Journal of Materials and
Environmental Sciences
ISSN : 2028-2508
CODEN : JMESCN

Copyright © 2017,
University of Mohammed Premier
Oujda Morocco

J. Mater. Environ. Sci., 2018, Volume 9, Issue 1, Page xxx-xxx

Corrected Proof

http://www.jmaterenvironsci.com

Quantitative structure-toxicity relationship studies of aromatic aldehydes to *Tetrahymena pyriformis* based on electronic and topological descriptors

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, Moulay Ismail University, Meknes, Morocco
[2]MEM, ESTM, Moulay Ismail University, Meknes, Morocco

Received 17 Jun 2017,
Revised 19 Aug 2017,
Accepted 25 Aug 2017

Keywords
- ✓ *QSTR model,*
- ✓ *DFT study,*
- ✓ *Tetrahymena pyriformis,*
- ✓ *Aromatic aldehydes.*

abdellahousaa@gmail.com
Tel. +212671310939

Abstract

To establish a quantitative structure-toxicity relationship (QSTR) of a series of 77 aromatic aldehydes for their acute toxicity against *Tetrahymena pyriformis*, were used the principal component analysis, the multiple linear regression and the multiple nonlinear regression analysis. We proposed linear and nonlinear models and interpreted the toxicity of the compounds by multivariate statistical analysis. The proposed models have been validated using internal validation and external validation techniques and an agreement between experimental and predicted values was verified. The applicability domains of MLR models were investigated using William's plot to detect outliers and outsides compounds. It is interesting to note that some of the selected electronic and topological descriptors in our models are better for the prediction of new similar molecules.

1. Introduction

With the advent of modern science, technology and industrialization, the use of aromatic aldehydes is getting increased accompanied with an increased number of new chemicals [1]. 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 [2-3]. 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 aromatic aldehydes towards *Tetrahymena pyriformis* using several statistical tools, principal components analysis (PCA), multiple linear regression (MLR) and multiple non-linear regression (MNLR).

2. Material and methods

2.1. Experimental data

To determine a quantitative structure-toxicity relationship, we studied a series of 77 selected aromatic aldehydes for their acute toxicity against the protozoan ciliate *Tetrahymena pyriformis* [4]. 66 molecules were selected to propose the quantitative model (training set) as well as 11 compounds that were not used in the training set were selected randomly served to test the performance of the proposed model (test set). The following table shows the studied compounds and the corresponding experimental toxicties pIC_{50} (Table 1).

Table 1: The range of the toxicity data varies between -1.50 and 2.63 (µM)

N°	Name (IUPAC)	pIC_{50}	N°	Name (IUPAC)	pIC_{50}
1	4-Nitrobenzaldehyde	0.203	**40**	2-Chloro-3-hydroxy-4-methoxybenzaldehyde	0.204
2	1-Naphthaldehyde	0.423	**41**	6-Chloro-2-fluoro-3-methylbenzaldehyde	1.238
3	4-Biphenylcarboxaldehyde	1.119	**42**	3-Chloro-2-fluoro-5-(trifluoromethyl)benzaldehyde	1.723
4	4-Bromobenzaldehyde	0.587	**43**	2,3,5-Trichlorobenzaldehyde	1.499
5	4-Cyanobenzaldehyde	0.043	**44**	2-Fluorenecarboxaldehyde	1.499
6	Benzaldehyde	-0.196	**45**	2-Methyl-1-naphthaldehyde	1.231
7	p-Tolualdehyde	-0.057	**46**	4-Methyl-1-naphthaldehyde	1.123
8	4-Fluorobenzaldehyde	-0.127	**47**	Phenanthrene-9-carboxaldehyde	1.708
9	4-Chlorobenzaldehyde	0.400	**48**	5-Hydroxy-2-nitrobenzaldehyde	0.329
10	4-Ethylbenzaldehyde	0.291	49	3-Hydroxy-4-nitrobenzaldehyde	0.273
11	Terephthaldicarboxaldehyde	-0.086	**50**	3-Hydroxybenzaldehyde	0.085
12	4-Anisaldehyde	-0.047	**51**	3-Hydroxy-4-methoxybenzaldehyde	-0.142
13	4-Ethoxybenzaldehyde	0.073	**52**	3,4-Dimethoxy-5-hydroxycarboxaldehyde	-0.390
14	4-Acetamidobenzaldehyde	-0.224	**53**	2,3-Dihydroxybenzaldehyde	0.111
15	2-Tolualdehyde	0.011	**54**	2,5-Dihydroxybenzaldehyde	0.277
16	3-Tolualdehyde	0.081	**55**	3,4-Dihydroxybenzaldehyde	0.107
17	2-Chlorobenzaldehyde	0.487	**56**	3,4,5-Trihydroxybenzaldehyde	-0.196
18	3-Chlorobenzaldehyde	0.406	**57**	2,3,4-Trihydroxybenzaldehyde	0.001
19	2-Nitrobenzaldehyde	0.167	**58**	2,4,6-Trihydroxybenzaldehyde	0.128
20	3-Nitrobenzaldehyde	0.178	**59**	2,4-Dihydroxybenzaldehyde	0.515
21	Phenyl-1,3-dialdehyde	0.183	**60**	3-Ethoxy-2-hydroxycarboxaldehyde	0.850
22	2-Anisaldehyde	0.148	**61**	3-Methoxysalicylaldehyde	0.377
23	3-Anisaldehyde	0.232	**62**	3,5-Dibromosalicylaldehyde	1.648
24	3-Bromobenzaldehyde	0.506	**63**	4,6-Dimethoxy-2-hydroxybenzaldehyde	0.617
25	3-Fluorobenzaldehyde	0.154	**64**	2-Hydroxy-3-nitrocarboxaldehyde	0.870
26	2,4-Dichlorobenzaldehyde	1.036	**65**	2-Chloro-4-hydroxycarboxaldehyde	0.890
27	2,4-Dimethoxybenzaldehyde	-0.056	**66**	4-Hydroxy-3-nitrobenzaldehyde	0.614
28	2,4,5-Trimethoxybenzaldehyde	-0.101	**67**	4-Hydroxybenzaldehyde	0.266
29	4-(Dimethylamino)benzaldehyde	0.231	**68**	2-Hydroxy-1-naphthaldehyde	1.320
30	4-Phenoxybenzaldehyde	1.257	**69**	5-Bromovanillin	0.617
31	2-Bromobenzaldehyde	0.477	**70**	4-Hydroxy-1-naphthaldehyde	1.050
32	2-Fluorobenzaldehyde	0.079	**71**	5-Bromosalicylaldehyde	1.107
33	4-Butoxybenzaldehyde	0.716	**72**	5-Chlorosalicylaldehyde	1.009
34	4-(Pentyloxy)benzaldehyde	1.179	**73**	2-Hydroxybenzaldehyde	0.424
35	4-Isopropylbenzaldehyde	0.67	**74**	3-Bromo-4-hydroxycarboxaldehyde	0.610
36	Pentafluorobenzaldehyde	0.815	**75**	3-Methoxy-4-hydroxybenzaldehyde	-0.030
37	2-Chloro-5-nitrobenzaldehyde	0.527	**76**	3,5-Dibromo-4-hydroxycarboxaldehyde	0.890
38	2-Chloro-6-fluorobenzaldehyde	0.155	**77**	3-Ethoxy-4-hydroxybenzaldehyde	0.015
39	3-Cyanobenzaldehyde	-0.020			

* Test set

2.2. Computational methods

An attempt has been made to correlate the toxicity of these compounds with various physicochemical parameters. DFT (density functional theory) and ChemSketch program methods were used in this study. 3D structures of the molecules were generated using the Gauss View 3.0 and then, all of the calculations were performed using the Gaussian 03 W program series. Geometry optimization of the 77 compounds was carried out by a B3LYP function employing a 6–31G (d) basis set [5,6]. The geometry of all of the species under investigation was determined by optimizing all of the geometrical variables without any symmetry constraints [7].

2.3. Calculation of the molecular descriptors

From the results of the DFT calculations, 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 [8] 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*).

2.4. Statistical analysis

To explain the structure-toxicity relationship, these 14 descriptors were calculated for the 77 molecules using the Gaussian 03W and ChemSketch program software. The study that we conducted consists of multiple linear regression (MLR) and non-linear regression (MNLR), which are available in the XLSTAT software [9]. The multiple linear regression statistical techniques used to study the relationship between one dependent variable and several independent variables. It is a mathematical technique that minimizes the differences between actual and predicted values. It has also served to select descriptors that are used as input parameters in multiple non-linear regression (MNLR).The MLR and MNLR techniques was employed to model the structure-toxicity relationships. The equations were justified by the correlation coefficient (R), the Mean Squared Error (MSE), the Fisher F-statistic (F), and the significance level (p-value) [10].

3. Results and discussion

3.1. Data set for analysis

QSTR analysis was performed using the pIC_{50} of 77 selected aromatic aldehydes to *Tetrahymena pyriformis* as reported in [4], the values of the 14 chemical descriptors are shown in table 2.

Table 2: The values of the fourteen chemical descriptors

N	pIC_{50}	MW	MR	MV	Pc	n	γ	D	α	E_T	E_{HOMO}	E_{LUMO}	ΔE	μ	log *P*
1*	0.203	151.12	39.55	112.90	307.80	1.62	55.10	1.34	15.67	-14978.44	7.03	-3.02	-10.05	2.27	1.303
2	0.423	156.18	50.84	135.20	356.10	1.68	48.10	1.16	20.15	-13593.56	-6.21	-1.99	4.22	3.55	2.463
3	1.119	182.22	57.59	166.30	425.60	1.61	42.80	1.10	22.83	-15701.67	-6.43	-1.85	4.57	3.86	2.777
4	0.587	185.02	40.69	117.20	302.80	1.61	44.40	1.58	16.13	-79421.13	-7.02	-1.99	5.03	2.07	2.231
5	0.043	131.13	36.28	113.50	300.20	1.55	48.90	1.15	14.38	-11921.64	-7.52	-2.63	4.89	2.62	1.814
6	-0.196	106.12	33.00	101.00	252.30	1.57	38.80	1.05	13.08	-9409.96	-6.95	-1.71	5.24	3.30	1.343
7	-0.057	120.15	37.83	117.30	289.90	1.56	37.20	1.02	14.99	-10480.62	-6.85	-1.60	5.25	3.78	1.758
8	-0.127	124.11	32.99	105.30	259.40	1.54	36.80	1.18	13.08	-12112.10	-7.06	-1.76	5.30	2.31	1.479
9	0.400	140.57	37.90	113.00	288.20	1.59	42.20	1.24	15.02	-21924.75	-7.16	-1.99	5.17	1.98	1.962
10	0.291	134.18	42.55	133.90	328.90	1.55	36.30	1.00	16.87	-11551.08	-6.82	-1.57	5.25	3.96	2.214
11	-0.086	134.13	39.75	112.70	297.30	1.62	48.30	1.19	15.76	-12495.73	-7.29	-2.67	4.62	1.41	0.82
12	-0.047	136.15	39.68	125.10	309.00	1.55	37.20	1.09	15.73	-12528.49	-6.36	-1.41	4.95	4.02	1.474
13	0.073	150.17	44.31	141.60	348.70	1.54	36.80	1.06	17.56	-13599.10	-6.29	-1.37	4.92	4.25	1.818
14	-0.224	163.17	47.27	134.10	356.00	1.62	49.50	1.22	18.74	-15088.10	-6.64	-3.61	3.03	3.70	0.759
15	0.011	120.15	37.83	117.30	289.90	1.56	37.20	1.02	14.99	-10480.51	-6.83	-1.71	5.12	3.34	1.758
16	0.081	120.15	37.83	117.30	289.90	1.56	37.20	1.02	14.99	-10480.60	-6.88	-1.65	5.23	3.71	1.758
17	0.487	140.57	37.90	113.00	288.20	1.59	42.20	1.24	15.02	-21924.65	-7.11	-2.01	5.09	3.30	1.962
18	0.406	140.57	37.90	113.00	288.20	1.59	42.20	1.24	15.02	-21924.73	-7.13	-2.03	5.10	1.79	1.962
19*	0.167	151.12	39.55	112.90	307.80	1.62	55.10	1.34	15.67	-14978.05	-7.39	-2.60	4.79	6.52	1.303
20	0.178	151.12	39.55	112.90	307.80	1.62	55.10	1.34	15.67	-14978.44	-7.51	-2.75	4.76	5.38	1.303
21*	0.183	134.13	39.75	112.70	297.30	1.62	48.30	1.19	15.76	-12495.71	-7.14	-2.16	4.99	5.41	0.82
22	0.148	136.15	39.68	125.10	309.00	1.55	37.20	1.09	15.73	-12528.27	-6.29	-1.44	4.85	3.87	1.474
23	0.232	136.15	39.68	125.10	309.00	1.55	37.20	1.09	15.73	-12528.45	-6.34	-1.67	4.66	2.18	1.474
24	0.506	185.02	40.69	117.20	302.80	1.61	44.40	1.58	16.13	-79421.11	-7.00	-2.03	4.97	3.57	2.231
25	0.154	124.11	32.99	105.30	259.40	1.54	36.80	1.18	13.08	-12112.06	-7.15	-1.95	5.19	3.49	1.479
26	1.036	175.01	42.79	125.00	324.00	1.60	45.10	1.40	16.96	-34439.40	-7.30	-2.26	5.05	1.63	2.581
27	-0.056	166.17	46.36	149.10	365.60	1.53	36.10	1.11	18.37	-15629.12	-6.13	-1.08	5.05	6.36	1.605
28	-0.101	196.20	53.04	173.10	422.30	1.52	35.40	1.13	21.02	-18765.17	-5.45	-1.18	4.27	5.28	1.736
29	0.231	149.19	47.31	139.00	354.30	1.60	42.10	1.07	18.75	-13058.02	-5.50	-1.06	4.43	6.32	2.177
30	1.257	198.22	59.44	171.60	443.40	1.61	44.50	1.15	23.56	-17749.58	-6.36	-1.49	4.87	4.70	3.318
31	0.477	185.02	40.69	117.20	302.80	1.61	44.40	1.58	16.13	-79421.07	-7.03	-2.00	5.03	3.25	2.231
32*	0.079	124.11	32.99	105.30	259.40	1.54	36.80	1.18	13.08	-12112.04	-7.03	-1.88	5.15	3.40	1.479
33	0.716	178.23	53.58	174.60	428.30	1.53	36.20	1.02	21.24	-15740.19	-6.33	-1.36	4.97	5.13	2.73
34	1.179	192.25	58.21	191.10	468.10	1.52	35.90	1.01	23.07	-16810.66	-6.32	-1.35	4.97	5.16	3.186
35	0.67	148.20	47.19	151.10	367.30	1.54	34.80	0.98	18.70	-12621.39	-2.58	-1.41	1.17	4.85	2.418
36	0.815	196.07	32.97	122.10	287.90	1.45	30.80	1.61	13.07	-22919.39	-7.44	-2.41	5.04	2.00	2.572
37*	0.527	185.56	44.44	124.80	343.60	1.63	57.30	1.49	17.61	-27493.09	-7.64	-2.95	4.69	4.33	1.922
38	0.155	158.56	37.89	117.20	295.30	1.56	40.20	1.35	15.02	-24607.69	-7.15	-3.36	3.79	3.98	2.098
39*	-0.020	131.13	36.28	113.50	300.20	1.55	48.90	1.15	14.38	-11921.67	-7.49	-2.37	5.11	2.15	1.814
40	0.204	186.59	46.46	135.40	359.80	1.60	49.70	1.38	18.41	-27091.24	-6.22	-2.01	4.20	5.96	1.704
41	1.238	172.58	42.71	133.50	332.90	1.55	38.60	1.29	16.93	-25697.21	-6.85	-2.01	4.84	4.82	2.513
42	1.723	226.56	42.87	150.70	352.50	1.48	29.80	1.50	16.99	-33628.15	-7.61	-2.63	4.98	1.26	3.418
43	1.499	209.46	47.69	136.90	359.90	1.61	47.70	1.53	18.90	-46953.96	-7.27	-2.52	4.75	1.52	3.2

44	1.499	194.23	60.54	159.90	431.80	1.68	53.00	1.21	24.00	-16739.62	-4.13	-1.91	2.22	3.71	2.801
45	1.231	170.21	55.67	151.40	393.80	1.66	45.60	1.12	22.07	-14664.09	-6.13	-1.91	4.22	3.17	2.878
46	1.123	170.21	55.67	151.40	393.80	1.66	45.60	1.12	22.07	-14664.17	-6.10	-1.92	4.17	3.92	2.878
47	1.708	206.24	68.69	169.30	459.90	1.75	54.40	1.22	27.23	-17776.88	-4.98	-1.82	3.16	2.16	3.583
48	0.329	167.12	41.43	111.30	322.80	1.67	70.50	1.50	16.42	-17026.21	-7.05	-2.52	4.53	7.29	0.914
49*	0.273	167.12	41.43	111.30	322.80	1.67	70.50	1.50	16.42	-17026.81	-7.17	-3.30	3.87	1.41	0.914
50	0.085	122.12	34.88	99.50	267.30	1.62	52.00	1.23	13.83	-11458.09	-6.43	-1.69	4.75	4.17	0.954
51	-0.142	152.15	41.56	123.50	324.00	1.59	47.30	1.23	16.47	-14576.64	-6.00	-1.40	4.59	5.18	1.085
52	-0.390	226.18	55.17	160.00	442.80	1.61	58.50	1.41	21.87	-22828.98	-4.03	-2.83	1.20	8.71	0.696
53	0.111	138.12	36.76	97.90	282.30	1.67	69.00	1.41	14.57	-13506.24	-6.13	-1.51	4.62	5.72	0.565
54	0.277	138.12	36.76	97.90	282.30	1.67	69.00	1.41	14.57	-13506.12	-5.91	-1.61	4.30	6.05	0.565
55	0.107	138.12	36.76	97.90	282.30	1.67	69.00	1.41	14.57	-13506.36	-6.19	-1.48	4.71	2.30	0.565
56	-0.196	154.12	38.65	96.30	297.30	1.73	90.50	1.60	15.32	-15554.32	-5.97	-1.49	4.48	1.15	0.176
57	0.001	154.12	38.65	96.30	297.30	1.73	90.50	1.60	15.32	-15554.27	-6.34	-1.27	5.07	4.51	0.176
58	0.128	154.12	38.65	96.30	297.30	1.73	90.50	1.60	15.32	-15554.24	-6.13	-1.05	5.08	5.65	0.176
59*	0.515	138.12	36.76	97.90	282.30	1.67	69.00	1.41	14.57	-13506.30	-6.31	-1.32	5.00	4.57	0.565
60	0.850	210.18	53.12	152.50	425.90	1.61	60.70	1.38	21.06	-20781.76	-6.86	-2.39	4.47	5.50	1.160
61	0.377	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	1.085
62	1.648	279.91	50.26	131.80	368.30	1.69	60.80	2.12	19.92	-151480.4	-6.66	-2.15	4.51	3.56	2.73
63*	0.617	182.17	48.24	147.50	380.60	1.57	44.30	1.23	19.12	-17695.53	-6.07	-1.37	4.70	4.66	1.216
64	0.870	211.13	48.36	123.90	384.90	1.71	93.10	1.70	19.17	-17027.11	-6.45	-3.13	3.32	3.94	0.271
65	0.890	200.58	46.71	124.00	365.30	1.68	75.30	1.62	18.51	-29106.95	-7.05	-2.03	5.02	5.34	1.507
66	0.614	167.12	41.43	111.30	322.80	1.67	70.50	1.50	16.42	-17026.92	-7.26	-3.09	4.17	1.09	0.914
67	0.266	122.12	34.88	99.50	267.30	1.62	52.00	1.23	13.83	-11458.17	-6.50	-1.45	5.04	3.38	0.954
68	1.320	172.18	52.72	133.60	371.10	1.72	59.40	1.29	20.90	-15641.51	-5.96	-1.78	4.19	4.43	2.074
69	0.617	231.04	49.25	139.70	374.50	1.62	51.60	1.65	19.52	-84587.80	-6.31	-1.65	4.66	3.07	1.973
70	1.050	172.18	52.72	133.60	371.10	1.72	59.40	1.29	20.90	-15641.78	-5.88	-1.74	4.14	3.66	2.074
71	1.107	201.02	42.57	115.70	317.80	1.66	56.90	1.74	16.87	-81469.21	-6.47	-1.90	4.57	5.11	1.842
72	1.009	156.57	39.78	111.40	303.20	1.63	54.70	1.40	15.77	-23972.82	-6.55	-1.91	4.64	5.19	1.573
73	0.424	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	0.954
74	0.610	201.02	42.57	115.70	317.80	1.66	56.90	1.74	16.87	-81469.42	-6.69	-1.73	4.96	2.90	1.842
75	-0.030	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	1.085
76	0.890	279.91	50.26	131.80	368.30	1.69	60.80	2.12	19.92	-151480.4	-6.83	-1.96	4.87	3.50	2.73
77	0.015	166.17	46.19	140.00	363.70	1.57	45.50	1.19	18.31	-15647.07	-6.34	-1.44	4.90	5.13	1.429

*Test set

3.2. Principal component analysis

The principal component analysis (PCA) was performed to the 14 descriptors of 77 molecules. The first three axes F1, F2 and F3 represent respectively (35.50%; 20.64% and 14.92%) of the total variance and they estimate 71.06% of the total information.

The principal component analysis (PCA) [11] was conducted to identify the link between the different descriptors. Bold values are different from 0 at a significance level of p= 0.05. 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 ≥ 0.9), were removed.

Table 3: Correlation matrix between different obtained descriptors

	pIC_{50}	**MW**	**MR**	**MV**	**Pc**	**N**	□	**D**	□	**E_T**	**E_{HOMO}**	**E_{LUMO}**	□ **E**	□	**log *P***
pIC_{50}	**1**														
MW	**0.622**	**1**													
MR	**0.591**	**0.639**	**1**												
MV	**0.474**	**0.534**	**0.874**	**1**											
Pc	**0.527**	**0.665**	**0.965**	**0.929**	**1**										
N	**0.229**	**0.239**	**0.264**	**-0.228**	0.109	**1**									
γ	0.001	0.182	0.006	**-0.383**	-0.027	**0.847**	**1**								
D	**0.298**	**0.668**	-0.025	**-0.262**	-0.040	**0.507**	**0.591**	**1**							
α	**0.591**	**0.639**	**1.000**	**0.874**	**0.965**	**0.264**	0.006	-0.025	**1**						
E_T	**-0.367**	**-0.711**	-0.143	-0.038	-0.099	-0.203	-0.083	**-0.763**	-0.143	**1**					
E_{HOMO}	-0.026	0.012	0.175	0.131	0.168	0.095	0.059	-0.090	0.175	0.102	**1**				
E_{LUMO}	-0.078	-0.214	0.037	0.100	0.012	-0.073	-0.148	**-0.295**	0.037	0.084	0.003	**1**			
ΔE	0.000	-0.078	-0.154	-0.093	-0.156	-0.113	-0.103	-0.008	-0.154	-0.071	**-0.949**	**0.313**	**1**		
μ	-0.179	0.048	0.192	0.198	**0.258**	0.029	0.105	-0.103	0.191	0.134	0.067	**0.265**	0.020	**1**	
log *P*	**0.699**	**0.436**	**0.542**	**0.656**	**0.497**	**-0.271**	**-0.570**	-0.094	**0.542**	**-0.318**	-0.042	0.018	0.045	**-0.243**	**1**

3.3. Multiple linear regressions MLR

Based on the 11 remaining descriptors, a mathematical linear model was proposed to predict quantitatively the physicochemical effects of substituents on the toxicity of the 66 molecules by using backward selection and stepwise selection in the multiple regression analysis.

The study of the descendant MLR multiple linear regression based on the elimination of descriptors until a valid model was obtained and the stepwise multiple linear regression procedures based on the forward selection and backward elimination methods were employed to determine the best regression models.

The QSAR models built using descendant and stepwise multiple linear regression methods are represented by the following equations:

For the descendant MLR:

$$\mathbf{pIC_{50}} = -8.176 + 9.649\ 10^{-03}\ \mathbf{MW} - 2.648\ 10^{-02}\ \mathbf{MR} + 4.591\ \mathbf{n} + 7.221\ 10^{-06}\ \mathbf{E_T} + 0.598\ \mathbf{log}\ \boldsymbol{P} \quad \text{(Equation 1)}$$

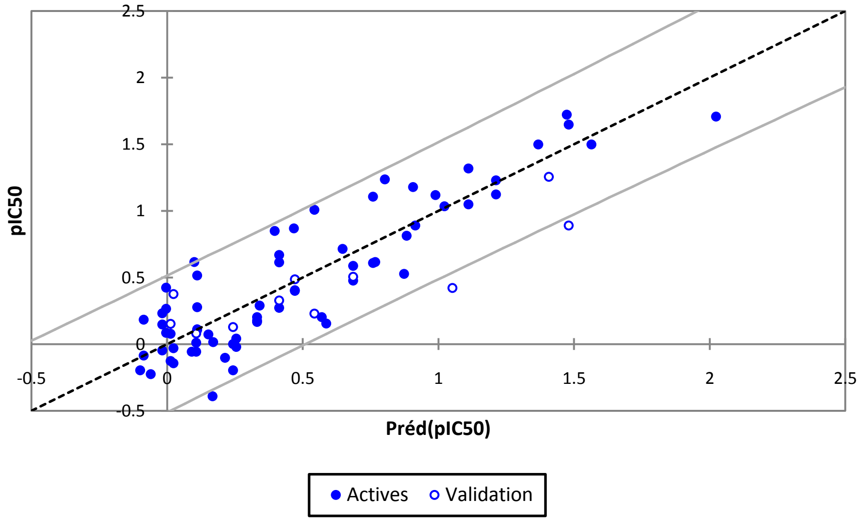


Figure 1: Graphical representation of calculated and observed toxicity by descendant MLR

For our 66 compounds, the correlation between experimental and calculated toxicity based on this model is quite significant (Figure 1) as indicated by statistical values:

$$\mathbf{N = 66 \quad R^2 = 0.799 \quad R^2_{CV} = 0.743 \quad MSE = 0.064 \quad F = 47.672 \quad p\text{-value} < 0.0001}$$

The elaborated QSTR model reveals that the toxicity of 66 aromatic aldehydes to *Tetrahymena pyriformis* could be explained by a number of electronic factors (**MW, n, $\mathbf{E_T}$** and **log *P***). The positive correlation of these factors with the value of the **$\mathbf{pIC_{50}}$** in equation 1 shows that an increase in the values of these factors implies an increase in the value of the **$\mathbf{pIC_{50}}$**, whereas a negative correlation of the **MR** shows that an increase in the value of this factor indicates a decrease in the value of the **$\mathbf{pIC_{50}}$**. For the stepwise MLR:

$$\mathbf{pIC_{50}} = -1.928 + 0.024\ \gamma + 0.709\ \mathbf{log}\ \boldsymbol{P} \quad \text{(Equation 2)}$$

For our 66 compounds, the correlation between experimental toxicity and calculated on based on this model is quite significant (Figure 2) as indicated by statistical values:

$$\mathbf{N = 66 \quad R^2 = 0.760 \quad R^2_{CV} = 0.732 \quad MSE = 0.073 \quad F = 99.483 \quad p\text{-value} < 0.0001}$$

The elaborated QSTR model reveals that the toxicity of 66 aromatic aldehydes to *Tetrahymena pyriformis* may be explained by the two selected descriptors in equation 2. The positive correlation of the **γ** and **log *P*** with the **$\mathbf{pIC_{50}}$** shows that an increase in the values of these factors implies a increase in the value of the **$\mathbf{pIC_{50}}$**.

The figures 1 and 2 show a very regular distribution of toxicity values depending on the experimental values. In the equation, N is the number of compounds, R^2 is the determination coefficient, MSE is the mean squared error, F is the Fisher's criterion and p-value 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 models expressed by equation 1 and equation 2 are cross-validated by its noticeable R^2_{cv} values (R^2_{cv} = 0.743 to a descendant MLR model and R^2_{cv} = 0,732 to a stepwise MLR model) obtained by the leave-one-out (LOO) method. A value of R^2_{cv} is greater than 0.5 is the important criterion for qualifying a QSTR model as valid [12]. The correlation coefficients between descriptors in the descendant MLR model were calculated by variance inflation factor (VIF) as shown in table 4.

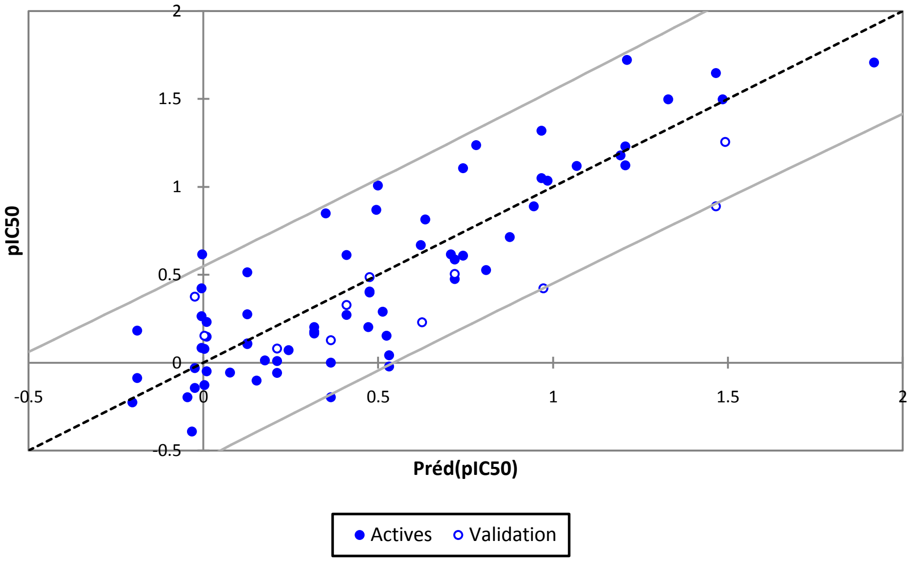


Figure 2: Graphical representation of calculated and observed toxicity by stepwise MLR

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

Statistique	**MW**	**MR**	**n**	**E_T**	**log *P***
Tolérance	0.216	0.243	0.573	0.311	0.427
VIF	4.628	4.110	1.746	3.221	2.344

The VIF was defined as 1/(1-R2), 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 5 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. With the MLR models, the values of predicted pIC_{50} calculated from equation 1 and equation 2 and the observed values are given in table 6.

3.4. Multiple nonlinear regression (MNLR)

We have used also the technique of nonlinear regression model to improve the structure-toxicity relationship to quantitatively evaluate the effect of the substituents and they have applied to the data matrix constituted obviously from the descriptors proposed by MLR corresponding to the 66 molecules (Training set).

The coefficients R^2, MSE are used to select the best regression performance. We used a pre-programmed function of XLSTAT following:

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

Where a, b, c, d...: represent the parameters and X1, X2, X3, X4….: represent the variables.

The proposed descriptors in equation 1 and equation 2 by MLR models are used as the input parameters in the MNLR method. The QSTR models built using multiple non-linear regression method are represented by the following equations:

The MNLR model using selected descriptors by descendant selection:

$$\mathbf{pIC_{50}} = -22.973 + 7.394\ 10^{-03}\ \mathbf{MW} + 2.923\ 10^{-02}\ \mathbf{MR} + 21.783\ \mathbf{n} + 1.295\ 10^{-05}\ \mathbf{E_T} + 0.456\ \mathbf{log}\ \boldsymbol{P} + 7.250\ 10^{-06}\ \mathbf{MW}^2 - 6.328\ 10^{-04}\ \mathbf{MR}^2 - 5.298\ \mathbf{n}^2 + 3.489\ 10^{-11}\ \mathbf{E}^2_{\mathbf{T}} + 0.053\ (\mathbf{log}\ \boldsymbol{P})^2 \quad \text{(Equation 3)}$$

$$\mathbf{N = 66 \quad R^2 = 0.810 \quad R^2_{CV} = 0.713 \quad MSE = 0.066}$$

The MNLR model using selected descriptors by stepwise selection:

$$\mathbf{pIC_{50}} = -2.093 + 4.111\ 10^{-02}\ \gamma + 0.423\ \mathbf{log}\ \boldsymbol{P} - 1.706\ 10^{-04}\ \gamma^2 + 7.150\ 10^{-02}\ (\mathbf{log}\ \boldsymbol{P})^2 \quad \text{(Equation 4)}$$

$$\mathbf{N = 66 \quad R^2 = 0.768 \quad R^2_{CV} = 0.732 \quad MSE = 0.073}$$

The higher values of R^2 of two MNLR models and the lower mean squared errors MSE indicate that the two proposed models are predictive and reliable. The obtained models were internally validated by the leave-one-out cross-validation technique. The values of R^2_{cv} for two MNLR models are higher than 0.5, indicate the better predictivity of MNLR models. The toxicity values pIC_{50} predicted by this model are almost similar to that observed. The correlations of predicted and observed pIC_{50} values are illustrated in figure 3.

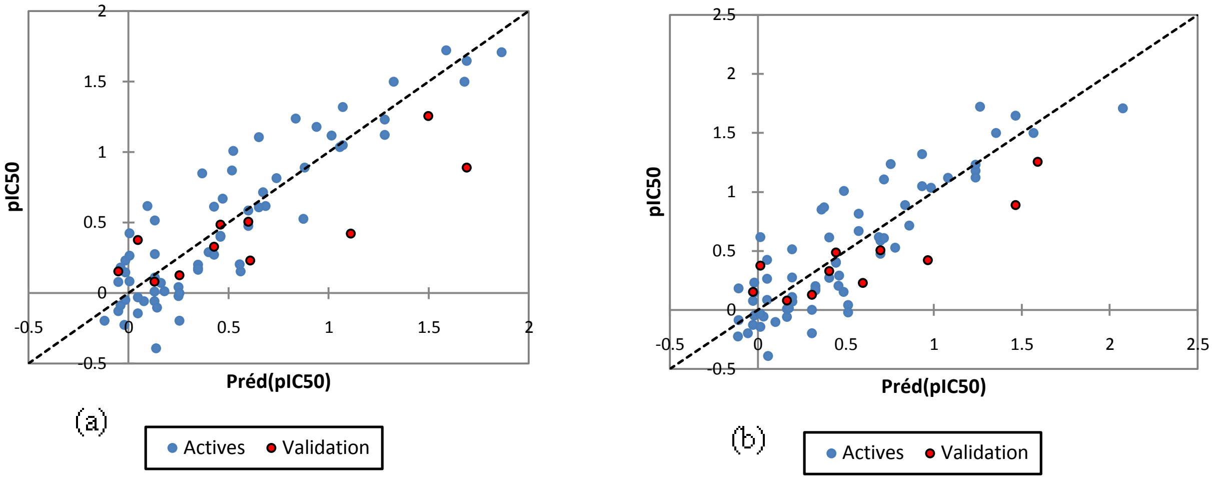


Figure 3: Graphical representation of calculated and observed toxicity by MNLR ((a): proposed descriptors by descendant selection and (b): by stepwise selection)

3.5. External validation

To estimate the predictive power of developed models, we must use a set of compounds that have not been used for training set to establish the QSTR models. The established models in the computation procedure using the 66 aromatic aldehydes are used to predict the toxicity of the remaining 11 compounds. The comparison of the values of pIC_{50}-test and pIC_{50}-obs shows that a good prediction has been obtained for the 11 compounds (R_{test} and R^2_{test} showed in table 5).

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

Model	**Training set**			**Test set**		
	R^2	**R^2cv**	**MSE**	**R ext**	**R^2ext**	**MSE**
MLR **descendant**	0.799	0,743	0.064	0.852	0.726	0.212
MLR **stepwise**	0.760	0,732	0.073	0.846	0.716	0.144
MNLR **descendant**	0.810	0.713	0.066	0.840	0.707	0.299
MNLR **stepwise**	0.768	0.732	0.073	0.875	0.766	0,139

The true predictive power of these models can be tested from their ability to predict perfectly the pIC_{50} of compounds from an external test set. The activities of the remaining set of 11 compounds are deduced from the quantitative proposed models in training set. The observed and calculated pIC_{50} values are given in table 6. These models were able to predict the activities of test set molecules in agreement with the experimentally determined value. The higher values of R^2test (R^2test = 0.726 for the descendant MLR model, R^2test = 0.716 for the stepwise MLR model, R^2test = 0.707 for MNLR model (with descriptors proposed by descendant MLR), and R^2test = 0.766 for MNLR model (with descriptors proposed by descendant MLR)) indicate the improved predictivity of these models.

A comparison of the quality of MLR and MNLR models shows that four approaches have the good predictive capability; which is sufficient to conclude the performance of these models and to establish a satisfactory relationship between selected descriptors and toxicity. Furthermore, the results obtained by MNLR are relatively better than those obtained by MLR, but the latter approach is more transparent and gives the most interpretable results and a good explanation of the descriptors associated with toxicities.

3.6. 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 [13]. The predicted compounds that fall within this domain may be considered as reliable. The applicability domain was discussed with the Williams graph in figures 4 and 5, which the standardized residuals and the leverage values (h_i) are plotted.

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

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

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.

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

		pIC50 (calc.)						**pIC50 (calc.)**			
N°	**pIC_{50} (obs.)**	**MLR_{step}**	**$NMLR_{step}$**	**MLR_{desc}**	**$NMLR_{desc}$**	**N°**	**pIC_{50} (obs.)**	**MLR_{step}**	**$NMLR_{step}$**	**MLR_{desc}**	**$NMLR_{desc}$**
1	0.203	0.316	0.327	0.331	0.346	**40**	0.204	0.472	0.457	0.569	0.554
2*	0.423	0.972	0.966	1.051	1.111	**41**	1.238	0.780	0.754	0.802	0.834
3	1.119	1.068	1.080	0.988	1.013	**42**	1.723	1.211	1.262	1.473	1.588
4	0.587	0.719	0.696	0.686	0.598	**43**	1.499	1.485	1.566	1.564	1.677
5	0.043	0.531	0.512	0.254	0.250	**44**	1.499	1.329	1.353	1.368	1.325
6	-0.196	-0.046	-0.058	-0.100	-0.121	**45**	1.231	1.207	1.237	1.212	1.280
7	-0.057	0.210	0.165	0.107	0.129	**46**	1.123	1.207	1.237	1.212	1.280
8	-0.127	0.003	-0.029	0.012	-0.052	**47**	1.708	1.918	2.073	2.023	1.864
9	0.400	0.475	0.443	0.470	0.459	**48***	0.329	0.409	0.404	0.413	0.427
10	0.291	0.513	0.462	0.341	0.400	**49**	0.273	0.409	0.404	0.413	0.427
11	-0.086	-0.189	-0.110	-0.086	-0.041	**50**	0.085	-0.006	0.052	-0.004	0.003
12	-0.047	0.009	-0.021	-0.018	-0.017	**51**	-0.142	-0.025	0.013	0.022	0.045
13	0.073	0.243	0.194	0.151	0.161	**52**	-0.390	-0.033	0.057	0.167	0.137
14	-0.224	-0.204	-0.114	-0.060	-0.022	**53**	0.111	0.126	0.193	0.110	0.131
15	0.011	0.210	0.165	0.107	0.129	**54**	0.277	0.126	0.193	0.110	0.131
16*	0.081	0.210	0.165	0.107	0.129	**55**	0.107	0.126	0.193	0.110	0.131
17*	0.487	0.475	0.443	0.470	0.459	**56**	-0.196	0.365	0.307	0.243	0.254
18	0.406	0.475	0.443	0.470	0.459	**57**	0.001	0.365	0.307	0.243	0.254
19	0.167	0.316	0.327	0.331	0.346	**58***	0.128	0.365	0.307	0.243	0.254
20	0.178	0.316	0.327	0.331	0.346	**59**	0.515	0.126	0.193	0.110	0.131
21	0.183	-0.189	-0.110	-0.086	-0.041	**60**	0.850	0.349	0.361	0.396	0.368
22	0.148	0.009	-0.021	-0.018	-0.017	**61***	0.377	-0.025	0.013	0.022	0.045
23	0.232	0.009	-0.021	-0.018	-0.017	**62**	1.648	1.466	1.464	1.479	1.688
24*	0.506	0.719	0.696	0.686	0.598	**63**	0.617	-0.004	0.014	0.099	0.093
25*	0.154	0.003	-0.029	0.012	-0.052	**64**	0.870	0.494	0.376	0.467	0.516
26	1.036	0.984	0.982	1.022	1.055	**65**	0.890	0.945	0.835	0.914	0.879
27	-0.056	0.076	0.032	0.091	0.076	**66**	0.614	0.409	0.404	0.413	0.427
28	-0.101	0.152	0.098	0.213	0.142	**67**	0.266	-0.006	0.052	-0.004	0.003
29*	0.231	0.625	0.595	0.543	0.608	**68**	1.320	0.967	0.932	1.110	1.070
30*	1.257	1.492	1.590	1.407	1.497	**69**	0.617	0.708	0.687	0.766	0.684
31	0.477	0.719	0.696	0.686	0.598	**70**	1.050	0.967	0.932	1.110	1.070
32	0.079	0.003	-0.029	0.012	-0.052	**71**	1.107	0.742	0.716	0.758	0.650
33	0.716	0.876	0.860	0.647	0.672	**72**	1.009	0.498	0.488	0.543	0.523
34	1.179	1.193	1.237	0.906	0.938	**73**	0.424	-0.006	0.052	-0.004	0.003
35	0.670	0.621	0.572	0.413	0.470	**74**	0.610	0.742	0.716	0.758	0.650
36	0.815	0.634	0.572	0.883	0.739	**75**	-0.030	-0.025	0.013	0.022	0.045
37	0.527	0.808	0.780	0.873	0.873	**76***	0.890	1.466	1.464	1.479	1.688
38	0.155	0.524	0.486	0.586	0.559	**77**	0.015	0.176	0.175	0.169	0.179
39	-0.020	0.531	0.512	0.254	0.250						

*Test set

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. From figure 4, five compounds are identified as outliers and one compound among five outliers is considered as outside for the descendant MLR model, which represents 6.49% of the total of studied compounds. Therefore, the predicted toxicity by the developed MLR model is reliable. The Williams plot for the stepwise MLR model is shown in figure 5.

From this plot, the leverage values (h_i) of any compound in the training and test sets are less than the critical value ($h^* = 0.136$) excepting the compounds 40 and 54. Also, the standardized residuals of all compounds in the training and test sets are less than three standard deviation units (±3σ). Therefore, the predicted toxicity by the developed stepwise MLR model is reliable.

3.7. Proposed novel compounds

Consequently, with MLR $_{descendant}$ and $MLR_{stepwise}$ approach, we can design new compounds with different and improved values of toxicity than the studied compounds. Taking into account the above results, we added suitable substitutions and then calculated the toxicities of the new compounds using the proposed model in equations 1 and 2. The leviers *h* of new compounds X_1, X_6, X_9 and X_{10} for the stepwise model and descendant model are defined as outliers, because they have a higher leverage which is greater than h^* (0.272 for descendant model and 0.136 for the stepwise model). We can suggest for the six remaining are regarded reliable compounds for design new compounds with different and improved values of toxicity than the studied compounds.

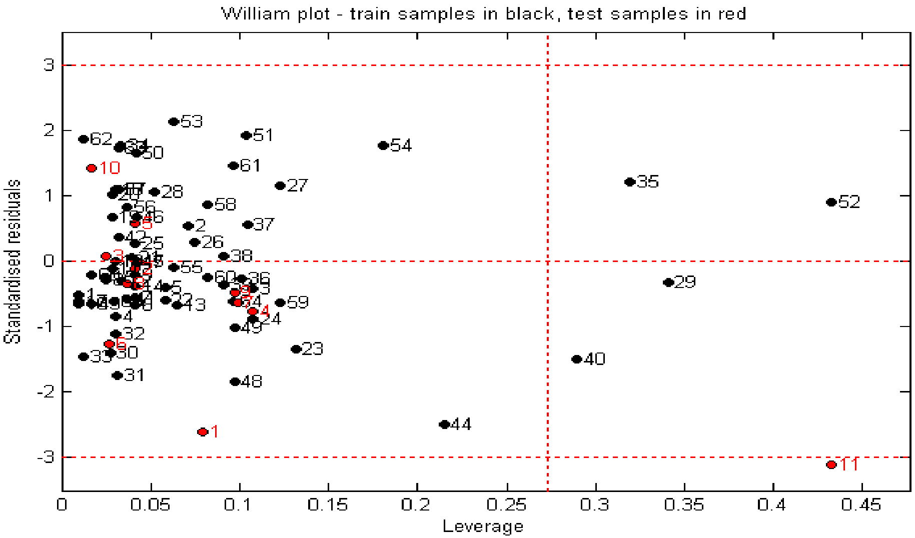


Figure 4: Williams plot for the descendant MLR model (with h^* = 0.272 and residual limits = ±3σ)

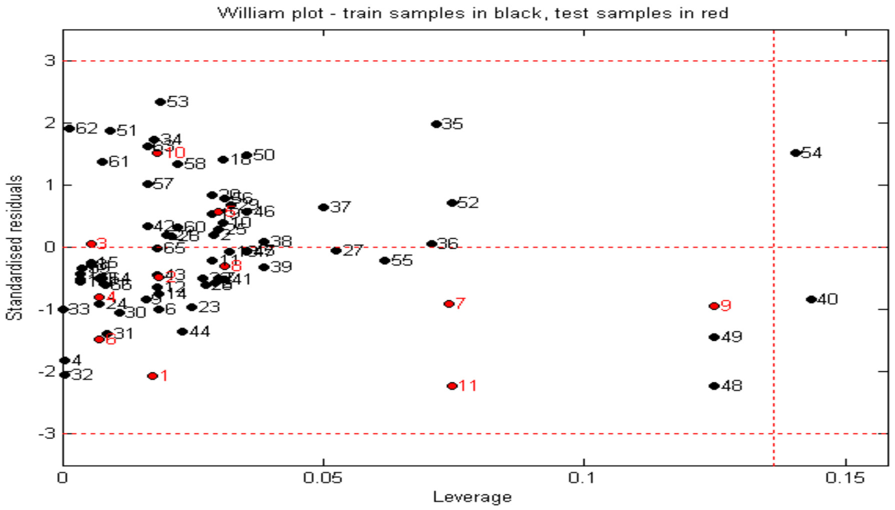
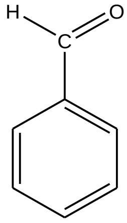


Figure 5: Williams plot for the stepwise MLR model (with h^* = 0.136 and residual limits = ±3σ)

Table 7: Proposed compounds, value of calculated descriptors, and predicted values of pIC_{50} using MLR $_{stepwise}$ model



	2	3	4	5	6	E_T	γ	MW	MR	n	Log *P*	pIC_{50} RLM (step)	*h*	pIC_{50} RLM (desc)	*h*
X_1	NH_2	NH_2	H	H	NH_2	-511.63	87.2	151.16	45.71	1.789	-1.096	-0.617	0.199	-0.373	0 .309
X_2	NH_2	H	NH_2	H	H	-456.28	68.5	136.15	41.47	1.714	-0.286	-0.490	0.087	-0.265	0 .147
X_3	NH_2	CH3	NH_2	CH3	CH3	-574.23	53	178.23	55.95	1.644	0.962	0.024	0.035	0.180	0.142
X_4	F	F	H	F	F	-742.47	32.1	178.08	32.98	1.471	2 .413	0.553	0.033	0.859	0.233
X_5	NO_2	H	NO_2	H	H	-754.55	71.8	196.11	46.09	1.66	1.263	0.688	0.036	0.866	0.064
X_6	OH	OH	OH	OH	H	-646.44	117.8	170.11	40.53	1.799	-0.213	0.743	0.233	0.519	0.127
X_7	CH_3	CH_3	H	CH_3	CH_3	-502.83	34.5	162.22	52.3	1.541	3.003	1.029	0.056	0.871	0.085
X_8	Cl	Cl	Cl	Cl	H	-2183.92	49.9	243.90	52.58	1.624	3 .819	1.977	0.083	2.508	0.175
X_9	H	C_6H_5	H	C_6H_5	H	-807.68	44.7	258.31	82.18	1.627	4.211	2.131	0.116	2.122	0.436
X_{10}	Br	Br	Br	Br	H	-10629.9	56.3	421.70	63.76	1.695	5.097	3.037	0.166	4.957	1.900

Conclusion

In this study two different modeling methods, multiple linear regression (MLR) and multiple non linear regression (MNLR), were used for predicting the toxicity of aromatic aldehydes to *Tetrahymena pyriformis*. The accuracy and predictability of the proposed models were proven by the comparison of key statistical terms of models. The good results obtained with the internal and external validations show that the proposed models in this paper are able to predict activities with a great performance and that the selected descriptors are pertinent. The applicability domains (AD) of the MLR models were defined.

The resulting models have shown that we have established a relationship between some descriptors and the activities in satisfactory manners. The MNLR results have substantially better predictive capability than the MLR results, but the latter gives the most important interpretable results.

The selected descriptors in the QSAR models can illustrate the contributing electronic and steric properties that are responsible for the toxicity of aromatic aldehydes to *Tetrahymena pyriformis*. By interpreting the molecular descriptors for the stepwise MLR model, we conclude that the increase octanol/water partition coefficient (log P) and γ as well are responsible for the greater activity of the studied compounds, presence of electronegative substituents (like O, N, F, Br, Cl), lipophilic substituents, e.g., chlorine. The aldehydic oxygen was also important for toxicity.

Finally, the accuracy and predictability of the proposed models were illustrated by comparing key statistical indicators such as shown in table 6, the models reported here may be used more conveniently than the previously reported models, with better confidence of prediction accuracy.

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

References

1. J.R. Seward, E.L.Hamblen, T.W. Schultz, Regression comparisons of *Tetrahymena pyriformis* and *Poecilia reticulate* toxicity, *Chemosphere* 47 (2002) 93–101.
2. J.W. Chen, W.J. Peijnenburg, X. Quan, *Chemsphere* 40 (2000) 1319-1326.
3. M.J. Zhu, F. Ge, R.L. Zhu, *Chemosphere* 80 (2010) 46-52.
4. T.I. Netzeva, T.W. Schultz, QSARs for the aquatic toxicity of aromatic aldehydes from *Tetrahymena* data, Chemosphere 61 (2005) 1632–1643.
5. (a) C. Adamo, V. Barone, *J. Chem. Phys. Lett.* 330 (2000) 152;
   (b) M. Parac, S. Grimme, *J. Phys. Chem.* 106 (2003) 6844;
   (c) Y. Yamaguchi, S. Yokoyama, S. Mashiko, *J. Chem. Phys.* 116 (2002) 6541.
6. (a) L . Becker, K. Hinrichs, U. Finke, A new algorithm for computing joins with grid files, in: Proc. of the 9th International Conference on Data Engineering, Vienna, Austria, (1993) 190–197;
   (b) S.J. Lee, J. Fink, A.B. Balantekin, M.R. Strayer, A.S. Umar, P.G. Reinhard, J.A. Maruhn, W. Greiner, *Phys. Rev. Lett.* 60 (1988) 163.
7. S. Chtita, M. Ghamali, M. Larif, A. Adad, R. Hmammouchi, M. Bouachrine, T. Lakhlifi, Prediction of biological activity of imidazo[1,2-a]pyrazine derivatives by combining DFT and QSAR results, *IJIRSET* 2 (12) (2013) 7962.
8. Advanced Chemistry Development Inc., Toronto, Canada (2009).
9. XLSTAT 2009 Add-in software (XLSTAT Company).www.xlstat.com.
10. M. Ghamali, S. Chtita, A. Ousaa, B. Elidrissi, M. Bouachrine, T. Lakhlifi, QSAR analysis of the toxicity of phenols and thiophenols using MLR and ANN, *Journal of Taibah University for Science* 11 (2017) 1-10.
11. A. Ousaa, B. Elidrissi, M. Ghamali, S. Chtita, M. Bouachrine, T. Lakhlifi, *Comp. J. Meth. Mol. Des.* 4(3) (2014) 10-18.
12. A. Golbraikh, A. Tropsha, *J. Mol. Graphics Model*. 20 (2002) 269–276.
13. L. Eriksson et al., Methods for reliability and uncertainty assessment and for applicability evaluations of classification and regression-based QSARs, *Environ. Health. Perspect.* 111 (2003) 1361-1375.

(2018) ; http://www.jmaterenvironsci.com