J. Mater. Environ. Sci. 7 (8) (2016) 3027-3034
ISSN : 2028-2508
CODEN: JMESC

Combining DFT and QSAR computation for predicting the soil sorption coefficients of substituted phenols and anilines

M. Ghamali [1*], S. Chtita [1], A. Adad [1], R. Hmamouchi [1], M. Bouachrine [2], T. Lakhlifi [1]

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

Received 22 May 2015, Revised 05 Apr 2016, Accepted 16 Apr 2016
**Corresponding author. E-mail: ghamalimounir86@gmail.com ; Tel. +212 670301669*

Abstract
The soil sorption coefficient (K_{oc}) is a key physicochemical parameter to assess the environmental risk of organic compounds. To predict the soil sorption coefficient in the more effective and economical way, here the QSAR model is applied to the set of 42 substituted phenols and anilines. This study was conducted using the principal component analysis (PCA), multiple linear regression (MLR), nonlinear regression (RNLM) and artificial neural network (ANN). We propose a quantitative model according to these analyses, and we interpreted the soil sorption coefficient of the compounds based on the multivariate statistical analysis. 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. This study shows that the MRA and MNLR have served to predict the soil sorption coefficient, but compared to the results of the ANN model, we conclude that the predictions fulfilled by the latter are more effective and better than other models.

Keywords: QSAR model, DFT study, substituted anilines and phenols, soil sorption coefficient, artificial neural network (ANN).

1. Introduction
The soil sorption coefficient K_{oc}, that determine the partitioning of an organic chemical between the soil/ sediment and aqueous solution, is an important environmental parameter. It is defined as the ratio of the concentration of a chemical adsorbed by the soil to the concentration of the chemical dissolved in the water, usually normalized to the organic carbon content. Thus, K_{oc} is a frequently used parameter to indicate the physical movements of pollutants, chemical degradation, and biodegradation activity of a given species in environment [1-5], and it is of great use for the environmental risk assessment of organic chemicals.
Halogenated aromatic compounds have been used for many years in chemical industry. They are used as solvents, propellants, additives, cooling agents and other polymers, for pesticides and organic syntheses [6]. Many of these chemicals were released into the environment and accumulated in nearly all environmental compartments, especially in aquatic systems, so it is beneficial to make a deep study of their potential hazard to aquatic organism. For these chemicals, quantitative structure–activity relationship (QSAR) modeling is a useful technique to correlate their physical, chemical, biological or environmental activities to their physicochemical property descriptors. With the rapid development of computer science and theoretical quantum chemical study, it can speedily and precisely obtain the quantum chemical parameters of compounds by computation. Moreover, these parameters, which have definite physical meaning, along with the introduction of the QSAR model can increase the interpretability. Nowadays, many QSAR models have been developed to predict the soil sorption coefficient of organic chemicals [7-9].
The objective of this study is to develop predictive QSAR models of the soil sorption coefficient (log K_{oc}) of substituted phenols and anilines using several statistical tools, such as principal components analysis (PCA),

multiple linear regression (MLR), nonlinear regression (RNLM) and artificial neural network (ANN) calculations. To test the performance and the stability of this model we have used the method validation.

2. Material and methods

2.1. Data sources

The observed soil sorption coefficients log K_{oc} for 42 substituted phenols and anilines was taken from a literature [10]. The table 1 shows the chemical compounds studied and their soil sorption coefficients. For the proper validation of our data set with a QSAR model, the 42 compounds data were divided into training and test sets. 33 molecules are considered as training set to build QSAR models while remaining 9 molecules is taken as test set. The division was performed by random selection.

Table 1: Substituted phenols and anilines their soil sorption coefficients

N°	Name (IUPAC)	log K_{oc}	N°	Name (IUPAC)	log K_{oc}
1	phenol	1.430	**22**	Catechol	2.030
2	2,3-dichlorophenol	2.650	**23**	Aniline	1.410
3	2,4-dichlorophenol	2.750	**24**	3-methylaniline	1.650
4	2,4,6-trichlorophenol	3.020	**25**	4-methylaniline	1.900
5	2,4,5-trichlorophenol	3.360	**26**	4-chloroaniline	1.960
6	3,4,5-trichlorophenol	3.560	**27**	4-bromoaniline	1.960
7	2,3,4,6-tetrachlorophenol	3.350	**28**	3-trifluoromethylaniline	2.360
8	Pentachlorophenol	3.730	**29***	3-chloro-4-methoxyaniline	1.930
9	4-bromophenol	2.410	**30**	3-methyl-4-bromoaniline	2.260
10	4-nitrophenol	2.370	**31**	2,4-dichloroaniline	2.720
11*	2-chlorophenol	2.600	**32**	2,6-dichloroaniline	3.250
12	3-chlorophenol	2.540	**33**	3,5-dichloroaniline	2.110
13	3,4-dichlorophenol	3.090	**34**	3,4-dichloroaniline	2.290
14*	3,5-dimethylphenol	2.830	**35**	2,3,4-trichloroaniline	2.600
15*	2,3,5-trimethylphenol	3.610	**36**	2,3,4,5-tetrachloroaniline	3.030
16	4-methylphenol	2.700	**37**	2,3,5,6-tetrachloroaniline	3.940
17*	2-methoxyphenol	1.560	**38***	Pentachloroaniline	4.620
18	3-methoxyphenol	1.500	**39**	3,5-dinitroaniline	2.550
19	3-hydroxyphenol	0.980	**40***	N-methylaniline	2.280
20	4,5,6-trichloroguaiacol	2.800	**41***	N,N-dimethylaniline	2.260
21	Tetrachloroguaiacol	2.850	**42***	Diphenylamine	2.780

* Test set

2.2. Molecular descriptors

Currently, there are a large number of molecular descriptors used in QSAR studies. After validation, the findings can be used to predict the activity of untested compounds.

The computation of electronic descriptors was performed using the Gaussian03W package [11]. The geometries of the 42 substituted phenols and anilines were optimized with DFT method with the B3LYP functional and 6-31G (d) base set. Then, several related structural parameters were selected from the results of quantum computation as follows: highest occupied molecular orbital energy (E_{HOMO}), lowest unoccupied molecular orbital energy (E_{LUMO}), dipole moment (μ), total energy (E_T), absolute hardness (η), absolute electronegativity (χ) and reactivity index (ω) [12].

The η, χ and ω were determined using the following equations:

$$\eta = (E_{LUMO} - E_{HOMO})/2 \qquad \chi = (E_{LUMO} + E_{HOMO})/2 \qquad \omega = \chi^2/2\eta$$

ACD/ChemSketch program [13] was used to calculate the topological descriptors, as follows: molar volume (MV), molecular weight (MW), molar refractivity (MR), parachor (Pc), density (D), refractive index (n) and surface tension (γ).Thus, in order to improve the estimate quality for these compounds, molecular descriptor which reflect other specific interactions should be also included as octanol/water partition coefficient (log *P*). The hydrophobic parameter, octanol/water partition coefficient is commonly used to predict the soil sorption coefficient [14, 15].

2.3. Statistical analysis

The objective of quantitative structure-activity relationship (QSAR) analysis is to derive empirical models that relate the biological activity of compounds to their chemical structure. In this QSAR analysis, quantitative descriptors are used to describe the chemical structure and analysis results in a mathematical model describing the relationship between the chemical structure and soil sorption coefficient log K_{oc}. To explain the structure-activity relationship, these 15 descriptors are calculated for the 42 molecules using the Gaussian03W and ChemSketch programs.

The quantitative descriptors of the substituted phenols and anilines are studied using statistical methods based on the principal component analysis (PCA) [16] with the software XLSTAT version 2013 [17]. PCA is a useful statistical technique for summarizing all of the information encoded in the structures of the compounds. It is also helpful for understanding the distribution of the compounds [18]. This is an essentially descriptive statistical method that aims to present, in graphic form, the maximum information contained in the data, as shown in table 2.

Multiple linear regression (MLR) analysis with descendent selection and elimination of variables was used to model the structure-activity relationships. It is a mathematical technique that minimizes the difference between actual and predicted values. Additionally, it selects the descriptors used as the input parameters in the multiple nonlinear regression (MNLR) and artificial neural network (ANN). MLR and MNLR are generated using the software XLSTAT version 2013. To predict the log K_{oc}, equations are justified by the coefficient of determination (R^2), the mean squared error (MSE), Fisher's criterion (F) and the significance level (P).

The ANN analysis is performed using the Matlab software version 2009a Neural Fitting tool (nftool) toolbox on a data set of our compounds [19]. A number of individual models of ANN were designed, built and trained. Three components constitute 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. Although there many different ANN models, the most frequently used type of ANN in QSAR is the three-layered feed forward network [20]. In this type of network, the neurons are arranged in layers as an input layer, one hidden layer and an output layer. Each neuron in any layer is fully connected with the neurons of a succeeding layer and no connections are between neurons belonging to the same layer.

According to the supervised learning adopted here, the networks are taught by providing them examples of input patterns and the corresponding target outputs. Through an iterative process, the connection weights are modified until the network gives the desired results for the training set of data. A backpropagation algorithm is used to minimize the error function. This algorithm was described previously with a simple example of an application [21] and the details of this algorithm are provided elsewhere [22].

Testing the stability, predictive power and generalization ability of the models is a very important step in a QSAR study. For the validation of the predictive power of a QSAR model, two basic principles, internal validation and external validation, are available. Cross-validation is one of the most popular methods for internal validation. In this study, the internal predictive capability of the model was evaluated using leave-one-out cross-validation (R^2_{cv}). A good R^2_{cv} often indicates a good robustness and high internal predictive power of a QSAR model. However, recent studies [23] indicate that there is no evident correlation between the value of R^2_{cv} and actual predictive power of a QSAR model, suggesting that the R^2_{cv} remains inadequate as a reliable estimate of the model's predictive power for all new chemicals. To determine both the generalizability of QSAR models for new chemicals and the true predictive power of the models, statistical external validation is applied during the model development step by properly employing a prediction set for validation.

3. Results and discussion

3.1. Data set for analysis

A QSAR study was performed for a series of 42 substituted phenols and anilines, as reported previously [10], to determine a quantitative relationship between the structure and soil sorption coefficient log K_{oc}. The values of the 15 chemical descriptors are shown in table 2.

3.2. Principal component analysis

The total of the 15 descriptors coding the 42 molecules was submitted to principal components analysis (PCA) [24]. The first three principal axes are sufficient to describe the information provided by the data matrix. Indeed, the percentages of the variance are 53.90%; 20.09% and 10.09% for the axes F1, F2 and F3, respectively. The total information is estimated as 84.08%.

Table 2: Values of the obtained parameters of the studied substituted phenols and anilines

N	log K_{oc}	E_T	E_{HOMO}	E_{LUMO}	μ	ω	η	χ	MW	MR	MV	Pc	n	γ	D	log *P*
1	1.430	-8372.091	-6.481	0.003	1.598	1.618	3.242	-3.239	94.111	28.130	87.800	222.200	1.553	40.900	1.071	1.475
2	2.650	-33401.772	-6.491	-0.668	1.328	2.201	2.911	-3.580	163.001	37.920	111.700	294.000	1.593	47.800	1.458	2.852
3	2.750	-33401.857	-6.357	-0.756	1.070	2.258	2.800	-3.556	163.001	37.920	111.700	294.000	1.593	47.800	1.458	2.972
4	3.020	-45916.494	-6.573	-1.054	1.421	2.635	2.759	-3.813	197.446	42.810	123.700	329.800	1.608	50.500	1.596	3.391
5	3.360	-45916.480	-6.559	-1.031	2.045	2.605	2.764	-3.795	197.446	42.815	123.700	329.800	1.608	50.500	1.596	3.601
6	3.560	-45916.139	-6.946	-1.007	2.984	2.662	2.969	-3.976	197.446	42.815	123.700	329.800	1.608	50.500	1.596	3.811
7	3.350	-58430.997	-6.681	-1.260	1.002	2.908	2.710	-3.971	231.891	47.710	135.600	365.700	1.620	52.800	1.709	3.997
8	3.730	-70945.451	-6.820	-1.446	1.906	3.179	2.687	-4.133	266.337	52.605	147.600	401.600	1.631	54.700	1.804	4.714
9	2.410	-78383.265	-6.436	-0.428	2.136	1.960	3.004	-3.432	173.007	35.827	104.000	272.700	1.604	47.200	1.662	2.635
10	2.370	-13940.916	-6.925	-2.223	5.341	4.449	2.351	-4.574	139.109	34.666	99.700	277.700	1.612	60.200	1.395	-0.285
11•	2.600	-20887.116	-6.255	-0.356	0.933	1.852	2.950	-3.305	128.556	33.026	99.800	258.100	1.575	44.700	1.287	2.155
12	2.540	-20887.070	-6.290	-0.371	1.109	1.874	2.960	-3.331	128.556	33.026	99.800	258.100	1.575	44.700	1.287	2.485
13	3.090	-3339.788	-6.336	-0.705	2.606	2.201	2.816	-3.521	163.001	37.921	111.700	294.000	1.593	47.800	1.458	3.182
14•	2.830	-10513.531	-5.786	0.123	1.348	1.357	2.954	-2.832	122.164	37.769	120.400	297.500	1.540	37.200	1.014	2.473
15•	3.610	-11584.085	-5.665	0.203	1.635	1.271	2.934	-2.731	136.191	42.613	136.600	335.200	1.535	36.100	0.996	2.872
16	2.700	-9442.881	-5.746	0.070	1.334	1.385	2.908	-2.838	108.138	32.950	104.100	259.900	1.545	38.800	1.038	1.974
17•	1.560	-11490.758	-5.533	0.316	2.734	1.164	2.925	-2.609	124.137	34.817	111.800	278.900	1.534	38.600	1.109	1.324
18	1.500	-11490.731	-5.735	0.221	1.743	1.276	2.978	-2.757	124.137	34.817	111.800	278.900	1.534	38.600	1.109	1.574
19	0.980	-10420.421	-5.781	0.197	1.356	1.304	2.989	-2.792	110.111	29.998	86.200	237.300	1.612	57.100	1.275	0.808
20	2.800	-49034.733	-6.124	-0.625	5.404	2.071	2.749	-3.375	227.472	49.501	147.700	386.500	1.584	46.800	1.539	3.470
21	2.850	-61549.174	-6.563	-1.139	3.698	2.734	2.712	-3.851	261.917	54.396	159.600	422.400	1.596	48.900	1.640	3.952
22	2.030	-10420.421	-5.627	0.220	2.518	1.250	2.923	-2.704	110.111	29.998	86.200	237.300	1.612	57.100	1.275	0.878
23	1.410	-7831.149	-6.546	0.035	1.354	1.611	3.290	-3.256	93.126	30.478	91.700	233.100	1.579	41.700	1.015	0.915
24	1.650	-8901.774	-6.359	0.038	1.644	1.561	3.199	-3.160	107.153	35.297	107.900	270.700	1.567	39.500	0.992	1.414
25	1.900	-8902.005	-5.233	0.268	1.512	1.120	2.751	-2.483	107.153	35.297	107.900	270.700	1.567	39.500	0.992	1.414
26	1.960	-20345.952	-6.573	-0.379	2.402	1.951	3.097	-3.476	127.572	35.372	103.600	268.900	1.598	45.300	1.230	1.908
27	1.960	-77842.335	-6.470	-0.387	2.315	1.933	3.041	-3.429	172.023	38.173	107.800	283.600	1.625	47.700	1.594	2.058
28	2.360	-17008.960	-5.846	-0.513	3.759	1.895	2.666	-3.179	161.124	35.473	125.200	290.300	1.478	28.800	1.286	2.288
29•	1.930	-23464.313	-5.982	-0.256	1.546	1.699	2.863	-3.119	157.598	42.058	127.600	325.600	1.572	42.300	1.234	1.699
30	2.260	-78913.253	-5.472	-0.041	3.253	1.399	2.716	-2.756	186.049	42.992	124.100	321.200	1.609	44.700	1.498	2.557
31	2.720	-32860.748	-6.823	-0.703	1.106	2.313	3.060	-3.763	162.017	40.267	115.600	304.800	1.613	48.300	1.401	2.719
32	3.250	-32860.660	-6.805	-0.663	1.918	2.270	3.071	-3.734	162.017	40.267	115.600	304.800	1.613	48.300	1.401	2.719
33	2.110	-32860.717	-6.891	-0.739	2.356	2.366	3.076	-3.815	162.017	40.267	115.600	304.800	1.613	48.300	1.401	2.719
34	2.290	-32860.614	-6.747	-0.693	2.535	2.286	3.027	-3.720	162.017	40.267	115.600	304.800	1.613	48.300	1.401	2.599
35	2.600	-45375.259	-6.968	-0.936	1.975	2.589	3.016	-3.952	196.462	45.162	127.500	340.700	1.626	50.800	1.540	3.226
36	3.030	-57889.853	-7.019	-1.185	2.138	2.884	2.917	-4.102	230.907	50.056	139.500	376.500	1.636	53.000	1.655	3.831
37	3.940	-57889.877	-6.920	-1.214	1.359	2.899	2.853	-4.067	230.907	50.056	139.500	376.500	1.636	53.000	1.655	3.951
38•	4.620	-70404.319	-7.061	-1.402	1.946	3.164	2.830	-4.232	265.352	54.951	151.400	412.400	1.645	54.900	1.751	4.549
39	2.550	-18967.534	-7.735	-2.489	4.279	4.981	2.623	-5.112	183.122	43.572	115.300	344.000	1.679	79.000	1.586	-2.605
40•	2.280	-8901.749	-5.178	0.311	1.774	1.079	2.744	-2.433	107.153	35.852	108.800	266.100	1.572	35.600	0.984	1.641
41•	2.260	-9972.040	-5.019	0.366	1.857	1.005	2.693	-2.327	121.180	40.570	127.400	309.300	1.549	34.700	0.950	2.307
42•	2.780	-14122.889	-5.086	-0.120	0.872	1.365	2.483	-2.603	169.222	55.632	155.400	400.600	1.634	44.000	1.088	3.620

• Test set

The principal component analysis (PCA) [25] was conducted to identify the link between the different variables. Correlations between the fourteen descriptors are shown in table 3 as a correlation matrix.
The obtained matrix provides information on the high or low interrelationship between variables. In general, the co-linearity (r>0.5) was observed between most of the variables, and between the variables and log K_{oc}. A high interrelationship was observed between MR and Pc (r = 0.991), and a low interrelationship was observed between log K_{oc} and μ (r = -0.013). Additionally, to decrease the redundancy existing in our data matrix, the descriptors that are highly correlated ($R \geq 0.9$), were excluded.

3.3. Multiple linear regressions (MLR)

Many attempts have been made to develop a relationship with the indicator variable of soil sorption coefficient, log K_{oc}, but the best relationship obtained using this method is only one corresponding to the linear combination of several descriptors selected, the energy E_{LUMO}, the octanol/water partition coefficient (log *P*).
The resulting equation is as follows:

$$\log K_{oc} = 1.365 - 0.645 \times E_{LUMO} + 0.305 \times \log P \quad (1)$$

N = 33 R^2 = 0.763 R^2_{cv} = 0.660 MSE = 0.125 F = 48.176 P < 0.0001

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

Table 3: Correlation matrix between different obtained descriptors

	log K_{oc}	E_T	E_{HOMO}	E_{LUMO}	μ	ω	η	χ	MW	MR	MV	Pc	n	γ	D
E_T	-0.493	**1**													
E_{HOMO}	-0.376	0.432	**1**												
E_{LUMO}	-0.542	0.459	0.830	**1**											
μ	**-0.013**	-0.135	-0.212	-0.424	**1**										
ω	0.471	-0.391	-0.845	-0.989	0.441	**1**									
η	-0.344	0.106	-0.167	0.412	-0.402	-0.367	**1**								
χ	-0.483	0.466	0.953	0.960	-0.337	-0.962	0.140	**1**							
MW	0.729	-0.814	-0.520	-0.691	0.251	0.610	-0.372	-0.637	**1**						
MR	0.710	-0.609	-0.273	-0.509	0.135	0.432	-0.453	-0.414	0.879	**1**					
MV	0.675	-0.505	-0.100	-0.357	0.145	0.272	-0.466	-0.244	0.802	0.954	**1**				
Pc	0.718	-0.588	-0.289	-0.541	0.197	0.467	-0.484	-0.439	0.894	**0.991**	0.964	**1**			
n	0.363	-0.532	-0.613	-0.655	0.078	0.649	-0.155	-0.663	0.558	0.503	0.224	0.451	**1**		
γ	0.258	-0.333	-0.687	-0.769	0.301	0.799	-0.236	-0.763	0.452	0.292	0.043	0.302	0.857	**1**	
D	0.578	-0.840	-0.698	-0.775	0.273	0.714	-0.229	-0.772	0.876	0.583	0.427	0.594	0.686	0.660	**1**
log *P*	0.662	-0.573	-0.075	-0.090	-0.263	-0.031	-0.036	-0.086	0.649	0.630	0.662	0.606	0.111	-0.191	0.451

A higher correlation coefficient and lower mean squared error indicates that the model is more reliable. A **P** that is smaller than 0.05 indicates that the regression equation has statistically significant. The QSAR model expressed by Eq. (1) is cross validated by its noticeable R^2_{cv} value (R^2_{cv} = 0.660) obtained by the leave-one-out (LOO) method. A value of R^2_{cv} is greater than 0.5 is the essential condition for qualifying a QSAR model as valid [23]. The correlation coefficients between variables 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. Models with a VIF greater than 5 were unstable and were eliminated, models with a VIF values between 1and 4 means the models can be accepted. As can be seen from table 4, the VIF values of the two descriptors are all smaller than 5.0, indicating that there is no collinearity among the selected descriptors and the resulting model has good stability.

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

Statistic	E_{LUMO}	log *P*
Tolerance	1.000	1.000
VIF	1.000	1.000

The elaborated QSAR model reveals that the soil sorption coefficient may be explained by a number of molecular descriptors. The negative correlation of the energy E_{LUMO} with the log k_{oc} shows that an increase in the value of this factor indicate a decrease in the value of the log k_{oc}, whereas a positive correlation of the octanol/water partition coefficient (log *P*) with the log k_{oc} reveals that an increase in the value of log K_{oc}.
The correlations of the predicted and observed log K_{oc} and the residual graph of the absolute numbers are illustrated in figure 1 (a) and (b) respectively.
The descriptors proposed in Eq. (1) by MLR are, therefore, used as the input parameters in the multiple nonlinear regressions (MNLR) and artificial neural network (ANN).

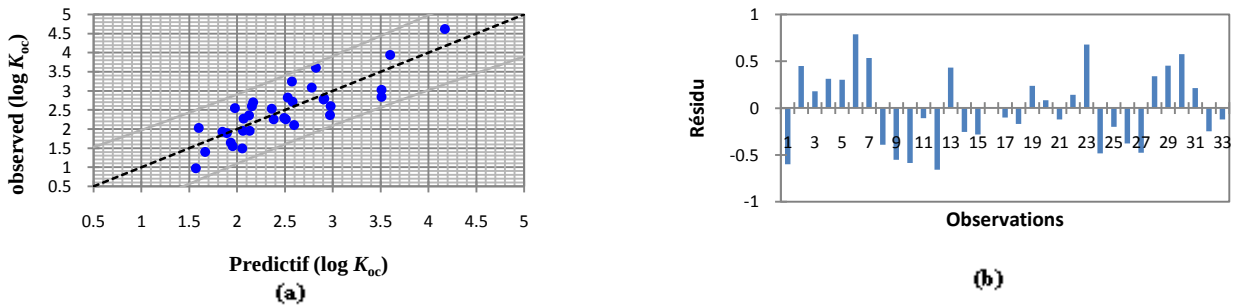

Figure 1: (a) Graphical representation of calculated and observed log K_{oc}; (b) The residual graph calculated by MLR

3.4. Multiple nonlinear regressions (MNLR)

We also used the nonlinear regression model to improve the soil sorption coefficient in a quantitative manner, taking into account several parameters. This is the most common tool for the study of multidimensional data. We applied this to the data matrix constituted from the descriptors proposed by the MLR corresponding to the 33 compounds training set.

The resulting equation is as follows:

$$\log K_{oc} = 1.405 - 0.273 \times E_{LUMO} + 0.208 \times \log P + 9.323\ 10^{-2} \times E^2_{LUMO} + 4.601\ 10^{-2} \times (\log P)^2 \qquad (2)$$

$$N = 33 \qquad R^2 = 0.786 \qquad R^2_{cv} = 0.707 \qquad MSE = 0.121$$

The QSAR model expressed by Eq. (2) is cross validated by its appreciable R^2_{cv} values (R^2cv =0.707) obtained by the leave-one-out (LOO) method. A value of R^2_{cv} greater than 0.5 is the basic requirement for qualifying a QSAR model as valid [23].

The correlations of the predicted and observed log K_{oc} and the residual graph of the absolute numbers are illustrated in figure 2 (a) and (b) respectively.

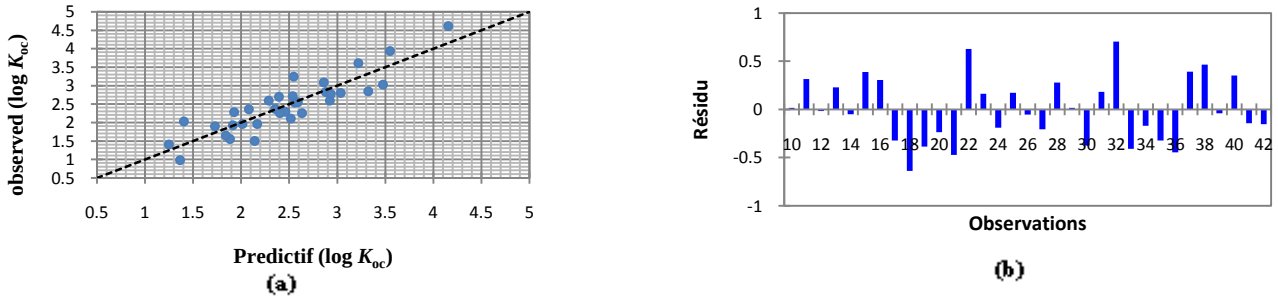

Figure 2: (a) Graphical representation of calculated and observed log K_{oc}; (b) The residual graph calculated by MNLR

3.5. Artificial neural networks (ANN)

Neural networks (ANN) can generate predictive models of the quantitative structure–activity relationships (QSAR) between a set of molecular descriptors obtained from the MLR and observed soil sorption coefficients. The ANN calculated log K_{oc} model was developed using the properties of several studied compounds. The correlation of the predicted and observed log K_{oc} and the residual graph of the absolute numbers are illustrated in figure 3 (a) and (b) respectively.

$$N = 33 \qquad R^2 = 0.868 \qquad R^2_{cv} = 0.718 \qquad MSE = 0.086$$

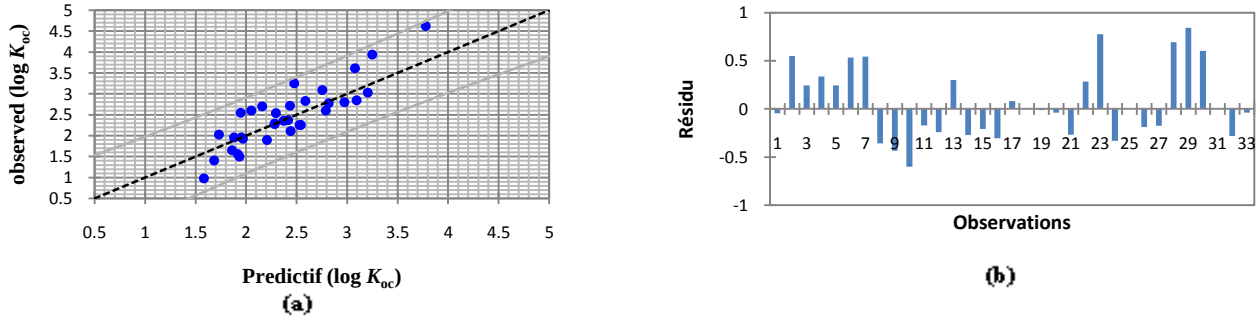

Figure 3: (a) Graphical representation of calculated and observed log K_{oc}; (b) The residual graph calculated by ANN

The obtained coefficient of determination (R^2) value is 0.868 for this data set of substituted phenols and anilines. This confirms that the artificial neural network (ANN) results were the best to build the quantitative structure-activity relationship model. Furthermore, the high R^2_{cv} value (R^2_{cv} = 0.718) observed also supports the suitability of the QSAR model for soil sorption coefficients of chemicals.

3.6. External validation

To estimate the predictive power of MLR, MNLR and ANN models, we must use a set of compounds that have not been used as the training set to establish the QSAR model. The models established in the computation procedure using the 33 substituted phenols and anilines are used to predict the soil sorption coefficients of the remaining 9 compounds. The main performance parameters of the three models are

shown in table 5. As seen from this table, the statistical parameters of ANN model are better than the others.

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	0.763	0.660	0.125	0.763	0.732	0.510
MNLR	0.786	0.707	0.121	0.786	0.785	0.457
ANN	0.868	0.718	0.086	0.868	0.748	0.232

We assessed the best linear QSAR regression equations established in this study. Based on this result, a comparison of the quality of the MLR and MNLR models shows that the ANN model has a significantly better predictive capability because the ANN approach yields better results than those of MLR and MNLR. ANN establishes a satisfactory relationship between the molecular descriptors and the soil sorption coefficients of the studied compounds.

The accuracy and predictability of the proposed models were illustrated by the comparing key statistical indicators, such as 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 log K_{oc} according to different methods

N°	log K_{oc} (obs.)	log K_{oc} (calc.)			N°	log K_{oc} (obs.)	log K_{oc} (calc.)		
		MLR	NMLR	ANN			MLR	NMLR	ANN
1	1.430	1.813	1.811	1.729	**26**	1.960	2.191	2.086	1.936
2	2.650	2.665	2.596	2.854	**27**	1.960	2.242	2.147	2.178
3	2.750	2.758	2.688	2.911	**28**	2.360	2.393	2.286	1.868
4	3.020	3.078	3.030	2.776	**30**	2.260	2.171	2.249	2.564
5	3.360	3.127	3.131	3.362	**31**	2.720	2.647	2.548	2.591
6	3.560	3.176	3.235	3.500	**32**	3.250	2.621	2.532	2.693
7	3.350	3.396	3.463	3.421	**33**	2.110	2.670	2.563	2.460
8	3.730	3.734	3.997	4.065	**34**	2.290	2.604	2.490	2.354
9	2.410	2.444	2.406	2.524	**35**	2.600	2.952	2.891	2.896
10	2.370	2.710	2.416	2.431	**36**	3.030	3.297	3.331	3.339
12	2.540	2.361	2.320	2.429	**37**	3.940	3.352	3.414	3.501
13	3.090	2.789	2.771	2.997	**39**	2.550	2.175	2.431	2.572
16	2.700	1.922	1.976	2.120	**11***	2.280	2.251	1.794	2.310
18	1.500	1.702	1.790	1.898	**14***	2.260	2.040	2.042	2.767
19	0.980	1.484	1.553	1.250	**15***	2.780	2.110	2.795	3.087
20	2.800	2.826	2.887	2.753	**17***	2.600	1.565	2.175	1.747
21	2.850	3.304	3.377	3.652	**29***	2.830	2.048	2.168	2.291
22	2.030	1.491	1.567	1.280	**38***	3.610	3.655	2.330	3.982
23	1.410	1.621	1.624	1.230	**40***	1.560	1.665	1.684	2.218
24	1.650	1.771	1.781	1.674	**41***	1.930	1.833	1.967	2.983
25	1.900	1.623	1.724	1.772	**42***	4.620	2.546	3.868	3.118

*Test set

Conclusion

In this study, we investigated the QSAR regression to predict the soil sorption coefficient of phenols and anilines.

The studies regarding the quality of the three models constructed in the study have good stabilities and great predictive powers. Moreover, compared to the MLR, MNLR models, the ANN model is better and is an effective tool to predict the soil sorption coefficient of phenols and anilines. Furthermore, using the ANN

approach, we established a relationship between several descriptors and log K_{oc} values of several organic compounds based on the substituted phenols and anilines in a satisfactory manner.
Finally, we conclude that studied descriptors, which are sufficiently rich in chemical, electronic and topological information to encode the structural features may 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. Moss G. P., Dearden J. C., Patel H., Cronin M. T. D., Toxicol. In Vitro 16 (3) (2002) 299–317.
2. Hodson J., Williams N. A., Chemosphere 17 (1988) 66–77.
3. Meylan W., Howard P. H., Boething R. S., Environ. Sci. Technol. 26 (1992) 1560–1567.
4. Muller M., Kordel W., Chemosphere 32 (12) (1996) 2493–2504.
5. Kortvelyesi T., Gorgenyi M., Heberger K., Anal. Chim. Acta 428 (2001) 1773–1782.
6. Sixt S., Altschuh J., Bruggemann R., Chemosphere 30 (1995) 2397–2414.
7. Abdul A. S., Gibon T. L., Rai D. N., Hazardous Waste Hazardous Mater. 4 (1987) 211–222.
8. Bakul H. R., Shyam R. A., Water Res. 35 (14) (2001) 3391–3401.
9. Sabljic A., Protic M., Bull. Environ. Contam. Toxicol. 28 (1982) 162–165.
10. Sabljic A., Gusten H., Verhaar H., Hermens J., Chemosphere 31 (1995) 4489–4514.
11. Frisch M. J. and al., Gaussian 03, Revision B.01, Gaussian, Inc., Pittsburgh, PA, 2003.
12. Sakar U., Parthasarathi R., Subramanian V., Chattaraji P. K., J. Mol. Des. IECMD, (2004) 1-24.
13. Advanced Chemistry Development Inc., Toronto, Canada, 2009. (www.acdlabs.com/resources/freeware/chemsketch/).
14. Liu G., Yu J., Water Res. 39 (2005) 2048–2055.
15. Wen Y., Su L. M., Qin W. C., Fu L., He J., Zhao Y. H., Chemosphere 86 (2012) 634–640.
16. Larif M., Adad A., Hmamouchi R., Taghki A. I., Soulaymani A., Elmidaoui A., Bouachrine M., Lakhlifi T., Arabian Journal of Chemistry, in press, 2013.
17. XLSTAT 2013 software (XLSTAT Company). http://www.xlstat.com.
18. Chtita S., Larif M., Ghamali M., Bouachrine M., Lakhlifi T., Journal of taibah university for chemistry 9(2) (2015) 143-154.
19. Adad A., Larif M., Hmamouchi R., Bouachrine M., Lakhlifi T., J. Comp. Meth. Mol. Des., 4(3) (2014) 72-83.
20. Hmamouchi R., Larif M., Adad A., Bouachrine M., Lakhlifi T., J. Comp. Meth. Mol. Des., 4(3) (2014) 61-71.
21. Cherqaoui D., Villemin D., J. Chem. Soc. Faraday. Trans. 90 (1994) 97-102.
22. Freeman J. A., Skapura D. M., Addition Wesley Publishing Company, Reading, 1991.
23. Golbraikh A., Tropsha A., J. Mol. Graphics Model. 20 (2002) 269–276.
24. STATITCF Software, Technical Institute of cereals and fodder, Paris, France, 1987.
25. Ousaa A., Elidrissi B., Ghamali M., Chtita S., Bouachrine M., Lakhlifi T., *J. Comp. Meth. Mol. Des.*, 4(3) (2014) 10-18.