Journal of materials and Environmental Sciences
ISSN : 2028-2508

JMES, 2017 Volume 8, Issue 2, Page 476-484

http://www.jmaterenvironsci.com /

QSTR analysis and combining DFT of the toxicity of heterogeneous phenols

A. Ousaa[1*], B. Elidrissi[1], M. Ghamali[1], S. Chtita[1], A. Aouidate, M. Bouachrine[2], T. Lakhlifi[1]

[1]*Molecular Chemistry and Natural Substances Laboratory, Faculty of Science, University Moulay Ismail, Meknes, Morocco*
[2]*MEM, ESTM, University Moulay Ismail, Meknes, Morocco*

Received 21 Jul 2016,
Revised 07 Nov 2016,
Accepted 13 Nov 2016

Keywords
- ✓ QSTR model,
- ✓ DFT study,
- ✓ heterogeneous phenols,
- ✓ *Tetrahymena pyriformis*

abdellahousaa@gmail.com
Tel. +21271310939

Abstract

Quantitative structure–toxicity relationship (QSTR) models are useful to understand how chemical structure relates to the toxicity of natural and synthetic chemicals. The chemical structures of 70 heterogeneous phenols have been characterized by electronic and physic-chemical descriptors. Density functional theory (DFT) with Beck's three parameter hybrid functional using the LYP correlation functional (B3LYP/6-31G(d)) calculations have been carried out in order to get insights into the structure chemical and property information for the study compounds. The present study was performed using principal component analysis (PCA) method, multiple linear regression method (MLR), multiple non-linear regression (MNLR) and artificial neural network (ANN). The quantitative model of the toxicity of these compounds was accordingly proposed and interpreted based on the multivariate statistical analysis. The statistical quality of the MLR and MNLR models was found to be efficient for the predicting of the toxicity, but when compared to the obtained results by ANN model, we realized that the predictions achieved by this latter one were more effective. This model provided statistically significant results and showed good internal stability and powerful predictability. The squared correlation coefficients were 0.801, 0.802 and 0.824 for MLR, MNLR and ANN models respectively. The obtained results suggested that the proposed descriptors could be useful to predict the toxicity of heterogeneous phenols to *Tetrahymena pyriformis.*

1. Introduction

Heterogeneous phenols have been used in chemical industry for many years ago. They are used as solvents, propellants, additives, cooling agents, insecticide, herbicides and organic syntheses [1-2]. Many of these chemicals were released into the environment and accumulated in nearly all natural environments, especially in aquatic systems, so it is beneficial to study seriously their potential hazard to aquatic organism.

Experiment is a direct way to obtain the toxicity data of organic compounds, but it has many deficiencies, such as requirement of enormous number of trial organisms, expensive cost, long time, the difference in measured value between different researchers. Consequently, it would be very difficult to obtain the toxicity data of all organic compounds by experiment, as new compounds are springing up, other difficulties will follow. So it is necessary to use the theoretical research to make up for disadvantages of the experiment and to predict the toxicity data of compounds quickly and exactly.

With the rapid development of computational science and theoretical chemistry, it can quickly and precisely obtain the quantum chemical parameters of organic compounds. Quantitative structure-activity relationship (QSAR) can predict the bioactivity such as toxicity, mutagenicity and carcinogenicity based on structural parameters of compounds and appropriate mathematical models.

At present, there are a large number of molecular descriptors that can be used in QSAR studies [3-4]. Once validated, the findings can be used to predict activities of untested compounds.

The aim of this study is to develop predictive QSTR models for the acute toxic effects of phenol compounds toward *Tetrahymena pyriformis* using several statistical tools, principal components analysis (PCA), multiple linear regression (MLR), multiple non-linear regression (MNLR) and artificial neural network (ANN) methods.

2. Material and Methods

2.1. Data sources

Acute toxicity data of 70 heterogeneous phenols to *Tetrahymena pyriformis* were taken from a literature [5]. **IC_{50}** here means the millimolar concentration causing 50% inhibition of growth about heterogeneous phenols to *Tetrahymena pyriformis*. The bigger the value of **–log IC_{50} (pIC_{50})**, the higher is toxicity of compounds, and vice versa. For the proper validation of our data set with a QSTR model, the 70 substituted phenols were divided into training and test sets. A total of 60 molecules were placed in the training set to build the QSTR models, whereas the remaining 10 molecules composed the test set. The division was carried out by random selection.

The following table shows the studied compounds and the corresponding experimental toxicties pIC_{50} (table 1). The range of the toxicity data varies between -1.50 and 2.63 (µM).

Table 1: heterogeneous phenol derivatives and their observed toxicities against *Tetrahymena pyriformis*

N°	Name (IUPAC)	pIC_{50}	N°	Name (IUPAC)	pIC_{50}
1*	4-Hydroxyphenylacetic acid	− 1.50	**36**	Salicylaldoxime	0.25
2	3-Hydroxybenzyl alcohol	− 1.04	**37***	2-Hydroxy-5-methylacetophenone	0.31
3	4-Carboxyphenol	− 1.02	**38**	3-Methoxysalicylaldehyde	0.38
4	3-Hydroxy-4-methoxybenzyl alcohol	− 0.99	**39***	Salicylhydroxamic acid	0.38
5	4-Hydroxy-3-methoxybenzyl amine	− 0.97	**40**	4-Allyl-2-methoxyphenol	0.42
6	4-Hydroxyphenethyl alcohol	− 0.83	**41**	2-Hydroxybenzaldehyde	0.42
7	3-Carboxyphenol	− 0.81	**42**	Ethyl-3-hydroxybenzoate	0.48
8	4-Hydroxybenzamide	− 0.78	**43**	4-Cyanophenol	0.52
9	4-Hydroxy-3-methoxybenzylalcohol−	0.70	**44**	4-Propyloxyphenol	0.52
10	2,6-Dimethoxyphenol	− 0.60	**45**	Ethyl-4-hydroxybenzoate	0.57
11	2,4,6-Tris(dimethylaminomethyl) phenol	− 0.52	**46**	5-Methyl-2-nitrophenol	0.59
12	Salicylic acid	− 0.51	**47**	Methyl-4-methoxysalicylate	0.62
13	2-Methoxyphenol	− 0.51	**48**	4-Butoxyphenol	0.70
14	5-Methylresorcinol	− 0.39	**49***	2-Methoxy-4-propenyIphenol	0.75
15	3-Hydroxyacetophenone	− 0.38	**50**	2,2'-Biphenol	0.88
16	2-Ethoxyphenol	− 0.36	**51**	2,2',4,4'-Tetrahydroxybenzophenone	0.96
17	4-Acetylphenol	− 0.30	**52**	4-sec-Butylphenol	0.98
18	3-Ethoxy-4-methoxyphenol	− 0.30	**53**	3-Hydroxydiphenylamine	1.01
19*	2-Hydroxybenzamide	− 0.24	**54**	4-Hydroxybenzophenone	1.02
20	4-Hydroxy-3-methoxyphenethylalcohol	− 0.18	**55**	Benzyl-4-hydroxyphenyl ketone	1.07
21*	3-Acetamidophenol	− 0.16	**56**	2-Phenylphenol	1.09
22	3-Hydroxy-4-methoxybenzaldehyde	− 0.14	**57**	2-Hydroxybenzophenone	1.23
23	4-Hydroxy-3-methoxyacetophenone	− 0.12	**58**	2-Hydroxydiphenylmethane	1.31
24	3,5-Dimethoxyphenol	− 0.09	**59***	Butyl-4-hydroxybenzoate	1.33
25	2-Hydroxyethylsalicylate	− 0.08	**60**	n-Pentyloxyphenol	1.36
26	3-Methoxy-4-hydroxybenzaldehyde	− 0.03	**61**	2-Hydroxy-4-methoxybenzophenone	1.42
27	4-Hydroxy-3-methoxybenzonitrile	− 0.03	**62**	Isoamyl-4-hydroxybenzoate	1.48
28	3-Ethoxy-4-hydroxybenzaldehyde	0.01	**63***	4-Heptyloxyphenol	2.03
29	2-Cyanophenol	0.03	**64**	Nonyl-4-hydroxybenzoate	2.63
30	2-Hydroxyacetophenone	0.08	**65**	2,4,6-Trinitrophenol	− 0.16
31	Methyl-4-hydroxybenzoate	0.08	**66**	3,4-Dinitrophenol	0.27
32*	4'-Hydroxypropiophenone	0.12	**67**	2,6-Dinitrophenol	0.54
33	Syringaldehyde	0.17	**68***	2,5-Dinitrophenol	0.95
34	Salicylhydrazide	0.18	**69**	2,4-Dinitrophenol	1.08
35	4-Hydroxy-2-methylacetophenone	0.19	**70**	2,6-Dinitro-4-cresol	1.23

* Test set

2.2. Molecular descriptors

The computation of electronic descriptors was performed using the Gaussian 03W program [6]. The geometries of all 70 theoretically heterogeneous phenols were optimized with DFT method at the B3LYP functional and 6-31G (d) base set. Then some related structural descriptors from the results of quantum computation were chosen: the highest occupied molecular orbital energy E_{HOMO} (eV), the lowest unoccupied molecular orbital energy E_{LUMO} (eV), the energy gap ΔE (eV), the dipole moment µ (Debye), the total energy E_T (eV).

ChemSketch program [7] was employed to calculate the others molecular descriptors such as: the molar volume MV (cm^3), the molecular weight MW (g/mol) , the molar refractivity MR (cm^3), the parachor Pc (cm^3), the density D (g/cm^3), the refractive Index n, the surface tension γ (Dyne/cm) and the polarizability α (cm^3). To

improve the estimate quality of toxicity of these compounds, molecular descriptor which reflect other specific interactions should be also included as octanol/water partition coefficient (log *P*).

2.3. Statistical analysis

The structures of 70 heterogeneous phenols toward *Tetrahymena pyriformis* were studied by statistical methods based on the principal component analysis (PCA) [8] using the software XLSTAT version 2013 [9]. PCA is a statistical method useful to summarize all the information encoded in the structures of the compounds. It is also very helpful for understanding the distribution and classification of the data set [10]. This is an important descriptive statistical method which aims to present, in graphic forms, the maximum of information contained in the data table 1 and table 2.

The multiple linear regression (MLR) analysis with backward selection was employed to model the structure toxicity relationships. It is a mathematic technique that minimizes differences between actual and predicted values. It has served also to select the descriptors that will be exploited as input parameters in the multiples nonlinear regression (MNLR) and artificial neural network (ANN).

The MLR and MNLR were performed using the software XLSTAT version 2013 [9], to predict toxic effects pIC_{50}. Equations were justified by the determination coefficient (R^2), mean squared error (MSE), Fisher's criterion (F) and significance level (P) [9]. The ANN is an artificial system that is simulating the function of the human brain. Three components form a neural network: the processing elements or nodes, the topology of the connections between the nodes, and the learning rule by which new information is encoded in the network. While there are a many different ANN types, the most commonly used in QSAR is the three-layered feed forward network [11]. In this type of network, the neurons are arranged in layers (an input layer, one hidden layer and an output layer). Each neuron in any layer is fully connected with the other neurons of a next layer and no connections are between neurons belonging to the same layer.

According to the supervised learning adopted, the networks are taught by giving them examples of input patterns and the corresponding target outputs. Through an iterative procedure, the connection weights are modified until the network gives the desired results for the training set of data. A back propagation algorithm is used to minimize the error function. This algorithm has been described previously with a simple example of application [12] and a detail of this algorithm is given elsewhere [13].

The ANN analysis was performed using Matlab software version 2009a Neural Fitting tool (nftool) toolbox [14-15].

Cheking the stability, predictibility and generalization ability of the proposed models are very important steps in a QSTR study. For the validation of the prediction ability of a QSTR model, two principal methods, internal and external validations are available. Cross-validation is one of the most common methods that are carried out for internal validation. In this study, the internal predictive ability of every model was evaluated using leave-one-out cross-validation (R^2cv). A good R^2cv often indicates good robustness and high internal predictive capacity of a QSTR model. However, recent studies [16] indicate that there is no evident correlation between the value of R^2cv and the actual predictive capacity of a QSTR model, suggesting that the R^2cv remains inadequate as a reliable estimate of the model predictive ability for all new chemicals. To determine both the generalizability of QSTR models for new chemicals and the true predictive ability of the models, statistical external validation is used during the model development step by properly using a prediction set for validation.

3. Results

3.1. QSTR models and analysis

QSTR analysis was performed using the pIC_{50} of 70 heterogeneous phenols to *Tetrahymena pyriformis* as reported in [5], the values of the 14 chemical descriptors are shown in table 2.

The principle objective is to perform in the first time, a principal component analysis (PCA), which allows us to eliminate descriptors that are highly correlated (dependent), then an MLR analysis was performed on the remaining descriptors using the backward method until a valid model.

1.1. Principal component analysis

The set of descriptors coding the 70 heterogeneous phenols, electronic and physico-chemical descriptors are submitted to PCA analysis [17]. The first three principal axes are sufficient to encode the information provided by the data matrix. Indeed, the percentages of variance are 44.66%, 29.29% and 11.96% for the axes F1, F2 and F3, respectively. The total information is estimated to a percentage of 85.92%.

Table 2: The values of the fourteen chemical descriptors

N	pIC_{50}	MW	MR	MV	Pc	n	□	D	□	E_T	E_{HOMO}	E_{LUMO}	□ E	□	log *P*
1*	-1.50	152.14	39.24	115.30	320.60	1.60	59.80	1.32	15.55	-14577.48	-5.99	-0.14	5.84	2.79	1.31
2	-1.04	124.14	34.58	101.60	275.80	1.60	54.10	1.22	13.71	-11490.68	-6.01	-0.14	5.87	1.72	0.90
3	-1.02	138.12	35.06	100.30	284.40	1.62	64.40	1.38	13.90	-13507.18	-6.42	-1.04	5.38	1.85	1.33
4	-0.99	154.16	41.26	125.60	332.40	1.57	48.90	1.23	16.35	-14608.89	-5.60	-0.24	5.36	1.90	0.74
5	-0.97	153.18	43.26	131.80	344.80	1.57	46.70	1.16	17.15	-14068.26	-5.48	0.21	5.69	2.30	0.64
6	-0.83	138.16	39.21	118.10	315.60	1.58	50.80	1.17	15.54	-12561.18	-5.68	0.19	5.87	2.46	1.19
7	-0.81	138.12	35.06	100.30	284.40	1.62	64.40	1.38	13.90	-13507.13	-6.34	-1.33	5.01	0.64	1.33
8	-0.78	137.14	37.06	106.50	296.70	1.61	60.10	1.29	14.69	-12966.07	-6.26	-0.74	5.52	2.79	0.52
9	-0.70	154.16	41.26	125.60	332.40	1.57	48.90	1.23	16.35	-14609.19	-5.78	0.14	5.92	2.66	0.74
10	-0.60	154.16	41.49	135.80	335.60	1.52	37.20	1.13	16.44	-14609.04	-5.26	0.57	5.83	2.22	1.35
11	-0.52	265.39	81.84	257.50	648.40	1.55	40.10	1.03	32.44	-22519.75	-3.06	-2.28	0.78	14.91	1.49
12	-0.51	138.12	35.06	100.30	284.40	1.62	64.40	1.38	13.90	-13507.16	-6.43	-1.60	4.84	4.27	1.98
13	-0.51	124.13	34.81	111.80	278.90	1.53	38.60	1.11	13.80	-11490.76	-5.53	0.32	5.85	2.73	1.51
14	-0.39	124.14	34.84	102.50	274.90	1.59	51.60	1.21	13.81	-11491.06	-5.71	0.26	5.97	1.37	1.88
15	-0.38	136.14	38.16	119.30	307.40	1.55	43.90	1.14	15.12	-12528.88	-6.32	-1.51	4.81	1.74	1.23
16	-0.36	138.16	39.44	128.30	318.70	1.53	38.00	1.08	15.63	-12561.37	-5.49	0.30	5.79	2.87	1.87
17	-0.30	136.15	38.16	119.30	307.40	1.55	43.90	1.14	15.12	-12528.92	-6.35	-1.23	5.12	2.61	1.23
18	-0.30	168.19	46.12	152.30	375.40	1.52	36.80	1.10	18.28	-15679.57	-5.32	0.14	5.46	3.07	1.71
19*	-0.24	137.14	37.06	106.50	296.70	1.61	60.10	1.29	14.69	-12966.42	-6.06	-1.19	4.87	3.99	1.17
20	-0.18	168.19	45.89	142.10	372.20	1.56	46.90	1.18	18.19	-15679.62	-5.39	0.49	5.88	1.45	1.03
21*	-0.16	151.16	42.40	120.90	326.00	1.62	52.80	1.25	16.81	-14036.43	-6.02	-0.40	5.62	4.47	0.91
22	-0.14	152.14	41.56	123.50	324.00	1.59	47.30	1.23	16.47	-14576.64	-6.05	-1.40	4.65	6.43	1.22
23	-0.12	166.17	44.84	143.30	364.10	1.54	41.50	1.16	17.77	-15647.31	-6.08	-1.01	5.07	4.01	1.07
24	-0.09	154.16	41.49	135.80	335.60	1.52	37.20	1.13	16.44	-14609.17	-5.56	0.62	6.17	2.09	1.35
25	-0.08	182.17	46.07	139.70	383.60	1.57	56.70	1.30	18.26	-17696.24	-6.27	-1.47	4.79	2.05	1.63
26	-0.03	152.15	41.56	123.50	324.00	1.59	47.30	1.23	16.47	-14576.66	-6.05	-1.39	4.65	5.05	1.22
27	-0.03	149.15	39.21	119.70	326.80	1.57	55.40	1.24	15.54	-14002.61	-6.19	-1.02	5.17	5.02	1.37
28	0.01	166.17	46.19	140.00	363.70	1.57	45.50	1.19	18.31	-15647.33	-6.02	-1.38	4.64	2.82	1.58
29	0.03	119.12	32.84	97.20	268.20	1.59	57.80	1.22	13.02	-10884.03	-6.64	-1.30	5.34	5.75	1.53
30	0.08	136.15	38.16	119.30	307.40	1.55	43.90	1.14	15.12	-12529.25	-6.41	-0.95	5.46	3.17	1.88
31	0.08	152.15	39.90	125.70	327.10	1.55	45.70	1.21	15.82	-14576.96	-6.41	-0.95	5.46	4.35	1.67
32*	0.12	150.17	42.79	135.90	347.20	1.54	42.60	1.10	16.96	-13599.44	-6.33	-1.21	5.12	2.46	1.93
33	0.17	182.17	48.24	147.50	380.60	1.57	44.30	1.23	19.12	-17694.98	-5.77	-1.36	4.42	3.64	1.07
34	0.18	152.15	40.65	115.30	324.60	1.62	62.70	1.32	16.11	-14472.69	-6.12	-1.26	4.86	2.56	0.87
35	0.19	150.17	42.98	135.60	345.10	1.55	41.80	1.11	17.04	-13599.24	-6.52	-2.61	3.91	4.68	1.74
36	0.25	137.14	37.01	115.80	300.90	1.55	45.50	1.18	14.67	-12963.71	-6.37	-1.47	4.91	3.32	1.39
37*	0.31	150.17	42.98	135.60	345.10	1.55	41.80	1.11	17.04	-13599.87	-5.99	-1.65	4.35	3.48	2.39
38	0.38	152.15	41.56	123.50	324.00	1.59	47.30	1.23	16.47	-14576.58	-6.02	-1.43	4.59	6.16	1.87
39*	0.38	153.14	38.65	109.20	312.30	1.63	66.90	1.40	15.32	-15012.64	-6.08	-1.16	4.92	5.24	1.17
40	0.42	164.20	48.72	156.20	384.30	1.54	36.50	1.05	19.31	-14668.48	-5.69	0.06	5.75	2.00	2.61
41	0.42	122.12	34.88	99.50	267.30	1.62	52.00	1.23	13.83	-11458.08	-6.50	-1.58	4.92	4.77	2.03
42	0.48	166.17	44.54	142.20	366.90	1.54	44.20	1.17	17.65	-15648.09	-6.22	-1.18	5.04	2.56	2.03
43	0.52	119.12	32.84	97.20	268.20	1.59	57.80	1.22	13.02	-10884.10	-6.60	-1.09	5.51	5.19	1.53
44	0.52	152.19	44.07	144.80	358.50	1.52	37.50	1.05	17.47	-13631.83	-5.33	0.03	5.37	2.39	2.39
45	0.57	166.17	44.54	142.20	366.90	1.54	44.20	1.17	17.65	-15647.63	-6.38	-0.94	5.44	4.18	2.03
46	0.59	153.14	39.50	115.90	315.40	1.60	54.70	1.32	15.66	-15008.17	-5.60	-2.60	3.00	3.20	2.12
47	0.62	182.17	46.58	149.70	383.80	1.53	43.10	1.22	18.46	-17696.16	-6.00	-0.93	5.07	2.25	2.17
48	0.70	166.22	48.71	161.30	398.30	1.52	37.10	1.03	19.31	-14702.27	-5.30	0.00	5.30	0.61	2.84
49*	0.75	166.22	48.99	161.10	395.30	1.52	36.10	1.03	19.42	-14702.40	-5.39	0.33	5.72	2.76	2.91
50	0.88	186.21	54.60	151.50	410.60	1.64	53.80	1.23	21.64	-16712.14	-5.98	-0.76	5.21	1.84	3.01
51	0.96	246.21	63.57	161.20	486.90	1.72	83.10	1.53	25.20	-23894.67	-5.89	-1.46	4.43	3.17	3.52
52	0.98	150.22	46.95	154.40	377.10	1.52	35.50	0.97	18.61	-12651.02	-4.68	-2.14	2.55	4.92	3.36
53	1.01	185.22	57.50	153.90	415.60	1.67	53.10	1.20	22.79	-16171.04	-5.08	-0.11	4.97	0.87	3.11
54	1.02	198.22	57.92	165.90	441.90	1.62	50.20	1.19	22.96	-17749.89	-6.27	-1.54	4.73	2.34	3.13
55	1.07	212.24	62.64	180.00	479.60	1.61	50.20	1.18	24.83	-18820.33	-6.34	-1.28	5.06	2.68	3.06
56	1.09	170.21	52.72	153.10	395.60	1.60	44.50	1.11	20.90	-14663.98	-5.81	-0.69	5.12	1.74	3.32
57	1.23	198.22	57.92	165.90	441.90	1.62	50.20	1.19	22.96	-17750.19	-6.14	-1.93	4.21	3.48	3.78
58	1.31	184.23	57.44	167.20	433.30	1.60	45.00	1.10	22.77	-15734.31	-5.82	-0.15	5.66	1.81	3.76
59*	1.33	194.23	53.80	175.20	446.40	1.53	42.10	1.11	21.33	-17789.18	-6.30	-0.89	5.40	1.36	3.00
60	1.36	166.21	48.71	161.30	398.30	1.52	37.10	1.03	19.31	-14702.30	-5.35	0.03	5.39	2.36	3.28
61	1.42	228.24	64.60	189.90	498.50	1.60	47.40	1.20	25.61	-20868.75	-6.00	-1.69	4.31	2.70	3.62
62	1.48	208.25	58.39	192.10	483.60	1.52	40.10	1.08	18.91	-18874.09	-6.33	-1.57	4.75	1.40	3.28
63*	2.03	208.30	62.61	210.90	517.60	1.51	36.20	0.99	24.82	-17913.77	-5.81	-1.10	4.71	2.08	4.17
64	2.63	264.36	76.97	257.80	645.40	1.51	39.20	1.03	30.51	-23141.55	-6.28	-0.87	5.41	1.15	5.22
65	-0.16	229.10	47.77	123.30	388.70	1.70	98.50	1.86	18.93	-25077.60	-8.24	-3.90	4.34	1.77	1.49
66	0.27	184.10	41.22	111.50	333.20	1.66	79.60	1.65	16.34	-19508.85	-7.43	-2.95	4.48	6.61	1.55
67	0.54	184.10	41.22	111.50	333.20	1.66	79.60	1.65	16.34	-19509.58	-7.63	-3.32	4.31	3.41	1.55
68*	0.95	184.10	41.22	111.50	333.20	1.66	79.60	1.65	16.34	-19509.49	-7.49	-3.64	3.85	1.17	1.55
69	1.08	184.10	41.22	111.50	333.20	1.66	79.60	1.65	16.34	-19509.58	-7.63	-3.32	4.31	3.41	1.55
70	1.23	198.13	46.05	127.80	370.80	1.64	70.80	1.55	18.25	-20579.90	-7.27	-3.27	4.00	4.57	2.06

*Test set

The principal component analysis (PCA) [18] was conducted to identify the link between the different descriptors. Bold values are different from 0 at a significance level of p= 0.05. Correlations between the fourteen descriptors are shown in table 3 as a correlation matrix. The Pearson correlation coefficients are listed in table 3. The obtained matrix provides information on the positive or negative correlation between descriptors. In general, the co-linearity (r>0.5) was observed between most of the variables, and between the variables and pIC_{50}. Additionally, to decrease the redundancy presented in our data matrix, the descriptors that are highly correlated ($R \geq 0.95$), were removed.

Table 3: Correlation matrix between different obtained descriptors

	pIC50	MW	MR	MV	Pc	N	□	D	□	E_T	E_{HOMO}	E_{LUMO}	□ E	□	log *P*
pIC_{50}	**1**														
MW	**0.588**	**1**													
MR	**0.617**	**0.907**	**1**												
MV	**0.592**	**0.804**	**0.951**	**1**											
Pc	**0.618**	**0.913**	**0.988**	**0.971**	**1**										
N	-0.026	0.161	-0.068	**-0.367**	-0.152	**1**									
□	-0.090	0.145	-0.199	**-0.438**	-0.218	**0.891**	**1**								
D	-0.115	0.136	**-0.252**	**-0.470**	**-0.264**	**0.831**	**0.963**	**1**							
□	**0.600**	**0.897**	**0.991**	**0.936**	**0.975**	-0.049	-0.189	**-0.242**	**1**						
E_T	**-0.495**	**-0.948**	**-0.727**	**-0.588**	**-0.740**	**-0.336**	**-0.392**	**-0.421**	**-0.719**	**1**					
E_{HOMO}	-0.133	0.022	**0.333**	**0.444**	**0.313**	**-0.506**	**-0.679**	**-0.723**	**0.343**	0.214	**1**				
E_{LUMO}	**-0.281**	**-0.342**	-0.082	0.058	-0.088	**-0.508**	**-0.618**	**-0.644**	-0.076	**0.503**	**0.599**	**1**			
□ E	**-0.235**	**-0.443**	**-0.385**	**-0.305**	**-0.375**	-0.201	-0.190	-0.186	**-0.386**	**0.443**	-0.106	**0.733**	**1**		
□	-0.131	0.098	0.112	0.094	0.107	0.044	0.035	0.027	0.128	-0.079	0.232	**-0.327**	**-0.603**	**1**	
log *P*	**0.845**	**0.576**	**0.709**	**0.701**	**0.699**	-0.128	**-0.248**	**-0.323**	**0.697**	**-0.404**	0.116	-0.029	-0.134	-0.220	**1**

1.2. Multiple Linear Regressions

To generate the quantitative relationships between toxicity pIC_{50} and selected descriptors, our data set were subjected to the MLR and MNLR. Only variables with significant coefficients were retained.

1.3. Multiple linear regression of the variable toxicity (MLR)

Many attempts have been made to develop a relationship with the indicator variable of toxicity pIC_{50}, but the best relationship obtained by this method is only one corresponding to the linear combination of two descriptors selected, the energy E_{LUMO} and the octanol/water partition coefficient (log *P*).

The resulting equation is:

$$\mathbf{pIC_{50}} = -1{,}275 - 0{,}170\times\mathbf{E_{LUMO}} + 0{,}675\times\mathbf{log\ \textit{P}} \qquad (1)$$

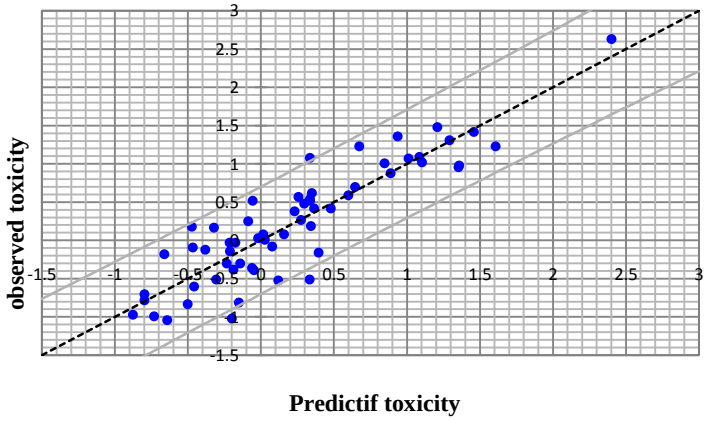

Figure 1: Graphical representation of calculated and observed toxicity by MLR

For our 60 compounds, the correlation between experimental toxicity and calculated on based on this model is quite significant (figure 1) as indicated by statistical values:

$$\mathbf{N = 60 \quad R^2 = 0.801 \quad R^2_{CV} = 0.777 \quad MSE = 0.120 \quad F = 114.457 \quad P<0.0001}$$

The figure 1 shows a very regular distribution of toxicity values depending on the experimental values.

In the equation, **N** is the number of compounds, $\mathbf{R^2}$ is the determination coefficient, **MSE** is the mean squared error, **F** is the Fisher's criterion and **P** is the significance level.

A higher correlation coefficient and lower mean squared error indicate that the model is more reliable. A **P** that is smaller than 0.05 exhibits that the regression equation is statistically significant. The QSTR model expressed by Eq. (1) is cross-validated by its noticeable R^2cv value (R^2cv =0.777) obtained by the leave-one-out (LOO) method. A value of R^2cv is greater than 0.5 is the important criterion for qualifying a QSTR model as valid [16]. The correlation coefficients between descriptors in the model were calculated by variance inflation factor (**VIF**) as shown in table 4. The **VIF** was defined as $1/(1-R^2)$, where R was the multiple correlation coefficients for one

independent variable against all the other descriptors in the model. If **VIF** greater than 5, it mean that models were unstable and must be rejected, models with a **VIF** values between 1 and 4 can be accepted. As can be seen from table 4, the **VIF** values of the two descriptors are all smaller than 5.0, resulting that there is no-collinearity between the selected descriptors and the obtained model has good stability.

Table 4: The variance inflation factors (VIF) of descriptors in QSAR model

	E_{LUMO}	log *P*
Tolerance	0.997	0.997
VIF	1.003	1.003

The elaborated QSTR model reveals that the toxicity of 60 heterogeneous phenols to *Tetrahymena pyriformis* may be explained by the two selected descriptors in Eq (1). The negative correlation of the energy E_{LUMO} with the pIC_{50} shows that an increase in the values of this factor indicates a decrease in the value of the pIC_{50}, whereas a positive correlation of the octanol/water partition coefficient (log *P*) with the pIC_{50} reveals an increase in the value of the pIC_{50}.

1.4. Multiple nonlinear regression of the variable toxicity (MNLR)

The nonlinear regression method was also used to improve the structure toxicity in a quantitative way, taking into account several parameters. We have applied it to table 2 containing 60 molecules associated with fourteen variables. We used a pre-programmed function of XLSTAT following:

$$\mathbf{Y = a + (b\ X1 + c\ X2 + d\ X3 + e\ X4 \ldots)}$$

Where a, b, c, d...: represent the parameters and X1, X2, X3, X4,...: represent the variables.
The resulting equation is:

$$\mathbf{pIC_{50}} = -1{,}195 - 0{,}151\times\mathbf{E_{LUMO}} + 0{,}595\times\mathbf{log}\ \boldsymbol{P} + 8{,}116\ 10^{-3}\times\mathbf{E^2_{LUMO}} + 1{,}690\ \ 10^{-2}\times\mathbf{log^2}\boldsymbol{P} \quad (2)$$

The obtained parameters describing the topological and the electronic aspects of the studied molecules are:

$$\mathbf{N = 60 \quad R^2 = 0.802 \quad R^2_{CV} = 0.751 \quad MSE = 0.124}$$

The toxicity values pIC_{50} predicted by this model are almost similar to that observed. The figure 2 shows a very regular distribution of toxicity values based on the observed values.
The obtained coefficient of determination in equation (2) is quite very interesting (**0.802**). The QSTR model expressed by Eq. (2) is cross-validated by its appreciable R^2cv values (R^2cv =0.751) obtained using the leave-one-out (LOO) method. A value of R^2cv is greater than 0.5 is the important criterion for qualifying a QSTR model as valid [16].
To optimize the error standard deviation and to improve our model, we involve in the next part artificial neural networks (ANN).

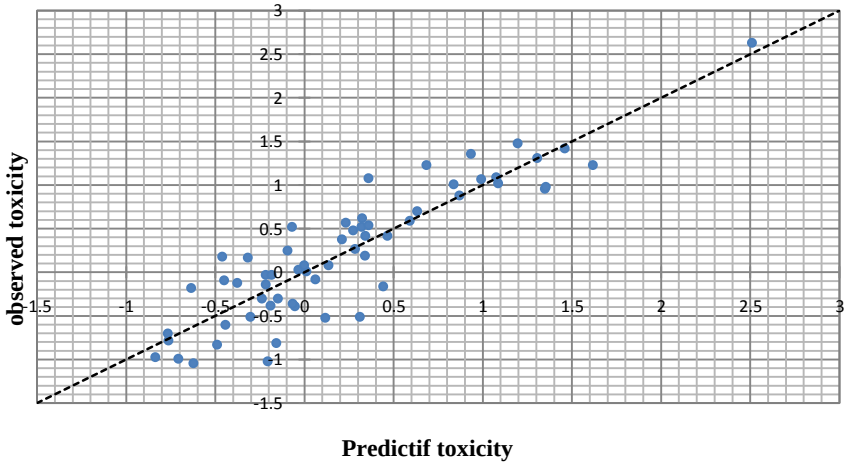

Figure 2: Graphical representation of calculated and observed toxicity by MNLR

1.5. Artificial neural networks ANN

In order to increase the probability of good characterization of studied compounds, neural networks (ANN) can be used to establish predictive models of quantitative structure–toxicity relationships (QSTR) between a set of molecular descriptors obtained from the MLR and observed toxicity. The ANN calculated toxicity model was developed using the parameters of the studied compounds. The correlation between ANN calculated and experimental toxicity values are very significant as illustrated in figure 3.

$$\mathbf{N = 60 \quad R^2 = 0.824 \quad R^2_{CV} = 0.704 \quad MSE = 0.100}$$

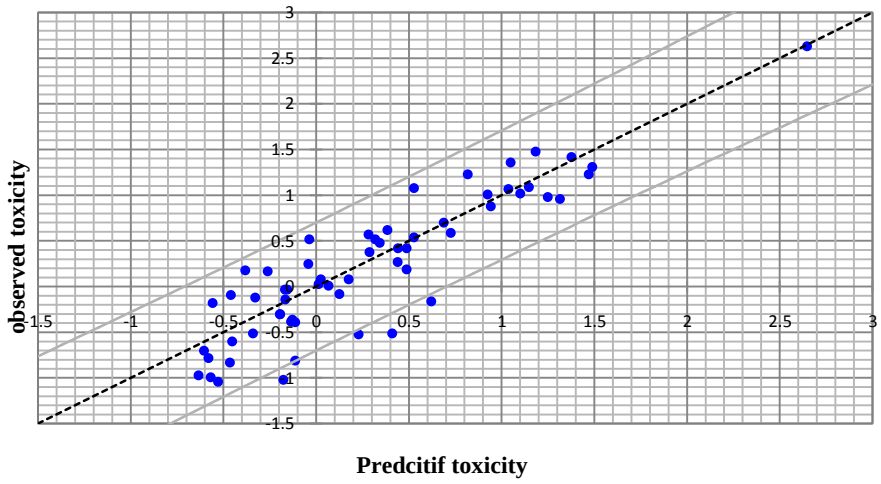

Figure 3: Correlations of observed and predicted activities calculated using ANN

The obtained determination coefficient (R^2) value is 0.824 for this data set of the heterogeneous phenol derivatives. This confirms that the artificial neural network (ANN) results are the best to predict the quantitative structure-activity relationship model. Furthermore, the R^2cv value shows that the ANN model is the high predictive power.

1.6. External validation

To estimate the predictive power of the MLR, MNLR and ANN models, we must use a set of compounds that have not been used for training set to establish the QSTR model. The models established in the computation procedure by using the 60 heterogeneous phenols are used to predict the toxicity of the remaining 10 compounds. The principal performance metrics of the three models are shown in table 5. As seen from this table, the statistical indictors of the ANN model are more significant than the other models.

Table 5: Performance comparison between models obtained by MLR, RNLM and ANN

Model	Training set			Test set		
	R^2	R^2cv	MSE	R^2	R^2ext	MSE
MLR	0.801	0.777	0.120	0.801	0.708	0.132
MNLR	0.802	0.751	0.124	0.802	0.716	0.115
ANN	0.824	0.704	0.100	0.824	0.773	0.109

We assessed the best linear QSTR regression equations developed in this study. Based on this result, a comparison of the quality of the MLR and MNLR models indicates that the ANN model has a significantly better predictive ability because the ANN approach gives better results than those of MLR and MNLR. ANN establishes a satisfactory relationship between the molecular descriptors and the toxicity of the studied compounds.

The accuracy and predictability of the proposed models were illustrated by comparing key statistical parameters, such as the R or R^2 of different models obtained using different statistical tools and different descriptors, as shown in Table 6.

Table 6: Observed values and calculated values of pIC_{50} according to different methods

N°	pIC_{50} (obs.)	pIC_{50} (calc.) MLR	NMLR	ANN	N°	pIC_{50} (obs.)	pIC_{50} (calc.) MLR	NMLR	ANN
1*	− 1.50	-0.365	-0.365	-0.362	**36**	0.25	-0.087	-0.097	-0.043
2	− 1.04	-0.642	-0.624	-0.530	**37***	0.31	0.619	0.593	0.698
3	− 1.02	-0.200	-0.208	-0.178	**38**	0.38	0.230	0.208	0.287
4	− 0.99	-0.733	-0.708	-0.569	**39***	0.38	-0.287	-0.290	-0.247
5	− 0.97	-0.877	-0.838	-0.635	**40**	0.42	0.478	0.464	0.487
6	− 0.83	-0.502	-0.491	-0.466	**41**	0.42	0.364	0.340	0.440
7	− 0.81	-0.151	-0.159	-0.114	**42**	0.48	0.296	0.271	0.341
8	− 0.78	-0.797	-0.764	-0.581	**43**	0.52	-0.057	-0.072	-0.038
9	0.70	-0.799	-0.767	-0.606	**44**	0.52	0.334	0.318	0.317
10	− 0.60	-0.459	-0.444	-0.454	**45**	0.57	0.256	0.231	0.282
11	− 0.52	0.119	0.115	0.228	**46**	0.59	0.598	0.588	0.725
12	− 0.51	0.333	0.310	0.409	**47**	0.62	0.348	0.322	0.383
13	− 0.51	-0.308	-0.305	-0.341	**48**	0.70	0.643	0.630	0.686
14	− 0.39	-0.049	-0.056	-0.113	**49***	0.75	0.635	0.630	0.681
15	− 0.38	-0.189	-0.192	-0.137	**50**	0.88	0.888	0.868	0.941
16	− 0.36	-0.062	-0.068	-0.129	**51**	0.96	1.351	1.346	1.313
17	− 0.30	-0.235	-0.240	-0.198	**52**	0.98	1.357	1.353	1.248
18	− 0.30	-0.144	-0.149	-0.196	**53**	1.01	0.845	0.835	0.923
19*	− 0.24	-0.283	-0.285	-0.241	**54**	1.02	1.101	1.084	1.100
20	− 0.18	-0.662	-0.636	-0.559	**55**	1.07	1.009	0.989	1.036
21*	− 0.16	-0.593	-0.578	-0.494	**56**	1.09	1.085	1.074	1.146
22	− 0.14	-0.214	-0.218	-0.167	**57**	1.23	1.606	1.616	1.469
23	− 0.12	-0.381	-0.379	-0.329	**58**	1.31	1.291	1.304	1.487
24	− 0.09	-0.467	-0.451	-0.461	**59***	1.33	0.903	0.882	0.952
25	− 0.08	0.076	0.059	0.124	**60**	1.36	0.935	0.932	1.047
26	− 0.03	-0.214	-0.218	-0.169	**61**	1.42	1.457	1.458	1.375
27	− 0.03	-0.176	-0.186	-0.158	**62**	1.48	1.208	1.195	1.182
28	0.01	0.026	0.010	0.066	**63***	2.03	1.728	1.754	1.818
29	0.03	-0.020	-0.035	0.012	**64**	2.63	2.399	2.507	2.646
30	0.08	0.157	0.134	0.174	**65**	− 0.16	0.393	0.440	0.620
31	0.08	0.015	-0.004	0.024	**66**	0.27	0.273	0.283	0.438
32*	0.12	0.233	0.210	0.276	**67**	0.54	0.335	0.358	0.527
33	0.17	-0.322	-0.320	-0.263	**68***	0.95	0.389	0.423	0.600
34	0.18	-0.474	-0.462	-0.383	**69**	1.08	0.335	0.358	0.527
35	0.19	0.342	0.339	0.486	**70**	1.23	0.672	0.682	0.817

* Test set

1.7. Domain of applicability

To estimate the reliability of any QSTR model and its ability to predict new compounds, the domain of applicability must be essentially defined. The predicted compounds that fall within this domain may be considered as reliable. The applicability domain was discussed with the Williams graph in figure 4, which the standardized residuals and the leverage values (h_i) are plotted.

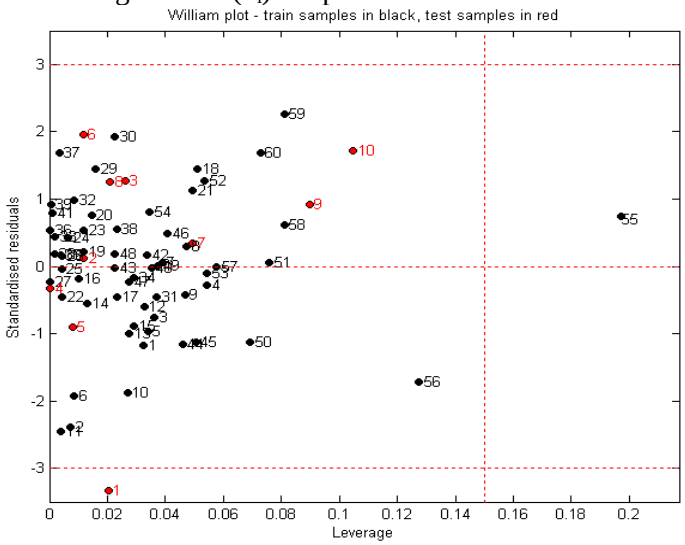

Figure 4: Williams plot for the presented MLR model

It is based on the calculation of the leverage h_i for each molecule, for which QSAR model is used to predict its toxicity:

$$h_i = x_i (X^T X)^{-1} x_i^T \qquad (i = 1, \dots, n) \qquad (3)$$

Where x_i is the row vector of the descriptors of compound i and X is the variable matrix deduced from the training set variable values. The index T refers to the matrix/vector transposed. The critical leverage h^* is, generally, fixed at 3(k+1)/N, where N is the number of training molecules, and k is the number of model descriptors. If the leverage value h of molecule is higher than the critical value (h^*) i.e., $h > h^*$, the prediction of the compound can be considered as not reliable.

The Williams plot for the presented MLR model is shown in figure 4. From this plot, the leverage values (h_i) of any molecule in the training and test sets are less than the critical value ($h^* = 0.15$) excepting the compounds 1 and 55 as outliers. Also, the standardized residuals of all molecules in the training and test sets are less than three standard deviation units (±3σ). Thus, the predicted toxicity by the developed MLR model is reliable.

Conclusion

In this study, three different methods, MLR, MNLR and ANN were used to generate the QSTR models for predicting the toxicity of heterogeneous phenols to *Tetrahymena pyriformis* and the resulting models were compared. It was shown the artificial neural network (ANN) results have substantially better predictive ability than the MLR and MNLR, yields a regression model with improved predictive power, we have established a relationship between several descriptors and the pIC_{50} values of the studied the heterogeneous phenol derivatives in a satisfactory manner.

Finally, we conclude that the studied descriptors, which are sufficiently rich in chemical, electronic and physic-chemical information to encode the structural features, might be used with other descriptors for the development of predictive QSAR models.

ACKNOWLEDGMENT-We are grateful to the "Association Marocaine des Chimistes Théoriciens" (AMCT) for its pertinent help concerning the programs.

References

1. Zang Y.B., *Chinese Agricultural Science Bulletin* 28 (2012) 282-285.
2. Zhan P.R., Wang H.T., Chen Z.X., Journal of Agro-Environment Science 27 (2008) 801-804.
3. Chen J.W., Peijnenburg W.J., Quan X., Chemsphere 40 (2000) 1319-1326.
4. Zhu M.J., F. Ge, Zhu R.L., Chemosphere 80 (2010) 46-52.
5. Duchowicz P.R., Mercader A.G., Fernández F.M., Castro E.A., Chemometrics and Intelligent Laboratory Systems 90 (2008) 97-107.
6. Frisch M.J. et al, Gaussian 03, Revision B.01, Gaussian, Inc., Pittsburgh, PA, (2003).
7. Advanced Chemistry Development Inc., Toronto, Canada (2009).
8. Ousaa A., Elidrissi B., Ghamali M., Chtita S., Bouachrine M., Lakhlifi T., Comp J. Meth. Mol. Des. 5(3) (2015) 16-24.
9. Larif M., Adad A., Hmammouchi R., Taghki A.I., Soulaymani A., Elmidaoui A., Bouachrine M., Lakhlifi T., Arab. J. Chem. (2016), http://dx.doi.org/10.1016/j.arabjc.2012.12.033 (in press).
10. XLSTAT 2013 software (XLSTAT Company). http://www.xlstat.com.
11. Hmamouchi R., Taghki A.I., Larif M., Adad A., Abdellaoui A., Bouachrine M., Lakhlifi T., Chem J.. Pharm. Research 5(9) (2013) 198-209.
12. Zupan J., Gasteiger J., VCH Publishers Weinheim, (1993).
13. Cherquaoui D., Villemin D., Chem J. Soc. Faraday. Trans. 90 (1994) 97-102.
14. Freeman J.A., Skapura D.M., Addison-Wesley, Reading, MA, (1991).
15. Ghamali M., Chtita S., Adad A., Hmamouchi R., Bouachrine M., Lakhlifi T., *IJARCSSE* 4 (2014) 536-546
16. Golbraikh A., Tropsha A., J. Mol. Graphics Model. 20 (2002) 269–276.
17. STATITCF Software, Technical Institute of cereals and fodder, Paris, France1987.
18. Ousaa A., Elidrissi B., Ghamali M., Chtita S., Bouachrine M., Lakhlifi T., *Comp J. Meth. Mol. Des.* 4(3) (2014) 10-18.