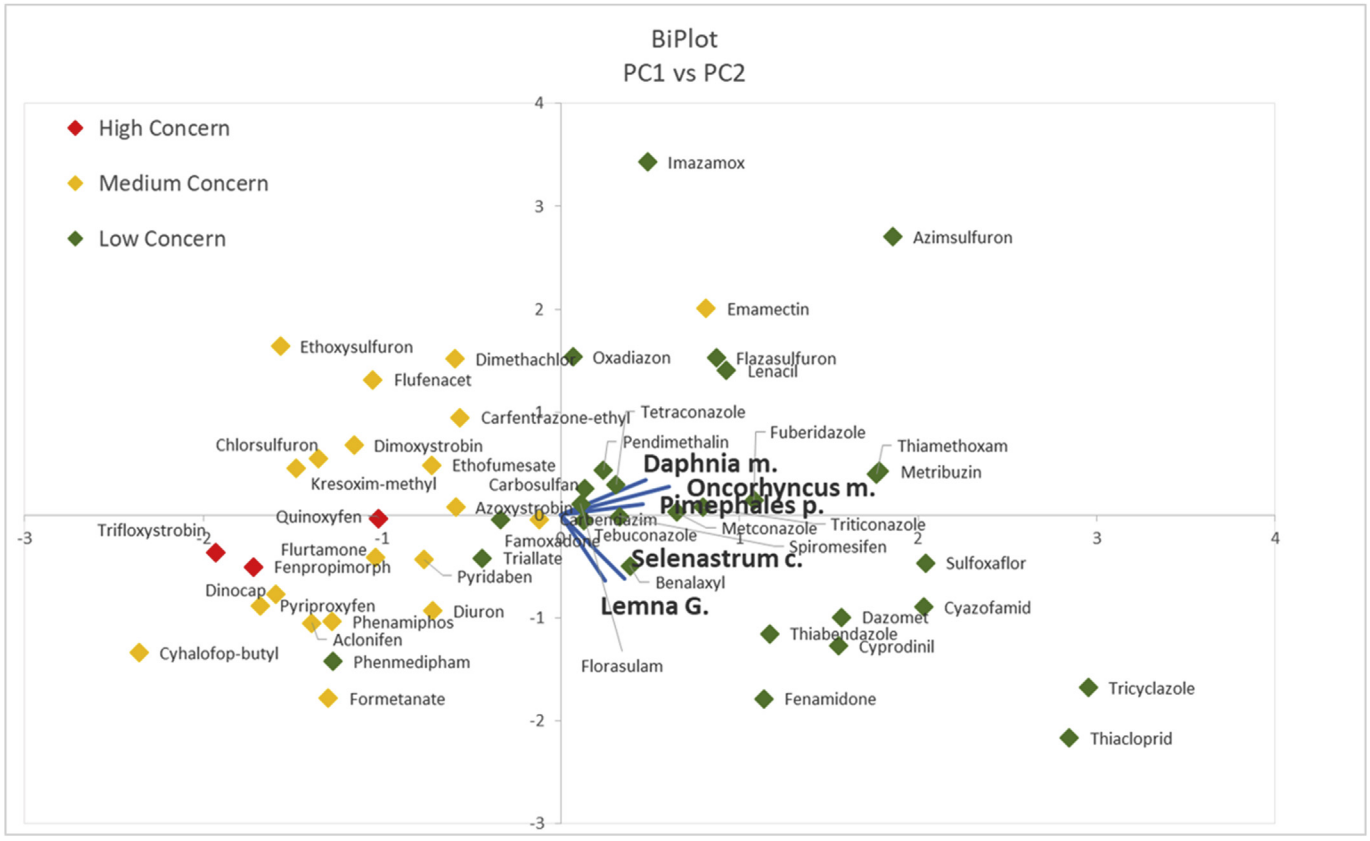


Fig. 11. PC1 E.V.% 34.7 – PC2 E.V.% 27.0. As expected moderate and high concern chemicals are grouped on the left side PC1.

for *Selenastrum capricornutum* (green alga), *Daphnia magna* (aquatic invertebrates), *Lemna gibba* (aquatic plant), *Pimephales promelas* and *Oncorhynchus mykiss* (fresh water fish) were collected. The EC_{50} were derived for specific combinations of pesticides, species, toxicological effects and biological parameters selected to measure the different effects.

The present study extends the completeness of the set of EC_{50} data for all the above cited monitored species and for all the 70 selected pesticides. The here proposed QSAR models, which were adequately built and tested by internal and external validations, had good statistical performances as well as satisfactory applicability domains. In particular, the good coverage of the dataset demonstrated that the predictions could be used reliably to predict Log (EC_{50}) data for several categories and different families of pesticides whose toxicity for aquatic organisms is completely unknown. The Principal Component Analysis of the interpolated predictions was performed to assess the toxicological profile of the studied pesticides in a simplified aquatic scenario described by experimental and predicted toxicity data for 5 species. Some interesting information were derived from this multivariate analysis such as the separation of the toxicity of the studied compounds for autotrophic and heterotrophic aquatic organisms. Finally, a priority list of the potentially most hazardous compounds on the basis of combined measures of toxicity for the aquatic environment was compiled on the basis of the combination of QSARs predictions and multivariate analysis.

4.1. Recommendations

Due to the importance of the ecotoxicological endpoints in the pesticide environmental risk assessment process, the present work does not have any conceit to substitute the risk assessment itself but it sets some predictions and a priority list, which may be of use in the general a priori assessment of the potential hazard of the pesticides.

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.

Acknowledgements

The author wish to thank Dr. Mara Luini and Dr. Flavio Marchetto for the useful discussion of models results and Prof. Paola Gramatica for authorizing the use of the software QSARINS.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.watres.2020.115583.

References

COUNCIL DIRECTIVE of 27 July 1976 on the Approximation of the Laws of the Member States Relating to Cosmetic Products (76/768/EEC) and REGULATION (EC) No 1223/2009 of the EUROPEAN PARLIAMENT and of the COUNCIL of 30 November 2009 on Cosmetic Products.

Benfenati, E., Manganaro, A., Gini, G., 2017. VEGA-QSAR: AI inside a platform for predictive toxicology. CEUR Workshop Proceedings 1107.

Burden, F.R., 1989. Molecular identification number for substructure searches. J. Chem. Inf. Comput. Sci. 29 (3), 225–227.

Burden, F.R., 1997. A chemically intuitive molecular index based on the eigenvalues of a modified adjacency matrix. Quant. Struct.-Act. Relat. 16 (4), 309–314. https://doi.org/10.1002/qsar.19970160406.

Cherkasov, A., Muratov, E.N., Fourches, D., 2014. QSAR Modeling: where have you been? Where are you going to? J. Med. Chem. 57 (12), 4977–5010. https://doi.org/10.1021/jm4004285.

Chirico, N., Gramatica, P., 2011. Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. J. Chem. Inf. Model. 51 (9), 2320–2335. https://doi.org/10.1021/ci200211n.

Chirico, N., Gramatica, P., 2012. Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. J. Chem. Inf. Model. 52, 2044–2058. https://doi.org/10.1021/ci300084j.

Consonni, V., Ballabio, D., Todeschini, R., 2009. Comments on the definition of the Q(2) parameter for QSAR validation. J. Chem. Inf. Model. 49, 1669–1678. https://doi.org/10.1021/ci900115y.

Damalas, C.A., Eleftherohorinos, I.G., 2011. Pesticide exposure, safety issues, and risk assessment indicators. Int. J. Environ. Res. Publ. Health 8 (5), 1402–1419. https://doi.org/10.3390/ijerph8051402.

Directive 2009/128/Ec, 2009. Off. J. Eur. Union L309, 71–86. https://doi.org/10.3000/17252555.L_2009.309.eng.

Doke, S.K., Dhawale, S.C., 2015. Alternatives to animal testing: a review', Saudi Pharmaceutical Journal. King Saud University 23 (3), 223–229. https://doi.org/10.1016/j.jsps.2013.11.002.

Echa and Qsar Toolbox. no date, Available at: https://www.qsartoolbox.org/.

Efsa, 2013. Guidance on tiered risk assessment for edge-of-field surface water. EFSA Journal. https://doi.org/10.2903/j.efsa.2013.3290.

Efsa Ppr, 2013. Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters. EFSA Journal 11 (7), 3290. https://doi.org/10.2903/j.efsa.2013.3290.

Epa, 2015. Sustainable Futures/P2 Framework Manual - Introduction, Table of Contents, Sections 1-14 and Appendices A-H. https://www.epa.gov/sustainable-futures/sustainable-futures-p2-framework-manual.

European Chemicals Agency, 2008. Guidance on Information Requirements and Chemical Safety Assessment: QSARs and Grouping of Chemicals, p. 134. https://doi.org/10.2823/43472. Guidance for the implementation of REACH, R.6(May).

Galimberti, F., Marchetto, F., 2015. Comparison of NOEC values to EC10/EC20 values, including confidence intervals, in aquatic and terrestrial ecotoxicological risk assessment. EFSA Journal. https://doi.org/10.2903/sp.efsa.2015.EN-906. GP/EFSA/PRAS/2013/01. http://www.efsa.europa.eu/en/supporting/pub/906e.

Gramatica, P., Worth, A., 2004. Evaluation of Different Statistical Approaches for the Validation of Quantitative Structure-Activity Relationship.

Gramatica, P., et al., 2013. QSARINS: a new software for the development, analysis, and validation of QSAR MLR models. J. Comput. Chem. 34 (24), 2121–2132. https://doi.org/10.1002/jcc.23361.

Gramatica, P., Cassani, S., Chirico, N., 2014. QSARINS-chem: insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. J. Comput. Chem. 35 (13), 1036–1044. https://doi.org/10.1002/jcc.23576.

Hao Zhu, H., Tropsha, A., Fourches, D., Varnek, A., Papa, E., Gramatica, P., Ö berg, P., Dao, P., Cherkasov, A., Tetkol, V., 2008. 'Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis' -. J. Chem. Inf. Model. 48, 766–784, 2008.

Haupt, R.L., Haupt, S.E., 2004. Practical Genetic Algorithms.

Jaworska, J.S., et al., 2003. Summary of a Workshop on Regulatory Acceptance of (Q) SARs for Human Health and Environmental Endpoints, vol. 111. Environmental Health Perspectives. NUMBER 10 August 2003.

Kang, Y.K., Jhon, M.S., 1982. Additivity of atomic static polarizabilities and dispersion coefficients. Theor. Chim. Acta 61 (1), 41–48. https://doi.org/10.1007/BF00573863.

Kolesnyk, S.D., 2017. Risk assessment of chemicals in food and in silico toxicology (short overview)' Theoretical and experimental medicine. Inter Collegas. ISSN: 2409-9988 4 (1), 2017.

Lin, L., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45 (1), 255–268. https://doi.org/10.2307/2532051.

Oecd, 2004. OECD principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationships models. Biotechnology 1–2. November. www.oecd.org/dataoecd/33/37/37849783.pdf.

Oecd, 2007. Guidance document on the validation of (quantitative) structure-activity relationship [(Q)sar] models. Transport 2, 1–154. https://doi.org/10.1787/9789264085442-en. February.

Oecd, 2015. G20/OECD Principles of Corporate Governance. OECD Publishing, Paris. https://doi.org/10.1787/9789264236882-en.

Pearlman, R.S., Smith, K.M., 1999. Metric validation and the receptor-relevant subspace concept. J. Chem. Inf. Comput. Sci. 39 (1), 28–35.

Pubchem Fingerprints, 2018. ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt.

PubChem Project. https://pubchem.ncbi.nlm.nih.gov/.

Qsarins, Qsar INSubria. no date. www.qsar.it.

Ranganatha, N., Kuppast, I.J., 2012. A review on alternatives to animal testing methods in drug development. Int. J. Pharm. Sci. 4, 28–32.

COMMISSION REGULATION (EU) No 283/2013 of 1 March 2013 Setting Out the Data Requirements for Active Substances, in Accordance with Regulation (EC) No 1107/2009 of the European Parliament and of the Council Concerning the Placing of Plant Protection Products on the Market

Regulation (EC) No 1107/2009 of the European Parliament and of the Council of 21 October 2009 Concerning the Placing of Plant Protection Products on the Market and Repealing Council Directives 79/117/EEC and 91/414/EEC

Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 Concerning the Registration, Evaluation, Authorisation and

Restriction of Chemicals (REACH), Establishing a European Chemicals Agency, Amending Directive 1999/45/EC and Repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as Well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC

Rücker, C., Rücker, G., Meringer, M., 2007. y-Randomization and its variants in QSPR/QSAR. J. Chem. Inf. Model. 47 (6), 2345–2357. https://doi.org/10.1021/ci700157b.

Sangion, A., Gramatica, P., 2016. Hazard of pharmaceuticals for aquatic environment: prioritization by structural approaches and prediction of ecotoxicity. Environ. Int. 95, 131–143. https://doi.org/10.1016/j.envint.2016.08.008. Elsevier Ltd.

Scholz, S., et al., 2013. A European perspective on alternatives to animal testing for environmental hazard identification and risk assessment. Regul. Toxicol. Pharmacol. 67 (3), 505–530. https://doi.org/10.1016/j.yrtph.2013.10.003.

Schüürmann, G., et al., 2008. 'External validation and prediction employing the predictive squared correlation coefficient — test set activity mean vs training set activity mean'. J. Chem. Inf. Model. 48 (11), 2140–2145. https://doi.org/10.1021/ci800253u.

Shi, L.M., et al., 2001. QSAR models using a large diverse set of estrogens. Chem. Inf. Comput. Sci. 41 (1), 186–195. https://doi.org/10.1021/ci000066d.

Smarts Theory, 2017. http://www.daylight.com/SMARTS Theory.

Svante, W., Esbensen, K., Geladi, P., 1987. Principal component analysis. Chemometr. Intell. Lab. Syst. 2 (1–3), 37–52. August 1987.

Technical Working Group On Pesticides (Twg), 2012. (Q)uantitative Structure Activity Relationship [(Q)SAR] Guidance Document, pp. 1–186.

Todeschini, R., Consonni, V., 2000. Handbook of Molecular Descriptors. https://doi.org/10.1002/9783527613106.

Todeschini, R., Consonni, V., 2010. Molecular Descriptors for Chemoinformatics.

Todeschini, R., et al., 2009. Molecular Descriptors for Chemoinformatics: Volume I: Alphabetical Listing/Volume II: Appendices, References, 2 Volume Set, 2nd. Revised and Enlarged Edition.

Triebe, J., Worth, A., Janusch Roi, A., Coe, A., 2017. JRC QSAR Model Database: EURL ECVAM DataBase Service on ALternative Methods to Animal Experimentation: to Promote the Development and Uptake of Alternative and Advanced Methods in Toxicology and Biomedical Sciences: User Support & Tutorial, EUR 28713 EN, 2017. Publications Office of the European Union, Luxembourg, ISBN 978-92-79-71406-1. https://doi.org/10.2760/905519. JRC107491.

Tropsha, A., 2010. Best practices for QSAR model development, validation, and exploitation. Molecular informatics 29, 476–488. https://doi.org/10.1002/minf.201000061, 6-7 July 12, 2010.

Verma, Rajeshwar P., Hansch, C., 2005. An approach toward the problem of outliers in QSAR. Bioorg. Med. Chem. 12 (15), 4597–4621. https://doi.org/10.1016/j.bmc.2005.05.002.

Villaverde, J.J., Sevilla-Moran, B., Lopez-Goti, C., Alonso-Prados, J.L., Sandín-España, P., 2017. Computational methodologies for the risk assessment of pesticides in the European union. J. Agric. Food Chem. 65, 2017–2018. https://doi.org/10.1021/acs.jafc.7b00516, 2017.

Villaverde, J.J., Sevilla-Morán, B., López-Goti, C., Alonso-Prados, J.L., Sandín-España, P., 2019. QSAR/QSPR models based on quantum chemistry for risk assessment of pesticides according to current European legislation. SAR QSAR Environ. Res. 31 (1), 49–72. https://doi.org/10.1080/1062936X.2019.1692368, 2020.

Wehrens, R., Putter, H., Buydens, L.M.C., 2000. The bootstrap: a tutorial. Chemometr. Intell. Lab. Syst. 54, 35–52.

Yap, C.W., 2011. PaDEL-Descriptor: an open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. 32 (7), 1466–1474. https://doi.org/10.1002/jcc.21707.