Journal of Computational Methods in Molecular Design, 2016, 6 (2):29-38

Scholars Research Library

ISSN : 2231- 3176
CODEN (USA): JCMMDA

Qunatitative Structure-Toxicity Models for Halogenated Phenols using Electrophilicity and Hydrophobicity Indexes

Khadidja Bellifa and Sidi Mohamed Mekelleche*

Laboratory of Applied Thermodynamics and Molecular Modeling, Department of Chemistry, Faculty of Science, University of Tlemcen, BP 119,Tlemcen, 13000, Algeria
Corresponding Email: sm_mekelleche@mail.univ-tlemcen.dz

ABSTRACT

Phenols and especially halogenated phenols represent a substantial part of the chemicals produced worldwide and are known as aquatic pollutants. Quantitative structure–toxicity relationship (QSTR) models are useful for understanding how chemical structure relates to the toxicity of chemicals. In the present study, the acute toxicities of 45 halogenated phenols to Tetrahymena Pyriformis were estimated using no cost semi-empirical AM1, PM3, and PM6 quantum chemistry methods. QSTR models were established using the multiple linear regression technique and the predictive ability of the models was evaluated by the internal cross-validation, the Y-randomization and the external validation. Their structural chemical domain has been defined by the leverage approach. The results show that that the best QSTR model is obtained with the AM1 method (R^2= 0.91, R^2_{CV}= 0.90, SD= 0.20 for the training set and R^2= 0.96, SD= 0.11 for the test set). Moreover, all the Tropsha' criteria for a predictive QSTR model are checked. The obtained QSTR models were developed with a few number of meaningful descriptors and put in evidence the importance of the transport factor expressed by the hydrophobicity parameter and the electronic effect expressed by the Parr's electrophilicity index in the interpretation and the prediction of the toxicity of halogenated phenols.

Keywords: Halogenated phenols; Toxicity; Electrophilicity index; Hydrophobicity index, Quantitative Structure-Toxicity Relationships; Semi-empirical methods.

INTRODUCTION

A variety of organic compounds can be environmental pollutants and toxicants. Therfore, it is vital to protect the environment and prevent occupational poisoning by studying the toxicity of these pollutants. The impact of the potential hazard of unstead chemicals, a challenge confroting international regulatory agencies [1-4], can be measured by experimental investigations,but this approach is both quite expensive and time-consuming [5]. Because of this a great deal of effort has been put into the use of theoretical and computational methods to make up for the disadvantages of the experiment. An alternative is to rely on QSTR (Quantitative Structure-Toxicity Relationship) models that describe a mathematical relationship between the structural feature of a set of chemicals and the particular toxicity assoociated with them [6,7].

Phenols represent a substantial part of the chemicals produced worldwide. They have been widely used as basic materials in medicine, industry and agriculture [8]. They can speared through air and water, with strong carcinogenecity and mutagenicity [9-10], which causes great damage to environment. The environmental hazards of phenolic compound have led to wide concern by researchers, and many works have been done for their QSTR models in recent years [11-15]. Cronin et al. [14] obtained QSTR models for a series of phenols using multiple linear regression (MLR) and neural netwok (NN) methods and their obtained results show the ability of the elaborated models to predict the two non-covalent mechanisms(polar narcosis and respiratory uncoupling) and their inability to estimate the toxicity of the electrophilic mechanism. Pasha et al. [16] studied the toxicity of a series of phenol derivatives using semi-empirical and DFT methods. However, the elaborated QSTR models involve several correlated molecular descriptors and the calculated values of the electrophilicity (see Tables 2-5 of Ref. 16) are erroneous and senseless. Recently, Ertürk et al. [17]studied the toxicity of a series of phenols to marine alga *Dunaliellatertiolecta* using the consensus MLR and NN approaches. Their QSTR models, elaborated on the basis of molecular descriptors calculated using CODESSA [18] and DRAGON [19] softwares, provided acceptable predictions although the physical meaning of the involved descriptors and their correlation with toxicity are not always clear and rationally explained. Ertürk et al.[20] also modelled the toxicity of a series of phenols to *Chlorella vulgaris* using the MLR approach and their results revealed that the established QSTR models provide acceptable predictions ($R^2 < 0.84$, SD <0.20) for polar narcotics and respiratory uncouplers, but they lack to predict the toxicity of reactive phenols exhibiting an electrophilic mechanism.

Halogenated phenols and specially chlorophenols are the most widespread and the largest group of phenols and these compounds are generally polar narcotics [21]. According to Schultz [14], it is difficult to well model the whole phenols in the reason of the existence of many modes of action. It is often difficult to determine whether or not a chemical possesses a particular mechanism of action. For this reason QSTRs were usually developed using compounds of a single chemical class (e.g. halogenated phenols) on the assumption that such a congeneric series has a common mechanism of action.

Several theoretical studies on the prediction of the toxicity of halogenated phenols can be found in the literature [22-24]. However, several elaborated QSTR models do not fully meet the OECD (Organisation for Economic Co-operation and Development) principles for QSAR validation [25]. For instance, the external validation is not systematically carried out or the model descriptors are highly correlated making it hard to know the external predictive power. Furthermore, the Y-randomization and the applicability domain of the model are not constantly evaluated and discussed. On the other hand, the halogenated phenols are generally polar narcosis, so it exists a flow of electron between the molecule toxic and the organism. This electronic effect has been expressed in QSTR modelling by different descriptors such as the energy of the highest occupied molecular orbital E_{HOMO} [24], the lowest unoccupied molecular orbital E_{LUMO} [14,26], and the super-electrophilic-delocalizability A_{max} [26-28]. Since E_{HOMO} expresses the trend of system to furnish electrons, i.e. the nucleophilic character, this descriptor cannot be used to express the electrophilicity behaviour. On the other hand, E_{LUMO} and A_{max} are only approximate definitions of the electrophilicity concept. Thereby, these quantities are not suitable quantum chemical parameters to express the electrophilicity power. Recently, Parr et al. [29] proposed a precise and rigorous definition of the electrophilicity power, denoted ω, based on the energy lowering associated with a maximum amount of electron flow between two species. The Parr’s electrophilicity index is of great interest in analysing several and diverse areas of chemistry. Indeed, it has been shown that the electrophilicity possesses adequate information regarding structure, stability, reactivity, toxicity, bonding, interactions and dynamics [30]. The ω descriptor has been used for the study of the toxicity of chlorinated phenols by Chattaraj et al. [31]. However, it has been used alone and the penetration factor, namely the lipophilicity parameter log *P*, has not been taken into account. In the present work, both the electronic and transport factors would be considered. Two objectives were targeted for the present study: i) To elaborate predictive models for the toxicity of a series of halogenated phenols via *Tetrahymena Pyriformis* using a few number of descriptors that explain both the measured toxicity and the mode of action of these compounds. The most important advantage of the present contribution is the use of no consuming computational semi empirical methods to establish reliable and satisfactory QSTR models involving a few and meanigful molecular descriptors. ii) To evaluate the influence of semiempirical methods (AM1, PM3 and PM6) on the quality of the elaborated QSTR models for halogenated phenols.

MATERIALS AND METHODS

2.1. *Dataset and biological data*

The database consists of 45 halogenated and alkyl halogenated phenols taken from the reference [11] and listed in Table 1. Their biological data are considered to be of high quality since they refer to the same endpoint measured under the same experimental conditions.Toxicities were converted into the corresponding $-\log IGC_{50}$ values ($pIGC_{50}$), where IGC_{50} here means the millimolar concentration causing 50% inhibition of growth about halogenated phenols to *Tetrahymena Pyriformis.*

Table 1 Chemical abstracts service (CAS) number, chemical name, values of descriptors, observed and predicted toxicity and residuals

Compound				AM1			PM3			PM6		
	CAS. N	Exp.Tox	logP	ω	Pred.Tox	Resid	ω	Pred.Tox	Resid	ω	Pred.Tox	Resid
4-fluorophenol	371-41-5	0.017	1.915	1.114	0.222	-0.205	1.165	0.269	-0.252	1.274	0.325	-0.308
2-chlorophenol	95-57-8	0.183	2.155	1.124	0.375	-0.192	1.119	0.350	-0.167	1.233	0.396	-0.213
2-bromophenol	95-56-7	0.33	2.355	1.160	0.549	-0.219	1.155	0.526	-0.196	1.261	0.549	-0.219
3-fluorophenol	372-20-3	0.381	1.915	1.162	0.301	0.080	1.205	0.327	0.054	1.245	0.280	0.101
2-chloro-5-methylphenol	615-74-7	0.393	2.654	1.113	0.640	-0.247	1.106	0.640	-0.247	1.153	0.549	-0.156
4-chlorophenol	106-48-9	0.545	2.485	1.105	0.532	0.013	1.107	0.537	0.008	1.255	0.612	-0.067
2-bromo-4-methylphenol	6627-55-0	0.599	2.854	1.141	0.800	-0.201	1.266	0.999	-0.400	1.223	0.768	-0.169
2,4-difluorophenol	367-27-1	0.604	1.947	1.286	0.524	0.080	1.346	0.553	0.051	1.419	0.565	0.039
2-chloro-4,5-dimethylphenol	1124-04-5	0.688	3.103	1.102	0.878	-0.190	1.116	0.933	-0.245	1.132	0.766	-0.078
4-chloro-2-methylphenol	1570-64-5	0.701	2.984	1.098	0.804	-0.103	1.103	0.840	-0.139	1.191	0.790	-0.089
2,6-dichlorophenol	87-65-0	0.735	2.627	1.272	0.889	-0.154	1.245	0.827	-0.092	1.412	0.932	-0.197
2,6-dichloro-4-fluorophenol	392-71-*2*	0.804	2.797	1.404	1.205	-0.401	1.394	1.150	-0.346	1.604	1.322	-0.518
3-chlorophenol	108-43-0	0.871	2.485	1.159	0.622	0.249	1.167	0.626	0.245	1.293	0.671	0.200
2,4-dichlorophenol	120-83-2	1.036	2.957	1.254	1.047	-0.011	1.231	1.010	0.026	1.426	1.138	-0.102
2,5-dichlorophenol	583-78-8	1.125	2.957	1.275	1.083	0.042	1.231	1.012	0.113	1.418	1.125	0.000
3-chloro-4-fluorophenol	2613-23-2	1.131	2.717	1.272	0.942	0.189	1.294	0.955	0.176	1.444	1.032	0.099
2,4,6-trichlorophenol	88-06-2	1.41	3.367	1.376	1.398	0.012	1.317	1.391	0.019	1.573	1.592	-0.182
4-bromo-2,6-dimethylphenol	2374-05-2	1.167	3.633	1.093	1.165	0.002	1.138	1.294	-0.127	1.143	1.077	0.090
2,3,5,6-tetrafluorophenol	769-39-1	1.167	2.068	1.547	1.219	-0.052	1.716	1.168	-0.001	1.687	1.045	0.122
4-chloro-3,5-dimethylphenol	88-04-0	1.201	3.483	1.139	1.156	0.045	1.078	1.113	0.088	1.087	0.908	0.293
2,3-dichlorophenol	576-24-9	1.276	2.837	1.270	1.006	0.270	1.228	0.933	0.343	1.402	1.034	0.242
4-bromo-6-chloro-2-methylphenol	7530-27-0	1.276	3.606	1.241	1.397	-0.121	1.261	1.457	-0.181	1.375	1.419	-0.143
2,4-dibromophenol	615-58-7	1.398	3.307	1.307	1.336	0.062	1.425	1.511	-0.113	1.444	1.359	0.039
Pentafluorophenol	771-61-9	1.638	2.213	1.825	1.572	0.066	1.882	1.499	0.139	1.836	1.356	0.282
3,4-dichlorophenol	95-77-2	1.745	3.167	1.251	1.162	0.583	1.215	1.118	0.627	1.407	1.224	0.521
4-bromo-2,6-dichlorophenol	3217-15-0	1.778	3.517	1.388	1.589	0.189	1.365	1.554	0.224	1.577	1.681	0.097
2,4,6-tribromophenol	118-79-6	2.03	3.917	1.442	1.907	0.123	1.530	2.042	-0.012	1.600	1.938	0.092
Pentachlorophenol	87-86-5	2.049	4.323	1.618	2.431	-0.382	1.475	2.214	-0.165	1.806	2.481	-0.432
2,4,5-trichlorophenol	95-95-4	2.097	3.577	1.399	1.641	0.456	1.333	1.544	0.553	1.572	1.706	0.391
2,3,5-trichlorophenol	933-78-8	2.373	3.577	1.410	1.661	0.712	1.328	1.537	0.836	1.572	1.707	0.666
3,4,5,6-tetrabromo-2-methylphenol	576-55-*6*	2.574	4.967	1.563	2.706	-0.132	1.613	2.814	-0.240	1.683	2.649	-0.075
Pentabromophenol	608-71-9	2.664	4.853	1.741	2.937	-0.273	1.710	2.886	-0.222	1.889	2.903	-0.239
3-iodophenol	626-02-8	1.119	2.895	1.196	0.885	-0.382	1.308	1.037	-0.165	1.586	1.280	-0.432
4-iodophenol	540-38-5	0.854	2.895	1.146	0.803	0.456	1.286	0.975	0.553	1.499	1.153	0.391

Test set												
2-fluorophenol	367-12-4	0.185	1.715	1.132	0.139	0.046	1.190	0.176	0.009	1.232	0.150	0.035
2,6-difluorophenol	28177-48-2	0.471	1.747	1.308	0.450	0.021	1.360	0.456	0.015	1.372	0.378	0.093
4-bromophenol	106-41-2	0.680	2.630	1.141	0.676	0.004	1.174	0.687	-0.007	1.266	0.691	-0.011
4-chloro-3-methylphenol	59-50-7	0.796	2.984	1.097	0.805	-0.009	1.098	0.777	0.019	1.164	0.727	0.069
4-chloro-3-ethylphenol	14143-32-9	1.081	3.513	1.094	1.102	-0.021	1.101	1.090	-0.009	1.171	1.021	0.060
3-bromophenol	591-20-8	1.145	2.635	1.198	0.774	0.371	1.206	0.738	0.407	1.321	0.777	0.368
4-bromo-3,5-dimethylphenol	7463-51-6	1.268	3.633	1.092	1.166	0.102	1.144	1.227	0.041	1.102	0.981	0.287
3,5-dichlorophenol	591-35-5	1.569	3.287	1.303	1.320	0.249	1.283	1.239	0.33	1.483	1.371	0.198
4-chloro--2-isopropyl-5-methylphenol	89-68-9	1.854	4.411	1.064	1.565	0.289	1.077	1.580	0.274	1.113	1.415	0.439
2,3,5,6-tetrachlorophenol	935-95-5	2.222	3.848	1.547	2.046	0.176	1.432	1.796	0.426	1.718	2.027	0.195
2,3,4,5-tetrachlorophenol	4901-51-3	2.712	4.058	1.484	2.060	0.652	1.371	1.825	0.887	1.664	2.058	0.654

2.2. *Geometry optimization and molecular descriptors*

The structures were drawn using chem-office package [32]. First, we carried out a preliminary molecular mechanics geometry optimization calculations for each compound of this study, and then AM1 [33], PM3[34], and PM6 [35] semi-empirical methods included in Gaussian 09 software [36] are used for the final geometry optimization. A large number of descriptors were calculated for each compound, representing structural, steric, electronic and electrostatic properties that may be related to the toxicity of halogenated phenols to *Tetrahymena pyriformis*. Logarithms of the octanol/water partition coefficient (log *P*) and other descriptors (for instance, polarzability, total positive charge, total absolute charge, surface area, total charge of halogenated atoms, total charge of carbons of the aromatic ring,…) were calculated using different software the ACD/Labs [37], Hyperchem[38] andMolinspiration [39] softwares. The log P values, were also taken from reference 11.

According to previous works [26-28, 24], the toxicity can be explained in terms of the electrophilicity power which has been expressed by A_{max}, the E_{LUMO} or the E_{HOMO}. Unfortunately, in our opinion all these definitions are not precise. However, recent studies show that the electrophilicity concept is more suitably defined within the conceptual density functional theory (CDFT). According to CDFT, the chemical potential and chemical hardness for the n-electron molecular system with total energy E and external potential are defined as the first and second derivatives of the energy with respect to n, respectively.

$$\mu = \frac{(E_{LUMO} + E_{HOMO})}{2} \qquad (1)$$

$$\eta = E_{LUMO} - E_{HOMO} \qquad (2)$$

Where E_{LUMO} is the lowest unoccupied molecular orbital's energy and E_{HOMO} is the highest occupied orbital's energy.

Using μ and η, Parr et al.[29] have defined the electrophilicity index, ω, which measures the propensity to absorb electrons and is defined as:

$$\omega = \frac{\mu^2}{2\eta} \qquad (3)$$

2.3. Statistical analysis

The multiple linear regression (MLR) was used to develop the QSAR models using the MINITAB (version15) software [40]. Testing the stability, predictive power and generalization ability of the models is a very important step in QSAR study. For the validation of predictive power of a QSAR model, two basic principles (internal and external validations) are mandatory.

2.4. Selection and validation of the best statistical model for predicting pIGC50

2.4.1. Cross-validation test.

Internal cross-validation is a popular method used to explore the reliability of statistical models. Based on this technique, a number of modified data sets are created by deleting in each case one molecule (leave-one-out) or many molecules (leave-many-out). For each data set, a model is developed, based on the utilized modeling technique. The model was evaluated by measuring its accuracy in predicting the responses of the remaining data. In the present study, the internal predictive capability of the model was evaluated by leave-one-out cross-validation (R^2_{cv}). A good R^2_{cv}indicates a good robustness and high internal predictive power of a QSAR model. However, recent studies of

Tropsha and co-workers [41], Gramatica et al.[42] indicate that there is no evident correlation between the value of R^2_{cv} and actual predictive power of a QSAR model, revealing that the $R^2_{cv}$is still insufficient for a reliable estimate of model's predictive power for all new chemicals.

2.4.2. Y-Randomization test

Y-randomization (randomization of response) is a widely used approach to establish model robustness. It consists of rebuilding the models using randomized activities (for instance, toxicity) of the training set and subsequent assessment of the model statistics. It is expected that models obtained from the training set with randomized activities should have significantly lower values of R^2_{cv} for the training set than the models built using training set with real activitiesor at least these models should not have satisfied some of the validation criteria defined in Eqs. 4-7 given below.If this condition is not satisfied, real models built for this training set are not reliable and should be discarded [43].

2.4. 3. Validation through the external validation

External validation is now a "must have" tool for evaluating the reliability of QSAR models [44]. In this procedure, typically the overall set is randomly divided into a training set and a test set. QSAR models were developed based on the training set and then were used to make predictions for the test set. In the present study, the toxicity data was sorted in ascending order, extracting one sample every three samples as test set, with training set retained.

According to Golbraikh andTropsha [45], a QSAR model has an acceptable predictive power if the following conditions were satisfied:

$$R^2 > 0.7 \qquad (4)$$

$$R^2_{cv} > 0.6 \qquad (5)$$

$$\frac{R^2 - R_0^2}{R^2} < 0.1 \quad \text{and } 0.85 \le k \le 1.15 \qquad (6a)$$

or

$$\frac{R^2 - R'^2_0}{R^2} < 0.1 \quad \text{and } 0.85 \le k' \le 1.15 \qquad (6b)$$

and

$$\left| R^2 - R_0^2 \right| \le 0.3 \,(7)$$

Where R^2 is the squared correlation coefficient between observed and predicted values for the test set; R^2_0 and k are the correlation coefficient and slopes of the linear regression between the observed and predicted values when intercept was set to zero. The predicted versus observed and observed versus predicted correlation coefficients and slopes are different and therefore the latter were designated as R'^2_0 and k', respectively.

2.5. Model applicability domain

In order to use a QSAR model for screening new compounds, its domain of application must be defined and predictions for only those compounds that fall into this domain may be considered reliable [46,47]. In this study, the leverage approach was used to visualize the applicability domain (AD) of the QSARs. Compounds with standardized residuals greater than 2.21 were identified as response outliers. The limit of structural outliers was determined by their critical hat values (h*) calculated by 3p/n, where p is the number of model variables plus one, and n is the number of compounds in the model. In this approach, the hat value of a particular compound was used as a measure to quantify the compound's distance from the structural space of a model and h>h* indicate that the compound in question is outside of the model's structural AD; thus, the prediction could be unreliable [42].

RESULTS AND DISCUSSION

The whole data set constituted by 45 halogenated phenols was divided into a training set formed by 34 compounds and test series formed by 11 compounds randomly chosen. Using the 'best subsets' method implemented in MINITAB, we built several models for both AM1, PM3, and PM6 semi-empirical methods. Next, we verify the non-collinearity of the descriptors appearing in each equation. If the descriptors in the MLR equation are highly correlated, the QSTR model is systematically rejected.

Since phenols in general and halogenated phenols specifically acting as polar narcosis [11-14], we have tested firstly the correlation between the measured toxicity and the octanol-water coefficient log *P*,which reflects the penetration of the toxicant into membrane lipids. The log P parameter has been calculated using several softwares. However, the values taken from the reference [11] are found to give the best simple linear regression (R^2 = 0.68, SD = 0.4). This result shows that the toxicity of this set of halogenated phenols can be partially explained solely by the log *P* descriptor. However, to improve the quality of the QSAR models, the inclusion of other descriptors is necessary. According to many studies of the literature, the majority of halogenated phenol act as polar-narcotics [11,13]. For this reason, the Parr's electrophilicity index, ω, has been calculated and used as a potential quantum chemistry descriptor. The two-parameter MLR models obtained using AM1, PM3, and PM6 semi-empirical methods are given in Eqs.(8-10); where n is the number of compounds included in the model, SD is the standard deviation of the regression, R^2 is the squared correlation coefficient, F is the Fischer ratio, R^2_{cv} is the square of the cross-validated correlation coefficient and P is the P-value.

AM1 method

$pIGC_{50} = - 2.69 + 1.65\ \omega + 0.57 \log P$ (8)

n= 34, R^2 = 0.87, R^2_{adj} = 0.86, R^2_{cv}= 0.84, SD = 0.26, F= 100.43, P=0.000

PM3 method

$pIGC_{50}= - 2.63 + 1.47\omega + 0.62 \log P$ (9)

n= 34, R^2= 0.84, R^2_{adj}= 0.83, R^2_{cv}= 0.81, SD = 0.28, F= 80.05, P =0.000

PM6 method

$pIGC_{50}= - 2.67 + 1.50\ \omega + 0.57 \log P$ (10)

n= 34, R^2 = 0.85, R^2_{adj}= 0.84, R^2_{cv}= 0.82, SD = 0.27, F= 90.20, P= 0.000

It turns out that a considerable improvement of the QSAR models is achieved by combining the log P parameter with the ω index. The three semi-empirical methods (AM1, PM3, PM6) gave satisfactory MLR models although the AM1 method seems to give the best statistical parameters. In Table 2, were reported the coefficient (Coef.), standard error of coefficients (SE Coef.), T-test, variance inflation factor (VIF) and the correlation coefficient (R_{cor}). The analysis of VIF values and R_{cor} shows that there is no correlation (collinearity) between the two descriptorsω and log *P*.

Table 2 Correlation coefficients and VIF values among the variables.

Predictor	Coef.	SE Coef.	T-test	VIF	R_{cor}
Constant	-2.69	0.31	-8.73		
ω	1.65	0.25	6.57	1.19	0.39
log *P*	0.57	0.06	8.91	1.19	0.39

The analysis of the predicted and the standardized residuals shows the existence of two outliers in the training set with standardized residual greater than 2.2 units of toxicity for the three models given in Eqs.(8-10). These outlier compounds are 2, 3, 5-trichlorophenol number (no. 25) and 3,4-dichlorophenol number (no. 30). For which the predicted toxicity are considerably less than the measured toxicity. After the elimination of these two outliers from the training set, the quality of the MLR models, given in Esq.(11-13) is remarkably improved.

AM1 method

$pIGC_{50}= - 2.64 + 1.64\ \omega + 0.54 \log P$ (11)

n= 30, R^2 = 0.91, R^2_{adj} = 0.91, R^2_{cv} = 0.90, SD = 0.20, F=151.22, P=0.000

PM3 method

$pIGC_{50}= - 2.67+ 1.55\ \omega + 0.58 \log P$ (12)

n= 30, R^2 = 0.90, R^2_{adj} = 0.89, R^2_{cv} = 0.89, SD = 0.21, F=132.92, P=0.000

PM6 method

$pIGC_{50}= - 2.60 + 1.46\ \omega + 0.54 \log P$ (13)

n= 30, R^2 = 0.89, R^2_{adj} = 0.89, R^2_{cv} = 0.86, SD = 0.22, F=121.20, P=0.000

Internal cross-validation

The internal stability of the established models to the inclusion/exclusion of compounds is measured by the correlation coefficient and standard deviation of the cross-validation. The statistics of leave one out cross-validation might be considered as a good measurement of the predictability of the models. The high values of the regression coefficients R^2_{cv} of leave-one-out cross-validation and the small values of standard deviations SDs (see Eqs. 11-13) proved the predictive power and the internal stability of the elaborated models.

Y-randomization

The QSAR models must be subjected to Y-randomization to ensure that the developed relationships are not 'chance' correlations. In this technique, only the dependent variable Y is randomly re-ordered while the independent variables are left untouched and a new fit (R^2_r) is obtained for the distorted relationship. In the present study, this procedure was repeated many times for each proposed QSAR and the mean R^2_r was reported for each model. In this so-called, model randomization, the resulting models were expected to have very lower squared correlation coefficient (R^2_r) compared to the original relationship (R^2) since the link between the structure and toxicity is served [25]. The results of the randomisation for the ten first iterations are presented in Table 3. It was found that all values of R^2_r of the randomized models are lower than the corresponding R^2 of the non-randomized (i.e. original) model. This finding indicated that the obtained relationships are not due to 'chance'.

Table3 R_r^2 and R^2_{cv} values after several Y-randomizations

Iteration	R_r^2(Eq.11)	R^2_{cv}(Eq.11)	R_r^2(Eq.12)	R^2_{cv}(Eq.12)	R_r^2(Eq.13)	R^2_{cv}(Eq.13)
1	0.040	0.000	0.031	0.000	0.020	0.000
2	0.078	0.000	0.016	0.000	0.027	0.000
3	0.082	0.000	0.008	0.000	0.037	0.000
4	0.034	0.000	0.124	0.000	0.003	0.000
5	0.138	0.000	0.041	0.000	0.200	0.077
6	0.213	0.000	0.000	0.000	0.010	0.000
7	0.073	0.000	0.012	0.000	0.044	0.000
8	0.132	0.000	0.038	0.000	0.230	0.000
9	0.079	0.000	0.005	0.000	0.006	0.000
10	0.019	0.000	0.139	0.000	0.104	0.000

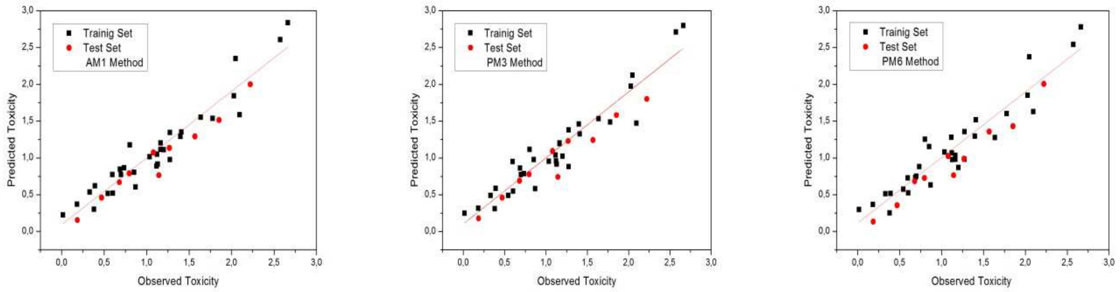
External cross-validation

It is well-known that the internal cross-validation is not sufficient to check the predictive power of a QSAR model and an external validation using a test series is necessary. All the MLR models given in Eqs. 8-13 were used to predict the toxicity for an external test set constituted by 11 halogenated phenols randomly chosen. The analysis of the residuals shows that the 2,3,4,5-tetrachlorophenol is an outlier compound and it is eliminated systematically. The results of the external validation using the remained ten molecules of the test set are presented in Table 4. As mentioned in section **2.3.2,** the predictive ability of a QSAR model can be verified using Tropsha's criteria (Esq. (4-7)). These results apparently show the good predictive ability of the MLR models only with AM1 and PM3 methods. For the MLR model obtained with the PM6 method, the Tropsha criteria are not verified since the k value is less than 0.85. At the end of the rigorous validation (internal and external) process, we are convinced that the proposed models presented by Eqs. (8-12) are stable and predictive. Moreover, the two descriptors, namely, log *P* and ω are not correlated.

Table 4 Internal and external validation of the QSARs models

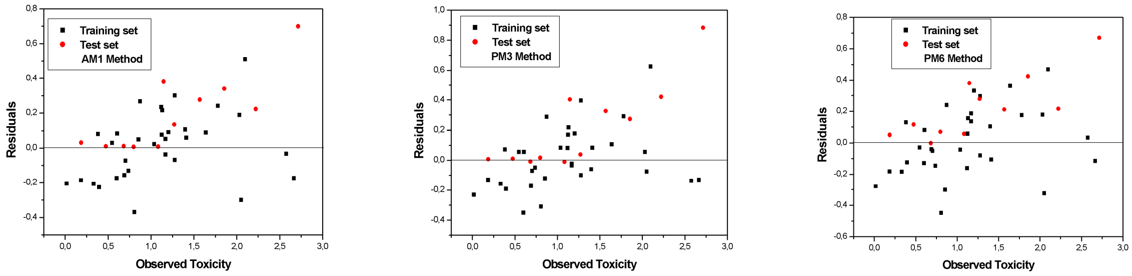
	Training set				**Test set**					**Tropsha' creteria**			
	n	**R^2**	**R^2_{cv}**	**SD**		**n**	**R^2**	**SD**		**R^2_0**	**k**	**$R^2 - R_0^2\ /R^2$**	**$\lvert R^2 - R_0^2\rvert$**
AM1													
Eq.(8)	34	0.86	0.85	0.26		10	0.96	0.11		0.96	0.90	0.00	0.00
Eq.(11)	32	0.91	0.90	0.20		10	0.96	0.11		0.96	0.86	0.00	0.00
PM3													
Eq.(9)	34	0.84	0.81	0.28		10	0.93	0.14		0.92	0.89	0.010	0.01
Eq.(12)	32	0.90	0.89	0.21		10	0.94	0.13		0.93	0.85	0.010	0.01
PM6													
Eq.(10)	34	0.85	0.82	0.27		10	0.96	0.12		0.96	0.86	0.000	0.00
Eq.(13)	32	0.89	0.86	0.22		10	0.96	0.11		0.96	0.84	0.000	0.00

The plots of the predicted versus experimental toxicity for both training and test sets of the models (11-13) are presented in figures (1a-1c)respectively.



Figs. (1a-1c)Predicted versus Observed Toxicity using Eqs (11-13)

Figures (2a-2c) present the plot of the residuals against the experimental values for the three models # 11-13. As most of the calculated residuals are distributed on two sides of the zero line, we can conclude that there is no systematic error in the development of the present models.



Figs. (2a-2c) Residuals versus Observed Toxicity using Eqs.(11-13)

Applicability domain

The goal of any QSAR is to develop reliable models that provide accurate predictions for as many chemical structures as possible in the universe, particularly for those that have not been tested or for which reliable experimental data is still not available. To this end, however, QSAR models must always be verified for their applicability with regard to chemical domain. In order to produce predicted data that can be considered reliable only for too structurally similar chemicals. A simple method to investigate the applicability domain for a prediction model is to carry out a leverage plot. The kind of leverage plots (plotting standardized residuals versus leverage of training compounds) for the best MLR model (Eq. 11), given in Figure 3, allows a graphical detection of both the outliers and the influential chemicals in a model. As observed in Figure 3, two of the data points (no. 1 and 10) moderately exceed critical leverage (h> h*=0.28). Both points can be kept in the model, but caution should be taken if similar compounds are predicted. The applicability of the model can be assessed with the descriptor ranges, minimum and maximum values for the modelled set of compounds, given in Table 1. Those ranges can be used while predicting unknown compounds.

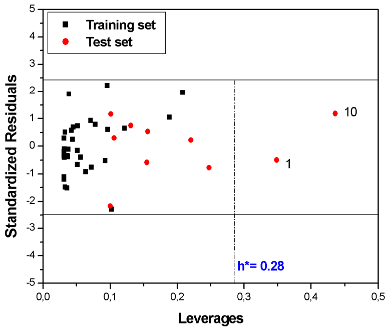


Fig.3 Standardized residuals versus leverages using Eq.11

Discussion of the toxicity mechanism

In the best MLR model given by Eq.(11), the main factors that could influence the toxicity are the hydrophobicity parameter, log *P*, and the Parr's electrophilicity index,ω, calculated using the AM1 method. The analysis of the T-test values (Table 3) and the standardized coefficient of each descriptor shows that the log *P* descriptor has the higher T-test and standardized coefficient values. Consequently, thehydrophobicity factor, as expressed by log *P* with a positive coefficient, is useful to describe transport to the site of action. So, if the compound has a high log *P* value, it will have good lipid solubility, and it can diffuse easily the cell membrane and concentrate on organisms, leading to an increase of the toxicity of the molecule. The contribution of electronic effects as expressed by the Parr's electrophilicity parameter,ω,is also important to predict the toxicity of the halogenated phenols. The positive coefficient of ω indicates that the increases of the electrophilicity power of a compound leads to the increase of its toxicity. Therefore, we can conclude that this descriptor is more appropriate to describe the electrophilic ability of halogenated phenols comparing with previous descriptors (E_{LUMO}, E_{HOMO} or A_{max}). On the other hand, halogenated phenols exhibit a polar narcosis mechanism so there is a non-covalent interaction with the lipid component. Classically, polar narcotic chemicals have been modelled in the framework of response-surface approach in the form of the two-parameter model including both transport and electronic effects [48]. In this approach, one independent variable, captures uptake of the chemical into the biophase, so called penetration characteristics of molecular structure. Another independent variable captures interaction with the site of action, i.e. electronic effects. The simple QSAR models, described in this work, combine both transport and electronic factors and explain adequately the polar narcosis mechanism of halogenated phenols.

CONCLUSION

In the present study, several QSTR models for the estimation of the toxicity of 45 halogenated phenols have been established using the MLR method. The models are constructed by the combination of the hydrophobicity parameter and the Parr's electrophilicity index calculated using both AM1, PM3, and PM6 semi-empirical methods. It turns out that AM1 MLR equation gives the best statistic parameters (R^2= 0.91, R^2_{CV}= 0.90, SD=0.13). Moreover, the elaborated MLR model is found to have good stability, robustness and high predictive power when verified by both internal and external validation and Y-randomization. The developed QSTR model shows that the toxicity increases with the increase of the log *P* parameter which explains the penetration of the halogenated phenols into *TetrahymenaPyriformis*cells. The toxicity of the compounds is also increased by the raise of the electrophilicity power of the molecules which explains the importance of the charge transfer between the toxicant, acting as an electrophile, and the living cell behaving as nucleophile. The present study put in evidence the relevance of the Parr's electrophilicity index in the rationalization of toxicity mechanism of halogenated phenols.

REFERENCES

[1] E Papa; F Villa; P Gramatica, *J. Chem. Inf. Model*, **2005**, 45, 1256-1266
[2] JD Wallker, *J. Mol. Struct-Theochem*, **2003**, 622, 167- 184
[3] S P Bradbury; C L Russom; G T Ankley; TW Schultz; JD Walker, *Environ. Toxicol. Chem*, **2003**, 228, 1789-1798
[4] European Commission. White Paper on a strategy for a future Community Policy for Chemicals, http://europa.eu.int:/comm/enterprise/reach/
[5] M W Toussaint; T R Shedd; W H Shalie; G R Leather, *Environ Toxicol Chem*, **1995**, 14, 907-915
[6] H Kubiniy, *Quantstrut-Act Rel*, **2002**, 21, 348-356
[7] http://www.epa.gov/nrmrl/std/qsar/qsar.html
[8] J Michałowicz; W.Duda, *Polish. J. Environ. Stud*, **2007**, 16, 347-362
[9] G M Della; P Monaco; G Pinto; A Pollio; L Previtera, F Temussi, *Bull. Environ Contam. Toxicol*, **2001**, 67,352-359
[10] R Garg; S Kapur, C Hansch, *Med. Res. Rev*, **2001**, 21,73-82
[11] T W Schultz; AP Bearden; J S Jaworska, *SAR&QSAR. Environ. Res*, **1996**, 5, 99-112
[12] A O Aptula; T I Netzeva; I V Valkova, M T D Cronin; T W Schultz; G Schuurmann Quant. *Struct-Act.Relat*, **2002**, 21, 12-22.
[13] M T D Cronin; A OAptula; J C Duffy; T I Netzeva; P H Rowe; I VValkova; T W Schultz, *Chemosphere*, **2002**, 49, 1201-1221
[14] S J Enoch, M T D Cronin, T W Schultz, J C Madden, *Chemosphere*, **2008**, 71, 1225–1232.
[15] C Maria, E G Guimarães, D G M Silva, P M Freitas, *Chemo. Int. Lab.Syst*, **2014**, 134, 53–57.
[16] F A Pasha; H KSrivestava, P P Singh, *Bioorg. Med. Chem*, **2005**, 13, 6823-6829.
[17] M DErtürka; M S Türker; A M Novicb; N Minovskib, *J. Mol. Graphics. Model*, **2012**, 38, 90–100.
[18] CODESSA PRO, University of Florida, www.codessa-pro.com.
[19] Dragon, TALETE, Italy, www.talete.mi.it/products/dragon_description.htm.
[20] M Erturka; S M Turker, *Ecotox. Environ. Saf*, **2013**, 90, 61–68.
[21] T W Schultz; D T Lin, S K Wesley, *Quality. Assur. Good.Pract.Regul. Law*, **1992**, 2, 132-143.
[22] J H Xing; Y T Zhang; *Comput. App. Chem*, **2007**, 24, 87-90.

[23] G Hea; L Fenga; H Chen, Procedia. *Engineering*, **2012**, 43, 204 – 209.
[24] A Ousaa1; B Elidrissi1; M Ghamali1; S Chtita1; M Bouachrine and T Lakhlifi, *J. Comput. Methods Mol. Des*, **2014**, 4, 10-18.
[25] http://www.oecd.org/document/23/0,2340,en_2649_201185_33957015_1_1_1_1,00.html
[26] M T D Cronin; N Manga; J R Seward; G D Sinks and T W Schultz, *Chem. Res. Toxicol*, **2001**, 14, 1498-1505.
[27] M T D Cronin; B W Gregory; TW Schultz, *Chem. Res. Toxicol*. **1998**, 11, 902-908
[28] S Ren, *Environ. Toxicol*, **2002**, 17, 119-127.
[29] R G Parr; L V Szentpaly and S Liu, *J. Am. Chem.Soc*, **1999**, 121, 1922-1924
[30] P K Chattaraj; S Giri and S Duley, *Chem. Rev*, **2011**, 111, PR43–PR75
[31] J Padmanabhan; R Parthasarathi; V Subramanian; P K Chattaraj, *Chem. Res. Toxicol*, **2006**, 19, 356-64
[32] Chemoffice **2004** (http://www.cambridgesoft.com)
[33] M J S Dewar; E G Zoebisch; E F Healy; J J P Stewart, *J. Am. Chem. Soc*, **1985**, 107, 3902–3909.
[34] J J P Stewart, *I. J Comput. Chem*, **1989**, 10, 209–220.
[35] J J P Stewart, *J. Mol. Model*, **2007**, 13, 1173–1213.
[36] Gaussian 09, Revision D.01, Gaussian, INC, Wallingford CT, **2009.**
[37] ACD/Labs, Release 12. **2009**, <http://www.acdlabs.com>
[38] HyperChem Release 7. HyperCube.Inc..http://www.hyper.com
[39] Molinspiration Cheminformatics, http: //www.molinspiration.com
[40] MINITAB, State College, PA Minitab, Inc. **2006.**
[41] A Tropsha; P Gramatica; V K Gombar, *QSAR. Comb. Sci*, **2003**, 22, 69-77.
[42] P Gramatica; E Giani; E Papa, *J. Mol. Graph. Model*, **2007**, 25, 755–766.
[43] L Zhang; H Zhu; T I Oprea; A Golbraikh; A Tropsha, *Pharm. Res*, **2008**, 25, 1902-1914.
[44] A Tropsha, *Mol. Inf.*, **2010**, 29, 476–488.
[45] A Golbraikh; A Tropsha, *Mol. Divers*, **2002**, 5, 231–243.
[46] L Eriksson; J Jaworska; A Worth; M D Cronin; R M McDowell, P Gramatica, *Environ. Health. Perspect*, **2003**, 111, 1361–1375.
[47] M, Martin; P Harten, D M Young; E N Muratov; A Golbraikh; H Zhu and A Tropsha, *J. Chem. Inf. Model*, **2012**, 52, 2570−2578.
[48]S D Dimitrov; O G Mekenyan; G D Sinks; TW Schultz, *J. Mol. Struct*, **2003**, 622, 63-70.