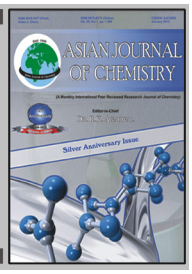
Asian Journal of Chemistry; Vol. 25, No. 1 (2013), 331-342

ASIAN JOURNAL OF CHEMISTRY

http://dx.doi.org/10.14233/ajchem.2013.13032



Chemometric Modeling to Predict Aquatic Toxicity of Benzene Derivatives Using Stepwise-Multi Linear Regression and Partial Least Square

MARYAM BORDBAR[1,*], JAHANBAKHSH GHASEMI[2], ALI YEGANEH FAAL[3] and RAZIEH FAZAELI[4]

[1]Department of Chemistry, Faculty of Science, University of Qom, P.O. Box 37185-359, Qom, Iran
[2]Department of Chemistry, K.N. Toosi University of Technology, P.O. Box 16167-45481, Tehran, Iran
[3]Department of Chemistry, Payame Noor University, P.O. Box 19395-3697, Iran
[4]Department of Chemistry, Shahreza Branch, Islamic Azad University, P.O. Box 86145-311, Shreza, Iran

*Corresponding author: Fax: +98 251 2916449; Tel: +98 251 2853311; E-mail: m.bordbar@qom.ac.ir

(*Received*: 7 November 2011; *Accepted*: 23 July 2012) AJC-11874

The aquatic toxicity of 392 benzene derivatives have been subjected to quantitative structure-activity relationship studies. Optimization of 3D structures of the molecules carried out by HyperChem using AM1 model. The molecular descriptors; constitutional, topological, molecular walk counts, aromaticity indices, geometrical, WHIM, functional group, empirical and properties were obtained by Dragon software. The models were constructed using 309 molecules as training set and predictive ability tested using 78 compounds. Modeling of log ($1/IGC_{50}$) of these compounds as a function of the theoretically derived descriptors was established by multiple linear regression (MLR) technique. This linear modeling method indicates the importance of different topological and electronic descriptors on the aquatic toxicity [log ($1/IGC_{50}$)]. The obtained model (stepwise MLR-PLS) was chosen based on highest external predictive R^2 value (0.81) and lowest RMSEP (2.41) values. It is observed that the Moriguchi octanol/water partition coefficient (log P) descriptor has great effect on the aquatic toxicity, which confirms its importance in mechanism of aquatic toxicity action of benzene derivatives.

Key Words: Quantitative structure-activity relationship, Multiple linear regression, Partial least square, Aquatic toxicity, Benzene derivatives.

INTRODUCTION

Structure-toxicity models exist at the intersection of biology, chemistry and statistics. The connection of these three subjects has permitted the development of structure-activity relationships as an accepted sub-discipline in toxicology. The next decade will see an increased use of (quantitative) structure-activity relationships (QSARs) to predict toxicity for new and existing chemicals. Much of the focus will be on their application to reduce or replace animal use in toxicological testing for the regulation of existing chemicals (*e.g.* in the REACH legislation)[1]. The official birth date of QSAR is considered to be 1962, when Hansch *et al.*[2] published a paper which showed a correlation between biological activity and octanol-water partition coefficient. Quantitative structure-activity relationship models have another ability which is obtaining a deeper knowledge about the mechanism of biological activity. Quantitative structure-activity relationships represent predictive models derived from application of statistical tools correlating biological activity (including therapeutic and toxic) of chemicals (drugs/toxicants/environmental pollutants) with descriptors representative of molecular structure and/or property. The concept that there exists a close relationship between bulk properties of similar compounds and their molecular structure allows one to provide a clear connection between the macroscopic and the microscopic properties of matter. Quantitative structure-property relationships are mathematical equations relating chemical structure to a wide variety of physical, chemical, biological and technological properties.

Nowadays, a wide range of descriptors has been used in QSAR modeling[3]. These descriptors have been classified into different categories according to Karelson approach including constitutional, geometrical, topological and quantum chemical, *etc.*[4].

The success of any QSAR model depends on the accuracy of input data, selection of appropriate descriptors that represent variations in structural property of molecules quantitatively and statistical tools and validation of the developed model[5]. The validation strategies check the reliability of the developed models for their possible application on a new set of data and confidence of prediction can thus be judged. For validation of QSAR models usually four strategies are adopted[6]: (a) Internal validation or cross-validation; (b) Validation by dividing the

data set into training and test compounds; (c) True external validation by application of model on external data and (d) Data randomization or Y-scrambling. As a result, a simple mathematical relationship is established:

Property = f (Structural descriptors)

Quantitative structure-activity relationship techniques include from chemical measurements and biological assays to the statistical techniques and interpretation of results[7].

In this work a QSAR study is performed; to develop model that relate the structures of 392 substituted benzenes to toxic action. The stepwise multiple linear regression were used to select the most informative descriptors from the calculated descriptors by Dragon (version 2.1) software[3]. The selected descriptors were used to develop a model for predicting the log (IGC_{50}) (decimal logarithm of the inverse 50 % growth inhibitory concentration) values. We have validated the models by dividing the data set into training (307 compounds) and test set (78 compounds) by K-means clustering technique. Different statistical techniques were used to develop the model to highlight the structural requirements for an ideal aquatic toxicity inhibitor. The two objectives of the present paper have been: (1) To explore the structure-activity relationships of aquatic toxicity of diverse compounds and (2) To select the best predictive model from among all comparable chemometric models for the aquatic toxicity.

EXPERIMENTAL

The QSAR model for the estimation of the log (IGC_{50}) of various substituted benzenes compounds is established in the following five steps: (i) the molecular structure input and generation of the files containing the chemical structures is stored in a computer-readable format; (ii) quantum mechanics geometry is optimized with a semi-empirical (AM1) method; (iii) structural descriptors are computed; (iv) structural descriptors are selected by stepwise multiple linear regression; (v) the structure-log (IGC_{50}) model is generated by partial least square calibration method and statistical analysis.

Chemical dataset selection: Central to the issues of quality, transparency and domain identification as they relate to toxicological QSAR is biological data. High-quality toxicity data, in a structurally diverse set of molecules are required to formulate and validate high-quality QSARs. Quality toxicity data typically come from standardized assays measured in a consistent manner, with a clear and unambiguous end point and lower experimental error[8]. Toxicity assessments which are made in a single laboratory by a single protocol tend to be the most precise. By taking these points into consideration, we select the database of inhibition of growth of the ciliated protozoan *T. pyriformis*. This database has been developed in a single laboratory over more than two decades and it has been recognized as a high-quality dataset[9].

The general dataset used in this study has been recently published by other researchers[8]. It consists of almost 400 substituted benzenes representing several mechanisms of toxic action (Table-1). Some compounds were reported by Schultz and Netzeva as not toxic at saturation. Hence these compounds were not used in the present work. A horizontal validation was performed using a training set, composed by 307 benzene derivatives, for models development and a validation set (78 compounds) to assess the predictive capability of the QSAR models. In order to split the database into training and prediction series, a k-means cluster analyses (k-MCA) was carried out for entire dataset to design, in a rational representative way, the training (training) and prediction (test) series[10].

Computer hardware and software: All calculations were run on a Pentium IV personal computer with windows XP operating system. The ChemDrawUltra version 9.0 (Chem Office 2005, CambridgeSoft Corporation) software was used for drawing the molecular structures. The optimizations of molecular structures were done by the HyperChem 7.5 (AM1 method) and descriptors were calculated by Dragon (Milano Chemometrics group, version 2.1) software. Stepwise multiple linear regression regression was performed by using SPSS version 11.5 software and partial least square calculations were performed in the MATLAB (version 7.5, MathWorks Inc.).

Molecular modeling and theoretical molecular descriptors: Molecular descriptors define the molecular structure and physicochemical properties of molecules by a single number. A wide variety of descriptors have been reported for using in QSAR analysis[3,11]. The structures were drawn in Chem Draw Ultra version 9.0 and exported in a file format suitable for HyperChem 7.5. The geometry optimization was performed with the semi-empirical quantum method Austin Method 1 (AM1)[12] incorporated in the HyperChem program. The gradient norm criterion 0.01 kcal/Aº was applied in the geometry optimization for all structures. The HyperChem mol files were used by the Dragon program to compute more than 1027 structural descriptors for the 392 benzene derivatives. Dragon computes 10 classes of structural descriptors: constitutional (number of various types of atoms and bonds, number of rings, molecular weight, *etc*.); topological (Wiener index, Randic indices, Kier-Hall shape indices, Balaban index, *etc*.); geometrical (moments of inertia, molecular volume, molecular surface area, *etc*.); electrostatic (minimum and maximum partial charges, polarity parameter, charged partial surface area descriptors, *etc*.); Molecular walk counts (molecular walk counts of order 1-10, total walk count, self-returning of order 1-10); aromaticity indices (harmonic oscillator model of aromaticity index, Jug RC index, aromaticity, HOMA total); WHIM descriptors (unweighted size, shape, symmetry and accessibility directional indices; size, shape, symmetry and accessibility directional indices weighted by atomic polarizability, atomic Sanderson electronegativity or atomic van der Waals volume; total size, shape symmetry and accessibility indices); functional group (numbers of different types of carbons, number of allenes groups, number of esters (aliphatic or aromatic), number of amides, number of different functional groups, number of CH3R, number of CR4, number of different halogens attached to different type of carbons, number of PX3, number of PR3); empirical (Unsaturation index, ydrophilic factor, aromatic ratio) and properties descriptors (molar refractivity, polar surface area, logP) were generated for each compound.

The generation of the descriptors is carried out without taking into account of the solvation of the molecules. It means that the generated descriptors are carried out using the

TABLE-1
EXPERIMENTAL AND PREDICTED VALUES [log ($1/IGC_{50}$)] FOR THE TRAINING AND TEST SET BY STEPWISE MLR-PLS MODEL

Compounds		CAS	log ($1/IGC_{50}$)	Training set	Test set
001	Benzene	71-43-2	-0.12	-0.09	
002	*p*-Xylene	106-42-3	0.25	0.37	
003	1-Phenyl-2-butanol	120055-09-6	-0.16	0.14	
004	Toluene	108-88-3	0.25	-0.04	
005	*n*-Butylbenzene	104-51-8	1.25	0.82	
006	*n*-Amylbenzene	538-68-1	1.79	1.19	
007	Benzylamine	100-46-9	-0.24	-0.68	
008	Isopropylbenzene	98-82-8	0.69	0.45	
009	6-Phenyl-1-hexanol	2430-16-2	0.87	0.95	
010	5-Phenyl-1-pentanol	10521-91-2	0.42		0.64
011	α,α-Ethylbenzenepropanol	103-05-9	-0.07	-0.09	
012	4-Phenyl-1-butanol	3360-41-6	0.12	0.23	
013	3-Phenyl-1-propanol	122-97-4	-0.21	-0.11	
014	Benzyl alcohol	100-51-6	-0.83		-0.41
015	*sec*-Phenethyl alcohol	98-85-1	-0.66	-0.46	
016	4-Ethylbenzyl alcohol	768-59-2	0.07	0.10	
017	3-Phenyl-1-butanol	2722-36-3	0.01		0.14
018	(R)-1-phenyl-1-butanol	22144-60-1	-0.01	0.08	
019	4-Biphenylmethanol	3597-91-9	0.92		0.66
020	4-Ethylbiphenyla	5707-44-8	1.97	1.44	
021	Biphenyl	92-52-4	1.05	1.11	
022	(±)-2-Phenyl-2-butanol	1565-75-9	0.06	-0.31	
023	(±)-1,2-Diphenyl-2-propanol	5342-87-0	0.80	0.62	
024	1,1-Diphenyl-2-propanol	29338-49-6	0.75		0.66
025	3,4-Dimethylaniline	95-64-7	-0.16	0.19	
026	3-Aminobenzyl alcohol	1877-77-6	-1.13	-0.03	
027	4-Butoxyaniline	4344-55-2	0.61	0.67	
028	4-Pentyloxyaniline	39905-50-5	0.97		0.99
029	4-Hexyloxyaniline	39905-57-2	1.38	1.29	
030	4-Methylaniline	106-49-0	-0.05	-0.13	
031	4-Isopropylaniline	99-88-7	0.22	0.12	
032	3-Ethylaniline	587-02-0	-0.03	0.00	
033	4-Ethylaniline	589-16-2	0.03	-0.07	
034	3-Methylaniline	108-44-1	0.28	-0.18	
035	4-Butylaniline	104-13-2	1.07	0.60	
036	(2-Bromoethyl)benzene	103-63-9	0.42	0.01	
037	2-Methylaniline	95-53-4	-0.16		-0.21
038	2,6-Diisopropylaniline	24544-04-5	0.76	0.99	
039	Aniline	62-53-3	-0.23		-0.24
040	2-Ethylaniline	578-54-1	-0.22	-0.16	
041	2,6-Diethylaniline	579-66-8	0.31	0.45	
042	Thioanisole	100-68-5	0.18	0.13	
043	4-Methoxyphenol	150-76-5	-0.14		-0.18
044	3,4,5-Trimethylphenol	527-54-8	0.93		0.77
045	Benzyl chloride	100-44-7	0.06	0.07	
046	4-Methylanisole	104-93-8	0.25	0.14	
047	2,3,5-Trimethylphenol	697-82-5	0.36	0.68	
048	2,4,6-Trimethylphenol	527-60-6	0.42	0.56	
049	4-*tert*-Butylpheno	98-54-4	0.91	0.43	
050	4-*tert*-Pentylphenol	80-46-6	1.23		0.63
051	2,3,6-Trimethylphenol	2416-94-6	0.28	0.54	
052	Phenetole	103-73-1	-0.14	0.14	
053	Anisole	100-66-3	-0.10		-0.21
054	2,4-Dimethylphenol	105-67-9	0.14	0.34	
055	2-Phenyl-3-butyn-2-ol	127-66-2	-0.18	-0.24	
056	*p*-Cresol	106-44-5	-0.16	0.00	
057	4-Ethylphenol	123-07-9	0.21	0.09	
058	4-Propylphenol	645-56-7	0.64	0.40	
059	3-Ethylphenol	620-17-7	0.29	0.13	
060	Nonylphenol	104-40-5	2.47	2.59	
061	*m*-Cresol	108-39-4	-0.08	-0.04	
062	*o*-Cresol	95-48-7	-0.29	-0.06	
063	2-Ethylphenol	90-00-6	0.16	0.00	
064	Phenol	108-95-2	-0.35	0.03	

Compounds		CAS	log (1/IGC$_{50}$)	Training set	Test set
065	2-Allylphenol	1745-81-9	0.33		0.11
066	Iodobenzene	591-50-4	0.36	0.43	
067	4-Chloroaniline	106-47-8	0.05	-0.14	
068	2-Tolunitrile	529-19-1	-0.24	-0.06	
069	4-Hydroxyphenethyl alcohol	501-94-0	-0.83	-0.30	
070	2-Chloro-4-methylaniline	615-65-6	0.18		0.17
071	2-Chloroaniline	95-51-2	-0.17	-0.21	
072	5-Pentylresorcinol	500-66-3	1.31	1.59	
073	3-Methoxyphenol	150-19-6	-0.33	-0.14	
074	4-Hexylresorcinola	136-77-6	1.80		1.66
075	4-Chloro-3,5-methylphenol	88-04-0	1.20	0.84	
076	4-Bromotoluene	106-38-7	0.47	0.59	
077	1-Bromo-4-ethylbenzene	1585-07-5	0.67	0.72	
078	4-Chloroanisole	623-12-1	0.6	0.41	
079	4-Chloro-3-methylphenol	59-50-7	0.8	0.53	
080	1,3-Dihydroxybenzene	108-46-3	-0.65	0.02	
081	Bromobenzene	108-86-1	0.08	0.44	
082	4-Chlorophenol	106-48-9	0.54	0.48	
083	4-Iodophenol	540-38-5	0.85	0.00	
084	2-(4-Chlorophenyl)ethylamine	156-41-2	0.14	-0.13	
085	4-Chlorobenzylamine	104-86-9	0.16	-0.28	
086	2,4-Dichloroaniline	554-00-7	0.56	0.18	
087	Chlorobenzene	108-90-7	-0.13	0.44	
088	3-Chloroaniline	108-42-9	0.22	-0.18	
089	1,2-Dimethyl-4-nitrobenzene	99-51-4	0.59		1.05
090	4-(Pentyloxy)benzaldehyde	5736-91-4	1.18	1.24	
091	4-Nitrotoluene	99-99-0	0.65	0.70	
092	4-Isopropylbenzaldehyde	122-03-2	0.67		0.59
093	1,2-Dimethyl-3-nitrobenzene	83-41-0	0.56	0.69	
094	3-Chlorophenol	108-43-0	0.87	0.40	
095	3-Nitrotoluene	99-08-1	0.42	0.67	
096	2-Nitrotoluene	88-72-2	0.26		0.46
097	1,4-Dibromobenzene	106-37-6	0.68	1.01	
098	Benzaldehyde	100-52-7	-0.2	-0.07	
099	3-Ethoxy-4-hydroxybenzaldehyde	121-32-4	0.02		0.41
100	3-Methoxy-4-hydroxybenzaldehyde	121-33-5	-0.03	-0.05	
101	4-Hydroxypropiophenone	70-70-2	0.12	0.12	
102	2,4-Dichlorophenol	120-83-2	1.04	1.03	
103	Valerophenone	1009-14-9	0.56	0.67	
104	Propiophenone	93-55-0	-0.07		-0.03
105	Butyrophenone	495-40-9	0.21	0.31	
106	2-Hydroxybenzaldehyde	90-02-8	0.42	0.00	
107	Heptanophenone	1671-75-6	1.56	1.52	
108	Acetophenone	98-86-2	-0.46	-0.17	
109	Nitrobenzene	98-95-3	0.14	0.29	
110	Octanophenone	1674-37-9	1.89	1.89	
111	2,5-Dichloroaniline	95-82-9	0.58		0.17
112	3,4-Dichlorotoluene	95-75-0	1.07	1.09	
113	3-Nitroaniline	99-09-2	0.03	0.2	
114	3,5-Dichloroaniline	626-43-7	0.71	0.17	
115	4-Bromo-6-chloro-o-cresol	7530-27-0	1.28		1.21
116	1,2-Dichlorobenzene	95-50-1	0.53		1.05
117	3-Nitroanisole	555-03-3	0.72	0.57	
118	Benzophenone	119-61-9	0.87	0.82	
119	3-Chloro-5-methoxyphenol	65262-96-6	0.76	0.43	
120	4-Nitrobenzyl chloride	100-14-1	1.18	0.81	
121	2,4-Dibromophenol	615-58-7	1.40		1.30
122	2-Amino-5-clorobenzonitrile	5922-60-1	0.44	0.33	
123	2-Hydroxy-4-acetophenone	552-41-0	0.55	0.03	
124	3,5-Dichlorophenol	591-35-5	1.56	0.90	
125	4-Chlorobenzaldehyde	104-88-1	0.40	0.39	
126	4-Chlorobenzophenone	134-85-0	1.5	1.38	
127	1,3,5-Trichlorobenzene	108-70-3	0.87	1.34	
128	2,4,5-Trichloroaniline	636-30-6	1.30	0.58	
129	4-Bromobenzophenone	90-90-4	1.26	1.47	
130	1,2,4-Trichlorobenzene	120-82-1	1.08	1.55	

Compounds		CAS	log (1/IGC_{50})	Training set	Test set
131	2,4,6-Trichlorophenol	88-06-2	1.41	1.37	
132	4-Ethoxy-2-nitroaniline	616-86-4	0.76	1.03	
133	5-Bromovanillin	2973-76-4	0.62	0.48	
134	4-Nitrophenetole	100-29-8	0.83	0.84	
135	4-Chloro-2-nitrotoluene	89-59-8	0.82	1.15	
136	1-Bromo-3-nitrobenzene	585-79-5	1.03	0.96	
137	4-Bromo-2,6-dichlorophenol	3217-15-0	1.78	1.80	
138	2-Chloro-6-nitrotoluene	83-42-1	0.68	1.00	
139	2,3,5,6-Tetrachloroaniline	3481-20-7	1.76		0.95
140	3-Nitrobenzonitrile	619-24-9	0.45	0.38	
141	2,4,5-Trichlorophenol	95-95-4	2.10	1.60	
142	1,2,4,5-Tetrachlorobenzene	95-94-3	2.00	2.05	
143	4-Methyl-2-nitroaniline	89-62-3	0.37	0.42	
144	1-Chloro-3-nitrobenzene	121-73-3	0.73	0.89	
145	2-Nitroaniline	88-74-4	0.08	-0.03	
146	2,3,4,5-Tetrachloroaniline	634-83-3	1.96	1.04	
147	2,4,6-Tribromophenol	118-79-6	1.91	2.03	
148	2-Bromo-5-nitrotoluene	7149-70-4	1.16	1.50	
149	1-Fluoro-3-iodo-5-nitrobenzene	3819-88-3	1.09		1.09
150	2-Nitrophenol	88-75-5	0.67	0.20	
151	2-Chloro-4-nitroaniline	121-87-9	0.75	0.74	
152	5-Hydroxy-2-nitrobenzaldehyde	42454-06-8	0.33	0.28	
153	3,4,5,6-Tetrabromo-*o*-cresol	576-55-6	2.57	*	
154	2,3,4,6-Tetrachlorophenol	58-90-2	2.18	1.91	
155	1-Fluoro-4-nitrobenzene	350-46-9	0.1	0.71	
156	Pentafluoro aniline	771-60-8	0.26	0.66	
157	1-Bromo-2-nitrobenzene	577-19-5	0.75	0.60	
158	3,5-Dibromo-salicylaldehyde	90-59-5	1.65	1.29	
159	3,5-Dichloro-nitrobenzene	618-62-2	1.13	1.48	
160	4-Chloro-3-nitrophenol	610-78-6	1.27		0.89
161	2,3,4,5-Tetrachlorophenol	4901-51-3	2.72	2.02	
162	Thiobenzamide	2227-79-4	0.09		-0.04
163	1-Chloro-4-nitrobenzene	100-00-5	0.43	0.81	
164	4,4,4,4-Tetrafluoro-*m*-toluidine	2357-47-3	0.77	0.59	
165	1-Chloro-2-nitrobenzene	88-73-3	0.68	0.76	
166	4-Chloro-6-nitro-*m*-cresol	7147-89-9	1.63	1.09	
167	Pentachlorophenol	87-86-5	2.07	2.33	
168	1,3-Dinitrobenzene	99-65-0	0.76	1.07	
169	2,4-Dinitrotoluene	121-14-2	0.87	1.39	
170	4,5-Dichloro-2-nitroaniline	6641-64-1	1.66	1.36	
171	Pentafluorophenol	771-61-9	1.63	1.58	
172	Pentabromophenol	608-71-9	2.66	*	
173	3-Chloro-4-fluoronitrobenzene	350-30-1	0.8	1.20	
174	1,4-Dinitrobenzene	100-25-4	1.30	1.05	
175	3,4-Dichloronitrobenzene	99-54-7	1.16	1.44	
176	2,5-Dichloronitrobenzene	89-61-2	1.13	1.38	
177	2,4-Dichloro-6-nitroaniline	2683-43-4	1.26	1.17	
178	3,4-Dinitrobenzyl alcohol	79544-31-3	1.09	0.74	
179	2,4-Dichloronitrobenzene	611-06-3	0.99		1.25
180	2,3-Dichloronitrobenzene	3209-22-1	1.07	1.06	
181	1,2-Dinitrobenzene	528-29-0	1.25	0.69	
182	Phenyl isothiocyanate	103-72-0	1.41	0.25	
183	3-Trifluoromethyl-4-ntrophenol	88-30-2	1.65	1.19	
184	2,6-Iodo-4-nitrophenol	305-85-1	1.81		1.01
185	2,4-Chloro-6-nitrophenol	609-89-2	1.75	1.29	
186	1,3,5-Trichloro-2-nitrobenzene	18708-70-8	1.43	1.52	
187	1,2,4-Trichloro-5-nitrobenzene	89-69-0	1.53	1.93	
188	1,2,3-Trichloro-4-nitrobenzene	17700-09-3	1.51	1.72	
189	2-Chloro-5-nitrobenzaldehyde	6361-21-3	0.53	0.94	
190	Pentafluorobenzaldehyde	653-37-2	0.82	1.06	
191	2,4-Dinitro-1-iodobenzene	709-49-9	2.12		1.90
192	2,3,5,6-Tetrachloronitrobenzene	117-18-0	1.82		2.00
193	2,5-Dinitrophenol	329-71-5	1.04	0.95	
194	2,4-Dinitroaniline	97-02-9	0.72	0.73	
195	2,3,4,5-Tetrachloronitrobenzene	879-39-0	1.78	2.30	
196	1,2,3-Trifluoro-4-nitrobenzene	771-69-7	1.89	1.25	

Compounds		CAS	log (1/IGC_{50})	Training set	Test set
197	1,2-Dichloro-4,5-dinitrobenzene	6306-39-4	2.21	2.17	
198	2,6-Dinitroaniline	606-22-4	0.84	0.62	
199	4,6-Dinitro-2-methylphenol	534-52-1	1.73		1.18
200	4-*tert*-Butyl-2,6-dinitrophenol	4097-49-8	1.8	1.97	
201	1-Bromo-2,4-dinitrobenzene	584-48-5	2.31	1.54	
202	2,4-Dinitrophenol	51-28-5	1.06	0.99	
203	1,5-Dichloro-2,3-dinitrobenzene	28689-08-9	2.42	1.98	
204	6-Chloro-2,4-dinitroaniline	3531-19-9	1.12		1.36
205	2-Bromo-4,6-dinitroaniline	1817-73-8	1.24	1.61	
206	2,3,4,6-tetrafluoronitrobenzene	314-41-0	1.87	1.23	
207	2,6-Dinitrophenol	573-56-8	0.83		0.67
208	1-Chloro-2,4-dinitrobenzene	97-00-7	2.16		1.54
209	2,4-Dinitro-1-fluorobenzene	70-34-8	1.71	1.34	
210	Pentafluoronitrobenzene	880-78-4	2.43		2.15
211	1,4-dinitrotetrachlorobenzene	20098-38-8	2.82	2.96	
212	1,5-Difluoro-2,4-dinitrobenzene	327-92-4	2.08	1.57	
213	1,3-Dinitro-2,4,5trichlorobenzene	2678-21-9	2.60	2.32	
214	1,3,5-Trichloro-2,4dinitrobenzene hemihydrate	6284-83-9	2.19	2.2	
215	4-Chloro-3,5-dinitrobenzaldehydea	1930-72-9	2.66	1.47	
216	1-Phenyl-2-propanol	14898-87-4	-0.62	-0.26	
217	4-Methylbenzyl alcohol	589-18-4	-0.49	-0.07	
218	(±)1-Phenyl-2-pentanol	705-73-7	0.16	0.45	
219	4-Isopropylbenzyl alcohol	536-60-7	0.18		0.39
220	2-(*p*-Tolyl)ethylamine	3261-62-9	-0.04	-0.12	
221	4-Methyl benzylamine	104-84-7	-0.01	-0.28	
222	3-Methylbenzyl alcohol	587-03-1	-0.24	-0.08	
223	3-Phenyl-2-propen-1-ol	104-54-1	-0.08		0.06
224	4-*tert*-Buthylbenzyl alcohol	877-65-6	0.48	0.47	
225	4-Methylphenetyl alcohol	699-02-5	-0.26	0.01	
226	1-Phenylethylamine	618-36-0	-0.18		-0.53
227	2-Methylbenzyl alcohol	89-95-2	-0.43		-0.39
228	2-Methyl-1-phenyl-2-propanol	100-86-7	-0.41	-0.5	
229	*N*-Methylphenethylamine	589-08-2	-0.41	-0.27	
230	b-Methylphenethylamine	582-22-9	-0.28	-0.17	
231	(±)-1-Phenyl-1-butanol	22135-49-5	-0.09	0.08	
232	(±)-1-Phenyl-1-propanol	93-54-9	-0.43	-0.28	
233	Phenetyl alcohol	60-12-8	-0.59	-0.37	
234	2-Phenyl-1-propanol	1123-85-9	-0.4	-0.26	
235	2-Phenyl-2-propanol	617-94-7	-0.57	-0.51	
236	2-Phenyl-1-butanol	89104-46-1	-0.11		0.11
237	Benzhydrol	91-01-0	0.5	0.41	
238	Benzaldoxime	622-32-2	-0.11	0.07	
239	3,5-Dimethylaniline	108-69-0	-0.36	0.17	
240	4-*tert*-Buthylaniline	769-92-6	0.36	0.25	
241	2,4-Dimethylaniline	95-68-1	-0.29	0.18	
242	4-Phenylbutyronitrile	2046-18-6	0.15	0.31	
243	2,4,6-Trimethylaniline	88-05-1	-0.05	0.44	
244	3-Phenylpropionitrile	645-59-0	-0.16	-0.01	
245	4-*sec*-Butylaniline	30273-11-1	0.61	0.51	
246	2,3-Dimethylaniline	87-59-2	-0.43	0.14	
247	Benzyl cyanide	140-29-4	-0.36	-0.21	
248	2,5-Dimethylaniline	95-78-3	-0.33		0.17
249	α-Methylbenzyl cyanide	1823-91-2	0.01		0.00
250	2-Isopropylaniline	643-28-7	0.12	0.06	
251	2,6-Dimethylaniline	87-62-7	-0.43	0.03	
252	*N*-ethylaniline	103-69-5	0.07		-0.02
253	2-Propylaniline	1821-39-2	0.08		0.12
254	*N*-Methylaniline	100-61-8	0.06	-0.31	
255	2-Amino-4-*tert*-butylaniline	1199-46-8	0.37		0.25
256	2-Methoxyaniline	90-04-0	-0.69	-0.43	
257	3-Phenylpyridine	1008-88-4	0.47		0.58
258	2-Aminobenzyl alcohol	5344-90-1	-1.07	-0.16	
259	2-Benzylpyridine	101-82-6	0.38	0.52	
260	3,5-Di-*tert*-butylphenol	1138-52-9	1.64	1.97	
261	Phenyl propargyl sulfide	5651-88-7	0.54	0.5	
262	4-Ethoxyphenol	622-62-8	0.01		0.16

Compounds		CAS	log (1/IGC_{50})	Training set	Test set
263	4-Butoxyphenol	122-94-1	0.70	0.86	
264	4-Benzylpyridine	2116-65-6	0.63		0.66
265	2-Phenylpyridine	1008-89-5	0.27	0.8	
266	3,4-Dimethylphenol	95-65-8	0.12	0.36	
267	3-*tert*-Buthylphenol	585-34-2	0.74	0.54	
268	3,5-Dimethylphenol	108-68-9	0.11		0.31
269	6-*tert*-Buthyl-2,4-dimethylphenol	1879-09-0	1.16	0.98	
270	4-Isopropylphenol	99-89-8	0.47	0.42	
271	3-Isopropylphenol	618-45-1	0.61		0.57
272	2,3-Dimethylphenol	526-75-0	0.12	0.27	
273	2,5-Dimethylphenol	95-87-4	0.14		0.30
274	4-Hydroxy-3-methoxybenzyl alcohol	498-00-0	-0.7	-0.19	
275	2-Isopropylphenol	88-69-7	0.61	0.39	
276	3-Amino-2-cresol	53222-92-7	-0.55	-0.02	
277	4-Chloro-2-methylaniline	95-69-2	0.35	0.18	
278	2-Methoxy-4-propenylphenol	97-54-1	0.75		0.69
279	2,4,6-*tris*(Dimethylaminomethyl)phenol	90-72-2	-0.52	*	
280	2-Fluoroaniline	348-54-9	-0.37	0.01	
281	4-Aminobenzyl cyanide	3544-25-0	-0.76	-0.4	
282	3-Iodoaniline	626-01-7	0.65		0.44
283	3-Cinnamonitrile	4360-47-8	0.16		0.52
284	3-Fluorobenzyl alcohol	456-47-3	-0.39	-0.16	
285	3-Cyanoaniline	2237-30-1	-0.47	-0.24	
286	4-Fluorophenol	371-41-5	0.02	-0.00	
287	2-Iodoaniline	615-43-0	0.35	0.49	
288	3-Fluoroaniline	372-19-0	-0.10	0.02	
289	4-Chloro-2-methylphenol	1570-64-5	0.7	0.56	
290	4-Chloro-3-ethylphenol	14143-32-9	1.08		0.72
291	2-Chloro-4,5-dimethylphenol	1124-04-5	0.69	0.96	
292	3,5-Dimethoxyphenol	500-99-2	-0.09	-0.04	
293	4-Hydroxybenzyl cyanide	14191-95-8	-0.38	-0.23	
294	4-Bromo-2,6-dimethylphenol	2374-05-2	1.16		0.70
295	2-Bromobenzyl alcohol	18982-54-2	0.10		0.07
296	2-Chloro-5-methylphenol	615-74-7	0.54	0.56	
297	2-Fluorophenol	367-12-4	0.19	-0.06	
298	4-(Dimethylamino)benzaldehyde	100-10-7	0.23	0.15	
299	4-Bromophenol	106-41-2	0.68	0.50	
300	3-Chloro-2-methylaniline	95-79-4	0.50	0.14	
301	3-Chloro-4-methylaniline	95-74-9	0.39		0.19
302	3-Chloro-2-methylaniline	87-60-5	0.38	0.14	
303	4-Chlorophenethyl alcohol	1875-88-3	0.32	0.12	
304	4-Chlorobenzyl alcohol	873-76-7	0.25	0.10	
305	2-Bromo-4-methylphenol	6627-55-0	0.60	0.57	
306	1,3,5-Trimethyl-2-nitrobenzene	603-71-4	0.86	1.08	
307	3-Chlorobenzyl alcohol	873-63-2	0.15	0.10	
308	2-Bromophenol	95-56-7	0.33		0.51
309	4-Hydroxy-3-methoxybenzonitrile	4421-08-3	-0.03	-0.14	
310	3-Nitrobenzyl alcohol	619-25-0	-0.22		0.25
311	4-Bromophenyl acetonitrile	16532-79-9	0.6	0.26	
312	4-Methoxybenzonitrile	874-90-8	0.10	-0.14	
313	2-Hydroxy-4,5-dimethylacetophenone	36436-65-4	0.71		0.57
314	2-Anisaldehyde	135-02-4	0.15		-0.03
315	4-Chlororesorcinol	95-88-5	0.13	0.59	
316	Methyl-4-methylaminobenzoate	18358-63-9	0.31	0.19	
317	4-Phenoxybenzaldehyde	67-36-7	1.26	1.26	
318	3-Hydroxy-4-methoxybenzaldehyde	621-59-0	-0.14	-0.08	
319	4-Biphenylcarboxaldehyde	3218-36-8	1.12	0.95	
320	2,4,5-Trimethoxybenzaldehyde	4460-86-0	-0.10	0.36	
321	4-Benzoylaniline	1137-41-3	0.68	0.94	
322	3-Anisaldehyde	5991-31-1	0.23	-0.08	
323	*n*-Propyl cinnamate	7778-83-8	1.23	1.12	
324	(*trans*)-Ethyl cinnamate	103-36-6	0.99	0.71	
325	Hexanophenone	942-92-7	1.19	1.11	
326	*n*-Butyl cinnamate	538-65-8	1.53	1.75	
327	4-Chlorobenzyl cyanide	140-53-4	0.66	0.21	
328	(*trans*)-Methyl cinnamate	103-26-4	0.58	0.38	

Compounds		CAS	log (1/IGC_{50})	Training set	Test set
329	Ethyl-4-methoxybenzoate	94-30-4	0.77	0.6	
330	Phenylacetic acid hydrazide	937-39-3	-0.48	-0.66	
331	3-Hydroxybenzaldehyde	100-83-4	0.08	-0.02	
332	2,6-Dichlorophenol	87-65-0	0.73	0.86	
333	Benzyl methacrylate	2495-37-6	0.65		0.57
334	Isoamyl-4-hydroxybenzoate	6521-30-8	1.48	1.48	
335	Benzyl-4-hydroxyphenyl ketonea	2491-32-9	1.07	*	
336	Benzyl benzoate	120-51-4	1.45	1.08	
337	4-Benzoylphenol	1137-42-4	1.02	0.99	
338	2-Methyl-5-nitrophenol	5428-54-6	0.66	0.71	
339	3-Acetoamidophenol	621-42-1	-0.16		0.16
340	4-Cyanobenzamide	3034-34-2	-0.38		-0.16
341	2-Nitrobiphenyl	86-00-0	1.30		1.38
342	5-Chloro-2-hydroxybenzamide	7120-43-6	0.59	0.26	
343	3-Nitrophenol	554-84-7	0.51	0.37	
344	Phenyl-1,3-dialdehyde	626-19-7	0.18		0.37
345	Ethyl-4-bromobenzoate	5798-75-4	1.33	0.86	
346	2,4-Dihydroxyacetophenone	89-84-9	0.25	-0.17	
347	3-Chlorobenzophenone	1016-78-0	1.55	1.35	
348	Phenyl-4-hydroxybenzoate	17696-62-7	1.37	1.38	
349	Phenyl benzoate	93-99-2	1.35	1.29	
350	2-Hydroxy-4-methoxybenzophenone	131-57-7	1.42	1.33	
351	Benzylidene malononitrile	2700-22-3	0.64	0.20	
352	4-Nitrophenyl phenyl ether	620-88-2	1.58	1.93	
353	Resorcinol monobenzoate	136-36-7	1.11		1.41
354	4-Bromophenyl-3-pyridyl ketonea	14548-45-9	0.82	*	
355	3-Nitroacetophenone	121-89-1	0.32		0.51
356	3-Nitrobenzaldehyde	99-61-6	0.11	-0.02	0.43
357	Ethyl phenylcyanoacetate	4553-07-5	-0.02		
358	2-Nitroanisole	91-23-6	-0.07		0.41
359	3-Methyl-2-nitrophenol	4920-77-8	0.61	0.27	
360	2,5-Diphenyl-1,4-benzoquinone	844-51-9	1.48	1.78	
361	2-Nitrobenzamide	610-15-1	-0.72	0.06	
362	Methyl-2,5-dichlorobenzoate	2905-69-3	0.81	1.07	
363	2-Nitrobenzaldehyde	552-89-6	0.17	0.18	
364	4-Methyl-2-nitrophenol	119-33-5	0.57	0.69	
365	2,2',4,4'-Tetrahydroxybenzophenone	131-55-5	0.96	1.15	
366	4-Nitrobenzaldehyde	555-16-8	0.20	0.43	
367	5-Methyl-2-nitrophenol	700-38-9	0.59	0.6	
368	3,5-Dichlorosalicylaldehyde	90-60-8	1.55	0.97	
369	2-(Benzylthio)-3-nitropyridine	69212-31-3	1.72	1.46	
370	Ethyl-4-nitrobenzoate	99-77-4	0.71		1.04
371	2,4-Dichlorobenzaldehyde	874-42-0	1.04		0.88
372	2',3',4'-Trichloroacetophenone	13608-87-2	1.34	1.12	
373	2,20-Dihydroxybenzophenone	835-11-0	1.16		0.82
374	Methyl-4-nitrobenzoate	619-50-1	0.39	0.65	
375	2-Chloromethyl-4-nitrophenol	2973-19-5	0.75	0.86	
376	α,α,α-Trifluoro-*p*-cresol	402-45-9	0.62		
377	Dimethylnitroterephthalate	5292-45-5	0.43		0.36
378	Thioacetanilide	637-53-6	-0.01	1.52	
379	2-Nitro resorcinol	601-89-8	0.66	0.24	
380	3,5-Dibromo-4-hydroxybenzonitrile	1689-84-5	1.16	0.18	
381	Pentafluorobenzyl alcohol	440-60-8	-0.20	1.41	
382	Methyl-4-chloro-2-nitrobenzoate	42087-80-9	0.82	0.44	
383	1-Fluoro-2-nitrobenzene	1493-27-2	0.23	1.21	
384	α,α,α-Tetrafluoro-*o*-toluidine	393-39-5	-0.02	0.46	
385	3-Hydroxy-4-nitrobenzaldehyde	704-13-2	0.27	0	
386	2,5-Dibromonitrobenzene	3460-18-2	1.37	0.32	
387	Benzoyl cyanide	613-90-1	0.31	1.28	
388	4,5-Difluoro-2-nitroaniline	78056-39-0	0.75	-0.09	
389	2,5-Difluoronitrobenzene	364-74-9	0.33	0.85	
390	2,4-Dibromo-6-nitroaniline	827-23-6	1.62	0.93	
391	4-Hydroxy-3-nitrobenzaldehyde	3011-34-5	0.61	1.86	0.49
392	Benzoyl isothiocyanate	532-55-8	0.10	0.29	

gas-phase geometry calculation of AM1. Twenty three constant and near constant variables exclude from descriptors and then 193 descriptors have low correlations (< 0.1) with log (IGC_{50}) and 677 descriptors with pair correlations 0.98 exclude. The total remaining descriptors were 134 and we use these descriptors for stepwise variable selection.

Cluster analysis: The cluster analysis (CA) is the name of a group of methods used to recognize similarities among cases (objects) or among variables and to single out some categories as a set of similar cases (or variables)[13]. This cluster analysis comprehends a number of different 'classification algorithms' and it allows organizing the data into subsystems. These algorithms are grouped into two categories: hierarchical clustering and partitional (non-hierarchical) clustering. Hierarchical clustering rearranges objects in a tree structure (joining clustering) in an agglomerative (bottom-up) procedure. On the other hand, partitional clustering assumes that the objects have non-hierarchical characters[14]. The most used cluster algorithms are the k-means cluster analysis (k-MCA) and Jarvis-Patrick algorithm (also known as k-nearest neighbor cluster analysis; k-NNCA); in our case, in order to design the training and test series to guarantee structural and toxicity variability in both series of the present database, we carried out k-MCA for the entire dataset of compounds. This approach (clustering) ensures that the similarity principle can be employed for the activity prediction of the test set. The number of members in each cluster and the standard deviation of the variables in the cluster (kept as low as possible) were taking into account, to have an acceptable statistical quality of data partition into clusters. The values of the standard deviation between and within clusters, those of the respective Fisher-ratio and their *p*-level of significance were also examined[14,15]. Finally, before carrying out the cluster processes, all the variables were standardized. In standardization, all values of selected variables (molecular descriptors) were replaced by standardized values, which are computed as follows:

$$Z = (x\text{-}mu)/\sigma$$

where, Z and x are normal distribution of z-scores and each raw score, mu and σ are normal mean and standard deviation of a set of scores.

Stepwise multiple linear regression 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 compound log (IGC_{50}). The log (IGC_{50}) value for each benzene derivative would then be calculated as a composite of each molecular descriptor weighted by the respective coefficients. A multilinear model can be represented as:

$$y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3x_3 + \cdots + \beta_kX_k \quad (1)$$

where, k is the number of independent variables, $\beta_1 \ldots \beta_k$, the regression coefficients and y is the dependent variable. Regression coefficients represent the independent contributions of each calculated molecular descriptor.

A single multiple linear regression model was developed for benzene derivative compounds using the SPSS version 11.5 software. The multiple linear regression model was built using a training set and validation using an external prediction set. Multiple linear regression techniques based on least-squares procedures are very often used for estimating the coefficients involved in the model equation[16].

The stepwise multiple linear regression is a commonly used variant of multiple linear regression. In this case, also a multiple-term linear equation is produced, but not all independent variables are used. Each variable is added to the equation at a time and a new regression is performed. The new term is retained only if equation passes a test for significance. This regression method is especially useful when the number of variables is large and when the key descriptors are not known. Usually, molecular descriptor matrices cannot be directly used as independent variables in the multiple linear regression analysis due to their lack of homogeneity, the high correlation between descriptors is larger than the number of compounds, some of them maybe redundant. Thus previous to the multiple linear regression analysis, normally a reduction of variables is necessary in order to obtain a concentrated set of significant underlying variables, not correlated between them, losing the minimum amount of information.

Partial least square: Partial least square is a generalization of regression, which can handle data with strongly correlated and/or noisy or numerous X variables[17]. It gives a reduced solution, which is statistically more robust than multiple linear regression. The linear partial least square model finds new variables (latent variables or X scores), which are linear combinations of the original variables. The latent variables in partial least square are also linear combinations of the descriptive variables in the data set, but instead of maximizing the variance in the matrix with descriptive variables like in principle component analysis (PCA), the covariance with the response variable is maximized. The scores on the partial least square factors are used as input for multiple linear regression after selection of the optimal number of partial least square-factors to be considered[18].

To avoid overfitting, a strict test for the significance of each consecutive partial least square component is necessary and then stopping when the components are non-significant. Cross-validation is a reliable and commonly used method for testing this significance[19]. However, recently it has been shown that from the viewpoint of external predictability, choice of variables for partial least square based on internal validation may not be optimum[15]. Application of partial least square allows the construction of larger QSAR equations while still avoiding overfitting and eliminating most variables. Partial least square is normally used in combination with cross-validation to obtain the optimum number of components. This ensures that the QSAR equations are selected based on their ability to predict the data rather than to fit the data[16]. Based on the standardized regression coefficients, the variables with smaller coefficients were removed from the partial least square regression, until there was no further improvement in Q^2 value, irrespective of the components.

RESULTS AND DISCUSSION

Similarity analysis and design of training and test sets: Principle component analysis study of 392 benzene derivatives X-matrix data, showed compounds 153, 172, 279, 335 and 354 (3,4,5,6-tetrabromo-*o*-cresol, pentabromophenol, 2,4,6-

tris(dimethylaminomethyl) phenol benzyl-4-hydroxyphenyl ketone and 4-bromophenyl-3-pyridyl) are as statistical outlieres, once rejected the statistical outliers, in order to split the whole group into two datasets (training and predicting ones), we perform a k-MCA. The main idea of this procedure consists in making a partition of chemicals in several statistically representative classes of compounds. This procedure ensures that any chemical class (as determined by the clusters derived from k-MCA) will be represented in both compounds' series. This rational design of the training and predicting series allowed us to design both sets: that are representative of the whole experimental universe. This procedure split the dataset of benzene derivatives into 9 clusters. Afterward, the selection of the training and prediction sets was performed by taking, in a random way, compounds belonging to each cluster. From these 387 compounds, 309 were chosen at random to form the training set. The remaining subset, composed of 78 compounds, was prepared as test set for the external set validation of the models. These compounds were never used in the development of the classification models. Fig. 1 illustrates graphically the above-described procedure, where a cluster analyses was performed to select a representative sample for the training and test sets.

Stepwise multiple linear regression and partial least square: Partial least square was applied to the data set after selection of descriptors by the stepwise multiple linear regression. The stepwise multiple linear regression algorithm was applied to the data set using the decimal logarithm as the log (IGC_{50}) values as response variables and the autoscaled calculated descriptors as independent variables. For evaluation of the predictive power of the generated partial least square, the optimized model was applied for prediction of log (IGC_{50}) values of 78 compounds in the prediction set, which were not used in the optimization procedure. Table-2 show the selected

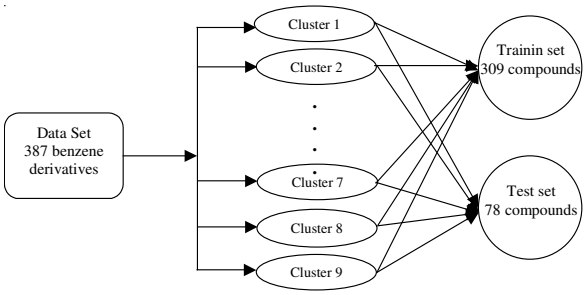


Fig. 1. General algorithm used for designing training and test sets throughout k-MCA

descriptors, their definition and class partial least square regression applied in the modeling procedure (stepwise MLR-PLS). For the constructed model, the predictive ability of the partial least square model was evaluated by calculation of statistical parameters. The root mean square error of calibration and prediction (RMSEC and RMSEP), predicted residual error sum of square (PRESS), standard error of prediction (SEP), the percent relative errors (% RE), square adjusted correlation coefficient for the training and prediction set (R^2 and R^2_{pred}) obtained by the partial least square method are presented in Table-3. The coefficient of determination equals 0.8104 and the RMSEC value is 0.3296. These values show that the stepwise MLR-PLS model fits the calibration data well and also has good predictive abilities ($R^2_{Pred} = 0.819$). The residual plot (Fig. 2) of the stepwise MLR-PLS model shows considerable small residuals. The plot of predicted log (IGC_{50}) *versus* experimental log (IGC_{50}) obtained by the stepwise MLR-PLS modeling is shown in Fig. 3. Also the agreement observed between the predicted and experimental log (IGC_{50}) values in Fig. 4 confirms a good predictive ability of stepwise MLR-PLS modeling. The partial least square latent variable for this data by this model was obtained[7].

TABLE-2
SELECTED DESCRIPTORS IN THE STEPWISE MLR MODEL AND STANDARDIZED COEFFICIENT FOR STEPWISE MLR-PLS MODEL

Descriptor	Definition	Descriptor class	Standardized coefficient (stepwise MLR-PLS model)
MLOGP	Moriguchi octanol-water partition coefficient (logP)	Molecular properties	0.2232
RARS	R matrix average row sum	GETAWAY descriptors	-0.1846
AM	A total size index/weighted by atomic masses	WHIM descriptors	0.2852
H-046	H attached to C0(sp^3) no X attached to next C^b	Atom-centered fragments	0.3278
BELM1	Lowest eigenvalue n. 1 of Burden matrix/weighted by atomic masses	Burden eigen values	-0.2294
PCWTE	Partial charge weighted topological electronic descriptor	Charge descriptors	0.1498
RDF040M	Radial Distribution Function - 4.0/ weighted by atomic masses	RDF descriptors	0.2372
FDI	Folding degree index	Geometrical descriptors	0.0964
MOR10P	3D-MoRSE -signal 01/weighted by atomic polarizabilities	3D-MoRSE descriptors	0.0556
MOR06U	3D-MoRSE -signal 06/unweighted	3D-MoRSE descriptors	-0.2570
SHP2	Average shape profile index of order 2	Randic molecular profiles	0.2274
JGI3	Mean topological charge index of order3	Topological charge indices	-0.2120
RDF050M	Radial Distribution Function - 5.0/weighted by atomic masses	RDF descriptors	0.1470
MOR18M	3D-MoRSE -signal 18/weighted by atomic masses	3D-MoRSE descriptors	-0.0976
RDF010M	Radial Distribution Function-1.0/weighted by atomic masses	RDF descriptors	-0.1406
MOR05M	3D-MoRSE -signal 05/weighted by atomic masses	3D-MoRSE descriptors	-0.1180

TABLE-3
OVERVIEW OF STEPWISE MLR-PLS MODEL

Method	RMSEC	RMSEP	REP (%)	SEP	PRESS	R^2(adjusted)	R^2_{Pred}(adjusted)
Stepwise MLR-PLS	0.32	0.29	50.72	0.29	6.55	0.80	0.81

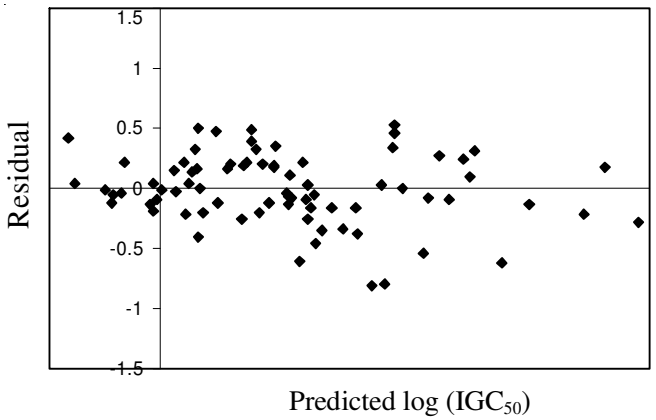


Fig. 2. Residual plot for the stepwise MLR-PLS model

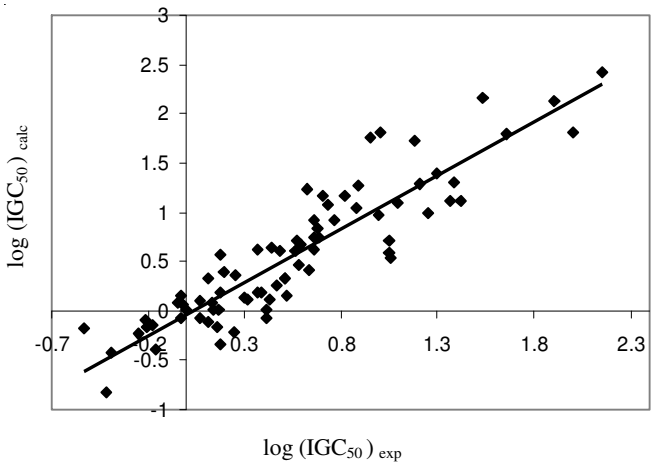


Fig. 3. Predicted log (IGC_{50}) by stepwise MLR-PLS modeling *versus* experimental log (IGC_{50}) for test molecules in prediction

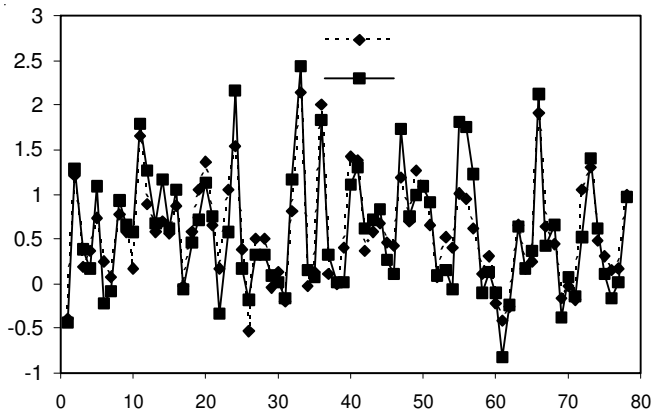


Fig. 4. Plots of experimental and predicted log (IGC_{50}) values by stepwise MLR-PLS modeling versus sample number in the prediction set

The descriptor H-046 is based on number of hydrogen attached to CO (sp^3) (carbon atom with formal oxidation number 0 and hybridization sp^3) with X (O, N, S, P, Se or halogens) attached to next C. The MLOGP descriptor is a measure for the Moriguchi octanol-water partition coefficient (log P). Most of the other descriptors can be related to the two-dimensional (BELM1) or three-dimensional (RARS, AM, PCWTE, RDF040M). Positive values in the regression coefficients indicate that the indicated descriptor contributes positively to the value of log (IGC_{50}), whereas negative values indicate that the greater the value of the descriptor the lower the value of log (IGC_{50}).

Our method favourably compares with other approaches implemented in the Dragon software and atom-based non-stochastic and stochastic linear indices for this data set[20]. All these results are summarized in Table-4, where a detailed comparison can be more easily performed. The results were obtained on different sets of molecules, since they are based on different test sets. As it can be seen, our model has statistical parameter better than models obtained with reference[20]. In this sense, the present approach showed the greater values of squared adjusted correlation coefficient of 0.8104 (R^2_{Pred} = 0.8195), with stepwise MLR-PLS, correspondingly.

Statistical parameters: For evaluation of the predictive power of the generated partial least square, the optimized model was applied for prediction of log(IGC_{50}) values of 78 compounds in the prediction set which were not used in the optimization procedure. For the constructed model, five general statistical parameters were selected to evaluate the predictive ability of the model for log(IGC_{50}) values. In this case, the predicted log(IGC_{50})'s of each sample in prediction step were compared with the experimental aquatic toxicity. The PRESS (predicted residual sum of squares) statistic appears to be the most important parameter accounting for a good estimate of the real predictive error of the models. Its small value indicates that the model predicts better than chance and can be considered statistically significant.

$$\text{Press} = \sum_{i=1}^{n} (\hat{y} - y_i)^2 \quad (2)$$

Root mean square error of prediction or calibration (RMSEP or RMSEC) is a measurement of the average difference between predicted and experimental values, at the prediction step. Root mean square error of prediction can be interpreted as the average prediction error, expressed in the same units as the original response values. The RMSEP was obtained by the following formula:

TABLE-4
STATISTICAL PARAMETERS OF THE QSAR MODELS OBTAINED USING DIFFERENT MOLECULAR DESCRIPTORS TO PREDICT AQUATIC TOXICITY[17]

Index		R^2
Non-stochastic linear indices	${}^{p}f^{H}_{0L}(x_E), {}^{k}f_0(x), {}^{p}f^{H}_{1L}(x_E), {}^{p}f_{3L}(x_E), {}^{pp}f_3(x), {}^{M}f^{H}_{9L}(x_E)$	0.721
Stochastic linear indices	${}^{Ms}f_{15}(x), {}^{ks}f^{H}_{0}(x), {}^{ks}f^{H}_{5}(x), {}^{Gs}f^{H}_{0}(x), {}^{Vs}f^{H}_{4L}(x_E), {}^{Vs}f^{H}_{2L}(x_E)$	0.733
2D autocorrelations	ATS3v, ATS8v, ATS3e, ATS8e, MATS1e, GATS1m	0.609
BCUT	BEHm7, BELm4, BELm6, BELv4, BELe6, BEHp6	0.690
Gálvez topological charge indices	GGI2, GGI6, GGI8, JGI2, JGI5, JGI8	0.516
Topological descriptors	ISIZ, X2sol, S2K,PW2, TIC1, pilD	0.716
Molecular walk count	MWC08, MWC09, TWC, SRW10	0.346

$$\text{RMSEP (or RMSEC)} = \left[\frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y_i)^2\right]^{0.5} \quad (3)$$

The third statistical parameter was relative error of prediction (REP) that shows the predictive ability of each component and is calculated as:

$$\text{REP}(\%) = \frac{100}{n}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y_i)^2\right]^{0.5} \quad (4)$$

The predictive applicability of a regression model is described in various ways. The most general expression is the standard error of prediction (SEP) which is given in the following formula:

$$\text{SEP} = \left[\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n-1}\right]^{0.5} \quad (5)$$

The square of the adjusted correlation coefficient, which is, indicated the quality of fit of all the data to a straight line is calculated for the checking of test set and is calculated as:

$$R^2_{\text{Pred}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (6)$$

where, y_i is the experimental log (IGC_{50}) of the benzene derivatives in the sample i, $\hat{y}_i$ represented the predicted log (IGC_{50}) of the benzene derivatives in the sample i, $\bar{y}$, is the mean of experimental log (IGC_{50}) in the prediction set and n is the total number of samples used in the prediction set. The statistical results (PRESS, RMSEC, RMSEP, REP %, SEP and R^2) are summarized in Table-3.

Conclusion

We have developed here a useful QSAR equation derived from quantum chemical descriptors associated with aquatic toxicity properties of 392 benzene derivatives. For each compound 134 descriptors, 11 classes of Dragon descriptors, calculated. The dataset was carefully split into training and test sets, guaranteeing enough molecular diversity in each subset, by using k-MCA cluster analysis. Then the best set of calculated descriptors was selected by stepwise multiple linear regression. Model was obtained with partial least square regression. Stepwise MLR-PLS is successfully presented for prediction aquatic toxicity [log ($1/IGC_{50}$)] of various benzene derivatives ($R^2_{\text{Pred}} = 0.8195$) with diverse chemical structures using a linear quantitative structure-activity property relationship. This model with high statistical quality and low prediction errors was obtained. In general, it can be concluded that, for this data set, the combinations of linear modeling techniques result in an improvement of the linear models. The results indicate that four descriptors, Moriguchi octanol-water partition coefficient (log P) (M log P), H attached to C_0 (sp^3) number X attached to next C (H-046), A total size index/ weighted by atomic masses (AM) and R matrix average row sum (RARS) were selected and play an important role on the aquatic toxicity of benzene derivatives structure.

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 aquatic toxicity in benzene derivatives that predetermine these activity.

REFERENCES

1. A. Worth, A. Bassan, J. De Bruijn, A. Saliner, T. Netzeva, G. Patlewicz, M. Pavan, I. Tsakovska and S. Eisenreich, *SAR QSAR Environ. Res.*, **18**, 111 (2007).
2. C. Hansch, P.P. Maloney, T. Fujita and R.M. Muir, *Nature*, **194**, 178 (1962).
3. R. Todeschini and V. Consonni, Handbook of Molecular Descriptors, Weinheim: Wiley-VCH (2000).
4. M. Karelson, Molecular Descriptors in QSAR/QSPR, Wiley-Interscience (2000); B. Hemmateenejad, M. Safarpour and F. Taghavi, *J. Mol. Struct. (Theochem.)*, **635**, 183 (2003); I. Moriguchi, S. Hirono, Q. Liu, I. Nakagome and Y. Matsushita, *Chem. Pharm. Bull*. (Tokyo), **40**, 127 (1992).
5. W. Tong, H. Hong, Q. Xie, L. Shi, H. Fang and R. Perkins, *Curr. Comput. Aided Drug. Des.*, **1**, 195 (2005); L. He and P. Jurs, *J. Mol. Graphics Model.*, **23**, 503 (2005); T. Ghafourian and M. Cronin, *SAR QSAR Environ. Res.*, **16**, 171 (2005); A. Tropsha, P. Gramatica and V. Gombar, *QSAR Comb. Sci.*, **22**, 69 (2003); A. Golbraikh and A. Tropsha, *J. Mol. Graphics Model.*, **20**, 269 (2002).
6. S. Wold and L. Eriksson, Chemometric Methods in Molecular Design, VCH, Weinheim, Vol. **195** (1995).
7. S.D. Brown, S.T. Sum, F. Despagne and B.K. Lavine, *Ana. Ch.*, **68**, 21 (1996); K. Roy and J. Leonard, *Bioorg. Med. Chem.*, **13**, 2967 (2005).
8. T. Schultz, T. Netzeva, in Predicting Chemical Toxicity and Fate, p. 265 (2004).
9. S. Bradbury, C. Russom, G. Ankley, T. Schultz and J. Walker, *Environ. Toxicol. Chem.*, **22**, 1789 (2009).
10. R. Johnson and D. Wichern, Applied Multivariate Statistical Analysis, Prentice Hall Englewood Cliffs, NJ, (1998); J. Mc Farland and D. Gans, Chapter Cluster Significance Analysis, VCH, Weinheim, 295 (1995).
11. P. Broto, G. Moreau and C. Vandycke, *Eur. J. Med. Chem.*, **19**, 61 (1984); J. Gálvez, R. García, M. Salabert and R. Soler, *J. Chem. Inf. Comput. Sci.*, **34**, 520 (1994); L. Kier and L. Hall, Molecular Connectivity in Structure-Activity Analysis, Wiley, New York (1986); E. Konstantinova, *J. Chem. Inf. Comput. Sci.*, **36**, 54 (1996); G. Ruecker and C. Ruecker, *J. Chem. Inf. Comput. Sci.*, **33**, 683 (1993).
12. M. Dewar, E. Zoebisch, E. Healy and J. Stewart, *J. Am. Chem. Soc.*, **107**, 3902 (1985).
13. J. Xu and A. Hagler, *Molecules*, **7**, 566 (2002).
14. R. Leardi and A. Lupiáñez González, *Chemometrics Intell. Lab. Syst.*, **41**, 195 (1998).
15. P. Geladi and B. Kowalski, *Anal. Chim. Acta*, **185**, 1 (1986).
16. R. Leardi, R. Boggia and M. Terrile, *J. Chemometrics*, **6**, 267 (2005).
17. D. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning Addision Wesley Longman Inc, MA, USA (1989).
18. R. Yu, Introduction to Chemometrics, Changsha, Hunan Education House, Vol. 67 (1991).
19. B. Dayal and J. MacGregor, *J. Chemometrics*, **11**, 73 (1998).
20. J. Castillo-Garit, Y. Marrero-Ponce, J. Escobar, F. Torrens and R. Rotondo, *Chemosphere*, **73**, 415 (2008).