Quantitative structure-property relationship study of alcohols water solubility based on a new model combined modified autocorrelation method and electro-topological indices

Chaymae Jermouni (chaymae.jermouni@gmail.com)
Universite Hassan II Casablanca Faculte des Sciences Techniques Mohammedia
https://orcid.org/0000-0002-7535-3658
Mohamed Nohair
Universite Hassan II Casablanca Faculte des Sciences Techniques Mohammedia
https://orcid.org/0000-0002-7553-4205

Research article

Keywords: Water solubility, aliphatic alcohols, modified autocorrelation method, electro-topological indices, Multiple Linear Regression, Artificial Neural Networks, Principal Components Analysis

Posted Date: April 2nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-20088/v1

Abstract

In this study, structure water solubility modeling was performed to describe a set of 50 of aliphatic alcohols in a Quantitative Structure-Property Relationship model by developing of two descriptors types based on multifunctional autocorrelation method, which gives a general description of whole molecule; and electro-topological descriptors. The index combines the topological nature with electronic state of the atom. The Modified Autocorrelation Method was used in structure–property relationships to describe the local environment of the hydroxyl group. For the statistical studies, Multiple Linear Regression, Artificial Neural Networks and Principal Components Analysis were used. The approach efficiency approach was evaluated through the predictive ability of models by leave-p-Out cross-validation method. The coefficient of determination and errors of descriptors combination calculated respectively by multiple linear regression and artificial neural networks were r= 0.99, s = 0.18 and r = 0.99, s = 0.32. In order to simplify components computation, the molecules were coded by means of SMILES system and stored as input files. The results showed that aliphatic alcohols solubility is dominated by the shape and molecule branching, also the electro-topological descriptors had a considered model effect.

1. Introduction

The present study is aimed at estimating water solubility of Aliphatic Alcohols, because it is considered as an important property of organic compounds [1]. Since alcohols are toxic materials, therefore it represents dangerous environmental pollutants; the first step in polluting alcohols action is their solubility in water. Various quantitative structure-property relationship (QSPR) models were proposed for estimating water solubility of Aliphatic Alcohols. In General, QSPR methods are based on the assumption that chemical compound property related to its structure through a certain mathematical algorithm. This relationship can be used in the prediction, interpretation, and new assessment compounds with desired properties reducing time, efforts and synthesis cost, and rationalizing of new product development [2–7].

Several factors influencing the physical property of molecule. Such us, the shape, the electronic structure and the molecular size. These factors are associated with different intermolecular interaction aspects, as though Van Der Waals forces, the polarity and the molecule ability to participate in hydrogen bonding caused by heteroatoms and which considered a very interesting factor influencing the physical properties of alcohols.

A significant approach is based on various group contribution methods using atomic types, bonds, or molecular fragments as descriptors with respect to different properties. [8]Among the many methods that describe the chemical structure for a set of molecules is that based on the molecular graph. There are now many different descriptions, among the methods used to describe molecular structures, many are based on the topological indices (TI)[1, 9–11] derived by the molecular graph. In fact, these conventional indices don’t take into account the each contributions of the individual types or groups of atoms to properties. For these reasons, the multifunctional autocorrelation method was used to describe the influence of the local environment of the molecule.

for the global molecule and the sum calculated for all pairs of atoms in the graph separated by a distance d of k bonds is considered as the component of order k (P_k) of an autocorrelation vector P corresponding to the specific property, and to describe the influence of the local environment of the molecule we used the modified autocorrelation method.

On the other hand, to increase the efficiency of topological indices, Kier and Hall [12] set up the electro-topological indices that boil down in the valence connectivity indices, including information not present in the graph such as the valence electron and the number of electrons in an atom with respect largely to their use as descriptors in QSAR/QSPR modeling.

Three-layer feedforward neural network was used for statistical method trained by the back-propagation algorithm and Multiple Linear Regression (MLR).

Also predictive linear and non-linear models were obtained by forward stepwise multi-linear regression analysis (MRA) and Artificial Neural Network (ANN) approaches respectively.

2. Method

MAM method originate from Moreau Autocorrelation method [13], elected to implement structure-property relationships based on the autocorrelation function. This method can be useful for database characterize action and encoding various physicochemical properties [7,14-16]. Moreau and Broto [8] consider the molecule as a graph, in which the nodes are atoms (hydrogen atoms are not taken into account) linked by bonds. Properties *f* of atoms (surface, volume, electronegativity…) are defined on every node of the graph. The Components P, of the autocorrelation vector corresponding to that property are defined by the relation bellow:

$$Pk = \sum_{k=0}^{n}((fi)(fj)). \quad (1)$$

Where P_k is autocorrelation component corresponding to topological distance of k bonds (smallest number of bonds between i and j) to the specific property f_i

To facilitate the computation for Big Data, a computation program on Matlab which generates the connection matrices, distances matrices and the variables of the MAM was developed in the following way:

The first program was used for the calculation of global environment descriptors of the molecule.

```
W=zeros(1,10);
i=1:1:
j=i+1:1:
D(i,j)==1
W(1,1)=W(1,1)+ VA(i)*VA(j);
D(i,j)==2
W(1,2)=W(1,2)+ VA(i)*VA(j);
D(i,j)==3
.
.
.
D(i,j)==n
W(1,n)=W(1,n)+ VA(i)*VA(j);
```

A new component Pik was specified to characterize the local environment of the atoms in molecules by the formula bellow:

$$Pik = \sum_{k=0}^{n}((fi)(fj))^{x} \qquad (2)$$

The principle of Modified Autocorrelation Method proposed by Nohair et al [16], comes from setting carbon atom i and P_{ik} the sum of *f*(i)**f*(j) of all chemical bonds existing between all pairs of carbon atoms i(fixed atom) and j separated by a topological distance k, and x=1.

The second program was applied to study the local environment of molecule, while fixing the atom concerned:

```
i=1:1:
D(1,i)==1
W(1,6)=W(1,6)+ VA(i)*VA(1);
D(1,i)==2
W(1,7)=W(1,7)+ VA(i)*VA(1);
D(1,i)==3
.
.
.
D(1,i)==n
W(1,n)=W(1,n)+ VA(i)*VA(1);
```

Fig. 1 shows the graph form of 2,2-Dimethyl-3-pentanol molecule. Several matrices such as connection matrix and distance matrix (Fig. 2) can be determined from this form.

Table 1 presents an example of variables calculated by developed program for 2,2-Dimethyl-3-pentanol. Noting that, the first two variables represent the global environment descriptors of a molecule (Van der Waals volume) and the other two represent the local environment descriptors of the molecule, while fixing the atom concerned.

Table1. MAM descriptors for 2,2-Dimethyl-3-pentanol

V_1	V_2	V_3	V_{11}	V_{12}	V_{13}
54.77	103.93	116,70	9.63	18.57	74,74

The second model is based on the use of electro-topological descriptors of Kier and hall [17]. The molecular shape indices of Kier and Hall kappa aims to compare the molecular graph for minimum and maximum and intended to exploit the different aspects of molecular form. The molecular operating environment software (MOE) was used to generate the Kier and hall descriptors.

To determine the components, a program based on an algorithm to build the connectivity matrix for all molecules by the SMILES form was used and stored as input. Some example molecules are presented in the table 2.

Table2 . Example of some molecules in SMILES

CCO	éthanol CH_3CH_2OH
CC(C)CCO	3-Methylbutan-1-ol $(CH3)_2CHCH_2CH_2OH$
OC(C)C(C)(C)C	3,3-Dimethylbutan-2-ol (CH3)3CCH(CH3)OH

3. Results And Discussion

3.1 Experimental Data set

A series of 50 Aliphatic Alcohols reported by Amic et al studied [18]. With short and long carbon chain with different category (linear and branched), primary, secondary and tertiary carbon.

The QSPR analysis was performed using the experimental solubility of aliphatic alcohols in water (expressed in lnSol) which depends on two factors: influence of hydrophobic carbon chain and hydrophilic hydroxyl group.

Table 3. The Autocorrelation Vectors and experimental water solubility Of Aliphatic Alcohols

Compound	SMILES code	Autocorrelation vector				Ln Sol		
		V_1	V_2	V_{11}	V_{12}	Exp	MLR	ANN
1	occ(c)c	39.41	46.95	11.83	9.63	0.0227	0.15	-0.17
2	oc(c)cc	39.41	46.95	9.63	25.5	0.658	0.63	0.57
3	occc(c)c	49.64	57.47	11.83	11.83	-1.168	-1.19	-1.17
4	occ(c)cc	49.93	53.14	11.83	9.63	-1.0584	-1.09	-1.13
5	oc(c)ccc	49.64	57.47	9.63	25.5	-0.6349	-0.71	-0.60
6	oc(cc)cc	49.93	53.14	9.63	23.65	-0.4861	-0.61	-0.47
7	oc(c)c(c)c	45.29	65.85	9.63	23.3	-0.405	-0.52	-0.50
8	oc(c)(c)cc	37.9	83.23	6.75	39.17	0.3386	0.33	0.33
9	oc(cc)ccc	60.16	63.66	9.63	23.65	-1.326	-1.95	-1.68
10	oc(c)(cc)cc	48.82	84.7	6.75	37.32	-0.8301	-0.81	-0.66
11	oc(c)(c)ccc	48.13	94.15	6.75	39.17	-1.1178	-1.02	-0.80
12	oc(c(c)c)cc	55.82	72.33	9.63	21.45	-1.6094	-1.77	-1.66
13	oc(c)c(c)cc	55.82	72.33	9.63	23.3	-1.6399	-1.77	-1.66
14	oc(c)(c)c(c)c	44.25	97.06	6.75	36.97	-0.851	-0.72	-0.49
15	occc(c)(c)c	48.13	94.15	11.83	11.83	-2.5903	-2.13	-2.30
16	oc(c)c(c)(c)c	44.25	97.06	9.63	20.42	-1.4106	-1.35	-1.32
17	occcc(c)c	59.87	67.7	11.83	11.83	-2.2828	-2.52	-2.42
18	oc(c)cc(c)c	55.17	81.42	9.63	25.5	-1.814	-1.98	-1.93
19	occ(cc)cc	60.46	59.57	11.83	9.63	-2.7871	-2.34	-2.35
20	oc(c)(c)cccc	58.36	104.38	6.75	39.17	-2.4734	-2.35	-2.10
21	oc(c)(cc)ccc	59.05	95.61	6.75	37.32	-2.2634	-2.16	-1.94
22	oc(cc)(cc)cc	59.73	86.41	6.75	35.48	-1.9173	-1.96	-1.76
23	oc(c)(c)c(c)cc	54.77	103.93	6.75	36.97	-2.0025	-1.98	-1.71
24	oc(c)(cc)c(c)c	55.16	98.81	6.75	35.12	-1.9379	-1.87	-1.62
25	oc(c)(c)cc(c)c	53.66	118.56	6.75	39.17	-2.1456	-2.30	-2.00
26	oc(c(c)c)c(c)c	61.7	91.88	9.63	19.25	-2.8018	-2.94	-2.92
27	oc(cc)c(c)(c)c	54.77	103.93	9.63	18.57	-2.6437	-2.61	-2.65
28	oc(cc)cccc	70.39	73.89	9.63	23.65	-3.1942	-3.28	-3.02
29	oc(ccc)ccc	70.39	74.19	9.63	23.65	-3.1966	-3.29	-3.03
30	oc(c)(cc)c(c)(c)c	54.73	124.65	6.75	32.24	-2.9318	-2.59	-2.29
31	oc(c)cccccc	80.33	88.16	9.63	25.5	-4.756	-4.70	-4.57
32	occ(cc)cccc	80.92	80.33	11.83	9.63	-4.9967	-5.01	-4.63
33	oc(c)ccccccc	90.56	98.39	9.63	25.5	-6.32	-6.03	-6.13
34	oc(cc)cccccc	90.85	94.35	9.63	23.65	-6.1193	-5.94	-6.17
35	oc(ccc)ccccc	90.85	94.65	9.63	23.65	-5.9522	-5.95	-6.17
36	oc(cccc)cccc	90.85	94.65	9.63	23.65	-5.7446	-5.95	-6.17

37	oc(cc(c)c)cc(c)c	81.45	122.08	9.63	23.65	-5.7764	-5.83	-5.40
38	oc(c(c)cc)c(c)cc	82.75	104.85	9.63	19.25	-5.2983	-5.44	-5.21
39	occ(cc)(cc)ccc	80.88	99.28	11.83	6.75	-5.5728	-5.57	-4.75
40	occccccc(c)c	90.56	98.39	11.83	11.83	-5.7446	-6.51	-5.65
41	occc(c)cc(c)(c)c	74.42	135.27	11.83	11.83	-5.7699	-5.99	-5.39
42	occcc	44.11	33.88	11.83	11.83	-0.0953	0.07	-0.59
43	occccc	54.34	44.11	11.83	11.83	-1.3471	-1.26	-1.57
44	occ(c)(c)c	37.9	83.23	11.83	6.5	-0.6463	-0.78	-0.72
45	occcccc	64.57	54.34	11.83	11.83	-2.7181	-2.59	-2.83
46	oc(c)cccc	59.87	67.7	9.63	25.5	-1.9951	-2.04	-1.83
47	occccccc	74.8	64.57	11.83	11.83	-4.0745	-3.93	-4.08
48	occcccccc	85.03	74.8	11.83	11.83	-5.4015	-5.26	-5.41
49	occcccccc	95.26	85.03	11.83	11.83	-6.9078	-6.59	-6.91
50	occcccccccc	105.49	95.26	11.83	11.83	-8.2208	-7.92	-8.17

3.2 molecular descriptors Determination

In the data processing part, the descriptors were obtained by the MAM method,(table 3) and the electro-topological indices were generated by the MOE (Molecular Operating Environment (table 8) software in order to set up a model that combines these two types of descriptors for the QSPR modeling.

3.3 Statistical analysis

To determine the structure-property relationship by the four descriptors (table1) calculated by the Multifunctional (V_1, V_2) and the Modified Autocorrelation Method (V_{11}, V_{12}), we used Multiple linear regression and Artificial Neural Networks as methods of modeling. The firsts represent the global description to the environment of the molecule in space while the seconds describe the hydroxyl group in the molecule. The component V_1, takes into account the branching and the size of the molecule. it is calculated by the sum f(i)*f(j) of all the chemical interactions between all the pairs of adjacent carbon atoms in the molecule. A study has already been done to analyze the behavior of V_k components. It was based on Van Der Waals volumes (V_0 - V_5). The model provided the following statistical indicators for the solubility of aliphatic alcohols (n = 50, r = 0.97, s = 0.38) [16]. In our model, we add components that have a relationship with the description of the hydroxyl group Vik (k = 1, 2) and we add descriptors of Kiers, we observe the significant progression and we got for n=50, s=0.18; r^2=0.99.

3.4 Neural network

Artificial neural networks (ANNs) are very common for QSPR-type mathematical models that convert structural features into different properties of chemical compounds. In this study, we demonstrate that

the multifunctional autocorrelation and kier methods are effective for modeling the water solubility of aliphatic alcohols. The main benefit of using neural networks in QSPR models is their ability to offer non-linear mapping of descriptors that describes a structure-property relationship.

A three-layer (input-hidden-output) was used by the computational neural network completely connected with feed-forward network. For each structural component. The input layer contains one node. for the description n of the molecule's environment, The size of the input layer of the network is calculated by the length of the code used and until now, there are no theoretical or empirical rules to determine the number of hidden layers or the number of neurons in the layer, we must choose a number of hidden neurons that is sufficient to ensure that the information contained in the description data is correctly represented. For the hidden layer, four types of function were used: the Gaussian or a sigmoid or hyperbolic tangent function and a sigmoid or hyperbolic tangent, for the output layer. The weights of the connections between the neurons were initially assigned to the uniforms of the random values using the standard method of back propagation [20-23].

After the choice of the significant descriptors, we used the components (V_1, V_{11},) as input data. We changed the number of hidden layers up to the number 7. We examine the change of the RMS error (RMS stands for root mean square, that is the square root of the average residual) for the total set and also for both the training and the test sets. Training was stopped when there was no further improvement in the test set RMS error. We also obtained the correlation coefficient between the observed and predicted data. (s= 0.27; r=0.99).

We made a comparison between the three different training algorithms to find out which one gives the best results with high performance . Table 4 generalizes the results obtained:

Table 4. Comparison of the three different training algorithms

	Levenberg-Marquardt	Bayesian Regularization	Scaled Conjugate Gradient
S	0,66	0,53	0,80
R^2	0,95	0,97	0,93

Levenberg-Marquardt

This algorithm typically requires more memory but less time. Training automatically stops when generalization stops improving, as indicated by an increase in the mean square error of the validation samples.

Bayesian Regularization

This algorithm typically requires more time, but can result in good generalization for difficult, small or noisy datasets. Training stops according to adaptive weight minimization (regularization).

Scaled Conjugate Gradient

This algorithm requires less memory. Training automatically stops when generalization stops improving, as indicated by an increase in the mean square error of the validation samples.

Based on these results, (table4) it can be concluded that Bayesian Regularization has better capability of a short-term forecast, however in the long run, it loses its accuracy and follows similar performance to that of Levenberg-Marquardt.

There are various techniques of cross-validation, and in our case the leave-20% Out technique has been applied using p observations as the validation set and the remaining observations as the training set. This is repeated on all ways to cut the original sample on a validation set 20% and training set 80%.the results are shown in the table 5.The true error S_{cv} is estimated as the average error rate on test examples:

$$\boldsymbol{Scv} = 1/\mathrm{N} \sum_{i=1}^{N} \boldsymbol{Si} \quad (3)$$

S_{cv}= 0.20 by the Multiple linear regression modeling.

S_{cv}= 0.70 by the Artificial Neural network modeling.

3.5 Variables Selection

In this study, the approach followed for determining the subset of variable selection problem is the forward stepwise regression [24,25,26], this method is simple to define. We began with no variables in the model; we selected the variable that had the highest R-Squared. We selected, at each step, the variable that increased R-Squared the most, and then we stopped adding variables when none of the remaining variables are significant. We confirmed our choice of relevant variables by AIC: Akaike Information Criterion [27- 30]. The introduction of this criterion is widely used for good modeling selection, which is the most important aspect of statistical inference; also, AIC is the basis of a paradigm for the foundations of statistics.

A formula for least squares regression type analyses:

AIC = n log(RSS) + 2K **(5)** .[13]

Where:

- RSS = Residual Sum of Squares/n,
- n = sample size,

- K is the number of model parameters.

The model containing MAM descriptors performed better than previous model, and that good statistical indicators were obtained when the two components V_{11} and V_{I2} were added. Excellent MLR regression: n = 50; s = 0.22; r^2 = 0.98 was obtained with the three components V_1, V_{11}, V_{12}. The matrix correlation (Table 2) shows that components V_{11} and V_{12} are strongly correlated (-0.97) which mean, they give the same information, this was confirmed by the forward stepwise regression (Table 5) by the selection of the most significant term to each step, the operation when added variable stops improving the model.

Table 5. Stepwise regression

	V_1	V_2	V_3	V_4
R^2	0.97	0.51	0.30	0.29
RSS	14.78	172.70	212.97	214.44
p-value	2.03E-30	1.53E-04	3.66E-02	4.44E-02

The results of Akaike information criterion Table 6 show that the (The higher the number, the better the fit. This was obtained from statistical output) that the two criterion of the model V_1, V_2, V_{11} and V_1, V_2, V_{11}; V_{12} are close: -147.83 and -144.39 respectively(table 4).we conclude that the model wasn't improved when we added the component V_{12} (r = 0.98, s = 0.23).

RMS: Root Mean Square error.

Table 6. Akaike Information Criterion

Descriptor considered	r	RMS	AIC
V_1 ; V_2	0.98	0.39	-89.65
V_1 ; V_3	0.94	0.56	-55.08
V_2 ; V_3	0.58	1.45	39.78
V_3 ; V_4	0.05	2.13	77.54
V_1 ; V_2 ; V_3	0.98	0.22	-144.83
V_1 ; V_2 ; V_3 ; V_4	0.98	0.23	-147.39

After dividing our data series into two parts 80% for calibration and 20% for validation, we obtained the results of table 7 with five distributions, we observe the significant progression of the model by contributing the edge OC with respect to the model describing just the global environment of each

molecule, and we got for n=50 , s=0.22; r=0.98. By the multiple linear regression and s=0.70; r=0.96 for Artificial Neural Network.

Table 7. Cross Validation: Leave 20%-Out

	r^2_{MLR}	r^2_{ANN}	S_{MLR}	S_{ANN}
1	0.99	0.98	0.10	0.47
2	0.99	0.94	0.13	0.74
3	0.99	0.95	0.13	0.71
4	0.98	0.97	0.29	0.64
5	0.97	0.93	0.35	0.94

3.6 Results of Electro-topological indices by Kier and Hall

We have Followed the same approach for Kier and Hall descriptors (table8) and the same techniques used in the first type of descriptors, we got the statistical indices s=0.42; r^2= 0.96 by the MLR, and s=0.44; r^2= 0.98 by ANN.

Table 8. Results of Kier Descriptors

N	Smiles Code	Kier (1, 2, 3, A1, A2, A3, flexibility)									
		K1	K 2	K 3	KA1	KA2	K3	Kflex	Exp	MLR	ANN
1	occ(c)c	5	2.25	4.00	5.00	2.25	4.00	2.25	0.0227	0.53	0.03
2	oc(c)cc	5	2.25	4.00	5.00	2.25	4.00	2.25	0.658	0.53	0.03
3	occc(c)c	6	3.20	5.33	6.00	3.20	5.33	3.20	-1.168	-0.82	-0.96
4	occ(c)cc	6	3.20	3.00	3.62	1.69	1.38	1.02	-1.0584	-0.64	-0.78
5	oc(c)ccc	6	3.20	5.33	4.62	2.31	3.60	1.78	-0.6349	-0.82	-0.96
6	oc(cc)cc	6	3.20	3.00	6.00	3.20	3.00	3.20	-0.4861	-0.64	-0.78
7	oc(c)c(c)c	6	2.22	3.00	6.00	2.22	3.00	2.22	-0.405	-0.34	-0.56
8	oc(c)(c)cc	6	1.63	5.33	6.00	1.63	5.33	1.63	0.3386	-0.34	-0.77
9	oc(cc)ccc	7	4.17	3.84	7.00	4.67	3.84	4.17	-1.326	-1.96	-1.93
10	oc(c)(cc)cc	7	2.34	2.67	7.00	2.34	2.67	2.34	-0.8301	-1.31	-1.38
11	oc(c)(c)ccc	7	2.34	6.00	7.00	2.34	6.00	2.34	-1.1178	-1.57	-1.54
12	oc(c(c)c)cc	7	3.06	2.67	7.00	3.06	2.67	3.06	-1.6094	-1.53	-1.61
13	oc(c)c(c)cc	7	3.06	2.67	7.00	3.06	2.67	3.06	-1.6399	-1.53	-1.61
14	oc(c)(c)c(c)c	7	1.85	2.67	7.00	1.85	2.67	1.85	-0.851	-1.16	-1.23
15	occc(c)(c)c	7	2.34	6.00	6.05	1.97	4.88	1.70	-2.5903	-1.57	-1.54
16	oc(c)c(c)(c)c	7	1.85	2.67	7.00	1.85	2.67	1.85	-1.4106	-1.16	-1.23
17	occcc(c)c	7	4.17	6.00	7.00	4.17	6.00	4.17	-2.2828	-2.12	-2.11
18	oc(c)cc(c)c	7	3.06	6.00	7.00	3.06	6.00	3.06	-1.814	-1.79	-1.74
19	occ(cc)cc	7	4.17	2.67	4.81	2.65	1.55	1.82	-2.7871	-1.87	-1.94
20	oc(c)(c)cccc	8	3.11	7.20	8.00	3.11	7.20	3.11	-2.4734	-2.85	-2.65
21	oc(c)(cc)ccc	8	3.11	3.67	8.00	3.11	3.67	3.11	-2.2634	-2.58	-2.35
22	oc(cc)(cc)cc	8	3.11	2.22	5.91	2.18	1.49	1.61	-1.9173	-2.47	-2.28
23	oc(c)(c)c(c)cc	8	2.52	2.81	8.00	2.52	2.81	2.52	-2.0025	-2.34	-2.15
24	oc(c)(cc)c(c)c	8	2.52	2.22	8.00	2.52	2.22	2.52	-1.9379	-2.29	-2.15
25	oc(c)(c)cc(c)c	8	2.52	7.20	8.00	2.52	7.20	2.52	-2.1456	-2.67	-2.48
26	oc(c(c)c)c(c)c	8	3.11	2.81	8.00	3.11	2.81	3.11	-2.8018	-2.52	-2.29
27	oc(cc)c(c)(c)c	8	2.52	2.81	8.00	2.52	2.81	2.52	-2.6437	-2.34	-2.15
28	oc(cc)cccc	8	5.14	5.00	8.00	5.14	5.00	5.14	-3.1942	-3.30	-3.08
29	oc(ccc)ccc	8	5.14	5.00	5.91	3.61	3.36	2.67	-3.1966	-3.30	-3.08
30	oc(c)(cc)c(c)(c)c	9	2.32	2.00	9.00	2.32	2.00	2.32	-2.9318	-3.17	-2.70
31	oc(c)cccccc	9	6.13	8.00	6.23	4.01	4.97	2.78	-4.756	-4.79	-4.41
32	occ(cc)cccc	10	7.11	8.00	7.21	4.30	5.00	3.10	-4.9967	-6.05	-5.64
33	oc(c)ccccccc	10	7.11	9.14	7.32	5.00	6.22	3.66	-6.32	-6.14	-5.77
34	oc(cc)cccccc	10	7.11	7.00	6.76	4.57	4.31	3.09	-6.1193	-5.98	-5.62

35	oc(ccc)ccccc	10	7.11	7.00	7.21	4.92	4.68	3.55	-5.9522	-5.98	-5.62
36	oc(cccc)cccc	10	7.11	7.00	10.00	7.11	7.00	7.11	-5.7446	-5.98	-5.62
37	oc(cc(c)c)cc(c)c	10	4.76	7.00	10.00	4.76	7.00	4.76	-5.7764	-5.26	-5.06
38	oc(c(c)cc)c(c)cc	10	4.76	3.11	6.27	2.81	1.74	1.76	-5.2983	-4.96	-4.25
39	occ(cc)(cc)ccc	10	7.11	9.14	7.91	5.46	6.85	4.32	-5.5728	-6.14	-5.77
40	occccccc(c)c	10	4.00	7.00	8.00	5.00	6.70	3.90	-5.7446	-5.03	-5.33
41	occc(c)cc(c)(c)c	10	4.00	7.00	9.18	3.64	6.31	3.34	-5.7699	-5.03	-5.33
42	occcc	5	4.00	4.00	5.00	4.00	4.00	4.00	-0.0953	0.00	-0.20
43	occccc	6	5.00	5.33	6.00	5.00	5.33	5.00	-1.3471	-1.37	-1.41
44	occ(c)(c)c	6	1.63	5.33	6.00	1.63	5.33	1.63	-0.6463	-0.34	-0.77
45	occcccc	7	6.00	6.00	7.00	6.00	6.00	6.00	-2.7181	-2.68	-2.70
46	oc(c)cccc	7	4.17	6.00	7.00	4.17	6.00	4.17	-1.9951	-2.12	-2.11
47	occccccc	8	7.00	7.20	8.00	7.00	7.20	7.00	-4.0745	-4.04	-3.95
48	occcccccc	9	8.00	8.00	9.00	8.00	8.00	8.00	-5.4015	-5.36	-5.14
49	occcccccc	10	9.00	9.14	10.00	9.00	9.14	9.00	-6.9078	-6.72	-6.86
50	occcccccccc	11	10.0	10.00	11.00	10.00	10.00	10.00	-8.2208	-8.05	-8.30

Forward stepwise (Table 9) is a tractable approach and it gives a good sequence of models .We observe the progression of the model by contributing the most significant variables Kier 1; Kier 2; Kier3 (K1; K2 and K3) with the lowest value of AIC (table 10) -82.42 (table 10). For the cross validation (table 11), we obtained by summing validation errors S_{cv}= 0.41 by MLR, and S_{cv}= 0.80 by ANN.

Table 9. Forward Stepwise regression of kier's descriptors

	Kier 1	Kier 2	Kier 3	KierA1	KierA2	KierA3	Kier flex
R^2	0.90	0.68	0.47	0.44	0.53	0.28	0.35
RSS	23.87	72.28	121.86	127.31	107.61	164.97	148.48
AIC	-29.98	25.42	51.54	53.73	45.33	66.69	61.42

Table 10. AIC of kier descriptors

Descriptorconsidered	r	RMS	AIC
K1;K2	0.96	0.44	-80.84
K1;K3	0.93	0.57	-54.89
K2;K3	0.69	1.21	20.37
K1;Kf	0.94	0.55	-58.15
K2;Kf	0.69	1.21	20.37
K3;Kf	0.49	1.56	45.65
K1;K2;K3	0.96	0.43	-82.42
K1;K2;Kf	0.96	0.44	-79.73
K2;K3;Kf	0.70	1.19	19.51
K1;K2;K3;KF	0.96	0.43	-81.03

Table 11. Leave 20%-Out cross validation

	r_{MLR}	r_{ANN}	S_{MLR}	S_{ANN}
1	0.96	0.95	0.38	0.78
2	0.95	0.98	0.41	0.40
3	0.99	0.96	0.28	0.67
4	0.97	0.92	0.44	1.20
5	0.94	0.97	0.51	0.94

3.7.1 Principal Component Analysis (PCA)

All the 11 descriptors encoding the 50 molecules were submitted to a Principal Component Analysis (PCA) and 11 components were obtained (Fig. 3). The objective consists in transforming the correlated variables into new variables uncorrelated from each other. It reduces the number of variables and makes the information less redundant. This transformation is defined in such a way that the first component has the largest possible variance that is, accounts for as much of the variability in the data as possible.

The first three axes, F1; F2 and F3 contributed respectively 55.74%; 23.87% and 8.42 to the total variability, and the total information 88,04%.

Table 12 summarized the Pearson correlation coefficients and the matrix provides information on the positive and negative correlations between variables; in Fig. 4, these variables (descriptors) are represented in a correlation circle.

In the correlation matrix, we have V11 and V12 are perfectly correlated (r= 0,98), the second descriptors of the Modified Autocorrelation Method is removed.

(Kier1; KierA1), (Kier2; KierA2), (Kier3;KierA3) are correlated and consequently Kier A1, KierA2, KierA3 are removed. These results confirm the choice of descriptors by forward stepwise regression and by Akaike Information Criterion.

Table 12. Pearson correlation coefficients of combination descriptors

Variables	V1	V2	V11	V12	Kier1	Kier 2	Kier 3	KierA1	KierA2	KierA3	Kierflex
V1	**1,00**										
V2	0,36	**1,00**									
V11	0,29	0,42	**1,00**								
V12	-0,26	0,36	**-0,98**	**1,00**							
Kier 1	**0,90**	0,70	0,03	-0,05	**1,00**						
Kier 2	**0,89**	0,02	0,47	-0,41	0,68	**1,00**					
Kier 3	0,68	0,19	0,39	-0,30	0,56	0,74	**1,00**				
KierA1	0,61	0,65	-0,10	0,08	**0,74**	0,47	0,44	**1,00**			
KierA2	0,80	0,01	0,45	-0,39	0,58	**0,91**	0,70	0,63	**1,00**		
KierA3	0,53	0,16	0,34	-0,24	0,42	0,60	**0,90**	0,59	0,74	**1,00**	
Kierflex	0,66	0,05	0,40	-0,34	0,45	0,80	0,62	0,67	**0,97**	0,75	**1,00**

The link of these two types of descriptors table (13) was in order to know the complementary role of the electro-topological indices kier and hall on our first model.

So we notice the obvious improvement of our model which decreased of an error from s_{MLR}= 0.22 and s_{ANN}= 0.42 to 0.17 and 0,32 by MAM descriptor's and Kier descriptor's respectively with a very low value of AIC compared to the other two models (-173.61).

The regression model developed by combining the descriptors of the MAM and the descriptors of Kier for the water solubility of alcohols is excellent. The standard deviation, 0.18, is better than those obtained for the other two data sets. It is this statistical indicator that is most important because it relates directly to the interests of the experimental scientist who wishes to do so. It is also the first report for this particular data series that includes different forms of aliphatic alcohols, as described.

Table 13. combination of MAM and KIER descriptors

	V_1	V_2	V_3	K1	K2	K3	EXP	MLR	ANN
1	39.41	46.95	11.83	5	2.25	4.00	0.0227	0.23	0.03
2	39.41	46.95	9.63	5	2.25	4.00	0.658	0.71	0.69
3	49.64	57.47	11.83	6	3.20	5.33	-1.168	-1.10	-1.17
4	49.93	53.14	11.83	6	3.20	3.00	-1.0584	-1.04	-1.27
5	49.64	57.47	9.63	6	3.20	5.33	-0.6349	-0.63	-0.13
6	49.93	53.14	9.63	6	3.20	3.00	-0.4861	-0.56	-0.45
7	45.29	65.85	9.63	6	2.22	3.00	-0.405	-0.46	-0.35
8	37.9	83.23	6.75	6	1.63	5.33	0.3386	0.29	0.18
9	60.16	63.66	9.63	7	4.17	3.84	-1.326	-1.92	-1.54
10	48.82	84.7	6.75	7	2.34	2.67	-0.8301	-0.83	-0.84
11	48.13	94.15	6.75	7	2.34	6.00	-1.1178	-1.03	-0.94
12	55.82	72.33	9.63	7	3.06	2.67	-1.6094	-1.70	-1.62
13	55.82	72.33	9.63	7	3.06	2.67	-1.6399	-1.70	-1.62
14	44.25	97.06	6.75	7	1.85	2.67	-0.851	-0.84	-1.05
15	48.13	94.15	11.83	7	2.34	6.00	-2.5903	-2.14	-2.56
16	44.25	97.06	9.63	7	1.85	2.67	-1.4106	-1.47	-1.36
17	59.87	67.7	11.83	7	4.17	6.00	-2.2828	-2.46	-2.92
18	55.17	81.42	9.63	7	3.06	6.00	-1.814	-1.88	-1.79
19	60.46	59.57	11.83	7	4.17	2.67	-2.7871	-2.31	-2.63
20	58.36	104.38	6.75	8	3.11	7.20	-2.4734	-2.31	-2.55
21	59.05	95.61	6.75	8	3.11	3.67	-2.2634	-2.15	-2.13
22	59.73	86.41	6.75	8	3.11	2.22	-1.9173	-1.89	-2.12
23	54.77	103.93	6.75	8	2.52	2.81	-2.0025	-2.03	-2.06
24	55.16	98.81	6.75	8	2.52	2.22	-1.9379	-1.88	-2.03
25	53.66	118.56	6.75	8	2.52	7.20	-2.1456	-2.36	-2.22
26	61.7	91.88	9.63	8	3.11	2.81	-2.8018	-2.87	-2.92
27	54.77	103.93	9.63	8	2.52	2.81	-2.6437	-2.66	-2.62
28	70.39	73.89	9.63	8	5.14	5.00	-3.1942	-3.26	-3.15
29	70.39	74.19	9.63	8	5.14	5.00	-3.1966	-3.27	-3.16
30	54.73	124.65	6.75	9	2.32	2.00	-2.9318	-2.76	-2.92
31	80.33	88.16	9.63	9	6.13	8.00	-4.756	-4.67	-5.61
32	80.92	80.33	11.83	10	7.11	8.00	-4.9967	-5.04	-5.01
33	90.56	98.39	9.63	10	7.11	9.14	-6.32	-6.01	-6.54
34	90.85	94.35	9.63	10	7.11	7.00	-6.1193	-5.95	-5.74
35	90.85	94.65	9.63	10	7.11	7.00	-5.9522	-5.96	-5.75
36	90.85	94.65	9.63	10	7.11	7.00	-5.7446	-5.96	-5.75
37	81.45	122.08	9.63	10	4.76	7.00	-5.7764	-5.73	-5.76

38	82.75	104.85	9.63	10	4.76	3.11	-5.2983	-5.29	-5.29
39	80.88	99.28	11.83	10	7.11	9.14	-5.5728	-5.76	-6.45
40	90.56	98.39	11.83	10	4.00	7.00	-5.7446	-5.79	-5.74
41	74.42	135.27	11.83	10	4.00	7.00	-5.7699	-6.00	-6.39
42	44.11	33.88	11.83	5	4.00	4.00	-0.0953	-0.06	-0.13
43	54.34	44.11	11.83	6	5.00	5.33	-1.3471	-1.39	-1.34
44	37.9	83.23	11.83	6	1.63	5.33	-0.6463	-0.82	-0.64
45	64.57	54.34	11.83	7	6.00	6.00	-2.7181	-2.76	-3.03
46	59.87	67.7	9.63	7	4.17	6.00	-1.9951	-1.98	-1.95
47	74.8	64.57	11.83	8	7.00	7.20	-4.0745	-4.10	-4.79
48	85.03	74.8	11.83	9	8.00	8.00	-5.4015	-5.45	-5.96
49	95.26	85.03	11.83	10	9.00	9.14	-6.9078	-6.80	-6.84
50	105.49	95.26	11.83	11	10.00	10.00	-8.2208	-8.15	-7.28

The values of experimental water solubility of Aliphatic Alcohols and predicted values are illustrated in Fig. 5 for Multiple Linear Regression and in Fig. 6 for Artificial Neural Network, which shows a strong correlation between the predicted and experimental data.

4. Conclusion

Multiple Linear Regression and Artificial Neural Network were used to construct a quantitative structure proprety relation model of 50 Aliphatic Alcohols for their water solubility; the study of this property is interesting because the toxic action of these compounds belongs primarily on it. the two modeling methods were compared and had significantly better predictive capability with a greater power, it is concluded that the solubility seems to be largely determined by the components (V_1, $V_{2,}$ and V_{11}), which represent the size and the branching of the molecule without neglecting the influence of hydroxyl group caused by C–OH, also the addition of kier and hall descriptors (Kier1, Kier2, Kier3) resulted in a good meaningful model. The efficiency of the model is demonstrated through the combination of the two types of descriptors. The high quality of QSPR model has been obtained for the property our data sets of alcohols.

Abbreviations

ANN
Artificial Neural Networks
CV
Cross Validation
MAM
Modified Autocorrelation Method

MLR
Multiple Linear Regressions
QSPR
Quantitative Structure-Property Relationship
PCA
Principal Components Analysis

Declarations

Availability of data and materials

The database described in this Data note can be freely and openly accessed. Please see Table 3 and Refs. [18] and table 8 for details to the data.

Declaration of competing interests

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). She is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs.

We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author and which has been configured to accept email from:
chaymae.jermouni@gmail.com

Funding

This study is funded by the CNRST (Centre National pour la Recherche Scientifique et Technique - Morocco) and MESRSFC (Ministère de l'Enseignement Supérieur et de la Recherche Scientifique et de la Formation des Cadres – Morocco) for their financial support.

Authors' Contributions

CJ participated in the writing of the manuscript, the statistical analysis, the development of the algorithms used and the interpretation of the results. MN was involved in the data collection process and the development of methods and algorithms. All authors have read and approved the final manuscript.

Corresponding author

Correspondence with Chaymae Jermouni

Acknowledgements : 'Not applicable'

References

[1] P. Solvents and A. Alcohols, “Solubility of Nonelectrolytes in Polar Solvents 11: Solubility of Aliphatic Alcohols in Water,” no. c.

[2] P. Gramatica and A. Sangion, “A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics and Terminology,” *J. Chem. Inf. Model.*, vol. 56, no. 6, pp. 1127–1131, 2016.

[3] D. A. Winkler, “The role of quantitative structure–activity relationships (QSAR) in biomolecular discovery.,” *Brief. Bioinform.*, vol. 3, no. 1, pp. 73–86, 2002.

[4] A. M. Kauppi, C. D. Andersson, H. A. Norberg, C. Sundin, A. Linusson, and M. Elofsson, “Inhibitors of type III secretion in Yersinia: Design, synthesis and multivariate QSAR of 2-arylsulfonylamino-benzanilides,” *Bioorganic Med. Chem.*, vol. 15, no. 22, pp. 6994–7011, 2007.

[5] K. Roy, *Advances in QSAR Modeling: Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences*, vol. 24, no. May 2017. 2017.

[6] M. Javaid and J. Cao, “Computing topological indices of probabilistic neural network,” *Neural Comput. Appl.*, vol. 30, no. 12, pp. 3869–3876, 2018.

[7] B. Ren, “Novel atomic-level-based AI topological descriptors: Application to QSPR/QSAR modeling,” *J. Chem. Inf. Comput. Sci.*, vol. 42, no. 4, pp. 858–868, 2002.

[8] Moreau G, Broto P (1980) *Nouv J Chim* 4:757–764

[9] D. H. Rouvray, “The modeling of chemical phenomena using topological indices,” *J. Comput. Chem.*, vol. 8, no. 4, pp. 470–480, 1987.

[10] S. Matsutani, “The physical meaning,” vol. 26, no. 1, pp. 5133–5143, 1993.

[11] M. Randić, “On molecular branching,” *Acta Chim. Slov.*, vol. 44, no. 1, pp. 57–77, 1997.

[12] I. . Nowell, “Molecular connectivity in structure-activity analysis,” *Endeavour*, vol. 10, no. 4, pp. 216–217, 1986.

[13] M. Chastrettep and D. Cretin, “Structure-property relationships-determination of the vapor pressure of hydrocarbons and oxygenated compounds using multifunctional autocorrelation method (MAM),” *SAR QSAR Environ. Res.*, vol. 3, no. 2, pp. 131–149, 1995.

[14] Chastrette M, Tiyal F, Peyraud JF . C R *Acad Sci Paris* Ser II 310:514–515,1990.

[15] Zakarya D, Tiyal F, Chastrette M. *J Phys Org* 6:574–582,1993.

[16] M. Nohair, D. Zakarya, and A. Berrada, “Autocorrelation method adapted to generate new atomic environments: Application for the prediction of 13-C chemical shifts of alkanes,” *J. Chem. Inf. Comput. Sci.*, vol. 42, no. 3, pp. 586–591, 2002.

[17] L. H. Hall and L. B. Kier, “The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling,” vol. 2, pp. 367–422, 2007.

[18] D. Amić, S. C. Basak, B. Lučić, S. Nikolić, and N. Trinajstić, “Structure-water solubility modeling of aliphatic alcohols using the weighted path numbers,” *SAR QSAR Environ. Res.*, vol. 13, no. 2, pp. 281–295, 2002.

[19] D. Weininger, “J. Chem. Inj Comput.,” *J. Chem. Inf. Comput. Sci.*, vol. 28, pp. 31–36, 1988.

[20] Rumelhart DE, Hinton GE, Williams RJ, 323:533–536,1986.

[21] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” pp. 1–18, 2012.

[22] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[23] S. P. Siregar and A. Wanto, “Analysis of Artificial Neural Network Accuracy Using Backpropagation Algorithm In Predicting Process (Forecasting),” *IJISTECH (International J. Inf. Syst. Technol.*, vol. 1, no. 1, p. 34, 2017.

[24] Z. Zhang, “Variable selection with stepwise and best subset approaches,” *Ann. Transl. Med.*, vol. 4, no. 7, pp. 1–6, 2016.

[25] J. Lv, Y. Liu, Z. Zhang, and J. Dai, “Factorial kriging and stepwise regression approach to identify environmental factors influencing spatial multi-scale variability of heavy metals in soils,” *J. Hazard. Mater.*, vol. 261, pp. 387–397, 2013.

[26] M. Templ, A. Kowarik, and P. Filzmoser, “Iterative stepwise regression imputation using standard and robust methods,” *Comput. Stat. Data Anal.*, vol. 55, no. 10, pp. 2793–2806, 2011.

[27] “As is well known , a fundamental difficulty in statistical analysis is the choice of an appropriate model , estimating and determining the order or dimension of a model . This is a common problem when a statistical model contains many parameters . The mai,” vol. 52, no. 3, pp. 345–370, 1987.

[28] H. Akalke, “ANew Look at the Statistical Model Identification,” vol. 19, pp. 716–723, 1974.

[29] T. Ingram and D. L. Mahler, “SURFACE : detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike Information Criterion,” pp. 416–425, 2013.

[30] L. Dirick, G. Claeskens, and B. Baesens, “An Akaike information criterion for multiple event mixture cure models,” pp. 1–28, 2014.

Figures

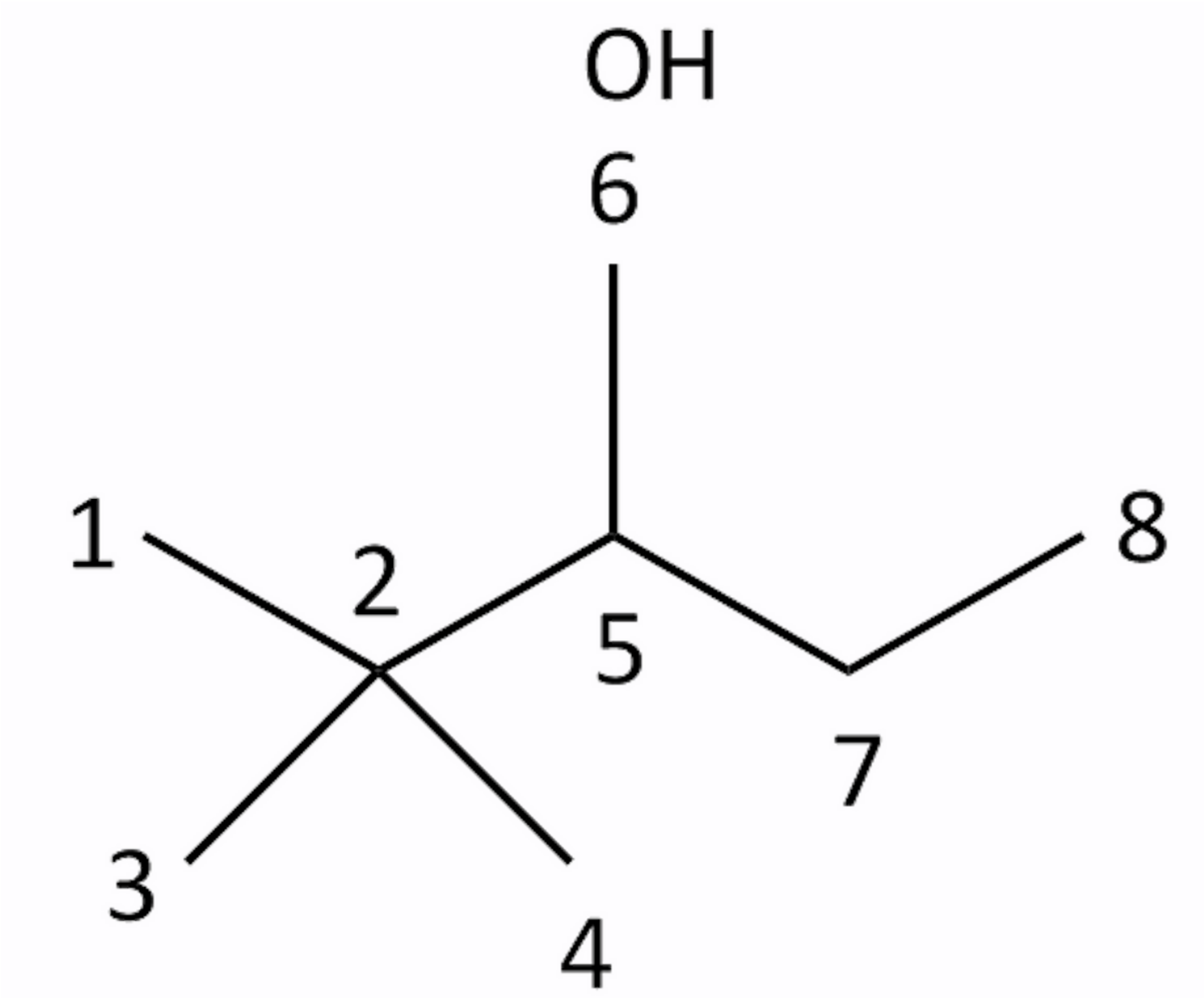

Figure 1

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \qquad \begin{pmatrix} 0 & 1 & 2 & 3 & 2 & 3 & 3 & 3 \\ 1 & 0 & 1 & 2 & 1 & 2 & 2 & 2 \\ 2 & 1 & 0 & 1 & 2 & 3 & 3 & 3 \\ 3 & 2 & 1 & 0 & 3 & 4 & 4 & 4 \\ 2 & 1 & 2 & 3 & 0 & 1 & 1 & 1 \\ 3 & 2 & 3 & 4 & 1 & 0 & 2 & 2 \\ 3 & 2 & 3 & 4 & 1 & 2 & 0 & 2 \\ 3 & 2 & 3 & 4 & 1 & 2 & 2 & 0 \end{pmatrix}$$

Figure 4

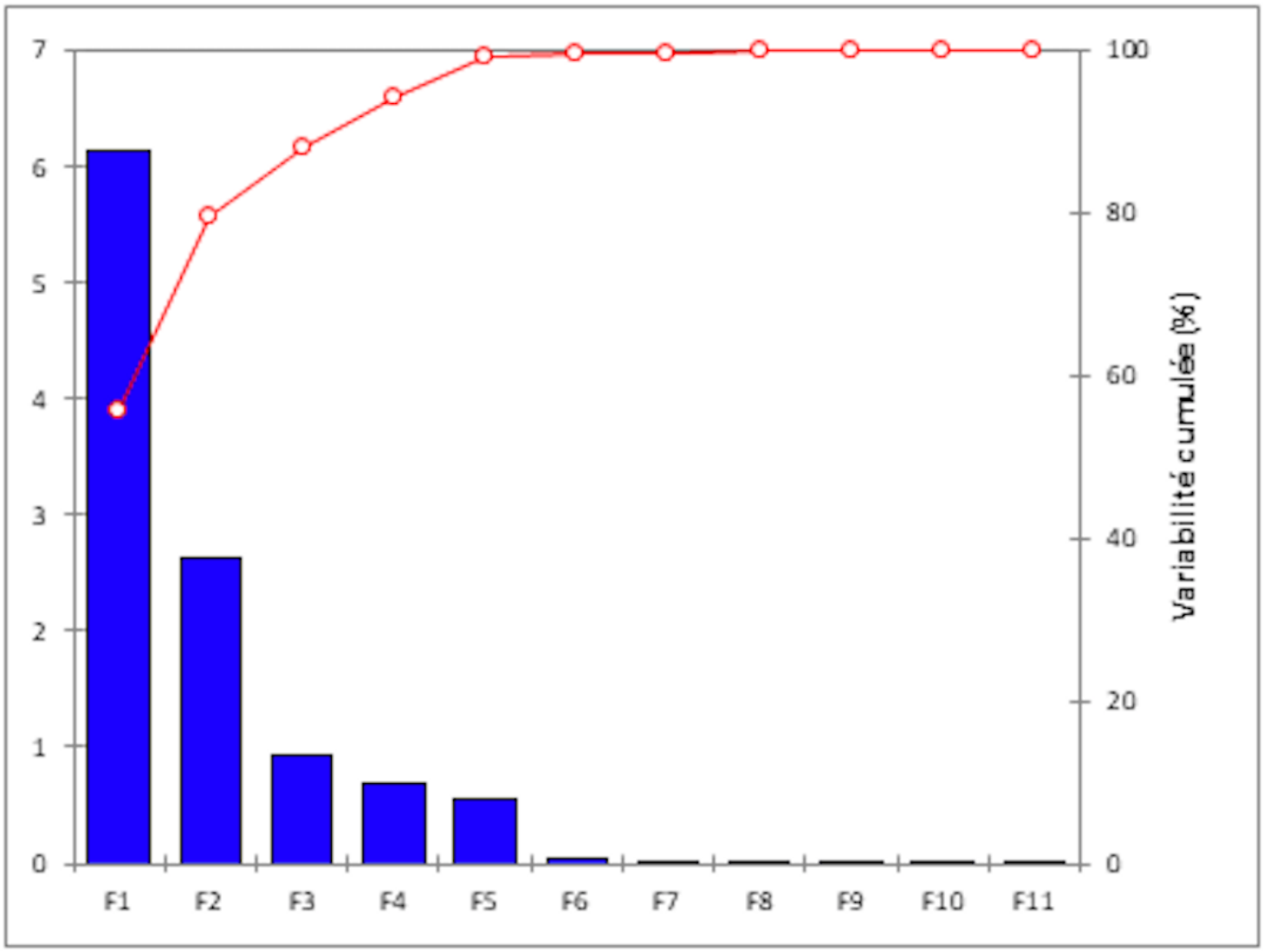

Figure 7

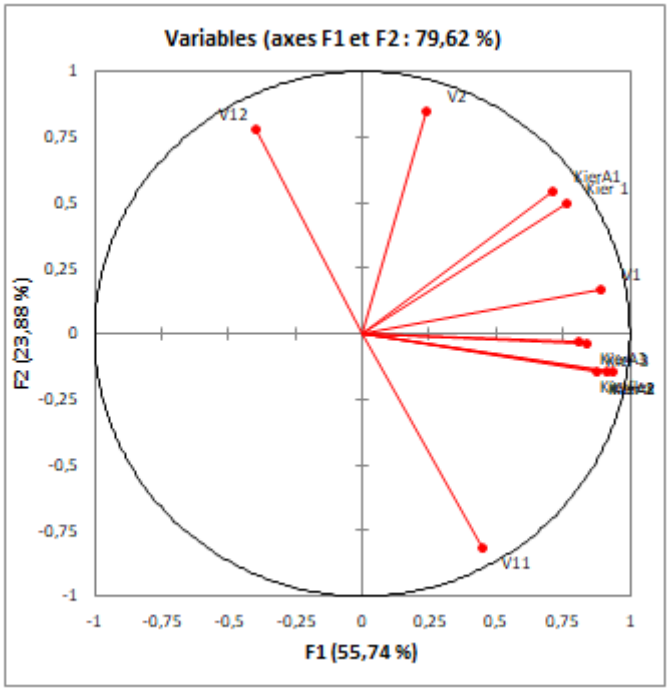
Variables (axes F1 et F2 : 79,62 %)
F2 (23,88 %)
F1 (55,74 %)
V12
V2
V1
V11
1
0,75
0,5
0,25
0
-0,25
-0,5
-0,75
-1

Figure 10

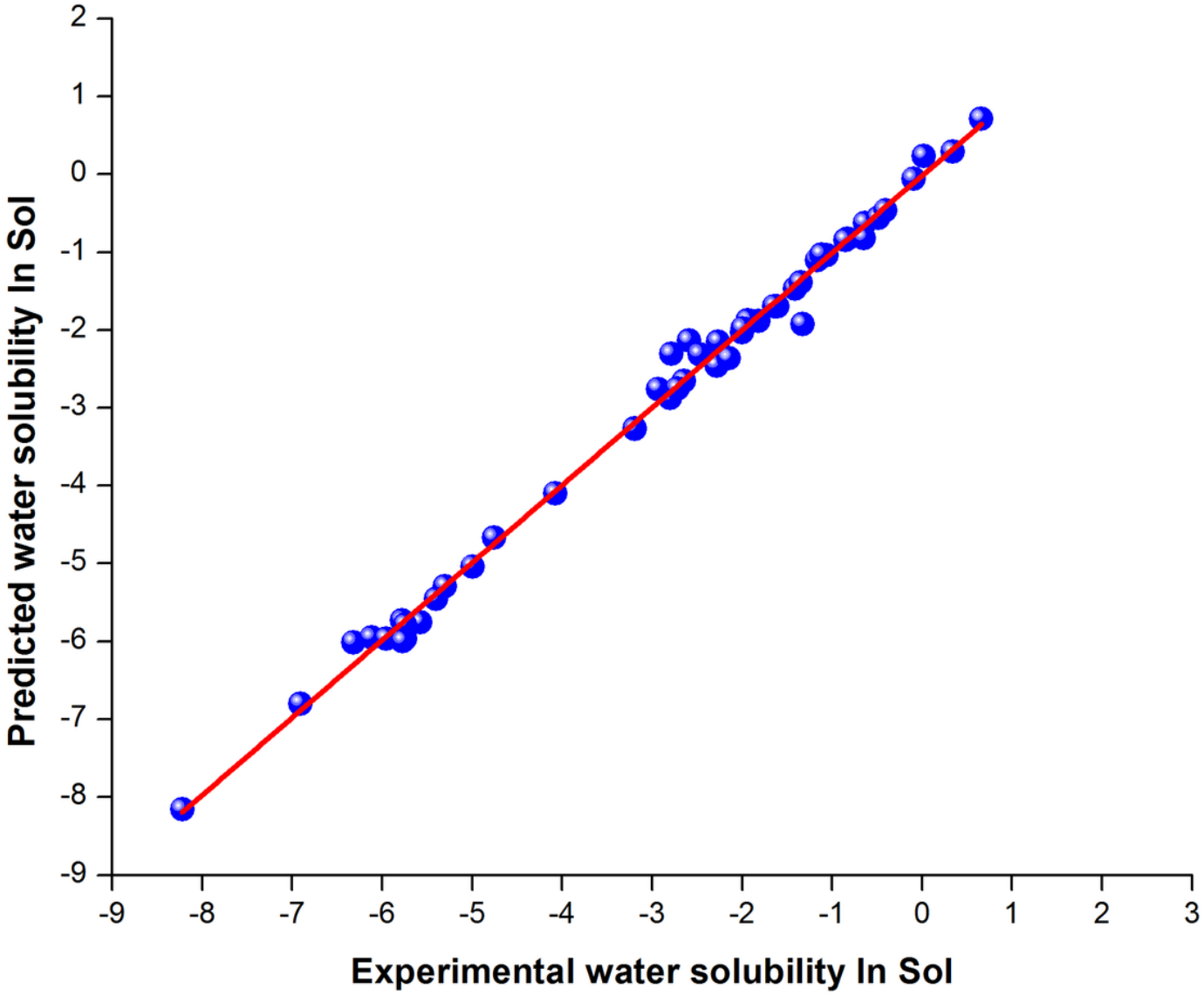
Predicted water solubility In Sol
Experimental water solubility In Sol
2
1
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
-9
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3

Figure 13

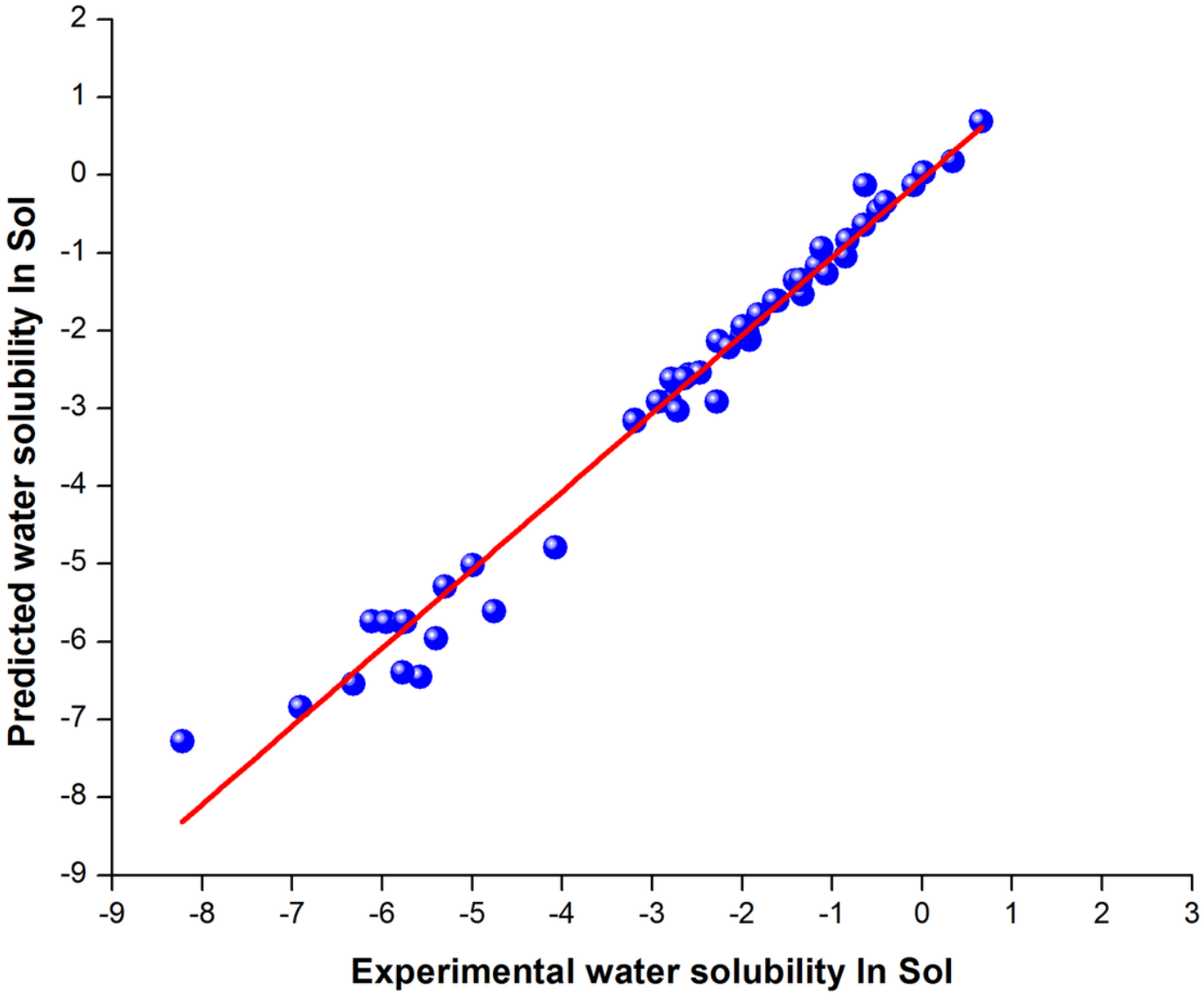

Figure 16

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- GraphicalAbstract.jpg
- GraphicalAbstract.jpg
- GraphicalAbstract.jpg