Pak. J. Chem. 2(1): 6-17, 2012 Full Paper
ISSN (Print): 2220-2625
ISSN (Online): 2222-307X DOI: 10.15228/2012.v02.i01.p02

QSPR study of molar diamagnetic susceptibility of diverse organic compounds using multiple linear regression analysis

*S. Saaidpour, S. A. Zarei and F. Nasri
Department of Chemistry, Sanandaj Branch, Islamic Azad University, Sanandaj, Iran
Email: *saadisaaidpour@gmail.com

ABSTRACT
The multiple linear regression (MLR) was used to build the linear quantitative structure-property relationship (QSPR) model for the prediction of the molar diamagnetic susceptibility (χ_m) for 140 diverse organic compounds using the three significant descriptors calculated from the molecular structures alone and selected by stepwise regression method. Stepwise regression was employed to develop a regression equation based on 100 training compounds, and predictive ability was tested on 40 compounds reserved for that purpose. The stability of the proposed model was validated using Leave-One-Out cross-validation and randomization test. Application of the developed model to a testing set of 40 organic compounds demonstrates that the new model is reliable with good predictive accuracy and simple formulation. By applying MLR method we can predict the test set (40 compounds) with Q^2_{ext} of 0.9894 and average root mean square error (RMSE) of 2.2550. The model applicability domain was always verified by the leverage approach in order to propose reliable predicted data. The prediction results are in good agreement with the experimental values.

Keywords: Molar diamagnetic susceptibility, Molecular descriptors, MLR, Prediction, QSPR

1. INTRODUCTION
Quantitative structure-activity relationships (QSARs) and quantitative structure-property relationships (QSPRs) are scientific fields in which the use of chemometric methods is of outstanding importance. Indeed, chemometric methods, as well as statistics and chemoinformatics, are the basic tools for finding mathematical meaningful relationships between the molecular structure and biological activities, physicochemical, toxicological, and environmental properties of chemicals. The diamagnetic susceptibility (χ) of compounds is an important physicochemical property. If a substance has no permanent magnetic dipole, but has one induced in it by an external field, this induced magnetic field will oppose the applied field. This effect is known as diamagnetism and is a universal property that is shown by most inorganic compounds. It is most perceptible when all electrons are paired, that is, when they have no permanent spin moment. For a diamagnetic substance χ is negative, small, independent of the magnetic field intensity, and independent of temperature. Molecules with a permanent magnetic dipole will behave like small bar magnets; they will align themselves with an applied field, thus reinforcing it. This effect is known as paramagnetism. Salts and certain complexes of transition elements, "odd" electron molecules like NO_2, O_2, and free radicals such as tri-phenyl methyl exhibit this effect, an effect sufficiently large to mask the underlying diamagnetism. For a paramagnetic substance χ is positive, small, independent of the magnetic field intensity, and decreases with increasing temperature. If the permanent magnetic dipoles in a substance are so close together as to interact and support each other, the result is a group or cooperative effect known as ferromagnetism. For a ferromagnetic substance χ is positive, large, and dependent on the magnetic field and temperature, and dependent on previous history. Beyond a certain temperature (the Curie point), magnetism drops and the material shows paramagnetic behavior. For an anti-ferromagnetic substance χ is small and positive, is dependent on previous history, and has complex temperature dependence. Up to a critical temperature, magnetization increases, then decreases past the transition temperature (known as the Néel point) as the material becomes diamagnetic.[1]

When a material is placed in a magnetic field **H**, a magnetization **M** is induced in the material which is related to **H** by $\mathbf{M} = \kappa\mathbf{H}$, where κ is called the volume susceptibility. Since **H** and **M** have the same dimensions, κ is dimensionless. A more useful parameter is the molar susceptibility χ_m, defined by:

$$\chi_m = \kappa V_m = \frac{\kappa M}{\rho} \quad (1)$$

where V_m is the molar volume of the substance, M the molar mass, and ρ the mass density. When the cgs system is used, the customary unit for χ_m is $cm^3\ mol^{-1}$; the corresponding SI unit is $m^3\ mol^{-1}$. Substances with no unpaired electrons are called diamagnetic; they have negative values of χ_m. Their molar susceptibility varies only slightly with temperature. Substances with unpaired electrons, which are termed paramagnetic, have positive χ_m and show much stronger temperature dependence, varying roughly as 1/T.[2]

Quantitative structure–property/activity relationships (QSPR/QSAR) are tools of modeling property/activity as defined by mathematical functions of molecular structure. The QSPR can be used to predict physicochemical properties of organic compounds by using theoretical descriptors. To develop a QSPR, molecular structures are often represented using molecular descriptors which encode much structural information. After the calculation of molecular

*Corresponding Author Received 10th September 2011, Accepted 15th November 2011

descriptors, linear methods, such as multiple linear regression (MLR), principal component regression (PCR) and partial least squares (PLS) or non-linear methods, i.e. neural networks (NN) and support vector machine (SVM) can be used in the development of a mathematical relationship between the structural descriptors and the property. There are many reports about the applications of different modeling approaches to predict the diamagnetic property of inorganic and organic compounds.[3-11]

In our previous papers, we reported on the application of quantitative structure–property/activity relationships (QSPR/QSAR) techniques in the development of a new, simplified approach to prediction of compounds properties using different models.[12-17]

Our goal here is to develop an accurate, simple, fast, and less expensive method for calculation of χ_m values. A stepwise regression (SR) procedure was used for selection of descriptors. Multiple linear regression (MLR) method is utilized to establish quantitative relationships between molar diamagnetic susceptibility and molecular descriptors. Compared with the previous work, the data set used in our investigation is more diverse and the model developed is more general and practical. The predictive power of the resulting model is demonstrated by testing them on unseen data that were not used during model generation. A physicochemical explanation of the selected descriptors is also given.

2. MATERIAL AND METHODS

The methodology applied in our study involved the following five steps: (i) collecting experimental data and splitting the compounds, for which the data was available, into a training set and a validation test set; (ii) molecular geometry optimization,(iii) calculating molecular descriptors for all compounds and selecting the optimal pool of the descriptors to be utilized in the QSPR model development; (iv) training and, simultaneously, internal validating the QSPR model, (v) externally validating the developed model with use of the validation test set.

2.1 *Data set*

All diamagnetic susceptibilities data of the present investigation were obtained from the CRC Handbook of Physics and Chemistry 2010.[2]Diamagnetic susceptibility range was from (-132.2×10^{-6}) to (-30.50×10^{-6}) cm3mol^{-1}. A complete list (140 compounds) of the compound names and corresponding experimental diamagnetic susceptibilities are shown in Table1. The data set was randomly divided into two subsets: a training set of 100 compounds and a validation set of 40 compounds. The training set was used to adjust the parameters of the MLR and the test set was used to evaluate its prediction ability.

Table-1: Molecular Descriptors, Experimental ($-\chi_m \times 10^{-6}$), Predicted ($-\chi_m \times 10^{-6}$), residuals and percent relative error values for training and test sets

NO.	Name	Sv	$1X_v$	MR	χ_m(Exp)	χ_m(Pred)	Residual	%RE
1	Ethylene oxide	3.71	1.08	13.40	30.50	29.87	0.63	-2.08
2	Pyrazine	5.65	1.46	18.30	37.80	40.76	-2.96	7.82
3	Cyclopropane	4.79	1.50	16.70	39.20	38.41	0.79	-2.01
4	Methyloxirane	5.31	1.51	17.40	42.50	39.89	2.61	-6.14
5	Pyrimidine	6.59	1.70	20.20	43.10	46.07	-2.97	6.90
6	Furan	5.71	1.47	18.80	43.10	41.42	1.68	-3.90
7	Furfural	7.22	1.92	21.54	47.20	50.26	-3.06	6.49
8	Succinic anhydride	6.73	1.86	20.33	47.50	47.88	-0.38	0.80
9	Pyridine	7.19	1.85	23.90	48.70	51.83	-3.13	6.43
10	1,4-Cyclohexadiene	7.33	1.97	22.70	48.70	51.98	-3.28	6.74
11	Cyclopentanone	7.71	2.09	22.50	51.60	53.42	-1.82	3.53
12	1,4-Dioxane	7.41	2.16	22.09	52.20	53.22	-1.02	1.95
13	1,2-Epoxybutane	7.27	2.05	22.80	54.80	52.72	2.08	-3.80
14	Benzene	7.79	2.00	26.06	54.80	56.11	-1.31	2.38
15	Morpholine	7.90	2.12	24.00	55.00	55.40	-0.40	0.72
16	Cyclopentane	7.99	2.24	23.01	56.20	55.68	0.52	-0.93
17	Thiophene	7.77	2.20	23.08	57.30	55.08	2.22	-3.88
18	Cyclohexene	8.99	2.36	27.09	58.00	62.07	-4.07	7.01
19	Fluorobenzene	7.90	2.10	26.27	58.40	57.39	1.01	-1.73
20	Phenol	8.31	2.13	27.75	60.60	59.68	0.92	-1.52
21	Furfuryl alcohol	7.82	2.07	26.50	61.00	57.17	3.83	-6.28
22	Cyclohexanone	9.50	2.43	26.80	62.00	63.14	-1.14	1.84
23	Aniline	8.79	2.20	30.00	62.40	63.09	-0.69	1.10
24	Piperidine	8.98	2.51	29.10	64.20	65.36	-1.16	1.81
25	Resorcinol	8.82	2.27	29.45	67.20	63.25	3.95	-5.88
26	o-Nitroaniline	10.21	2.71	32.40	67.40	72.01	-4.61	6.84
27	Styrene	10.39	2.61	33.99	68.20	72.89	-4.69	6.87
28	o-Nitrophenol	9.73	2.64	35.08	68.90	73.31	-4.41	6.40

29	Tetrahydrofurfuryl alcohol	9.01	2.66	30.00	69.40	67.62	1.78	-2.56
30	Chlorobenzene	9.33	2.48	30.86	69.50	67.23	2.27	-3.26
31	m-Nitroaniline	10.21	2.70	33.80	69.70	73.29	-3.59	5.16
32	Methylcyclopentane	9.59	2.70	31.10	70.20	69.87	0.33	-0.47
33	m-Phenylenediamine	9.78	2.40	35.46	70.40	71.52	-1.12	1.59
34	p-Phenylenediamine	9.78	2.40	35.46	70.70	71.52	-0.82	1.16
35	Benzyl alcohol	9.90	2.58	32.87	71.80	70.88	0.92	-1.28
36	Benzamide	10.30	2.65	34.64	72.00	73.79	-1.79	2.49
37	m-Cresol	9.90	2.55	32.79	72.20	70.48	1.72	-2.38
38	Anisole	9.90	2.52	31.17	72.20	68.73	3.47	-4.81
39	p-Cresol	9.90	2.55	32.79	72.40	70.48	1.92	-2.65
40	o-Phenylenediamine	9.78	2.41	35.46	72.50	71.57	0.93	-1.28
41	p-Methylaniline	10.38	2.61	35.80	72.50	74.62	-2.12	2.92
42	2,6-Dimethylpyridine	10.38	2.69	33.70	72.50	73.36	-0.86	1.18
43	m-Nitrotoluene	10.81	2.91	35.20	72.70	77.41	-4.71	6.47
44	o-Cresol	9.90	2.55	32.79	73.30	70.54	2.76	-3.77
45	Cyclohexanol	10.10	2.80	32.40	73.40	72.74	0.66	-0.91
46	Cycloheptane	11.18	2.82	32.21	73.90	74.22	-0.32	0.43
47	N-Methylaniline	10.38	2.66	36.25	74.10	75.52	-1.42	1.92
48	Salicylic acid	10.33	2.73	34.51	75.00	74.42	0.58	-0.78
49	m-Chloroaniline	10.32	2.68	35.56	76.60	74.93	1.67	-2.18
50	Benzeneacetonitrile	10.79	2.84	36.55	76.90	78.03	-1.13	1.47
51	p-Xylene	10.99	2.82	36.14	77.00	77.73	-0.73	0.95
52	Ethylbenzene	10.99	2.97	35.70	77.30	78.69	-1.39	1.80
53	Methylcyclohexane	11.18	3.06	34.30	78.90	78.44	0.46	-0.58
54	o-Chloroaniline	10.87	2.68	35.56	79.50	75.74	3.76	-4.73
55	Isopropenylbenzene	11.62	3.01	38.42	80.00	82.56	-2.56	3.20
56	p-Chlorotoluene	11.23	2.89	35.90	80.30	78.45	1.85	-2.30
57	Methyl benzoate	11.41	2.98	37.59	81.60	81.13	0.47	-0.57
58	p-Dichlorobenzene	10.78	2.96	35.67	81.70	78.23	3.47	-4.25
59	o-Chlorotoluene	11.42	2.89	35.90	82.40	78.77	3.63	-4.40
60	p-Toluic acid	11.41	3.00	37.86	82.40	81.59	0.81	-0.98
61	m-Toluic acid	11.41	3.00	37.86	83.00	81.59	1.41	-1.69
62	Indene	11.39	3.21	38.42	83.00	84.07	-1.07	1.28
63	Phthalic acid	11.84	3.18	39.58	83.60	85.53	-1.93	2.31
64	Isoquinoline	11.79	3.25	40.35	83.90	86.86	-2.96	3.52
65	o-Toluic acid	11.41	3.01	37.86	84.30	81.65	2.65	-3.14
66	Phenetole	11.50	3.11	35.92	84.50	80.90	3.60	-4.26
67	Indole	11.70	2.99	37.15	85.00	81.21	3.79	-4.46
68	N,N-Dimethylaniline	11.98	3.03	40.49	85.10	85.17	-0.07	0.08
69	N-Ethylaniline	11.98	3.22	41.00	85.60	87.43	-1.83	2.14
70	Paraldehyde	11.74	3.15	38.00	86.10	83.54	2.56	-2.97
71	Methyl salicylate	11.93	3.12	39.28	86.60	84.76	1.84	-2.13
72	Propylbenzene	12.59	3.47	40.30	89.10	89.91	-0.81	0.91
73	Isopropylbenzene	12.59	3.35	40.25	89.50	88.78	0.72	-0.80
74	Naphthalene	12.39	3.41	42.51	91.60	91.14	0.46	-0.50
75	1,3,5-Trimethylbenzene	12.59	3.44	41.18	92.30	90.47	1.83	-1.98
76	1-Naphthalenamine	13.38	3.61	43.90	92.50	95.72	-3.22	3.49
77	Benzyl acetate	13.01	3.46	42.03	93.20	92.05	1.15	-1.24
78	Ethyl benzoate	13.01	3.57	42.33	93.80	93.30	0.50	-0.53
79	1-Naphthol	12.90	3.55	44.20	96.20	94.75	1.45	-1.50
80	2-Naphthol	12.90	3.54	44.20	96.80	94.70	2.10	-2.17
81	Safrole	14.01	3.75	43.60	97.50	97.60	-0.10	0.10
82	2-Naphthalenamine	13.38	3.60	45.80	98.00	97.49	0.51	-0.52
83	a-Pinene	14.78	3.92	43.65	100.70	100.27	0.43	-0.42
84	1,2,4,5-Tetramethylbenzene	14.18	3.66	46.22	101.20	99.46	1.74	-1.72
85	Dimethyl terephthalate	15.04	3.95	47.50	101.60	104.63	-3.03	2.98
86	b-Pinene	14.78	4.00	43.65	101.90	101.01	0.89	-0.87
87	2-Methylnaphthalene	13.99	3.82	47.55	102.70	101.95	0.75	-0.73
88	p-Cymene	14.18	3.77	45.29	102.80	99.59	3.21	-3.13
89	Camphor, (+)	15.29	4.08	47.20	103.00	105.85	-2.85	2.77
90	Benzil	14.88	4.19	50.70	106.80	109.65	-2.85	2.67
91	1-Chloronaphthalene	14.66	4.03	49.20	107.60	106.43	1.17	-1.09
92	Acenaphthene	14.99	4.25	50.79	109.90	110.44	-0.54	0.49
93	Acenaphthylene	15.38	4.15	50.27	111.60	109.55	2.05	-1.84

94	9,10-Anthracenedione	16.64	4.45	52.80	113.00	116.48	-3.48	3.08
95	Acridine	16.38	4.54	56.06	118.80	120.07	-1.27	1.07
96	Carbazole	16.56	4.48	54.50	119.90	118.27	1.63	-1.36
97	1-Bromonaphthalene	16.79	4.72	55.10	123.60	121.38	2.22	-1.80
98	Phenanthrene	16.99	4.82	58.96	127.60	126.22	1.38	-1.08
99	1,2-Diphenylethane	18.18	4.87	60.40	127.80	129.74	-1.94	1.52
100	Benzyl benzoate	18.61	5.12	62.20	132.20	134.38	-2.18	1.65

Table-1: (continued)

NO.	Name	Sv	$1X_v$	MR	χ_m(Exp)	χ_m(Pred)	Residual	%RE
1	Maleic anhydride	5.35	1.39	17.40	35.80	38.84	-3.04	8.48
2	Cyclobutane	6.39	1.61	18.40	40.00	43.26	-3.26	8.15
3	Pyrrole	6.19	1.76	20.82	48.60	46.68	1.92	-3.96
4	Pyrrolidine	7.38	2.21	23.50	54.80	55.01	-0.21	0.38
5	4-Methylpyridine	8.79	2.26	28.94	59.80	62.64	-2.84	4.75
6	Cyclopentanol	8.50	2.58	28.00	64.00	64.25	-0.25	0.39
7	p-Hydroquinone	8.82	2.27	29.45	64.70	63.25	1.45	-2.25
8	Benzonitrile	9.19	2.38	31.80	65.20	67.07	-1.86	2.86
9	Toluene	9.39	2.41	31.10	65.60	66.92	-1.32	2.02
10	Salicylaldehyde	9.82	2.58	34.34	66.80	72.13	-5.33	7.98
11	Cyclohexane	9.59	2.71	29.60	68.00	68.53	-0.53	0.77
12	Pyrocatechol	8.82	2.28	29.45	68.20	63.30	4.90	-7.18
13	2,4-Dimethylpyridine	10.38	2.68	33.84	71.30	73.40	-2.10	2.95
14	m-Methylaniline	10.38	2.61	35.80	74.60	74.62	-0.02	0.02
15	o-Methylaniline	10.38	2.62	35.80	74.90	74.67	0.23	-0.30
16	m-Xylene	10.99	2.82	36.14	76.40	77.73	-1.33	1.74
17	p-Chloroaniline	10.43	2.68	35.56	76.70	75.08	1.62	-2.11
18	o-Xylene	10.99	2.83	36.14	77.70	77.79	-0.09	0.11
19	o-Methoxyaniline	10.90	2.73	35.87	79.10	76.49	2.61	-3.30
20	m-Chlorotoluene	11.03	2.89	35.90	79.70	78.18	1.52	-1.91
21	(Chloromethyl)benzene	11.11	3.07	35.93	81.60	79.95	1.65	-2.02
22	Benzeneacetic acid	11.41	3.05	37.37	82.40	81.56	0.84	-1.02
23	2,4,6-Trimethylpyridine	11.98	3.10	38.74	83.10	84.18	-1.08	1.29
24	Terephthalic acid	11.84	3.18	39.58	83.50	85.47	-1.97	2.36
25	m-Dichlorobenzene	11.51	2.96	35.67	84.10	79.23	4.87	-5.79
26	Isophthalic acid	11.84	3.18	39.58	84.60	85.47	-0.87	1.03
27	Cyclooctane	12.45	3.25	36.81	85.30	84.34	0.96	-1.13
28	Quinoline	11.79	3.26	39.98	86.10	86.59	-0.49	0.57
29	p-Bromotoluene	11.88	3.30	38.72	88.70	85.87	2.83	-3.19
30	d-Limonene	13.75	3.81	45.61	98.00	99.71	-1.71	1.75
31	Butylbenzene	14.18	3.97	44.90	100.70	101.12	-0.42	0.41
32	Isobutylbenzene	14.18	3.89	44.85	101.70	100.34	1.36	-1.34
33	tert-Butylbenzene	14.18	3.66	44.72	101.80	98.08	3.72	-3.65
34	1-Methylnaphthalene	13.99	3.82	47.55	102.90	102.00	0.90	-0.87
35	N,N-Diethylaniline	15.18	4.18	49.98	107.90	109.29	-1.39	1.29
36	Diphenylmethane	16.59	4.53	54.40	116.00	118.66	-2.66	2.29
37	Diphenylacetylene	16.99	4.57	54.70	116.00	119.89	-3.89	3.36
38	Hexamethylbenzene	17.38	4.58	56.31	122.50	122.05	0.45	-0.37
39	Diethyl phthalate	18.23	4.91	58.61	127.50	128.47	-0.97	0.76
40	Anthracene	17.75	4.81	58.96	129.80	127.21	2.59	-2.00

2.2 *Descriptor generation*

The molecular structures of all compounds were drawn into the HyperChem7.5 program (Hypercube, Inc., Gainesville, 2002) and pre-optimized using MM+ molecular mechanics method (Polak–Ribiere algorithm). The final geometries of the minimum energy conformation were obtained by more precise optimization with the semi-empirical PM3 method, applying a root mean square gradient limit of 0.01 Kcal/(mol. Å) as a stopping criterion for optimized structures. Then a total of 1195 molecular descriptors were calculated for each polymer by the DRAGON software (Taletesrl, Milan, 2006) on the minimal energy conformations. These descriptors are classified as (a) 0D-constitutional (atom and group counts); (b) 1D-functional groups and atom centered fragments; (c) 2D-topological, BCUTs, walk and path counts, autocorrelations, connectivity indices, information indices, topological charge indices, and eigenvalue-based indices; and (d) 3D-Randic molecular profiles from the geometry matrix, geometrical, WHIM, and GETAWAY descriptors. In order to reduce redundant and non-useful information, constant or near constant values and descriptors found to be highly correlated pair-wise (one of any two descriptors with a correlation greater than

0.99)[18] were excluded in a pre-reduction step, therefore 145 molecular descriptors underwent subsequent variable selection.

2.3 *Stepwise regression for descriptor selection*

After the calculation of molecular descriptors, a stepwise regression routine implemented in SPSS 15.0 software package (SPSS Inc., 2006, Chicago, IL) was used to develop the linear QSPR model using calculated descriptors. The selection of relevant descriptors, which relate the molar diamagnetic susceptibility to the molecular structure, is an important step to construct a predictive model. In order to select the subset of descriptors that best explain compounds χ_m, we have used stepwise regression.[19-21]The stepwise regression was applied to the input set of 145 molecular descriptors for each chemical of the studied data sets and the related response, in order to extract the best set of molecular descriptors, which are, in combination, the most relevant variables in modeling the response of the training set chemicals. Stepwise regression(SR), included in the SPSS software, was used for variables selection (based on the training set). Finally we obtained a three significant descriptor subset, which keeps most interpretive information for χ_m. A total of three descriptors were calculated for each organic in the data set. The selected descriptors are Ghose-Crippen molar refractivity, MR (Steric molecular properties), [22] valence connectivity index chi-1, $^1\chi^v$ (topological descriptors) [23] and sum of atomic vander Waals volumes (scaled on Carbon atom),S_v (constitutional descriptors).[24]

2.4 *Linear modeling*

The general purpose of multiple regressions is to quantify the relationship between several independent or predictor variables and a dependent variable. A set of coefficients defines the single linear combination of independent variables (molecular descriptors) that best describes molar diamagnetic susceptibility. The molar diamagnetic susceptibility value for each compound would then be calculated as a composite of each molecular descriptor weighted by the respective coefficients. A multi-linear model can be represented as:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + ... + b_kx_k \qquad (2)$$

Where k is the number of independent variables (descriptors), $b_1, \ldots, b_k$ are the regression coefficients and y is the dependent variable (χ_m). Regression coefficients represent the independent contributions of each calculated molecular descriptor. The algebraic MLR model is defined in Eq. (2) and in matrix notation:

$$\mathbf{y = Xb + e} \qquad (3)$$

When **X** is of full rank the least squares solution is: $\hat{\mathbf{b}} = (\mathbf{X^T X})^{-1}\mathbf{X^T y}$ where $\hat{\mathbf{b}}$ is the estimator for the regression coefficients in $\hat{\mathbf{b}}$.

A MLR model was developed for organic compounds using The Unscrambler version 9.7 software (CAMO Software AS, 2007; Norway). MLR model was constructed with remaining descriptors based on stepwise feature selection. The MLR model was built using a training set and validation using an external prediction set. MLR techniques based on least-squares procedures are very often used for estimating the coefficients involved in the model equation.[25]

2.5 *Validation of the model*

Model validation is of crucial importance to QSPR modeling. The training and predictive capability of a QSPR model should be tested through model validation.[26-29]

Leave one out cross validation (LOO-CV) is one of the QSPR model internal validation. The predictability of the QSPR model is determined using the LOO-CV method. The cross validated explained variance (Q^2_{LOO}) is calculated by the following equation:

$$Q^2_{LOO} = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i)^2}{\sum_{i=1}^{n}(yi - y)^2} \qquad (4)$$

Where y_i, y_i and y are, respectively, the measured, predicted, and averaged (over the entire training set) values of the dependent variable, respectively; the summations cover all the compounds in the training set. The LOO-CV approach is not sufficient to assess robustness and predictivity. The QSPR model developed using only training set chemicals is then applied to the external validation set chemicals to verify, more reliably, the predictive ability of the model.
The formula for the calculation of Q^2_{ext} is:

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{test}(y_i - y_i)^2}{\sum_{i=1}^{test}(y_i - y_{tr})^2} (5)$$

Where y_i and y_i are respectively the measured and predicted (over the test set) values of the dependent variable, and y_{tr} is the averaged value of the property for the training set; the summations cover all the compounds in the validation set. The Q^2 value is good tests for evenly distributed data, but they are not always reliable for unevenly distributed datasets; instead RMSEs (Root Mean Squared Errors)provide a more reliable indication of the fitness of the model, independently of the applied splitting. Other useful parameter to be considered are the RMSEs calculated on different sets: on training (RMSEV) and prediction (RMSEP). RMSE is calculated as in Eq. (6):

$$\mathrm{RMSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - y_i)^2 \quad ^{0.5} \quad (6)$$

Where y_i and y_i are respectively the measured and predicted values of the property; n is the number of compounds in each set of data.

Another method for validation of the model is randomization testing or Y-scrambling. Randomization testing is a technique for checking the robustness of a QSPR model and the statistical significance of the estimated predicted power. In this test, the dependent variable vector (χ_m), Y-vector, is randomly shuffled and a new QSPR model is developed using the original independent variable matrix. The process is repeated several times. It is expected that the resulting QSPR models will generally have low R^2, low $Q^2_{LOO}$and high RMSE values. If the new models developed from the data set with randomized responses have significantly lower R^2 and Q^2 than the original model, then this is strong evidence that the proposed model is well founded, and not just the result of chance correlation. In contrast, if all the QSPR models obtained in the Y-randomization test have relatively high R^2 and Q^2_{LOO}, then it implies that, for the given data set, the current modeling method is unable to give an acceptable QSPR model.[30,31]

2.6 *Applicability domain of the model*

A crucial problem of a QSPR model is the applicability domain (AD).As even a robust, significant and validated QSPR cannot be expected to reliably predict the modeled property for the entire universe of chemicals, its domain of application must be defined, and the predictions for only those chemicals that fall in this domain can be considered reliable. The chemical domain of applicability is a theoretical region in the space defined by the modeled response and the descriptors of the model, for which a given QSPR should make reliable predictions. This region is defined by the nature of the chemicals in the training set, and can be characterized in various ways. Away of defining the AD of a QSPR model is according to the leverage of a compound. The leverage (h) of a compound measures its influence on the model.[32,33] The leverage of a compound in the original variable space is defined as:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}} \quad (7)$$

Where the**X** is the model matrix derived from the training set descriptor values and the leverage values of training set are diagonal elements of the Hat or Influence matrix **H (**$h_i = \mathrm{diag}\ \mathbf{H}$. The leverage values are always between 0 and 1. The warning leverage h* is defined as follows:

$$\boldsymbol{h^* = 3 \times \frac{\sum_i h_i}{n} = 3 \times \frac{p'}{n} \qquad (i = 1, \ldots, n)} \quad (8)$$

Where n is the number of training set compounds and p' is the number of model parameters plus one.Observations with standardized residuals greater than (-2; +2) range, which lie outside the horizontal reference lines on the plot, are outlier's responses in The Unscrambler 9.7

Standardized residual (SR_i) for each sample is calculated as in Eq. (9):

$$SR_i = \frac{(y_i - \hat{y}_i)}{\sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}} \qquad (9)$$

Where y_i and y_i are respectively the measured and predicted values of the property; n is the number of compounds in each set of data.

In the standardized residuals plot all values are within the (-2; +2) range, which confirms that there are no outliers. Furthermore, there is no clear pattern in the residuals, so nothing seems to be wrong with the model. To visualize the AD of a QSPR model, the plot of standardized residuals versus leverage values (h) (Williams plot) can be used for an immediate and simple graphical detection of both the response outliers and structurally influential chemicals in a model (h>h*).Samples with high leverages have a stronger influence on the model than other samples;

they may or may not be outliers, but they are influential. An influential outlier (high residual + high leverage) is the worst case; it can however easily be detected using an influence plot. Leverages are useful for the detection of samples which are far from the center within the space described by the model. If a sample has a very large leverage, it may be different from the rest and can be considered to be an outlier. Large leverage indicates a high influence on the model.

3. RESULTS AND DISCUSSION

3.1 *MLR analysis*

The software package used for conducting MLR analysis was Unscrambler 9.7. Multiple linear regression (MLR) analysis has been carried out to derive the best QSPR model. The MLR technique was performed on the molecules of the training set shown in Table-1:**.** After regression analysis, a few suitable models were obtained among which the best model was selected and presented in Eq. (10). A small number of molecular descriptors (S_V, $^1\chi^v$and MR) proposed were used to establish a QSPR model. Multiple linear regression analysis provided a useful equation that can be used to predict theχ_m of organic compound based upon these parameters. The best equation obtained for the molar diamagnetic susceptibility of the organic compounds is:

$$\chi_m = 2.00198 + 1.37564 \ \ S_v \ + 9.23318\ [^1\chi^v] + 0.956377 \ \ MR \qquad (10)$$

$$n = 100, R^2 = 0.9892, R^2_{adj} = 0.9888, Q^2_{LOO} = 0.9884\ , s = 2.2705, F = 2915.341$$

Where n is the number of compounds used for regression, R^2 is the squared correlation coefficient,$R^2_{adj}$is the adjusted squared correlation coefficient,Q^2_{LOO} is the cross-validated squared correlation coefficient, s is the standard error of the regression, and F is the Fisher ratio for the regression. The squared correlation coefficient, R^2=0.9892, is a measure of the fit of the regression model. Correspondingly, it represents the part of the variation in the experimental data that is explained by the model. The squared correlation coefficient values closer to 1 represents the better fit of the model. Eq. (10) has an adjusted R^2 value of 0.98880, which indicates very good agreement between the correlation and the variation in the data. The cross-validated squared correlation coefficient Q^2_{LOO}=0.0.988359 illustrates the robustness and stability of the model by focusing on the sensitivity of the model to the elimination of any single data point. The **s** is the standard error measured by the error mean square, which expresses the variation of the residuals or the variation about the regression line. Thus, the standard error measures the model error. In general, the larger the magnitude of the F ratio, the better the model predicts the property values in the training set. The large F ratio of 2915.341 indicates that Eq. (10) does an excellent job of predicting the χ_m values of the training set. The F-test reflects the ratio of the variance explained by the model and the variance due to the error in the model, and high values of the F-test indicate the model is statistically significant. Positive values in the regression coefficients indicate that the indicated descriptor contributes positively to the value of χ_m. In other words, increasing the Sv, $^1\chi^v$andMR will increase absolute value (more negative) χ_m of the organic compounds. The predicted values of χ_m, residuals and the percent relative errors (%RE) of prediction obtained by the MLR method are presented in Table-1:. The plot of predicted χ_m versus experimental χ_m and the residuals (experimental χ_m - predicted χ_m) versus experimental χ_m values, obtained by the MLR modeling, and the random distribution of residuals about zero mean are shown in Figure1**.**

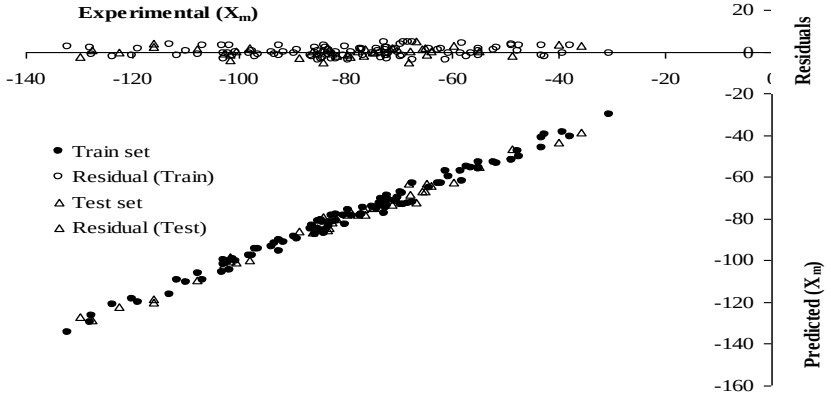

Fig-1: Plot of predicted χ_m and residuals estimated by MLR modeling versus experimental χ_m of training and test sets

The distributions of residuals for the whole dataset are also shown graphically in Figure1**.**The predicted values are in good agreement with the experimental values. The robustness of each model was expressed by the cross-validated (leave-one-out technique, LOO) validation coefficient (Q^2_{LOO}) and the root mean square errors of LOO cross validation (RMSECV). Successfully validated QSPR model with confirmed predictive abilities was used to predict χ_m for all 40 compounds. The internal predictive capability of a model was evaluated by leave-one-out cross-validation(Q^2_{LOO} =

0.9884) on the training set, and the predictive capability of a model on external test set can be expressed by $Q^2_{ext} =$ 0.9894.Moreover, both useful parameters the root means square error (RMSE) and the average percent relative error ($\%RE$) calculated on both the training and test sets were employed to evaluate the performance of the developed model (see Table-2). The statistical results are listed in Table-2.

Table-2: The main statistical parameters of the obtained MLR model

Statistical parameters	Training set	Test set
N	100	40
R^2	0.9892	-
R^2_{adj}	0.9888	-
Q^2_{LOO}	0.9884	-
Q^2_{EXT}	-	0.9894
RMSE	2.2246	2.2550
$\%RE$	0.1217	0.3333

The model was subsequently validated using the response permutation test, also known as Y-scrambling. This procedure involves fitting several models, on the same dependent variables but on a permutated response. It gave the following results: the random models, performed using a scrambled order of the χ_m values, were found to have significantly lower $R^2$and $Q^2_{LOO}$and higher RMSE than the original model (R^2 range: 0.0013 – 0.0239; Q^2_{LOO} range: 0.0003-0.0205; RMSE range: 21.0931-28.4233) corroborating the statistical reliability of the actual model.

To visualize the AD of a QSPR model, the plot of standardized residuals versus leverage values (h) (the Williams plot) can be used for an immediate and simple graphical detection of both the response outliers and structurally influential chemicals in a model. In the Williams plot for AD (see Fig. 2), sample 98 (Phenanthrene) in the training set is to the right of the vertical line, which indicate it has slightly high leverage value (h>h*=0.12) and low standardized residual, it is belong to the model AD. Samples 26 (o-Nitroaniline), 27 (Styrene), 43 (m-Nitrotoluene) in the training set and 10 (Salicylaldehyde), 12 (Pyrocatechol), 25 (m-Dichlorobenzene) in the test set are outliers, indicated by their position above and below the horizontal reference lines, but they have low leverage values. Sample10 (Salicylaldehyde) is wrongly predicted, but in this case it belongs to the AD of the model because in this area there are three compounds belong to the training set. As can be found there is no influential chemical in the test set used in this study.

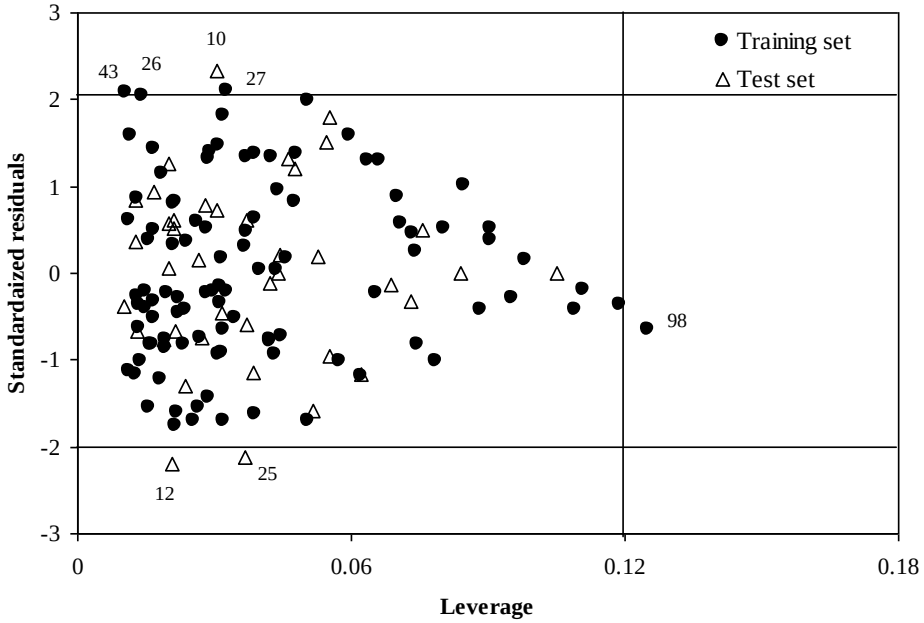

Fig-2: Williams plot for the model with four variables. The χ_m values for the training and test set chemicals are labeled differently, the response outliers and structurally influential chemicals are numbered. The solid lines are, respectively, the $\pm 2\sigma$ limit and the warning value of hat (h* = 0.12).

Chemicals 12 and 25 are wrongly predicted, but in this case they belong to the model AD, being within the cutoff value of Hat (h*).This erroneous prediction could probably be attributed to wrong experimental data rather than to molecular structure.

3.2 *Interpretation of descriptors*

The first selected significant descriptor involved in the Eq. (10) is sum of atomic vander Waals volumes (scaled on carbon atom), S_V. This parameter is a measure of the size of a molecule. The constitutional descriptors depend on

atomic constitution of the chemical structure (molecule). They also include the descriptors related to the types of bonds and the presence of rings in the molecule and etc. The vander Waals radius, r_w, of an atom is the radius of an imaginary hard sphere which can be used to model the atom for many purposes. The vander Waals volume, V_w, also called the atomic volume or molecular volume, is the atomic property most directly related to the vander Waals radius. It is the volume "occupied" by an individual atom (or molecule). The vander Waals volume may be calculated if the vander Waals radii (and, for molecules, the inter-atomic distances and angles) are known. For a spherical single atom, it is the volume of a sphere whose radius is the vander Waals radius of the atom:

$$V_W = \frac{4}{3}\pi r_W^3 (11)$$

For a molecule, it is the volume enclosed by the vander Waals surface. The vander Waals volume of a molecule is always smaller than the sum of the vander Waals volumes of the constituent atoms: the atoms can be said to "overlap" when they form chemical bonds. The vander Waals volume of an atom or molecule may also be determined by experimental measurements on gases, notably from the vander Waals constant (b), the polarizability (α) or the molar refractivity (MR). In all three cases, measurements are made on macroscopic samples and it is normal to express the results as molar quantities. To find the vander Waals volume of a single atom or molecule, it is necessary to divide by the Avogadro constant (N_A).[34]When the size of atomic vander Waals volume a molecule increases the χ_m of that molecule increases. In other words, the molar diamagnetic susceptibility of a molecule increases when its size increases.

The second selected significant descriptor involved in the Eq. (10) is first order valence connectivity ($^1\chi^v$).As the name suggests, topological descriptors consider the topology of a molecule. That is, in the most general case, only the connections between the atoms in a hydrogen suppressed molecule, effectively converting it into a mathematical graph. Certain topological descriptors consider the type or certain properties of atoms involved in the connections as weights. Topological descriptors characterize features such as path lengths and connectivity.
Topological molecular descriptors calculated from the vertex degree of the atoms in the H-depleted molecular graph.

Hall and Kier [23,35] have developed molecular connectivity indices (Chi) that reflect the atom identities, bonding environments and number of bonding hydrogen's. These Kier indices are consequently useful in a wider variety of applications. Molecules that are drawn without hydrogen atoms can be decomposed into fragments of length m, which may be divided into different categories. Hall and Kier defined four series of fragment categories: Path, Cluster, Path/Cluster, and Ring. The spread and numbers of fragment membership for each category is determined by molecule connectivity. Hall and Kier defined groups of Chi (χ) and ChiV (χ^v) indices based on these fragment categories, also incorporating information about the bonding environment. Molecular graph can be denoted by G and having $v_1, v_2, v_3, \ldots, v_n$ as its vertices. The connectivity index $\chi=\chi(G)$ of a graph G is defined by Randic[36]as under:

$$\chi = \chi(G) = \sum_{ij} \frac{1}{\sqrt{\delta_i \delta_j}} \quad (12)$$

Where δi and δj are the valence of a vertex i and j, equal to the number of bonds connected to the atoms i and j, in G. In the case of hetero-systems the connectivity is given in terms of valence delta values δ_i^v and δ_j^v of atoms i and j and is denoted by χ^v. This version of the connectivity index is called the valence connectivity index and is defined [36]as under:

$$\chi^v = \chi^v(G) = \sum_{ij} \frac{1}{\sqrt{\delta_i^v . \delta_j^v}} \quad (13)$$

where the sum is taken over all bonds i–j of the molecule. Valence delta values are given by the following expression:

$$\delta_i^v = \frac{Z_i^v - H_i}{Z_i - Z_i^v - 1} \quad (14)$$

Where Z_i is the atomic number of atom i, Z_i^v is the number of valence electron of the atom i and H_i is the number of hydrogen atoms attached to atom i. Now-a-days, the connectivity and the valence connectivity indices expressed by Eq.(12)and (13) are termed as first-order connectivity and first-order valence connectivity indices, respectively. The molecular connectivity index is a good descriptor of molecular bulk. [37, 38]The compounds with the highest first-order valence connectivity indices have the highest molar diamagnetic susceptibility. The results indicate that the first order valence connectivity increases as χ_m increases. With increasing the number of atoms and the number of valence

electron of the atom in compounds the molecular weight and intermolecular forces increases. Finally with increasing the first valence connectivity index χ_m increases.

The third descriptor in the QSPR model is molar refractivity (MR).The molar refractivity is a constitutive-additive property calculated by the Lorenz-Lorentz formula:

$$MR = \frac{n^2-1}{n^2+2}.\frac{M}{\rho} \quad (15)$$

Where M is the molecular weight, n is the refraction index and ρ the density, and its value depends only of the wavelength of the light used to measure the refraction index.[39]For a radiation of infinite wavelength, the molar refractivity represents the real volume of the molecules and its polarizability. Then, the molar refractivity is related, not only to the volume of the molecules but also to the London dispersive forces that act in the intermolecular interaction. The atomic contribution to molecular refractivity calculated by Ghose and Crippen method. Ghose and Crippen defined 110 atom types, representing most commonly occurring atomic states of carbon, hydrogen, oxygen, nitrogen, halogens, and sulphur in organic molecules to split the molar refractivity.[40,41] They stated that this classification partially differentiates the polarizing effects of heteroatom and the effect of overlapping with non-hydrogen atoms, although they accepted that this classification might be weak in differentiating the conjugation effects. The authors stated that the classification may not completely cover all organic molecules, and that addition of atom types is always feasible. They assumed that the sum of the atomic values (a_i) is the molecular value of the molar refractivity (Eq.(16)):

$$MR_{calc.} = \quad n_i\, a_i \quad (16)$$

The results indicate that the molar refractivity increases as χ_m increases. Finally we see the S_V, $^1X^v$ and MR have the additive atomic relationships. As molar diamagnetic susceptibility is essentially an additive property. According to Eq. (1), the selected significant descriptors are interpretable and meaningful.

4. CONCLUSION

The herein presented QSPR three-parameter model allows the prediction of molar diamagnetic susceptibilities of structurally diverse cyclic and aromatic compounds with average percent relative error of 0.33%. The model is theoretically justified and provides significant additional insight into the relationship between the structure and the molar diamagnetic susceptibilities of the compounds. The aim of this work is the development, using theoretical molecular descriptors, and the proposal of externally validated general QSPR models for the prediction of molar diamagnetic susceptibilities for a wide and heterogeneous set of organic compounds. The great advantage of theoretical descriptors is that they can be calculated homogeneously by defined software for all chemicals, even those not yet synthesized, the only need being a hypothesized chemical structure. The results indicate that the stepwise regression (SR) is a very effective variable selection approach for QSPR analysis. Multiple linear regression (MLR) has been used for structure–property relationship analysis for a set of 140 organic compounds. The results obtained from this study indicate that three descriptors, S_V, $^1\chi^v$and MR play an important role on the molar diamagnetic susceptibility of organic structures. Application of the developed model to a testing set of 40 compounds demonstrates that the new model is reliable with good predictive accuracy and simple formulation. Since the QSPR was developed on the basis of theoretical molecular descriptors calculated exclusively from molecular structure, the proposed model could potentially provide useful information about the χ_m of organic compounds. This procedure allowed us to achieve a precise and relatively fast method for determination of χ_m of different series of organic compounds and to predict with sufficient accuracy the χ_m of new organic derivatives. The macroscopic (bulk) activities/properties of chemical compounds clearly depend on their microscopic (structural) characteristics. Development of quantitative structure property/ activity relationships (QSPR/QSAR) on theoretical descriptors is a powerful tool not only for prediction of the chemical, physical and biological properties/activities of compounds, but also for deeper understanding of the detailed mechanisms of interactions in complex systems that predetermine these properties/activities.

REFERENCE

1. Dean, J. A., Analytical chemistry handbook, McGraw-Hill, New York, **(1995)**.
2. Lide, D. R., CRC Handbook of Chemistry and Physics, 90th Edition, CRC Press/Taylorand Francis, Boca Raton, FL, (**2010**).
3. Afantitis, A., Melagraki, G., Sarimveis, H., Koutentis, P. A., Markopoulos, J., Markopoulou, O. I., Development and Evaluation of a QSPR Model for the Prediction of Diamagnetic Susceptibility. QSAR Comb. Sci. **(2008)** 27:432-436, http://dx.doi.org/10.1002/qsar.200730083.
4. Estrada, E., Modelling the Diamagnetic Susceptibility of Organic Compounds by a Sub-Structural Graph-Theoretical Approach. J. Chem. Soc. Faraday Trans. **(1998)** 94:1407-140, http://dx.doi.org/10.1039/a709032c.

5. Estrada, E., Gutierrez, Y., Gonzalez, H., Modeling Diamagnetic and Magnetooptic Properties of Organic Compounds with the TOSS-MODE Approach. J. Chem. Inf.Compute. Sci. **(2000)**, 40: 1386-1399, http://dx.doi.org/10.1021/ci000041e.
6. Zhokhova, N. I., Baskin, I. I., Palyulin, V., Zefirov, N., Zefirov, N. S., Fragment descriptors in qspr:application to magnetic susceptibility calculations. J. Struct. Chem. **(2004)**, 45: 626-635, http://dx.doi.org/10.1007/s10947-005-0037-2.
7. Mu, L., He, H. M., Feng, C. J., Quantitative Structure Property Relations (QSPR) for Predicting Molar Diamagnetic Susceptibilities, χ_m, of Inorganic Compounds. Chin. J. Chem. **(2007)**, 25: 743-750, http://dx.doi.org/10.1002/cjoc.200790138.
8. Mu, L., Feng, C. J., He, H. M., Modeling diamagnetic susceptibilities of organic compounds with a novel connectivity index. Ind. Eng. Chem. Res. **(2008)**, 47: 2428-2433, http://dx.doi.org/10.1021/ie071232a.
9. Mu, L., Feng, C. J., He, H. M., Topological research on diamagnetic susceptibilities of organic compounds. J. Mol. Model. **(2008)**, 14: 109-134, http://dx.doi.org/10.1007/s00894-007-0256-x.
10. Mu, L., He, H., Yang, W., Improved QSPR study of diamagnetic susceptibilities for organic compounds using two novel molecular connectivity indexes. Chin. J. Chem. **(2009)**, 27: 1045-1054, http://dx.doi.org/10.1002/cjoc.200990175.
11. Mu, L., He, H., Yang, W., Feng, C., Variable molecular connectivity indices for predicting the diamagnetic susceptibilities of organic compounds. Ind. Eng. Chem. Res. **(2009)**, 48: 4165-4175, http://dx.doi.org/10.1021/ie801252j.
12. Ghasemi, J., Saaidpour, S., Quantitative structure-property relationship study of n-octanol-water partition coefficients of some of diverse drugs using multiple linear regression. Anal. Chim. Acta **(2007)**, 604:99-106, http://dx.doi.org/10.1016/j.aca.2007.10.004.
13. Ghasemi, J., Saaidpour, S., QSPR prediction of aqueous solubility of drug-like organic compounds.Chem. Pharm. Bull. **(2007)**, 55:669-674, http://dx.doi.org/10.1248/cpb.55.669.
14. Ghasemi, J., Saaidpour S., Brown, S. D., QSPR study for estimation of acidity constants of some aromatic acids derivatives using multiple linear regression (MLR) analysis. J. Mol. Struct. (Theochem) **(2007)**, 805: 27-32, http://dx.doi.org/10.1016/j.theochem.2006.09.026.
15. Ghasemi, J., Saaidpour, S., QSPR modeling of stability constants of diverse 15-crown-5 ethers complexes using best multiple linear regression. J. Incl. Phenom. Macrocycl. Chem. **(2008)**, 60: 339-351, http://dx.doi.org/10.1007/s10847-007-9383-3.
16. Ghasemi, J., Saaidpour, S., Artificial Neural Network Based Quantitative Structural Property Relationship for Predicting Boiling Points of Refrigerants.QSAR Comb. Sci. **(2009)**, 28: 1245-1254, http://dx.doi.org/10.1002/qsar.200810101.
17. Ghasemi, J., Saaidpour, S., QSRR prediction of the chromatographic retention behavior of painkiller drugs. J. Chromatogr. Sci. **(2009)**, 47:156-163, http://dx.doi.org/10.1093/chromsci/47.2.156.
18. Liu, H., Gramatica, P., QSAR study of selective ligands for the thyroid hormone receptor beta. Bioorgan. Med. Chem. **(2007)**, 15: 5251-5261, http://dx.doi.org/10.1016/j.bmc.2007.05.016.
19. Massart, D. L., Vandeginste, B. G. M., Buydens, L. M. C., Jong, S. D., Lewi, P. J., Verbeke, J. S., Handbook of Chemometrics and Qualimetrics, Part A. Elseiver, Amsterdam, **(1997)**.
20. Darlington, R. B., Regression and Linear Models.McGraw-Hill Higher Education, New York **(1990)**.
21. Xu, L., Zhang, W. J., Comparison of different methods for variable selection, Anal. Chim. Acta **(2001)**, 446: 475-481, http://dx.doi.org/10.1016/S0003-2670(01)01271-5.
22. Viswanadhan, V. N., Ghose, A. K., Revankar, G. R., Robins, R. K., Ghose-Crippen molar refractivity. J. Chem. Inf. Comput. Sci. **(1989)**, 29:163-172, http://dx.doi.org/10.1021/ci00063a006.
23. Kier, L. B., Hall, L. H., Molecular connectivity in Structure-Activity Analysis. RSP-Wiley, Chichetser, **(1986)**.
24. Todeschini, R., Consonni, V., Handbook of Molecular Descriptors. Wiley-VCH, Weinheim, Germany, **(2000)**, http://dx.doi.org/10.1002/9783527613106.
25. Gemperline, P., Practical guide to chemometrics. Taylor & Francis Group, LLC, Boca Raton, **(2006)**, http://dx.doi.org/10.1201/9781420018301.
26. Golbraikh, A., Tropsha, A., Beware of Q^2.J. Mol. Graph. Model**. (2002)**, 20: 269-276, http://dx.doi.org/10.1016/S1093-3263(01)00123-1.
27. Gramatica, P., Pilutti, P., Papa, E., Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting into training-test sets and consensus modeling. J. Chem. Inf. Comput. Sci. **(2004)**, 44: 1794-1802, http://dx.doi.org/10.1021/ci049923u.
28. Gramatica, P., Principles of QSAR models validation: internal and external. QSAR Comb. Sci. **(2007)**, 26: 694-701, http://dx.doi.org/10.1002/qsar.200610151.
29. Gramatica, P., Giani, E., Papa, E., Statistical external validation and consensus modeling: a QSPR case study for K_{OC} prediction. J. Mol. Graph. Model. **(2007)**, 25:755-766, http://dx.doi.org/10.1016/j.jmgm.2006.06.005.

30. Tropsha, A., Gramatica, P., Gombar, V. K., The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, QSAR Comb. Sci. **(2003)**, 22: 69-76, http://dx.doi.org/10.1002/qsar.200390007.
31. Eriksson, L., Jaworska, J., Worth, A., Cronin, M., McDowell, R. M., Gramatica, P., Methods for Reliability, Uncertainty Assessment, and Applicability Evaluations of Regression Based and Classification QSARs. Environ. Health Perspect. **(2003)**, 111:1361-1375, http://dx.doi.org/10.1289/ehp.5758.
32. Atkinson. A. C., Plots, Transformations and Regression. Clarendon Press, Oxford, **(1985)**.
33. Shacham, M., Brauner, N., Cholakov, G. S., Stateva, R. P., Identifying applicability domains for quantitative structure property relationships, in Computer Aided Chemical Engineering, eds. P. Valentin and A. Paul Serban, Elsevier, (**2007**).
34. Bondi, A.,Vanderwaals volumes and radii. J. Phys. Chem. **(1964)**, 68: 441-451, http://dx.doi.org/10.1021/j100785a001.
35. Hall, L. H., Kier, L. B., The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling. VCH Publishers, Inc., **(1991)**.
36. Randic, M., Characterization of Molecular Branching. J. Am. Chem. Soc. **(1975)**, 97: 6609-6615, http://dx.doi.org/10.1021/ja00856a001.
37. Kier, L. B., Hall, L. H., Murray, W., Molecular Connectivity I: Relationship to local anasthesia, J. Pharm. Sci.**(1975)**, 64: 1971-1974, http://dx.doi.org/10.1002/jps.2600641214.
38. Kier, L. B., Hall, L. H., Molecular Connectivity VII: Specific Treatment to Heteroatom's. J. Pharm. Sci. **(1976)**, 65:1806-1809, http://dx.doi.org/10.1002/jps.2600651228.
39. Livingstone, D. J., The characterization of chemical structures using molecular properties - a survey. J. Chem. Inform. Comput. Sci. **(2000)**, 40: 195-209, http://dx.doi.org/10.1021/ci990162i.
40. Ghose, A. K., Crippen, G. M., Physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships. i. partition coefficients as a measure of hydrophobicity. J. Comput. Chem. **(1986)**, 7: 565-577, http://dx.doi.org/10.1002/jcc.540070419.
41. Ghose, A. K., Crippen, G. M., Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. J. Comput. Chem. **(1987)**, 27: 21-35.