Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2013, 5(7):28-41

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Atmospheric half-lives of persistent organic pollutants (POPs) study combining DFT and QSPR results

Azeddine Adad[a], Rachid Hmammouchi[a], Abdelhafid Idrissi Taghki[a], Abdelaziz Abdellaoui[b], Mohammed Bouachrine[c] and Tahar Lakhlifi[a*]

[a]Molecular Chemistry and Natural Substances Laboratory, Faculty of Sciences, University Moulay Ismail, Meknes, Morocco
[b]Laboratory of Chemical Biology Applied to the Environment, Faculty of Sciences, University Moulay Ismail, Meknes, Morocco
[c]ESTM, University Moulay Ismail, Meknes, Morocco

ABSTRACT

Quantitative Structure-Property Relationship (QSPR) study was applied to the prediction of the characteristic of Persistent Organic Pollutants (POPs) screening for atmosphere persistence. The mean and maximum half-life estimations for degradation in air of 45 Nations Environment Program (UNEP) POPs and possible POPs were modeled using the parameters from quantum-chemical calculations at density functional theory (DFT) level. It was expected that the main contribution to the degradation rate was given by the E_{HOMO}. Parameter Principal Component Analysis (PCA) method, the Multiple Linear Regression method (MLR), Partial Least Square analysis (PLS) and the Artificial Neural Network (ANN), showed a determination coefficient (R^2) more than 0,9. The prediction results were in excellent agreement with the experimental value.

Keywords: half-life; PCA; MLR; PLS; ANN; DFT

INTRODUCTION

Persistent Organic Pollutants (POPs), and their transformation products, are the most investigated organic environmental contaminants within the past five decades. Organochlorines have been found in virtually all environment compartments on the globe. Under certain meteorological and geographic conditions, high altitude environments can serve as “cold condensers” for atmospheric POP loadings. In addition to this several persistent organic pollutants are suspected to contribute to the increasing prevalence and risk of type 2 diabetes. The 2nd and 3rd tertiles of adipose tissue concentrations of some POPs were positively associated with the risks of diabetes. Also the risk of diabetes increased with tertiles of exposure in a linear manner in non-obese subject but not in the obese. In whom an inverted U-shape pattern was observed. Congener-specific blubber/milk partition coefficients indicated that lower-halogenated POPs were selectively offloaded into milk and changes in adult female dolphin contaminant profiles likely resulted from the offloading of POPs during the first reproductive event and their gradual re-accumulation thereafter.

Quantitative structure-activity relationship (QSAR) as an important area of chemo metrics has been the subject of a series of investigations [1,2]. The main aim of QSAR studies is to establish an empirical rule or function relating the structural descriptors of compounds under investigation to bioactivities. This rule or function is then utilized to predict the same bioactivities of the compounds not involved in the training set from their structural descriptors. Whether the bioactivities can be predicted with satisfactory accuracy depends to a great extent on the performance

of the applied multivariate data analysis method, provided the property being predicted is related to the descriptors. Many multivariate data analysis methods such as principal components analysis (PCA), multiple linear regression (MLR) and artificial neural net-work (ANN) have been used in QSAR studies. MLR, as almost commonly used chemo metric method, has been extensively applied to QSAR investigations. However, the practical usefulness of MLR in QSAR studies is rather limited, as it provides relatively poor accuracy. ANN offers satisfactory accuracy in most cases but tends to overfit the training data.

QSAR [1,2] has been widely used for years to provide quantitative analysis of structure and biological activity relationships of compounds. Different QSAR studies were reported to identify important structural features responsible for the antiamoebic activity [3-5] and to develop air half-lives persistence models for diverse chemicals by different workers [6-10].

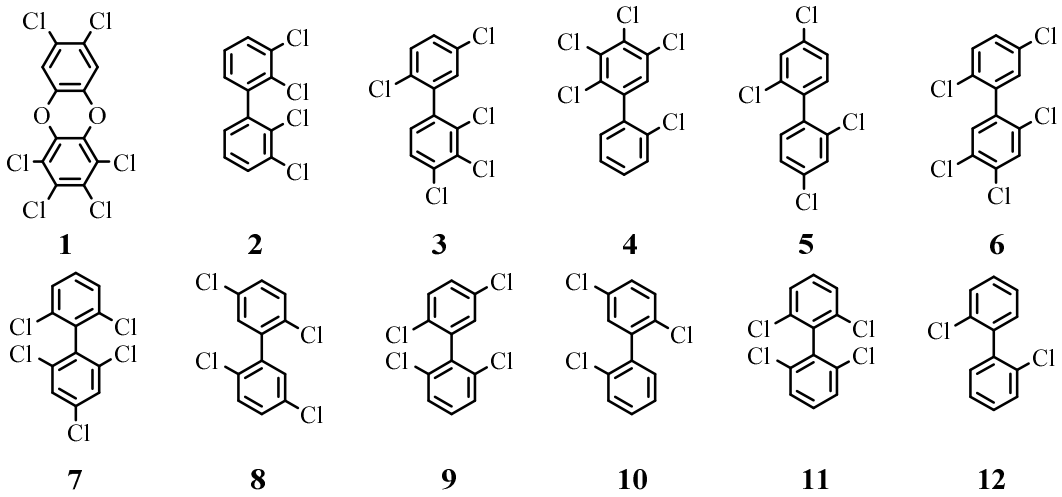
At present, there are a large number of molecular descriptors that can be used in QSAR studies. Once validated, the findings can be used to predict activities of untested compounds. Recently, computer-assisted drug design based on QSAR has been successfully employed to develop new drugs for the treatment of cancer, AIDS, SARS, and other diseases.

In this study, we have modeled air half-lives persistence of several organic compounds based on POPs (fig.1) using several statistical tools, principal components analysis (PCA), multiple linear regression (MLR) and artificial neural network (ANN) calculations. The objectives of this work are to develop predictive QSPR models for the air half-life persistence of our studied molecules. On the other hand, several quantum chemical methods and quantum-chemistry calculations have been performed in order to study the molecular structure and electronic properties [11-14]. The geometry as well as the nature of their molecular orbital, HOMO (highest occupied molecular orbital) and LUMO (lowest unoccupied molecular orbital) is involved in the properties of biological activity of organic compounds. The more relevant molecular properties were calculated. These properties are the highest occupied molecular orbital energy E_{HOMO}, the lowest unoccupied molecular orbital energy E_{LUMO}, energy gap ΔE, dipole moment μ, total energy E_T, activation energy E_a, absorption maximum λ_{max} and factor of oscillation $f_{(SO)}$.

EXPERIMENTAL SECTION

Antecedent studies [15,16] had established a quantitative model of molecular-structure air persistance and long-range potential for screening POPs. Further works on the same subject were produced by Lu Yuying [17].

The property under investigation is the screening for atmosphere persistence of 45 United Nations Environment Program (UNEP) POPs. The following figure shows the chemicals structures of studied compound.



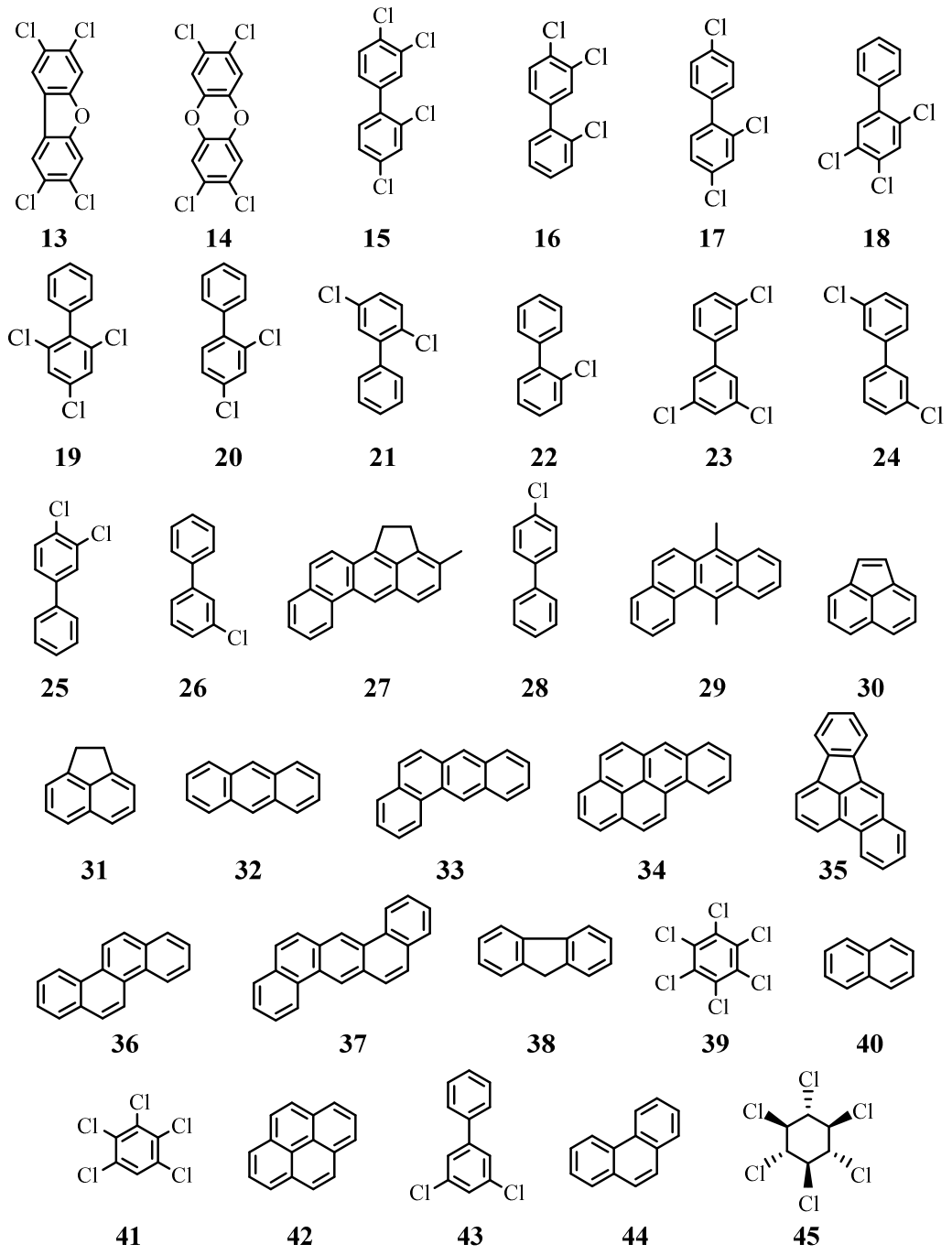


Fig.1. Chemical structures of the studied POPs.

The experimental air half-lives of the studied compounds have been collected from recent work [17] (Table1). The range of the air half-life persistence data varies from - 0,14 to 4,57 (Log units).

Table 1. Observed air mean and maximum half-lives of the studied POPs [17]

N°	IUPAC Name	Log air mean half-life value (h)	Log air maximum half-life value (h)
1	1,2,3,4,7,8-hexachlorodibenzo[b,e]dioxine	1,770	1,910
2	2,2',3,3'-tetrachloro-1,1'-biphenyl	2,720	2,860
3	2,2',3,4,5'-Pentachlorobiphenyl	3,330	3,460
4	2,2',3,4,5-Pentachlorobiphenyl	3,010	3,160
5	2,2',4,4'-Tetrachlorobiphenyl	3,010	3,160
6	2,2',4,5,5'-Pentachlorobiphenyl	3,330	3,460
7	2,2',4,6,6'-Pentachlorobiphenyl	3,330	3,460
8	2,2',5,5'-tetrachlorobiphenyl	3,010	3,160
9	2,2',5,6'-tetrachlorobiphenyl	3,010	3,160
10	2,2',5-Trichlorobiphenyl	2,480	2,610
11	2,2',6,6'-Tetrachlorobiphenyl	3,010	3,160
12	2,2'-Dichlorobiphenyl	2,280	2,420
13	2,3,7,8-Tetrachloro-dibenzofuran	2,190	2,420
14	2,3,7,8-Tetrachloro-dibenzo-p-dioxin	1,860	2,350
15	2,3',4,4'-Tetrachlorobiphenyl	3,010	3,160
16	2,3',4'-Trichlorobiphenyl	2,720	2,860
17	2,4,4'-Trichlorobiphenyl	2,720	2,860
18	2,4,5-Trichlorobiphenyl	2,720	2,860
19	2,4,6-Trichlorobiphenyl	2,720	2,860
20	2,4-Dichlorobiphenyl	2,480	2,610
21	2,5-Dichlorobiphenyl	2,480	2,610
22	2-Chlorobiphenyl	1,770	2,040
23	3,3',5-Trichlorobiphenyl	2,720	2,860
24	3,3'-Dichlorobiphenyl	2,480	2,610
25	3,4-Dichlorobiphenyl	2,480	2,610
26	3-Chlorobiphenyl	2,280	2,420
27	3-Methylchlolanthrene	0,240	0,500
28	4-Chlorobiphenyl	2,280	2,420
29	7,12-Dimethylbenz[a]anthracene	0,250	0,510
30	Acenaphthylene	-0,140	0,100
31	Acenaphthene	0,680	0,940
32	Anthracene	0,060	0,230
33	Benz[a]anthracene	0,300	0,480
34	Benzo[a]pyrene	-0,130	0,040
35	Benzo[b]fluoranthene	0,900	1,160
36	Chrysene	0,640	0,900
37	Dibenz[a,h]anthracene	0,370	0,630
38	Fluorene	1,570	1,830
39	Hexachlorobenzen	4,310	4,570
40	Naphthalene	1,210	1,470
41	Penthachlorobenzene	3,780	4,040
42	Pyrene	0,130	0,310
43	3,5-dichloro-1,1'-biphenyl	2,480	2,610
44	Phenanthrene	1,040	1,300
45	γ-Hexachlorocyclohexane	3,330	3,460

Principal Components Analysis (ACP)

The molecules of POPs (1 to 45) were studied by statistical methods based on the principal component analysis (PCA) [18] using the software XLSTAT 2009.

This is an essentially descriptive statistical method which aims present, in graphic form, the maximum of information contained in a data Table 1.

PCA is a statistical technique useful for summarizing all the information encoded in the structures of compounds. It is also very helpful for understanding the distribution of the compounds.

Multiple Linear Regressions (RLM)

The multiple linear regression statistic technique is used to study the relation between one dependent variable and several independent variables. It is a mathematic technique that minimizes differences between actual and predicted values. The multiple linear regression model (MLR) was generated using the software SYSTAT, version 12, to predict air half-live Log HL. It has served also to select the descriptors used as the input parameters for a back propagation network (ANN).

Partial Least Square analysis (PLS)

The PLS has two objectives: to approximate the matrix X of molecular structure descriptors to the matrix Y of dependent variables and to maximize the correlation between them. The leave-one-out (LOO) method was used to perform the cross-validated analysis. The cross-validated coefficient, Q^2, is calculated using the following equation:

$$Q^2 = 1 - \frac{\sum (Yi - Yipred)^2}{\sum (Yi - Ymean)^2} \quad (1)$$

Where *Yi* is the i[th] experimental Log HL value, *Yipred* is the i[th] predicted Log HL *Ymean* is the mean of the experimental Log HL.

The optimal number of components (N) is employed to do non-validation PLS analysis to get the final model parameters such as determination coefficient R^2, standard deviation (S) and Fischer test value (F).

Artificial Neural Networks (ANNs)

The ANNs analysis was performed with the use of Matlab software v 2009a Neural Fitting tool (nftool) toobox on a data set of POPs air mean and maximum half-live property [17].

A number of individual models of ANN were designed built up and trained. Generally the network was built for tree layers; one input layer, one hidden layer and one output layer were considered [19]. The input layer was consisted of 8 artificial neurons of linear activation function (Fig.2). The number of artificial neural in the hidden layer was adjusted experimentally. The hidden layer consisted of 10 artificial neural. Two neurons formed the output layer of sigmoid function activation. The architecture of the applied ANN models is presented in Figure 2.

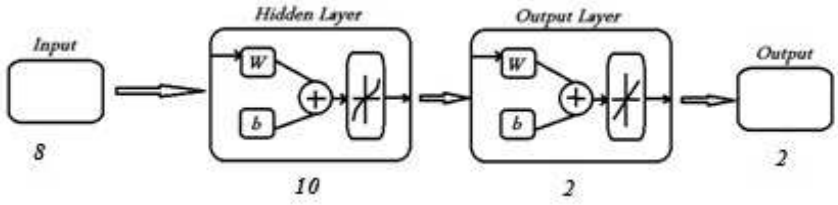


Fig. 2. The ANNs architecture.

The data subjected to ANN analysis was randomly divided into three sets: a learning set, a validation set and a testing set. Prior to that, the whole data set was scaled within the 0 to 1 range.

The set of POPs air half-live propriety [17] were subjected to the ANN analysis. First, for the learning set of compounds, i.e., 31 POPs derivatives were used. ANN models were designed, built and trained. The learning set of data is used in ANNs to recognize the relationship between the input and output data. Then for the revision of ANN model designed and selected, the validation set of 7 compounds was used. Testing set with 7 compounds was provided to be an independent evaluation of the ANN model performance for the finally applied network.

In this study, we selected the Sigmoid as a basis function [20]. The operation of the output layer is linear, which is given as below:

$$y_k(X) = \sum_{j=1}^{n_k} w_{kj} h_j(X) + b_k \quad (2)$$

Where y_k is the k[th] output layer unit for the input vector X, w_{kj} is the weight connection between the k[th] output unit and the j[th] hidden layer unit and b_k is the bias allows a transfer function "non-zero" given by the following equation:

$$Bias = \sum (\bar{y} - y) \quad (3)$$

Where y is the measured value and $\bar{y}$ is the value predicted by the model.

The accuracy of the model was mainly evaluated by Root Mean Square Error (RMSE). Formula as follows:

$$RMSE = \sqrt{\frac{1}{n} . \sum_{i=1}^{n} (p_{exp} - p_{pred})^2} \quad (4)$$

Where n = number of compounds, p_{exp} = experimental value, p_{pred} = predicted value and summation is over all patterns in the analysed data set [21,22]. The scripts were run on a personal PC.

DFT calculations

DFT (density functional theory) methods were used in this study. These methods have become very popular in recent years because they can reach similar precision to other methods in less time and less cost from the computational point of view. In agreement with the DFT results, energy of the fundamental state of a polyelectronic system can be expressed through the total electronic density, and in fact, the use of electronic density instead of wave function for calculating the energy constitutes the fundamental base of DFT [23-25] using the B3LYP functional [26,27] and a 6-31G* basis set. The B3LYP, a version of DFT method, uses Becke's three-parameter functional (B3) and includes a mixture of HF with DFT exchange terms associated with the gradient corrected correlation functional of Lee, Yang and Parr (LYP). The geometry of all species under investigation was determined by optimizing all geometrical variables without any symmetry constraints.

RESULTS AND DISCUSSION

A QSPR study was carried for a series of 45 of POPs, in order to determine a quantitative relationship between structure and half-lives.

Table 2. Values of the parameters obtained by DFT/B3LYP 6-31G* optimization of studied POPs

Mol	E_T (Ua)	E_{HOMO} (eV)	E_{LUMO} (eV)	ΔE (eV)	μ (D)	E_a (eV)	λ_{max} (nm)	$f_{(SO)}$	LogHL me(h)	LogHL mx(h)
1	-3370,067	-1,591	4,684	-6,275	0,238	4,317	287,230	0,001	1,770	1,910
2	-2301,670	-0,896	6,021	-6,917	3,725	3,427	361,780	0,032	2,720	2,860
3	-2761,263	-1,149	5,756	-6,905	2,688	3,515	352,750	0,066	3,330	3,460
4	-2761,258	-1,252	5,733	-6,985	3,633	3,605	343,930	0,070	3,010	3,160
5	-2301,679	-1,207	5,606	-6,813	1,690	3,570	347,280	0,115	3,010	3,160
6	-2761,242	-2,025	4,720	-6,745	1,474	4,044	306,610	0,002	3,330	3,460
7	-2761,267	-1,098	5,894	-6,992	1,935	1,734	714,98	0,001	3,330	3,460
8	-2301,678	-1,058	5,780	-6,837	0,111	3,550	349,260	0,033	3,010	3,160
9	-2301,573	-4,141	1,648	-5,789	3,219	1,566	791,54	0,001	3,010	3,160
10	-1841,921	-1,166	5,795	-6,961	2,495	4,146	299,040	0,053	2,480	2,610
11	-2301,674	-0,809	6,038	-6,847	0,000	1,738	713,18	0,008	3,010	3,160
12	-1382,490	-0,712	5,912	-6,624	2,011	3,571	347,200	0,039	2,280	2,420
13	-2375,697	-1,869	4,698	-6,567	0,616	4,178	296,780	0,019	2,190	2,420
14	-2450,894	-1,285	4,750	-6,035	0,000	4,254	291,470	0,151	1,860	2,350
15	-2301,678	-1,434	5,236	-6,670	1,770	3,989	310,810	0,456	3,010	3,160
16	-1842,084	-1,214	5,354	-6,568	3,514	4,098	302,540	0,357	2,720	2,860
17	-1842,088	-1,246	5,272	-6,518	1,430	4,029	307,710	0,480	2,720	2,860
18	-1842,082	-1,234	5,345	-6,579	2,262	4,086	303,450	0,257	2,720	2,860
19	-1842,082	-0,971	5,948	-6,919	0,986	3,400	364,690	0,044	2,720	2,860
20	-1382,492	-1,019	5,417	-6,436	2,168	4,147	298,980	0,351	2,480	2,610
21	-1382,492	-1,037	5,453	-6,490	0,728	4,201	295,100	0,200	2,480	2,610
22	-922,897	-0,748	5,575	-6,323	1,683	4,269	290,410	0,284	1,770	2,040
23	-1842,092	-1,481	5,268	-6,749	1,952	4,159	298,130	0,101	2,720	2,860
24	-1382,647	-1,497	5,288	-6,785	1,206	4,227	293,310	0,255	2,480	2,610
25	-1382,493	-1,173	5,177	-6,350	3,203	4,162	297,900	0,504	2,480	2,610
26	-922,902	-0,965	5,334	-6,299	2,101	4,162	297,900	0,504	2,280	2,420
27	-809,918	-1,377	3,647	-5,025	1,151	4,273	290,140	0,404	0,240	0,500
28	-922,902	-0,947	5,219	-6,166	2,147	4,152	298,590	0,462	2,280	2,420
29	-771,791	-1,537	3,589	-5,126	0,161	2,978	416,410	0,101	0,250	0,510
30	-462,088	-1,888	3,923	-5,811	0,334	1,541	804,530	0,010	-0,140	0,100
31	-463,315	-0,758	4,714	-5,472	0,818	4,100	302,380	0,089	0,680	0,940
32	-539,531	-1,633	3,596	-5,229	0,000	2,962	418,580	0,060	0,060	0,230
33	-693,179	-1,551	3,775	-5,325	0,066	3,120	397,450	0,067	0,300	0,480
34	-769,414	-1,738	3,368	-5,106	0,046	2,956	419,420	0,300	-0,130	0,040
35	-769,399	-1,717	4,009	-5,726	0,382	3,465	357,780	0,011	0,900	1,160
36	-693,182	-1,267	4,248	-5,515	0,000	3,567	347,620	0,151	0,640	0,900
37	-846,827	-1,479	3,899	-5,379	0,000	3,199	387,630	0,091	0,370	0,630
38	-501,423	-0,714	5,045	-5,759	0,482	4,550	272,500	0,155	1,570	1,830
39	-2989,783	-1,751	5,558	-7,309	0,000	4,632	267,680	0,003	4,310	4,570
40	-385,893	-0,959	4,833	-5,792	0,000	4,237	292,630	0,070	1,210	1,470
41	-2530,201	-1,476	5,695	-7,171	0,940	4,730	262,120	0,005	3,780	4,040
42	-615,773	-1,482	3,848	-5,330	0,000	3,541	350,160	0,310	0,130	0,310
43	-1382,497	-1,237	5,301	-6,538	2,428	4,758	260,600	0,002	2,480	2,610
44	-539,539	-0,994	4,740	-5,734	0,042	3,886	319,060	0,002	1,040	1,300
45	-2993,428	-1,757	6,589	-8,345	0,000	5,959	208,070	0,004	3,330	3,460

The table 2 shows the values of the calculated parameters obtained by DFT/B3LYP 6-31G* optimization of the studied POPs.

Principal component analysis (Training Set Selection)

The selection of the training set is one of the most important steps in QSPR modeling, since the establishment and optimization of a QSPR model are based on this training set. Predictability and applicability of a QSPR model also depend on the training set selection.

In this part, PCA was applied to select a training set from among 45 POPs derivatives.

The set of descriptors encoding the 45 POPs and electronic and energetic parameters are submitted to PCA analysis [28]. The first three principal axes are sufficient to describe the information provided by the data matrix. Indeed, the percentages of variance are 44,38% ; 21,76% and 11,52% for the axes PC1, PC2 and PC3 respectively. The total information is estimated to a percentage of 79,65%.

The principal component analysis (PCA) [29] was conducted to identify the link between the different variables. Bold values are different from 0 at a significance level of p = 0,05. Correlations between the eight descriptors are shown in table 3 as a correlation matrix and in figure 4 these descriptors are represented in a correlation circle.

The Pearson correlation coefficients are summarized in the following table 3. The obtained matrix provides information on the negative or positive correlation between variables.

Bold values are different from 0 at a level significant for p < 0,05
At a very significant for p < 0,01
At a highly significant to p < 0,001

Table 3. Correlation matrix (Pearson (n)) between different obtained descriptors

Variables	E_T	E_{HOMO}	E_{LUMO}	ΔE	μ	E_a	λ_{max}	$f_{(SO)}$	LogHLme	LogHLmx
E_T	**1**									
E_{HOMO}	0,213	**1**								
E_{LUMO}	**-0,456**	**0,667**	**1**							
ΔE	**0,772**	-0,120	**-0,819**	**1**						
μ	-0,290	-0,003	0,265	**-0,356**	**1**					
E_a	-0,022	0,274	**0,332**	-0,231	-0,003	**1**				
λ_{max}	-0,215	-0,029	0,073	-0,120	-0,015	**-0,806**	**1**			
$f_{(SO)}$	0,292	0,227	-0,015	0,194	0,273	0,253	-0,240	**1**		
LogHLme	**-0,806**	-0,003	**0,668**	**-0,893**	**0,492**	0,138	0,173	-0,088	**1**	
LogHLmx	**-0,808**	-0,002	**0,664**	**-0,887**	**0,471**	0,147	0,168	-0,098	**0,998**	**1**

** The Log HL is well correlated with the energy E_{LUMO} for (r = 0,66 and p < 0,05) at a significant level.*
** The E_{HOMO} energy is positively correlated with the dipole moment (r = 0,61 and p < 0,05) and λ_{max} (r = 0,635 and p < 0,05) at a significant level.*
** The Log HL is negatively correlated with the gap energy **ΔE** (eV) (r = – 0,88 and p < 0,05) and with E_T (eV) r = – 0,8 p < 0,05 at a significant level.*
** The Log HLme is strongly correlated with Log HLmx for r = 0,99 and p < 0,001 at a high level.*

Correlation circle

The Principal Component Analysis (PCA) was also performed to detect the connection between the different variables. The principal component analysis revealed from the correlation circle (Fig.3) shows that the PC1 axis (49,76% of the variance) is mainly due to the HUMO energy, while the axis PC2 (23,32% of the variance) is located by the other parameters of energy.

On the other hand, the correlation circle and the 3D plot of PCA (Fig. 3) shows that there is a strong correlation between air half-life persistence and HOMO energy.

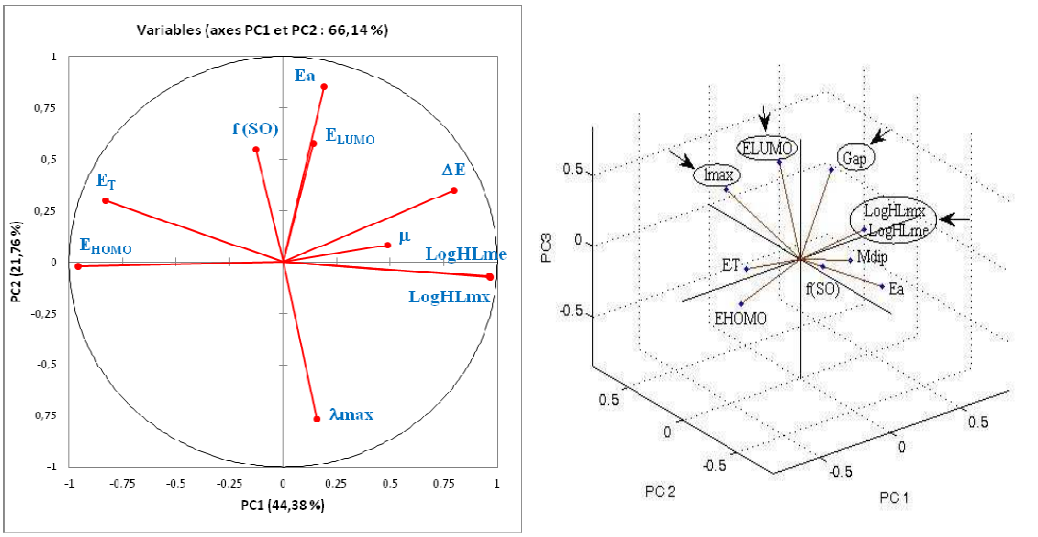


Fig. 3. Correlation circle and 3D-PCA plot

On the other hand, the projection PC1-PC2 (66,14% of the total variance) also shows that we can discern six groups of molecules with special structures propriety.

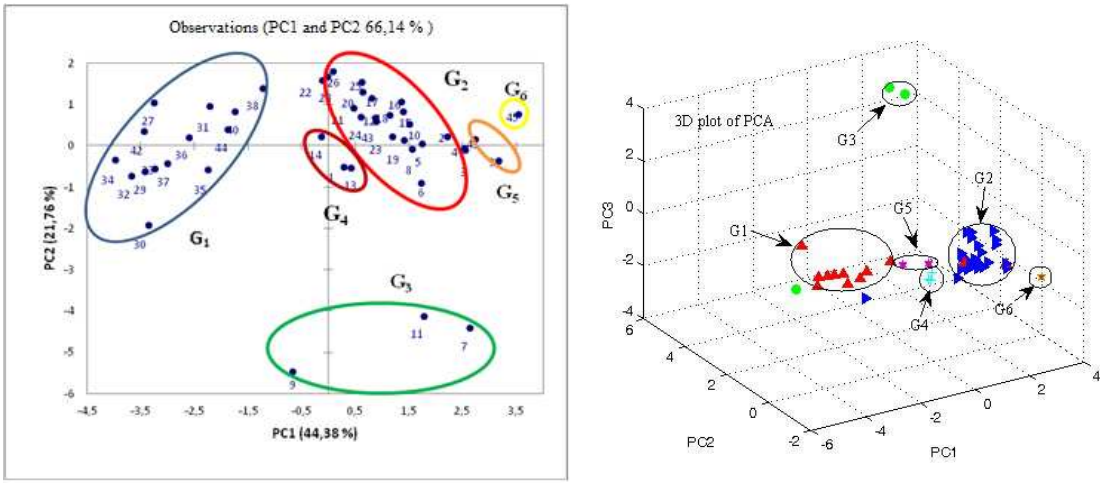


Fig. 4. Cartesian diagram according to PC1, PC2 and PC3: Separation between six groups.

The figure 4 shows a distribution of molecules in six groups: the group 1 (G_1) containing the flat benzo-condensed hemolytic compounds, the group 2 (G_2) variously substituted biphenyls but not more than two ortho, the group 3 (G_3) which 3,4-orthosubstituted biphenyl derivatives with λ_{max} maximum, the group 4 (G_4) flat benzo-condensed heterolytic, the group 5 (G_5) contains one aromatic ring, the group 6 (G_6) not aromatic compound.

Multiple linear regressions

To establish quantitative relationships between air half-live and selected descriptors, our array data were subjected to a multiple regression linear and nonlinear. Only variables whose coefficients are significant were retained.

Multiple linear regressions of the variable means and maximum half-live (MLR)

Many attempts have been made to develop a relationship with the indicator variable of mean half-live LogHLme , but the best relationship obtained by this method is only one corresponding to the linear combination of several descriptors: the total energy, energy E_{HOMO}, energy E_{LUMO}, activation energy E_a, the dipole moment μ and the factor of oscillation $f_{(SO)}$.

$$\text{LogHLme(h)} = -5{,}999 - 3{,}504.10^{-4}\,E_T - 1{,}149\,E_{LUMO} + 1{,}004\,\Delta E + 0{,}185\,\mu + 0{,}186\,E_a + 2{,}610.10^{-4}\,\lambda_{max} + 0;556\,f_{(SO)}$$ Equation (5)

N = 45 R^2 = 0,87 R^2_{adj} = 0,85 S = 1.21 F = 31,60 RMSE = 0,41

$$\text{LogHLmx(h)} = -5{,}463 - 3{,}701.10^{-4}\,E_T - 1{,}068\,E_{LUMO} + 0{,}934\,\Delta E + 0{,}168\,\mu + 0{,}165\,E_a + 2{,}759.10^{-4}\,\lambda_{max} + 0{,}469\,f_{(SO)}$$ Equation (6)

N = 45 R^2 = 0,86 R^2adj = 0,83 S = 1.18 F = 28,20 RMSE = 0,42

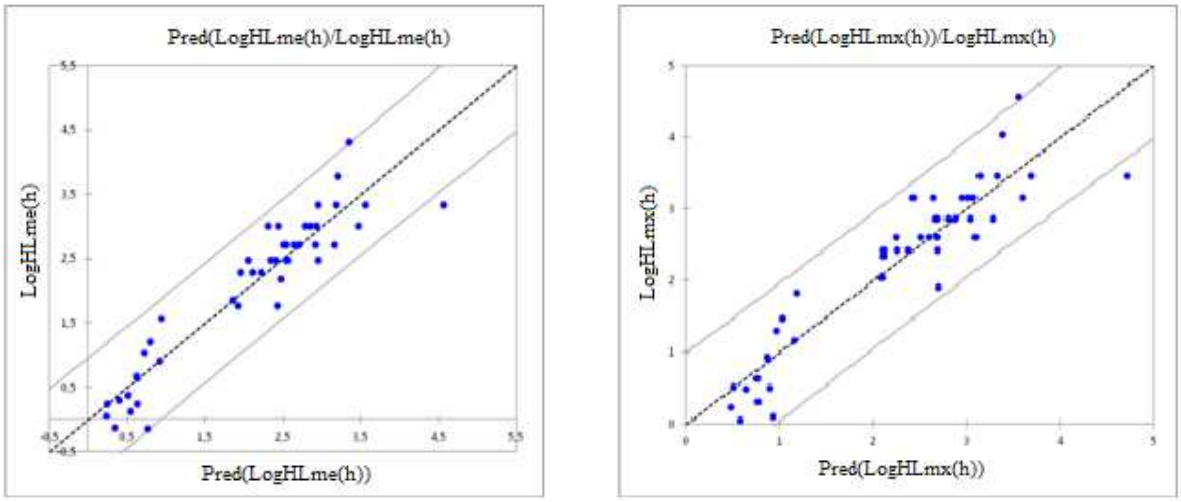


Fig. 5. Graphical representation of calculated and observed Log HL(h).

The figure 6 shows a very regular distribution of air half-lives persistence values depending on the experimental values.

Multiple nonlinear regressions of the variable means and maximum half-live (MNLR)

We have used also the technique of nonlinear regression model to improve the structure-mean half-live in a quantitative way. It takes into account several parameters. This is the most common tool for the study of multidimensional data. We have applied to the table 2 containing 45 molecules associated with 8 variables. The resulting equations are:

$$\begin{aligned}\text{LogHLme(h)} = {} & -17{,}907 - 5{,}920.10^{-4}\,E_T - 3{,}073\,E_{LUMO} + 0{,}497\,\Delta E + 2{,}051\,\mu \\ & + 5{,}675\,E_a + 1{,}122.10^{-2}\,\lambda_{max} - 0{,}828\,f_{(SO)} - 1{,}038.10^{-7}\,E_T^2 \\ & + 0{,}462\,E_{LUMO}^2 - 0{,}233\,\Delta E^2 + 0{,}247\,E_{HOMO}^2 - 6{,}782\,\mu^2 - 6{,}615\,E_a^2 \\ & + 1{,}932.10^{-6}\,\lambda_{max}^2 + 2{,}794\,f_{(SO)}^2\end{aligned}$$ Equation (7)

N = 45 R^2 = 0,95 R^2adj = 0,94 S = 1,27 F = 91,70 RMSE = 0,25

$$\begin{aligned}\text{LogHLmx(h)} = {} & -18{,}059 - 8{,}050.10^{-4}\,E_T + 4{,}195\,E_{LUMO} + 0{,}177\,\Delta E + 3{,}051.10^{-4}\,\mu \\ & + 6{,}235\,E_a + 1{,}353.10^{-2}\,\lambda_{max} - 0{,}846\,f_{(SO)} - 1{,}502.10^{-7}\,E_T^2 \\ & + 0{,}485\,E_{LUMO}^2 - 0{,}301\,\Delta E^2 + 0{,}318\,E_{HOMO}^2 - 4{,}381\,\mu^2 - 0{,}600\,E_a^2 \\ & - 2{,}400.10^{-6}\,\lambda_{max}^2 + 2{,}634\,f_{(SO)}^2\end{aligned}$$ Equation (8)

N = 45 R^2 = 0,95 R^2adj = 0,94 S = 1,24 F = 82,06 RMSE = 0,26

The air half-lives persistence value Log HL predicted by this model is somewhat similar to that observed. The figure 7 shows a very regular distribution of air half-lives persistence values based on the observed values.

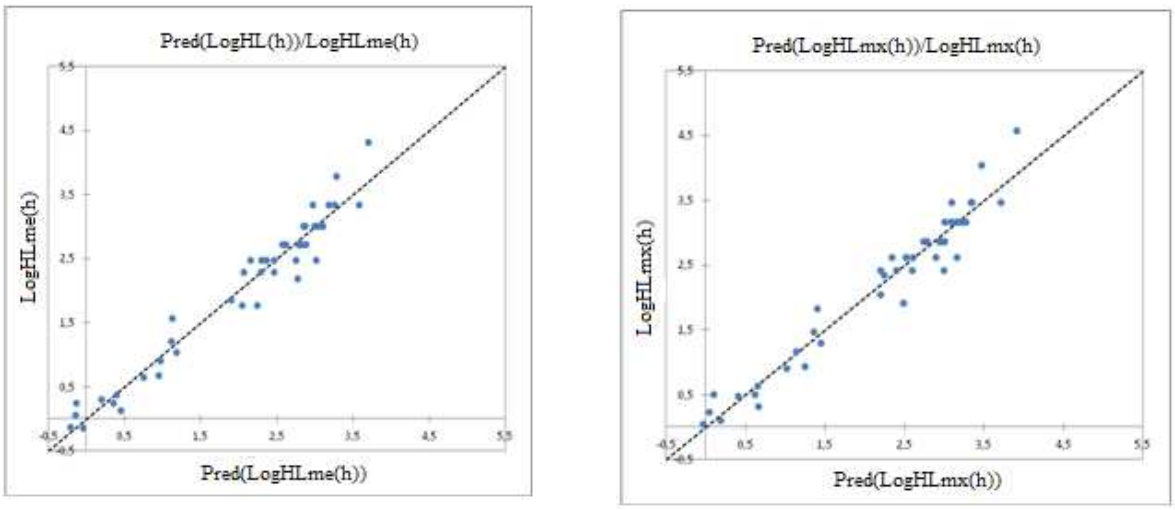


Fig. 6. Graphical representation of calculated and observed LogHL(h)

The obtained coefficient of determination in equation (8) is quite interesting (0,953).

Partial least square regression (PLS)

To linearly correlate the molecule descriptors: the total energy, energy E_{HOMO}, energy E_{LUMO}, energy gap ΔE, the dipole moment μ, activation energy E_a, absorption maximum λ_{max} and the factor of oscillation $f_{(SO)}$ to Log HL, the following equations was used:

$$\begin{aligned}\text{LogHLme(h)} = &-5{,}999 - 2{,}479.10^{-4}\,E_T - 0{,}213\,E_{LUMO} + 0{,}495\,\Delta E - 0{,}722\,E_{HOMO} \\ &+ 9{,}320.10^{-2}\,\mu - 4{,}219.10^{-2}\,E_a + 1{,}187.10^{-4}\,\lambda_{max} + 0{,}155\,f_{(SO)}\end{aligned}$$ Equation (9)

N = 45 R^2 = 0,83 R^2adj = 0,79 Q^2 = 0,12 S = 1,23 F = 21,82 RMSE = 0,50

$$\begin{aligned}\text{LogHLmx(h)} = &-5{,}255 - 2{,}428.10^{-4}\,E_T - 0{,}209\,E_{LUMO} + 0{,}484\,\Delta E - 0{,}722\,E_{HOMO} \\ &+ 9{,}428.10^{-2}\,\mu - 4{,}131.10^{-2}\,E_a + 1{,}163.10^{-4}\,\lambda_{max} + 0{,}152\,f_{(SO)}\end{aligned}$$ Equation (10)

N = 45 R^2 = 0,82 R^2adj = 0,78 Q^2 = 0,11 S = 1,20 F = 20,51 RMSE = 0,50

The figure 8 shows a very regular distribution of Log HL values depending on the experimental values.

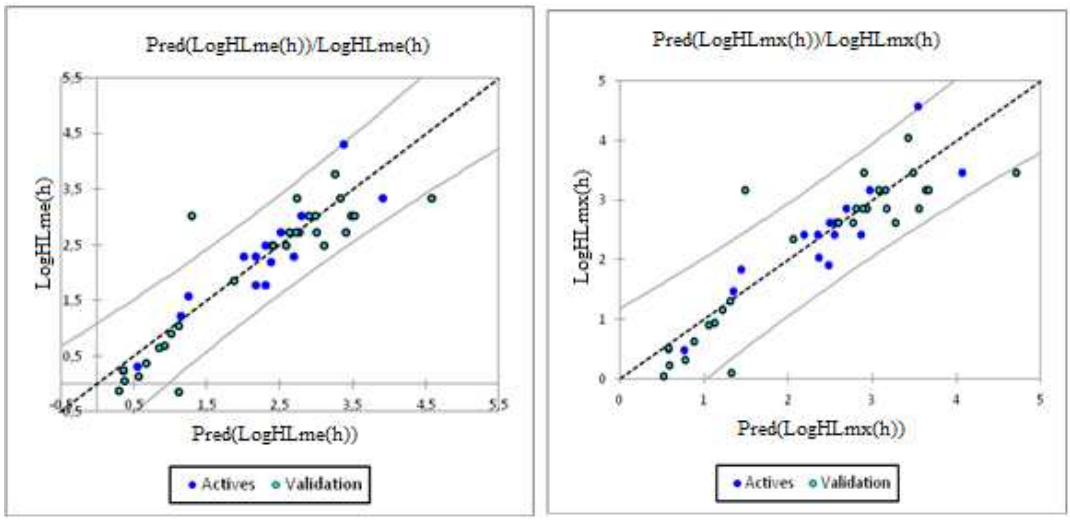


Fig. 7. Graphical representation of calculated and observed Log HL

The obtained coefficient of determination in equation (5) is quite interesting (0,8). To optimize the error standard deviation and better finish building our model, we involve in the next part artificial neural networks (ANN).

As part of this conclusion, we can say that the LogHL values obtained from past least square regression are highly correlated to that of the observed Log HL.

To optimize the error standard deviation and a better finish building our model, we involve in the next part artificial neural networks (ANN).

As part of this conclusion, we can say that the air half-lives persistence values obtained from nonlinear regression are highly correlated to that of the observed air half-life persistence comparing to results obtained by MLR method.

Artificial neural network (ANN)

In order to increase the probability of good characterization of studied compounds, neural networks (ANN) can be used to generate predictive models of quantitative structure-propriety relationships (QSPR) between a set of molecular descriptors obtained from the MLR and observed air half-lives persistent. The ANN calculated air half-life persistence model were developed using the properties of several studied compounds. The correlation between ANN calculated and experimental air half-lives persistence values is very significant as illustrated in figure 9 and as indicated by R and R^2 values.

N = 45 R = 0,991 R^2 = 0,982 RMSE = 0,138

These values show that the relationship between the estimated values of LogHL and their residues established by artificial neural networks are illustrated in figure 10.

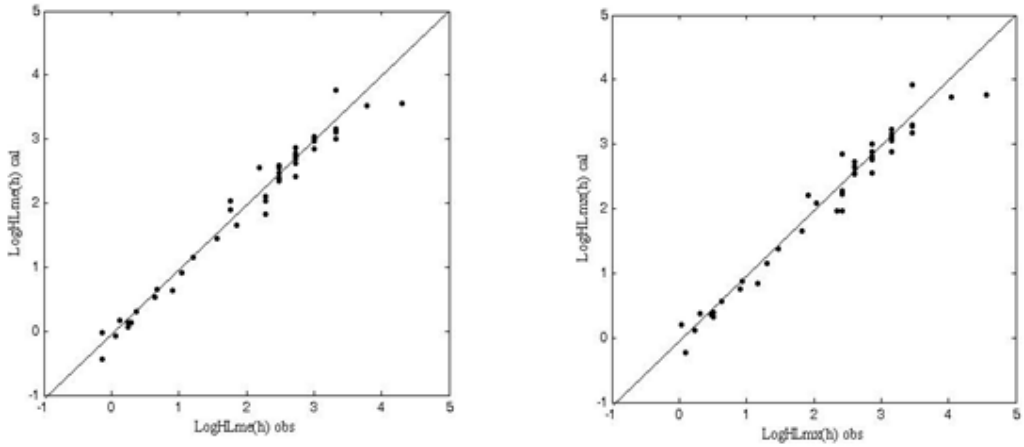


Fig. 8. Correlation between the calculated and experimental Log HL.

The statistic of the tree steps of the calculated by the ANNs: Training, validation and test are illustrated in table 4.

Table 4. Values obtained by ANNs.

	Samples	RMSE	R	R^2
Training	31	0,050	0,981	0,962
Validation	7	0,049	0,992	0,984
Test	7	0,049	0,984	0,968

R: correlation coefficient; R^2: determination coefficient; RMSE: root mean square error.

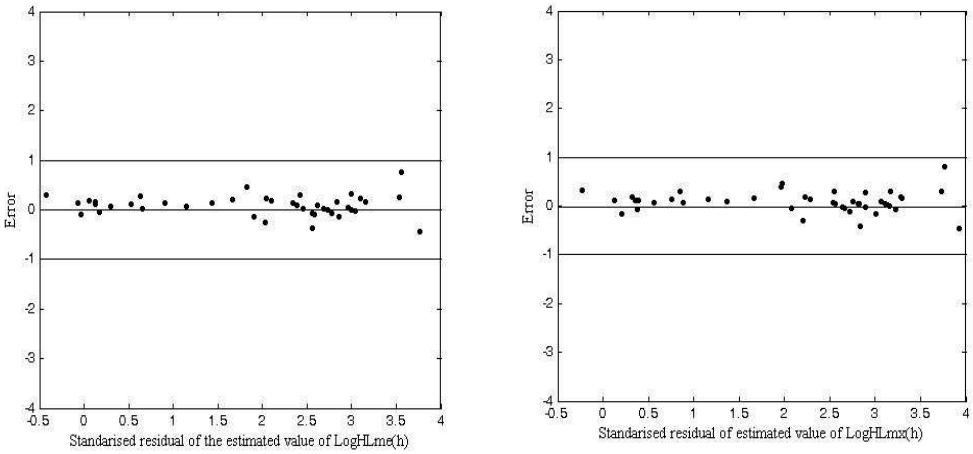


Fig.9. Relationship between the estimated values of Log HL and their residues established by artificial neural networks

The obtained squared correlation coefficient (R^2) value is 0,982 for this data set of POPs. It confirms that the artificial neural network results were the best to build the quantitative structure activity relationship models.

In this part, we investigated the best linear QSAR regression equations established in this study. Based on this result, a comparison of the quality of de PCA, MLR and ANN models shows that the ANN models have substantially better predictive capability because the ANN approach gives better results than MLR. ANN was able to establish a satisfactory relationship between the molecular descriptors and the activity of the studied compounds Table 5.

Table 5. Statistical results of comparative all models based on the N = 45 compounds

Statistical results	Log air mean half-life value (calc)				Log air maximum half-life value (calc)			
	MLR	**MNLR**	**PLS**	**ANN**	**MLR**	**MNLR**	**PLS**	**ANN**
F	31,61	91,70	21,82	133,25	28,21	82,06	20,51	107,39
S	1,22	1,27	1,23	1,31	1,18	1,24	1,20	1,27
R^2	0,88	0,95	0,83	0,97	0,86	0,95	0,82	0,96
R^2adj	0,85	0,94	0,79	0,96	0,83	0,94	0,78	0,95
RMSE	0,41	0,25	0,50	0,21	0,42	0,26	0,50	0,23
Q^2	---	---	0,12	---	---	---	0,11	---

R^2: determination coefficient; R^2adj: adjusted determination coefficient; Q^2: the cross-validated value; S: standard error of estimated; F: Fischer test value; RMSE: root mean square error.

CONCLUSION

In this work we have investigated the QSPR regression to predict air half-life persistence of several compounds based on POPs.

Comparison of key statistical terms like R or R^2 of different models obtained by using different statistical tools and different descriptors has been shown in table 6.

The study of the quality of the MLR and ANN models shown that the ANN results have substantially better predictive capability than the other methods. With ANN approach we have establish a relationship between several descriptors (E_T, E_{HOMO}, E_{LUMO}, ΔE, E_a, λ_{max} , μ and $f_{(SO)}$) and air half-life persistence in satisfactory manners.

Finally, we can conclude that one studied descriptors (E_T, E_{HOMO}, E_{LUMO}, ΔE, E_a, λ_{max}, μ and $f_{(SO)}$.), which are sufficiently rich in chemical and electronic information to encode the structural features, may be used with other topological descriptors for development of predictive QSPR models.

Table 6. Observed values and calculated of LogHL (mean and maximum) according to different methods.

N°	Obs.	Log air mean half-life value(calc)				Obs.	Log air maximum half-life value(calc)			
		MLR	NMLR	PLS	ANN		MLR	NMLR	PLS	ANN
1	1,77	2,43	2,25	2,30	2,03	1,91	2,70	2,48	2,48	2,20
2	2,72	3,16	2,88	3,40	2,62	2,86	3,28	3,01	3,56	2,76
3	3,33	3,18	2,97	3,33	3,16	3,46	3,33	3,09	3,49	3,29
4	3,01	3,47	3,10	3,49	2,96	3,16	3,59	3,23	3,64	3,13
5	3,01	2,79	2,85	3,00	3,04	3,16	2,94	3,00	3,16	3,16
6	3,33	2,95	3,18	2,73	3,00	3,46	3,14	3,34	2,90	3,17
7	3,33	3,56	3,26	3,92	3,10	3,46	3,69	3,35	4,07	3,28
8	3,01	2,45	2,87	2,90	3,00	3,16	2,64	3,09	3,07	3,23
9	3,01	2,31	2,99	1,29	2,96	3,16	2,43	3,14	1,49	3,11
10	2,48	2,95	3,02	3,11	2,56	2,61	3,09	3,16	3,27	2,66
11	3,01	2,85	3,08	3,52	2,84	3,16	3,02	3,27	3,68	2,89
12	2,28	2,22	2,47	2,70	1,83	2,42	2,38	2,60	2,87	1,97
13	2,19	2,48	2,77	2,38	2,56	2,42	2,69	2,99	2,56	2,84
14	1,86	1,86	1,90	1,87	1,66	2,35	2,11	2,24	2,06	1,96
15	3,01	2,93	3,03	2,80	3,00	3,16	3,07	3,17	2,97	3,06
16	2,72	2,91	2,62	2,77	2,73	2,86	3,04	2,79	2,94	2,81
17	2,72	2,54	2,78	2,52	2,69	2,86	2,69	2,94	2,69	2,82
18	2,72	2,64	2,57	2,64	2,78	2,86	2,79	2,74	2,81	2,89
19	2,72	2,51	2,83	3,01	2,42	2,86	2,67	2,97	3,17	2,55
20	2,48	2,35	2,36	2,41	2,39	2,61	2,50	2,52	2,59	2,54
21	2,48	2,06	2,30	2,31	2,46	2,61	2,25	2,51	2,49	2,66
22	1,77	1,92	2,04	2,18	1,90	2,04	2,09	2,20	2,36	2,08
23	2,72	2,71	2,80	2,72	2,86	2,86	2,87	2,97	2,89	3,01
24	2,48	2,54	2,75	2,59	2,58	2,61	2,69	2,90	2,77	2,72
25	2,48	2,56	2,47	2,39	2,46	2,61	2,69	2,60	2,57	2,64
26	2,28	2,11	2,29	2,17	2,10	2,42	2,25	2,39	2,35	2,28
27	0,24	0,64	0,36	0,36	0,13	0,5	0,89	0,62	0,58	0,39
28	2,28	1,96	2,07	2,01	2,04	2,42	2,11	2,20	2,19	2,23
29	0,25	0,25	-0,13	0,35	0,06	0,51	0,50	0,10	0,57	0,32
30	-0,14	0,76	-0,04	1,12	-0,43	0,1	0,93	0,19	1,32	-0,23
31	0,68	0,62	0,95	0,92	0,65	0,94	0,86	1,25	1,13	0,88
32	0,06	0,23	-0,14	0,37	-0,07	0,23	0,47	0,04	0,59	0,12
33	0,3	0,40	0,20	0,55	0,13	0,48	0,65	0,41	0,77	0,36
34	-0,13	0,34	-0,21	0,29	-0,03	0,04	0,58	-0,03	0,51	0,21
35	0,9	0,92	0,97	1,01	0,63	1,16	1,14	1,13	1,22	0,85
36	0,64	0,63	0,75	0,84	0,53	0,9	0,87	1,01	1,05	0,76
37	0,37	0,51	0,40	0,67	0,30	0,63	0,75	0,65	0,88	0,56
38	1,57	0,94	1,13	1,24	1,44	1,83	1,18	1,40	1,45	1,66
39	4,31	3,35	3,70	3,38	3,56	4,57	3,55	3,91	3,54	3,76
40	1,21	0,79	1,11	1,14	1,15	1,47	1,03	1,36	1,35	1,37
41	3,78	3,20	3,29	3,26	3,53	4,04	3,39	3,47	3,42	3,73
42	0,13	0,54	0,45	0,56	0,17	0,31	0,77	0,67	0,78	0,38
43	2,48	2,41	2,15	2,42	2,34	2,61	2,59	2,34	2,60	2,56
44	1,04	0,72	1,18	1,11	0,91	1,3	0,97	1,45	1,32	1,16
45	3,33	4,56	3,57	4,58	3,76	3,46	4,72	3,72	4,71	3,92

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

REFERENCES

[1] Hansch. C, Muir. R. M, Fujita. T, Maloney. P. P, Geiger. F, Streich. M, *J. Am. Chem. Soc.*, **1963**, 85, 2817-2825.
[2] Bodor. N, *Current Medicinal Chemistry*, **1988**, 5, 353-380. From book: Biochemistry of Redox Reactions, by Bernard Testa, editor: London [u, a], Acad. Press, **1995**.
[3] Sabljic. A, Güsten. H, Verhaar. H, Hermens. J, *Chemosphere*, **1995**, 31, 4489–4514.
[4] Sabljic. A, *Chemosphere*, **2001**, 43, 363–375.
[5] Yang Wen. Li M, Su Wie. C, Qin. Ling Fu. Jia He. Yuan. H, Zhao, *Chemosphere*, **2012**, 86, 634–640.
[6] Benigni. R, Zito. R, *Mutat. Res.*, **2004**, 566, 49–63.
[7] Zakarya. D, Larfaoui. E. M, Boulaamail. A, Tollabi. M. and Lakhlifi. T, *Chemosphere*, **1998**, Vol. 36, N° 13, 2809-2818.

[8] Elhallaoui. M, Elasri. M, Ouazzani. F, Mechaqrane. A. and Lakhlifi. T, *Int. J. Mol. Sci.*, **2003**, 4, 249-262.
[9] Papa. E, Battaini. F, Gramatica. P, *Chemosphere*, **2005**, 58, 559-570.
[10] Zhang. L, Hao. G. F, Tan. Y, Xi. Z, Huang. M. Z, Yang. G. F, *Bioorganic & Medicinal Chemistry*, **2009**, 17, 4935-4942.
[11] Laarej. K, Bouachrine. M, Radi. S, Kertit. S and Hammouti. B, *E-Journal of Chemistry*, **2010**, 7(2), 419-424.
[12] Zarrok. H, Oudda. H, Zarrouk. A, Salghi. R, Hammouti. B, Bouachrine. M, *Der Pharma Chemica*, **2011**, 3 (6): 576-590.
[13] Larif. M, Adad. A, Hmammouchi. R, Taghki. A. I, Soulaymani. A, Elmidaoui. A, Bouachrine. M, Lakhlifi. T, Biological activities of triazine derivatives. Combining DFT and QSAR results, article in press in *Arabian Journal of Chemistry*, **2013**, http://dx.doi.org/10.1016/j.arabjc.2012.12.033
[14] Adad. A, Larif. M, Hmammouchi. R, Taghki, A. I, Bouachrine. M, Lakhlifi. T, Submitted in *Journal of Chemical Acta*., **2013**.
[15] Matthies. M, Beyer. A, Mackay. D, *Organochalogen Compd*, **1999**, 41: 347–349.
[16] Beyer. A, Mackay. D, Matthies. M, Wania. F, Webster. E, *Environ Sci Technol*, **2000**, 34(4): 699–703.
[17] Lu Yuyin., Yin Chunsheng, Liu Hongyan, Yi Zhongsheng, Wang Yang, *Journal of Environmental Sciences*, **2008**, 20, 1433-1438.
[18] Hogarh. J. N, Seike. N, Kobara. Y, Habib. A, Namd. J. J, Lee. J. S, Qilu. Li. Liu. X, Jun Li. Zhang. G, Masunaga. S, *Chemosphere*, **2012**, 86: 718–726.
[19] Zupan. J. Gasteiger. J, Neural Networks for Chemistry and Drug Design: An Introduction, second ed., VCH, Weinheim, **1999**.
[20] Turkkan. N, *Revue de l'Université de Moncton*, **1993**, 26 (1), 205-221.
[21] Lee. P. Y, Chen. C. Y. J, *Hazard. Mater.*, **2009**, 165, 156-161.
[22] Jing. G, Zhou. Z, Zhuo. J, *Chemosphere*, **2012**, 86, 76-82.
[23] Adamo. C, Barone. V, *Chem. Phys. Lett.*, **2000**, 330, 152–160; Abdel Aziz Abu-Yamin, Mahmoud Salman and Ibrahim Saraireh, *J. Chem. Pharm. Res*, **2013**, 5(3):33-4; Vasishta. D, Bhatt, Smita Srivastava and Ketul Patel, *J. Chem. Pharm. Res.*, **2010**, 2(3):559-566.
[24] Parac. M, Grimme. S, *J. Phys. Chem.*, **2003**, A 106, 6844–6850; Narpat Raj Nenival and Gangotri, K. M, *J. Chem. Pharm. Res.*, **2011**, 3(6): 553-561.
[25] Gaussian 03, Revision B.01, M. J. Frisch, M. J. and al., Gaussian, Inc., Pittsburgh, PA, **2003**.
[26] Becke. A. D, *J. Chem. Phys.*, **1993**, 98, 1372.
[27] Lee. C, Yang. W, Parr. R G, *Phys. Rev.*, **1988**, B. 37, 785-789; Mahesh. C and Meena. R.C, *J. Chem. Pharm. Res.*, **2011**, 3(3):264-270.
[28] STATITCF Software, **1987**. Technical Institute of cereals and fodder, Paris, France.
[29] Jonathan. N, Nobuyasu. H, Yuso. S, Ahsan. K, Jae-Jak. H, Jong-Sik. N.N, Xiang. L Q L, Jun. L L, Gan. Z and Shigeki. M, *Chemosphere*, **2012**, 86, 718–726.