Neonicotinoids Activity Against Cowpea Aphids by Computational Estimation

L. Crisan[+], A. Borota[+], A. Bora[+] and S. Funar−Timofei[•]

"Coriolan Dragulescu" Institute of Chemistry, Romanian Academy, Bul. Mihai Viteazu 24, 300223 Timisoara, Romania

ARTICLE INFO	ABSTRACT
	In this study, the insecticidal activity against Cowpea aphids (*Aphis craccivora*) of a series of 23 phenylazo, pyrrole-, dihydropyrrole-fused and chain-opening nitromethyleneneonicotinoids was evaluatedby using the multiple linear regression (MLR) and pharmacophore modelling. Conformer insecticide ensembles were modeled using the MMFF94s force field. Minimum energy conformers were employed to calculate structural parameters, which were related to the experimental pLC_{50} values. Several statistical criteria of goodness of fit and predictivity were checked to validate the models. Robust and predictable MLR models were obtained. Further, the Phase module from Schrodinger suite was engaged in the generation of the ligand-based pharmacophore models. The atom-based 3D-QSAR module from the aforementioned software was used for the validation of a best four-point pharmacophore model. The obtained significant statistical parameters attested the pharmacophore model validity. The MLR and pharmacophore models are useful for the prediction of new insecticides with activity against Cowpea aphids.

1   INTRODUCTION

Neonicotinoid insecticides, first introduced in the mid-1990s, are chemicals with a major impact in the economy and the ecosystem of any country. Their high efficiency, low toxicity, broad insecticidal spectra and unique mode of action has turned them into key players for the development of safe insecticides in a short

---

time [1–3]. Insecticides are employed for the eradication of pests such as insects, mites, rodents, birds, etc. Controlling action of pests is very important in order to increase the quality and quantity of the products. Insects, e.g. Cowpea aphids (*Aphis craccivora*), are one of these pests very difficult to control. *Cowpea aphid* is considered to be one of the most serious pests of legumes that cause direct or indirect damages by feeding on the plant's sap [4,5]. The significant increase in resistance to insecticides alongside with the detrimental effect on the bee health and the environment has led to the urgent need for the development of new control strategies and more potent insecticidal agents [2]. The design of new insecticides involves various methods, ranging from the analysis of the already known chemical scaffolds to the high-throughput screening of the candidates in order to find new chemical skeletons with improved activity profile [2,6].

The structure of the first-generation neonicotinoid, imidacloprid [7], was the starting point for the generation of the second (e.g. thiamethoxam [8], clothianidin [9]) and third (e.g. dinotefuran [10], sulfoxaflor [11]) classes of insecticide analogs. The aromatic heterocycle (e.g. pyridine), a flexible linkage, a hydroheterocycle or guanidine/amidine fragment, and an electron-withdrawing part are essential features to synthesize new neonicotinoids. The pyridine unit is considered to be a significant intermediate for agrochemicals and pharmaceuticals, being present in more than 70 products available on the market [11–13].

Computational methods, such as Quantitative Structure-Activity Relationship (QSAR), pharmacophore modeling, protein-structure prediction, virtual screening, molecular docking, are considered as effective and indispensable tools for guiding the design of novel insecticides by reducing time, resources and costs. These methods are also an alternative solution to minimize the demand for animal test and to provide a rapid assessment of the potential impacts of chemicals on human health and the environment. In this context, the toxicity prediction of insecticides remains of continuous interest in QSAR modeling [14–16]. The literature survey shows a large number of QSAR approaches [14, 16–18] dedicated to the toxicity of neonicotinoids, but only a few refer to neonicotinoids with activity against Cowpea aphids [19–23].

In this regard, a data set consisting of lethal concentration, 50% (LC_{50}) of 23 phenylazo, pyrrole-, dihydropyrrole-fused and chain-opening nitromethyleneneonicotinoids to Cowpea aphids was used to establish the multiple linear regression (MLR) and pharmacophore models. The developed models were evaluated using various statistical parameters. Based on the analysis of the developed models, important information in connection with toxicity was obtained and helped us to better understand the neonicotinoids activity against Cowpea aphids. Furthermore, these QSAR and pharmacophore models may provide a

better way to evaluate and predict the toxicity of other untested neonicotinoid analogs, before they manifest side effects on both human and the ecosystem.

2 METHODS

2.1. DATASET AND DESCRIPTORS

A series of 23 phenylazo, pyrrole-, dihydropyrrole-fused and chain-opening nitromethyleneneonicotinoid derivatives (Table 1) having insecticidal activity against the Cowpea aphids (*Aphis craccivora*) was collected from literature [24, 25]. Experimental insecticidal lethal concentration, 50% (LC_{50}) values were converted to pLC_{50} values, further used as the dependent variable.

Table 1. Insecticide structures, experimental ($pLC_{50 \, EXP}$) insecticidal activity and descriptors of the neonicotinoids used in the MLR1.
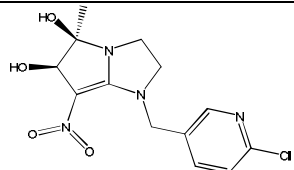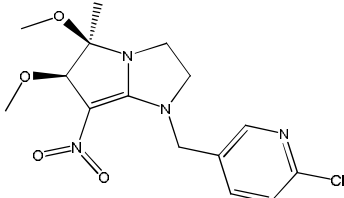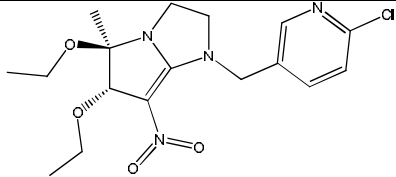
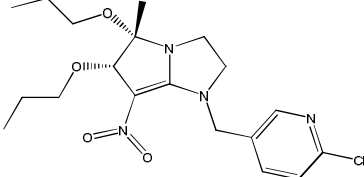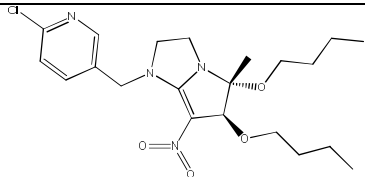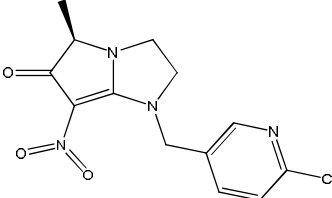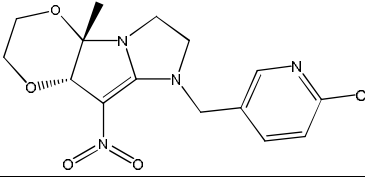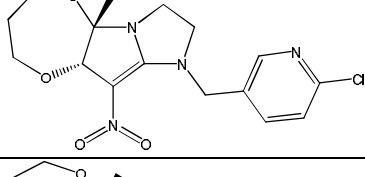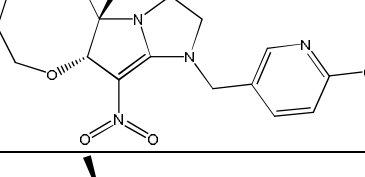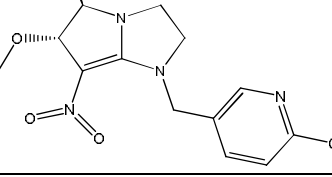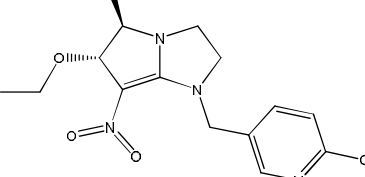| No	Structure	$pLC_{50 \, EXP}$	JGI2[a]	BLI[b]	Mor32v[c]
1		5.21	0.096	0.89	-0.197
2*		5.70	0.097	0.911	-0.198
3		5.80	0.092	0.972	-0.181
4**		5.71	0.088	1.007	-0.27
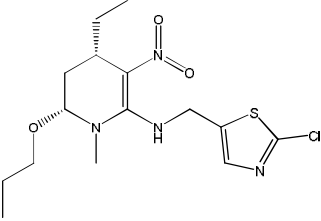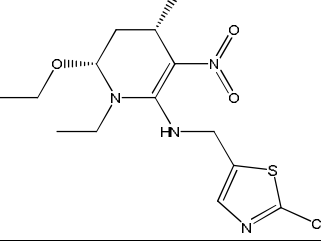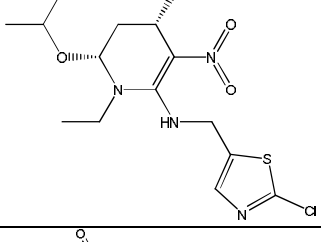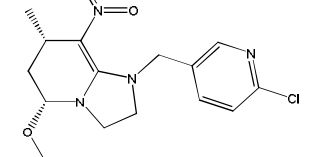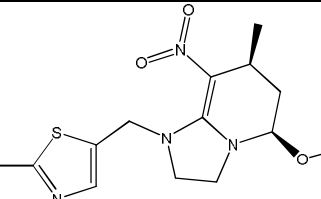
Table 1 (Continued).

No	Structure	pLC$_{50\ EXP}$	JGI2[a]	BLI[b]	Mor32v[c]
5		5.11	0.084	1.038	-0.276
6*		3.85	0.078	0.902	-0.101
7		4.55	0.087	0.906	-0.227
8*,**		4.52	0.085	0.927	-0.24
9**		4.41	0.083	0.947	-0.131
10		4.35	0.076	0.93	-0.195
11**		3.96	0.074	0.963	-0.185

Table 1 (Continued).

No	Structure	pLC$_{50\ EXP}$	JGI2[a]	BLI[b]	Mor32v[c]
12		4.16	0.072	0.984	-0.214
13*,**		3.97	0.07	1.003	-0.229
14		3.79	0.083	1.07	0.062
15*,**		4.25	0.081	1.088	0.026
16**		4.07	0.078	1.074	0.008
17		3.91	0.083	1.063	0.07
18		3.98	0.081	1.092	0.065

Table 1 (Continued).

19**		4.41	0.078	1.109	0.036
20		3.82	0.078	1.118	0.056
21**		3.86	0.074	1.12	0.053
22		3.58	0.074	0.951	-0.027
23		3.72	0.07	1.029	-0.076

* test compounds in the MLR models

** test compounds in the pharmacophore model.

[a]JGI2 - mean topological charge index of order 2.

[b]BLI - Kier benzene-likeliness index.

[c]Mor32v- 3D-MoRSE - signal 32 / weighted by atomic van der Waals volumes.

The neonicotinoid structures were first modeled using molecular mechanics calculations, with the MMFF94s force field included in the OMEGA (version 2.5.1.4, OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com)

software [26, 27]. Starting structures were generated using SMILES notation. A maximum of 400 conformers per compound, an energy cutoff of 10 kcal/mol relative to a global minimum identified from the search were employed for the conformer ensemble generation. Any conformer having an RMSD fit outside 0.5 Å to another conformer was removed to avoid redundant structures.

Conformers of minimum energy thus obtained were further used to calculate structural 0D, 1D, 2D, and 3D descriptors. Thus, 1611 parameters were computed using the DRAGON software (Dragon Professional 5.5 (2007), Talete S.R.L., Milano, Italy), which includes 22 types of descriptors (constitutional, functional groups counts, topological descriptors: BLI (Kier benzene-likeliness index), Burden eigenvalues, eigenvalue-based indices, Galvez descriptors (topological charge indicies): JGI2 (mean topological charge index of order2), Getaway descriptors: R4u (R autocorrelation of lag 4 / unweighted) and HGM (geometric mean on the leverage magnitude), Randić descriptors (Randić molecular profiles), RDF descriptors (radial distribution function descriptors; MWC (Molecular walk counts, path counts – atomic and molecular descriptors) and 3D-MoRSE (3D-molecule representation of structure based on electron diffraction descriptors): Mor26p (3D-MoRSE signal 26 / weighted by atomic polarizabilities), Mor32v (3D-MoRSE - signal 32 / weighted by atomic van der Waals volumes), Mor15p (3D-MoRSE - signal 15 / weighted by atomic polarizabilities), Mor21u (3D-MoRSE - signal 21 / unweighted), information indices, edge adjacency indices, topological charge indices, connectivity indices, 2D-autocorrelations, molecular properties, 2D binary fingerprints, and 2D frequency fingerprints.). The InstantJchemsoftware was used for structure database management, search and prediction (InstantJchem 15.10.0, 2012, ChemAxon (http://www.chemaxon.com), and to calculate additional structural descriptors.

2.2. MULTIPLE LINEAR REGRESSION (MLR)

Multiple linear regression (MLR) [28] calculations were performed using the QSARINS v.2.2 program [29].

In the MLR approach, one experimental variable y_k, or dependent variable (e.g. the insecticidal activity), is correlated with one or several independent variables x_i (e.g. molecular descriptors), using the equation:

$$y_i = b_0 + \sum_{j=1}^{m} b_j \cdot x_{ij} + e_i \tag{1}$$

where b_j represents the partial regression coefficients and e_i the deviations and residuals that account for the disagreement between the observed responses y_i and the predicted results [30] (with *n* being the number of compounds (i = 1 ... n) and

m number of predictors). The regression coefficients b_j are calculated by minimizing the sum of the squared residuals, using a least squares procedure, to give the smallest possible sum of squared differences between the values of real and predicted dependent variable.

The Genetic Algorithm [31] (GA) was used for the 1611 structural descriptors calculated for the 23 neonicotinoid compounds to select variables in the multiple linear regression models. In the QSARINS package the following parameters were used: the RQK fitness function [32] with leave-one-out cross-validation [33] correlation coefficient as a constrained function to be optimized, a crossover/mutation trade-off parameter of T = 0.5 and a model population size of P = 50.

2.3. PHARMACOPHORE

The atom-based pharmacophore models were realized whit the aid of PHASE [34, 35] software from Schrödinger package (http://www.schrödinger.com). For this purpose, the dataset of 23 neonicotinoids with lethal concentration, 50% (LC_{50}) insecticidal activity determined against *Aphis craccivora* were united from two literature studies [24, 25] For the preparation and optimization of these structures, the LigPrep program (LigPrep, Schrödinger, LLC, New York, NY, 2014, http://www.schrödinger.com) incorporated within PHASE (default settings) was used. The threshold for activity was set at > 5 for the active compounds and < 3.9 for inactive ones.

All the pharmacophore characteristics available in the PHASE [34, 35] (hydrophobic, hydrogen bond donor, hydrogen bondacceptor, aromatic rings, positive and negative ionisable feature) were taken in consideration for the generation of the common pharmacophore hypotheses. The 3D QSAR module was involved in statistical validation of the pharmacophore hypotheses obtained. The QSAR analysis was carried out using two partial least-squares (PLS) factors, and the split into training and test set was done randomly, using 60% of compounds for the training set.

2.3. MODEL VALIDATION

The MLR models were internally validated using the following robustness parameters (Eqs. 2 to 9): squared correlation coefficient for fitting (r_{tr}^2), adjusted r^2 (r_{adj}^2), leave-one-out cross-validation (q_{LOO}^2), Y-scrambling parameters (r_{scr}^2 & q_{scr}^2) [36], root-mean-square errors for training set ($RMSE_{tr}$) and mean absolute error for training set (MAE_{tr}) [37], concordance correlation coefficient for training set (CCC_{tr}) [38], Fischer test for the training set (F), standard deviation of regression

for the training set (SD) and q_{LMO}^2 leave-more-out (LMO) cross-validation (carried out for 30% of data out of training, each run). In Y-scrambling the process was randomly mixed 2000 times.

$$r_{tr}^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$

$$r_{adj}^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)/(n-m-1)}{\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)} \tag{3}$$

$$q_{LOO}^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_{i/i} - y_i)}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

$$RMSE_{tr} = \sqrt{\frac{\sum_{i=1}^{n_{tr}}(y_i - \hat{y}_i)^2}{n_{tr}}} \tag{5}$$

$$MAE_{tr} = \frac{\sum_{i=1}^{n_{tr}}|y_i - \hat{y}_i|}{n_{tr}} \tag{6}$$

$$CCC_{tr} = \frac{2\sum_{i=1}^{n_{tr}}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sum_{i=1}^{n_{tr}}(y_i - \bar{y})^2 + \sum_{i=1}^{n_{tr}}(\hat{y}_i - \bar{\hat{y}})^2 + n_{tr}(\bar{y} - \bar{\hat{y}})^2} \tag{7}$$

$$F = \frac{ssy/df_1}{ssy/df_2}; \; ssy = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}); \; df_1 = m+1; \; df_2 = n_{tr} - m - 2 \tag{8}$$

$$SD = \sqrt{sse/df_2}; \; sse = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{9}$$

where: y_i - the experimental values of the dependent variable; $\hat{y}_i$ - the calculated dependent variable values; $\bar{y}$ - the average of the experimental dependent variable values; $\bar{\hat{y}}$ - the average of the predicted dependent variable values; $\hat{y}_{i/i}$ - the predicted value of the response calculated excluding the ith element from the model during computation; $test$ refers to prediction set and tr refers to training set; n - the number of objects; m – the number of predictor variables; ssy - the variance of the model; df_1 - the degrees of freedom of the model; df_2 - the degrees of freedom of the input data; sse - the sum of squared errors.

The domain of applicability was checked using the Williams plots (standardized cross-validated residuals versus leverage (Hat diagonal) values) [29]. A threshold of residual value greater than 3 times the value of standard error in calculation was employed for outlier detection.

The model's predictive power was tested using the following external parameters (eqs. 10 to 15): Q_{F1}^2 [39]; Q_{F2}^2 [40]; Q_{F3}^2 [41] (with acceptable values higher than 0.7) and the concordance correlation coefficient for the test set (CCC_{test}) [38] (with a minimum threshold of 0.85), root-mean-square errors for the test set ($RMSE_{test}$) and mean absolute error for the test set (MAE_{test}).

$$Q_{F1}^2 = 1 - \frac{\sum_{i=1}^{n_{test}}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{test}}(y_i - \bar{y}_{tr})^2} \tag{10}$$

$$Q_{F2}^2 = 1 - \frac{\sum_{i=1}^{ntest}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{ntest}(y_i - \bar{y}_{test})^2} \tag{11}$$

$$Q_{F3}^2 = 1 - \frac{[\sum_{i=1}^{ntest}(y_i - \hat{y}_i)^2]/n_{test}}{[\sum_{i=1}^{ntr}(y_i - \bar{y}_{tr})^2]/n_{tr}} \tag{12}$$

$$CCC_{test} = \frac{2\sum_{i=1}^{ntest}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sum_{i=1}^{ntest}(y_i - \bar{y})^2 + \sum_{i=1}^{ntest}(\hat{y}_i - \bar{\hat{y}})^2 + n_{test}(\bar{y} - \bar{\hat{y}})^2} \tag{13}$$

$$RMSE_{test} = \sqrt{\frac{\sum_{i=1}^{ntest}(y_i - \hat{y}_i)^2}{n_{test}}} \tag{14}$$

$$MAE_{test} = \frac{\sum_{i=1}^{ntest}|y_i - \hat{y}_i|}{n_{test}} \tag{15}$$

In addition, other statistical measures were used to check the model predictivity [42] (eqs. 16 to 20): 1) the squared correlation coefficient (r_{test}^2) between the predicted and observed activities, as well as squared correlation coefficient by cross-validation (q_{LOO}^2); 2) the coefficient of determination for linear regressions with intercepts set to zero, i.e. r_0^2 (predicted versus observed activities), and $r_0^{'2}$ (observed versus predicted activities); 3) slopes k and k' of the above mentioned two regression lines. All these measures were applied over the test set compounds.

$$q_{LOO}^2 > 0.5 \tag{16}$$

$$r_{test}^2 > 0.6 \tag{17}$$

$$\frac{(r^2 - r_0^2)}{r^2} < 0.1 \quad \text{and} \quad 0.85 \le k \le 1.15 \tag{18}$$

$$\frac{(r^2 - r_0^{'2})}{r^2} < 0.1 \quad \text{and} \quad 0.85 \le k' \le 1.15 \tag{19}$$

$$\left| r_0^2 - r_0^{'2} \right| < 0.3 \tag{20}$$

and r_m^2 [43] (eq. 21) (values higher than 0.6 were considered as acceptable).

$$r_m^2 = r_{test}^2 \left(1 - \sqrt{r_{test}^2 - r_0^2} \right) \tag{21}$$

To test model collinearity variance inflation factors (VIF) [44] were calculated. It was considered that if VIF shows values > 10, or if the tolerance remains below 0.10, then the model present multicollinearity [45]. For VIF < 5, no significant collinearity is present.

3   RESULTS AND DISCUSSION

3.1. MLR RESULTS

The multiple linear regression approach was used to correlate the insecticidal activity of the series of the neonicotinoidanalogues with their calculated structural parameters.

The structural data were normalized based on the autoscaling method, which can be described as:

$$XT_{mj} = \frac{X_{mj} - \overline{X}_m}{S_m} \tag{10}$$

where for each variable *m*, XT_{mj} and X_{mj} are the values *j* for the variable *m* after and before scaling respectively, $\overline{X}_m$ is the mean and S_m the standard deviation of the variable.

The neonicotinoid derivatives were divided into training and test sets randomly. Five compounds were taken out of the total number of compounds: **2**, **6**, **8**, **13**, **15**.

Variables were selected using the genetic algorithm, with the leave-one-out fit criterion as a constrained function to be optimized. Five MLR models having acceptable statistical results and the predictive power are listed in Tables 2 to 4. The internal and external validation criteria show that they are satisfactory in the fitting, and have good predictive power. Among them, model MLR1 with good fitting and predictivity results is the most stable one, from a statistical point of view. Similar RMSE values were observed for fitting, cross-validation, and test sets. For this last model, the correlation matrix, variance of inflation factors, tolerances, and standardized coefficients are presented in Table 5. The experimental versus predicted pLC_{50} values plots for the MLR1 model fitting are included in Figure 1a.

The applicability domain of the MLR models was checked using the Williams plot. For the best MLR1 model, for the model fitting and leave-one-out cross-validation, the plots are presented in Figure 2. No outliers or influential points are present in this model.

The internal and external validation criteria were checked for MLR1 model. The small difference of CCC values between the training and test sets of 0.1% was noticed, which demonstrates that this model is able to predict the response for chemicals not used in the model development (validation set) just as they do for chemicals used to find the relationship (training set).
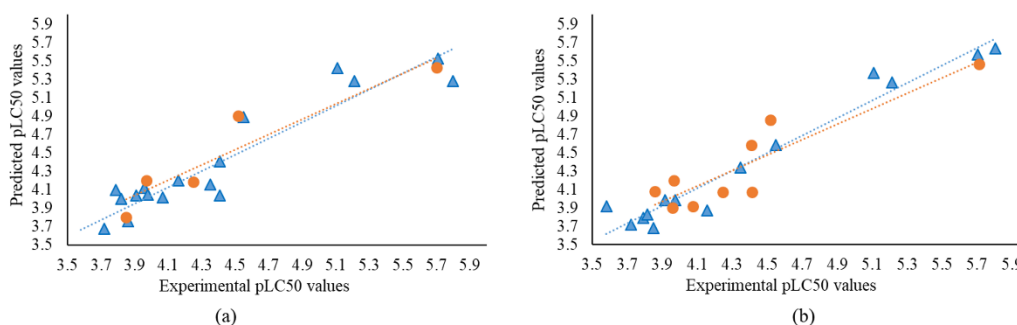
**Figure 1.** Experimental versus predicted pLC$_{50}$ values for the final MLR1 model (a) and atom-based QSAR for AHHR.62 hypothesis (b). Blue triangles and orange circles indicate training and test set compounds, respectively.

Table 2. Calculated internal validation criteria of the MLR and pharmacophore (AHHR.62) models*.

Model	r_{tr}^2	q_{LOO}^2	q_{LMO}^2	r_{adj}^2	RMSE$_{tr}$	MAE$_{tr}$	CCC$_{tr}$
MLR1	0.882	0.797	0.766	0.857	0.226	0.179	0.937
MLR2	0.902	0.803	0.783	0.881	0.206	0.170	0.949
MLR3	0.924	0.875	0.859	0.908	0.181	0.157	0.961
MLR4	0.876	0.786	0.754	0.850	0.231	0.211	0.934
MLR5	0.873	0.786	0.756	0.845	0.234	0.192	0.932
AHHR.62	0.954	0.808	-	-	0.158	0.1099	0.976

Model	RMSE$_{CV}$	MAE$_{CV}$	CCC$_{CV}$	r_{scr}^2	q_{scr}^2	SD	F
MLR1	0.296	0.234	0.895	0.179	-0.380	0.256	34.876
MLR2	0.292	0.229	0.902	0.175	-0.409	0.233	42.977
MLR3	0.233	0.201	0.935	0.174	-0.390	0.205	57.086
MLR4	0.304	0.276	0.886	0.180	-0.385	0.262	32.993
MLR5	0.304	0.250	0.889	0.172	-0.386	0.266	31.977
AHHR.62	-	-	-	-	-		

* r_{tr}^2- correlation coefficient; q_{LOO}^2 - leave-one-out correlation coefficient; q_{LMO}^2 - leave-more-out correlation coefficient; RMSE$_{tr}$-training root-mean-square errors; MAE$_{tr}$- training mean absolute error; CCC$_{tr}$- training the concordance correlation coefficient; RMSE$_{CV}$-leave-one-out cross-validation root-mean-square errors; MAE$_{CV}$- leave-one-out cross-validation mean absolute error; CCC$_{CV}$- leave-one-out cross-validation the concordance correlation coefficient; r_{scr}^2 - scrambled r^2; q_{scr}^2 - scrambled cross-validated q^2; SD-standard error of estimates; F-Fischer test.
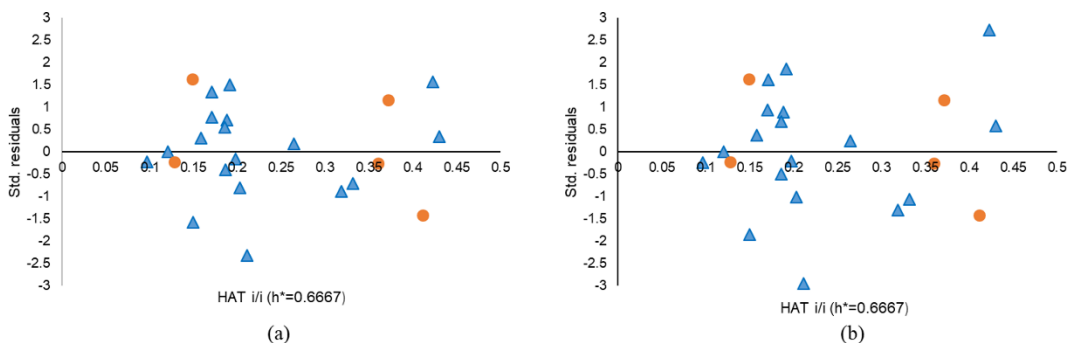
Figure 2. Williams plot predicted by the final MLR1 (a) and leave-one-out cross-validation procedure (b). Blue triangles and orange circles indicate training and test set compounds, respectively.

Table 3. Calculated external validation criteria calculated for the MLR models and the descriptors selected in the MLR models*.

Model	Q_{F1}^2	Q_{F2}^2	Q_{F3}^2	RMSE$_{ext}$	MAE$_{ext}$	CCC$_{ext}$	r_m^2	Descriptors included in the model
MLR1	0.874	0.871	0.868	0.239	0.201	0.927	0.817	BLI JGI2 Mor32v
MLR3	0.858	0.855	0.851	0.253	0.190	0.925	0.829	JGI2 Mor26p R4u
MLR2	0.841	0.837	0.833	0.269	0.237	0.923	0.892	JGI2 Mor26p HGM
MLR4	0.868	0.865	0.861	0.245	0.199	0.934	0.796	BLI JGI2 Mor15p
MLR5	0.816	0.812	0.807	0.288	0.248	0.903	0.783	JGI2 Mor21u Mor32v
AHHR.62	0.807	0.806	0.899	0.233	0.217	0.896	0.780	-

* $Q_{F1}^2, Q_{F2}^2, Q_{F3}^2, r_m^2$-external validation parameter; RMSE$_{ext}$-root-mean-square errors; MAE$_{ext}$-mean absolute error; CCC$_{ext}$-the concordance correlation coefficient; JGI2 - mean topological charge index of order2; Mor26p - 3D-MoRSE - signal 26 / weighted by atomic polarizabilities; R4u - R autocorrelation of lag 4 / unweighted; HGM - geometric mean on the leverage magnitude; BLI - Kier benzene-likeliness index; Mor32v - 3D-MoRSE - signal 32 / weighted by atomic van der Waals volumes; Mor15p - 3D-MoRSE - signal 15 / weighted by atomic polarizabilities; Mor21u - 3D-MoRSE - signal 21 / unweighted.

Table 4. Other external validation criteria [42].

Model	r_{test}^2	$\dfrac{r^2 - r_0^2}{r^2}$	$\dfrac{r^2 - r_0^{'2}}{r^2}$	k	k'	$\left\lvert r_0^2 - r_0^{'2} \right\rvert$
MLR1	0.880	0.006	0.052	0.991	1.006	0.041
MLR2	0.840	0.017	0.000	1.030	0.969	0.036
MLR3	0.911	0.001	0.005	1.036	0.963	0.004
MLR4	0.873	0.009	0.002	0.999	0.998	0.006
MLR5	0.821	0.003	0.024	0.988	1.008	0.018
AHHR.62	0.807	0.001	0.037	1.000	0.997	0.029

Table 5. Correlation matrix, standardized coefficients (Std. coeff.), variance inflation factors (VIF) and tolerances of the descriptors of the MLR1 model.

	BLI	JGI2	Mor32v	Std. coeff.	VIF	Tolerance
BLI	1.000			0.429	2.247	0.445
JGI2	-0.100	1.000		0.661	1.148	0.871
Mor32v	0.607	0.156	1.000	-0.782	2.224	0.450

The LOO validation highlights that the model is stable, not obtained by chance; in fact, the difference between $r_{training}^2$ and q_{LOO}^2 is small: 8.5 % in case of the MLR1 model. This model is internally predictive with differences between q_{LMO}^2 and q_{LOO}^2 of 3.1%. The absence of chance correlation in the MLR1 model is proved by the low values of the Y-scramble parameters (Table 2).

Topological indices used are mainly based on distances between atoms, calculated by the number of separating bonds and are thus considered through-bond indices [46]. Topological charge indices were proposed to evaluate the charge transfer between pairs of atoms, and therefore the global charge transfer in the molecule [47].

Topological charge indices are derived from the eigenvalues of adjacency/square matrix with a series of weighting schema, including weighted by atomic masses, atomic van der Waals volumes, atomic Sanderson electronegativities, and atomic polarizabilities. The JGI2 (mean topological charge index of order 2) descriptor in the MLR1 linear equation has a positive coefficient value. The increase of its value will increase the pLC_{50} values.

Table 6. Score of different parameters of the best 20 hypotheses*.

No	ID	Survival	Survival -inactive	Site	Vector	# Matches	Activity	Inactive
1	AHHR.62	3.492	1.587	0.85	0.951	5	5.111	1.905
2	AAHH.344	3.443	1.515	0.79	0.944	5	5.705	1.928
3	AAAH.59	3.535	1.690	0.82	0.954	5	5.711	1.844
4	AAHH.337	3.601	1.734	0.87	0.972	5	5.805	1.868
5	AAHH.340	3.484	1.692	0.82	0.937	5	5.111	1.792
6	AAHH.290	3.559	1.644	0.80	0.967	5	5.805	1.915
7	AHHR.54	3.640	2.163	0.90	0.959	5	5.805	1.478
8	AHHR.56	3.539	1.668	0.90	0.951	5	5.111	1.871
9	AAHH.118	3.589	2.104	0.82	0.979	5	5.805	1.486
10	AAHR.33	3.439	1.430	0.76	0.954	5	5.805	2.009
11	AAHR.34	3.439	1.430	0.76	0.954	5	5.805	2.009
12	AAHH.293	3.586	1.694	0.84	0.942	5	5.805	1.892
13	AAHR.37	3.179	1.155	0.66	0.867	5	5.705	2.024
14	AAHH.289	3.564	1.695	0.84	0.941	5	5.711	1.868
15	AHHR.47	3.603	1.736	0.87	0.951	5	5.711	1.867
16	AAAH.126	3.483	1.693	0.81	0.945	5	5.111	1.790
17	AAHR.21	3.617	1.791	0.86	0.948	5	5.805	1.826
18	AAHR.12	3.595	1.785	0.85	0.946	5	5.711	1.81
19	AAAH.123	3.535	1.674	0.78	0.968	5	5.805	1.861
20	AAHR.14	3.599	1.786	0.85	0.95	5	5.711	1.813

A - Acceptor; H – Hydrophobic and R - Aromatic ring. Survival score represents the "Weighted combination of the vector, site, volume, and survival scores, and a term for the number of matches"; Survival – inactive score is " Survival score for actives with a multiple of the survival score for inactives subtracted"; Site score measures "how closely the site points are superimposed in an alignment to the pharmacophore of the structures that contribute to this hypothesis, based on the RMS deviation of the site points of a ligand from those of the reference ligand"; Vector alignment score measure "how well the vectors for acceptors, donors, and aromatic rings are aligned in the structures that contribute to this hypothesis, when the structures themselves are aligned to the pharmacophore"; #Matches denotes the number of actives that match every hypothesis; Activity represents the activity in logarithm units for the reference ligand and Inactivity is survival score for inactive copounds[34,35].

The Kier benzene-likeliness index [48] is an aromaticity index calculated from molecular topology [49]. High values of the BLI (Kier benzene-likeliness index) descriptor would be expected to increase the insecticidal activity.

Table 7. The statistical parameters attained for the best atom based 3D-QSAR models using 2 PLS factors

No	ID	SD*	r_{tr}^2 *	F*	P*	RMSE[#]	q2[#]	Pearson-R[#]
1	AHHR.62	0.178	0.954	113.400	4.56E-08	0.232	0.808	0.899
2	AAHH.344	0.175	0.955	117.000	3.86E-08	0.244	0.788	0.918
3	AAAH.59	0.199	0.942	89.300	1.58E-07	0.268	0.744	0.882
4	AAHH.337	0.216	0.931	74.700	3.98E-07	0.279	0.722	0.910
5	AAHH.340	0.191	0.947	97.700	9.95E-08	0.287	0.705	0.902
6	AAHH.290	0.257	0.903	51.300	2.66E-06	0.301	0.676	0.831
7	AHHR.54	0.173	0.956	119.200	3.50E-08	0.303	0.671	0.833
8	AHHR.56	0.165	0.960	131.800	2.07E-08	0.304	0.669	0.858
9	AAHH.118	0.236	0.918	61.700	1.05E-06	0.307	0.663	0.832
10	AAHR.33	0.192	0.946	95.900	1.09E-07	0.318	0.639	0.823
11	AAHR.34	0.192	0.946	95.900	1.09E-07	0.318	0.639	0.823
12	AAHH.293	0.270	0.893	45.800	4.64E-06	0.320	0.635	0.813
13	AAHR.37	0.207	0.937	82.400	2.41E-07	0.321	0.633	0.806
14	AAHH.289	0.269	0.894	46.300	4.38E-06	0.325	0.622	0.818
15	AHHR.47	0.191	0.947	97.700	9.94E-08	0.325	0.622	0.805
16	AAAH.126	0.255	0.905	52.200	2.44E-06	0.326	0.619	0.891
17	AAHR.21	0.232	0.921	64.000	8.76E-07	0.327	0.617	0.806
18	AAHR.12	0.243	0.914	58.100	1.43E-06	0.328	0.616	0.801
19	AAAH.123	0.279	0.886	42.700	6.57E-06	0.330	0.611	0.818
20	AAHR.14	0.240	0.915	59.500	1.26E-06	0.334	0.601	0.803

The parameters with the * symbol refer to the training set and the parameters with the [#]symbol refer to the test set. SD is the standard deviation of regression; r_{tr}^2 represents the regression coefficient; F is the Fisher test, defined as ratio of the model variance to the observed activity variance (variance ratio); P denotes the significance level of variance ratio; RMSE is the RMS error in the test set predictions; q^2 depicts the leave-N-out cross-validated correlation coefficient for the test set (the default N is 1).

3D-MoRSE (Molecule Representation of Structure based on Electron) [50] are geometrical descriptors, being the sums of atom weights with different angular scattering function. They extract information from the 3D atomic coordinates by using the same transform as in electron diffraction studies. They give an idea of how the weighting property is distributed in space. The Mor32v (3D-MoRSE - signal 32 / weighted by atomic van der Waals volumes) descriptor in the MLR1

linear equation has negative coefficient value. Therefore, the increase in its value will decrease the pLC_{50} values.
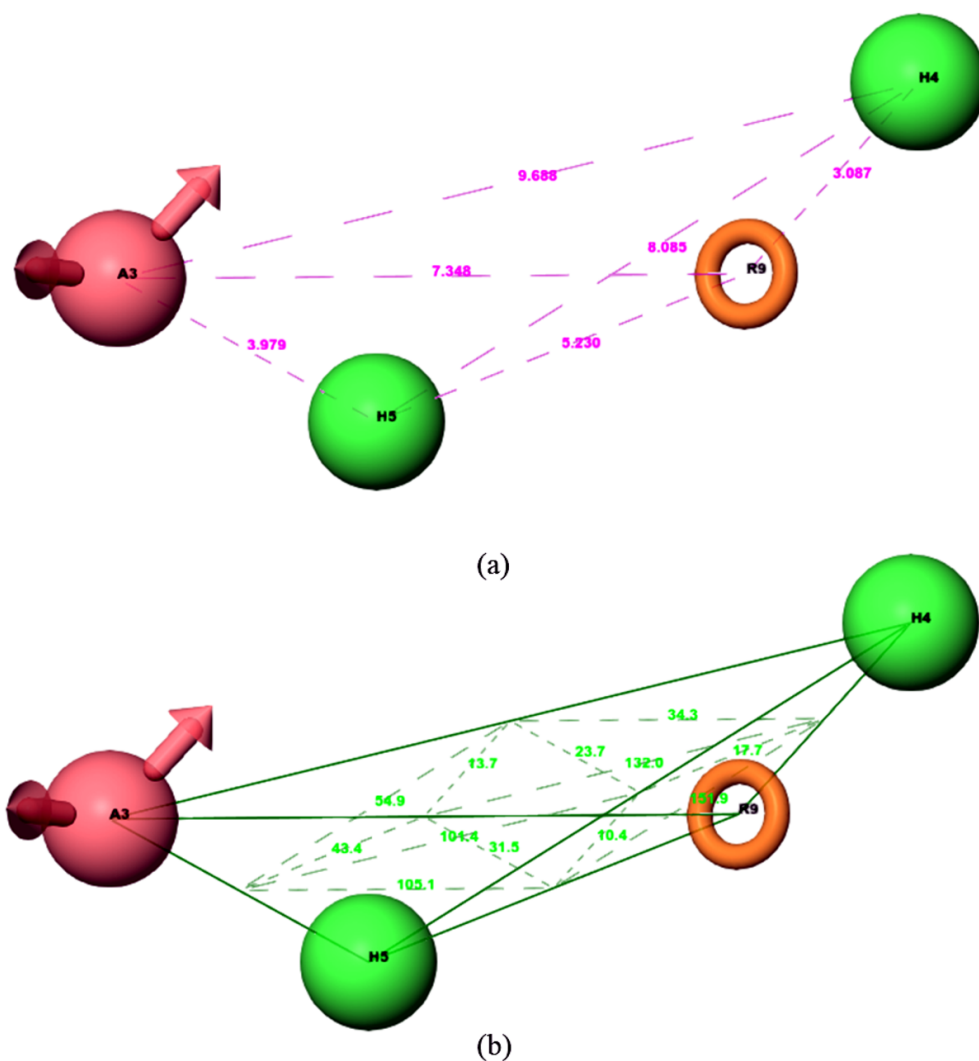
3.2. PHARMACOPHORE RESULTS

The PHASE module has generated 40 four-point pharmacophore hypotheses, which are common for all selected neonicotinoid derivatives. Of them, 20 hypotheses (see Tables 6 and 7) were externally validated using the atom-based 3D QSAR approach. Pharmacophore sites for 23 neonicotinoid derivatives (Table 7) consist of a set of chemical features of PHASE: hydrogen bond acceptors (A), hydrophobic (H), and an aromatic ring (R).

The model AHHR.62 with the highest statistical significance ($r^2 = 0.954$, $q^2_{LOO} = 0.808$ and Person-R [34, 35] = 0.899) contains: one hydrogen bond acceptor, two hydrophobic and one aromatic ring features (Figures 3 and 4) with high survival score (3.492). The predictive abilities of the atom-based 3D QSAR model of AHHR.62 hypothesis are statistically significant, as shown in Tables 3 and 4. The statistical values obtained for the test set (compounds **4, 8, 9, 11, 13, 15, 16, 19** and **21**) proved that the selected QSAR model is stable and predictive. The plot of observed versus predicted pLC_{50} insecticidal activities for the training and test sets obtained for atom-based 3D QSAR model of AHHR.62 hypothesis is portrayed in Figure 1b.
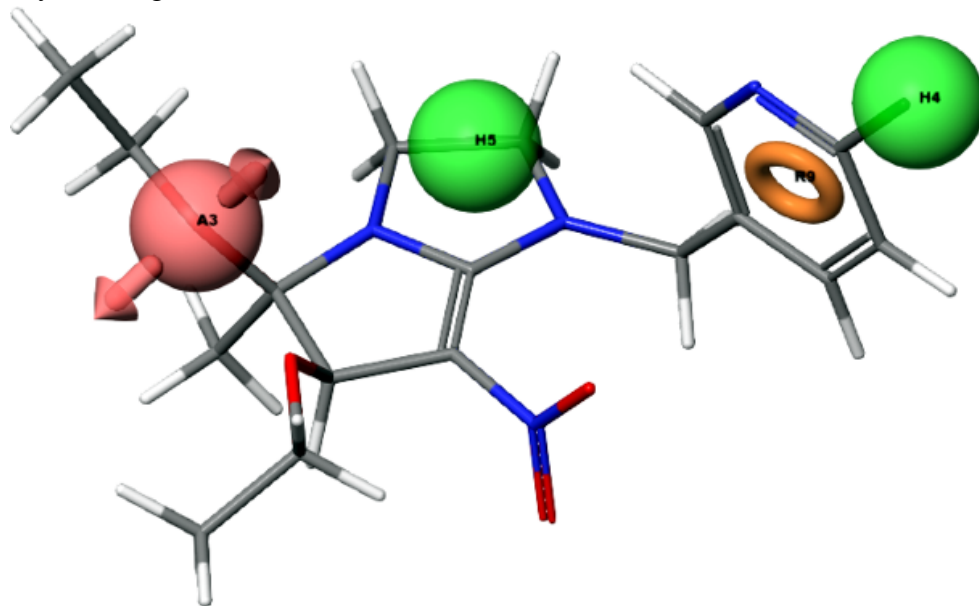

Figure 3. The most active compound **3** on AHHR.62 hypothesis. Hydrogen bond acceptor (A3) is shown by red sphere, hydrophobic features are indicated by green spheres (H4 and H5), and aromatic ring (R9) is displayed by orange circle.

(a)

(b)

Figure 4. Intersite distances (a) and angles (b) between the pharmacophoric points of the best model AHHR.62.

The representation of the hydrophobic and electron withdrawing maps resulted from 3D-QSAR model in the perspective of the best and the least active compound is shown in Figure 5. The positive coefficients for activity are depicted by blue cubes, while the negative contributions are represented in red, indicating the areas were the structural disposition decreases the biological activity. The hydrophobicity, an important feature for our pharmacophore model, as one can see from Figure 5a is better represented by blue cubes around the hydrophobic moieties (the hydrocarbon chains) of the compound **3** compared with those of compound **22**.
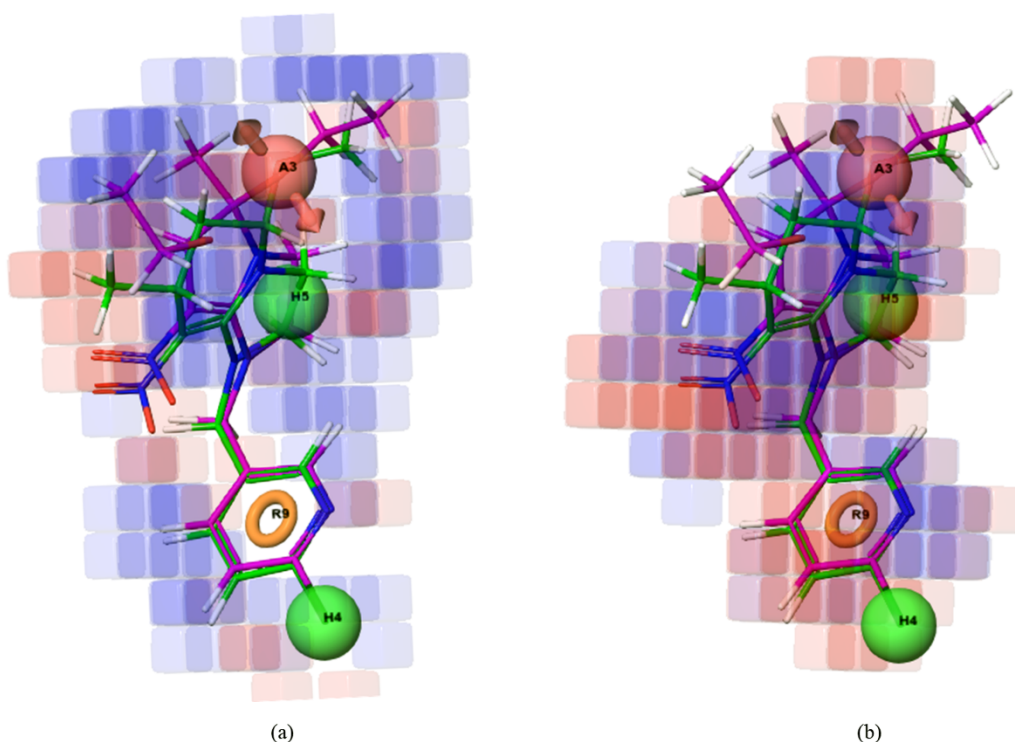
Figure 5. The atom-based 3D-QSAR model visualization for the most active (compound **3** -carbon atoms colored in magenta) and the least active neonicotinoid (compound **22** - carbon atoms represented in green) in the context of hydrophobic (a), and electron withdrawing maps (b).

The blue cubes around the positions 3, 4 and 6 of the pyridine ring show that substitutions here with hydrophobic moieties would be favorable for improving the insecticidal activity. Comparing the disposition of the NO_2 group (from compound **3** and **22**) on the electron withdrawing map (Figure 5b) is noticing differences. The NO_2 group belonging to compound **3** is placed in the favourable region, while the NO_2 group of compound **22** migrates to the undesirable area of red cubs. The nitrogen atom of the pyridine ring, which has an electron acceptor character, seems to be an important feature for the biological activity of this series of insecticides.

4   CONCLUSIONS

The multiple linear regression (MLR) approach and pharmacophore modeling of a series of 23 phenylazo, pyrrole-, dihydropyrrole-fused and chain-opening nitromethyleneneonicotinoids were applied to study their insecticidal activity against the Cowpea aphids. Structures were optimized using molecular mechanics

calculations; the derived structural parameters were correlated with the experimental pLC$_{50}$ values. Stable models with good fitting results and predictive power were obtained. New insecticides active against the Cowpea aphids can be predicted based on the best MLR model, including topological charge index, aromaticity index and 3DMorSE descriptors and the pharmacophore model with one hydrogen bond acceptor, two hydrophobic and one aromatic ring features.

REFERENCES

1. P. Jeschke, R. Nauen, M. Schindler and A. Elbert, Overview of the status and global strategy for neonicotinoids, *J. Agric. Food Chem.* **59** (7) (2011) 2897–2908.
2. J. E. Casida and K. A. Durkin, Neuroactive insecticides: targets, selectivity, resistance, and secondary effects, *Annu. Rev. Entomol*. **58** (2013) 99–117.
3. J. E. Casida, Neonicotinoids and other insect nicotinic receptor competitive modulators: progress and prospects, *Annu. Rev. Entomol*. **63** (2018) 125–144.
4. A. Whittaker, Datasheet – Aphis craccivora, In: Invasive species compendium – CABI, Wallingford, UK, (2018) https://www.cabi.org/isc/datasheet/6192. Accessed 25/02/2018.
5. S. I. Bishara, E. Z. Farm, A. A. Attia and M. A. El-Hariry, Yield losses of faba bean due to aphid attack, *Fabis News Letter* **1** (1984) 16–18.
6. M. Akamatsu, Importance of physicochemical properties for the design of new pesticides, *J. Agric. Food Chem.* **59** (2011) 2909–2917.
7. S. Kagabu, Discovery of imidacloprid and further developments from strategic molecular designs, *J. Agric. Food Chem.* **59** (2010) 2887–2896.
8. P. Maienfisch, H. Huerlimann, A. Rindlisbacher, L. Gsell, H. Dettwiler, J. Haettenschwiler, E. Sieger and M. Walti, The discovery of thiamethoxam: a second-generation neonicotinoid, *Pest. Manag. Sci.* **57**(2001) 165–176.
9. H. Uneme, Chemistry of clothianidin and related compounds, *J. Agric. Food Chem.* **59** (2010) 2932–2937.

10. T. Wakita, K. Kinoshita, E. Yamada, N. Yasui, N. Kawahara, A. Naoi, M. Nakaya, K. Ebihara, H. Matsuno and K. Kodaka, The discovery of dinotefuran: a novel neonicotinoid, *Pest. Manag. Sci.* **59** (2003) 1016–1022.

11. M. R. Loso, Z. Benko, A. Buysse, T. C. Johnson, B. M. Nugent, R. B. Rogers, T. C. Sparks, N. X. Wang, G. B. Watson and Y. Zhu, SAR studies directed toward the pyridine moiety of the sap-feeding insecticide sulfoxaflor (Isoclast™ active), *Bioorg. Med. Chem.* **24** (2016) 378–382.

12. J. E. Casida, Golden age of RyR and GABA-R diamide and isoxazoline insecticides: Common genesis, serendipity, surprises, selectivity, and safety, *Chem. Res. Toxicol.* **28** (2015) 560–566.

13. P. Schäfer, G. Hamprecht, M. Puhl, K.-O. Westphalen and C. Zagar, Synthesis and herbicidal activity of phenylpyridines – A new lead, *Chimia* **57** (2003) 715–719.

14. N. Basant and S. Gupta, QSAR modeling for predicting mutagenic toxicity of diverse chemicals for regulatory purposes, *Environ. Sci. Pollut. Res.* **24** (16) (2017) 14430–14444.

15. M. Hamadache, O. Benkortbi, S. Hanini and A. Amrane, QSAR modeling in ecotoxicological risk assessment: application to the prediction of acute contact toxicity of pesticides on bees (Apis mellifera L), *Environ. Sci. Pollut. Res.* **25** (1) (2018) 896–907.

16. M. Hamadache, O. Benkortbi, S. Hanini, A. Amrane, L. Khaouane and C. S. Moussa, A quantitative structure activity relationship for acute oral toxicity of pesticides on rats: validation, domain of application and prediction, *J. Hazard. Mater.* **303** (2016) 28–40.

17. Y. Xie, W. Peng, F. Ding, S. J. Liu, H. J. Ma and C. L. Liu, Quantitative structure-activity relationship (QSAR) directed the discovery of 3-(pyridin-2-yl)benzenesulfonamidederivatives as novel herbicidal agents, *Pest. Manag. Sci.* **74** (1) (2018) 189–199.

18. S. Kagabu, K. Nishimura, Y. Naruse and I. Ohno, Insecticidal and neuroblocking potencies of variants of the thiazolidine moiety of thiacloprid and quantitative relationship study for the key neonicotinoid pharmacophore, *J. Pestic. Sci.* **33** (1) (2008) 58–66.

19. Z. Tian, X. Shao, Z. Li, X. Qian and Q. Huang, Synthesis, insecticidal activity, and QSAR of novel nitromethylene neonicotinoids with tetrahydropyridine fixed cis configuration and exo-ring ether modification, *J. Agric. Food Chem.* **55** (6) (2007) 2288–2292.

20. M. J. Wang, X. B. Zhao, D. Wu, Y. Q. Liu, Y. Zhang, X. Nan, H. Liu, H. T. Yu, G. F. Hu and L. T. Yan, Design, synthesis, crystal structure, insecticidal activity, molecular docking, and QSAR studies of novel N3-substituted imidacloprid derivatives, *J. Agric. Food Chem*. **62** (24) (2014) 5429−5442.

21. S. Xia, J. Cheng, Y. Feng, X. Shao, H. Luo, Z. Xu, X. Xu and Z. Li, Computational investigations about the effects of hetero-molecular aggregation on bioactivities: a case of neonicotinoids and water, *Chin. J. Chem*. **32** (2014) 324−334.

22. R. Xu, M. Luo, R. Xia, X. Meng, X. Xu, Z. Xu, J. Cheng, X. Shao, H. Li and Z. Li, Seven-membered azabridged neonicotinoids: synthesis, crystal structure, insecticidal assay, and molecular docking studies, *J. Agric. Food Chem*. **62** (46) (2014) 11070−11009.

23. A. Bora, T. Suzuki and S. Funar-Timofei, Neonicotinoid insecticide design: molecular docking, multiple chemometric approaches and toxicity relationship with Cowpea aphids, *Environ. Sci. Pollut. Res*. (2019), DOI: 10.1007/s11356-019-04662-9.

24. S. Lu, X. Shao, Z. Li, Z. Xu, S. Zhao, Y. Wu and X. Xu, Design, synthesis, and particular biological behaviors of chain-opening nitromethylene neonicotinoids with cis configuration, *J. Agric. Food Chem.* **60** (2012) 322−330.

25. Z. Ye, L. Shi, X. Shao, X. Xu, Z. Xu and Z. Li, Pyrrole- and dihydropyrrole-fused neonicotinoids: Design, synthesis, and insecticidal evaluation, *J. Agric. Food Chem.* **61** (2013) 312−319.

26. P. C. D. Hawkins and A. Nicholls, Conformer generation with OMEGA: Learning from the data set and the analysis of failures, *J. Chem. Inf. Model.* **52** (2012) 2919−2936.

27. P. C. D. Hawkins, G. A. Skillman, G. L. Warren, B. A. Ellingson and M. T. Stahl, Conformer generation with OMEGA: Algorithm and validation using high quality structures from the protein databank and Cambridge structural database, *J. Chem. Inf. Model.* **50** (2010) 572–584.

28. S. Wold and W. J. Dunn III, Multivariate quantitative structure-activity relationships (QSAR): conditions for their applicability, *J. Chem. Inf. Comput. Sci.* **23** (1983) 6−13.

29. P. Gramatica, N. Chirico, E. Papa, S. Cassani and S. Kovarich, QSARINS: A new software for the development, analysis, and validation of QSAR MLR models, *J. Comput. Chem.* **34** (2013) 2121 – 2132.

30. P. Gramatica, On the development and validation of QSAR models, *computational toxicology. Methods in molecular biology* (*Methods and Protocols*), Vol. 930, Humana Press, Totowa, 2013, pp. 499‒526.

31. U. Depczynski, V. J. Frost and K. Molt, Genetic algorithms applied to the selection of factors in principal component regression, *Anal. Chim. Acta* **420** (2000) 217‒227.

32. R. Todeschini, V. Consonni, A. Mauri and M. Pavan, Detecting "bad" regression models: multicriteria fitness functions in regression analysis. *Anal. Chim. Acta* **515** (2004) 199‒208.

33. D. M. Hawkins, S. C. Basak and D. Mills, Assessing model fit by cross-validation, *J. Chem. Inf. Comput. Sci.* **43** (2003) 579‒586.

34. S. L. Dixon, A. M. Smondyrev, E. H. Knoll, S. N. Rao, D. E. Shaw and R. A. Friesner, PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results, *J. Comput. Aided. Mol. Des.* **20** (2006) 647–671.

35. Small-Molecule Drug Discovery Suite 2014–1: Phase (2014) version 3.8, Schrodinger, LLC, New York, NY, 2014 Phase, version 3.1 Schrödinger, LLC, New York. http://www.schrödinger.com.

36. R. Todeschini, V. Consonni and A. Maiocchi, The K correlation index: theory development and its applications in chemometrics, *Chemometr. Intell. Lab.* **46** (1999) 13‒29.

37. M. Goodarzi, S. Deshpande, V. Murugesan, S. B. Katti and Y. S. Prabhakar, Is feature selection essential for ANN modeling, *QSAR combinatorial science* **28** (2009) 1487‒1499.

38. N. Chirico and P. Gramatica, 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** (2011) 2320–2335.

39. L. M. Shi, H. Fang, W. Tong, J. Wu, R. Perkins, R. M. Blair, W. S. Branham, S. L. Dial, C. L. Moland and D. M. Sheehan, QSAR models using a large diverse set of estrogens, *J. Chem. Inf. Model.* **41** (2001) 186–195.

40. G. Schüürmann, R. U. Ebert, J. Chen, B. Wang and R. Kuhne, External validation and prediction employing the predictive squared correlation coefficient test set activity mean vs training set activity mean. *J. Chem. Inf. Model.* **48** (2008) 2140–2145.

41. V. Consonni, D. Ballabio and R. Todeschini, Comments on the definition of the Q2 parameter for QSAR validation. *J. Chem. Inf. Model.* **49** (2009) 1669−1678.

42. D. L. J. Alexander, A. Tropsha and D. A. Winkler, Beware of R2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *J. Chem. Inf. Model.* **55** (2015) 1316−1322.

43. P. Pratim Roy, S. Paul, I. Mitra and K. Roy, On two novel parameters for validation of predictive QSAR models, *Molecules* **14** (2009) 1660−1701.

44. S. Chatterjee and B. Price, Regression Analysis by Example, 2nd ed., *John Wiley & Sons*, New York, 1991.

45. I. E. Frank and R. Todeschini, *The Data Analysis Handbook*, *Elsevier*, Amsterdam, 1994.

46. M. V. Diudea, D. Horvath, A. Graovac, Molecular topology. 15. 3D distance matrices and related topological indices. *J. Chem. Inf. Comput. Sci.* **35** (1995) 129–135.

47. J. Gálvez, R. Garcìa, M. T. Salabert and R. Soler, Charge indexes: new topological descriptors, *J. Chem. Inf. Comput. Sci.* **34** (1994) 520–525.

48. L. B. Kier and L. H. Hall, *Molecular Connectivity in Structure–Activity Analysis, Research Studies*, *John Wiley & Sons*, Chichester, UK 1986, p. 262.

49. R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, *Alphabetical Listing*, Vol. I, Wiley-VCH, Weinheim, 2009, pp. 191.

50. J. Schuur, P. Selzer and J. Gasteiger, The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure–spectra correlations and studies of biological activity, *J. Chem. Inf. Comput. Sci.* **36** (1996) 334–344.