Nereis. Revista Iberoamericana Interdisciplinar de Métodos, Modelización y Simulación	11	15-30	Universidad Católica de Valencia San Vicente Mártir	Valencia (España)	ISSN 1888-8550

Environmental toxicity prediction using computational tools: prediction of potential hazardous effects of chemicals in *Lactuca sativa* seed germination

Predicción de la toxicidad ambiental usando herramientas computacionales: predicción de efectos tóxicos de químicos en la germinación de semillas de *Lactuca sativa*

Fecha de recepción y aceptación: 15 de noviembre de 2018, 3 de diciembre de 2018

Juan Alberto Castillo-Garit[1,2,*†], Yuleidis González Pérez[3 †], Eberts M. Albear[3], Elizabeth Rodríguez[3], Virginia Pérez-Doñate[4] and Facundo Pérez-Giménez[2]

[1] Unidad de Toxicología Experimental. Universidad de Ciencias Médicas de Villa Clara. Santa Clara. Cuba.
[2] Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular. Departamento de Química Física. Facultad de Farmacia. Universitat de València.
[3] Laboratorio de Ecotoxicología del Centro de Toxicología y Biomedicina de Santiago de Cuba.
[4] Departamento de Microbiología. Hospital Universitario de la Ribera (Valencia).
[*] Correspondencia: Universidad de Ciencias Médicas de Villa Clara. Unidad de Toxicología Experimental. Santa Clara. Villa Clara 50200. Cuba. *E-mail*: jacgarit@yahoo.es
[†] Both authors contribute equally to the paper.

ABSTRACT

The main aim of the study was to develop quantitative structure-activity relationship (QSAR) models for the prediction of phytotoxicity effects of chemical compounds on the *Lactuca sativa* seeds germination. A database of 73 compounds, assayed against *L. sativa* and Dragon's molecular descriptors are used to obtain a QSAR model for the prediction of the phytotoxicity. The model is carried out with QSARINS software and validated according to OECD principles. The best model showed good value for the determination coefficient ($R^2 = 0.917$) and others parameters appropriate for fitting ($s = 0.256$ and $RMSE_{tr} = 0.236$). The validation results confirmed that the model has good robustness and stability ($Q^2_{LOO} = 0.874$ and $Q^2_{LMO} = 0.875$), an excellent predictive power ($R^2_{ext} = 0.896$) and was product of a non-random correlation ($R^2_{Y\text{-}scr} = 0.130$ and $Q^2_{Y\text{-}scr} = -0.265$). Finally, we can say that this model is a good predictor tool to predict the toxicity over *L. sativa* of chemical compounds.

KEYWORDS: *ecotoxicity*, *Lactuca sativa*, phytotoxicity*, QSARINS software.*

RESUMEN

El objetivo principal del estudio fue desarrollar modelos cuantitativos de relación estructura-actividad (QSAR) para la predicción de los efectos fitotóxicos de compuestos químicos, en la germinación de las semillas de *Lactuca sativa*. Se utiliza una base de datos de 73 compuestos, ensayados contra *L. sativa* y los descriptores moleculares del programa Dragon para obtener un modelo QSAR para la predicción de la fitotoxicidad. El modelo se lleva a cabo con el software QSARINS y se valida de acuerdo con los principios de la OCDE. El mejor modelo mostró buen

valor para el coeficiente de determinación (R^2 = 0.917) y otros parámetros apropiados para el ajuste (s = 0.256 and $RMSE_{tr}$ = 0.236). Los resultados de la validación confirmaron que el modelo tiene una buena robustez y estabilidad (Q^2_{LOO} = 0.874 and Q^2_{LMO} = 0.875), un excelente poder predictivo (R^2_{ext} = 0.896) y que no fue producto de una correlación casual ($R^2_{Y\text{-}scr}$ = 0.130 and $Q^2_{Y\text{-}scr}$ = -0.265). Finalmente, podemos decir que el modelo es una buena herramienta de predicción para predecir la toxicidad de compuestos químicos sobre *L. sativa*.

PALABRAS CLAVE: *ecotoxicidad*, *fitotoxicidad*, *Lactuca sativa*, *programa QSARINS.*

INTRODUCTION

As a consequence of progress in the chemical and pharmaceutical industries in the last decades, the quantity and the number of xenobiotics have considerably increased. The U.S. agencies, such as the Environmental Protection Agency (U.S. EPA) and the U. S. Food and Drug Administration (U.S. FDA), reported that in recent years people have been exposed to about 63000 chemicals, many of which show a high toxic activity [1]. The detrimental effects caused by these compounds act on both people's health and on the balance of natural and manipulated ecosystems [2].

The environment is regularly exposed to organic chemicals (e. g. phenols, anilines, benzenes derivatives, etc.) as well as mixtures of several chemical elements; through their use in industrial processes, chemical assays in research institutions and domestic use of different substances. Wastewaters usually contain a variety of organic and inorganic chemicals, and the joint actions of the individual components can lead to complicated integrated toxic effects to the environment [3]. Several ecological perturbations are associated with wastewater discharge into fluvial ecosystems, such as the introduction of micropollutants to stream flow or the loss of invertebrate biodiversity [4, 5]. Consequently, the development of tools able to assess potential hazardous effects of chemicals on living organisms needs to receive attention. Therefore, information about the toxicity of industrial organic chemicals to the environment is of interest [6].

The experimental tests provide most reliable data on the effects of chemicals, but they involve much time consumption and extensive resources, which makes it difficult to research great numbers of potential toxic compounds [7]. Therefore, quantitative structure-activity/toxicity relationships (QSAR/QSTR) studies provide an invaluable tool in the prediction of environmental toxicity directly from the molecular structure of compounds [8]. The QSAR/QSTR studies offer the advantages of higher speed and lower cost, especially when compared to experimental testing [9], and, in recent years, the predictions from computer models have been widely used in modern toxicological research, as they are an important alternative for obtaining experimental evidence and play an important role in evaluating the toxicity of chemicals [10].

Among the different endpoints used to evaluate the potential toxicological impact of chemical exposure to terrestrial organisms, *Lactuca sativa* was chosen because is a very used test species; due to its sensitivity to different organic compounds, its high-frequency use in phytotoxicological testing and its handiness [11, 12]. Also, lettuce (*L. sativa*) has agricultural importance and is one of the species recommended by the EPA, FDA and OECD (Organization for Economic Cooperation and Development) [11].

Furthermore, our research group have a large experience, of more than a decade, developing QSAR models to predict different physicochemical, chemical, pharmacokinetical, toxicological as well as biological properties of different compounds [13-19]. The main objective of the present study was to propose a new and better mathematical model for prediction of potential hazardous effects of chemicals in *L. sativa* seed germination process using QSAR INSubria software (QSARINS) [20-22]. This approach offers an alternative to conventional methods used to predict the toxicity by identifying the relationship between the chemical structure and its toxicity.

MATERIALS AND METHOD

Chemical Database and Descriptor Calculation

The general dataset used in this study assay was obtained from previously published papers [23, 24]; it consists of 73 compounds, assayed against *L. sativa*. The dataset was divided in training set (56 compounds) to fit the model the external prediction set (18 compounds), also known as test set. The experimental values of the molecules of the entire dataset are shown in Table 1.

For this study, we employed Dragon software [25] to calculate the molecular descriptors (MDs). This program allows calculating several families (0-2) of molecular descriptors such as: constitutional indices, functional group counts, 2D-autocorrelations, topological indices and molecular properties among others. After that, those MDs with values constant or near to constant were removed, thus they were not used in further analysis. Consequently, we finally used 313 descriptors to perform the genetic algorithm (implemented in QSARINS software) for variable selection.

QSAR-MLR Model Development

For the development of the multiple linear regression (MLR) model QSARINS software version 2.2.2 was used, which has been developed in the QSAR Research Unit, University of Insubria, Italy. It allows obtaining MLR models by ordinary least squares (OLS) method [21]. The model was developed based on criteria such as fitting (highest R^2), robustness (greatest Q^2_{LOO}), stability (lowest R^2–Q^2_{LOO}), and low correlation of descriptors (lowest K_{XX}), amongst others, so that it has a small difference between fitting, cross-validation and external validation parameters.

Table 1. Structures and experimental values[a] of the data set

No	Smile	Log CL_{50}	No	Smile	Log CL_{50}
1	ClC1=CC=CC=C1O	1.818	45	CC(O)C	2.826
2	ClC1=CC(O)=CC=C1	1.650	46	CC(C)CO	2.485
3	ClC1=CC=C(O)C(Cl)=C1	1.381	47	O=C(O)C1=CC=C(CC)C(CC)=C1C(O)=O	1.664
4	ClC1=CC(Cl)=CC(O)=C1	1.112	48	O=C(O)C1=CC=C(CCCC)C(CCCC)=C1C(O)=O	0.584
5	ClC1=CC(Cl)=C(Cl)C(O)=C1	1.002			
6	ClC1=C(O)C(Cl)=CC(Cl)=C1	1.079	49	COP(OC)(SCN2C(C1=CC=CC=C1N=N2)=O)=S	1.530
7	CC1=CC(Cl)=CC=C1O	1.516			
8	CC1=CC(O)=CC=C1Cl	1.362	50	CCOP(OCC)(OC1=NN(C2=CC=CC=C2)C=N1)=S	1.146
9	OC1=CC=CC=C1O	2.428			
10	OC2=CC=C1C=CC=CC1=C2	1.554	51	NCCN	3.829
11	CC1=CC=CC=C1O	1.914	52	CCCNCCC	2.048
12	CC1=CC(O)=CC=C1	1.909	53	CCCCNCCCC	1.492
13	ClC1=CC=CC=C1	1.487	54	C=CC(N)=O	3.172
14	ClC1=CC=C(Cl)C=C1	1.160	55	CC(NC1=CC=CC=C1)=O	2.293
15	ClC1=C(Cl)C(Cl)=CC=C1	0.863	56	C1=NC=CC=C1	2.538
16	ClC1=CC=C(Cl)C(Cl)=C1	0.920	57	NC1=CC=CC=C1N	2.778
17	ClC1=CC(Cl)=CC(Cl)=C1	0.858	58	O=CC1=CC=CC=C1	2.140
18	ClC1=CC=C(Cl)C(Cl)=C1Cl	0.623	59	O=C(O)CCCCC	2.020
19	ClC1=C(Cl)C=C(Cl)C(Cl)=C1	0.536	60	O1C=CC=C1	2.207
20	ClC1=CC(Cl)=C(Cl)C(Cl)=C1Cl	0.368	61	C1(C2=CC=CC=C2)=CC=CC=C1	0.887
21	ClC1=CC=CC=C1[N+]([O-])=O	1.775	62	C1=CC=CS1	1.981
22	ClC1=CC=CC([N+]([O-])=O)=C1	1.693	63	CC3(O)C2CC1C(N(C)C)C(C(C(N)=O)=C(O)C(O)1C(C2=C(O)C4=C3C=CC=C4O)=O)=O	1.158
23	ClC1=C(Cl)C([N+]([O-])=O)=CC=C1	1.386			
24	NC1=CC=CC=C1	2.418			
25	NC1=CC=CC=C1Cl	1.938	64	CC1=CC(C)=NC(NS(=O)(C2=CC=C(N)C=C2)=O)=N1	2.196
26	NC1=CC=CC(Cl)=C1	1.948			
27	NC1=CC=C(Cl)C=C1Cl	1.516	65	CCN2C=C(C(O)=O)C(C1=CC(F)=C(N3CCNCC3)C=C12)=O	1.694
28	NC1=CC=C(Cl)C(Cl)=C1	1.559			
29	NC1=CC(Cl)=CC(Cl)=C1	1.458	66	CCC1C(C)(O)C(O)C(C)C(C(C)CC(C)(O)C(OC3C(O)C(N(C)C)CC(C)O3)C(C)C(OC2CC(C)(OC)C(O)C(C)O2)C(C)C(O1)=O)=O	1.838
30	NC1=CC(Cl)=C(Cl)C=C1Cl	1.194			
31	NC1=C(Cl)C=C(Cl)C=C1Cl	1.160			
32	NC1=CC(Cl)=C(Cl)C(Cl)=C1Cl	0.959			
33	NC1=C(Cl)C(Cl)=CC(Cl)=C1Cl	0.882	67	OCC(NC(C(Cl)Cl)=O)C(O)C1=CC=C([N+]([O-])=O)C=C1	2.310
34	CC1=CC=CC=C1	1.540			
35	C12=CC=CC=C1C=CC=C2	1.266	68	O=C(O)CC1=CC=CC=C1NC2=C(Cl)C=CC=C2Cl	2.943
36	C23=C1C(CC3)=CC=CC1=CC=C2	0.968			
37	CC1=CC=CC=C1C	1.352	69	O=C(C)NC1=CC=C(O)C=C1	3.450
38	C=CC1=CC=CC=C1	1.434	70	OC1=CC=CC=C1C(O)=O	2.230
39	CCCCCCC	0.738	71	OC(COC2=C1C=CC=CC1=CC=C2)CNC(C)C	2.344
40	CC(Cl)(Cl)Cl	1.655			

No	Smile	Log CL_{50}	No	Smile	Log CL_{50}
41	ClC1(Cl)C(Cl)(Cl)CCCC(Cl)1Cl	1.117	72	CCCC(NC1=CC(C(C)=O)=C(OCC(O)CNC(C)C)C=C1)=O	3.676
42	Cl/C(Cl)=C(Cl)/Cl	1.218			
43	ClC1=C(Cl)C(Cl)=C(Cl)C(Cl)1Cl	0.431	73	CC(C)(NCC(O)COC1=NSN=C1N2CCOCC2)C	3.753
44	CC(O)CO	3.292			

[a] the values of Log CL_{50} are expressed in ug/ml.

Validation of model

An important aspect of the QSAR model is validation. Consequently, the obtained model must be carefully checked and thoroughly validated. The QSARINS offers several tools to confirm that the model meets the standards, set by the OECD for the development, validation, acceptance and use of QSAR models in order to increase the confidence in the reliability of data predicted by it. Based on

these principles, the QSAR-MLR model must meet the following criteria: 1) a defined end point; 2) an unambiguous algorithm; 3) a defined domain of applicability; 4) appropriate measures of goodness-of-fit, robustness and predictivity; and 5) a mechanistic interpretation, if possible [26]. In a previous paper we give a complete explanation of all these aspects; for details see reference [27].

In this work we use some cross-validation techniques for internal validation. First, *leave-one-out* (LOO) exercises were done because, with the disturbance of a single compound in a small database, one gets criteria of its robustness. Later, a *leave-many-out* (LMO) technique was developed, which allows studying the behavior of the model when a greater number of compounds are excluded. In order to demonstrate that the model is not the result of a casual correlation, the randomization (*Y-Scrambling*) procedure was also applied. In this process, the answers are randomly located, so that there is no correlation with the descriptors and, as a result, the model performance should decay dramatically. In order to prove the model predictivity, an external validation was performed. For this, we use the prediction set, which has never been used in the model calculation.

Acute Toxicity Test with *L. sativa* Seeds

The acute toxicity test was carried out in the ecotoxicology laboratory of the Center of Toxicology and Biomedicine (TOXIMED) in Santiago de Cuba. We followed the procedure previously described by Sobrero and Ronco in 2004 [28], to develop acute toxicity test (120-h exposure) with *L. sativa* seeds; it was used for assessing the phytotoxic effects of the chemical compounds on the seed germination process; the toxicity was assessed as inhibition of the elongation of the radicle and hypocotyl. This test has been recommended and implemented by different agencies for environmental protection for the ecotoxicological assessment of environmental samples and pure compounds and also for evaluating the phytotoxic effect of pesticides on non-target species for the registration of these compounds [29, 30].

RESULTS AND DISCUSSION

Develop of the QSAR-MLR Model

Several models were obtained while exploring the best combinations of molecular descriptors which show a high correlation with the response variable (Log CL_{50}).Therefore, taking into account the principle of parsimony (describe the maximum information with the least number of descriptors), an analysis of the model's parameters was done. Based on it, a QSAR-MLR model was developed to evaluate the toxic effects with seven variables. In the development of the model compounds the number 64 was detected as statistical outlier and removed; once rejected the model was obtained. The equation and the statistical parameters of the best model are the following:

$$\textbf{Log } CL_{50} = -17.4396(\pm3.288) +19.910(\pm2.67)*\textbf{Mi} -0.773(\pm0.162)*\textbf{RCl} -3.699(\pm0.502)*\textbf{X0Av} +0.048(\pm0.007)*\textbf{P\_VSA\_s\_5} -0.264(\pm0.028)*\textbf{SssCH2} +1.414(\pm0.194)*\textbf{B02[C-S]} -0.743(\pm0.202)*\textbf{B09[N-O]} \quad (1)$$

R^2= 0.917	R^2_{adj}= 0.904	R^2-R^2_{adj}= 0.013	LOF= 0.101
K_{XX}= 0.315	Delta K= 0.052	$RMSE_{tr}$= 0.236	MAE_{tr}= 0.190
RSS_{tr}= 3.003	CCC_{tr}= 0.957	*s*= 0.256	F= 72.231

The number of compounds used as the training set was 56 compounds: R^2 is the coefficient of determination, R^2_{adj} is adjusted R^2, *s* is standard error of estimate, F is variance ratio, LOF is Friedman lack of fit [31, 32] and K_{xx} is the correlation among descriptors [33]. Delta K is the difference of the correlation between the descriptors (K_x) and the descriptors plus the responses (K_{xy}), $RMSE_{tr}$ is Root-Mean-Square Error in fitting (for the training set), MAE_{tr} is Mean Absolute Error in fitting (calculated on the training set), RSS_{tr} is the Residual Sum of Squares in the fitting (also for the training set) and CCC_{tr} is the concordance correlation coefficient calculated over the training set [34-36].

The model developed in this study showed a R^2 value of 0.917, indicating an appropriate fit to model the toxicity in *L. sativa*; this means that the model explains almost the 92 % of the experimental variance. The LOF parameter showed a low value (0.101), indicating that there was no overfitting risk. The correlation between the descriptors of the model is low because the K_{xx} is small (0.315), so we can say that there is little redundant information in the selected descriptors. In addition, the correlation between the descriptors and the modeled response is appropriate, in accordance with the Delta K parameter (0.052), with a small error on the calculations of training and parameters estimation ($RMSE_{tr}$= 0.236; MAE_{tr}= 0.190; s= 0.256). The Figure 1A shows the scatter plot of the predicted response against experimental, in which it is evident that the compounds are located next to the diagonal line.

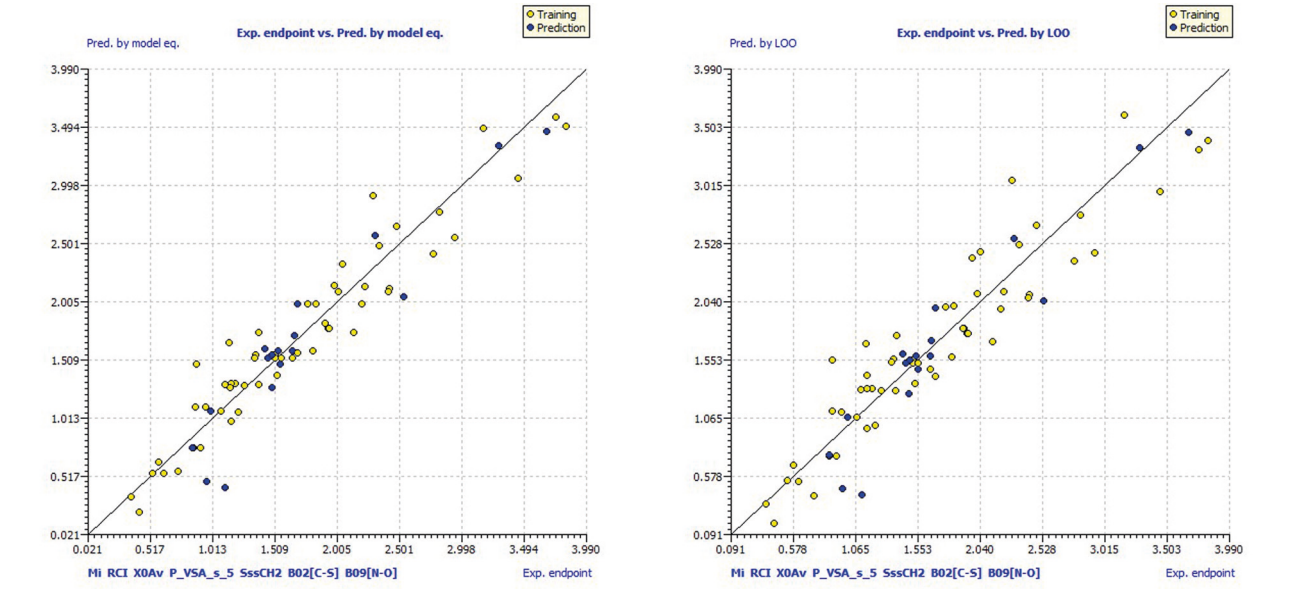

Figure 1. Scatter plot of experimental Log (CL50) versus predicted by model. **A**) predicted by the model (Eq. 1) and **B**) predicted in by LOO croos validation..

Validation of the developed Model

The principal importance of the validation is to prove the predictability and the robustness of the model. In this work we performed both: internal (LOO, LMO and *Y*-scrambling) and external (prediction set) validation experiments only for the final model. According to the results of the *LOO cross-validation*, it can be stated that internal predictions are good and $Q^2_{LOO} = 0.874$ has a value very close to R^2 value, so the model is considered internally stable and robust, with a small error in the predictions ($RMSE_{cv} = 0.289$ and $MAE_{cv} = 0.233$). The Figure 1B shows the values predicted by LOO *vs*. experimental Log **(**CL_{50}**)** values for the training and test sets.

The model performance of internal validation by LMO ($Q^2_{LMO} = 0.875$) is very similar to the obtained with LOO, although it should be noticed that this technique is more useful with larger databases. The Figure 2 shows Q^2_{LMO} *vs.* K_{XY} (correlation between the descriptors and toxicity). Notice that Q^2_{LMO} values are similar to each other and comparable to K_{XY} values, corroborating the good fit and stability of the model.

The last experiment of the internal validation was the *Y-scrambling* procedure, as we pointed out before; it was done to demonstrate that the model is not the result of a casual correlation. Here, the answers were placed at random, so that there is no correlation with the descriptors. As a consequence of this, the model performance decays dramatically. The values of R^2 and Q^2 of every iteration, and their averages ($R^2_{Y\text{-}scr}$ and $Q^2_{Y\text{-}scr}$) offer the criteria that the model is good, since these parameters are ever lower with regard to the values of the model ($R^2_{Y\text{-}scr} = 0.130$ and $Q^2_{Y\text{-}scr} = -0.265$). The $R^2_{Y\text{-}scr}$ and

$Q^2_{Y\text{-}scr}$ values against R^2 and Q^2 of the model are shown in Figure 3. Notice that the values of R^2 and Q^2 of the model are far from the values obtained for those parameters in the *Y-scrambling* experiment, which indicates that the model is not obtained as a result of a random correlation.

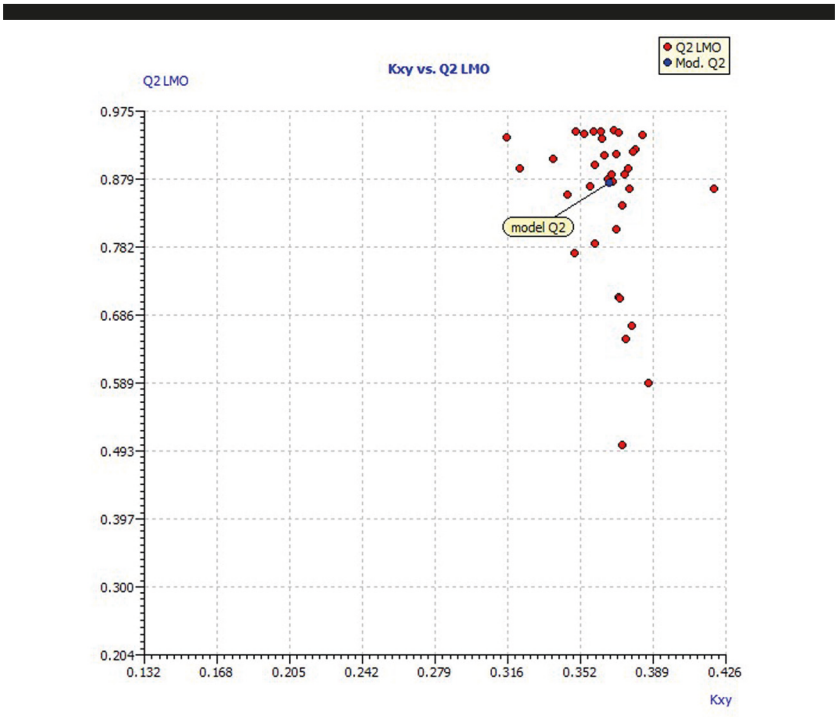

Figure 2. Plot of LMO validations compared with the original model.

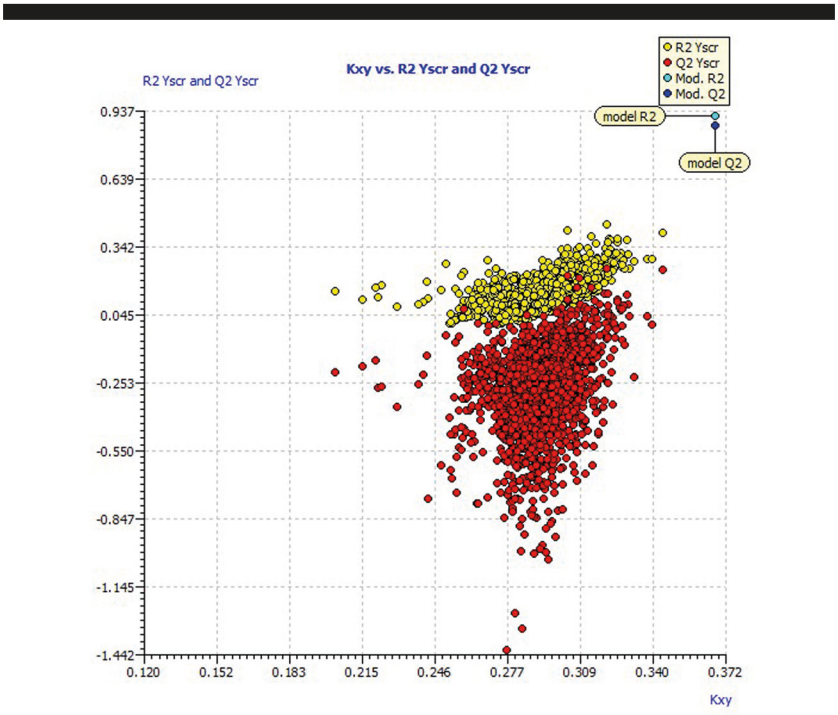

Figure 3. Plot of *Y-scrambled* model compared with the original model.

Table 2. Experimental and predicted values (expressed as Log CL_{50}) of the toxicity in *L. Sativa*

No.	Experimental[a]	Prediction[b]	Residual[c]	No.	Experimental[a]	Prediction[b]	Residual[c]
1	1.818	1.588	-0.231	38*	1.434	1.608	0.174
2*	1.650	1.588	-0.063	39	0.738	0.563	-0.175
3	1.381	1.303	-0.078	40	1.655	1.533	-0.122
4	1.112	1.303	0.191	41*	1.117	0.427	-0.690
5*	1.002	1.079	0.077	42	1.218	1.067	-0.151
6	1.079	1.079	0.000	43	0.431	0.221	-0.210
7	1.516	1.558	0.042	44*	3.292	3.340	0.048
8	1.362	1.558	0.196	45	2.826	2.780	-0.046
9	2.428	2.125	-0.303	46	2.485	2.656	0.171
10*	1.554	1.482	-0.072	47*	1.664	1.719	0.055
11	1.914	1.826	-0.088	48	0.584	0.647	0.063
12	1.909	1.826	-0.083	49	1.530	1.379	-0.151
13*	1.487	1.279	-0.208	50	1.146	1.660	0.514
14	1.160	0.991	-0.169	51	3.829	3.504	-0.325
15	0.863	0.763	-0.100	52	2.048	2.334	0.286
16	0.920	0.763	-0.157	53*	1.492	1.556	0.064
17*	0.858	0.763	-0.095	54	3.172	3.487	0.315

No.	Experimental[a]	Prediction[b]	Residual[c]	No.	Experimental[a]	Prediction[b]	Residual[c]
18	0.623	0.545	-0.078	55	2.293	2.915	0.622
19	0.536	0.545	0.009	56*	2.538	2.050	-0.488
20	0.368	0.350	-0.019	57	2.778	2.421	-0.357
21	1.775	1.994	0.219	58	2.140	1.745	-0.395
22*	1.693	1.994	0.301	59	2.020	2.096	0.076
23	1.386	1.752	0.366	60	2.207	1.995	-0.212
24	2.418	2.092	-0.326	61	0.887	1.483	0.596
25*	1.938	1.783	-0.155	62	1.981	2.148	0.167
26	1.948	1.783	-0.165	63	1.158	1.276	0.118
27	1.516	1.530	0.014	64	2.196	*-outlier-*	*-outlier-*
28	1.559	1.530	-0.029	65	1.694	1.576	-0.118
29*	1.458	1.530	0.072	66	1.838	1.989	0.151
30	1.194	1.310	0.116	67*	2.310	2.576	0.266
31	1.160	1.310	0.150	68	2.943	2.560	-0.384
32	0.959	1.112	0.153	69	3.450	3.065	-0.385
33	0.882	1.112	0.230	70	2.230	2.137	-0.094
34*	1.540	1.588	0.048	71	2.344	2.490	0.146
35	1.266	1.296	0.030	72*	3.676	3.467	-0.209
36*	0.968	0.477	-0.491	73	3.753	3.586	-0.167
37	1.352	1.531	0.179				

[a] Experimental values (See Table 1). [b] Predicted values using Eq. 1. [c] Residual values: log(Obsd)–log(Pred). * Compounds used as prediction set.

However, it has been recognized that the only way to establish the real predictivity power of any model is through an external validation [37]. The use of an external prediction set is the best way of validating a QSAR model. The result of our model for the prediction showed excellent results (R^2_{ext} = 0.896, $RMSE_{ext}$ = 0.269, MAE_{ext} = 0.199, $PRESS_{ext}$ = 1.300, Q^2-F1 = 0.869, Q^2-F2 = 0.868, Q^2-F3 = 0.892, CCC_{ext} = 0.939, r^2m_aver = 0.797, r^2m_delta = 0.097). Where, R^2_{ext} is the external determination coefficient [37]; $RMSE_{ext}$ is the Root-Mean-Square Error in external prediction; MAE_{ext} is the Mean Absolute Error in external prediction; $PRESS_{ext}$ is the Predictive Residual Sum of Squares (external validation); Q^2-F1 [38], Q^2-F2 [39], Q^2-F3 [40] are the explained variances in external prediction; CCC_{ext} is the Concordance Correlation Coefficient [34-36] and r^2_m_aver and r^2_m_delta are the Roy criteria: average and delta [41]. The predictions of compounds in the external set can be seen in Figure 1 (blue circles) and Table 2.

Aplicability Domain

A critical aspect in chemometrics and QSAR studies, for either classification or regression models, is the definition of the applicability domain (AD), because only predictions for compounds that fall

within the domain of applicability may be considered reliable. In this work we use the *leverage* (*h*) and standardized residual approach, described in the literature [42]. The Figure 4 shows the graph of William for the training and prediction sets; as can be seen in this figure, most compounds are within the AD of the model. Only rather few chemicals of the training set showed values of leverage reater than the *critical leverage* value (h^*=0.444), although they showed values of standard deviation within the limits, which means that they should be considered as influential compounds instead of *outliers*. Therefore, we can say that these models, for their applicability domain, can be used with high accuracy.

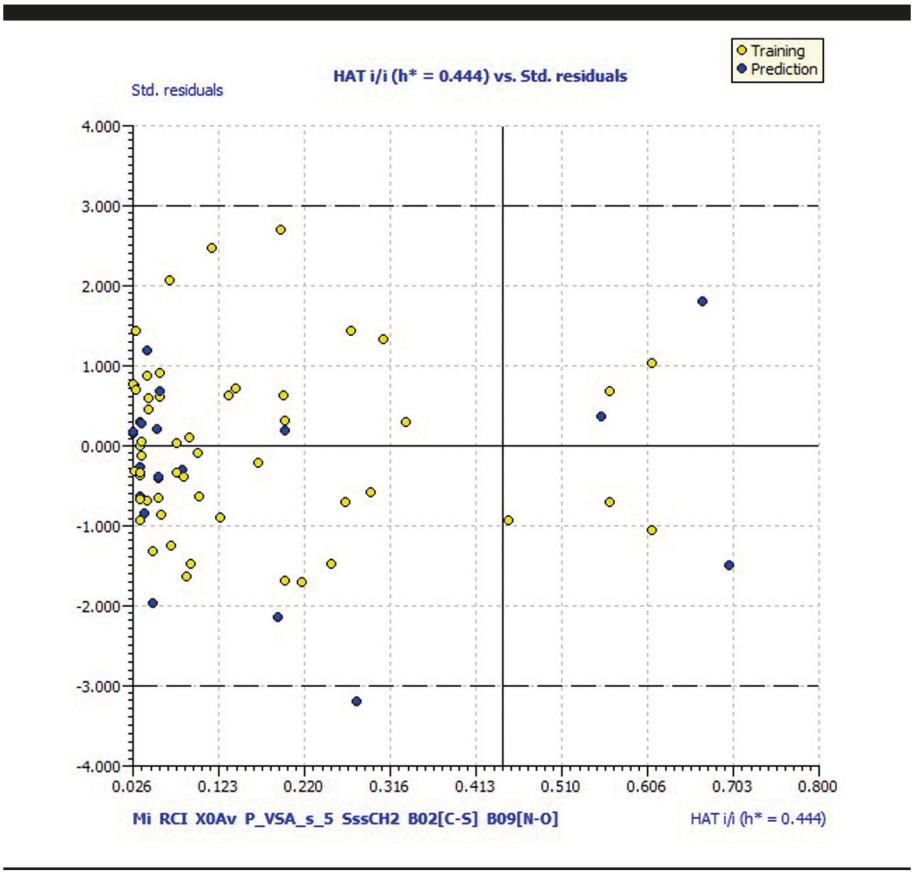

Figure 4. Williams plot. Hat diagonal values versus standardized residuals.

In addition, QSARINS software provides a new approach to determine the application domain: the Insubria graph. This method is based on the *leverage* and model predictions. The Insubria graph is useful to evaluate the position of the molecules, lacking experimental response compared to the structural application domain, and to compare their predictions to those of the compounds having experimental values. We detect that the same compounds that have values over the h^* in the previous graph showed the same behavior with this approach. The Figure 5 shows the Insubria graph for the applicability domain.

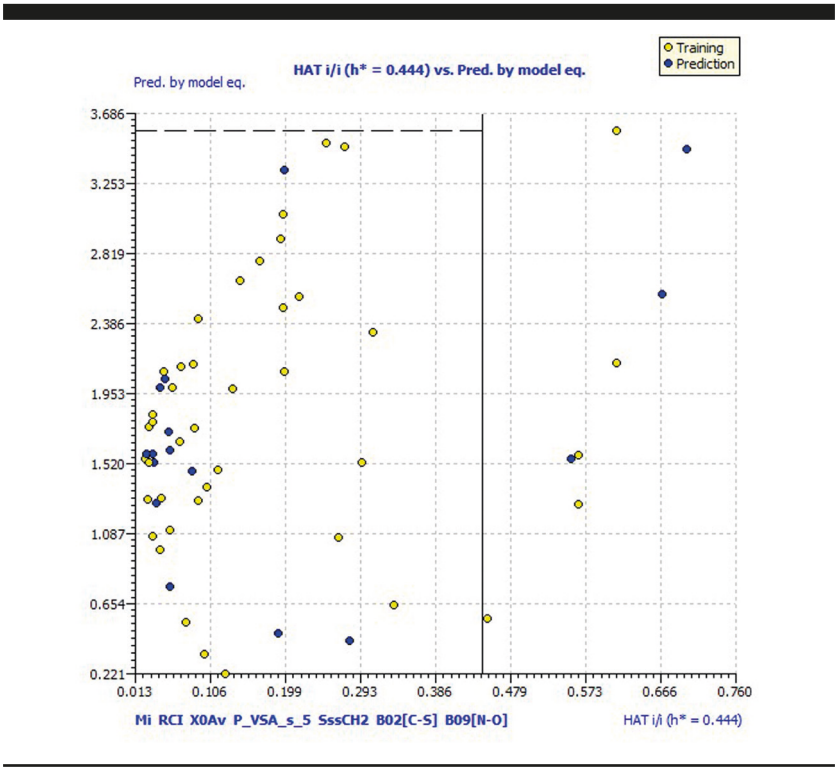

Figure 5. Insubria graph. Hat diagonal values versus predicted data.

Experimental Evaluation

Finally, some chemical compounds, used routinely in the ecotoxicology assays of the laboratory of the TOXIMED, were experimentally tested against *L. sativa*. We carried out the assay as explained above and the results are presented in the following table, as well as the prediction obtained by the model (Eq. 1) for each compound.

Table 3. Experimental results of the evaluated compounds

Compound	Experimental[a]	Prediction[b]
Methanol	4.670	3.244
Ethanol	4.170	2.967
Chromotropic acid	2.673	3.147
Phenol	1.974	1.900
p-nitroaniline	1.419	2.587

[a] Experimental values obtained in the assay in our lab.
[b] Predicted values using Eq. 1.

As can be seen, the compounds were sorted in decreasing order of toxicity (according to the experimental results). We can classify as more toxic those compounds with value of Log CL_{50}<3.0 and as less toxic those with values of Log CL_{50}>3.0; taking this into account, the computational model predicts the more toxic compounds better. Notice that the prediction of two compounds, phenol and chromotropic acid are good and showed low residual values; it also predicts the p-nitroaniline acceptably well (with a residual about 1.1). However, the prediction of the least toxic compounds (ethanol and methanol) has not been so good (residuals being greater than 1.5). For both compounds the model overestimates the toxicity when compared to the experimental values; the increment in the difference could be caused because in the development of the experiment perhaps there was a loss by evaporation (notice that both compounds are volatile) and the value of the assay could underestimate the real toxicity of the compound. Thus, the experiment should be carried out in the future taking into account this inconvenience, as well as one should evaluate other chemical of interest for the above mentioned laboratory. In general, we can say that this model is a good predictor tool to predict the toxicity over *L. sativa* of chemical compounds and that, for future works, it could become an important tool for the work in the ecotoxicology laboratory of TOXIMED.

CONCLUSION

In the present work, a QSAR-MLR model was developed by using molecular descriptors calculated using the Dragon software, which adequately predicts the acute toxicity against seeds of *L. sativa*, so as to assess the phytotoxic effects of the chemical compounds on the seed germination. The model was preformed with QSARINS software and extensively validates following the principles established by the OECD; the robustness of the model was tested through internal validation techniques (*LOO*, *LMO*, and *Y-scrambling*), and its predictability was checked through an external prediction set (external validation). The obtained model showed good statistical parameters for training and a prediction set concluding that the proposed computational tools are efficient to predict the acute toxicity against *L. sativa*.

LITERATURE CITED

[1] Radisky DC, Levy DD, Littlepage LE, Liu H, Nelson CM, Fata JE, *et al*. Rac1b and reactive oxygen species mediate MMP-3-induced EMT and genomic instability. Nature. 2005;436:123.

[2] D'Abrosca B, Fiorentino A, Izzo A, Cefarelli G, Pascarella MT, Uzzo P, *et al*. Phytotoxicity evaluation of five pharmaceutical pollutants detected in surface water on germination and growth of cultivated and spontaneous plants. J Environ Sci Health A. 2008;43(3):285-94.

[3] Deng M, Zhang Y, Quan X, Na C, Chen S, Liu W, *et al*. Acute toxicity reduction and toxicity identification in pigment-contaminated wastewater during anaerobic-anoxic-oxic (A/A/O) treatment process. Chemosphere. 2017;168:1285-92.

[4] Sánchez-Morales M, Sabater F, Muñoz I. Effects of urban wastewater on hyporheic habitat and invertebrates in Mediterranean streams. Sci Total Environ. 2018;642:937-45.

[5] Ziajahromi S, Neale PA, Leusch FDL. Wastewater treatment plant effluent as a source of microplastics: review of the fate, chemical interactions and potential risks to aquatic organisms. Wat Sci Tech. 2016;74(10):2253-69.
[6] Castillo-Garit JA, Marrero-Ponce Y, Escobar J, Torrens F, Rotondo R. A novel approach to predict aquatic toxicity from molecular structure. Chemosphere. 2008;73:415-27.
[7] Dieguez-Santana K, Pham-The H, Villegas-Aguilar PJ, Le-Thi-Thu H, Castillo-Garit JA, Casañola-Martin GM. Prediction of acute toxicity of phenol derivatives using multiple linear regression approach for Tetrahymena pyriformis contaminant identification in a median-size database. Chemosphere. 2016;165:434-41.
[8] Salahinejad M, Ghasemi JB. 3D-QSAR studies on the toxicity of substituted benzenes to Tetrahymena pyriformis: CoMFA, CoMSIA and VolSurf approaches. Ecotoxicol Environ Safety. 2014;105:128-34.
[9] Netzeva TI, Schultz TW. QSARs for the aquatic toxicity of aromatic aldehydes from Tetrahymena data. Chemosphere. 2005;61(11):1632-43.
[10] Nicolau A, Mota M, Lima N. Effect of different toxic compounds on ATP content and acid phosphatase activity in axenic cultures of Tetrahymena pyriformis. Ecotox Environ Safe. 2004;57(2):129-35.
[11] Wang W. Root elongation method for toxicity testing of organic and inorganic pollutants. Environ Toxicol Chem. 1987;6(5):409-14.
[12] Adema DMM, Henzen L. A comparison of plant toxicities of some industrial chemicals in soil culture and soilless culture. Ecotox Environ Safe. 1989;18(2):219-29.
[13] Castillo-Garit JA, Marrero-Ponce Y, Torrens F, García-Domenech R, Rodríguez-Borges JE. Applications of Bond-Based 3D-Chiral Quadratic Indices in QSAR Studies Related to Central Chirality Codification. QSAR & Comb Sci. 2009;28:1465-77.
[14] Brito-Sánchez Y, Castillo-Garit JA, Le-Thi-Thu H, González-Madariaga Y, Torrens F, Marrero-Ponce Y, *et al*. Comparative study to predict toxic modes of action of phenols from molecular structures. SAR QSAR Environ Res. 2013;24(3):235-51.
[15] Castillo-Garit JA, del Toro-Cortés O, Vega MC, Rolón M, Rojas de Arias A, Casañola-Martin GM, *et al*. Bond-based bilinear indices for computational discovery of novel trypanosomicidal drug-like compounds through virtual screening. Eur J Med Chem. 2015;96:238-44.
[16] Cañizares-Carmenate Y, Mena-Ulecia K, Perera-Sardiña Y, Torrens F, Castillo-Garit JA. An approach to identify new antihypertensive agents using Thermolysin as model: In silico study based on QSARINS and docking. Arab J Chem. 2016. doi: http://dx.doi.org/10.1016/j.arabjc.2016.10.003.
[17] Le-Thi-Thu H, Cañizares-Carmenate Y, Marrero-Ponce Y, Torrens F, Castillo-Garit JA. Prediction of Caco-2 Cell Permeability Using Bilinear Indices and Multiple Linear Regression. Lett Drug Des Discov. 2016;13(2):161-9.
[18] Castillo-Garit JA, Abad C, Casañola-Martin GM, Barigye SJ, Torrens F, Torreblanca A. Prediction of Aquatic Toxicity of Benzene Derivatives to Tetrahymena pyriformis According to OECD Principles. Curr Pharm Des. 2016;22(33):5085-94.

[19] Castillo-Garit JA, Casañola-Martin GM, Le-Thi-Thu H, Pham-The H, Barigye SJ. A Simple Method to Predict Blood-Brain Barrier Permeability of Drug-Like Compounds Using Classification Trees. Med Chem. 2017;13(7):664-9.
[20] Gramatica P, Chirico N, Papa E, Cassani S, Kovarich S. QSARINS, software for QSAR MLR model development and validation: QSAR Res. Unit in Environ. Chem. and Ecotox. Varese (Italy): University of Insubria; 2013. Available at: http://www.qsar.it.
[21] Gramatica P, Chirico N, Papa E, Cassani S, Kovarich S. QSARINS: A new software for the development, analysis and validation of QSAR MLR models. J Comput Chem (Software News and Updates). 2013;34(24):2121-32.
[22] Gramatica P, Cassani S, Chirico N. QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. J Comput Chem (Software News and Updates). 2014;35(13):1036-44.
[23] Hulzebos EM, Adema DMM, Dirven-van Breemen EM, Henzen L, van Gestel CAM. QSARs in phytotoxicity. Sci Total Environ. 1991;109-110:493-7.
[24] Pan M, Chu LM. Phytotoxicity of veterinary antibiotics to seed germination and root elongation of crops. Ecotox Environ Safe. 2016;126:228-37.
[25] Kode SRL. Dragon (software for molecular descriptor calculation) version 7.0.10; 2017. Available at: https://chm.kode-solutions.net.
[26] Worth AP, Bassan A, De Bruijn J, Gallegos Saliner A, Netzeva T, Patlewicz G, *et al*. The role of the European Chemicals Bureau in promoting the regulatory use of (Q)SAR methods. SAR and QSAR in Environmental Research. 2007;18(1-2):111-25.
[27] Cañizares-Carmenate Y, Hernandez-Morfa M, Torrens F, Castellano G, Castillo Garit JA. Larvicidal activity prediction against *Aedes aegypti* mosquito using computational tools. J Vector Borne Dis. 2017;54(2):164-71.
[28] Castillo Morales G, editor. Ensayos toxicológicos y métodos de evaluación de calidad de aguas: Estandarización, intercalibración, resultados y aplicaciones. Canadá: IDRC; 2004.
[29] OECD. Test No. 208: Terrestrial Plant Test: Seedling Emergence and Seedling Growth Test2006.
[30] Greene J, Bartels C, Warren-Hicks W, Parkhurst B. G. L. Protocols for short term toxicity screening of hazardous waste sites. In: Agency USEP, editor. Washington D. C.; 1988.
[31] Friedman JH. Multivariate Adaptive Regression Splines. Ann Stat. 1991;19(1):1-67.
[32] Friedman JH. Rejoinder: Multivariate Adaptive Regression Splines. Ann Stat. 1991;19(1):123-41.
[33] Todeschini R, Consonni V, Maiocchi A. The K correlation index: theory development and its application in chemometrics. Chemom Intell Lab Syst. 1999;46(1):13-29.
[34] Chirico N, Gramatica P. 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. 2011;51(9):2320.
[35] Chirico N, Gramatica P. 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. 2012;52(8):2044.
[36] Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989;45(1):255-68.

[37] Golbraikh A, Tropsha A. Beware of q 2! J Mol Graphics Model. 2002;20(4):269-76.
[38] Shi LM, Fang H, Tong W, Wu J, Perkins R, Blair RM, *et al*. QSAR models using a large diverse set of estrogens. J Chem Inf Comput Sci. 2001;41(1):186-95.
[39] Schüürmann G, Ebert R, Chen J, Wang B, Kühne R. External Validation and Prediction Employing the Predictive Squared Correlation Coefficient - Test Set Activity Mean vs Training Set Activity Mean. J Chem Inf Model. 2008;48(11):2140-5.
[40] Consonni V, Ballabio D, Todeschini R. Comments on the definition of the Q2 parameter for QSAR validation. J Chem Inf Model. 2009;49(7):1669-78.
[41] Ojha PK, Mitra I, Das RN, Roy K. Further exploring rm 2 metrics for validation of QSPR models. Chemom Intell Lab Syst. 2011;107(1):194-205.
[42] Atkinson AC. Plots, Transformations, and Regression. An Introduction to Graphical Methods of Diagnostic Regression Analysis. Clarendon Press; 1985.

ACKNOWLEDGEMENTS

Castillo-Garit, J. A. thanks the program “Estades Temporals per a Investigadors Convidats” for a fellowship to work at Valencia University in 2018. The authors are grateful to Prof. Paola Gramatica and her group of the University of Insubria for providing the QSARINS software for this study.