


Article

Estimation of the Toxicity of Different Substituted Aromatic Compounds to the Aquatic Ciliate *Tetrahymena pyriformis* by QSAR Approach

Feng Luan ^{1,*}, Ting Wang ¹, Lili Tang ¹, Shuang Zhang ¹ and M. Natália Dias Soeiro Cordeiro ² 

¹ College of Chemistry and Chemical Engineering, Yantai University, Yantai 264005, China; 18865557652@163.com (T.W.); tangtang20160216@sina.com (L.T.); zs20160903@163.com (S.Z.)

² LAQV/REQUIMTE, Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal; ncordeir@fc.up.pt

* Correspondence: fluan@sina.com; Tel: +86-0535-690-2063

Received: 19 March 2018; Accepted: 21 April 2018; Published: 24 April 2018



Abstract: Nowadays, quantitative structure–activity relationship (QSAR) methods have been widely performed to predict the toxicity of compounds to organisms due to their simplicity, ease of implementation, and low hazards. In this study, to estimate the toxicities of substituted aromatic compounds to *Tetrahymena pyriformis*, the QSAR models were established by the multiple linear regression (MLR) and radial basis function neural network (RBFNN). Unlike other QSAR studies, according to the difference of functional groups ($-\text{NO}_2$, $-X$), the whole dataset was divided into three groups and further modeled separately. The statistical characteristics for the models are obtained as the following: MLR: $n = 36$, $R^2 = 0.829$, RMS (root mean square) = 0.192, RBFNN: $n = 36$, $R^2 = 0.843$, RMS = 0.167 for Group 1; MLR: $n = 60$, $R^2 = 0.803$, RMS = 0.222, RBFNN: $n = 60$, $R^2 = 0.821$, RMS = 0.193 for Group 2; MLR: $n = 31$, $R^2 = 0.852$, RMS = 0.192; RBFNN: $n = 31$, $R^2 = 0.885$, RMS = 0.163 for Group 3, respectively. The results were within the acceptable range, and the models were found to be statistically robust with high external predictivity. Moreover, the models also gave some insight on those characteristics of the structures that most affect the toxicity.

Keywords: substituted aromatic compounds; toxicity; quantitative structure–activity relationship (QSAR); multiple linear regression (MLR); radial basis function neural network (RBFNN)

1. Introduction

With the rapid development of science and technology, tens of thousands of new chemicals are synthesized and widely used in all walks of life every day. However, as we all know, if chemicals are used or handled incorrectly, they may enter the aquatic environment or bio-accumulate in the food chain, where they may adversely impact the people, ultimately. One of the current interests in medicinal chemistry, environmental sciences, and especially for toxicology, is to rank and establish the chemical substances with respect to their potential hazardous effects on humans, wildlife, and aquatic flora and fauna [1]. Among the vast organic matter, it is noteworthy that the substituted aromatic compounds [2–8] occupy important positions, since they are produced in large quantities and released into the environment as a result of their wide use in agriculture and industry, and are widely distributed in air, natural water, waste water, soil, sediment, and living organics [9,10]. In addition, recent studies have proved that the substituted aromatic compounds are also a kind of biotoxic environmental pollutant, and even have the effects of carcinogenesis and gene mutation on organisms [10,11]. Therefore, studies on the properties of substituted aromatics have important significance.

Up to now, both experimental [12–15] and theoretical methods [16,17] have been used to evaluate kinds of substituted aromatic compounds for their different toxicities. Also, it is well known that

the theoretical predictions of properties or activities by quantitative structure–activity relationship (QSAR) studies have been widely adopted and applied since the 90s, because of their advantages, such as rapidness, easiness, sensitiveness, and cheapness [11]. The QSAR method has been widely applied in different fields, including physical chemistry, pharmaceutical chemistry, environmental chemistry, toxicology, and other research fields [18]. It has been proven that the use of QSAR modeling for toxicological predictions would help to determine the potential adverse effects of chemical entities in risk assessment.

For a long time, a lot of meaningful research focusing on the toxicity of substituted aromatic compounds by QSAR approach have been carried out. In 1982, Schultz et al. tried to perform the QSAR study between the cellular response to *Tetrahymena pyriformis* and molecular connectivity indexes for a series of 24 mono- and dinitrogen heterocyclic compounds. In this study, the authors established a better model than before, and pointed out that toxicity increases with an increase in the number of atoms and degree of methylation per compound, and that toxicity decreases with an increase in nitrogen substitution [1]. In 1998, Cronin et al. established several QSAR models focusing on a dataset of 42 alkyl- and halogen-substituted nitro- and dinitrobenzenes to *Tetrahymena pyriformis* [19]. They found that the nitrobenzenes were thought to elicit their toxic response through multiple (and mixed) mechanisms by one or two molecule descriptor models. In 2001, in order to compare the differences among kinds of QSAR model-building methods, Cronin and Schultz developed QSAR studies for the toxicity of 268 aromatic compounds in the *Tetrahymena pyriformis* growth inhibition assay [16]. In their study, they not only compared the influence of different descriptors on the models, but also the Bayesian regularized neural network (BRANN) and partial least-squares (PLS) analysis to build the models. In the following year, the same authors performed also the same study on a dataset of phenolic toxicity data to *Tetrahymena pyriformis* [17]. The above works gave us some guidelines or directions on how to build better models on the toxicities to this group of compounds. Netzeva et al. developed relative simple QSARs models (one or two descriptors) for the acute toxicity of a dataset of 77 aromatic aldehydes to the ciliate *Tetrahymena pyriformis* using mechanistically interpretable descriptors [20]. They revealed that the octanol/water partition coefficient ($\log K_{OW}$) is the most important descriptor, and the models would be improved by using another electronic descriptor. Roy et al. performed a QSAR studies on the toxic potency to *Tetrahymena pyriformis* of a dataset of 174 aromatic compounds (phenols, nitrobenzenes, and benzonitriles) using electrophilicity index [21]. In this study, the compounds in the dataset were divided into the electron donor and acceptor group, and they stated that electrophilicity indices, along with the total Hartree–Fock energy, can be used to build the model perfectly. Later, the performances of the linear and nonlinear models were estimated by Devillers et al. using a structurally heterogeneous set of 200 phenol derivatives on *Tetrahymena pyriformis*. In this study, the authors pointed out the superiority of the nonlinear methods over the linear ones to find complex structure–toxicity relationships among large sets of structurally diverse chemicals [22]. Tetko et al. gave studies on the applicability domain and the influence of the overfitting in the QSAR model building process by the toxicity dataset against *Tetrahymena pyriformis* [23]. The hierarchical technology for QSAR was performed using 95 diverse nitroaromatic compounds against the ciliate *Tetrahymena pyriformis* [24]. Zarei et al. developed a model for the prediction of the toxicity of 268 substituted benzene compounds including phenols, monosubstituted nitrobenzenes, multiply substituted nitrobenzenes and benzonitriles to *T. pyriformis* using bee algorithm (BA) for selecting descriptors and adaptive neuro-fuzzy inference system (ANFIS) for building model [25]. A molecular structural characterization (MSC) method named molecular vertexes correlative index (MVCI) was successfully used to describe the structures of 30 substituted aromatic compounds, and the results suggested good stability and predictability of the QSAR models [26]. Comparative molecular field (CoMFA), molecular similarity index analysis (CoMSIA), and density functional theory (DFT) methods were used to establish QSAR models for analyzing and predicting the toxicities of 31 substituted thiophenols [27]. And later, Salahinejad et al. also used the CoMFA, CoMSIA, and VolSurf techniques to develop valid and predictive models

able to estimate the toxicity of substituted benzenes toward *T. pyriformis*. In the paper, they confirmed that in addition to hydrophobic effects, electrostatic and H-bonding interactions also play important roles in the toxicity of substituted benzenes, as well as that the information obtained from CoMFA and CoMSIA 3-D contour maps could be useful to explain the toxicity mechanism of substituted benzenes [28]. The linear (MLR) and nonlinear statistical (RBFNN) methods were used by us to build a reliable, credible, and fast QSAR model for the prediction of mixture toxicity of non-polar narcotic chemicals, including 9 PFCAs, 12 alcohols, and 8 chlorobenzenes and bromobenzenes. The predictive values are in good agreement with the experimental ones [18]. In the same way, recursive neural networks (RNN) and multiple linear regression (MLR) methods were also employed to build models for prediction of the toxicity values of 69 benzene derivatives, both methods provided good results as compared to other studies available in the literature [29]. To build a reliable and predictive QSAR model, a genetic algorithm along with partial least square (GA-PLS) was employed to select the optimal subset of descriptors that significantly contribute to the toxicity of 45 nitrobenzene derivatives to *Tetrahymena pyriformis* [30].

The goal of present study was to develop reliable and predictive QSAR models using both MLR and RBFNN methods to identify and predict the acute toxicity (the 50% growth inhibitory concentration IGC_{50}) of substituted aromatic compounds to the aquatic ciliate *Tetrahymena pyriformis*. For this purpose, the whole dataset was divided into three groups with respect to the important function group of the substituted aromatic compounds such as $-NO_2$, $-X$ etc. They were Group 1: Compounds with NO_2 group, etc. (46 compounds); Group 2: Compounds with $-X$, etc. (75 compounds); Group 3: Compounds with both $-NO_2$ and $-X$, etc. (39 compounds). In so doing, different accurate models were built to evaluate the toxicities of these aromatic compounds.

2. Materials and Methods

2.1. Datasets

For the aromatic compounds, Wei et al. have mentioned that the order of the contribution of the special substituents to the toxicity of the aromatic compound is: $-NO_2 > -Cl > -CH_3 > -NH_2 > -OH$ [31]. Based on the dataset given by Schultz et al. [32], we selected the typical compounds containing the most influential functional groups ($-NO_2$ and $-X$), and divided them into three subgroups. Group 1 includes 46 compounds whose chemical structures have the functional group $-NO_2$ without $-X$. Among them, 36 compounds were substituted by a $-NO_2$ and 10 compounds were substituted by two $-NO_2$. Group 2 contains 75 compounds which have functional groups $-X$ without $-NO_2$. Among them, the 54, 16, and 5 compounds were replaced by one, two, or three functional groups $-X$, respectively. Group 3 contains 39 compounds, in which both the $-NO_2$ and $-X$ functional groups are included, and the total number of substituents for $-NO_2$ and $-X$ is not more than 3.

In this study, compounds in each group were randomly divided into two subsets. One called training set was used to build a model, and there were 36, 60, 31 compounds in the training set for Group 1, 2, 3, respectively. The remaining compounds were used to verify the robustness and feasibility of the model as a test set which includes 10, 15, and 8 for the corresponding groups, respectively. The CAS number, name, and toxicity ($-\log IGC_{50}$) of the above compounds are all listed in Table 1.

Table 1. The CAS number, name, experimental $-\log \text{IGC}_{50}$ values, predicted $-\log \text{IGC}_{50}$ values and their corresponding residual for the three groups of compounds.

No.	CAS	Name	Experimental $-\log \text{IGC}_{50}$	Predicted $-\log \text{IGC}_{50}$				Set
				MLR	Residual	RBFNN	Residual	
Group 1. Compounds with the functional group $-\text{NO}_2$.								
1 *	619-25-0	3-Nitrobenzyl alcohol	-0.22	0.58	0.80	0.01	0.23	T
2	91-23-6	2-Nitroanisole	-0.07	0.22	0.29	0.00	0.07	A
3	99-09-2	3-Nitroaniline	0.03	0.46	0.43	0.34	0.31	B
4	88-74-4	2-Nitroaniline	0.08	0.43	0.35	0.35	0.27	C
5	99-61-6	3-Nitrobenzaldehyde	0.11	0.27	0.16	0.16	0.05	D
6 *	98-95-3	Nitrobenzene	0.14	0.31	0.17	-0.10	-0.24	T
7	552-89-6	2-Nitrobenzaldehyde	0.17	0.25	0.08	0.19	0.02	A
8	555-16-8	4-Nitrobenzaldehyde	0.2	0.38	0.18	0.16	-0.04	B
9	88-72-2	2-Nitrotoluene	0.26	0.48	0.22	0.32	0.06	C
10	704-13-2	3-Hydroxy-4-nitrobenzaldehyde	0.27	0.57	0.30	0.39	0.12	D
11 *	121-89-1	30-Nitroacetophenone	0.32	0.47	0.15	-0.02	-0.34	T
12	42454-06-8	5-Hydroxy-2-nitrobenzaldehyde	0.33	0.54	0.21	0.42	0.09	A
13	89-62-3	4-Methyl-2-nitroaniline	0.37	0.62	0.25	0.55	0.18	B
14	619-50-1	Methyl-4-nitrobenzoate	0.39	0.70	0.31	0.46	0.07	C
15	99-08-1	3-Nitrotoluene	0.42	0.53	0.11	0.43	0.01	D
16 *	5292-45-5	Dimethyl nitroterephthalate	0.43	1.51	1.08	0.09	-0.34	T
17	619-24-9	3-Nitrobenzotrile	0.45	0.70	0.25	0.67	0.22	A
18	554-84-7	3-Nitrophenol	0.51	0.58	0.07	0.49	-0.02	B
19	83-41-0	1,2-Dimethyl-3-nitrobenzene	0.56	0.67	0.11	0.66	0.10	C
20	119-33-5	4-Methyl-2-nitrophenol	0.57	0.63	0.06	0.62	0.05	D
21 *	99-51-4	1,2-Dimethyl-4-nitrobenzene	0.59	0.74	0.15	0.31	-0.28	T
22	700-38-9	5-Methyl-2-nitrophenol	0.59	0.78	0.19	0.76	0.17	A
23	4920-77-8	3-Methyl-2-nitrophenol	0.61	0.64	0.03	0.51	-0.10	B
24	3011-34-5	4-Hydroxy-3-nitrobenzaldehyde	0.61	0.47	-0.14	0.45	-0.16	C
25	99-99-0	4-Nitrotoluene	0.65	0.54	-0.11	0.52	-0.13	D
26 *	5428-54-6	2-Methyl-5-nitrophenol	0.66	0.84	0.18	0.60	-0.06	T
27	601-89-8	2-Nitroresorcinol	0.66	0.63	-0.03	0.55	-0.11	A
28	88-75-5	2-Nitrophenol	0.67	0.43	-0.24	0.32	-0.35	B
29	99-77-4	Ethyl-4-nitrobenzoate	0.7	0.76	0.06	0.67	-0.03	C
30	555-03-3	3-Nitroanisole	0.71	0.66	-0.05	0.48	-0.23	D
31 *	97-02-9	2,4-Dinitroaniline	0.72	1.33	0.61	0.97	0.25	T
32	616-86-4	4-Ethoxy-2-nitroaniline	0.76	1.09	0.33	0.84	0.08	A
33	99-65-0	1,3-Dinitrobenzene	0.76	1.07	0.31	0.94	0.18	B
34	100-29-8	4-Nitrophenetole	0.83	0.98	0.15	0.84	0.01	C
35	573-56-8	2,6-Dinitrophenol	0.83	1.25	0.42	1.30	0.47	D
36 *	606-22-4	2,6-Dinitroaniline	0.84	1.30	0.46	0.80	-0.04	T
37	603-71-4	1,3,5-Trimethyl-2-nitrobenzene	0.86	0.92	0.06	0.73	-0.13	A
38	121-14-2	2,4-Dinitrotoluene	0.87	1.30	0.43	1.07	0.20	B
39	329-71-5	2,5-Dinitrophenol	1.04	1.50	0.46	0.94	-0.10	C
40	528-29-0	1,2-Dinitrobenzene	1.25	1.06	-0.19	0.89	-0.36	D
41 *	100-25-4	1,4-Dinitrobenzene	1.3	1.23	-0.07	1.33	0.03	T
42	86-00-0	2-Nitrobiphenyl	1.3	1.20	-0.10	1.03	-0.27	A
43	620-88-2	4-Nitrophenyl phenyl ether	1.58	1.71	0.13	1.59	0.01	B
44	69212-31-3	2-(Benzylthio)-3-nitropyridine	1.72	1.98	0.26	1.71	-0.01	C
45	534-52-1	4,6-Dinitro-2-methylphenol	1.73	1.61	-0.12	1.70	-0.03	D
46 *	4097-49-8	4-(<i>tert</i>)-Butyl-2,6-dinitrophenol	1.8	2.00	0.20	1.71	-0.09	T
Group 2. Compounds with the functional group $-\text{X}$.								
1 *	348-54-9	2-Fluoroaniline	-0.37	-0.20	0.17	-0.17	0.20	T
2	95-51-2	2-Chloroaniline	-0.17	0.03	0.20	-0.04	0.13	A
3	108-90-7	Chlorobenzene	-0.13	0.11	0.24	0.05	0.18	B
4	372-19-0	3-Fluoroaniline	-0.1	-0.27	-0.17	-0.23	-0.13	C
5	371-41-5	4-Fluorophenol	0.02	-0.12	-0.14	-0.14	-0.16	D
6 *	106-47-8	4-Chloroaniline	0.05	-0.01	-0.06	0.06	0.01	T
7	100-44-7	Benzyl chloride	0.06	0.31	0.25	0.24	0.18	A
8	108-86-1	Bromobenzene	0.08	0.18	0.10	0.15	0.07	B
9	18982-54-2	2-Bromobenzyl alcohol	0.1	0.32	0.22	0.24	0.14	C
10	95-88-5	4-Chlororesorcinol	0.13	0.50	0.37	0.48	0.35	D
11 *	156-41-2	2-(4-Chlorophenyl)-ethylamine	0.14	0.43	0.29	0.46	0.32	T
12	873-63-2	3-Chlorobenzyl alcohol	0.15	0.56	0.41	0.55	0.40	A
13	104-86-9	4-Chlorobenzylamine	0.16	0.28	0.12	0.27	0.11	B
14	615-65-6	2-Chloro-4-methylaniline	0.18	0.46	0.28	0.41	0.23	C
15	367-12-4	2-Fluorophenol	0.19	0.17	-0.02	0.09	-0.10	D
16 *	108-42-9	3-Chloroaniline	0.22	-0.07	-0.29	0.04	-0.18	T
17	873-76-7	4-Chlorobenzyl alcohol	0.25	0.33	0.08	0.26	0.01	A

Table 1. Cont.

No.	CAS	Name	Experimental −log IGC ₅₀	Predicted −log IGC ₅₀				Set
				MLR	Residual	RBFNN	Residual	
18	1875-88-3	4-Chlorophenethyl alcohol	0.32	0.48	0.16	0.43	0.11	B
19	95-56-7	2-Bromophenol	0.33	0.50	0.17	0.45	0.12	C
20	95-69-2	4-Chloro-2-methylaniline	0.35	0.45	0.10	0.39	0.04	D
21 *	615-43-0	2-Iodoaniline	0.35	0.40	0.05	0.48	0.13	T
22	591-50-4	Iodobenzene	0.36	0.30	−0.06	0.32	−0.04	A
23	87-60-5	3-Chloro-2-methylaniline	0.38	0.35	−0.03	0.30	−0.08	B
24	95-74-9	3-Chloro-4-methylaniline	0.39	0.39	0.00	0.38	−0.01	C
25	104-88-1	4-Chlorobenzaldehyde	0.4	0.48	0.08	0.41	0.01	D
26 *	103-63-9	(2-Bromoethyl)-benzene	0.42	0.51	0.09	0.65	0.23	T
27	5922-60-1	2-Amino-5-chlorobenzonitrile	0.44	0.28	−0.16	0.22	−0.22	A
28	106-38-7	4-Bromotoluene	0.47	0.48	0.01	0.42	−0.05	B
29	95-79-4	5-Chloro-2-methylaniline	0.5	0.44	−0.06	0.40	−0.10	C
30	95-50-1	1,2-Dichlorobenzene	0.53	0.53	0.00	0.49	−0.04	D
31 *	106-48-9	4-Chlorophenol	0.54	0.06	−0.48	0.27	−0.27	T
32	615-74-7	2-Chloro-5-methylphenol	0.54	0.59	0.05	0.59	0.05	A
33	554-00-7	2,4-Dichloroaniline	0.56	0.49	−0.07	0.44	−0.12	B
34	95-82-9	2,5-Dichloroaniline	0.58	0.36	−0.22	0.29	−0.29	C
35	7120-43-6	5-Chloro-2-hydroxybenzamide	0.59	0.39	−0.20	0.43	−0.16	D
36 *	623-12-1	4-Chloroanisole	0.6	0.27	−0.33	0.36	−0.24	T
37	6627-55-0	2-Bromo-4-methylphenol	0.6	0.96	0.36	0.84	0.24	A
38	16532-79-9	4-Bromophenyl acetonitrile	0.6	0.66	0.06	0.63	0.03	B
39	2973-76-4	5-Bromovanillin	0.62	0.85	0.23	0.87	0.25	C
40	626-01-7	3-Iodoaniline	0.65	0.40	−0.25	0.35	−0.30	D
41 *	140-53-4	4-Chlorobenzyl cyanide	0.66	0.57	−0.09	0.72	0.06	T
42	1585-07-5	1-Bromo-4-ethylbenzene	0.67	0.92	0.25	0.94	0.27	A
43	106-37-6	1,4-Dibromobenzene	0.68	0.61	−0.07	0.58	−0.10	B
44	106-41-2	4-Bromophenol	0.68	0.48	−0.20	0.42	−0.26	C
45	1124-04-5	2-Chloro-4,5-dimethylphenol	0.69	0.90	0.21	0.83	0.14	D
46 *	1570-64-5	4-Chloro-2-methylphenol	0.7	0.54	−0.16	0.52	−0.18	T
47	626-43-7	3,5-Dichloroaniline	0.71	0.52	−0.19	0.49	−0.22	A
48	65262-96-6	3-Chloro-5-methoxyphenol	0.76	0.54	−0.22	0.49	−0.27	B
49	59-50-7	4-Chloro-3-methylphenol	0.8	0.49	−0.31	0.49	−0.31	C
50	2905-69-3	Methyl-2,5-dichlorobenzoate	0.81	1.13	0.32	1.13	0.32	D
51 *	14548-45-9	4-Bromophenyl-3-pyridyl ketone	0.82	1.25	0.43	1.20	0.38	T
52	540-38-5	4-Iodophenol	0.85	0.59	−0.26	0.56	−0.29	A
53	108-43-0	3-Chlorophenol	0.87	0.43	−0.44	0.37	−0.50	B
54	108-70-3	1,3,5-Trichlorobenzene	0.87	1.13	0.26	1.14	0.27	C
55	120-83-2	2,4-Dichlorophenol	1.04	0.89	−0.15	0.95	−0.09	D
56 *	874-42-0	2,4-Dichlorobenzaldehyde	1.04	0.97	−0.07	1.08	0.04	T
57	95-75-0	3,4-Dichlorotoluene	1.07	1.02	−0.05	0.95	−0.12	A
58	120-82-1	1,2,4-Trichlorobenzene	1.08	1.10	0.02	1.16	0.08	B
59	14143-32-9	4-Chloro-3-ethylphenol	1.08	0.70	−0.38	0.74	−0.34	C
60	2374-05-2	4-Bromo-2,6-dimethylphenol	1.16	1.04	−0.12	0.95	−0.21	D
61 *	1689-84-5	3,5-Dibromo-4-hydroxybenzoxonitrile	1.16	1.54	0.38	1.29	0.13	T
62	88-04-0	4-Chloro-3,5-dimethylphenol	1.2	0.70	−0.50	0.74	−0.46	A
63	90-90-4	4-Bromobenzophenone	1.26	1.37	0.11	1.36	0.10	B
64	7530-27-0	4-Bromo-6-chloro-2-cresol	1.28	1.37	0.09	1.34	0.06	C
65	636-30-6	2,4,5-Trichloroaniline	1.3	1.18	−0.12	1.31	0.01	D
66 *	5798-75-4	Ethyl-4-bromobenzoate	1.33	1.16	−0.17	1.23	−0.10	T
67	13608-87-2	20,30,40-Trichloroacetophenone	1.34	1.44	0.10	1.31	−0.03	A
68	615-58-7	2,4-Dibromophenol	1.4	1.07	−0.33	1.15	−0.25	B
69	88-06-2	2,4,6-Trichlorophenol	1.41	1.62	0.21	1.47	0.06	C
70	134-85-0	4-Chlorobenzophenone	1.5	1.30	−0.20	1.34	−0.16	D
71 *	1016-78-0	3-Chlorobenzophenone	1.55	1.27	−0.28	1.20	−0.35	T
72	90-60-8	3,5-Dichlorosalicylaldehyde	1.55	1.49	−0.06	1.52	−0.03	A
73	591-35-5	3,5-Dichlorophenol	1.56	1.28	−0.28	1.22	−0.34	B
74	90-59-5	3,5-Dibromosalicylaldehyde	1.65	1.55	−0.10	1.61	−0.04	C
75	456-47-3	3-Fluorobenzyl alcohol	−0.39	−0.05	0.34	−0.09	0.30	D
Group 3. Compounds with both −NO₂ and −X.								
1 *	89-59-8	4-Chloro-2-nitrotoluene	0.43	1.08	0.65	0.75	0.32	T
2	585-79-5	1-Bromo-3-nitrobenzene	0.53	0.48	−0.05	0.54	0.01	A
3	7149-70-4	2-Bromo-5-nitrotoluene	0.68	0.99	0.31	1.09	0.41	B
4	100-14-1	4-Nitrobenzyl chloride	0.68	0.70	0.02	0.71	0.03	C
5	610-78-6	4-Chloro-3-nitrophenol	0.73	1.08	0.35	1.03	0.30	D
6 *	7147-89-9	4-Chloro-6-nitro-3-cresol	0.73	1.20	0.47	1.12	0.39	T
7	364-74-9	2,5-Difluoronitrobenzene	0.75	0.66	−0.09	0.85	0.10	A

Table 1. Cont.

No.	CAS	Name	Experimental −log IGC ₅₀	Predicted −log IGC ₅₀				Set
				MLR	Residual	RBFNN	Residual	
8	6361-21-3	2-Chloro-5-nitrobenzaldehyde	0.75	0.86	0.11	0.75	0.00	B
9	83-42-1	2-Chloro-6-nitrotoluene	0.75	0.55	−0.20	0.53	−0.22	C
10	88-73-3	1-Chloro-2-nitrobenzene	0.75	0.69	−0.06	0.72	−0.03	D
11 *	121-73-3	1-Chloro-3-nitrobenzene	0.8	0.69	−0.11	0.83	0.03	T
12	87-65-0	2,6-Dichlorophenol	0.82	0.83	0.01	0.81	−0.01	A
13	121-87-9	2-Chloro-4-nitroaniline	0.82	0.79	−0.03	0.77	−0.05	B
14	577-19-5	1-Bromo-2-nitrobenzene	0.99	0.71	−0.28	0.80	−0.19	C
15	2973-19-5	2-Chloromethyl-4-nitrophenol	1.03	1.03	0.00	1.00	−0.03	D
16 *	78056-39-0	4,5-Difluoro-2-nitroaniline	1.06	1.13	0.07	0.86	−0.20	T
17	350-30-1	3-Chloro-4-fluoronitrobenzene	1.07	0.86	−0.21	0.93	−0.14	A
18	42087-80-9	Methyl-4-chloro-2-nitrobenzoate	1.09	1.25	0.16	1.30	0.21	B
19	611-06-3	2,4-Dichloronitrobenzene	1.12	1.23	0.11	1.27	0.15	C
20	51-28-5	2,4-Dinitrophenol	1.13	1.17	0.04	1.12	−0.01	D
21 *	3209-22-1	2,3-Dichloronitrobenzene	1.13	0.83	−0.30	1.04	−0.09	T
22	3819-88-3	1-Fluoro-3-iodo-5-nitrobenzene	1.16	0.90	−0.26	0.97	−0.19	A
23	618-62-2	3,5-Dichloronitrobenzene	1.16	0.96	−0.20	1.08	−0.08	B
24	89-61-2	2,5-Dichloronitrobenzene	1.18	1.19	0.01	1.24	0.06	C
25	99-54-7	3,4-Dichloronitrobenzene	1.24	0.87	−0.37	0.92	−0.32	D
26 *	2683-43-4	2,4-Dichloro-6-nitroaniline	1.26	1.28	0.02	1.14	−0.12	T
27	3460-18-2	2,5-Dibromonitrobenzene	1.27	1.27	0.00	1.31	0.04	A
28	827-23-6	2,4-Dibromo-6-nitroaniline	1.37	1.40	0.03	1.45	0.08	B
29	6641-64-1	4,5-Dichloro-2-nitroaniline	1.62	1.31	−0.31	1.38	−0.24	C
30	609-89-2	2,4-Chloro-6-nitrophenol	1.63	1.46	−0.17	1.50	−0.13	D
31 *	305-85-1	2,6-Iodo-4-nitrophenol	1.66	1.40	−0.26	1.47	−0.19	T
32	3531-19-9	6-Chloro-2,4-dinitroaniline	1.71	1.53	−0.18	1.64	−0.07	A
33	1817-73-8	2-Bromo-4,6-dinitroaniline	1.75	1.63	−0.12	1.73	−0.02	B
34	97-00-7	1-Chloro-2,4-dinitrobenzene	1.81	1.82	0.01	1.88	0.07	C
35	709-49-9	2,4-Dinitro-1-iodobenzene	2.12	1.78	−0.34	2.02	−0.10	D
36 *	70-34-8	2,4-Dinitro-1-fluorobenzene	2.16	1.67	−0.49	0.67	−1.49	T
37	350-46-9	1-Fluoro-4-nitrobenzene	0.1	0.21	0.11	0.32	0.22	A
38	1493-27-2	1-Fluoro-2-nitrobenzene	0.23	−0.07	−0.30	0.23	0.00	B
39	100-00-5	1-Chloro-4-nitrobenzene	0.33	0.46	0.13	0.51	0.18	C

* represents the compound in the test set.

2.2. Molecular Descriptors' Generation and Selection

To calculate the molecular descriptors of each compound, their structures were drawn using ISIS Draw 2.3 (MDL Information Systems, Inc., San Ramon, CA, USA) [33]. The MM⁺ molecular mechanics forcefield in the HyperChem 6.0 program (Hypercube, Inc.: Waterloo, ON, Canada) was then used to carry out the preliminary molecular geometry optimization [34]. The further optimization of the compound structure was done by semi-empirical PM₃ method utilizing the Polak–Ribiere algorithm until the root mean square gradient was 0.01 kcal/mol [35]. Finally, a more precise optimization was achieved by MOPAC 6.0 software package (Indiana University: Bloomington, IN, USA) [36]. Afterwards, the final optimized structures were converted to the CODESSA 2.63 program (University of Florida, Gainesville, FL, USA) for calculating the five classes of descriptors, namely constitutional, topological, geometrical, electrostatic, and quantum-chemical descriptors [37]. It was necessary to explain that the logP descriptor, which cannot be calculated by the CODESSA 2.63, but can be obtained by Hyperchem, was then added to the descriptors pool [34]. Through doing these, 494, 597, and 611 descriptors were gained for each of the studied compounds in Group 1, 2, and 3, respectively.

Before establishing the QSAR models, it is necessary to remove the insignificant descriptors, and the constant and highly intercorrelated descriptors (the intercorrelation of the descriptors should be lower than 0.8). In this paper, the heuristic method (HM) was used to achieve a thorough search for the best multilinear correlations with the computed descriptors in the framework of the program CODESSA 2.63 [37].

2.3. Multiple Linear Regressions (MLR)

Multiple linear regressions (MLR) are often accepted as a classical method for solving linear problems when there are two or more than two independent variables in QSAR modeling. The purpose of MLR is to find a mathematical function which best depicts the desired activity Y (here, $-\log \text{IGC}_{50}$ values) as a linear combination of the X -variables (the molecular descriptors), with the regression coefficients b_n . The equation is as follows:

$$Y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n.$$

Usually, the good fit alone does not guarantee that the model is useful for prediction purposes by the R^2 (coefficient of determination), $LOOq^2$ (leave-one-out correlation coefficient), RMS (root mean square error), F (Fisher's statistics), etc. [38]. Some statistical characteristics of the test set are also needed to be considered: R^2 (coefficient of determination), R_0^2 (the coefficients of determination, predicted vs observed activities, when the Y -intercept b_0 is set to zero), as well as by their corresponding slopes k and k' . The following conditions need to be fulfilled to adequately estimate the predictive ability of a model [39]:

$$\begin{aligned} q^2 &> 0.5 \\ R^2 &> 0.6 \\ \frac{(R^2 - R_0^2)}{R^2} &< 0.1 \text{ or } \frac{(R^2 - R_0'^2)}{R^2} < 0.1 \\ 0.85 &\leq k \leq 1.15 \text{ or } 0.85 \leq k' \leq 1.15 \end{aligned}$$

2.4. Radial Basis Function Neural Networks (RBFNN)

In general, RBFNN may have a better result than MLR, because it can take into account some nonlinear behavior between the molecular descriptors and the desired activities values ($-\log \text{IGC}_{50}$). The detailed introduction of RBFNN has been stated in previous studies [40,41], so we only make a simple statement of the key parts here.

The RBFNN is a typical feed forward neural network which is composed of three layers, which are the input layer, the hidden layer, and the output layer. The first layer is linear, and distributes the input values, while the next layer is nonlinear, and uses radial basis function. The third layer linearly combines the outputs. Each neuron in each layer is adequately linked to the next layer. However, there is no connection between neurons in a given layer. Each hidden layer unit stands for a single radial basis function, which is characterized by a center and a width. In this layer, each neuron uses a radial basis function as nonlinear transfer function to handle the input information from the previous layer. The most common use of RBF is the Gauss function, characterized by the center (c_j) and width (r_j) [42]. It is used to measure the Euclidean distance between the input vector (x) and the radial basis function center (c_j), and gain the nonlinear transformation within the hidden layer, defined as

$$h_j = \exp\left(-\|x - c_j\|^2 / r_j^2\right),$$

where h_j is the output of the j th RBF unit, while c_j and r_j are the center and width of such a unit, respectively. And the operation of the output layer is linear and is given by

$$y_k(x) = \sum_{j=1}^{n_h} w_{kj}h_j(x) + b_k$$

where y_k is the k th output unit for the input vector X , w_{kj} is the weight connection between the k th output unit and the j th hidden layer unit, and b_k is the respective bias.

In the present study, we used the MATLAB package (MathWorks, Natick, MA, USA) (www.mathworks.com/products/matlab/) to accomplish all the RBFNN calculations. The total functions of the RBFNN model can be evaluated by the same statistical parameters as the MLR method together with its reliability and robustness.

2.5. Applicability Domain (AD) of the Model

It is necessary to give the application domain (AD) of the model. The applicability domain (AD) of a QSAR model refers to a theoretical region in the space defined by the compounds in the training set. It demonstrates the nature of the compound molecules that can be utilized in the built model. That is to say, AD restricts a theoretical region, also for unknown chemicals without experimental data, with the lowest number of bad predictions (Y-outliers) and chemicals far from the training structural domain [43]. In this study, a William's plot, i.e., a plot of standardized residuals (R) vs leverages was used [44]. Here, a simple measure of a chemical being too far from the applicability domain of the model is its leverage, h_i [43], as follows:

$$h_i = x_i^T (x^T x)^{-1} x_i (i = 1, 2, \dots, n).$$

In the above equation, x_i represents the descriptor row vector of the studied compound, while x represents the $n \times k - 1$ matrix of k model descriptor values for the n training set compounds. The superscript "T" refers to the transpose of the matrix/vector. h_i characterizes the leverage of a compound, and is one of the coordinates of the William's plot (standardized residuals versus leverage).

3. Results and Discussion

3.1. MLR Results

As mentioned above, based on the structural differences among the molecules which are caused by the influential functional groups ($-\text{NO}_2$ and $-\text{X}$), Group 1, 2, and 3 have 46, 75, 39 compounds, respectively. The models of each group were established by the training sets. Before doing this, the heuristic method (HM) was used to conduct the descriptor selection. After the preselection of the descriptors, 178, 203, and 160 descriptors were left for each group by removing of the descriptors not obeyed the thumb rules [45].

Multilinear regression models were then developed in a stepwise procedure, that is, the descriptors and correlations were sorted by the values of the F -test and the correlation coefficients. Beginning with the top descriptor from the list, two-parameter correlations were calculated. Later, the descriptors were added one by one, until the preselected number of descriptors in the model is fulfilled. Finally, three descriptors were used to describe the relationship between molecule structure and toxicity for each group of compounds. The selected descriptors and their chemical meaning, along with the statistical parameters, are listed in Tables 2–4.

Table 2. Descriptors, Coefficients, Standard Error, and t -Test Values for the Best MLR Model of Group 1.

	Coefficients	Standard Errors	t -Test	Descriptors
0	73.123	20.547	3.559	Intercept
1	0.002	0.000	10.075	Gravitation index (all bonds) (G^2)
2	−1.016	0.188	−5.394	Max bond order of a O atom (P_o)
3	−1.082	0.510	−3.568	Max n−n repulsion for a C−H bond ($Enn(C-H)$)
$N = 36, R^2 = 0.829, LOOq^2 = 0.813, F = 51.697, RMS = 0.192$				

Table 3. Descriptors, Coefficients, Standard Errors, and *t*-Test Values for the Best MLR Model of Group 2.

	Coefficients	Standard Errors	<i>t</i> -Test	Descriptors
0	−18.030	3.703	−4.869	Intercept
1	0.438	0.045	9.808	LogP
2	−6.605	0.605	−10.918	FNSA-2 Fractional PNSA (PNSA-2/TMSA) [Zefirov's PC] (PNSA-2/TMSA)
3	17.066	3.794	4.498	Max SIGMA–SIGMA bond order(P _{SIGMA})

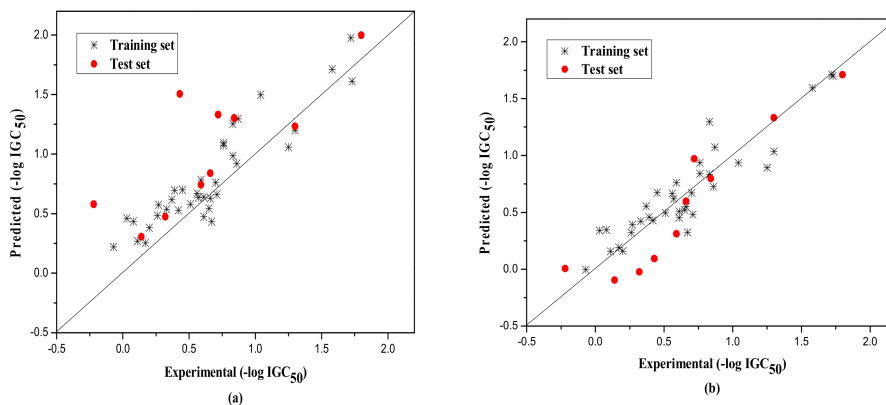
$N = 60, R^2 = 0.803, LOOq^2 = 0.792, F = 76.016, RMS = 0.222$

Table 4. Descriptors, Coefficients, Standard Errors, and *t*-Test Values for the Best MLR Model of Group 3.

	Coefficients	Standard Errors	<i>t</i> -Test	Descriptors
0	−16.640	3.092	−5.382	Intercept
1	−2.141	0.322	−6.650	Principal moment of inertia C (Ic)
2	0.151	0.024	6.292	Max e–e repulsion for a C–C bond (Enn(C–C))
3	−51.290	7.493	−6.845	RPCG Relative positive charge (QMPOS/QTPLUS) [Quantum-Chemical PC] (QMPOS/QTPLUS)

$N = 31, R^2 = 0.852, LOOq^2 = 0.835, F = 51.678, RMS = 0.193$

The external test set was also used to further evaluate the three models. The statistical parameters obtained are as follows: $N_{\text{ext}} = 10, R^2 = 0.917, q_{\text{ext}}^2 = 0.851, F = 13.820, RMS = 0.222$ for group 1; $N_{\text{ext}} = 15, R^2 = 0.789, q_{\text{ext}}^2 = 0.732, F = 13.720, RMS = 0.266$ for group 2; $N_{\text{ext}} = 8, R^2 = 0.733, q_{\text{ext}}^2 = 0.730, F = 260.404, RMS = 0.380$ for Group 3. Figures 1, 2 and 3a show the predicted vs observed $-\log IGC_{50}$ values for all the training and test set compounds. Thus, it can be seen that the model is reasonable in both statistical significance and predictive ability.

**Figure 1.** Plot of the predicted versus experimental $-\log IGC_{50}$ including the training and the test set compounds of Group 1 by MLR model (a) and by RBFNN model (b).

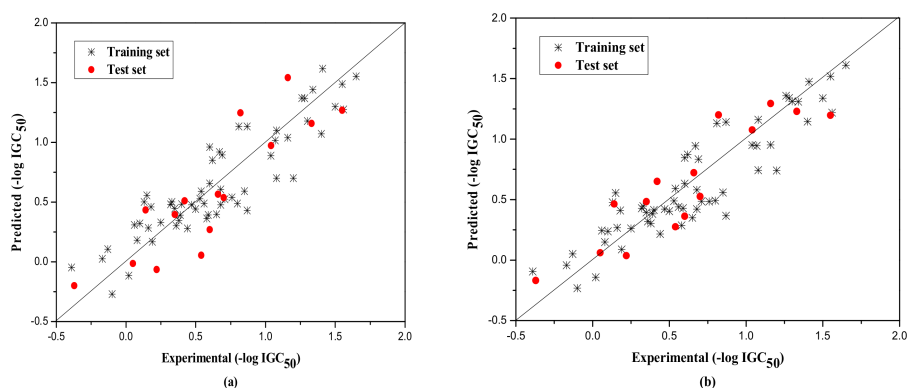


Figure 2. Plot of the predicted versus experimental $-\log \text{IGC}_{50}$ including the training and the test set compounds of Group 2 by MLR model (a) and by RBFNN model (b).

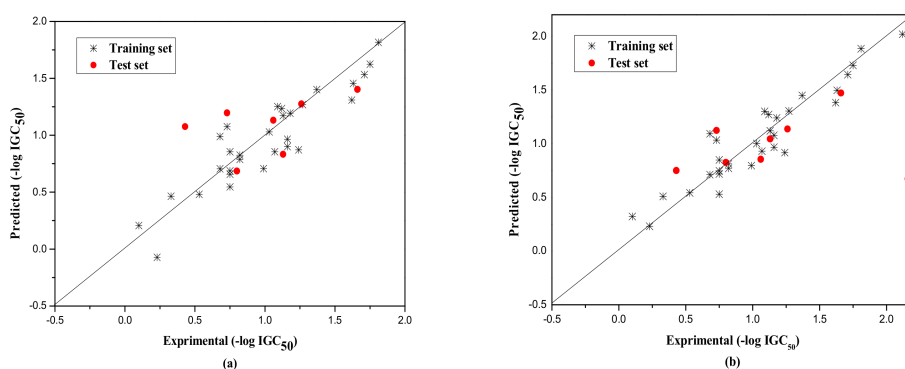


Figure 3. Plot of the predicted versus experimental $-\log \text{IGC}_{50}$ including the training and the test set compounds of Group 3 by MLR model (a) and by RBFNN model (b).

3.2. Model Applicability Domain Analysis and Improved MLR Model

It is also an important step to consider the possible outliers of the models. In order to visualize the AD, the plot of standardized cross-validated residuals versus leverage (the William's plot), which can provide an immediate and simple graphical detection, was used to find out the outliers from the models. In this plot, the horizontal and vertical straight lines represent the normal control values of Y-outliers and X-outliers, respectively. The limit of X-coordinate is $3m/n$, where m is the number of model parameters, and n is the number of samples belonging to the training set. In the present study, the normal control value for Y-outliers (RES) was set as $\pm 3\sigma$. Figures 4–6 show the William's plot based on the MLR models for the whole dataset compounds of group 1, 2, 3, respectively.

As can be judged from Figure 4, in the model for Group 1, there is one X-outlier (for Group 1: compound 2), which is 2-nitroanisole. In its structure, there are two functional groups, $-\text{NO}_2$ and methoxy. The former is in all of the compounds belonging to this group as a strong electron-withdrawing group. However, the methoxy group has oxygen lone pair electrons which are a strong electron donor moiety, compared to other ones in the group. Therefore, care should be taken when using the compounds with methoxy, since they can activate the benzene ring and exert an unusual influence on the toxicity. And from Figure 6, it can also be seen that there is a X-outlier (for group 3: compound 15), that is, 2-chloromethyl-4-nitrophenol. This compound has three electron-withdrawing moieties, including $-\text{Cl}$, $-\text{OH}$, and $-\text{NO}_2$, which has almost the strongest induction effect of the compounds in this group. Also, there seem to be another outlier (for Group 3: compound 36), which belongs to the test set. This may be due to variability in the measurement, or it may indicate experimental error.

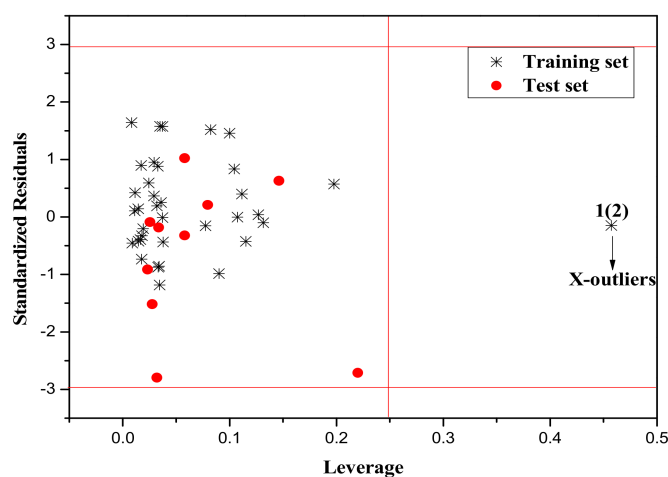


Figure 4. The William's plot for the training and test set compounds of Group 1 by MLR model.

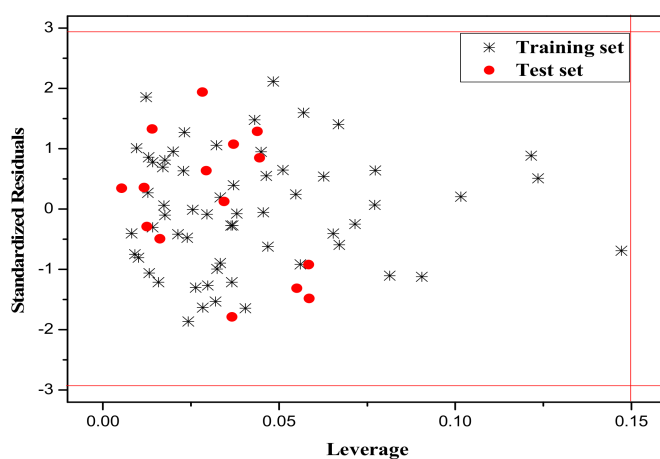


Figure 5. The William's plot for the training and test set compounds of Group 2 by MLR model.

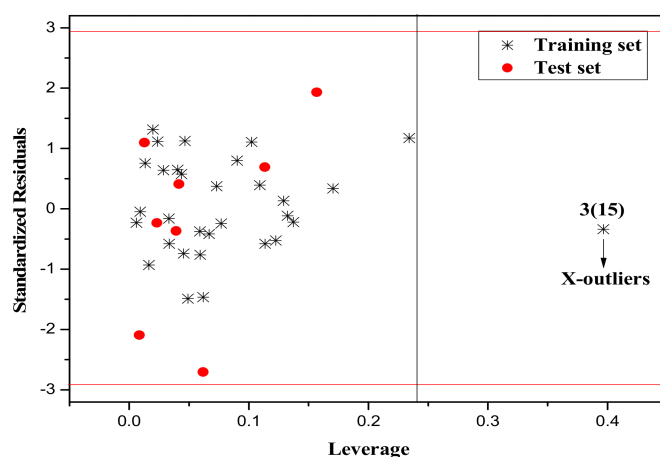


Figure 6. The William's plot for the training and test set compounds of Group 3 by MLR model.

If the handling of the outliers is unreasonable, the accuracy of the model will be affected. Thus, the quality and ability of the model prediction will be affected. Therefore, we removed the outliers from Group 1 and Group 3, set up the models anew, and the results were as follows: for the training

set of Group 1 (removing compound 2 in Group 1): $N = 35$, $R^2 = 0.926$, $LOOq^2 = 0.916$, $F = 94.266$, $RMS = 0.125$. For the training set of Group 3 (removing compound 15 in Group 3): $N = 30$, $R^2 = 0.852$, $LOOq^2 = 0.834$, $F = 49.710$, $RMS = 0.196$. The statistical parameters of the model are better after removing the escape values.

To further assess the predictive powers of the model established by the MLR method, parameters such as $\frac{(R^2 - R_0^2)}{R^2}$, k , k' , etc., were also calculated, and the results were shown in Table 5. From the table, we can see the statistical results were all within the acceptable ranges for the methods of MLR.

Table 5. The statistical results of the external test set for the three models of each group.

	Group 1		Group 2		Group 3	
	MLR	RBFNN	MLR	RBFNN	MLR	RBFNN
R^2	0.83	0.84	0.80	0.82	0.85	0.89
q_{ext}^2	0.92	0.88	0.79	0.81	0.73	0.63
R_0^2	0.81	0.84	0.79	0.82	0.84	0.88
$\frac{(R^2 - R_0^2)}{R^2}$	0.024	0.00	0.012	0.00	0.012	0.011
k	0.88	0.86	0.80	0.83	0.85	0.87
k'	1.11	0.97	0.93	0.91	0.93	0.98

3.3. Validation Results of the Models

Further, a fivefold cross-validation algorithm was applied for validation of the stability of the three models. The members selected for each group (i.e., groups A, B, C, D, and T) were shown in Table 1. The R^2 , F , and RMS values for each validation along with their average values were shown in Table 6 for the MLR models. As can be seen, both models are stable, judging from the obtained values for the average training quality and for the average predicting quality.

Table 6. Validation of the MLR models.

Training Set	R^2	F	RMS	Test Set	R^2	F	RMS
Group 1							
A + B + C + D	0.829	51.697	0.192	T	0.917	13.820	0.222
A + B + C + T	0.735	30.457	0.261	D	0.874	48.592	0.153
B + C + D + T	0.742	31.565	0.255	A	0.795	27.199	0.217
A + C + D + T	0.723	28.775	0.259	B	0.889	55.784	0.173
A + B + D + T	0.744	31.912	0.247	C	0.835	35.501	0.217
Average	0.755	34.881	0.243		0.862	36.180	0.196
Group 2							
A + B + C + D	0.803	76.016	0.222	T	0.852	51.678	0.193
A + B + C + T	0.802	75.661	0.225	D	0.748	38.559	0.256
B + C + D + T	0.784	67.805	0.235	A	0.857	78.133	0.193
A + C + D + T	0.786	68.459	0.233	B	0.818	58.416	0.221
A + B + D + T	0.780	66.184	0.235	C	0.831	63.981	0.218
Average	0.791	70.834	0.230		0.821	58.153	0.216
Group 3							
A + B + C + D	0.789	13.720	0.260	T	0.733	260.404	0.380
A + B + C + T	0.760	29.502	0.263	D	0.927	63.850	0.115
B + C + D + T	0.755	27.787	0.254	A	0.899	53.612	0.170
A + C + D + T	0.762	28.869	0.250	B	0.917	66.385	0.159
A + B + D + T	0.754	27.573	0.248	C	0.846	32.957	0.237
Average	0.764	25.490	0.255		0.864	95.442	0.212

3.4. RBFNN Results

In the field of QSAR research, RBFNN often shows better results than MLR because of its ability to consider some nonlinear relationships between the molecular structure and its activity. In order to confirm this view, RBFNN was utilized to build nonlinear predictive models using the same descriptors selected by the MLR models. The RBFNN can be traced as $i-n_k-1$ net to indicate the number of units in the three layers, respectively. Meanwhile, the width (r) of RBF was computed by systemically changing its value in the training step from 0.1 to 4.0 with increments of 0.1. For the three groups of compounds belonging to training sets in this study, the RBFNN models were 3-10-1, 3-9-1, and 3-9-1, along with widths of 0.8, 2.0, and 1.7, respectively.

Their statistical results of the training and the test set are as follows. Group1: for training set, $N = 36$, $R^2 = 0.843$, $LOOq^2 = 0.838$, $F = 182.306$, $RMS = 0.167$, and for the test set, $N_{ext} = 10$, $R^2 = 0.881$, $q_{ext}^2 = 0.867$, $F = 59.483$, $RMS = 0.210$; Group 2: for training set, $N = 60$, $R^2 = 0.821$, $LOOq^2 = 0.818$, $F = 265.898$, $RMS = 0.192$, and for the test set, $N_{ext} = 15$, $R^2 = 0.810$, $q_{ext}^2 = 0.796$, $F = 55.506$, $RMS = 0.232$; Group 3: for training set, $N = 31$, $R^2 = 0.885$, $LOOq^2 = 0.882$, $F = 224.261$, $RMS = 0.163$, and for the test set, $N_{ext} = 8$, $R^2 = 0.632$, $q_{ext}^2 = 0.622$, $F = 63.660$, $RMS = 0.298$. The corresponding predicted endpoint values of each compound in each group were shown in Table 1, and the plot of the predicted and experimental values of both training and test set were displayed in Figures 1b, 2b and 3b. Different from the original literature, [35], we selected and classified the original compounds according to the structural characteristics and further modeled, analyzed, and predicted the corresponding toxicity values. The models, thus established, are also more targeted for the particular compounds, and the statistical results of $\frac{(R^2 - R_0^2)}{R^2}$, k , k' , etc., as shown in Table 5 by RBFNN, also indicated the models to be statistically robust with high external predictivity.

3.5. Interpretation of Model Descriptors

In order to deepen the understanding of this study, more detailed explanations of the descriptors selected in each group were performed. For group 1, three descriptors were selected in the QSAR model, namely: G^2 , P_{AB} , and $Enn_{(C-H)}$. The positive sign of them indicated that the $-\log IGC_{50}$ values increased with its increase, and vice versa. G^2 refers to gravitation indexes for all bonded pairs of atoms, and it is defined as $G^2 = \sum_{(i>j)}^{N_B} \frac{m_i m_j}{r_{ij}^2}$ [46], where m_i and m_j are the atomic weights of atoms i and j , r_{ij} is the interatomic distance, N_B is the number of bonds in the molecule. P_o belongs to the valency-related descriptors, which relate to the strength of intermolecular bonding interactions and characterize the stability of the molecules, their conformational flexibility and other valency-related properties [47]. $E_{nn(C-H)}$ is Max n-n repulsion for a C-H bond, calculated as follows: $E_{nn}(CH) = \frac{Z_C Z_H}{R_{CH}}$, where Z_C and Z_H are the nuclear (core) charges of atoms C and H , respectively, and R_{CH} is the distance between them. This energy describes the nuclear repulsion driven processes in the molecule, and may be related to the conformational (rotational, inversional) changes or atomic reactivity in the molecule [46].

For Group 2, focusing on the compounds without the functional group $-\text{NO}_2$, but with $-X$, three descriptors were chosen. That is, $\text{Log } P$, PNSA-2/TMSA , and P_{SIGMA} . PNSA-2/TMSA is FNSA-2 fractional PNSA (PNSA-2/TMSA) [Zefirov's PC], which contributes to the calculation of atomic partial charges to the total molecular solvent-accessible surface area [46]. P_{SIGMA} represents the maximum bond order for a given pair of atomic species in the molecule, its values for a given pair of atomic species in the molecule with the lower limit $P_{\text{SIGMA}}(\text{min}) > 0.1$. $\text{Log } P$ stands for the solvational characteristic (hydrophobicity of chemicals) because it is closely related to the change in the Gibbs energy of solvation of a solute between two solvents.

For Group 3, three descriptors were selected to build the model, that is I_c , $Enn_{(C-C)}$, and RPCG . The chemical meaning of them can be seen in Table 4. I_c is a geometrical descriptor which relates to the atomic masses, the distance of the atomic nucleus from the main rotational axes, which characterizes the mass distribution in the molecule. $Enn(C-C) = Z_C Z_C / R_{C-C}$, where Z_C and Z_C are the nuclear (core) charges of atoms C , and R_{C-C} is the distance between them. This energy describes the nuclear

expulsion driven processes in the molecule, and may be related to the conformational (rotational, inversional) changes or atomic reactivity in the molecule [48]. RPCG, relative positive charge, belongs to electrostatic descriptors. From its coefficient, we can find that the relative positive charge of the molecule is negatively related to the endpoint values ($-\log \text{IGC}_{50}$).

In summary, we found that the repulsion between the two bonds and the local charge on the surface of the molecule appeared in different models, indicating that these two factors have a greater influence on the structure of the compound and should be relatively valued.

4. Conclusions

In the present study, the QSAR models were performed on the study of the acute toxicity of substituted aromatic compounds to the aquatic ciliate *Tetrahymena pyriformis* using the MLR and RBFNN methods, and by dividing the whole dataset into three groups based on the most influential functional group ($-\text{NO}_2$ and $-\text{X}$). Acceptable statistical results for each model indicated their good stability and good predictability. We can also see from the results of the MLR and RBFNN models that the MLR method can establish reasonable models for evaluating the activity of compounds, and the RBFNN method can provide better statistical parameters. Also, the selected descriptors are effective and feasible for evaluating the toxicity of this group of compounds. Lastly, the results of this study provided useful insights on the characteristics of the structures that most affect the toxicity.

Author Contributions: Feng Luan and M. Natália Dias Soeiro Cordeiro conceived and designed the experiments; Ting Wang and Lili Tang performed the experiments and wrote the paper; Feng Luan and Shuang Zhang reviewed the paper.

Acknowledgments: This work was financially supported by the National Natural Science Foundation of China (No. 21675138).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Schultz, T.W.; Kier, L.B.; Hall, L.H. Structure-toxicity relationships of selected nitrogenous heterocyclic compounds. III. Relations using molecular connectivity. *Bull. Environ. Contam. Toxicol.* **1982**, *28*, 373–378. [[CrossRef](#)] [[PubMed](#)]
2. Oren, A.; Gurevich, P.; Henis, Y. Reduction of nitrosubstituted aromatic compounds by the halophilic anaerobic eubacteria *Haloanaerobium praevalens* and *Sporohalobacter marismortui*. *Appl. Environ. Microbiol.* **1991**, *57*, 3367–3370. [[PubMed](#)]
3. Gooch, A.; Sizochenko, N.; Rasulev, B.; Gorb, L.; Leszczynski, J. In vivo toxicity of nitroaromatics: A comprehensive quantitative structure–activity relationship study. *Environ. Toxicol. Chem.* **2017**, *36*, 2227–2233. [[CrossRef](#)] [[PubMed](#)]
4. Finger, G.C.; Kruse, C.W. Aromatic fluorine compounds. VII. Replacement of aromatic-Cl and $-\text{NO}_2$ groups by $-\text{F}^{1,2}$. *J. Am. Chem. Soc.* **1956**, *78*, 6034–6037. [[CrossRef](#)]
5. Zhang, A.Q.; Chen, R.Q.; Wei, D.B.; Wang, L.S. QSAR research of chlorinated aromatic compounds toxicity to *Selenastrum capricornutum*. *China Environ. Sci.* **2000**, *20*. [[CrossRef](#)]
6. Gupta, A.K.; Chakraborty, A.; Giri, S.; Chattaraj, P. Toxicity of halogen, sulfur and chlorinated aromatic compounds. *Int. J. Chemoinform. Chem. Eng.* **2011**, *1*, 61–74. [[CrossRef](#)]
7. Lu, G.H.; Wang, C.; Yuan, X.; Lang, P.Z. Quantitative structure-activity relationships for the toxicity of substituted benzenes to *Cyprinus carpio*. *Biomed. Environ. Sci.* **2005**, *18*, 53–57. [[PubMed](#)]
8. Shintou, T.; Fujii, S.; Kubo, S. Process for Producing Iodinated Aromatic Compounds. U.S. Patent 6,437,203, 20 August 2002.
9. Arcangeli, J.P.; Arvin, E. Biodegradation rates of aromatic contaminants in biofilm reactors. *Water Sci. Technol.* **1995**, *31*, 117–128.
10. Jing, G.H.; Li, X.L.; Zou, Z.M. Quantitative structure-activity relationship (QSAR) study of toxicity of substituted aromatic compounds to *Photobacterium phosphoreum*. *Chin. J. Struct. Chem.* **2010**, *29*, 1189–1196.

11. Khadikar, P.V.; Mather, K.C.; Singh, S.; Phadnis, A.; Shrivastava, A.; Mandaloi, M. Study on quantitative structure toxicity relationships for benzene derivatives acting by narcosis. *Bioorg. Med. Chem.* **2002**, *10*, 1761–1766. [[CrossRef](#)]
12. Giddings, J.M. Acute toxicity to *Selenastrum capricornutum*, of aromatic compounds from coal conversion. *Bull. Environ. Contam. Toxicol.* **1979**, *23*, 360–364. [[CrossRef](#)] [[PubMed](#)]
13. Kuivasniemi, K.; Eloranta, V.; Knuutinen, J. Acute toxicity of some chlorinated phenolic compounds to *Selenastrum capricornutum*, and phytoplankton. *Arch. Environ. Contam. Toxicol.* **1985**, *14*, 43–49. [[CrossRef](#)]
14. Sverdrup, L.E.; Krogh, P.H.; Nielsen, T.; Kjaer, C.; Stenersen, J. Toxicity of eight polycyclic aromatic compounds to red clover (*Trifolium pratense*), ryegrass (*Lolium perenne*), and mustard (*Sinapsis alba*). *Chemosphere* **2003**, *53*, 993–1003. [[CrossRef](#)]
15. Kobetičová, K.; Bezchlebová, J.; Lána, J.; Sochová, I.; Hofman, J. Toxicity of four nitrogen-heterocyclic polyaromatic hydrocarbons (NPAHs) to soil organisms. *Ecotoxicol. Environ. Saf.* **2008**, *71*, 650–660. [[CrossRef](#)] [[PubMed](#)]
16. Cronin, M.T.; Schultz, T.W. Development of quantitative structure–activity relationships for the toxicity of aromatic compounds to *Tetrahymena pyriformis*: Comparative assessment of the methodologies. *Chem. Res. Toxicol.* **2001**, *14*, 1284–1295. [[CrossRef](#)] [[PubMed](#)]
17. Cronin, M.T.; Aptula, A.O.; Duffy, J.C.; Netzeva, T.I.; Rowq, P.H.; Valkova, I.V.; Schultz, T.W. Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to *Tetrahymena pyriformis*. *Chemosphere* **2002**, *49*, 1201–1221. [[CrossRef](#)]
18. Luan, F.; Xu, X.; Liu, H.T.; Cordeiro, M.N. Prediction of the baseline toxicity of non-polar narcotic chemical mixtures by QSAR approach. *Chemosphere* **2013**, *90*, 1980–1986. [[CrossRef](#)] [[PubMed](#)]
19. Cronin, M.T.; Gregory, B.W.; Schultz, T.W. Quantitative structure-activity analyses of nitrobenzene toxicity to *Tetrahymena pyriformis*. *Chem. Res. Toxicol.* **1998**, *11*, 902–908. [[CrossRef](#)] [[PubMed](#)]
20. Netzeva, T.I.; Schultz, T.W. QSAR for the aquatic toxicity of aromatic aldehydes from tetrahymena data. *Chemosphere* **2005**, *61*, 1632–1643. [[CrossRef](#)] [[PubMed](#)]
21. Roy, D.R.; Parthasarathi, R.; Subramanian, V.; Chattaraj, P.K. An electrophilicity based analysis of toxicity of aromatic compounds towards *Tetrahymena pyriformis*. *Mol. Inform.* **2006**, *25*, 114–122.
22. Devillers, J. Linear versus nonlinear QSAR modeling of the toxicity of phenol derivatives to *Tetrahymena pyriformis*. *SAR QSAR Environ. Res.* **2007**, *15*, 237–249. [[CrossRef](#)] [[PubMed](#)]
23. Tetko, I.V.; Sushko, I.; Pandey, A.K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: Focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733–1746. [[CrossRef](#)] [[PubMed](#)]
24. Artemenko, A.G.; Muratov, E.N.; Kuz'Min, V.E.; Muratov, N.N.; Varlamova, E.V.; Kuz'Mina, A.V.; Gorb, L.G.; Golius, A.; Hill, F.C.; Leszczynski, J.; et al. QSAR analysis of the toxicity of nitroaromatics in *Tetrahymena pyriformis*: Structural factors and possible modes of action. *SAR QSAR Environ. Res.* **2011**, *22*, 575–601. [[CrossRef](#)] [[PubMed](#)]
25. Zarei, K.; Atabati, M.; Kor, K. Bee algorithm and adaptive neuro-fuzzy inference system as tools for QSAR study toxicity of substituted benzenes to *Tetrahymena pyriformis*. *Bull. Environ. Contam. Toxicol.* **2014**, *92*, 642–649. [[CrossRef](#)] [[PubMed](#)]
26. Li, J.F.; Liao, L.M. Structural characterization and acute toxicity prediction of substituted aromatic compounds by using molecular vertexes correlative index. *Chin. J. Struct. Chem.* **2013**, *32*, 557–563. [[CrossRef](#)]
27. Shi, J.Q.; Cheng, J.; Wang, F.Y.; Flamm, A.; Wang, Z.Y.; Yang, X. Acute toxicity and *n*-octanol/water partition coefficients of substituted thiophenols: Determination and QSAR analysis. *Ecotoxicol. Environ. Saf.* **2012**, *78*, 134–141. [[CrossRef](#)] [[PubMed](#)]
28. Salahinejad, M.; Ghasemi, J.B. 3D-QSAR studies on the toxicity of substituted benzenes to *Tetrahymena pyriformis*: CoMFA, CoMSIA and VolSurf approaches. *Ecotoxicol. Environ. Saf.* **2014**, *105*, 128–134. [[CrossRef](#)] [[PubMed](#)]
29. Bertinetto, C.; Duce, C.; Solaro, R.; Tiné, M.R.; Micheli, A.; Héberger, K.; Miličević, A.; Nikolić, S. Modeling of the acute toxicity of benzene derivatives by complementary QSAR methods. *Match-Commun. Math. Comput. Chem.* **2013**, *70*, 1005–1021.
30. Wang, D.D.; Feng, L.L.; He, G.Y.; Chen, H.Q. QSAR studies for assessing the acute toxicity of nitrobenzenes to *Tetrahymena pyriformis*. *J. Serb. Chem. Soc.* **2014**, *79*, 1111–1125. [[CrossRef](#)]

31. Wei, D.B.; Zhai, L.H.; Dong, C.H.; Hu, H.Y. Determination and prediction of the acute toxicity of substituted benzene compounds to luminescent bacteria. *Chin. J. Environ. Sci.* **2002**, *S1*, 3–7.
32. Schultz, T.W.; Netzeva, T.I.; Cronin, M.T. Selection of data sets for QSARs: Analyses of tetrahymena toxicity from aromatic compounds. *SAR QSAR Environ. Res.* **2003**, *14*, 59–81. [[CrossRef](#)] [[PubMed](#)]
33. *ISIS Draw2.3*, MDL Information Systems, Inc.: San Ramon, CA, USA, 1990–2000.
34. *HyperChem 6.01*, Hypercube, Inc.: Waterloo, ON, Canada, 2000.
35. Dewar, M.J.; Storch, D.M. Development and use of quantum molecular models. 75. Comparative tests of theoretical procedures for studying chemical reactions. *J. Am. Chem. Soc.* **1985**, *107*, 3898–3902. [[CrossRef](#)]
36. Stewart, J.P.P. *MOPAC 6.0*, Quantum Chemistry Program Exchange, No. 455; Indiana University: Bloomington, IN, USA, 1989.
37. Katritzky, A.R.; Lobanov, V.S.; Karelson, M. *CODESSA 2.63: Training Manual*; University of Florida: Gainesville, FL, USA, 1995.
38. Golbraikh, A.; Tropsha, A. Beware of q²! *J. Mol. Graph. Model.* **2002**, *20*, 269–276. [[CrossRef](#)]
39. Tropsha, A.; Gramatica, P.; Gombar, V. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003**, *22*, 69–77. [[CrossRef](#)]
40. Xiang, Y.H.; Liu, M.C.; Zhang, X.Y.; Zhang, R.S.; Hu, Z.D.; Fan, B.T.; Doucet, J.P.; Panaye, A. Quantitative prediction of liquid chromatography retention of *N*-benzylideneanilines based on quantum chemical parameters and radial basis function neural network. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 592–597. [[CrossRef](#)] [[PubMed](#)]
41. Gharagheizi, F. QSPR analysis for intrinsic viscosity of polymer solutions by means of GA-MLR and RBFNN. *Comput. Mater. Sci.* **2007**, *40*, 159–167. [[CrossRef](#)]
42. Shahlaei, M.; Madadkar-Sobhani, A.; Fassihi, A.; Saghaie, L.; Arkan, E. QSAR study of some CCR5 antagonists as anti-HIV agents using radial basis function neural network and general regression neural network on the basis of principal components. *Med. Chem. Res.* **2012**, *21*, 3246–3262. [[CrossRef](#)]
43. Atkinson, A.C. Plots, transformations, and regression. An introduction to graphical methods of diagnostic regression analysis. *J. R. Stat. Soc.* **1985**, *152*, 1927–1934.
44. Gadaleta, D.; Mangiatordi, G.F.; Catto, M.; Carotti, A.; Nicolotti, O. Applicability domain for QSAR models: Where theory meets reality. *Int. J. QSPR* **2016**, *1*, 45–63. [[CrossRef](#)]
45. Luan, F.; Tang, L.L.; Zhang, L.H.; Zhang, S.; Monteagudo, M.C.; Cordeiro, M.N.D.S. A further development of the QNAR model to predict the cellular uptake of nanoparticles by pancreatic cancer cells. *Food Chem. Toxicol.* **2018**, *112*, 571–580. [[CrossRef](#)] [[PubMed](#)]
46. Katritzky, A.R.; Lobanov, V.S.; Karelson, M. *Comprehensive Descriptors for Structural and Statistical Analysis; Reference Manual, Version 2.0*; University of Florida: Gainesville, FL, USA, 1994.
47. Sannigrahi, A.B. AB initio molecular orbital calculations of bond index and valency. *Adv. Quantum Chem.* **1992**, *23*, 301–351.
48. Štrouf, O. *Chemical Pattern Recognition*; Research Studies Press: Baldock, UK, 1986; Volume 11.

Sample Availability: Samples of the compounds are available from the authors.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).