OPEN ACCESS

Conference Proceedings Paper – 21st International Electronic Conference on Synthetic Organic Chemistry (November 2017) ECSOC-21

QSAR Study of Neonicotinoid Insecticidal Activity Against Cowpea Aphids

Simona Funar-Timofei * and Alina Bora

Institute of Chemistry Timisoara of the Romanian Academy, Bul. Mihai Viteazu 24, 300223 Timisoara, Romania; E-Mails: timofei@acad-icht.tm.edu.ro (S.F.T.); alina.bora@gmail.com (A.B.)

* Author to whom correspondence should be addressed; E-Mail: timofei@acad-icht.tm.edu.ro (S.F.T.)
Tel.: +40-256-491-818; Fax: +40-256-491-824.

Abstract: A series of 30 neonicotinoid insecticides, bearing nitroconjugated double bond and five-membered heterocycles and nitromethylene compounds containing a tetrahydropyridine ring with exo-ring ether modifications, active against the cowpea aphids (*Aphis craccivora*), was analyzed using multiple linear regression (MLR) method. The semiempirical quantum chemical PM7 approach was employed for structure optimization. Structural descriptors were calculated for the minimum energy conformers and were related to the insecticidal activity (expressed as pLC_{50} values) through genetic algorithm, using the multiple linear regression (MLR) approach. Several parameters were applied for internal and external model validation. The final MLR models demonstrated good statistical results and predictive power. Fewer number of 6-membered rings, a reduced number of rings containing secondary C(sp3) atoms, and/or lower values of strongest basic pK_a in the core structure of neonicotinoids are considered to increase the insecticide activity.

Keywords: insecticide; MLR; PM7; cowpea aphids

1. Introduction

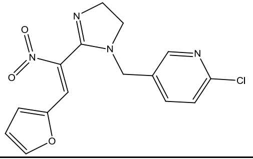
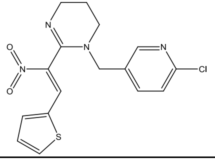
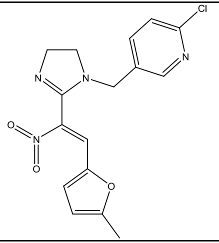
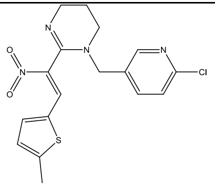
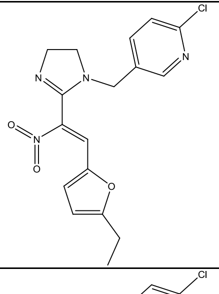
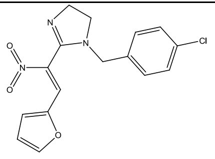
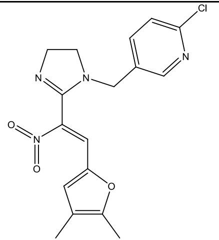
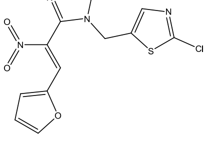
Neonicotinoids are considered to be one of the most important and relevant classes of insecticides used nowadays [1, 2].

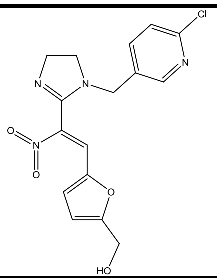
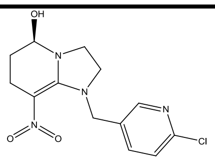
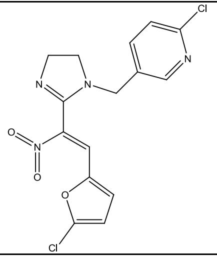
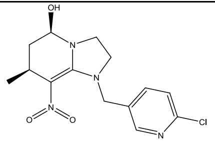
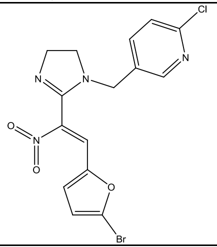
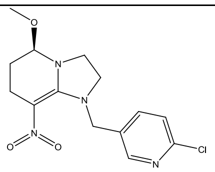
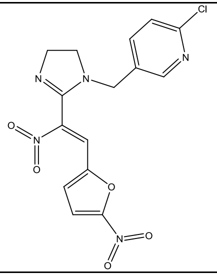
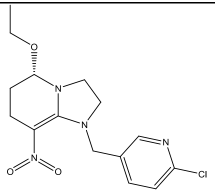
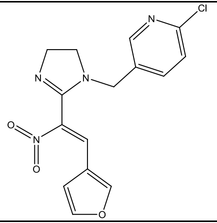
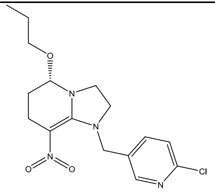
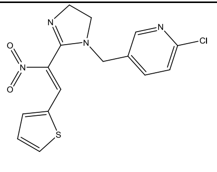
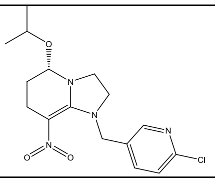
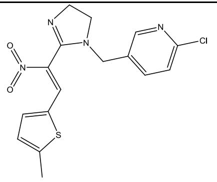
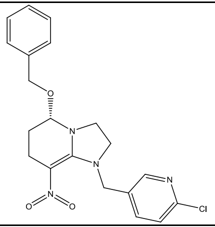
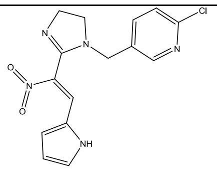
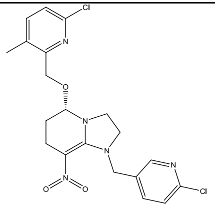
They are active on the insect postsynaptic nicotinic acetylcholine receptors (nAChRs) and still of current interest, despite their resistance and bee toxicity [3]. Several studies of computational chemistry and electrophysiology tried to model the neonicotinoid-receptor interactions. Electrostatic interactions and possibly hydrogen bond formation were found to be important for the insecticidal activity [4].

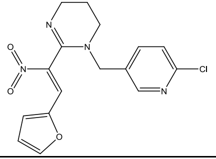
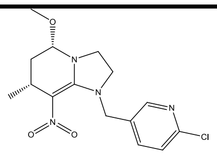
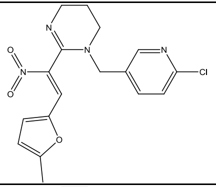
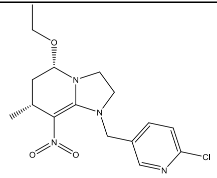
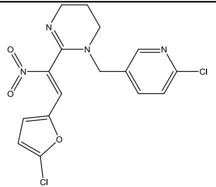
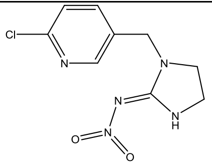
A number of 26 N3-substituted imidacloprid insecticides, active against the housefly *Musca Domestica*, were previously studied using multiple linear regression (MLR) [5]. Good correlation of compound structural features with the insecticide activity was noticed, but models with modest predictive power. It was found that high values of squared octanol-water partition coefficients and of tautomers were favorable for the insecticidal activity.

In the present study, a series of 30 neonicotinoid analogues tested against the cowpea aphids (*Aphis craccivora*) was modeled by molecular and quantum mechanics approaches. The structural descriptors derived from the minimum energy structures were correlated to the insecticidal activity using the multiple linear regression approach. Predictive models, useful to predict new insecticides, with improved activity were developed.

Table 1. The neonicotinoid structures,-their experimental insecticidal (pLC_{50}) and predicted ($pLC_{50\ pred}$) activity values for the best MLR1 model

No	Structure	pLC_{50exp}	pLC_{50pred}	No	Structure	pLC_{50exp}	pLC_{50pred}
1		5.43	4.92	**16**		4.69	4.31
2		5.20	5.36	**17**		4.61	4.49
3		5.74	5.35	**18****		3.63	
4		5.33	5.42	**19**		5.46	5.89

5		4.98	5.35	**20**		3.73	3.89
6		5.12	5.22	**21**		4.01	3.89
7*		5.14	4.83	**22***		3.88	3.91
8		4.96	4.87	**23**		4.02	3.95
9		5.35	4.91	**24**		3.98	3.98
10*		5.37	4.96	**25**		3.59	3.97
11		5.51	5.35	**26**		3.24	3.12
12*		4.95	5.01	**27**		2.94	3.08

13*		4.12	4.03	**28**		3.83	3.89
14**		3.16		**29**		3.73	3.96
15		4.22	4.46	**30****		4.46	

* Test compounds included in the final MLR1 data set

** Compounds excluded from the final MLR1 model

2. Methods

2.1. Definition of target property and molecular structures

A set of 30 neonicotinoid analogues bearing nitroconjugated double bond and five-membered heterocycles and nitromethylene neonicotinoids containing a tetrahydropyridine ring with exo-ring ether modifications with known insecticidal activity was analyzed [6, 7]. The insecticidal activity against cowpea aphids (*Aphis craccivora*) activity data, expressed as pLC_{50} values (where LC_{50} represents the median lethal concentration of the chemical in air that kills 50% of the test animals during the observation period) was used as dependent variable.

In the first step, the structures of the investigated molecules were pre-optimized using the conformer plugin (with MMFF94 as molecular mechanics force field) of the MarvinSketch (MarvinSketch 15.2.16.0, ChemAxon Ltd. http://chemaxon.com) package. In the next step, the lowest energy conformers were refined using the semiempirical PM7 Hamiltonian [8] implemented in MOPAC 2016 program (MOPAC2016, James J. P. Stewart, Stewart Computational Chemistry, Colorado Springs, CO, USA, HTTP://OpenMOPAC.net (2016)). For the geometry optimization a gradient norm limit of 0.01kcal/Å was set. Structural 0D, 1D, 2D and 3D molecular descriptors were calculated for the lowest energy structures using the DRAGON (Dragon Professional 5.5, 2007, Talete S.R.L., Milano, Italy) and InstanJChem (Instant JChem (2012) version 5.10.0, Chemaxon, http://www.chemaxon.com) software.

2.2. The Multiple Linear Regression (MLR) method

Because the number of computed descriptors is too high (1624 descriptors) compared to the number of compounds (N = 30), a proper variable selection method was compulsory. The Genetic Algorithm (GA) is a trustworthy and extensively used variable selection method [9]. The QSARINS v. 2.1 program [10] uses GAs to choose the meaningful descriptors that influence the variation of biologic activity of the compounds. The following parameters were employed: the RQK fitness function with leave-one-out cross-validation correlation coefficient as 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. Model validity

The neonicotinoid derivatives were divided into training and test sets by random split, taking out 18.5% of the total number of compounds (no. 7, 10, 12, 13, 22), while the remaining 81.5% were used as training set. The model's predictability was tested using the Q^2_{F1} [11]; Q^2_{F2} [12]; Q^2_{F3} [13] and the concordance correlation coefficient (CCC) [14] (having the thresholds values higher than 0.85, as they have been rigorously determined by a simulation study [15])-external validation parameters.

Moreover, the predictive power of the QSAR models was evaluated based on the predictive parameter r^2_m (with a lowest threshold value of 0.5 to be accepted) [16].

The model robustness (overfit) was tested using the Y-randomization test. The dependent variable is arbitrarily mixed and a model is built using the same X matrix of molecular descriptors. The obtained MLR models (after 2000 randomizations) must have minimal r^2 (correlation coefficient) and q^2 (cross-validation coefficient) values [17].

The data over fitting and model applicability was checked by comparing the root-mean-square errors (RMSE) and the mean absolute error (MAE) of the training and validation sets [18].

For internal validation results several measures of robustness were employed: Y-scrambling [19], adjusted correlation coefficient (r^2_{adj}) and q^2 (leave-one-out, q^2_{LOO}, and leave-more-out, q^2_{LMO}) cross-validation coefficient.

The Multi-Criteria Decision Making (MCDM) validation criteria [20], having values between 0 (the worst) and 1 (the best), is used to summarize the performance of MLR models. To every validation criteria a desirability function is associated, and MCDM values are calculated from the geometric average of all the desirability function values. In this study, the best MLR models were chosen from the ‚MCDM all' scores, based on the fitting, cross validated and external criteria.

3. Results and Discussion

The data was normalized using the autoscaling method:

$$XT_{mj} = \frac{X_{mj} - \overline{X}_m}{S_m} \qquad (1)$$

where for each variable m, XT_{mj} and X_{mj} are the j values for the m variable after and before scaling, respectively, $\overline{X}_m$ is the mean, and S_m is the standard deviation of the variable.
Three compounds (**14**, **18** and **30**) were found to be outliers, having the standardized residual values greater than 2.5 standard deviation units and were not included in the final MLR models.

Variable selection using the genetic algorithm was employed to build several MLR models. The statistical results for model fitting and predictivity are included in Tables 2-4.

Table 2. Fitting and cross-validation statistical results of the MLR models (training set)*

Model	$r^2_{training}$	q^2_{LOO}	q^2_{LMO}	r^2_{adj}	$RMSE_{tr}$	MAE_{tr}	CCC_{tr}	r^2_{scr}	q^2_{scr}	SEE	F
MLR1	0.896	0.853	0.845	0.885	0.261	0.216	0.945	0.095	-0.220	0.281	81.61
MLR2	0.887	0.851	0.841	0.876	0.271	0.220	0.940	0.095	-0.228	0.292	74.90
MLR3	0.808	0.770	0.763	0.799	0.354	0.302	0.894	0.045	-0.157	0.372	84.35
MLR4	0.824	0.786	0.779	0.815	0.340	0.294	0.904	0.049	-0.152	0.356	93.58

* $r^2_{training}$ -correlation coefficient; q^2_{LOO} - leave-one-out correlation coefficient; q^2_{LMO} leave-more-out correlation coefficient; r^2_{adj} -adjusted correlation coefficient; RMSEtr-root-mean-square errors; MAEtr-mean absolute error; CCCtr-the concordance correlation coefficient; r^2_{scr} and q^2_{scr} -Y-scrambling parameters; SEE-standard error of estimates; F-Fischer test.

Table 3. MLR predictivity results (test set)*

Model	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	$RMSE_{ext}$	MAE_{ext}	CCC_{ext}
MLR 1	0.851	0.840	0.916	0.235	0.179	0.907
MLR 2	0.805	0.790	0.890	0.269	0.244	0.913
MLR 3	0.876	0.867	0.930	0.214	0.207	0.934
MLR 4	0.820	0.806	0.898	0.258	0.236	0.921

* Q^2_{F1}; Q^2_{F2}; Q^2_{F3} -external validation parameters;
$RMSE_{ext}$-root-mean-square errors; MAE_{ext} -mean absolute error;
CCC_{ext}-the concordance correlation coefficient

Table 4. The r^2_m predictivity parameters, 'MCDM all' score values and descriptors in the final MLR models*

Model	r^2_m	MCDM all	Descriptors included in the model*
MLR1	0.810	0.878	nR06, E3m
MLR2	0.697	0.865	nCrs, C-003
MLR3	0.817	0.846	Strongest basic pKa
MLR4	0.656	0.840	nCrs

* nR06 – number of 6-membered rings, E3m- 3rd component accessibility directional WHIM index/weighted by atomic masses, nCrs- number of ring secondary C(sp3), C-003 - CHR3 (atom-centred fragments), strongest basic pKa- the basic pK_a value for the first strength index.

In order to verify the reliability of the developed equations, experimental versus predicted pLC_{50} values, Williams plots and Y-scramble plots for the MLR1 best model are presented in Figures 1, 2 and 3, respectively.

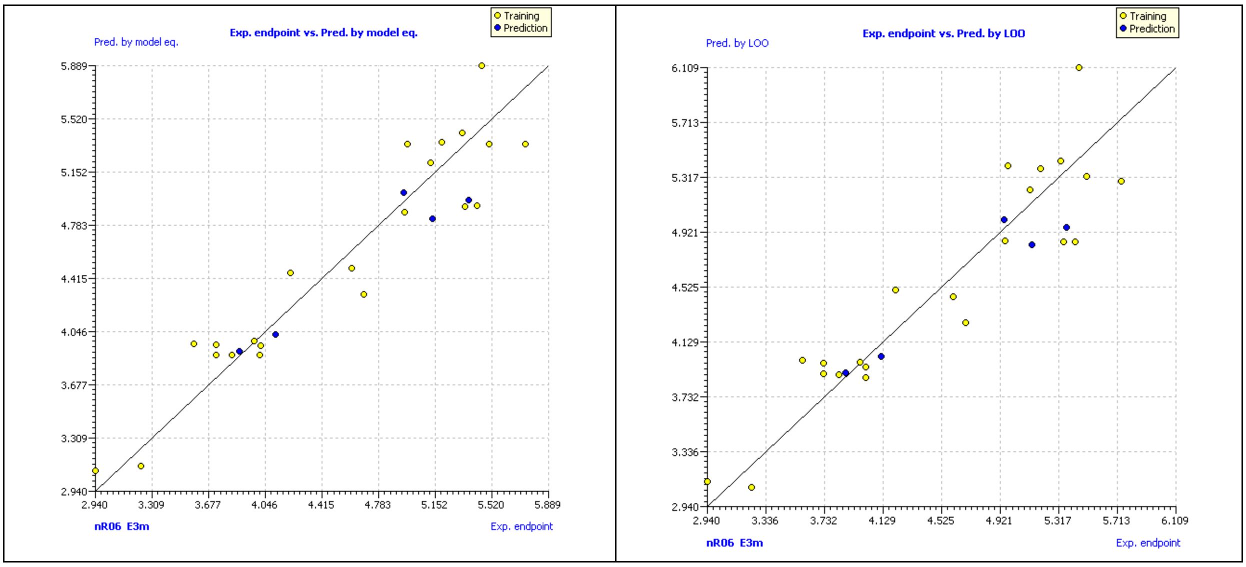


Figure. 1. Plots of experimental versus predicted pLC_{50} values for the MLR1 model predicted by the model (left) and by the leave-one-out (right) cross-validation approach (yellow circles-training compounds, blue circles-test compounds).

The Williams plot is used to identify compounds with the greatest structural influence ($h_i > h^*$; h_i =leverage of a given chemical; h^*= the warning leverage) in the MLR model. The applicability domain of the MLR models was considered in the range of ±2.5σ (the MLR1 leverage threshold h* = 0.409). All compounds in the dataset are within the applicability domain of the MLR1 model, as presented in Figure 2.

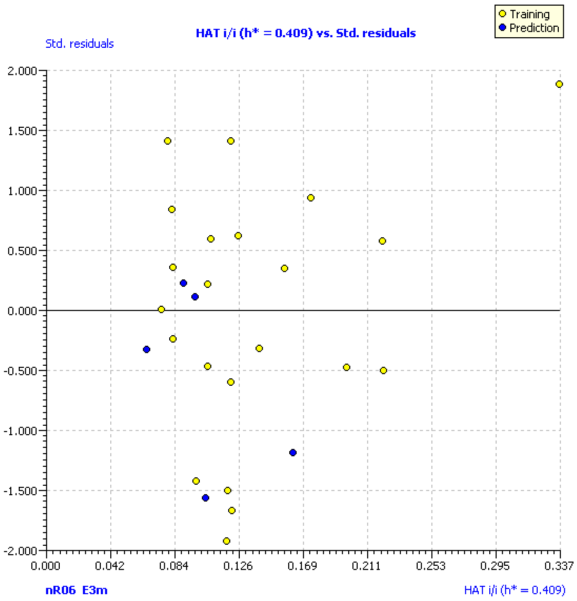


Figure. 2. Williams plot predicted by the MLR1 model (yellow circles-training compounds, blue circles-test compounds).

In the y-scrambling test performed for the MLR models, a significant low scrambled r^2 (r^2_{scr}) and cross-validated q^2 (q^2_{scr}) values were obtained for 2000 trials. Figure 3 suggest that in case of all the randomized models, the values of r^2_{scr} and q^2_{scr} for the MLR1 model were < 0.5 (r^2_{scr}/q^2_{scr} of 0.095/-0.220). The low calculated r^2_{scr} and q^2_{scr} values indicate no chance correlation for all MLR chosen models (Table 2).

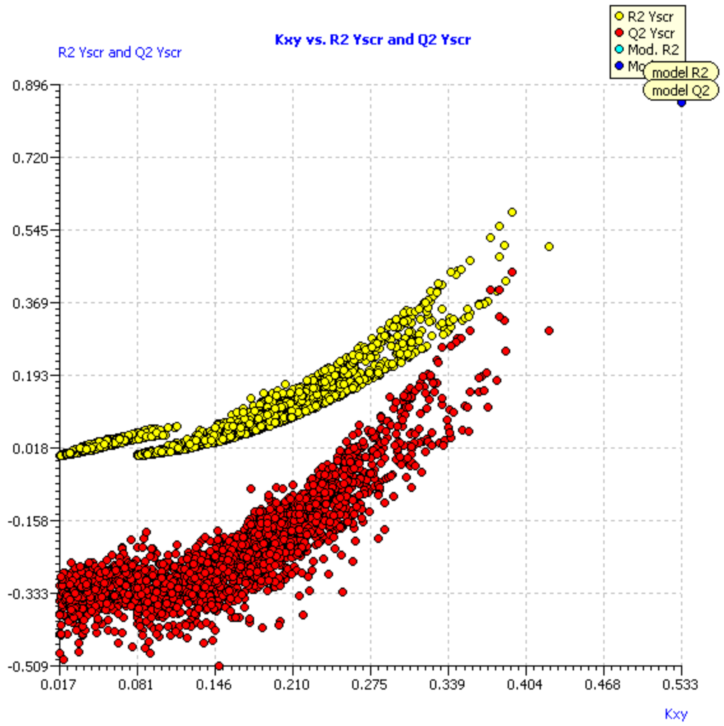


Figure. 3. Y-scramble plots for the MLR1 model.

A correlation matrix of the selected molecular descriptors from the MLR1 model is presented in Table 5. The two selected descriptors are not intercorrelated.

Table 5. Correlation matrix of the selected descriptors included in the best MLR1 model

	nR06	E3m
nR06	1	
E3m	0.247	1

The best MLR1 model has two descriptors: the constitutional nR06 descriptor, which represents the number of 6-membered rings, and the WHIM E3m descriptor (3rd component accessibility directional WHIM index / weighted by atomic masses). Increase of E3m is beneficial for the insecticidal activity. The presence of less 6-membered rings in the structure favors the insecticide action. Interestingly is that MLR models including only one descriptor, e.g. the number of ring secondary C(sp3) or strongest basic pKa descriptors (for both parameters lower values raise the insecticidal activity), gave good statistical results and models with predictive power.

The MLR models presented in this study can be used for prediction of new neonicotinoid structures, active as insecticides for the cowpea aphids.

4. Conclusions

In the current study, quantitative relationships between the molecular structure and cowpea aphids (*Aphis craccivora*) inhibitory activity of neonicotinoids analogues were verified by the MLR approach. The semiempirical quantum chemical PM7 method was employed for structure optimization. The genetic algorithm was used for variable selection. The final MLR models have good statistical parameters and predictive power. Structural features, such as the number of 6-membered rings, basic pKa capacity and the number of ring secondary C(sp3) are particulary significant in the design of novel neonicotinoids with scaffold containing nitroconjugated double bond and five-membered heterocycles and nitromethylene compounds containing a tetrahydropyridine ring with exo-ring ether modifications.

Acknowledgments

This project was financially supported by Project 1.1 of the Institute of Chemistry of the Romanian Academy. Access to the Chemaxon Ltd., QSARINS and MOPAC 2016 software are greatly acknowledged by the authors.

Author Contributions

S.F.T. and A.B. analyzed the data; A.B. contributed to molecular modeling calculations; S.F.T. performed the statistical analysis and wrote the paper.

Conflicts of Interest

The authors declare no conflict of interest.

References and Notes

1. Ren, L.; Lou, Y.; Chen, N.; Xia, S.; Shao, X.; Xu, X.; Li, Z. Synthesis And Insecticidal Activities Of Tetrahydroimidazo[1,2-A]Pyridinones: Further Exploration On Cis-Neonicotinoids. *Synthetic Commun.* **2014**, *44*, 858–867.
2. Nauen, R.; Denholm, I. Resistance of Insect Pests to Neonicotinoid Insecticides: Current Status and Future Prospects, *Arch. Insect Biochem.* **2005**, *58*, 200–215.
3. Matsuda, K.; Kanaoka, S.; Akamatsu, M.; Sattelle, D. B. Diverse actions and target-site selectivity of neonicotinoids: Structural insights, *Mol. Pharmacol.* **2009**, *76*, 1–10.
4. Matsuda, K.; Shimomura, M.; Ihara, M.; Akamatsu. M.; Sattelle, D.B. Neonicotinoids show selective and diverse actions on their nicotinic receptor targets: electrophysiology, molecular biology, and receptor modeling studies. *Biosci. Biotechnol. Biochem.*, **2005**, *69*, 1442-1452.
5. Suzuki, T.; Avram, S.; Borota, A.; Funar-Timofei, S. QSAR modeling of N3-substituted imidacloprid insecticides used against the housefly *Musca Domestica. Journal of Toyo University, Natural Science* **2014**, *58*, 83-95, http://jairo.nii.ac.jp/0236/00004962/en.

6. Tian, Z.; Shao, X.; Li, Z.; Qian, X.; Huang, Q. Synthesis, Insecticidal Activity, and QSAR of Novel Nitromethylene Neonicotinoids with Tetrahydropyridine Fixed cis Configuration and Exo-Ring Ether Modification. *J. Agric. Food Chem*. **2007**, *55*, 2288-2292.

7. Shao, X.; Li, Z.; Qian, X.; Xu, X. Design, Synthesis, and Insecticidal Activities of Novel Analogues of Neonicotinoids: Replacement of Nitromethylene with Nitroconjugated System. *J. Agric. Food Chem.* **2009**, *57*, 951–957.

8. Stewart, J.J.P. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Model*. **2013**, *19*, 1–32.

9. Depczynski, U.; Frost V.J.; Molt, K. Genetic algorithms applied to the selection of factors in principal component regression. *Anal. Chim. Acta* **2000**, *420*, 217-227.

10. 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.* **2013**, *34*, 2121–2132.

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

12. Schüürmann, G.; Ebert, R.U.; 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*, 2140–2145.

13. Consonni, V.; Ballabio, D.; Todeschini, R. Comments on the definition of the Q2 parameter for QSAR validation. *J. Chem. Inf. Model*. **2009**, *49*, 1669–1678.

14. 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*, 2320-2335.

15. 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*, 2044−2058.

16. Roy, K.; Mitra, I. On the Use of the Metric r_m^2 as an Effective Tool for Validation of QSAR Models in Computational Drug Design and Predictive Toxicology. *Mini-Rev. Med. Chem*. **2012**, *12*, 491−504.

17. Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Multi and megavariate data analysis: principles and applications*. Umetrics AB, Umea, 2001, pp. 92–97, pp. 489–491.

18. Goodarzi, M.; Deshpande, S.; Murugesan, V.; Katti, S.B.; Prabhakar, Y.S. Is Feature Selection Essential for ANN Modeling? *QSAR Comb. Sci*. **2009**, *28*, 1487–1499.

19. Todeschini, R.; Consonni, V.; Maiocchi, A. The K correlation index: Theory development and its application in chemometrics. *Chemometr. Intell. Lab*. **1999**, *46*, 13-29.

20. Keller, H.R.; Massart, D.L.; Brans, J.P. Multicriteria decision making: a case study. *Chemom. Intell. Lab. Syst*. **1991**, *11*, 175-189.