

NONEXPERIMENTAL SCREENING OF THE WATER SOLUBILITY, LIPOPHILICITY, BIOAVAILABILITY, MUTAGENICITY AND TOXICITY OF VARIOUS PESTICIDES WITH QSAR MODELS AID

O.G. Kolumbin^a, L.N. Ognichenko^b, A.G. Artemenko^b, P.G. Polischuk^b,
M.A. Kulinsky^b, E.N. Muratov^b, V.E. Kuz'min^b, V.A. Bobeica^c

^aTiraspol State University of T.G.Shevchenko, Moldova, Tiraspol, 3300, 25 October, 128,

^bDepartment of Molecular Structure and Chemometrics, A.V. Bogatsky Physical-Chemical Institute National Academy of Science of Ukraine, Ukraine, Odessa, 65080, LustdorfskayaDoroga 86,

^cMoldova State University, Department of Industrial and Ecological Chemistry, Republic of Moldova, Kishinev, 2009, 60 Mateevich Str. *Corresponding author: valentinbobeica@rambler.ru

Abstract: In our study the dataset containing 489 pesticides and their active substances of different classes of organic compounds was used for analysis. For compounds of analyzed dataset the values of lipophilicity, water solubility, toxicity, bioavailability and mutagenicity were predicted by developed QSAR models. The most environmentally hazardous substances were identified using prediction of QSAR models for pesticides. The satisfactory coincidence between the experimental values of investigated properties and their predicted values by QSAR models was obtained (coefficient of determination in the range 83-94%).

Keywords: pesticides, toxicological analysis, QSAR, lipophilicity, RF method

Introduction

The prognosis of dangerous xenobiotics and elaboration of new methods of their toxic evaluation is getting more actual when the growing amount of new pesticides outnumbers the possibilities of their experimental toxic evaluation. Therefore, different mathematical models based on the relationship between biological activity and physical-chemical properties of compounds are used for predicting toxic parameters of new pesticides, their teratogenesis and mutagenesis and for installing the calculation methods of the corresponding normatives. This approach of nonexperimental screening can considerably reduce the number of experimental researches of new pesticides and use of time and materials in these investigations. The goal of this study is to determine mathematical QSAR effective application for prognosis of ecological danger caused by pesticides on mammals using such properties as the lipophilicity, water solubility, bioavailability, mutagenicity and toxicity calculations.

Materials and Methodes

In this study a dataset¹ containing information about solubility, lipophilicity and toxicity for 489 pesticides and their active substances of different classes of organic compounds was used for research. Similar researches and analysis had been done using HiT QSAR software [1, 2]. Experimental values of lypophilicity are known for 334 molecules and are varied in the range of -4 till 8,39. Values for water solubility are known for 371 molecules of the analyzed database and are varied from -7,8 till 1,57.

All the 489 compounds were analyzed using previously developed models for lypophilicity [3], solubility [4], bioavailability [5], mutagenicity [6] and toxicity on *Tetrahymenapyritormis*[7]. All these 5 models were built using simplex representation of molecular structure (SiRMS) and the statistic method of random forest (RF) [8]. CAS numbers of investigated molecules are presented in Table 1.

In this study Simplex Representation of Molecular Structure (SiRMS) QSPR approach [1] was used for calculation of structural descriptors of molecules. Main concept of this approach is that any molecule can be represented as a system of different simplexes (tetraatomic fragments with fixed composition and topological structure). At the 2D level, the connectivity of atoms in simplex, atom type and bond nature (single, double, triple, aromatic) are taking into consideration. In the present study, only bounded 2D simplex descriptors were used for molecular structure representation. Not only atom type but also other physical-chemical characteristics of an atom such as: partial charge, lipophilicity, refraction and the ability for an atom to be a donor or acceptor in hydrogen bond formation were used for atom differentiation. The usage of sundry variants of differentiation of simplex vertexes (atoms) represents the principal feature of the proposed approach. LogP, molecular refraction, electronegativity of the molecule and its mass were used as integral descriptors in addition to calculated 2D simplex descriptors. Although the developed in our laboratory SiRMS method is novel, it has been employed successfully in several studies to differentiate "structure-activity" relationships [9-11]. The main advantages of SiRMS are the possibility of analysis of molecules with noticeable structural differences as well as the possibility to reveal individual molecular fragments (simplex combinations) promoting or interfering with investigated property.

¹ <http://chem.sis.nlm.nih.gov/chemidplus/chemidlite.jsp>

Such approach avoids additive contributions of structural fragments, because the contributions of atom/or structural fragment depend on their surrounding. SiRMS methodology does not have many of the restrictions of such well-known and widely used approaches as CoMFA, CoMSIA, HASL, in which the application is limited to a structurally homogeneous set of molecules only. SiRMS approach is similar to HQSPR approach but has not its limitations (consideration of atom type only) and deficiencies (an ambiguity of descriptor formation during the hashing of molecular holograms).

Results and Discussion

As early mentioned the experimental lipophilicity values were found only for 334 compounds (68% from all molecules). Using the lipophilicity RF model [3] LogP values were predicted for all 489 molecules. RF model for calculation of lipophilicity was built on the base of more than 10,000 molecules. Simplex descriptors were calculated for studied 489 molecules using such tuning parameters which were used for the construction of lipophilicity RF model. For molecules of analyzed dataset with known LogP values the model predictive ability was assessed. The coefficient of determination (R^2) between the observed experimental values of Log P and predicted Log P for RF model equals 0.94 and the standard error (SE) equals 0.46.

For water solubility (LogS_w) a model was built on dataset consisting of more than 1200 compounds [12]. Statistical characteristics of the RF model for water solubility are quite adequate: R^2 for the training set is 0.99 and for *out-of-bag* set equals 0.91. For 489 molecules simplex descriptors were calculated considering the same tuning parameters which were used to build the RF model for solubility. It should be noted that only 312 of the molecules (64% from the dataset) with known experimental values LogS_w get in the model domain applicability. The other 59 molecules (12%) are out the range of applicability. Using the molecules that are in the domain applicability of the model with known LogS_w values the predictive ability of the model was estimated. The coefficient of determination (R^2) between the observed experimental LogS_w values and the predicted LogS_w by RF model is 0.83 and the standard error (SE) = 0.85. Among the remaining 118 molecules for which experimental LogS_w values aren't known only 90 molecules are in the domain applicability of the model and 28 outside of the domain applicability.

Table 1

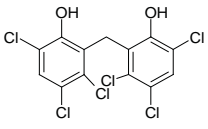
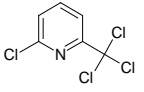
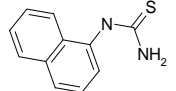
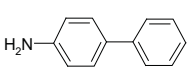
CAS – numbers of analysed compounds

100-02-7	108-34-9	116-29-0	126-22-7	1563-66-2	2310-17-0	327-19-5	520-45-6
100-52-7	108-35-0	116-52-9	126-61-4	1582-09-8	2312-35-8	327-98-0	524-40-3
101-05-3	1085-98-9	117-18-0	126-75-0	1596-84-5	2314-09-2	330-54-1	524-42-5
101-20-2	108-60-1	117-26-0	127-21-9	1610-17-9	2425-06-1	330-55-2	52-46-0
101-21-3	108-62-3	117-52-2	127-63-9	1610-18-0	2425-10-7	330-64-3	52-51-7
101-27-9	108-80-5	117-80-6	127-90-2	1646-87-3	2431-96-1	333-41-5	52-60-8
101-42-8	108-91-8	117-81-7	130-15-4	1646-88-4	2439-01-2	3337-71-1	52-68-6
1014-69-3	108-94-1	118-52-5	130-86-9	1689-83-4	2463-84-5	3347-22-6	52-85-7
1014-70-6	108-95-2	1186-09-0	131-11-3	1689-84-5	2497-07-6	3383-96-8	53-19-0
101-99-5	109-84-2	118-74-1	131-89-5	1689-99-2	2511-10-6	350-46-9	532-54-7
102-60-3	109-94-4	118-75-2	131-91-9	1696-17-9	2540-82-1	3566-00-5	533-74-4
102-71-6	109-97-7	118-96-7	132-66-1	1698-60-8	2587-90-8	3615-21-2	534-52-1
1031-47-6	110-12-3	119-12-0	133-06-2	1702-17-6	2593-15-9	366-18-7	535-89-7
103-17-3	110-44-1	119-27-7	133-07-3	1836-75-5	2595-54-2	3689-24-5	54-11-5
103-18-4	110-85-0	119-38-0	133-32-4	1836-77-7	2597-03-7	371-62-0	542-75-6
103-33-3	1113-02-6	1194-65-6	133-90-4	1861-32-1	262-20-4	371-86-8	545-06-2
1034-01-1	1113-14-0	119-89-1	133-91-5	1861-40-1	2631-37-0	3740-92-9	54-62-6
104-01-8	111-44-4	120-23-0	136-25-4	1897-45-6	2631-40-5	420-04-2	551-06-4
104-04-1	1114-71-2	120-32-1	136-45-8	1912-24-9	2636-26-2	470-90-6	55-38-9
105-13-5	112-12-9	120-36-5	137-18-8	1912-26-1	2642-71-9	477-27-0	555-37-3
105-28-2	1121-30-8	120-51-4	137-26-8	1918-00-9	2655-14-3	485-31-4	555-77-1
105-67-9	112-24-3	120-57-0	139-40-2	1918-02-1	2686-99-9	494-52-0	556-61-6
105-99-7	112-27-6	120-60-5	1397-94-0	1918-13-4	2797-51-5	495-73-8	558-25-8
106-24-1	112-30-1	120-62-7	139-94-6	1918-16-7	288-32-4	499-75-2	563-12-2
106-46-7	112-38-9	120-80-9	140-57-8	1929-77-7	2921-88-2	50-00-0	584-79-2
106-50-3	112-42-5	120-83-2	141-03-7	1929-82-4	297-97-2	500-28-7	60-51-5
106-51-4	112-56-1	120-93-4	141-27-5	1943-79-9	298-00-0	50-14-6	60-57-1
106-89-8	1129-41-5	121-21-1	141-43-5	1982-47-4	298-01-1	50-18-0	60-80-0

106-93-4	1134-23-2	121-29-9	141-66-2	1982-49-6	298-02-2	502-55-6	615-15-6
106-95-6	113-48-4	1214-39-7	141-78-6	2008-41-5	298-03-3	50-29-3	61-82-5
106-96-7	114-26-1	121-75-5	141-84-4	2032-59-9	298-04-4	50-31-7	62-56-6
107-02-8	1146-99-2	122-10-1	1420-06-0	2032-65-7	299-85-4	504-24-5	62-73-7
107-04-0	114-83-0	122-14-5	1420-07-1	2104-64-5	299-86-5	50-65-7	63-25-2
107-06-2	1149-23-1	122-15-6	142-46-1	2104-96-3	300-76-5	506-77-4	63-74-1
107-07-3	115-26-4	122-17-8	1444-64-0	2164-09-2	301-11-1	50-71-5	64-00-6
107-13-1	115-29-7	122-34-9	145-73-3	2164-17-2	301-12-2	510-15-6	640-15-3
1071-83-6	115-31-1	122-39-4	148-01-6	2212-67-1	305-85-1	51-03-6	645-05-6
107-18-6	115-32-2	122-42-9	148-24-3	2227-13-6	3060-89-7	51-14-9	65-85-0
107-19-7	115-90-2	122-88-3	148-79-8	2227-17-0	311-45-5	51-17-2	66-02-4
107-21-1	115-91-3	123-09-1	149-30-4	2255-17-6	311-47-7	51-18-3	66-27-3
107-22-2	115-92-4	123-33-1	1498-64-2	2274-67-1	312-73-2	512-56-1	66-76-2
107-31-3	115-93-5	123-54-6	149-91-7	2275-18-5	314-40-9	51-28-5	66-81-9
107-49-3	116-01-8	124-17-4	150-50-5	2303-16-4	314-42-1	513-49-5	67-63-0
1079-33-0	116-06-3	124-28-7	150-68-5	2303-17-5	321-54-0	518-20-7	67-66-3
1081-34-1	116-16-5	126-15-8	152-20-5	2307-68-8	3244-90-4	518-75-2	67-72-1
67-99-2	731-27-1	759-94-4	78-87-5	84-74-2	91-20-3	94-96-2	97-23-4
680-31-9	732-11-6	76-03-9	79-00-5	85-34-7	919-86-8	950-10-7	973-21-7
68-11-1	73-24-5	76-06-2	79-01-6	86-29-3	92-52-4	950-35-6	97-53-0
694-59-7	741-58-2	76-22-2	79-09-4	86-50-0	92-67-1	950-37-8	97-77-8
70-30-4	74-59-9	76-44-8	79-11-8	867-27-6	92-84-2	95-14-7	98-64-6
70-34-8	74-60-2	77-06-5	79-19-6	86-87-3	934-32-7	953-17-3	98-95-3
70-38-2	74-83-9	77-09-8	79-21-0	86-88-4	934-87-2	95-50-1	991-42-4
70-43-9	74-85-1	77-47-4	79-34-5	869-29-4	93-65-2	95-57-8	99-30-9
709-98-8	74-90-8	77-49-6	79-40-3	87-51-4	93-72-1	95-65-8	99-92-3
71-23-8	75-01-4	77-92-9	79-46-9	87-68-3	93-76-5	957-51-7	99-93-4
71-43-2	75-09-2	780-11-0	79-57-2	87-86-5	94-09-7	95-80-7	75-99-0
71-55-6	75-15-0	78-34-2	81-81-2	88-06-2	941-69-5	95-93-2	786-19-6
72-20-8	75-21-8	78-48-8	82-68-8	886-50-0	944-22-9	95-95-4	84-65-1
72-33-3	75-56-9	78-52-4	834-12-8	88-85-7	947-02-4	96-12-8	91-15-6
72-43-5	75-75-2	78-53-5	83-79-4	89-78-1	94-74-6	96-24-2	94-82-6
72-54-8	757-58-4	78-57-9	841-06-5	90-43-7	94-75-7	96-45-7	97-17-6
72-56-0							

Table 2

Predicted and observed values of lipophilicity, water solubility, toxicity and bioavailability for 14 the most environmentally hazardous soluble in water substances analysed with QSAR-models

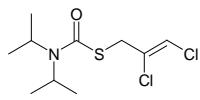
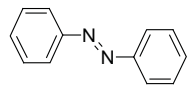
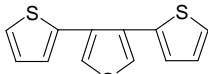
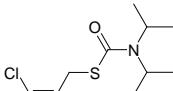
№	CAS number	Structure formula	LogP		Log (IGC50-1)	Bioavailability (Pred.)			LogS _w	
			Pred.	Exp.		Pred.	low	average	high	Pred.
1	70-34-8		1.67		1.44	1	2	2	-2.15	-2.67
2	1929-82-4		3.15	3.41	1.28	1	2	2	-3.63	-3.51
3	86-88-4		1.96	1.65	1.06	1	2	2	-3.15	-2.53
4	92-67-1		2.94	2.86	1.06	1	1	2	-3.66	-

5	118-96-7		1.82	1.60	0.97	1	2	2	-2.43	-3.24
6	330-55-2		3.09	3.20	0.94	1	1	2	-3.31	-3.52
7	1918-13-4		2.12	2.96	0.92	1	1	2	-2.78	-2.34
8	330-54-1		2.78	2.68	0.86	1	1	2	-3.16	-3.74
9	63-25-2		2.49	2.36	0.85	1	1	2	-3.20	-3.26
10	2797-51-5		1.88	2.12	0.43	1	1	1	-2.35	-
11	366-18-7		1.61	1.50	0.41	1	1	2	-1.60	-1.42
12	108-60-1		2.32	2.48	0.37	1	2	2	-1.57	-2.00
13	524-42-5		1.72		0.22	1	1	2	-2.11	-
14	130-15-4		1.77	1.71	0.20	1	1	2	-2.15	-

Table 3

Predicted and observed values of lipophilicity, water solubility, toxicity and bioavailability for 6 the most environmentally hazardous lipophilic substances from analysed dataset with QSAR-models

№	CAS number	Structure formula	LogP		Log (IGC50-1)	Bioavailability (Pred.)			LogS _w	
			Pred.	Exp.	Pred.	low	average	high	Pred.	Exp.
1	72-56-0		5.39		1.63	1	1	2	-6.78	-6.49
2	50-65-7		4.74		1.49	1	2	2	-4.65	-5.31

3	2303-16-4		4.25	4.49	1.02	1	1	2	-4.26	-4.29
4	103-33-3		3.80	3.82	1.34	1	1	2	-4.25	-4.45
5	1081-34-1		3.45	5.57	1.18	1	2	2	-4.37	
6	2303-17-5		4.42	4.60	1.04	1	1	2	-4.72	-4.88

In the result of the analysis of the 489 molecules of the studied dataset according to the developed three models for bioavailability [5], it was found that 116 of them (34%) are not bioavailable molecules, 25 molecules (5%) - with high bioavailability, 124 molecules (25%) - with average and 175 molecules (36%) - with low bioavailability. In these models, if the molecule is assigned 1 - the substance is bioavailable, and 2 – not bioavailable.

For 489 molecules of the dataset using the RF model for calculation of mutagenicity [6] it was predicted that 392 molecules (80% of all the molecules of dataset) are part of the class of non-mutagenic, while the remaining 97 molecules (20%) are dangerous mutagenic substances, which were used to find the most environmentally hazardous substances.

Using the RF model of toxicity on *Tetrahymena pyriformis* organisms [7] toxicity values (Log(IGC50-1)) for all 489 molecules of dataset were predicted. 339 molecules (69%) concern the area of applicability of the model, their predicted values are varied in the range of -1.6 till 2.4. The remaining 150 molecules (31%) are not part of domain of applicability of the model.

In the result of the data analysis the most environmentally hazardous substances which are bioavailable, mutagenic, toxic, having proper water solubility and have a high lipophilicity were selected. Mutagenic, bioavailable and toxic substances were selected for which the values of solubility in water and fats were analyzed. In particular, there were selected 14 compounds well soluble in water and 6 with large lipophilicity value. The predicted LogP, LogS_w, Log(IGC50-1) values and bioavailability for most environmentally hazardous molecules from the dataset are presented in Table 2 and 3.

Conclusion

Non- experimental screening of the lipophilicity, solubility in water, toxicity, bioavailability and mutagenicity was carried out for the database consisting of 489 pesticides and their active substances of different classes of organic compounds. For prediction of oral human bioavailability of pesticides the previously developed RF model was used. Using screening results the most environmentally hazardous substances which are mutagenic, bioavailable, toxic, having high water and fats solubility were identified.

The satisfactory coincidence between the experimental data and predicted by QSAR models was revealed (coefficient of determination in the range 83-94 %). Thus, this allows the use of these models as a non experimental, ecotoxicity prior prediction calculation tool for new pesticides.

Acknowledgment

The authors are thankful for Prof. J. Leszczynski (Interdisciplinary Center for Nanotoxicity, Department of Chemistry and Biochemistry, Jackson State University, Jackson, Mississippi, USA) and Prof. L. Gorb (Badger Technical Services, LLC, Vicksburg, MS, USA) for providing facilities of work.

References

- [1]. Kuz'min, V. E.; Artemenko, A. G.; Muratov, E. N.; Polischuk, P. G.; Ognichenko, L. N. Liahovsky, A. V.; Hromov, A. L.; Varlamova, E. V. In *Recent Advances in QSAR Studies*. Puzyn, T.; Cronin, M.; Leszczynski, J. (eds), **2009**, Springer, New York.
- [2]. Kuz'min, V. E.; Artemenko, A. G.; Muratov, E. N. *J. Comp. Aid. Mol. Des.* **2008**, *22*, pp. 403–421.
- [3]. Ognichenko, L. N.; Kuz'min, V. E.; Gorb, L.; Hill, F. C.; Artemenko, A. G.; Polischuk, P. G.; Leszczynski, J. *Mol. Inf.*, **2012**, *31*, pp. 273–280.
- [4]. Kovdienko, N. A.; Polischuk, P. G.; Muratov, E. N.; Artemenko, A. G.; Kuz'min, V. E.; Gorb, L.; Hill, F.; Leszczynski, J. *Mol. Inf.* **2010**, *29*, pp. 394 – 406.

- [5]. Golovenco, M.Ja.; Kuz'min, V. E.; Artemenko, A. G.; Kulinskii, M.A.; Polishchuk, P.G.; Borisiuk, J.Ju. Prognozirovanie biodostupnosti likarskih zasobiv metodom klasificatsiinih modelei. *Jurnal kliniceskaia informatica i telemeditsina*, **2011**, 8, 88-92 (in ukrainian).
- [6]. Sushko, Yu.; Novotarskyi, S.; Korner, R.; Pandey, A. K. et al. *J. Chem. Inf. Model.* **2010**, 50, pp. 2094–2111.
- [7]. Polishchuk, P. G.; Muratov, E. N.; Artemenko, A. G.; Kolumbin, O. G.; Muratov, N. N.; Kuz'min, V. E. Application of Random Forest Approach to QSAR Prediction of Aquatic Toxicity. *J. Chem. Inf. Model.* **2009**, 49, pp. 2481–2488.
- [8]. Breiman, L. *Machine Learning*. **2001**, 45, pp. 5-32.
- [9]. Artemenko, A. G.; Muratov, E. N.; Kuz'min, V. E.; Kovdienko, N. A.; Hromov, A. I.; Makarov, A. A.; Riabova, O. B.; Wutzler, P.; Schmidtke, M. *J. Antimicrob. Chemother.* **2007**, 60, pp. 68–77.
- [10]. Kuz'min, V. E.; Artemenko, A. G.; Muratov, E. N.; Volineckaya, I. L.; Makarov, V. A.; Riabova, O. B.; Wutzler, P.; Schmidtke, M. *J. Med. Chem.* **2007**, 50, pp. 4205–4213.
- [11]. Kuz'min, V. E.; Muratov, E. N.; Artemenko, A. G. Gorb, L.; Qasim, M.; Leszczynski, J. *J. Comp. Aid. Mol. Des.* **2008**, 22, pp. 747–759.
- [12]. Tetko, I. V.; Tanchuk, V. Y.; Villa, A. E. P. *J. Chem. Inf. Comp. Sci.* **2001**, 41(5), pp. 1407-1414.