Prediction of the Aquatic Toxicity of Phenols to *Tetrahymena Pyriformis* from Molecular Descriptors

Jiang, D. X. [1], Li, Y. [2], Li, J. [3] and Wang, G. X. [1*]

[1] Northwest A&F University, Xinong road 22th, Yangling, 712100, China

[2] Dalian University of Technology, Linggong Road 2, Dalian, 116024, China

[3] Freshwater Fisheries Sciences Institute of Liaoning Province, Liaoning, 111000, China

ABSTRACT: The purpose of this work is to develop robust and interpretable quantitative structure"activity relationship (QSAR) models for assessing the aquatic toxicity of phenols using a combined set of descriptors encompassing the logP and recently developed variables (Monconn-Z variables). The used dataset consists of 250 chemicals with toxicity data to the ciliate *Tetrahymena pyriformis*. For each compound, a total of 197 physico-chemical descriptors including logP and Molconn-Z descriptors were calculated. Multiple linear regression (MLR) and Partial least squares (PLS) were used to obtain QSARs and the predictive performance of the proposed models were verified using external statistical validations. The results of stepwise-MLR analysis reveal that the AlogP, MlogP and ClogP models were not successful for the prediction of aquatic toxicity for all the compounds. And by using the logP (AlogP and MlogP) with Molconn-Z descriptors, the obtained QSARs were not successful enough nutill removal of some outliers. Two optimal QSARs were built with R^2 of 0.71 and 0.70 for the training sets and the external validation Q^2 of 0.69 and 0.68 respectively. All these selected descriptors in the best models account for the hydrophobic (AlogP, MlogP) and other electrotopological properties like SHCsatu, Scarboxylicacid, SHBa, gmax and nelem. PLS produces a good model using all the calculated descriptors with R^2 of 0.78 and Q^2 of 0.64, and hydrophobic and electrotopological descriptors show importance for the prediction of phenolic toxicity.

Key word: QSAR, Molconn-Z descriptors, LogP descriptors, Aquatic toxicity, *Tetrahymena pyriformis*, Phenols

INTRODUCTION

Phenols represent a substantial part of the chemicals produced worldwide. They have been widely used as materials in medicine, industry and agriculture (Liu *et al.*, 2010). Despite their great importance, when released in the environment as wastewater streams, such organic pollutants are toxic to humans (Bukowska *et al.*, 2004), and serious threat to the ecosystems as well (Kušić, 2009). Even, some are persistent in the environment (Cunningham *et al.*, 2005). Due to these adverse effects on living species and environment, it is necessary to assess the toxicity and explore the acute toxicity mechanisms of such compounds, which is not only significant to preserve of environment but also helpful to propose a reasonable policy for the government to regulate those compounds.

However, assessment the toxicity for a given compound by performing a toxicological experiment is not an easy task because this can be costly, time-consuming and could potentially produce toxic side products from the experimental methods used today (Hill, 1972). Analysis of the toxicity should also consider multiple environments and all biological interactions with the living organisms of the ecosystems ,but data that quite often are not available (Duchowicz *et al.*, 2008). Therefore, it is unreality to obtain the accurate toxicity data by performing experiments. A generally accepted strategy for overcoming the shortage of experimental measurements is the analysis based on Quantitative Structure–Activity Relationships (QSAR) (Hansch and Leo, 1995; Cronin and Dearden, 1995; Vighi *et al.*, 2001). Thus, in the past years, many attempts have been made to develop QSARs for the prediction of the toxicity of phenols and its derivatives to *Tetrahymena pyriformis* with different methods and different descriptors based on the same data set (Cronin and Schultz, 1996; Garg *et al*, 2001; Cronin *et al*, 2002; Duchowicz *et al*, 2008). And these QSAR

*Corresponding author E-mail: wanggaoxue@126.com

studies having proved to be valuable in predict the toxicity and interpret the mechanism of toxic action for phenols. However, because of the complicate structure, diversiform groups and mother circus rings of phenols, it is impossible to ensure the accurate prediction of toxicity based on the limited descriptors. Generally, different descriptors represent different mechanisms of toxic action. Only well known the mechanism of toxic action we can give much more accurate prediction of toxicity. Therefore, it is an urgent need to develop newer descriptors to encode molecular features and chemical information from different dimensions. In this background, we have introduced a number of novel descriptors calculated by Molconn-Z software. To our knowledge, the data set used in this study was calculated by Molconn-Z software for the first time and there is still no model which is built based on the Molconn-Z molecular descriptors for predicting the toxicity of phenols. Further, using a combined set of descriptors encompassing the logP and Molconn-Z parameters for the assessment the toxicity of phenols was also novel.

In the present work, a dataset of toxicity values for 250 phenolic compounds and a total of 197 descriptors (including logP and Molconn-Z) were used to develop predictive QSAR models for the toxicity of diverse chemical to T. *pyriformis* by stepwise regression-multiple linear regression (MLR) and partial least squares (PLS) regression. To evaluate performance of logP and Molconn-Z parameters, the robust models obtained in the present study were compared with those models developed from corresponding method and other descriptors in literature. Significance of different parameters appearing in the final models in relation to the toxicity was discussed and the mechanism about the aquatic toxicity of phenols was interpreted.

MATERIALS & METHODS

The aqueous toxicities are expressed as $pIGC_{50} = \log (IGC_{50})^{-1}$, with IGC_{50} expressing the concentration (mmol/L) producing a 50% growth inhibition on T. *pyriformis* under a static regime. A dataset of toxicity values ($pIGC_{50}$) for 250 phenols compounds in this study was obtained from the literature (Cronin *et al*., 2002). The dataset was randomly divided into two groups, a training set and a validation set, with approximately one forth of the total compounds were assigned in the test set. The training set containing 187 compounds was used to develop prediction models, and the test set including 63 compounds was used for the assessment of these models. All these compounds with their chemical names, CAS#s, Smiles and values of experimental toxicity employed in the study are provided in Table 1 (supplementary information).

In this study, the molecular descriptors were calculated using Molconn-Z program (version 4.10) based on the SMILES format of all compounds. The Molconn-Z software is capable of calculating a wide range of topological indices of molecular structure, including the molecular connectivity chi indices, $m\chi t$ and $m\chi tv$; kappa shape indices, m_k and m_k; electrotopological state indices, S_i; hydrogen electrotopological state indices, HES_i; atom type and bond type electrotopological state indices; new group type and bond type electrotopological state indices; topological equivalence indices and total topological indices; several information indices, such as the Shannon and the Bonchev-Trinajstiç information indices; counts of graph paths, atoms, atoms types, bond types; and others. Another important descriptor, i.e., 1-octanol/water partition coefficient (Log P), MlogP and AlogP, was calculated by Dragon software. And ClogP was calculated by Hansch-Leo's logP calculation method (Hansch and Leo, 1979). In total, the most widely used 197 indices were calculated. The descriptors with > 99% zeros were excluded, and the remaining 95 descriptors (including 3 logP and 92 Molconn-Z descriptors) are used for further analysis.

QSAR models were developed from training set using stepwise-MLR and PLS method. The MLR models development consists two parts. Firstly, we developed QSARs for the assessment of aquatic toxicity among all data using only logP descriptors, i.e., AlogP, MlogP and ClogP, to allow an easier explanation of the aquatic toxicity mechanism of phenols. Next, the stepwise-MLR methodology was used to investigate QSARs to prediction of aquatic toxicity of phenols using logP descriptors combined with Molconn-Z descriptors, i.e., MlogP + Molconn-Z descriptors, AlogP + Molconn-Z descriptors, ClogP + Molconn-Z descriptors, and all the calculated descriptors. One model was developed by PLS based on all the calculated descriptors. All the models used the same training set for model development and the same test set for external validation. Fit of the resultant MLR models were quantified with the goodness of fit (R^2), the standard error of prediction for training set (S_{EP}), the goodness of prediction (Q^2) and the standard error of estimation for test set (S_{EE}), the Fisher ratio (F). The quantitative measure of model performance for PLS model is given by the R^2 and Q^2 values, which give the fraction explained variance for training and validation data, respectively.

Linear regression analysis provides one of the most widely used statistical methods largely because it is simple to apply, easy to interpret and always gives good results when solving a wide range of problems (Hand, 1981). In this study, a variable selection

Table 1.The chemical names, CAS#s, Smiles and values of experimental toxicity employed in the study (Continue)

ID	Name	CAS#	Smiles	Toxicity
1	4-Hydroxyphenylacetic acid	000156-38-7	O=C(O)Cc(ccc(O)c1)c1	-1.5
2	3-Hydroxybenzyl alcohol	000620-24-6	OCc1cccc(O)c1	-1.04
3	4-Carboxyphenol	000099-96-7	O=C(O)c(ccc(O)c1)c1	-1.02
4	3-Hydroxy-4-methoxybenzyl alcohol	004383-06-6	COc1ccc(CO)cc1O	-0.99
5	4-Hydroxy-3-methoxybenzyl amine	007149-10-2	COC1=C(O)C=CC(=C1)CN	-0.82
6	4-Hydroxyphenethyl alcohol	000501-94-0	OCCc(ccc(O)c1)c1	-0.83
7	3-Carboxyphenol	000099-06-9	O=C(O)c(cccc1O)c1	-0.81
8	4-Hydroxybenzamide	000619-57-8	NC(=O)c1ccc(O)cc1	-0.78
9	4-Hydroxy-3-methoxybenzyl alcohol	000498-00-0	COc1cc(CO)ccc1O	-0.7
10	2,6-Dimethoxyphenol	000033-51-2	COC1=CC=CC(=C1O)OC	-0.6
11	2,4,6-Tris(dimethylaminomethyl) phenol	000090-72-2	Oc(c(cc(c1)CN(C)C)CN(C)C)c1CN(C)C	-0.52
12	Salicylic acid	000069-72-7	O=C(O)c(c(O)ccc1)c1	-0.51
13	2-Methoxyphenol	000090-05-1	O(c(c(O)ccc1)c1)C	-0.51
14	5-Methylresorcinol	000504-15-4	Oc(cc(O)cc1C)c1	-0.39
15	4-Methylcyanophenol	000055-55-0	CC1=CC=C(O)C(=C1)C#N	-0.38
16	3-Hydroxyacetophenone	000121-71-1	O=C(c(cccc1O)c1)C	-0.38
17	2-Ethoxyphenol	000094-71-3	O(c(c(O)ccc1)c1)CC	-0.36
18	4-Acetylphenol	000099-93-4	O=C(c(ccc(O)c1)c1)C	-0.3
19	3-Ethoxy-4-methoxyphenol	000150-76-5	O(c(ccc(O)c1)c1)C	-0.3
20	2-Methylphenol	000095-48-7	Oc(c(cccc1)C)c1	-0.29
21	2-Hydroxybenzamide	000065-45-2	O=C(N)c(c(O)ccc1)c1	-0.24
22	Phenol	000108-95-2	Oc(ccccc1)c1	-0.21
23	4-Methylphenol	000106-44-5	Oc(ccc(c1)C)c1	-0.18
24	4-Hydroxy-3-methoxyphenethylalcohol	002380-78-1	CCOC1=CC(=CC=C1OC)O	-0.18
25	3-Acetamidophenol	000621-42-1	O=C(Nc(cccc1O)c1)C	-0.16
26	3-Hydroxy-4-methoxybenzaldehyde	000621-59-0	O=Cc(ccc(OC)c1O)c1	-0.14
27	4-Hydroxy-3-methoxyacetophenone	000498-02-2	O=C(c(ccc(O)c1OC)c1)C	-0.12
28	3,5-Dimethoxyphenol	000500-99-2	COc1cc(O)cc(OC)c1	-0.09
29	2-Hydroxyethylsalicylate	000087-28-5	OCCOC1=CC=CC=C1C([O-])=O	-0.08
30	3-Methylphenol	000108-39-4	Oc(ccccc1C)c1	-0.06
31	Methyl-3-hydroxybenzoate	019438-10-9	COC(=O)c1cc(O)ccc1	-0.05
32	3-Methoxy-4-hydroxybenzaldehyde	000121-33-5	O=Cc(ccc(O)c1OC)c1	-0.03
33	4-Hydroxy-3-methoxybenzonitrile	004421-08-3	Oc(ccc1C#N)c(c1)OC	-0.03
34	3-Ethoxy-4-hydroxybenzaldehyde	000121-32-4	O=Cc(ccc(O)c1OCC)c1	0.01
35	4-Fluorophenol	000371-41-5	Fc(ccc(O)c1)c1	0.02
36	2-Cyanophenol	000611-20-1	OC1=CC=CC=C1C#N	0.03
37	5-Fluoro-2-hydroxyacetophenone	000394-32-1	CC(=O)C2=C(O)C=CC(=C2)F	0.04
38	2,4-Dimethylphenol	000105-67-9	Oc(c(cc(c1)C)C)c1	0.07
39	2-Hydroxyacetophenone	000118-93-4	OCC(=O)C1=CC=CC=C1	0.08
40	2,5-Dimethylphenol	000095-87-4	Oc(c(cccc1C)C)c1	0.08
41	Methyl-4-hydroxybenzoate	000099-76-3	O=C(OC)c(ccc(O)c1)c1	0.08
42	3,5-Dimethylphenol	000108-68-9	Oc(cc(cc1C)C)c1	0.11

Table 1.The chemical names, CAS#s, Smiles and values of experimental toxicity employed in the study (Continues)

43	4'-Hydroxypropiophenone	000070-70-2	O=C(c(ccc(O)c1)c1)CC	0.12
44	2,3-Dimethylphenol	000526-75-0	Oc(c(c(cc1)C)C)c1	0.12
45	3,4-Dimethylphenol	000095-65-8	Oc(ccc(c1C)C)c1	0.12
46	2-Ethylphenol	000090-00-6	Oc(c(cccl)CC)c1	0.16
47	Syringaldehyde	000134-96-3	O=Cc(cc(OC)c(O)c1OC)c1	0.17
48	Salicylhydrazide	000936-02-7	c1cccc(O)c1C(=O)NN	0.18
49	2-Chlorophenol	000095-57-8	Oc(c(cccl)Cl)c1	0.18
50	4-Hydroxy-2-methylacetophenone	000875-59-2	Oc1cc(C)c(C(=O)C)cc1	0.19
51	4-Ethylphenol	000123-07-9	Oc(ccc(c1)CC)c1	0.2
52	3-Ethylphenol	000620-17-7	Oc1cc(CC)ccc1	0.23
53	Salicylaldoxime	000094-67-7	N(O)=Cc(c(O)ccc1)c1	0.25
54	2,3,6-Trimethylphenol	002416-94-6	Oc(c(cccl C)C)c1C	0.28
55	2,4,6-Trimethylphenol	000527-60-6	Oc(c(cc(c1)C)C)c1C	0.28
56	2-Hydroxy-5-methylacetophenone	001450-72-2	O=C(c(c(O)ccc1C)c1)C	0.31
57	2-Bromophenol	000095-56-7	Oc(c(cccl)Br)c1	0.33
58	5-Bromo-2-hydroxybenzyl alcohol	002316-64-5	c1c(O)c(CO)cc(Br)c1	0.34
59	2,3,5-Trimethylphenol	000697-82-5	Oc(c(c(cc1C)C)C)c1	0.36
60	3-Methoxysalicylaldehyde	000148-53-8	O=Cc(c(O)c(OC)ccl)c1	0.38
61	Salicylhydroxamic acid	000089-73-6	ONC(=O)c1c(O)cccc1	0.38
62	2-Chloro-5-methylphenol	000615-74-7	Oc(c(cccl C)Cl)c1	0.39
63	4-Allyl-2-methoxyphenol	000097-53-0	O(c(c(O)ccc1CC=C)c1)C	0.42
64	2-Hydroxybenzaldehyde	000090-02-8	O=Cc(c(O)ccc1)c1	0.42
65	2,6-Difluorophenol	028177-48-2	Fc1cccc(F)c1O	0.47
66	Ethyl-3-hydroxybenzoate	007781-98-8	c1c(O)ccccl C(=O)OCC	0.48
67	4-Cyanophenol	000767-00-0	C(#N)c(cc c(O)c1)c1	0.52
68	4-Propyloxyphenol	018979-50-5	O(c(ccc(O)c1)c1)CCC	0.52
69	4-Chlorophenol	000106-48-9	Oc(ccc(c1)Cl)c1	0.55
70	Ethyl-4-hydroxybenzoate	000120-47-8	O=C(OCC)c(ccc(O)c1)c1	0.57
71	5-Methyl-2-nitrophenol	000700-38-9	CC1=CC=C(C(=C1)O)[N+]([O-])=O	0.59
72	2-Bromo-4-methylphenol	006627-55-0	CC1=CC=C(O)C(=C1)Br	0.6
73	2,4-Difluorophenol	000367-27-1	Cl(F)=CC(F)=C(O)C=Cl	0.6
74	3-Isopropylphenol	000618-45-1	Oc(ccccl C(C)C)c1	0.61
75	5-Bromovanillin	002973-76-4	O=Cc(cc(OC)c(O)c1Br)c1	0.62
76	α,α,α-Trifluoro-4-cresol	000402-45-9	Oc1ccc(C(F)(F)F)cc1	0.62
77	Methyl-4-methoxysalicylate	005446-02-6	COC(=O)c1c(O)cc(OC)cc1	0.62
78	4-Bromophenol	000106-41-2	Oc(ccc(c1)Br)c1	0.68
79	2-Chloro-4,5-dimethylphenol	001124-04-5	Cc1cc(O)c(Cl)cc1C	0.69
80	4-Butoxyphenol	000122-94-1	O(c(ccc(O)c1)c1)CCCC	0.7
81	4-Chloro-2-methylphenol	001570-64-5	Oc(c(cc(c1)Cl)C)c1	0.7
82	3-tert-Butylphenol	000585-34-2	Oc(ccccl C(C)(C)C)c1	0.73
83	2,6-Dichlorophenol	000087-65-0	Oc(c(cccl)Cl)c1Cl	0.73
84	2-Methoxy-4-propenylphenol	005932-68-3	Oc(ccc1C=CC)c(c1)OC	0.75
85	3-Chloro-5-methoxyphenol	082477-68-7	COC1=CC(=CC(=C1)O)Cl	0.76
86	4-Chloro-3-methylphenol	000059-50-7	Oc(ccc(cl C)Cl)c1	0.8

**Table 1.The chemical names, CAS#s, Smiles and values of experimental toxicity employed in the study
(Continues)**

87	2-Isopropylphenol	000088-69-7	Oc(cccc1)C(C)C)c1	0.8
88	2,6-Dichloro-4-fluorophenol	000392-71-2	OC1=C(Cl)C=C(F)C=C1Cl	0.8
89	4-Iodophenol	000540-38-5	Oc(ccc(c1)I)c1	0.85
90	2,2'-Biphenol	001806-29-7	Oc(c(c(c(O)ccc1)c1)ccc2)c2	0.88
91	4-tert-Butylphenol	000098-54-4	Oc(ccc(c1)C(C)(C)C)c1	0.91
92	3,4,5-Trimethylphenol	000527-54-8	Oc(cc(c(c1C)C)C)c1	0.93
93	2,2',4,4'-Tetrahydroxybenzophenone	000131-55-5	O=C(c(c(c(O)cc(O)c1)c1)c(c(O)cc(O)c2)c2	0.96
94	4-sec-Butylphenol	000099-71-8	Oc(ccc(c1)C(CC)C)c1	0.98
95	3-Hydroxydiphenylamine	000101-18-8	Oc(cccc1Nc(cccc2)c2)c1	1.01
96	4-Hydroxybenzophenone	001137-42-4	O=C(c(cccc1)c1)c(ccc(O)c2)c2	1.02
97	2,4-Dichlorophenol	000120-83-2	Oc(c(cc(c1)Cl)Cl)c1	1.04
98	2,4,6-Tribromoresorcinol	002437-49-2	Oc1c(Br)cc(Br)c(O)c1Br	1.06
99	Benzyl-4-hydroxyphenyl ketone	002491-32-9	OC1=CC=C(C(=O)C2=CC=C(O)C=C2CC3=CC=CC=C3)C(=C1)CC4=CC=CC=C4	1.07
100	4-Chloro-3-ethylphenol	014143-32-9	Oc1cc(CC)c(Cl)cc1	1.08
101	2-Phenylphenol	000090-43-7	Oc(c(c(cccc1)c1)ccc2)c2	1.09
102	2,5-Dichlorophenol	000583-78-8	Oc(c(ccc1Cl)Cl)c1	1.13
103	3-Chloro-4-fluorophenol	002613-23-2	Oc1cc(Cl)c(F)cc1	1.13
104	3-Bromophenol	000591-20-8	Oc(cccc1Br)c1	1.15
105	6-tert-Butyl-2,4-dimethylphenol	001879-09-0	Oc(c(cc(c1)C)C(C)(C)C)c1C	1.16
106	4-Chloro-3,5-dimethylphenol	00088-04-0	Oc(cc(c(c1C)Cl)C)c1	1.2
107	2-Hydroxybenzophenone	000117-99-7	O=C(c(cccc1)c1)c(c(O)ccc2)c2	1.23
108	4-tert-Pentylphenol	000080-46-6	Oc(ccc(c1)C(CC)(C)C)c1	1.23
109	4-Bromo-3,5-dimethylphenol	007463-51-6	Oc1cc(C)c(Br)c(C)c1	1.27
110	4-Bromo-6-chloro-2-cresol	007530-27-0	Oc1c(Cl)cc(Br)cc1C	1.28
111	4-Cyclopentylphenol	001518-83-8	Oc1ccc(cc1)C2CCCC2	1.29
112	2-tert-Butylphenol	000088-18-6	Oc(c(cccc1)C(C)(C)C)c1	1.29
113	2-tert-Butyl-4-methylphenol	002409-55-4	Oc(c(cc(c1)C)C(C)(C)C)c1	1.3
114	2-Hydroxydiphenylmethane	028994-41-4	Oc(c(cccc1)Cc(cccc2)c2)c1	1.31
115	Butyl-4-hydroxybenzoate	000094-26-8	O=C(OCCCC)c(ccc(O)c1)c1	1.33
116	3-Phenylphenol	000580-51-8	Oc(cccc1c(cccc2)c2)c1	1.35
117	n-Pentyloxyphenol	018979-53-8	O(c(ccc(O)c1)c1)CCCCC	1.36
118	2,4-Dibromophenol	000615-58-7	Oc(c(cc(c1)Br)Br)c1	1.4
119	2,4,6-Trichlorophenol	000088-06-2	Oc(c(cc(c1)Cl)Cl)c1Cl	1.41
120	2-Hydroxy-4-methoxybenzophenone	000131-57-7	O=C(c(cccc1)c1)c(c(O)cc(OC)c2)c2	1.42
121	Isoamyl-4-hydroxybenzoate	006521-30-8	O=C(OCCC(C)C)c(ccc(O)c1)c1	1.48
122	3,5-Dichlorosalicylaldehyde	000090-60-8	O=Cc(c(O)c(cc1Cl)Cl)c1	1.55
123	4-Cyclohexylphenol	001131-60-8	Oc(ccc(c1)C(CCCC2)C2)c1	1.56
124	3,5-Dichlorophenol	000591-35-5	Oc1cc(Cl)cc(Cl)c1	1.57
125	3,5-Di-tert-butylphenol	001138-52-9	Oc(cc(cc1C(C)(C)C)C(C)(C)C)c1	1.64
126	3,5-Dibromosalicylaldehyde	000090-59-5	Oc1c(Br)cc(Br)cc1C=O	1.64
127	3,4-Dichlorophenol	000095-77-2	Oc1ccc(Cl)c(Cl)c1	1.75
128	4-Bromo-2,6-dichlorophenol	003217-15-0	Brc(cc(Cl)c1O)cc1Cl	1.78
129	2,6-Di-tert-butyl-4-methylphenol	000128-37-0	Oc(c(cc(c1)C)C(C)(C)C)c1C(C)(C)C	1.8

Table 1.The chemical names, CAS#s, Smiles and values of experimental toxicity employed in the study (Continues)

130	4-Chloro-2-isopropyl-5-methylphenol	000089-68-9	Oc(c(cc(c1C)Cl)C(C)C)c1	1.85
131	2,4,6-Tribromophenol	000118-79-6	Oc(c(cc(c1)Br)Br)c1Br	2.03
132	4-Heptyloxyphenol	013037-86-0	O(c(ccc(O)c1)c1)CCCCCCC	2.03
133	4-tert-Octylphenol	003294-03-9	CCCCC(C)(C)C1=CC=C(O)C=C1	2.1
134	4-(4-Bromophenyl)phenol	029558-77-8	OC1=CC=C(C=C1)C2=CC=C(Br)C=C2	2.31
135	3,5-Diiodosalicylaldehyde	002631-77-8	O=Cc1c(c(cc(c1)I)I)O	2.34
136	2,3,5-Trichlorophenol	000933-78-8	Oc1cc(Cl)cc(Cl)c1Cl	2.37
137	4-Nonylphenol	000104-40-5	Oc(ccc(c1)CCCCCCCCC)c1	2.47
138	Nonyl-4-hydroxybenzoate	038713-56-3	CCCCCCCCCOC(=O)C1=CC=C(O)C=C1	2.63
139	2,4,6-Trinitrophenol	000088-89-1	OC1=C(C=C(C=C1[N+]([O-])=O)[N+]([O-])=O)[N+]([O-])=O	-0.16
140	3,4-Dinitrophenol	000577-71-9	OC1=CC=C(C(=C1)[N+]([O-])=O)[N+]([O-])=O	0.27
141	2,6-Dinitrophenol	000573-56-8	OC1=C(C=CC=C1[N+]([O-])=O)[N+]([O-])=O	0.54
142	2,6-Dichloro-4-nitrophenol	000618-80-4	OC1=C(Cl)C=C(C=C1Cl)[N+]([O-])=O	0.63
143	2,5-Dinitrophenol	000329-71-5	OC1=CC(=CC=C1[N+]([O-])=O)[N+]([O-])=O	0.95
144	2,4-Dinitrophenol	000051-28-5	OC1=CC=C(C=C1[N+]([O-])=O)[N+]([O-])=O	1.08
145	2,6-Dinitro-4-cresol	000609-93-8	CC1=CC(=C(O)C(=C1)[N+]([O-])=O)[N+]([O-])=O	1.23
146	4-Bromo-2-fluoro-6-nitrophenol	000320-76-3	OC1=C(C=C(Br)C=C1F)[N+]([O-])=O	1.62
147	Pentafluorophenol	000771-61-9	OC1=C(F)C(=C(F)C(=C1F)F)F	1.64
148	4,6-Dinitro-2-methylphenol	000534-52-1	CC1=CC(=CC(=C1O)[N+]([O-])=O)[N+]([O-])=O	1.72
149	2,4-Dichloro-6-nitrophenol	000609-89-2	OC1=C(C=C(Cl)C=C1Cl)[N+]([O-])=O	1.75
150	Pentachlorophenol	000087-86-5	Oc(c(c(c(c1Cl)Cl)Cl)Cl)c1Cl	2.05
151	2,3,5,6-Tetrachlorophenol	000935-95-5	Oc1c(Cl)c(Cl)cc(Cl)c1Cl	2.22
152	Pentabromophenol	000608-71-9	Oc(c(c(c(c1Br)Br)Br)Br)c1Br	2.66
153	2,3,4,5-Tetrachlorophenol	004901-51-3	Oc1cc(Cl)c(Cl)c(Cl)c1Cl	2.71
154	4-Acetamidophenol	000103-90-2	O=C(Nc(ccc(O)c1)c1)C	-0.82
155	3-Aminophenol	00591-27-5	Oc(cccc1N)c1	-0.52
156	4-Aminophenol	000123-30-8	Oc(ccc(N)c1)c1	-0.08
157	3-Methylcatechol	000488-17-5	Oc(c(cccl)C)c1O	0.28
158	2-Amino-4-tert-butylphenol	001199-46-8	Nc1c(O)ccc(C(C)(C)C)c1	0.37
159	4-Methylcatechol	000452-86-8	Oc(c(O)cc(c1)C)c1	0.37
160	1,2,4-Trihydroxybenzene	000533-73-3	Oc(c(O)cc(O)c1)c1	0.44
161	Hydroquinone	000123-31-9	Oc(ccc(O)c1)c1	0.47
162	Catechol	000120-80-9	Oc(c(O)ccc1)c1	0.75
163	2-Amino-4-chlorophenol	000095-85-2	Oc(c(N)cc(c1)Cl)c1	0.78
164	1,2,3-Trihydroxybenzene	000087-66-1	Oc(c(O)cccl)c1O	0.85
165	2-Aminophenol	000095-55-6	Oc(c(N)cccl)c1	0.94
166	4-Chlorocatecol	002138-22-9	Oc1ccc(Cl)cc1O	1.06
167	Chlorohydroquinone	000615-67-8	Oc(c(cc(O)c1)Cl)c1	1.26
168	4-Amino-2-cresol	002835-96-3	Oc(c(cc(N)c1)C)c1	1.31
169	2,3-Dimethylhydroquinone	000608-43-5	c1c(O)c(C)c(C)c(O)c1	1.41
170	4-Amino-2,3-dimethylphenol	003096-69-3	Nc1c(C)c(C)c(O)cc1	1.44
171	Bromohydroquinone	000583-69-7	Oc(c(cc(O)c1)Br)c1	1.68

Table 1.The chemical names, CAS#s, Smiles and values of experimental toxicity employed in the study (Continues)

172	Tetrachlorocatechol	001198-55-6	ClOC1=CC=C(Cl)C(=Cl OCl)Cl	1.7
173	Phenylhydroquinone	001079-21-6	Oc(c(c(cccc1)c1)cc(O)c2)c2	2
174	3,5-Di-tert-butylcatechol	001020-31-1	Oc(c(cc(c1)C(C)(C)C)C(C)(C)C)c1O	2.11
175	Methoxyhydroquinone	000824-46-4	c1c(O)c(OC)cc(O)c1	2.2
176	3-Hydroxy-4-nitrobenzaldehyde	000704-13-2	OC1=CC(=CC=C1 [N+]([O-])=O)C=O	0.27
177	5-Hydroxy-2-nitrobenzaldehyde	042454-06-8	OC1=CC=C(C(=C1)C=O)[N+]([O-])=O	0.33
178	2-Amino-4-nitrophenol	061702-43-0	NC1=CC(=CC=C1O)[N+]([O-])=O	0.47
179	4-Methyl-2-nitrophenol	000099-53-6	CC1=CC=C(O)C(=C1)[N+]([O-])=O	0.57
180	4-Hydroxy-3-nitrobenzaldehyde	003011-34-5	OC1=CC=C(C=O)C=C1[N+]([O-])=O	0.61
181	4-Nitrosophenol	000104-91-6	O=Nc(ccc(O)c1)c1	0.65
182	2-Nitroresorcinol	000601-89-8	OC1=CC=CC(=C1 [N+]([O-])=O)O	0.66
183	4-Methyl-3-nitrophenol	002042-14-0	CC1=CC=C(O)C=C1[N+]([O-])=O	0.74
184	2-Chloromethyl-4-nitrophenol	002973-19-5	OC1=CC=C(C=C1 CCl)[N+]([O-])=O	0.75
185	2-Bromo-2'-hydroxy-5'-nitroacetanilide	003947-58-8	OC1=CC=C(C=C1 NC(=O)CBr)[N+]([O-])=O	0.87
186	4-Amino-2-nitrophenol	000119-34-6	NC1=CC=C(O)C(=C1)[N+]([O-])=O	0.88
187	2-Fluoro-4-nitrophenol	000403-19-0	OC1=CC=C(C=C1 F)[N+]([O-])=O	1.07
188	5-Fluoro-2-nitrophenol	000446-36-6	OC1=CC(=CC=C1 [N+]([O-])=O)F	1.13
189	4-Nitrocatechol	003316-09-4	OC1=CC=C(C=C1O)[N+]([O-])=O	1.17
190	2-Amino-4-chloro-5-nitrophenol	006358-07-2	NC1=CC(=C(C=C1O)[N+]([O-])=O)Cl	1.17
191	4-Fluoro-2-nitrophenol	000394-33-2	OC1=CC=C(F)C=C1[N+]([O-])=O	1.38
192	4-Nitrophenol	000100-02-7	OC1=CC=C(C=C1)[N+]([O-])=O	1.42
193	2-Chloro-4-nitrophenol	000619-08-9	OC1=CC=C(C=C1 Cl)[N+]([O-])=O	1.59
194	4-Chloro-6-nitro-3-cresol	007147-89-9	CC1=CC(=C(C=C1Cl)[N+]([O-])=O)O	1.64
195	3-Methyl-4-nitrophenol	002581-34-2	CC1=CC(=CC=C1[N+]([O-])=O)O	1.73
196	4-Bromo-2-nitrophenol	007693-52-9	OC1=CC=C(Br)C=C1[N+]([O-])=O	1.87
197	4-Chloro-2-nitrophenol	000089-64-5	OC1=CC=C(Cl)C=C1[N+]([O-])=O	2.05
198	Tetrabromocatechol	000488-47-1	Oc1c(O)c(Br)c(Br)c(Br)c1Br	0.98
199	Tetramethylhydroquinone	000700-13-0	Oc(c(cc(O)c1C)C)c1 C	1.28
200	Tetrachlorohydroquinone	000087-87-6	Oc(c(c(c(O)c1Cl)Cl)Cl)c1Cl	2.11
201	1,3,5-Trihydroxybenzene	006099-90-7	Oc1 cc(O)cc(O)c1	-1.26
202	2-Hydroxybenzylalcohol	000090-01-7	OCc(c(O)ccc1)c1	-0.95
203	Resorcinol	000108-46-3	Oc(cccc1 O)c1	-0.65
204	4-(4-Hydroxyphenyl)-2-butanone	005471-51-2	O=C(CCc(ccc(O)c1)c1)C	-0.5
205	3-Methoxyphenol	000150-19-6	O(c(cccc1 O)c1)C	-0.33
206	Ethyl-4-hydroxy-3-methoxyphenylacetate	060563-13-5	CCOC(=O)Cc1cc(OC)c(O)cc1	-0.23
207	4-Methoxyphenol	000150-76-5	O(c(ccc(O)c1)c1)C	-0.14
208	3-Cyanophenol	000873-62-1	Oc1 cccc(C#N)c1	-0.06
209	4-Ethoxyphenol	000622-62-8	O(c(ccc(O)c1)c1)CC	0.01
210	4-Hydroxypropiophenone	000070-70-2	O=C(c(ccc(O)c1)c1)CC	0.05
211	3-Hydroxybenzaldehyde	000100-83-4	O=Cc(cccc1O)c1	0.09
212	4-Chlororesorcinol	000095-88-5	Oc(c(ccc1 O)Cl)c1	0.13
213	2-Fluorophenol	000367-12-4	Fc(c(O)ccc1)c1	0.19
214	4-Hydroxybenzaldehyde	000123-08-0	O=Cc(ccc(O)c1)c1	0.27
215	2-Allylphenol	001745-81-9	Oc(c(cccc1)CC=C)c1	0.33

Table 1.The chemical names, CAS#s, Smiles and values of experimental toxicity employed in the study
(Continues)

216	3-Fluorophenol	000372-20-3	Oc1cccc(F)c1	0.38
217	4-Isopropylphenol	000099-89-8	Oc(ccc(c1)C(C)C)c1	0.47
218	2-Hydroxy-4-methoxyacetophenone	000552-41-0	COc1ccc(C(C)=O)c(O)c1	0.55
219	3-Methyl-2-nitrophenol	004920-77-8	CC1=C(C(=CC=C1)O)[N+]([O-])=O	0.61
220	4-Propylphenol	000645-56-7	Oc(ccc(c1)CCC)c1	0.64
221	2-Hydroxy-4,5-dimethylacetophenone	036436-65-4	Oc1cc(C)c(C)cc1C(=O)C	0.71
222	2-Methyl-3-nitrophenol	005460-31-1	CC1=C(C=CC=C1O)[N+]([O-])=O	0.78
223	3-Chlorophenol	000108-43-0	Oc(cccc1Cl)c1	0.87
224	4,6-Dichlororesorcinol	000137-19-9	Oc(c(cc(c1O)Cl)Cl)c1	0.97
225	4-Benzyloxyphenol	000103-16-2	O(c(ccc(O)c1)c1)Cc(cccc2)c2	1.04
226	3-Iodophenol	000626-02-8	Oc(cccc1I)c1	1.12
227	4-Bromo-2,6-dimethylphenol	002374-05-2	Oc(c(cc(c1)Br)C)c1C	1.17
228	2,3-Dichlorophenol	000576-24-9	Oc(c(c(cc1)Cl)Cl)c1	1.28
229	5-Pentylresorcinol	000500-66-3	CCCCCC1=CC(=CC(=C1)O)O	1.31
230	4-Phenylphenol	000092-69-3	Oc(ccc(c(cccc1)c1)c2)c2	1.39
231	Benzyl-4-hydroxybenzoate	000094-18-8	O=C(OCc(cccc1)c1)c(ccc(O)c2)c2	1.55
232	4-Hexyloxyphenol	018979-55-0	O(c(ccc(O)c1)c1)CCCCCC	1.64
233	4-Hexylresorcinol	000136-77-6	Oc(c(cccl O)CCCCCC)c1	1.8
234	2,4,5-Trichlorophenol	000095-95-4	Oc(c(cc(c1Cl)Cl)Cl)c1	2.1
235	2-Ethylhexyl-4'-hydroxybenzoate	005153-25-3	CCCCC(CC)CC1=CC(=CC=C1C([O-])=O)O	2.51
236	2,3-Dinitrophenol	000066-56-8	OC1=CC=CC(=C1[N+]([O-])=O)[N+]([O-])=O	0.46
237	2,3,5,6-Tetrafluorophenol	000769-39-1	Oc1c(F)c(F)cc(F)c1F	1.17
238	2,6-Diiodo-4-nitrophenol	000305-85-1	OC1=C(I)C=C(C=C1I)[N+]([O-])=O	1.71
239	3,4,5,6-Tetrabromo-2-cresol	000576-55-6	Oc(c(c(c(c1Br)Br)Br)C)c1Br	2.57
240	2,4-Diaminophenol	000137-09-7	NC1=CC=C(O)C(=C1)N	0.13
241	5-Amino-2-methoxyphenol	001687-53-2	Nc1cc(O)c(OC)cc1	0.45
242	6-Amino-2,4-dimethylphenol	041458-65-5	Oc(c(cc(c1)C)C)c1N	0.89
243	Trimethylhydroquinone	000700-13-0	Oc(c(cc(O)c1C)C)c1C	1.34
244	Methylhydroquinone	096937-50-7	CC1=CC(=CC=C1O)O	1.86
245	3-Nitrophenol	000554-84-7	OC1=CC=CC(=C1)[N+]([O-])=O	0.51
246	2-Nitrophenol	000088-75-5	OC1=CC=CC=C1[N+]([O-])=O	0.67
247	3-Fluoro-4-nitrophenol	000394-41-2	OC1=CC=C(C(=C1)F)[N+]([O-])=O	0.94
248	2,6-Dibromo-4-nitrophenol	000099-28-5	OC1=C(Br)C=C(C=C1Br)[N+]([O-])=O	1.36
249	4-Nitro-3-(trifluoromethyl)phenol	000088-30-2	OC1=CC=C(C(=C1)C(F)(F)F)[N+]([O-])=O	1.65
250	Tetrafluorohydroquinone	000771-63-1	Fc(c(F)c(O)c1F)c(O)c1F	1.84

approach was applied in the linear regression analysis process: Stepwise technique (Jenrich, 1960). The stepwise-MLR process builds up a model through stepwise addition of descriptors, where the inclusion of a given descriptor is based on the F statistic values. A deletion process is then employed where each independent variable is held out in turn, and a model is developed by using the remaining pool of descriptors. Then all pairs and triplets are held out, and the process is repeated. In this work, the stepwise-MLR was performed using a free R-statistical software.

Partial Least Squares (PLS), introduced by Wold et al. (1984), is a combination of MLR and PCA method which has received a great of attention in the field of chemometrics, bioinformatics, medicine, pharmacology and others (Nguyen and Rocke, 2002). It attempts to explain the variance in the independent variables and tries to obtain a good correlation between the dependent and the independent variables. The original descriptors was reduced to a smaller number of latent variables called PLS components (PC) that were used as independent variables while the toxicity served as the response variable. The PCs contained most of the information in the independent variables that was useful for predicting dependent variables, while reducing the dimensionality of the regression problem by using fewer components than the number of independent variables. Partial least squares (PLS) are considered as an especially useful method for

constructing predictive models when the factors are many and highly collinear.

The quality of the final optimized equations obtained via the MLR approach is judged by means of criteria: the Kubinyi function (FIT) (Kubinyi, 1994). The FIT (d) criterion has a low sensitivity to changes in small d values and a substantially increasing sensitivity for large d values. It is given by

$$FIT = \frac{R(d)^2 (N - d - 1)}{(N + d^2)(1 - R^2)} \qquad (1)$$

Where N is the number of molecules in the training set, d is the number of descriptors, $R(d)$ is the correlation coefficient for a model with d descriptors. The greater the FIT value the better the linear equation. The optimal number of molecular descriptors to be included in the linear regression model (d_{opt}) is deduced from the plot of FIT vs d, as the Kubinyi function achieves a maximum value (d_{max}). And the following criterion was adopted for determining d_{opt}:

1. Calculate $d_1 = [d_{max}/2] + 1$, where $[x]$ denotes the integer part of x.

2. If the slope of FIT at d_1 is greater than at d_1+1, then $d_{opt}=d_1$, otherwise, $d_{opt}=d_1+1$.

By means of this criterion, it is expected to obtain a d_{opt} value that reflects a "breaking point" beyond which the FIT improvement is negligible (Duchowicz *et al.*, 2008).

A crucial problem for the obtained QSAR model is the definition of its applicability domain (AD) (Xia *et al.*, 2009). For any QSAR model, only the predictions for chemicals falling within its AD can be considered reliable and not model extrapolations (Pan *et al.*, 2009). Several methods were reported for defining the AD of QSAR models, but the most common one is determining the leverage values for each compound (Gramatica, 2007). In the present work, the Williams plot, the plot of the standardized residuals versus the leverage, was exploited to visualize the AD of a QSAR model. The distance of the chemical from the centroid of its training set was measured by the leverage of a chemical. And the leverage of a compound in the original variable space is defined as (Netzeva *et al.*, 2005):

$$h_i = x_i^T (X^T X)^{-1} x_i \qquad (2)$$

Where x_i is the descriptor vector of the considered compound and X is the descriptor matrix derived from the descriptor values of training set. A "warning leverage" (h^*) is defined as (Eriksson *et al.*, 2003):

$$h^* = 3(p + 1)/n \qquad (3)$$

Where n is the number of training compounds, p is the number of predictor variables. A leverage value which is greater than the warning leverage is considered large.

Outliers to each of the models developed were identified on the basis of having a high standardised residual more than 2.5 times the standard error of estimate for a particular equation (Verma and Hansch, 2005). QSAR models that have no or very few outliers are considered as good models. Thus, in order to develop good models, outliers were removed and re-analysis. Then, the mechanisms of these removed outliers were given a reasonable interpretation from statistical data and structure of compound.

RESULTS & DISCUSSION

We first built three models (Table 2) using ALogP, MlogP and ClogP individually. ClogP model has poor prediction with a R2 of 0.03 and Q² of 0.04. MlogP and AlogP model, derived from MlogP and AlogP descriptor respectively, show better prediction capability than ClogP model. Interestingly, MlogP exhibts stronger modeling power than AlogP, and the prediction accuracies were 52% for MlogP and 39% for AlogP.

Then, 3 logP and 92 Molconn-Z descriptors were used to built QSARs by stepwise-MLR method and the criteria of Kubinyi function was applied for the optimal of molecular descriptors. The resultant QSARs were summarized in Table 1 and the definition of descriptors involved in the model were shown in Table 3.

The following equation was developed from a stepwise selection of ClogP + Molconn-Z descriptors and the d_{opt} was six.

$$(4)$$

pIGC$_{50}$ = − 0.59(±0.139)
− 0.38(±0.090) SsssN + 0.15(±0.027) SssCH2

− 0.08(±0.018) SssO + 0.05(±0.009) SsCl
− 0.05(±0.008) SHBint2 + 0.01(±0.001) molweight

$N_{tra} = 187, N_{test} = 63, R^2 = 0.60, S_{EP} = 0.56, F = 44.73, Q^2 = 0.48, S_{EE} = 0.53, FIT = 1.21$

Because of the low prediction accuracy (Q² = 0.48) and large gap between R² and Q² for Eq. 4, the model was not further analyzed.

The following relationship was found between the toxicity of the phenols to T. *pyriformis* and MlogP + Molconn-Z descriptors. The stepwise-MLR and sensitivity analysis result the following equation:

Table 2. Summary of all MLR models developed in this study and relative parameters

Equations	R^2	S_{EP}	F	Q^2	S_{EE}	d_{opt}
$pIGC_{50}=+0.74(\pm0.063 + 3.36\times10^{-6}(\pm0.000)ClogP$	0.03	0.85	5.66	0.04	0.70	—
$pIGC_{50}=-0.47(\pm0.104) + 0.54(\pm0.041)AlogP$	0.48	0.62	172.61	0.39	0.56	—
$pIGC_{50}=-0.55(\pm0.112) + 0.63(\pm0.049)MlogP$	0.47	0.63	164.12	0.52	0.50	—
Eq. 4	0.60	0.56	44.73	0.48	0.53	6
Eq. 5	0.66	0.51	58.76	0.60	0.47	6
Eq. 6	0.71	0.46	106.92	0.69	0.41	4
Eq. 7	0.69	0.49	66.30	0.67	0.43	6
Eq. 8	0.70	0.47	84.31	0.68	0.42	5

Table 3. The symbols and definitions of the molecular descriptors in this work

Class	Descriptor	Description
logP	AlogP	octanol-water partition coefficient based on Atom Type Classification
	ClogP	octanol-water partition coefficient based on the constants of these fragments and correction factors
	MlogP	octanol-water partition coefficient based on Molecular Type Classification
E-state	SsssN	Sum of atom-type E-State: >N-
	SssCH2	Sum of atom-type E-State: -CH2-
	SssO	Sum of atom-type E-State: -O-
	SsCl	Sum of atom-type E-State: -Cl
	SHBint2	E-State descriptors of potential internal H bond strength
	SHCsatu	E-State of Csp3 bonded to unsaturated C atoms
	SHsOH	Atom type electrotopological state index values for atom types
	Scarboxylicacid	E-State of carboxylic acid
	SwHBa	weak H bond acceptor index, sum of E-State values for >N-, -O-, =O, -S- along with -F, and –Cl
	gmax	Extreme atom level E-State values in molecule: Gmax—Largest E-State value
	SHBa	Acceptor descriptor for molecule (sum of E-state values for all hydrogen bond acceptors in the molecule). The following groups are classified as acceptors: -OH, =NH, -NH2, -NH-, >N-, -O-, =O, -S- along with -F and -Cl.
Molecular connectivity	nclass	number of classes of topologically (symmetry) equivalent graph vertices
	nelem	number of elements in molecule
Molecular weight	molweight	Molecular weight

$pIGC_{50} = -1.341(\pm0.215) + 0.69(\pm0.043)$ MlogP
$-0.32(\pm0.076)$ SHCsatu $+0.14(\pm0.031)$ ShsOH   (5)
$-0.09(\pm0.015)$ Scarboxylicacid $+0.07(\pm0.018)$ nclass
$-0.04(\pm0.009)$ SwHBa

$N_{tra} = 187, N_{test} = 63, R^2 = 0.66, S_{EP} = 0.51, F = 58.76, Q^2 = 0.60, S_{EE} = 0.47, FIT = 1.58$

From residuals (not shown) of Eq. 5, it can be seen that the model does not provide adequate predictive power for the assessment of toxicity. Eleven compounds with residual values over 2.5S, nine in training set and two in test set, were identified as statistical outliers (Table 4). Removal of these compounds and re-analysis reveals a significant equation with $d_{opt} = 4$:

$pIGC_{50} = +0.65(\pm0.323) + 0.73(\pm0.039)$ MlogP   (6)
$-0.22(\pm0.038)$ gmax

$-0.07(\pm0.013)$ Scarboxylicacid $+0.03(\pm0.004)$ SHBa

$N_{tra} = 178, N_{test} = 61, R^2 = 0.71, S_{EP} = 0.46, F = 106.92,$
$Q^2 = 0.69, S_{EE} = 0.41, FIT = 2.21$

The resulted correlation between experimental and predicted $pIGC_{50}$ values of this model is shown in Fig. 1 (a). In this figure, plots in the training and test data sets are all well distribution around the regression line. Table 1 presents the detailed statistics of the model we obtained, where it is clear that the R^2 (= 0.71) and Q^2 (= 0.69) have a significant improvement and the number of variables decrease after removing these outliers.

The Williams plot (Fig. 1 (b)) shows that six chemicals in the training set with h > h* (h* = 0.06) and the standardized residuals < 2.5S. However, all the six chemicals in the training set fit the model well. Another two compounds, one in the training set and the other

Table 4. The structures of outliers in MLR and PLS models

| Compound	structure	Outliers to equations	Compound	structure	Outliers to equations
Compound 2		Eq.(6)	compound 175		Eq.(5), Eq.(7), PLS
Compound 4		Eq.(5)	Compound 192		Eq.(5)
Compound 138		Eq.(6)	Compound 195		Eq.(5)
Compound 139		Eq.(5), Eq.(8)	Compound 198		Eq.(5), Eq.(7), PLS
Compound 142		Eq.(7)	ompound 201		Eq.(5), Eq.(7), PLS
Compound 168		Eq.(5)	Compound 238		Eq.(7)
Compound 169		Eq.(5)	compound 244		Eq.(5), Eq.(7), PLS
Compound 170		Eq.(5)			
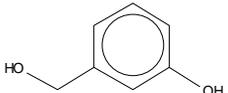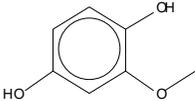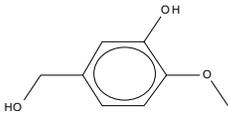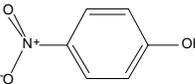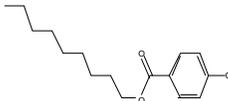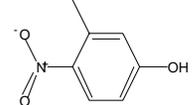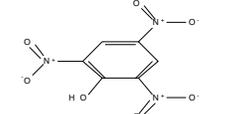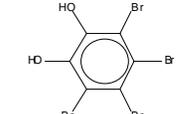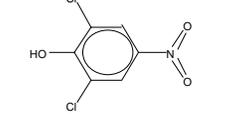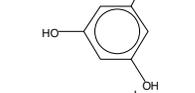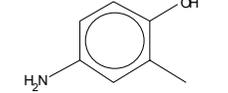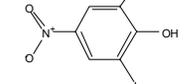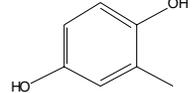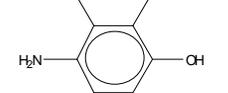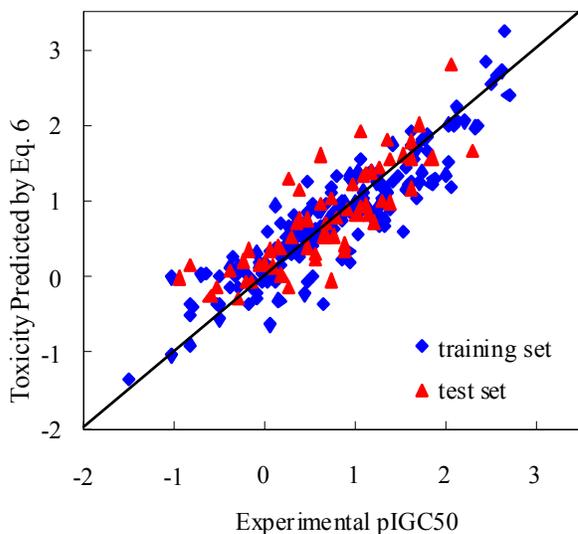
**Fig. 1(a). Plot of the experimental pIGC$_{50}$ to *T. pyriformis* against toxicity predicted by Eq. (6)**
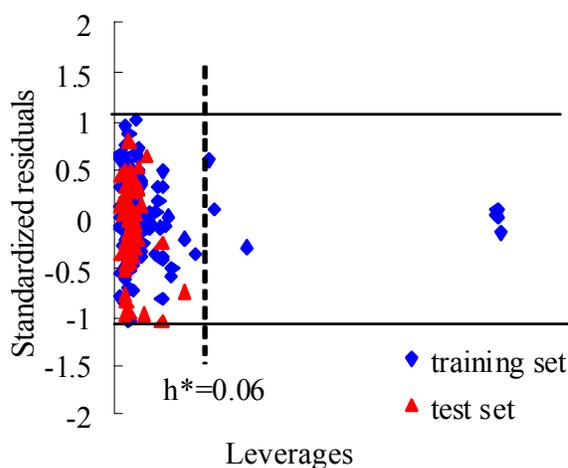
Fig. 1(b). Plot of standardized residuals versus leverages. Lines represent ±2.5 standardized residuals, dotted line represents warning leverage (h* = 0.06)

in the test set, with h < h* and the standardized residuals < 2.5S, were considered as outliers. These outliers were compound 2 (3-Hydroxybenzyl alcohol) and 138 (Nonyl-4-hydroxybenzoate).

Based on AlogP, MlogP, ClogP and Molconn-Z descriptors, stepwise-MLR and sensitive analysis result the following equation:

(7)

$$pIGC_{50} = -1.512(\pm0.306) + 0.580(\pm0.034) \, AlogP$$
$$- 0.314(\pm0.075) \, SHCsatu + 0.189(\pm0.066) \, nelem$$
$$+ 0.070(\pm0.017) \, nclass - 0.060(\pm0.014)$$
$$Scarboxylicacid - 0.043(\pm0.009) \, SwHBa$$

$N_{tra} = 187, N_{test} = 63, R^2 = 0.69, S_{EP} = 0.49, F = 66.37, Q^2 = 0.67, S_{EE} = 0.43, FIT = 1.79$

Intrestingly, a model developed from AlogP + Molconn-Z descriptors shared the same equation as Eq. (7). From the residuals point of view, the Eq. 7 was not an adequate model to estimate the pIGC$_{50}$ in this study. Six outliers (Table 3) with a residual exceeding the value 2.5S were observed. Removal of these outliers and re-analysis reveal a significant model with $d_{opt} = 5$:

(8)

$$pIGC_{50} = -1.45(\pm0.263) + 0.59(\pm0.036) \, AlogP$$
$$- 0.33(\pm0.080) \, SHCsatu + 0.20(\pm0.070) \, nelem$$
$$- 0.06(\pm0.015) \, Scarboxylicacid + 0.01(\pm0.004) \, SHBa$$

$N_{tra} = 184, N_{test} = 60, R^2 = 0.70, S_{EP} = 0.47, F = 84.31, Q^2 = 0.68, S_{EE} = 0.42, FIT = 2.02.$

The plot of the predicted pIGC$_{50}$ values based on Eq. (8) versus experimental ones is shown in Fig. 2 (a). Obviously, the predicted pIGC$_{50}$ values are in a good agreement with experimental ones. The 5-parameter model provides high statistical quality: $R^2 = 0.70$ and $Q^2 = 0.68$.
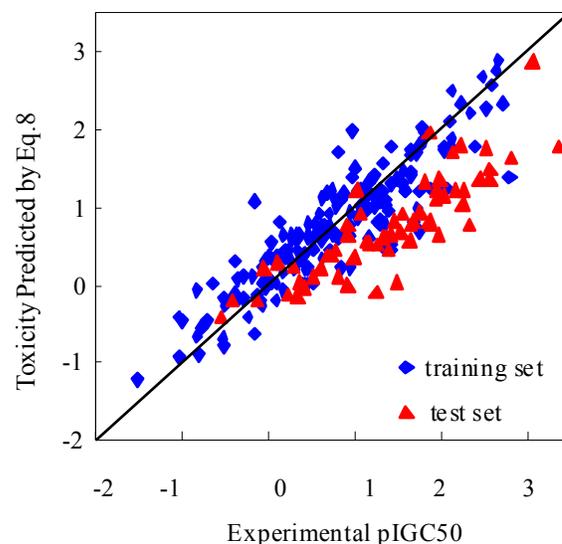

Fig. 2(a). Plot of the experimental pIGC$_{50}$ to *T. pyriformis* against toxicity predicted by Eq. (8)

The optimal model was subjected to the further applicability domain analysis. As can be seen from Williams plot of Eq. 8 shown in Fig. 2 (b), eight chemicals in the training set with h > h* (h* = 0.07), seven of which with the standardized residuals < 2.5S and the rest one with the standardized residuals > 2.5S. The distribution of residuals in Fig. 2 (b) shows that the best molecular descriptors given by Eq. 8 lead to

residuals that tend to follow a normal distribution for most of the phenols. Only one outlier, compound 139 (2, 4, 6-Trinitrophenol), were observed.
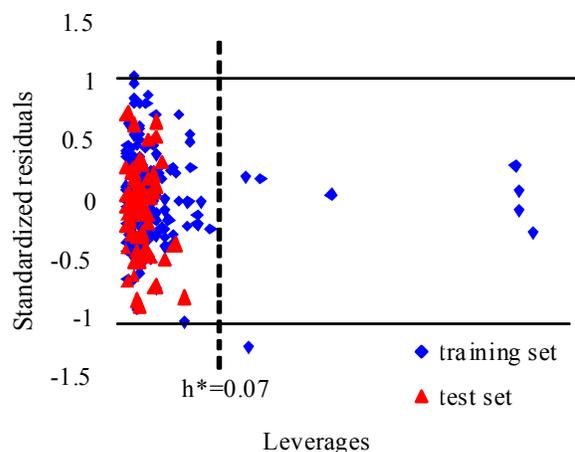

Fig. 2(b). Plot of standardized residuals versus leverages. Lines represent ±2.5 standardized residuals, dotted line represents warning leverage ($h^* = 0.07$)

The original 96 variables were compressed and analyzed by PLS, yielding 10 Latent Factor (PCs). The statistical results of the PLS were shown in Table 5 and the coefficients for these 10 PCs were shown in Table 6. As can be seen from Table 4, four major components described over 70% of the total variance. The first PC (PC1) explains 42.63% of the total variance, and each component is expected to contribute in an equal manner to PC1. The second PC accounted for about 18.70% of the total variance. The third PC explained about 8.40% of the total physical properties of descriptors. The remaining six PCs did not show any significant contributions, cumulative accounting for 5.61% of the variance. PLS utilized these reduced principal component provided a model with an R^2 of 0.78 and Q^2 of 0.64. The plot of predicted vs experimental toxicities was shown in Fig. 3. All the chemicals in this figure followed a regression line well, which indicated that PLS model has a high goodness-of-fit. Three compounds (Table 4), compound 175 (Methoxyhydroquinone), compound 198 (Tetrabromocatechol) and 244 (Methylhydroquinone) were treated as outliers of this model. Among them, compound 175 and 244 were in training set, and compound 198 was in test set. All these outliers in PLS model were also present in MLR model.

As demonstrated in the literature (Pontolillo and Eganhouse, 2001), the octanol-water coefficient for a given compound could be subject to high variability due to the applied experimental procedure or the selected calculation method. Thus the accuracy and quality of a QSAR model are often greatly affected by the specific logP used. For this reason, we first built three models using ALogP, MlogP and ClogP individually (Table 2). Obviously, models developed using logP descriptors alone are not satisfactory to predict the toxicity of phenols due to their low R^2 and Q^2 values. Therefore, more descriptors, 92 Molconn-Z descriptors, were added to build more powerful predictive models. Usually, the more variables chosen

Table 5. Proportion of Variance Explained (Percentage)

Latent Factor	Input variables (X)		Target Variables (Y)	
	Current X (%)	Cumulative X (%)	Current Y (%)	Cumulative Y (%)
1	19.001	19.001	42.626	42.626
2	9.387	28.388	18.696	61.322
3	4.176	32.564	8.395	69.716
4	6.792	39.357	2.524	72.240
5	6.136	45.492	1.756	73.996
6	3.425	48.917	1.142	75.138
7	2.746	51.663	0.912	76.050
8	3.369	55.032	0.577	76.627
9	2.445	57.477	0.661	77.288
10	2.728	60.206	0.566	77.854

Table 6. The coefficients of PLS model

| PCs	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
	0.43	0.19	0.08	0.03	0.02	0.01	0.01	0.01	0.01	0.01
Coefficients	(0.426)	(0.613)	(0.697)	0.722)	(0.740)	(0.751)	(0.761)	(0.766)	(0.773)	(0.779)
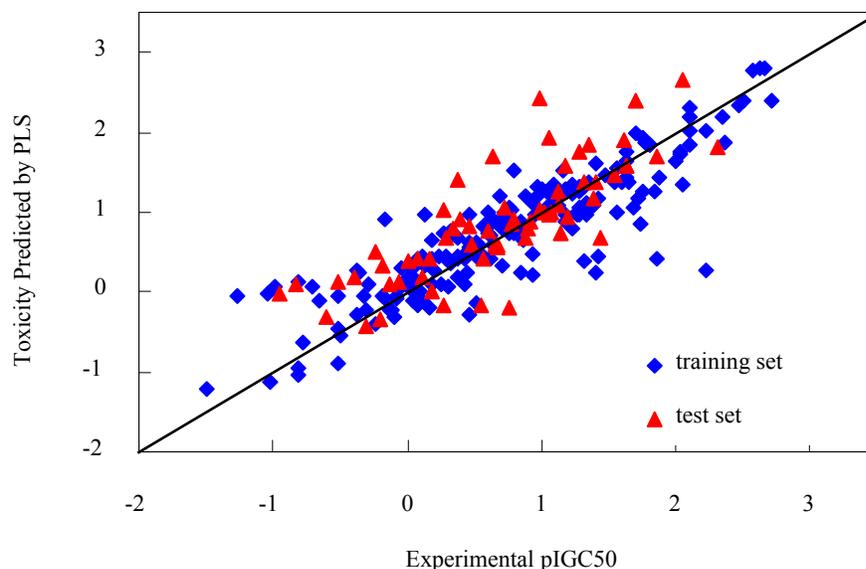
Fig. 3. Plot of the experimental pIGC$_{50}$ to *T. pyriformis* against toxicity predicted by PLS model

by the stepwise regression the higher correlation to some extent. But the inclusion of more descriptors indicates that there are more factors controlling toxicity in this data set, which may be hard for interpretation of the mechanisms of toxic action. Therefore, in the present study, a strategy of the principle of maximal parsimony, the Kubinyi function (FIT) method, was taken into account in the process of developing models to optimize the number of molecular descriptors. And to state whether the model's assumptions are met, the applicability domain of a QSAR model was investigated.

As we know, the molecular structures of the chemicals control their activities and descriptors directly encode particular features of molecular structure. Thus, it is possible to shed light on mechanisms of toxic action of the compounds by interpreting the descriptors in the regression model. And from coefficients of each variable in the model, the importance of each descriptor and the question that which of the independent variables has a greater effect on the dependent variable in the multiple regression analysis can also be interpreted. In Eq. 6, from the absolute size of the standardized regression coefficients, it can be concluded that LogP (MlogP) contribute the most significantly to pIGC$_{50}$ variation. LogP is a descriptor which can well describe the bioavailability of a chemical to organism. The penetration/solubility descriptors (like logP) reflect the ability of a compound to form non-covalent interactions with its environment, to dissolve and persist in water or in a lipidic environment, or to permeate the phase interfaces. Generally, larger logP indicates a stronger ability of the chemical to permeate the cell membrane of an organism and, therefore, to much more easily interact

with its target in the organism. This is especially true for aquatic toxicological assay where the target species is put into a solution with a given concentration of the molecule investigated. MlogP descriptor encodes information about the molecular hydrophobicity. The MlogP takes a positive coefficient, which indicates that increasing values of MlogP correlate with increasing the aquatic toxicity of phenols. The second significant molecular descriptors in the Eq. 6 are the E-state indices which contain gmax, Scarboxylicacid and SHBa descriptors. The E-state values represent the binding affinity for a compound. The negative coefficient of gmax and SHCsatu indicates that increasing values of gmax and SHCsatu correlate with decreasing the aquatic toxicity of phenols, which can be interpreted as that the high negative E-State values can lead to decreased binding affinity of a compound. The low binding affinity is potentially low toxic to aquatic organisms. Interestingly, the SHBa descriptor has the positive coefficient, indicating the larger SHBa the higher toxicity to organism. The group-type E-state descriptors SHBa relate to toxicophores that have hydrogen bond acceptor. It demonstrates that the more hydrogen bond acceptor can improve the ability of attaching to membrane or dissolving in body fluids, thus increases the aquatic toxicity.

From Eq. 8, the most significant descriptor in the equation continues to be logP (AlogP), whose weight takes a positive sign further indicating that the increasing values of AlogP correlate with increasing potential of aquatic toxicity. Besides AlogP, the other three important descriptors are directly related to the atom-type E-state descriptors, i.e., SHCsatu, Scarboxylicacid and SHBa (Table 3). Among, the latter two descriptors were also present in Eq. 6. As

compared to Eq. 6, both Scarboxylicacid and SHBa in Eq. 8 are less important than that in Eq. 6. Interestingly, it also takes a negative weight as it does in Eq. 6, which further implies that the increase of Scarboxylicacid and SHBa value makes a compound less toxic. The descriptor SHCsatu encodes E-State of Csp3 bonded to unsaturated C atoms. Here, molecules with larger SHCsatu values tend to have smaller predicted aquatic toxicity values, as revealed by its negative coefficient in Eq. 8. And the descriptor of nelem, a molecular connectivity descriptor, encodes the number of elements in molecule, which could explain the bimolecular accessibility. In this model, nelem plays a positive influence on the aquatic toxicity.

During the process of model developing, a lot of outliers (Table 4) have been observed. From their structural and physicochemical characteristics, all these outliers can be classed into three types, i.e., 1): quinones compounds, or capable of metabolism or oxidization to quinones; 2): aromatic-nitrogen-containing compounds, or capable of metabolism to aromatic-nitrogen-containing compounds and 3): extremely lipophilic or hydrophilic chemicals. Compound 169 (2, 3-Dimethylhydroquinone), 175 (Methoxyhydroquinone) and 244 (Methylhydroquinone) have the structure of 1- and 4-substituted hydroxyl, which is oxidised into quinones easily. Five nitroso compounds, Compound 139 (2, 4, 6-Trinitrophenol), 142 (2, 6-Dichloro-4-nitrophenol), 192 (4-Nitrophenol), 195 (3-Methyl-4-nitrophenol) and 238 (2, 6-Diiodo-4-nitrophenol) are aromatic-nitrogen-containing compounds. Compound 168 (4-Amino-2-cresol) and 170 (4-Amino-2, 3-dimethylphenol) are amino compounds, which may be metabolized to a nitroso group. Compound 4 (3-Hydroxy-4-methoxybenzyl alcohol), 198 (Tetrabromocatechol) and 201 (1, 3, 5-Trihydroxybenzene) have the structure of multiple-substituted hydroxyl groups which maybe contribute to their high hydrophilic.

To simplify the analysis and to attempt to allow some interpretation of the results, PCA was applied to the descriptor set. PCA identified 10 PCs as representative of the complete set of 96 descriptors. PLS analysis based on these components provided a model with Q^2 of 0.64. Though the high Q^2 of PLS model, three outliers were observed in this model such as, compound 175 (Methoxyhydroquinone), 198 (Tetrabromocatechol) and 244 (Methylhydroquinone). The prediction toxicity of compound 175 and 244 were lower than the experiments, and the prediction of compound 198 was higher than the experiments. All of those outliers were observed in MLR models. For assessment of the predictivity of the models, an external validation (test) set was used.

In the present study, three optimal models were obtained and they were compared with those calculated

in previous work. Cronin et al. (2002) using stepwise-MLR method obtained a seven-descriptor model (including logD, LUMO, PNEG, ABSQon, SsOH, MW, MaxHp) with $R^2 = 0.65$, $Q^2 = 0.63$ based on 108 descriptors, while removal of these outliers and re-analysis reveals a significant seven-variables equation with $R^2 = 0.83$ and $Q^2 = 0.82$. Enoch et al. (2008) employed 168 descriptors and stepwise-MLR method constructed a 4-variables model (including logP, AHard, NH_{Don}, SdssC) with R^2 of 0.66 and Q^2 of 0.64. In this work, two six-variables MLR models (Eq. 5 and Eq. 7) were built based on 95 descriptors. From statistics, Eq. 5 has the similar predictive power to Cronin model but poor to Enoch model. While Eq. 7 shows better predictive than both Cronin and Enoch model. Remove of outliers and reanalysis, two optimal models (Eq. 6 and Eq. 8) were obtained. Eq. 6 exhibits high predictive power ($R^2 = 0.71$, $Q^2 = 0.69$) using only four optimal variables and Eq. 8 contains only five descriptors with R^2 of 0.70 and Q^2 of 0.68, indicating that both Eq. 6 and Eq. 8 have stronger predictive than Enoch's model. Though the re-analysis Cronin's model shows high Q^2, it is harder to interpret the mechanism of toxic action because of its more variables than both Eq. 6 and Eq. 8. Cronin et al. (2002) used PLS approach with 200 compounds provided an 11-descriptors model with a R^2 of 0.60, remove of the outliers and re-analysis provided a model with a R^2 of 0.82. While the current PLS model was developed based on 10 PCs with 187 compounds, providing a R^2 of 0.78 without excluding any outliers. It is clear that our PLS model shows more statistical significant than the one obtained by Cronin. From the high predictive power of the obtained models, logP and Molconn-Z descriptors have been demonstrated significant variables for the prediction of the toxicity of phonels.

CONCLUSION

In this paper, a large data set of toxicity values for phenols to the ciliated protozaon *T. pyriformis* has been collected from literature and MLR and PLS methods were performed on the data set. The employment of logP descriptors and Molconn-Z descriptors made possible to achieve better statistical parameters that compare fairly well with others published previously based on stepwise-MLR and PLS method. By stepwise-MLR analysis, two robust models were obtained and the mechanism of toxic action was interpreted according to these descriptors involved in models. The results showed that Eq.6 and Eq.8 were highly predictive models for aquatic toxicity of phenols because of high in R^2 and Q^2 and concise in variables. Even, the models developed for training set are good and potential for assessment and regulation purpose because of the negligible difference between Q^2 and

R^2 value. The results also indicate that developed models in this study show better statistical significance than those corresponding models reported in the literature. The strong predictivity of final models shows that LogP and Molconn-Z descriptors are significant contribution to the prediction of toxicity and interpretation of mechanism of toxic action. The resulut of PLS analysis indicated that the PLS model was of potential to predict the toxicity of phenols.

REFERENCES

Bukowska, B. and Kowalska, S. (2004). Phenol and catechol induce prehemolytic and hemolytic changes in human erythrocytes. Toxicol. Lett., **152**, 73–84.

Cronin, M. T. D. and Dearden, J. C. (1995). Review, QSAR in Toxicology. 1. Prediction of Aquatic Toxicity. Quant. Struct. Act. Relat., **14**, 1–7.

Cronin, M. T. D. and Schultz, T. W. (1996). Structure–toxicity relationships for phenols to Tetrahymena pyriformis. Chemosphere, **32**, 1453–1468.

Cronin, M. T. D., Aptula, A. O., Duffy, J. C., Netzeva, T. I., Rowe, P. H., Valkova, I. V. and Schultz, T. W. (2002). Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to Tetrahymena pyriformis. Chemosphere, **49**, 1201–1221.

Cunningham, W. P., Cunningham, M. A. and Saigo, B. (2005). A Global Concern, Mc Graw-Hill Education. New York: Environmental Science.

Duchowicz, P. R., Mercader, A. G., Fernandez, F. M. and Castro, E. A. (2008). Prediction of aqueous toxicity for heterogeneous phenol derivatives by QSAR. Chemometr. Intell. Lab. Syst., **90**, 97–107.

Eriksson, L., Jaworska, J., Worth, A. P., Cronin, M. T. D., McDowell, R. M. and Gramatica, P. (2003). Methods for reliability and uncertainty assessment and for applicability evaluations of classification and regression-based QSARs. Environ. Health Persp., **111**, 1361–1375.

Enoch, S. J., Cronin, M. T. D., Schultz, T. W. and Madden, J. C. (2008). An evaluation of global QSAR models for the prediction of the toxicity of phenols to Tetrahymena pyriformis. Chemosphere, **71**, 1225–1232.

Garg, R. Kurup, A. and Hansch, C. (2001). Comparative QSAR: on the toxicology of the phenolic OH moiety. Crit. Rev. Toxicol., **31**, 223–245.

Gramatica, P. (2007). Principles of QSAR models validation: internal and external. QSAR Comb. Sci., **26**, 694–701.

Hill, D. L. (1972). The Biochemistry and Physiology of Tetrahymena, first ed (pp. 230). New York, Academic Press.

Hansch, C. and Leo, A. (1979). Substituent Constants for Correlation Analysis in Chemistry and Biology. New York: Wiley.

Hand, D. J. (1981). Discrimination and Classification. New York: John Wiley & Sons.

Hansch, C. and Leo, A. (1995). American Chemical Society. Washington, D. C.

Jenrich, R. I. (1960). Stepwise discriminant analysis (pp. 76–95). In K. Enslein, A. Ralston, and H. S. Wilf (ed.), Statistical methods for digital computers. New York: John Wiley & Sons.

Kubinyi, H. (1994). Variable Selection in QSAR Studies. II. A Highly Efficient Combination of Systematic Search and Evolution. Quant. Struct.-Act. Relatsh., **13**, 393–401.

Kušić, H. (2009). Prediction of rate constants for radical degradation of aromatic pollutants in water matrix: A QSAR study. Chemosphere, **75**, 1128–1134.

Liu, H., Tan, J., Yu, H. X., Liu, H. X., Wang, L. S. and Wang, Z. Y. (2010). Determination of the Apparent Reaction Rate Constants for Ozone Degradation of Substituted Phenols and QSPR/QSAR Analysis. Int. J. Environ. Res., **4 (3)**, 507-512.

Nguyen, D. V. and Rocke, D. M. (2002). Partial least squares proportional hazard regression for application to DNA microarray survival data. Bioinformatics, **18**, 1625-1632.

Netzeva, T. I., Worth, A. P., Aldenberg, T., Benigni, R., Cronin, M. T. D., Gramatica, P., Jaworska, J. S., Kahn, S., Klopman, G., Marchant, C. A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G. Y., Perkins, R., Roberts, D. W., Schultz, T. W., Stanton, D. T., van de Sandt, J. J. M., Tong, W., Veith, G. and Yang, C. (2005). Current status of methods for defining the applicability domain of (Quantitative) Structure–Activity Relationships: The report and recommendations of ECVAM Workshop 52. Alternatives to Laboratory Animals, **33**, 155–173.

Pontolillo, J. and Eganhouse, R. P. (2001). The Search for Reliable Aqueous Solubility (Sw) and Octanol-Water Partition Coefficient (Kow) Data for Hydrophobic Organic Compounds: DDT and DDE as a Case Study. U.S. Geological Survey. Water-Resources Investigations Report 01-4201, USGS: Reston, VA.

Pan, Y., Jiang, J., Wang, R., Cao, H. Y. and Cui, Y. (2009). A novel QSPR model for prediction of lower flammability limits of organic compounds based on support vector machine. J. Hazard. Mater., **168**, 962–969.

Verma, R. P. and Hansch, C. (2005). An approach toward the problem of outliers in QSAR. Bioorg. Med. Chem., **13**, 4597–4621.

Wold, S., Ruhe, H., Wold, H. and Dunn, W. J. (1984). The collinearity problem in linear regression. The PLS approach to generalized inverse. SIAM J. Sci. Statist. Comput., **5**, 735–743.

Xia, B., Liu, K., Gong, Z., Zheng, B., Zhang, X. and Fan, B. (2009). Rapid toxicity prediction of organic chemicals to Chlorella vulgaris using quantitative structure–activity relationships methods. Ecotox. Environ. Safe., **72**, 787–794.