Ecotoxicology and Environmental Safety 128 (2016) 52–60

Contents lists available at ScienceDirect

Ecotoxicology and Environmental Safety

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

Application of quantitative structure-property relationship analysis to estimate the vapor pressure of pesticides

Mohammad Goodarzi [a], Leandro dos Santos Coelho [b,c], Bahareh Honarparvar [d], Erlinda V. Ortiz [e], Pablo R. Duchowicz [f,*]

[a] Department of Biosystems, Faculty of Bioscience Engineering, Katholieke Universiteit Leuven – KULeuven, Kasteelpark Arenberg 30, B-3001 Heverlee, Belgium

[b] Department of Electrical Engineering, Federal University of Parana (UFPR), Rua Cel. Francisco Heraclito dos Santos, 100, 81531-980 Curitiba, PR, Brazil

[c] Industrial and Systems Engineering Graduate Program (PPGEPS), Pontifical Catholic University of Parana (PUCPR), Imaculada Conceição, 1155, 80215-901 Curitiba, PR, Brazil

[d] School of Pharmacy and Pharmacology, University of KwaZulu-Natal, Durban 4001, South Africa

[e] IMCoDeG (CONICET), Fac. de Tecnología y Cs. Aplicadas, Universidad Nacional de Catamarca, Maximio Victoria 55, Catamarca, Argentina

[f] Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas INIFTA (CCT La Plata-CONICET, UNLP), Diag. 113 y 64, Sucursal 4, C.C. 16, 1900 La Plata, Argentina

ARTICLE INFO

Article history:
Received 14 September 2015
Received in revised form
23 January 2016
Accepted 25 January 2016

Keywords:
Quantitative Structure-Property Relationships (QSPR)
Multivariable Linear Regression (MLR)
Pesticides
Vapor pressure (VP)

ABSTRACT

The application of molecular descriptors in describing Quantitative Structure Property Relationships (QSPR) for the estimation of vapor pressure (VP) of pesticides is of ongoing interest.

In this study, QSPR models were developed using multiple linear regression (MLR) methods to predict the vapor pressure values of 162 pesticides. Several feature selection methods, namely the replacement method (RM), genetic algorithms (GA), stepwise regression (SR) and forward selection (FS), were used to select the most relevant molecular descriptors from a pool of variables. The optimum subset of molecular descriptors was used to build a QSPR model to estimate the vapor pressures of the selected pesticides. The Replacement Method improved the predictive ability of vapor pressures and was more reliable for the feature selection of these selected pesticides. The results provided satisfactory MLR models that had a satisfactory predictive ability, and will be important for predicting vapor pressure values for compounds with unknown values. This study may open new opportunities for designing and developing new pesticide.

1. Introduction

In light of pesticide's role as environmental contaminants, considerable efforts have been made to monitor the maximum allowed concentration of pesticide residues in surface, ground and drinking water (Papadakis et al., 2015; Otieno et al., 2015). Vapor pressure (VP) is a significant chemical property in the environmental fate, transport and distribution of compounds in water, air and soil. VP is defined as the partial pressure of a chemical in the gas phase in equilibrium with the pure solid or liquid phase. The VP of pesticides determines the distribution of the portion remaining in atmosphere in the gas phase and that settling onto the soil in the liquid or solid phase. It is important for estimating the liquid viscosity, enthalpy of vaporization and other important physico-chemical properties. It also plays an important role in fire and explosion prevention, gas separation, and the design and optimization of the processes (process engineering) and process control (Lawson, 1980; Redeker, 1997; Sandler et al., 2002; Stoll, 2005). Due to the low vapor pressure values of the majority of plant-protection products, their experimental measurements are an ongoing challenge (Meulenberg et al., 1995). Techniques to predict the properties of pesticide compounds are therefore important in a range of fields.

The basis of quantitative structure-property/activity relationship (QSPR/QSAR) methods is that the different behavior of the compounds (*i.e.* physical or chemical properties) can be correlated with changes in the corresponding molecular features (descriptors) (Goodarzi and Freitas, 2008a, 2008b; Freitas et al., 2008; Goodarzi et al., 2009a, 2009c). Once a QSAR/QSPR model is developed, it can be used to predict the chemical or biological properties of new compounds. QSAR/QSPR models are based on physicochemical, geometrical, topological and electronic properties that are related to the molecular structures. The models capture the inherent information from expensive experimental data

* Corresponding author.
E-mail address: pabloducho@gmail.com (P.R. Duchowicz).

http://dx.doi.org/10.1016/j.ecoenv.2016.01.020
0147-6513/© 2016 Elsevier Inc. All rights reserved.

and thereby save significant resources in the process of designing new compounds. In addition, QSAR/QSPR techniques have several applications in designing virtual compound libraries and combinatorial ones with appropriate absorption, distribution, metabolism and excretion (ADMET) properties, as well as computationally optimizing compounds. Several studies found the implication of QSPR to predict the vapor pressure of organic compounds, but very few results have been published for pesticides (Mamy et al., 2015). Staikova et al. (2004) constructed the QSPR models to estimate vapor pressure, octanol–air partition coefficients (K_{OA}) of a various chlorinated organic compounds (not pesticides). It was found that molecular polarizability (α) was strongly correlated with log P_L (liquid vapor pressures) and log K_{OA}. The squared correlation coefficients (R^2) for plots of experimental log P_L and log K_{OA} values, as a function of molecular polarizability, were around 0.98. Puzyn and Falandysz (2005) built a QSPR model for all 75 congeners of chloronaphthalene, which was based on the logarithm of K_{OA} and sub-cooled liquid vapor pressure (log P_L), the R^2 of their linear regression being 0.994.

Ding et al. (2006) used partial least squares (PLS) approach as a linear QSPR based model on quantum chemical and topological descriptors for solid vapor pressure of PCDD/Fs. Two topological predictive descriptors, namely Kier symmetry index (S0K) and Kier flexibility index (PHI) were found to improve the prediction ability of the obtained models (R^2=0.972). Godavarthy et al. (2006) presented a scaled variable reduced coordinates (SVRC) model to a diverse dataset containing over 1221 molecules involving 73 classes of chemicals. The results for the 52,445 datasets indicated that SVRC model represent these saturated vapor pressure data with a 0.35% average absolute deviation (AAD), while the generalized SVRC–QSPR model predicted the saturated vapor pressures with 0.5% AAD. Zeng et al. (2007) constructed a linear QSAR model for 209 polychlorinated diphenyl ethers to predict subcooled liquid vapor pressure. The R^2 of the first linear model was 0.998, which improved to 0.991 when other subsets of descriptors were considered, such as the numbers of Cl substitutions.

Katritzky et al. (2007) constructed a QSPR model on vapor pressures of the 645 diverse organic compounds with a variety of molecular descriptors, such as topological, electronic, geometrical, and hybrid type series. These four descriptors had the R^2 of 0.937 and appeared in the best QSPR model. Wang et al. (2008) built a linear QSPR model on 209 polybrominated diphenyl ethers (PBDEs). The calculated descriptors were used to establish two QSPR models to predict the P_L and K_{OA} of PBDEs. This approach was based on the theoretical linear solvation energy relationship (TLSER) and the RSME values were 0.069 and 0.062, respectively.

Nakajoh et al. (2009) introduced a Conductor-like Screening Model for Real Solvents (COSMO-RS) to predict the liquid phase vapor pressure of chlorobenzenes (CBzs) and polychlorinated biphenyls (PCBs). The liquid phase vapor pressures of 12 CBzs, biphenyl and 26 PCBs, and their enthalpies of vaporization were derived from the temperature dependency of the predicted vapor pressure using the Clausius–Clapeyron equation. The RMSE values of the COSMO-RS predictions for CBzs, and the non-ortho, mono-ortho, di-ortho and tri-ortho congeners of PCBs were in the range of 0.035–0.539, 0.079–0.21, 0.28–0.58, 0.47–0.74 and 0.77–0.87, respectively. Lazzus (2009) presented Artificial Neural Networks (ANN) as a nonlinear method to estimate solid vapor pressures for 212 organic and inorganic compounds. His model was constructed using five physicochemical properties to discriminate among the different substances. The 152 compounds were used to train the network and evaluate his model.

The proposed method represents an alternative approach to those described elsewhere, the intention being to develop a model that estimates the solid vapor pressures that can be used with confidence for any substances. Goudarzi and Goodarzi (2009) developed QSPR models on 55 halogenated methyl-phenyl ether (anisole) compounds using linear and nonlinear methods. The RMSE values of the training set and the test set for the PC-ranking-LS-SVM model were 0.2912 and 0.2427, and the correlation coefficients were 0.9259 and 0.9112, respectively.

Many studies have been concerned with developing QSPR models for a number of organic compounds, with little emphasis on the development of the predictive QSPR models to estimate vapor pressure for various pesticide agents. Although QSPR methods have been successfully used to predict several effective physico-chemical descriptors, the extent of their applications was limited. The majority of these models were derived from very limited data sets, and most of scientists focused on constructing their QSPR models with little attention being paid to validate their model. In a recently published review paper by Mamy et al. (2015), almost no QSAR modeling has been reported to predict the vapor pressure of pesticides. This study aims to develop QSPR models to estimate the vapor pressure for libraries of pesticides. Several feature selection methods were used to probe the applicability of the obtained QSPR models as a powerful chemometrics tool.

2. Materials and methods

2.1. Dataset

In this QSPR studies, 162 pesticides representing several chemical classes were collected, which are widely identified pollutants in the atmosphere. The vapor pressure of these pesticides (Barcelo and Hennion, 1997) in logarithmic scale were used for QSPR analysis as the response variables (Table 1).

2.2. Feature selection methods

To generate optimal sets of molecular descriptors for a given structure, reasonable and statistically valid feature detection (or descriptor selection) methods were applied to construct the QSPR model. An ideal feature selection method aims to choose a small number of the most informative descriptors to facilitate the interpretation of the QSPR model (Eklund et al., 2014). The widely used feature selection methods of forward selection (FS), stepwise regression (SR), replacement method (RM), genetic algorithms (GA), for the libraries of datasets were used in this study, and are described.

2.2.1. Forward selection (FS) method

The forward selection method is an interesting technique due to its didactical point of view as well as its simplicity and accuracy. This technique is based on the stepwise addition of variables sequentially to find the one with the smallest values of fitness function. Each variable is assessed with all the remaining variables until the pair that minimizes the fitness function is found. This procedure is repeated with all the variables until the optimum subset is found or a stopping criterion is met (Chatterjee and Price, 1997; Goodarzi et al., 2009b). The standard deviation (S) as a fitness function is defined as:

$$S = \frac{1}{N-d-1}\left(\sum_{i=1}^{N}\Omega_i^2\right) \tag{4}$$

where N is the number of the training set molecules, d is the number of variables (descriptors) of the model and Ω stands for the residual of molecule i (difference between the experimental and predicted properties).

Table 1
The selected 162 pesticide compounds with their corresponding vapor pressure values.

Number	Compounds	Vapor pressure (Pa)	−log VP
1	Acephate	2.3E−4	3.638
2	Acifluorfen	1E−5	5.000
3	Aclonifen	1.6E−5	4.796
4	Alachlor	2.9E−3	2.537
5	Aldicarb	1.3E−2	1.886
6	Aldrin	3.6E−2	1.444
7	Aldoxycarb	0.012	1.921
8	Ametryn	3.65E−4	3.438
9	Amitrole	5.5E−8	7.259
10	Atrazine	3.9E−5	4.409
11	Azinphos-ethyl	3.2E−4	3.495
12	Azinphos-methyl	1.8E−4	3.745
13	Benalaxyl	6.7E−4	3.174
14	Benazolin	1E−7	7.000
15	Bendiocarb	4.6E−3	2.337
16	Bensulfuron	2.8E−12	11.553
17	Bentazone	4.6E−4	3.337
18	Bifenox	3.2E−4	3.495
19	Bromacil	4.1E−5	4.387
20	Bromofenoxim	1.3E−6	5.886
21	Bromophos-ethyl	6.1E−3	2.215
22	Bupirimate	0.1E−3	4.000
23	Butocarboxin	10.6E−6	4.975
24	Butachlor	0.6E−3	3.222
25	Butoxycarboxim	2.7E−4	3.568
26	Carbaryl	4.1E−5	4.387
27	Carbendazim	9E−5	4.046
28	Carbofuran	3.1E−5	4.508
29	Carboxin	2.5E−5	4.602
30	Chlorbromuron	5.3E−5	4.275
31	Chlordane	1.3E−3	2.886
32	Chlorfenvinphos	1E−3	3.000
33	Chlorobenzilate	1.2E−4	3.921
34	Chlornitrofen	3.2E−3	2.495
35	Chlorothalonil	7.6E−5	4.119
36	Chlorotoluron	1.7E−5	4.769
37	Chlorthal dimethyl	2.1E−4	3.678
38	Chlorthiamide	1.3E−4	3.886
39	Clofentezine	1.3E−7	6.886
40	Clopyralid	1.33E−3	2.876
41	Coumaphos	1.3E−5	4.886
42	Cyanazine	2.E−7	6.698
43	Cycloate	2.13E−3	2.672
44	Cyhalothrin	1E−6	6.000
45	Cymoxanil	0.8E−5	5.097
46	Cypermethrin	2.3E−7	6.638
47	Cyproconazole	3.5E−5	4.456
48	Desmedipham	4E−8	7.398
49	Desmetryn	1.33E−4	3.876
50	Diazinon	1.2E−2	1.921
51	Dicamba	4.5E−3	2.347
52	Dichlobenil	0.088	1.055
53	Diclofop	2.5E−3	2.602
54	Diethofencarb	8.4E−3	2.076
55	Difenoconazole	3.3E−8	7.481
56	Diflubenzuron	1.2E−7	6.921
57	Dimefuron	0.1E−3	4.000
58	Dimethoate	1.1E−3	2.958
59	Dinoterb	2E−2	1.698
60	Disulfoton	7.2E−3	2.143
61	Diuron	1.1E−3	2.958
62	DNOC	1.4E−2	1.854
63	Endosulfan	8.3E−4	3.081
64	EPTC	1E−5	5.000
65	Esfenvalerate	2E−7	6.698
66	Ethalfluralin	0.012	1.921
67	Ethiofencarb	4.5E−4	3.347
68	Ethion	2E−4	3.698
69	Ethirimol	2.67E−4	3.573
70	Ethofumesate	6.5E−4	3.187
71	Ethoprophos	0.046	1.337
72	Etofenprox	3.2E−2	1.494
73	Etridiazole	0.019	1.721
74	Fenamiphos	0.12E−3	3.921
75	Fenitrothion	1.8E−2	1.745
76	Fenoxaprop-P	5.3E−7	6.276
77	Fenoxycarb	8.67E−7	6.062
78	Fenpropathrin	7.3E−4	3.137
79	Fenpropidin	1.7E−2	1.769
80	Fenpropimorph	2.3E−3	2.638
81	Fenthion	7.4E−4	3.131
82	Fenuron	2.1E−2	1.678
83	Fipronil	3.7E−7	6.432
84	Fluazipop	5.5E−5	4.259
85	Fluometuron	1.25E−4	3.903
86	Fluridone	1.3E−5	4.886
87	Flusilazole	3.9E−5	4.408
88	Fonofos	2.8E−2	1.553
89	Heptachlor	0.053	1.276
90	Hexaconazole	1E−5	5.000
91	Hexazinone	8.5E−3	2.070
92	Iprodione	5E−7	6.301
93	Isofenphos	2.2E−4	3.657
94	Isoproturon	3.3E−5	4.481
95	Isoxaben	5.5E−4	3.259
96	Lenacil	0.2E−6	6.698
97	Lindane	5.6E−3	2.252
98	Linuron	5.1E−5	4.292
99	Malathion	5.3E−3	2.276
100	Mecoprop	0.31E−3	3.508
101	Metamitron	8.6E−7	6.065
102	Metazachlor	4.9E−5	4.309
103	Methabenzthiazuron	5.9E−6	5.229
104	Methidathion	2.5E−4	3.602
105	Methiocarb	1.5E−5	4.824
106	Methomyl	6.65E−3	2.177
107	Metobromuron	4E−4	3.398
108	Metolachlor	4.2E−3	2.377
109	Metoxuron	4.3E−3	2.366
110	Metribuzin	5.8E−5	4.236
111	Monolinuron	1.3E−3	2.886
112	Myclobutanil	2.13E−4	3.672
113	Napropamide	5.3E−4	3.276
114	Parathion	8.9E−4	3.051
115	Parathion-methyl	0.2E−3	3.698
116	Penconazole	2.1E−4	3.678
117	Pendimethalin	4E−3	2.398
118	Permethrin	4.5E−5	4.347
119	Phenthoate	5.3E−3	2.276
120	Phorate	8.5E−2	1.070
121	Phosmet	6.5E−5	4.187
122	Phoxim	2.1E−3	2.677
123	Picloram	8.2E−5	4.086
124	Pirimicarb	0.97E−3	3.013
125	Pirimiphos-ethyl	0.68E−3	3.167
126	Pirimiphos-methyl	2E−3	2.698
127	Prochloraz	1.5E−4	3.824
128	Procymidone	1.8E−2	1.745
129	Prometon	0.306E−3	3.514
130	Prometryn	0.169E−3	3.772
131	Propanil	2.6E−5	4.585
132	Propazine	3.9E−6	5.408
133	Propiconazole	5.6E−5	4.252
134	Propoxur	1.3E−3	2.886
135	Propyzamide	5.8E−5	4.236
136	Prosulfocarb	6.9E−5	4.161
137	Pyridate	1.3E−7	6.886
138	Simazine	2.94E−6	5.531
139	Sulfotep	1.4E−2	1.854
140	Tebuconazole	1.3E−6	5.886
141	Terbacil	6.25E−5	4.204
142	Tebuthiuron	0.27E−3	3.568
143	Teflutrin	8E−3	2.097
144	Terbufos	3.46E−2	1.461
145	Terbumeton	0.27E−3	3.568
146	Terbuthylazine	0.15E−3	3.824
147	Terbutryn	0.225E−3	3.648
148	Tetrachlor-vinphos	5.6E−6	5.252
149	Tetraconazole	1.6E−3	2.796
150	Tetramethrin	9.44E−4	3.025

Table 1 (*continued*)

Number	Compounds	Vapor pressure (Pa)	− log VP
151	Thiodicarb	5.7E − 3	2.244
152	Thiofanox	2.26E − 2	1.646
153	Thiometon	2.3E − 2	1.638
154	Thiram	2.3E − 3	2.638
155	Triallate	1.6E − 2	1.796
156	Triazophos	0.39E − 3	3.408
157	Trichlorfon	2.1E − 4	3.677
158	Triclopyr	2E − 4	3.698
159	Tricyclazole	2.7E − 5	4.568
160	Tridemorph	6.4E − 3	2.194
161	Trifluralin	9.5E − 3	2.022
162	Vinclozolin	1.6E − 5	4.796

2.2.2. Stepwise regression (SR) method

The Stepwise regression (SR) method is one of the common feature selection techniques applied in QSAR/QSPR studies, and is a combination of forward and backward procedures. A variable that enters the model in the earlier selection stages may be eliminated during the later steps. The resulting algorithm gradually adds new variables to the model, starting from the x variable, which has the largest empirical correlation with the dependent variable y, as in the forward selection method. In addition, the algorithm incorporates a mechanism for eliminating variables, as in the backward elimination method. The selection process of the most important variables is the same in the forward selection and backward procedures. The number of selected variables depend on the fitness function assumed for inclusion and exclusion of the variables from the model (Broersen, 1986).

2.2.3. Replacement method (RM)

A Full Search (FS) of optimal variables requires $D!/(D-d)!d!$ linear regressions and is impractical, where d is an optimal subset of d descriptors $d=\{X_1, X_2,\ldots, X_d\}$ and D referred to as the pool of descriptors. However, the Replacement Method (RM) (Duchowicz et al., 2005; Duchowicz et al., 2006; Duchowicz et al., 2006; Talevi et al., 2011; Mercader et al., 2011) generates linear regression QSAR models that are quite close to that of the FS. The basis of the RM algorithm is to choose an optimal subset of d descriptors $d=\{X_1, X_2,\ldots, X_d\}$, $(d<D)$ and descriptors from the pool of D descriptors with minimum standard deviation (S) (Duchowicz et al., 2013; Sun et al., 2009).

The RM technique consist of five steps, firstly, the d descriptors, $d=\{X_1, X_2,\ldots, X_d\}$ are selected randomly and a linear regression is performed. Secondly, one of the descriptors of this set is chosen, for instance X_i, and is replaced by each of the D descriptors of the pool (except itself), with the best obtaining set being kept. As any of the d descriptors in the initial model can be replaced, a regression equation with d variables has d possible paths to achieve the final result. Thirdly, the variables with greatest relative error in their coefficient (except the one replaced in the previous step) are chosen and replaced with all the D descriptors (except itself), with the best set being maintained. Fourthly, the rest of the remaining variables are replaced in the same way, by passing those replaced in previous steps. Finally, the variable with the greatest relative error in the coefficient is reconsidered and the whole process repeated. This iterative procedure is continued until the set of descriptors remains unchanged. The final stage presents the best model for the path i with the same process occurring for all possible paths $i=1, 2,\ldots, d$, with the obtained models being compared to keep the best one. RM feature selection method provides QSPR models with more suitable (improved) statistical parameters in comparison to the Forward Stepwise Regression method (Draper and Smith, 1981) and various elaborated Genetic Algorithms (Mercader et al., 2010).

2.2.4. Genetic algorithm (GA)

The Genetic algorithm (GA) is a commonly used optimization technique that is useful for image processing, designing complex networks (*e.g.* computers and integrated circuits), classifications, job scheduling, robotics and parameter fitting (Davis, 1991). GA is a stochastic method that has been widely used as a feature selection method. In QSAR/QSPR studies, the evolution of the population is simulated to select the most relevant descriptors (Hunger and Huttner, 1999; Waller and Bradley, 1999), with a chromosome of binary values delineating each of the population. In genetic terms, each variable is called a gene, and a set of variables is called a chromosome. It should be noted that the population of the first generation is selected randomly, and the state of each variable is represented by the value of 1 (selected variables in the model) or zero (not selected), with the genes take the values of 1 or 0. The selected variables (genes with a value of 1) are collected as a small subset of descriptors in a way that the probability of generating 0 for a gene is at least 60% greater than 1 (Aires-de-Sousa et al., 2002). Crossover and Mutations are operators used with the probabilities of 60% and 0.1%, respectively. The fitness value for each chromosome is calculated as Root Mean Square Error (RMSE). For different runs of genetic algorithm, the population size varies between 50 and 200.

2.3. Descriptor calculations and selection

The chemical structures of the molecules for the study were drawn using Hyperchem package (HyperChem Version 7.0., 2007), and the final geometries were obtained with the semi-empirical AM1 method. Optimization was preceded by the Polak–Rebiere algorithm to reach 0.01 root mean square gradient. The optimized geometry was transferred into the Gaussian 09 (Gaussian 09 RA et al., 2009) and Dragon (Mauri et al., 2006) programs to calculate molecular descriptors. The highest occupied molecular orbital (HOMO), lowest unoccupied molecular orbital (LUMO) energies and molecular dipole moment were calculated with Gaussian package. Quantum chemical indices including hardness (η), softness (S), electronegativity (χ) and electrophilicity (ω) were calculated using equations proposed by Thanikaivelan et al. (HyperChem Version 7.0., 2007). Some chemical parameters, namely molar volume (V), molecular surface area (SA), hydrophobicity ($\log P$), hydration energy (HE) and molecular polarizability (α), were calculated using Hyperchem software. DRAGON software was used to calculate various descriptors, such as constitutional, topological, geometrical, charge, GETAWAY (geometry, topology and Atoms-Weighted AssemblY), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D molecular Representation of Structure based on Electron diffraction), Molecular Walk Count, BCUT, 2D-Autocorrelation, Aromaticity Index, Randic molecular profile, Radial Distribution Function, Functional group and Atom-Centered Fragment classes (Mercader et al., 2011). Multiple linear regressions (MLR) were utilized as a linear technique with various RM, GA, SR-MLR, and FS feature selections methods. The best subset was selected and used for further analysis.

In MLR analysis, the calculated descriptors were first explored to identify the constant or approximately constant variables. The initial sets of descriptors were then reduced by eliminating all descriptors with insignificant variance to avoid random and irrelevant series in the model. At the same time, to decrease the redundancy amongst the descriptor data matrix, between the pairs of the two highly correlated descriptors, the one with the highest property correlation were kept, and the others were excluded from the data matrix. MLR analysis selects the most suitable models at each rank. The final model between them was chosen which sufficiently correlated and that prevented any model overfit or overparameterizations. All programs were written using Matlab software (Waller and Bradley, 1999). Several RM, GA, SR-MLR, and FS feature selection methods were compared to select the most relevant descriptors from the pool of options. The best subset was selected and used for further analysis of the constructed QSPR model.

3. Results and discussion

A feature selection method is considered an important preprocessing step in machine learning and data mining processes, particularly in QSAR/QSPR studies with high dimensional feature spaces. These tools can efficiently reduce the dimensionality of available datasets, and improve not only the classification, but also the regression accuracy by identifying relevant features. In this context, one of the most important steps, which affect the complexity of modeling methods in QSAR/QSPR studies, is the selection of the most effective descriptors from a pool of descriptors. It is also notable that a model's predictive capability for different pesticides depends on how well the training set represents these chemicals, and how robust the model is in extrapolating beyond the dimensional space defined by the training set. Consequently, the selection of the training set in QSPR analysis is of great importance.

In this work, the total data ($n=162$) were first divided into two parts, the so-called training set ($n=122$) and test (external validation) set ($n=40$). In each case, 75% of the total compounds were nominated as training sets while the remaining 25% of the datasets used as the test set. The division of the dataset was based on the property range of training set, which included the test set properties. All the feature selection methods were then performed on the training set to find the most important subset, the influences of the number of descriptors being investigated from one to ten to select the optimum number of descriptors. Fig. 1 shows the plot of correlation coefficients versus the different number of descriptors selected using the GA, FS, SR and RM feature selection methods. The selected descriptors by the GA, FS, SR and RM feature selection methods were all acceptable. However, based on the comparison of the R^2-values in Fig. 1, the selected descriptors using RM with R^2 between 0.7 and 0.9 appeared to be more reliable. The selected descriptors using RM method as a feature selection approach was used to construct the QSPR model which correlated the structural information to the vapor pressure of pesticides. This enabled the selection of the most important descriptors with heuristic search algorithms using a fitness function that minimized the difference between the original feature set and the one in the reduced feature subsets. The fitness function measures how well a solution fits the problem and should be fast to calculate. It can also be inferred from Fig. 1 that when the descriptors were added into the model, the correlation coefficient was improved to 0.9, and the model results for predicting the log VP were more suitable. The graph of calculated vapor pressure (−log VP) against experimental values using RM/MLR equation is provided in supplementary information (Fig. S1). The contribution of the different number of descriptors on the predictive potential of training set, the leave one out (LOO) cross-validation of training set and test set are shown in Fig. 2 and Table 2. According to Fig. 2, an insignificant improvement on the statistics of test set was observed for 6–10 descriptors. In other words, while the correlation coefficients of the training set were increased, the correlation coefficient of test set decreased. This implied that the optimum subset size could be achieved with a maximum five of descriptors.

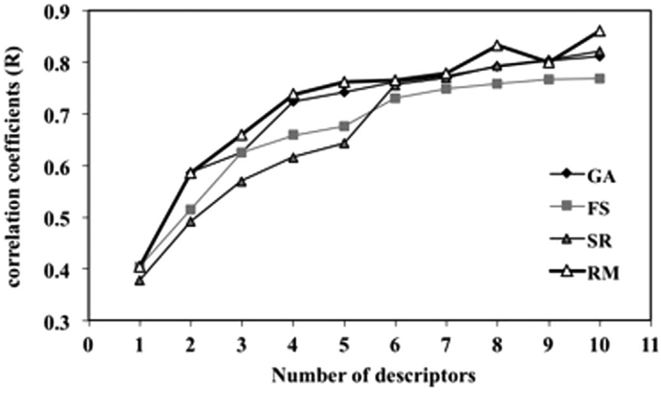

Fig. 1. Correlation coefficient (R) as a function of the number of 10 descriptors used with different feature selection methods.

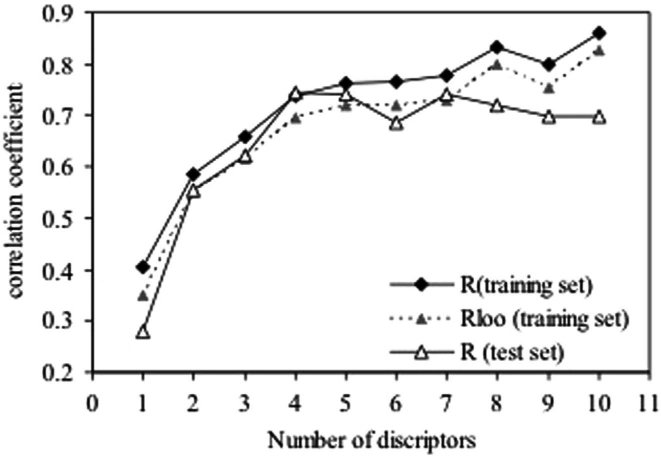

Fig. 2. Correlation coefficient of training set ($R_{\text{training set}}$), leave-one-out cross-validation correlation coefficient (R_{loo}) and external validation ($R_{\text{test set}}$) with different number of descriptors using replacement method (RM).

To select the most relevant descriptors and build a linear QSPR model, all mentioned feature selection methods were performed by assigning 122 compounds as training set and 40 as test sets (external sets). The following equations were obtained by different feature selection methods:

$$-\log \text{VP (RM/MLR)} = 8.225\,(\pm 1.449) + 0.8404\,(\pm 0.114) \times \text{SNar} - 2.719\,(\pm 0.514) \times \text{VRv2} + 3.084(\pm 0.343) \times \text{RDF010u} - 9.021(\pm 0.922) \times \text{RDF010p} + 1.709(\pm 0.538) \times \text{Mor32e}$$

where

$N_{train} = 122,\ \ R = 0.762,\ \ S = 1.093,\ \ p < 10^{-4}$
$R_{loo} = 0.720,\ \ S_{loo} = 1.172,\ \ R_{l-25\%-o} = 0.604,\ \ S_{l-25\%-o} = 1.371,\ \ S_{Rand} = 1.417$
$N_{test} = 40,\ \ R_{test} = 0.741,\ \ S_{test} = 1.108$

3.1. Topological descriptors

Narumi simple topological index (log) (SNar), Average Randic-type eigenvector-based index from van der Waals Weighted distance matrix (VRv2) *RDF* descriptors: Radial Distribution function-1.0/unweighted (RDF010u), Radial Distribution function-1.0/

Table 2
Statistical parameters for different subsets selected with replacement method (RM) using 1–10 descritores.

S	***R***	***S***S_{loo}	R_{loo}	S_{test}	R_{test}	**Descriptors**
1.518	0.404	1.557	0.349	1.493	0.279	Mor22m
1.350	0.586	1.392	0.551	1.346	0.555	HNar, IC1
1.257	0.659	1.319	0.616	1.262	0.621	RDCHI, RDF010u, RDF010p
1.135	0.737	1.208	0.696	1.102	0.744	SNar, VRv2, RDF010u, RDF010p
1.093	0.762	1.172	0.720	1.108	0.741	SNar, VRv2, RDF010u,RDF010p, Mor32e
1.091	0.765	1.179	0.719	1.226	0.686	X0, X1, RDF010u, RDF075m, RDF010p, Mor32e
1.071	0.777	1.166	0.730	1.142	0.742	nBT, X1, SEigv, MATS1m, RDF010u, RDF010p, R3e+
0.948	0.832	1.029	0.799	1.306	0.719	TI1, GGI2, GATS2e, RDF020m, RDF020p, Mor01m, Vp, MLOGP
1.031	0.799	1.129	0.754	1.275	0.697	Se, Eig1e, BELv2, RDF025u, RDF055u, RDF010e, Mor01m, Mor22m, MLOGP
0.879	0.860	0.969	0.828	1.383	0.699	TI1, GGI2, GATS4v, GATS2e, RDF020m, RDF075m, RDF020p, Mor01m,Vp, MLOGP

Note 1: ***S***: model's standard deviation from calibration; ***R***: correlation coefficient of the training set, and sub-index *loo* and *l-25%-O* stand for the Leave-One-Out and leave 25% out Cross Validation technique, respectively.

Note 2: **Mor22m**: weighted by atomic masses; **HNar**: Narumi harmonic topological index; **IC1**: Information content index neighborhood symmetry of 1-order; **RDCHI**: **RDF010u:** unweighted Radial Distribution function-1.0; **RDF010p**: Radial Distribution function-1.0/weighted by atomic polarizabilities; **SNar**: Narumi simple topological index (log); **VRv2**: Average Randic-type eigenvector-based index from van der Waals Weighted distance matrix; **RDF075m**: Radial Distribution function-8.5/weighted by atomic masses; **Mor32e**: 3D-MoRSE signal 32/weighted by atomic Sanderson electronegativities; **nBT:** number of bonds**; SEigv:** Eigenvalue sum from van der Waals weighted distance matrix**; MATS1m:** Moran autocorrelation-lag 1/weighted by atomic masses**; R3e+**: R maximal autocorrelation of lag 3/weighted by atomic Sanderson electronegativities; **TI1:** first Mohar index TI1; **GGI2:** topological charge index of order 2;GATS2e, **RDF020m**: Radial Distribution function-2/weighted by atomic masses; **RDF020p:** Radial Distribution Function-2.0/weighted by atomic polarizabilities; **Mor01m**: 3D-MoRSE signal 01/weighted by atomic masses; **Vp:** V total size index/weighted by atomic polarizabilities; **MLOGP**: Moriguchi octanol-water partition coefficient; **Se:** sum of atomic Sanderson electronegativities (scaled on Carbon atom)**; Eig1e:** Leading eigenvalue from electronegativity weighted distance matrix; **BELv2:** lowest eigenvalue n**; RDF025u:** Radial Distribution Function-2.5/unweighted**; RDF055u**: Radial Distribution Function-5.5/unweighted; **RDF010e:** Radial Distribution Function-1.0/weighted by atomic Sanderson electronegativities; **Mor01m:** 3D-MoRSE signal 01/weighted by atomic masses; **Mor22m**: 3D-MoRSE signal 22/weighted by atomic masses; **TI1:** first Mohar index TI1; **GGI2:** topological charge index of order 2**; GATS4v:** Geary autocorrelation-lag 4/weighted by van der Waal; **RDF020m**: Radial Distribution function-2/weighted by atomic masses, **RDF075m:** Radial Distribution Function-7.5/weighted by atomic masses; **RDF020p:** Radial Distribution function-8.5/weighted by atomic masses, **Mor01m**: 3D-MoRSE signal 01/weighted by atomic masses.

weighted by atomic polarizabilities (RDF010p). **3D-MoRSE descriptors**: 3D-MoRSE signal 32/weighted by atomic Sanderson electronegativities (Mor32e).

$$
\begin{aligned}
-\log \mathrm{VP}\,(\mathrm{GA/MLR}) = {} & 1.016\,(\pm 0.491) + 0.159\,(\pm 0.023) \times \mathrm{SCBO} \\
& -33.343\,(\pm 6.811) \\
& \times \mathrm{JGI5} + 2.919(\pm 0.364) \times \mathrm{RDF010u} \\
& -0.117(\pm 0.029) \\
& \times \mathrm{RDF085u} + 0.116(\pm 0.037) \times \mathrm{RDF075m} \\
& -7.518(\pm 1.007) \\
& \times \mathrm{RDF010p} - 0.248(\pm 0.083) \times \mathrm{MLOGP}
\end{aligned}
$$

$N_{train} = 122,\ R = 0.772,\ S = 1.080,\ p < 10^{-4}$

$R_{loo} = 0.728,\ S_{loo} = 1.167,\ R_{l-25\%-o} = 0.603,\ S_{l-25\%-o} = 1.408,\ S_{Rand} = 1.445$

$N_{test} = 40,\ R_{test} = 0.715,\ S_{test} = 1.183$

3.2. Constitutional descriptors

Sum of conventional bond orders (H-depleted) (SCBO) *Galvez topol.charge indices*: mean topological charge index of order 5 (JGI5). *RDF* descriptors: Radial Distribution function-1.0/unweighted (RDF010u), Radial Distribution function-8.5/unweighted (RDF085u), Radial Distribution function-8.5/weighted by atomic masses (RDF075m), Radial Distribution function-1.0/weighted by atomic polarizabilities (RDF010p).

3.3. Properties descriptors

Moriguchi octanol-water partition coefficient (MLOGP).

$$
\begin{aligned}
-\log \mathrm{VP}\,(\mathrm{SR/MLR}) = {} & -5.455\,(\pm 1.298) + 4.455\,(\pm 0.787) \times \mathrm{HNar} \\
& + 2.785\,(\pm 0.379) \\
& \times \mathrm{RDF010u} - 0.090(\pm 0.028) \times \mathrm{RDF085u} \\
& + 0.142(\pm 0.036) \\
& \times \mathrm{RDF075m} - 7.099(\pm 1.019) \times \mathrm{RDF010p} \\
& + 3.026(\pm 0.963) \\
& \times \mathrm{G1e} - 0.386(\pm 0.092) \times \mathrm{MLOGP} \\
& -0.271(\pm 0.111) \times \mathrm{RDF020m} \\
& + 0.004(\pm 0.001) \times \mathrm{Mor01m}
\end{aligned}
$$

$N_{train} = 122,\ R = 0.803,\ S = 1.021,\ p < 10^{-4}$

$R_{loo} = 0.760,\ S_{loo} = 1.116,\ R_{l-25\%-o} = 0.604,\ S_{l-25\%-o} = 1.391,\ S_{Rand} = 1.411$

$N_{test} = 40,\ R_{test} = 0.713,\ S_{test} = 1.223$

3.4. Topological descriptors

Narumi harmonic topological index (HNar). *RDF descriptors*: Radial Distribution function-1.0/unweighted (RDF010u), Radial Distribution function-8.5/unweighted(RDF085u), Radial Distribution function-8.5/weighted by atomic masses (RDF075m), Radial Distribution function-2/weighted by atomic masses (RDF020m), Radial Distribution function-1.0/weighted by atomic polarizabilities (RDF010p).

3.5. WHIM descriptors

1st component symmetry directional WHIM index/weighted by atomic Sanderson electronegativities (G1e). **Properties descriptors**: Moriguchi octanol-water partition coefficient (MLOGP). *3D-MoRSE descriptors*:3D-MoRSE signal 01/weighted by atomic masses (Mor01m).

$$
\begin{aligned}
-\log V_P(FS/MLR) = {} & -8.714\,(\pm 1.945) \\
& -1.363\,(\pm 0.454) \times \mathrm{Mor22m} \\
& + 5.116\,(\pm 0.910) \\
& \times \mathrm{HNar} + 1.577(\pm 0.335) \times \mathrm{IC1} \\
& -0.267(\pm 0.081) \times \mathrm{MLOGP}
\end{aligned}
$$

$$N_{train} = 122,\quad R = 0.659,\quad S = 1.264,\quad p < 10^{-4}$$
$$R_{loo} = 0.608,\quad S_{loo} = 1.335,\quad R_{l-25\%-o} = 0.493,\quad S_{l-25\%-o} = 1.508,\quad S_{Rand} = 1.411$$
$$N_{test} = 40,\quad R_{test} = 0.514,\quad S_{test} = 1.433$$

3D-MoRSE descriptors: 3D-MoRSE signal 22/weighted by atomic masses (Mor22m).

Topological descriptors: Narumi harmonic topological index (HNar), Information content index neighborhood symmetry of 1-order (IC1). **Properties descriptors**: Moriguchi octanol-water partition coefficient (MLOGP).

In these equations, *N* is the number of the selected compounds; *R* is the correlation coefficient of the training set, *S* stands for the model's standard deviation from calibration, *p* is the significance of the model, and subindex *loo* and *l-25%-O* stand for the Leave-One-Out and leave 25% out cross validation technique, respectively. The S_{Rand} represents the standard deviation according to the Y-Randomization technique (100,000 cases). The results show that RM is more reliable than the other feature selection methods. More information about these descriptors can be found in the handbook of molecular descriptors (Todeschini and Consonni, 2000). For further details, the published article is reported by Katritzky et al. (2007).

Fig. 3 shows their residual values against the predicted −log VP data for the training and test sets using the replacement method (RM). The six commonly used statistical parameters for probing the prediction ability of the constructed model are listed in Table 3. These parameters are: correlation coefficients (R^2), root mean square error of prediction (RMSEP) and percent of relative standard error of prediction (RSEP) and percent of mean absolute error (MAE), predictive residual sum of squares (PRESS) and ratio of PRESS/SST, where SST is the regression sum of squares values. The results confirm the reliability of the considered models. In general, the RM, GA and SR-MLR models, but particularly the RM-MLR case, were found to be efficient well-estimated methods for parameter selection (for both training sets and test sets) to improve the QSPR predictive capability for the selected series of pesticides. FS-MLR method, with the lowest correlation coefficients (R^2) and low values of other statistical parameters (RMSEP, RSEP, MAE, PRESS, and Ratio of PRESS/SST) reported in Table 3, appeared not to be satisfactory parameter selection method. This observation is consistent with our earlier publication, (Wong et al., 2014; Gharagheizi et al., 2012a, 2012b) which reported that the Replacement Method (RM) yields the linear regression QSAR models with much less computational work. We also described that RM gives models with better statistical parameters than the Forward Stepwise Regression procedure and Genetic Algorithm method (GA) (Mercader et al., 2010).

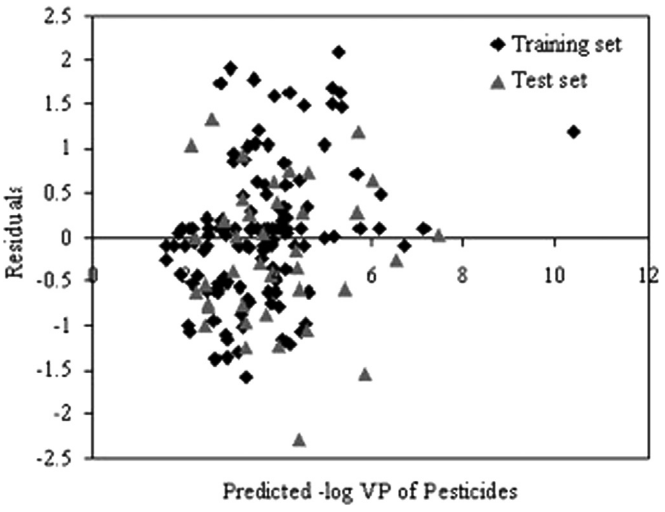

Fig. 3. Plot of residual values against the predicted vapor pressure (−log VP) of pesticides for training set and test set using replacement method (RM).

Table 3
Statistical parameters used to assess QSPR models using RM-MLR, GA-MLR, SR-MLR, and FS-MLR methods.

Parameters	Set	RM-MLR	GA-MLR	SR-MLR	FS-MLR
R^2	Training set	0.5804	0.5974	0.6461	0.4340
	Test set	0.5496	0.5119	0.5095	0.2644
RMSEP	Training set	1.0659	1.0441	0.9790	1.2381
	Test set	1.0220	1.0590	1.0598	1.3407
RSEP(%)	Training set	26.2851	25.7482	24.1409	30.5304
	Test set	25.3665	26.2847	26.3043	33.2762
MAE(%)	Training set	8.3180	8.1737	7.7564	8.9780
	Test set	14.8636	14.7034	14.9099	16.4353
PRESS	Training set	138.6127	133.0081	116.9208	187.0030
	Test set	41.7784	44.8574	44.9246	71.8946
Ratio PRESS/SST	Training set	0.4196	0.4026	0.3539	0.5660
	Test set	0.4660	0.5003	0.5011	0.8019

Note: R^2: square correlation coefficients; **RMSEP**: root mean square error of prediction; **RSEP**: percent of relative standard error of prediction; **MAE**: Percent of mean absolute error; **PRESS**: predictive residual sum of squares; and **Ratio of PRESS/SST** where SST is the regression sum of squares values.

To the best of our knowledge, this is first account to build QSPR model for this series of pesticides. Overall, the prediction performances of all the applied techniques were shown to be the same, and that all the methods evolution of the generation took the same fitness after the iteration as they converged to a similar R^2. The results can be also found in Table S1 in supplementary material.

4. Conclusion

Multiple linear regression (MLR) method was used to construct a quantitative relationship between the vapor pressure values of the selected pesticide agents and their calculated physico-chemical descriptors. Different RM, GA, SR, and FS-MLR feature selection methods were useful for these chemicals to identify the most contributing descriptors in QSPR models. Amongst the applied feature selection method, RM yielded more improved statistical parameters of QSPR models in comparison to the Forward Stepwise Regression method and various elaborated Genetic Algorithms.

Several RM, GA, SR, and FS-MLR validation techniques confirmed the accuracy of the produced QSPR model through calculating its fitness based on both training sets and by testing the prediction capabilities of the model. The encouraging results showed that the QSPR model was robust and could be used successfully to estimate the vapor pressure of pesticides compounds. It appears that a significant requirement for constructing a robust QSPR model should be that it reinforces the performance of reference models (a QSAR model that we are referring to validate our own model) in terms of fitness and prediction quality. The outcome of this study suggests that the predictive performance for physico-chemical properties, and the reliability and interpretability of a QSPR model, can be improved or maintained in comparison to that of the reference model using different RM, GA, SR, and FS feature selection methods.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

We thank the financial support provided by the National Research Council of Argentina (CONICET) PIP11220100100151 project and to Ministerio de Ciencia, Tecnología e Innovación Productiva for the electronic library facilities. EVO and PRD are members of the scientific researcher career of CONICET.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.ecoenv.2016.01.020.

References

Aires-de-Sousa, J., Hemmer, M.C., Gasteiger, J., 2002. Prediction of H-1 NMR chemical shifts using neural networks. Anal. Chem. 74 (1), 80–90.

Barcelo, D., Hennion, M.C., 1997. Trace Determination of Pesticides and their Degradation Products in Water. Elsevier Sciences BV, United States, pp. 21–33.

HyperChem Version 7.0., 2007. Hypercube I. In. Gainesville.

Broersen, P.M.T., 1986. Subset resgression with stepwise directed search applied statistics. J. R. Stat. Soc. Ser. C 35 (2), 168–177.

Chatterjee, S., Price, B., 1997. Regression Analysis by Example. Wiley, New York.

Davis, L., 1991. Handbook of Genetic Algorithm. Van Nostrand-Reinhold, London.

Ding, G.H., Chen, J.W., Qiao, X.L., Huang, L.P., Lin, J., Chen, X.Y., 2006. Quantitative relationships between molecular structures, environmental temperatures and solid vapor pressures of PCDD/Fs. Chemosphere 62 (7), 1057–1063.

Draper, N.R., Smith, H., 1981. Applied Regression Analysis. John Wiley & Sons, New York.

Duchowicz, P.R., Castro, E.A., Fernandez, F.M., 2006. Alternative algorithm for the search of an optimal set of descriptors in QSAR-QSPR studies. Match-Commun. Math. Comput. Chem. 55 (1), 179–192.

Duchowicz, P.R., Castro, E.A., Fernandez, F.M., Gonzalez, M.P., 2005. A new search algorithm for QSPR/QSAR theories: normal boiling points of some organic molecules. Chem. Phys. Lett. 412 (4–6), 376–380.

Duchowicz, P.R., Giraudo, M.A., Castro, E.A., Pomilio, A.B., 2013. Amino acid profiles and quantitative structure-property relationship models as markers for Merlot and Torrontes wines. Food Chem. 140 (1–2), 210–216.

Duchowicz, P.R., Fernandez, M., Caballero, J., Castro, E.A., Fernandez, F.M., 2006. QSAR for non-nucleoside inhibitors of HIV-1 reverse transcriptase. Bioorg. Med. Chem. 14 (17), 5876–5889.

Eklund, M., Norinder, U., Boyer, S., Carlsson, L., 2014. Choosing feature selection and learning algorithms in QSAR. J. Chem. Inf. Model. 54 (3), 837–843.

Freitas, M.P., da Cunha, E.F.F., Ramalho, T.C., Goodarzi, M., 2008. Multimode methods applied on MIA descriptors in QSAR. Curr. Comput.-Aided Drug Des. 4 (4), 273–282.

Gaussian 09 RA, Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Montgomery, Jr., J.A., Vreven, T., Kudin, K.N., Burant, J.C., Millam, J.M., Iyengar, S.S., Tomasi, J., Barone, V., Mennucci, B., Cossi, M., Scalmani, G., Rega, N., Petersson, G.A., Nakatsuji, H., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Klene, M., Li, X., Knox, J.E., Hratchian, H.P., Cross, J.B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R.E., Yazyev, O., Austin, A.J., Cammi, R., Pomelli, C., Ochterski, J.W., Ayala, P.Y., Morokuma, K., Voth, G.A., Salvador, P., Dannenberg, J.J., Zakrzewski, V. G., Dapprich, S., Daniels, A.D., Strain, M.C., Farkas, O., Malick, D.K., Rabuck, A.D., Raghavachari, K., Foresman, J.B., Ortiz, J.V., Cui, Q., Baboul, A.G., Clifford, S., Cioslowski, J., Stefanov, B.B., Liu, G., Liashenko, A., Piskorz, P., Komaromi, I., Martin, R.L., Fox, D.J., Keith, T., Al-Laham, M.A., Peng, C.Y., Nanayakkara, A., Challacombe, M., Gill, P. M.W., Johnson, B., Chen, W., Wong, M.W., Gonzalez, C., Pople, J. A., 2009. Gaussian, Inc., Wallingford CT.

Gharagheizi, F., Eslamimanesh, A., Ilani-Kashkouli, P., Mohammadi, A.H., Richon, D., 2012. QSPR molecular approach for representation/prediction of very large vapor pressure dataset. Chem. Eng. Sci. 76, 99–107.

Gharagheizi, F., Eslamimanesh, A., Ilani-Kashkouli, P., Mohammadi, A.H., Richon, D., 2012. Determination of vapor pressure of chemical compounds: a group contribution model for an extremely large database. Ind. Eng. Chem. Res. 51 (20), 7119–7125.

Godavarthy, S.S., Robinson Jr., R.L., Gasem, K.A.M., 2006. SVRC-QSPR model for predicting saturated vapor pressures of pure fluids. Fluid Phase Equilib. 246 (1–2), 39–51.

Goodarzi, M., Freitas, M.P., 2008. Feature selection and linear/nonlinear regression. QSAR Comb. Sci. 27, 1092–1098.

Goodarzi, M., Freitas, M.P., 2008. Predicting boiling points of aliphatic alcohols through multivariate image analysis applied to quantitative structure-property relationships. J. Phys. Chem. A 112 (44), 11263–11265.

Goodarzi, M., Freitas, M.P., Jensen, R., 2009. Feature selection and linear/nonlinear regression methods for the accurate prediction of glycogen synthase kinase-3 beta inhibitory activities. J. Chem. Inf. Model. 49 (4), 824–832.

Goodarzi, M., Freitas, M.P., Jensen, R., 2009. Ant colony optimization as a feature selection method in the QSAR modeling of anti-HIV-1 activities of 3-(3,5-dimethylbenzyl)uracil derivatives using MLR, PLS and SVM regressions. Chemom. Intell. Lab. Syst. 98; , pp. 123–129.

Goodarzi, M., Duchowicz, P.R., Wu, C.H., Fernandez, F.M., Castro, E.A., 2009. New hybrid genetic based support vector regression as QSAR approach for analyzing flavonoids-GABA(A) complexes. J. Chem. Inf. Model. 49 (6), 1475–1485.

Goudarzi, N., Goodarzi, M., 2009. Prediction of the vapor pressure of some halogenated methyl-phenyl ether (anisole) compounds using linear and nonlinear QSPR methods. Mol. Phys. 107 (15), 1615–1620.

Hunger, J., Huttner, G., 1999. Optimization and analysis of force field parameters by combination of genetic algorithms and neural networks. J. Comput. Chem. 20 (4), 455–471.

Katritzky, A.R., Slavov, S.H., Dobchev, D.A., Karelson, M., 2007. Rapid QSPR model development technique for prediction of vapor pressure of organic compounds. Comput. Chem. Eng. 31 (9), 1123–1130.

Lawson, D.D., 1980. Methods of calculating engineering parameters for gas separations. Appl. Energy 6 (4), 241–255.

Lazzus, J.A., 2009. Prediction of solid vapor pressures for organic and inorganic compounds using a neural network. Thermochim. Acta 489 (1–2), 53–62.

Mamy, L., Patureau, D., Barriuso, E., Bedos, C., Bessac, F., Louchart, X., Martin-Laurent, F., Miege, C., Benoit, P., 2015. Prediction of the fate of organic compounds in the environment from their molecular properties: a review. Crit. Rev. Environ. Sci. Technol. 45 (12), 1277–1377.

Mauri, A., Consonni, V., Pavan, M., Todeschini, R., 2006. Dragon software: an easy approach to molecular descriptor calculations. Match-Commun. Math. Comput. Chem. 56 (2), 237–248.

Mercader, A.G., Duchowicz, P.R., Fernández, F.M., Castro, E.A., 2010. Replacement method and enhanced replacement method versus the genetic algorithm approach for the selection of molecular descriptors in QSPR/QSAR theories. J. Chem. Inf. Model. 50, 1542–1548.

Mercader, A.G., Duchowicz, P.R., Fernández, F.M., Castro, E.A., 2011. Advances in the replacement and enhanced replacement method in QSAR and QSPR theories. J. Chem. Inf. Model. 51, 1575–1581.

Meulenberg, E.P., Mulder, W.H., Stoks, P.G., 1995. Immuno-assays for Pesticides. Environ. Sci. Technol. 29 (3), 553–561.

Nakajoh, K., Grabda, M., Oleszek-Kudlak, S., Shibata, E., Eckert, F., Nakamura, T., 2009. Prediction of vapour pressures of chlorobenzenes and selected polychlorinated biphenyls using the COSMO-RS model. J. Mol. Struct.: THEOCHEM 895 (1–3), 9–17.

Otieno, P., Owuor, P.O., Lalah, J.O., Pfister, G., Schramm, K.W., 2015. Monitoring the occurrence and distribution of selected organophosphates and carbamate pesticide residues in the ecosystem of Lake Naivasha, Kenya. Toxicol. Environ. Chem. 97 (1), 51–61.

Papadakis, E.N., Vryzas, Z., Kotopoulou, A., Kintzikoglou, K., Makris, K.C., 2015. Papadopoulou-mourkidou E: a pesticide monitoring survey in rivers and lakes of northern Greece and its human and ecotoxicological risk assessment. Ecotoxicol. Environ. Saf. 116, 1–9.

Puzyn, T., Falandysz, J., 2005. Computational estimation of logarithm of n-octanol/air partition coefficient and subcooled vapor pressures of 75 chloronaphthalene congeners. Atmos. Environ. 39 (8), 1439–1446.

Redeker, T., 1997. In Praxis der Sicherheitstechnik. DECHEMA, Freiberg.

Sandler, S.I., Lin, S., Sum, A.K., 2002. Molecular models for the prediction of thermophysical properties of pure fluids and mixtures. Fluid Phase Equilib. 194, 61–75.

Staikova, M., Wania, F., Donaldson, D.J., 2004. Molecular polarizability as a single-parameter predictor of vapour pressures and octanol-air partitioning coefficients of non-polar compounds: a priori approach and results. Atmos. Environ. 38 (2), 213–225.

Stoll, J., 2005. Molecular models for the prediction of thermophysical properties of pure fluids and mixtures. Verfahrenstechni 836, 1–231.

Sun, M., Zheng, Y., Wei, H., Chen, J., Cai, J., Ji, M., 2009. Enhanced replacement method-based quantitative structure-activity relationship modeling and support vector machine classification of 4-anilino-3-quinolinecarbonitriles as Src kinase inhibitors. QSAR Comb. Sci. 28 (3), 312–324.

Talevi, A., Goodarzi, M., Ortiz, E.V., Duchowicz, P.R., Bellera, C.L., Pesce, G., Castro, E. A., Bruno-Blanch, L.E., 2011. Prediction of drug intestinal absorption by new linear and non-linear QSPR. Eur. J. Med. Chem. 46 (1), 218–228.

Todeschini, Roberto, Consonni, Viviana, 2000. Handbook of Molecular Descriptors. Wiley-VCH, Weinheim.

Waller, C.L., Bradley, M.P., 1999. Development and validation of a novel variable selection technique with application to multidimensional quantitative structure-activity relationship studies. J. Chem. Inf. Comput. Sci. 39 (2), 345–355.

Wang, Z.-Y., Zeng, X.-L., Zhai, Z.-C., 2008. Prediction of supercooled liquid vapor pressures and n-octanol/air partition coefficients for polybrominated diphenyl ethers by means of molecular descriptors from DFT method. Sci. Total Environ. 389 (2–3), 296–305.

Wong, K.Y., Mercader, A.G., Saavedra, L.M., Honarparvar, B., Romanelli, G.P., Duchowicz, P.R., 2014. QSAR analysis on tacrine-related acetylcholinesterase inhibitors. J. Biomed. Sci. 21, 84–91.

Zeng, X., Wang, Z., Ge, Z., Liu, H., 2007. Quantitative structure-property relationships for predicting subcooled liquid vapor pressure (P-L) of 209 polychlorinated diphenyl ethers (PCDEs) by DFT and the position of Cl substitution (PCS) methods. Atmos. Environ. 41 (17), 3590–3603.