

# Research Journal of Pharmaceutical, Biological and Chemical Sciences

## Molecular Signature Descriptors Based Approach for Octanol/Water Partition Coefficient ( $K_{ow}$ ) Prediction.

Belgacem SOUYEI<sup>1,2</sup>, and Mourad KORICHI<sup>1\*</sup>.

<sup>1</sup>Université Ouargla, Fac. des Sciences Appliquées, Lab. Dynamique, Interactions et Réactivité des Systèmes, BP 511 Ouargla, Algérie

<sup>2</sup>Université d'El-Oued, Lab. Valorisation et Technologie des Ressources Sahariennes, B.P.789, El-Oued 39000 Algérie

### ABSTRACT

In this study, a new method for the prediction of octanol/water partition coefficient ( $\log k_{ow}$ ) for 123 volatile organic compounds belonging to 7 different chemical classes (hydrocarbons, alcohols, aldehydes, ketones, carboxylic acids, esters and halogen compounds) based on molecular signature descriptors. Our approach consisted of molecular descriptors calculation, then finding the correlation between the desired property ( $\log k_{ow}$ ) and the molecular structures. Both steps are facilitated by the use of signature. Application of the multi linear regression method using KNIME software led to equation for the totality of chemical classes. Such calculation gives us a model that gives results in remarkable correlations with the descriptors of these chemical classes ( $R^2 = 0.97$ ,  $SD = 0.25$ ).

**Keywords:**  $\log k_{ow}$ ; Signature descriptor; property prediction; QSAR; model. Lipophilic / hydrophilic; property estimation.

*\*Corresponding author*

## INTRODUCTION

Among the important physicochemical properties of chemical compounds, especially those being considered for drug development programs, is lipophilicity. This powerful property, which is expressed by the octanol–water partition coefficient ( $K_{ow}$ ), estimates the solubility in both aqueous and organic phases [1]. Although several methods are available for the estimation of  $K_{ow}$  of organic compounds [2,3,4], the values of  $K_{ow}$  generated using these various methods may vary by several orders of magnitude hence  $K_{ow}$  is usually expressed in the logarithmic form [5].  $\log K_{ow}$  is essential for understanding the transport mechanisms and distribution of compounds in vivo [6]. This is of particular interest to many fields including pharmacology, medicine, foods, chemical industry, fragrances, and environmental protection [7]. For example,  $\log K_{ow}$  is widely employed in predicting the specificity of organic molecules towards their target proteins, their metabolism and more commonly their efficiency in crossing cellular membranes [8-9].

The motivation to produce more sustainable and environmentally friendly chemicals that meet the consumer needs has increased considerably over the last decade [10]. Therefore, it is important to have a systematic methodology that allows the design of chemicals that possess both the consumer specified attributes and acceptable environmental characteristics. Most biological and environmental properties are structure dependent and functional group contribution techniques are not available or reliable for the determination of these properties [11]. However, a lot of work has already been done to categorize atoms or molecules systematically based on their structure and to relate these assignments to their biological activities and physicochemical properties [12].

The Quantitative Structure- Property Relationships (QSPR) is a viable tool in the determination of many properties from molecular structure information. Quantitative structure–property relationships (QSPR) remain the focus of many studies aimed at modeling and predicting the physicochemical and biological properties of molecules. A powerful tool that helps in this task is chemometrics, which uses statistical and mathematical methods to extract maximum information from data sets [13], and it has provided new insight in to the philosophy and theory behind QSPR modeling [14–15]. QSPR has received significant contributions from various research schools [16–17]. QSPR models represent powerful tools already successfully used for biological [18], toxicological [19-20], pharmaceutical [21-22] and physico-chemical applications [23-24]. It uses chemometric methods to describe how a given physicochemical property varies as a function of molecular descriptors describing the chemical structure of molecules.

The aim of the present study is to investigate the possibility of estimating the lipophilicity of organic compounds, namely  $\log K_{ow}$  of 123 volatile organic compounds from 7 different classes (hydrocarbons, alcohols, aldehydes, ketones, carboxylic acids, esters and halogen compounds).

In order to obtain a relationship suitable for structurally diverse sets, using a technique to describe molecules on a computer, called signature. This relationship is represented in QSPR model, which is a mathematical equation that relates the variation of the  $\log K_{ow}$  in a series of structurally similar compounds to the variation in their chemical structure. The molecular descriptors used in the QSPR model building are based on the concept of molecular signatures and their calculation method needs to only count the appearance of a particular signature in a molecule as a descriptor.

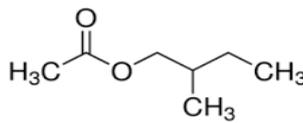
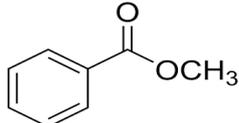
### Methodology

#### Molecular signature descriptor :

The molecular signature is a powerful molecular descriptor with previous success in molecular structure representation. It is a type of topological index that encodes all atoms in a pre-defined height  $h$  away from the root atom [25]. The molecular signature for a compound is the sum of each atomic signature multiplied by the occurrence vector of that atomic signature in the given compound and it can be calculated using the following equation [25]:

$${}^h\sigma(G) = \sum_{x \in V_G} {}^h\sigma_G(x) = {}^h\alpha_G \sum$$

Where the elements of  $V_G$  (matrix of the vertices) are the atoms (X),  ${}^h\Sigma$  is the basis set of all atomic signatures of height h, and  ${}^h\alpha_G$  is the vector of occurrence number of atomic h-signatures of graph G. Examples of molecular signatures for methyl butyl acetate and methyl benzoate are given in Figure 1.

Methyl butyl acetate	Methyl Benzoate
	
$\sigma^1(\text{C}) = [\text{C}][[\text{C}][\text{O}][\text{H}][\text{H}]] + 3[\text{C}][[\text{C}][\text{H}][\text{H}][\text{H}]] + [\text{C}][[\text{C}][\text{C}][\text{H}][\text{H}]] + [\text{C}][[\text{C}][\text{O}][\text{O}]] + [\text{C}][[\text{C}][\text{C}][\text{C}][\text{H}]] + [\text{O}][[\text{C}][\text{C}]] + [\text{O}][[\text{C}]] + 14[\text{H}][[\text{C}]]$	$\sigma^1(\text{C}) = [\text{C}][[\text{O}][\text{H}][\text{H}][\text{H}]] + [\text{C}][[\text{C}][\text{O}][\text{O}]] + 5[\text{C}][\text{p}[\text{C}]\text{p}[\text{C}][\text{H}]] + [\text{C}][\text{p}[\text{C}]\text{p}[\text{C}][\text{C}]] + [\text{O}][[\text{C}][\text{C}]] + [\text{O}][[\text{C}]] + 8[\text{H}][[\text{C}]]$

**Figure 1: Graphical depiction of the 1-Signatures for Methyl butyl acetate and Methyl Benzoate**

The aim of this study is to explore the utility and robustness of molecular signatures as descriptors in QSPR. To this end we have chosen a data set to explore the various features of signature for QSPR modeling.

The work was divided in two phases. The first devoted to the account of molecular signatures as descriptors for all compounds belonging to the database. This phase contained the calculation of unique height 1 atomic signatures Table 1 and construction of constraint equations, and their solutions are showed in Table 2 followed by molecular signatures (molecular descriptors) calculation examples Table 3. In the second phase, volatile organic compounds structures, octanol – water partition coefficient ( $\log K_{ow}$ ) experimental values and multiple linear regression analysis techniques were combined together to generate predictive quantitative structure property relationship QSPR model(s).

#### Calculation of the unique height 1 atomic signatures:

We made the calculation of unique height 1 atomic signatures of seven chemical families, and then we surveyed these signatures (similar signatures are represented by only one). The results of this survey gave us 32 atomic signatures ( $X_1$  to  $X_{32}$ ). The unique height 1 atomic signatures for the volatile organic compounds dataset are given in Table 1. As an example, the first atomic signature  $X_1$ , encodes a carbon atom to chloride atom, and two aromatic carbon atoms.

**Table 1: height 1 atomic Signature database**

The atoms ( $X_i$ )	Unique height -1 atomic signatures	The atoms ( $X_i$ )	Unique height-1 atomic signatures
$X_1$	$[\text{C}][\text{p}[\text{C}]\text{p}[\text{C}][\text{Cl}]]$	$X_{17}$	$[\text{C}][[\text{C}] = [\text{C}][\text{H}]]$
$X_2$	$[\text{C}][[\text{C}] = [\text{O}][\text{H}]]$	$X_{18}$	$[\text{C}][[\text{C}] = [\text{C}][\text{H}][\text{H}]]$
$X_3$	$[\text{C}][\text{p}[\text{C}]\text{p}[\text{C}][\text{C}]]$	$X_{19}$	$[\text{C}][[\text{C}][\text{O}] = [\text{O}]]$
$X_4$	$[\text{C}][\text{p}[\text{C}]\text{p}[\text{C}][\text{H}]]$	$X_{20}$	$[\text{C}][[\text{O}][\text{H}][\text{H}][\text{H}]]$
$X_5$	$[\text{C}][[\text{C}][\text{H}][\text{H}][\text{H}]]$	$X_{21}$	$[\text{C}][[\text{C}][\text{O}][\text{H}][\text{H}]]$
$X_6$	$[\text{C}][[\text{C}][\text{H}][\text{H}][\text{Cl}]]$	$X_{22}$	$[\text{C}][[\text{C}][\text{C}][\text{O}][\text{H}]]$
$X_7$	$[\text{C}][[\text{C}][\text{C}][\text{H}][\text{I}]]$	$X_{23}$	$[\text{C}][[\text{C}][\text{C}][\text{C}][\text{O}]]$
$X_8$	$[\text{C}][[\text{H}][\text{H}][\text{H}][\text{I}]]$	$X_{24}$	$[\text{O}][[\text{C}][\text{C}]]$
$X_9$	$[\text{C}][[\text{C}][\text{H}][\text{H}][\text{I}]]$	$X_{25}$	$[\text{O}][[\text{C}][\text{H}]]$
$X_{10}$	$[\text{C}][[\text{C}][\text{C}][\text{H}][\text{Br}]]$	$X_{26}$	$[\text{O}][[\text{C}]]$

X <sub>11</sub>	[C]([C][H][H][F])
X <sub>12</sub>	[C]([C][H][H][Br])
X <sub>13</sub>	[C]([C][C][H][H])
X <sub>14</sub>	[C]([C][C] = [O])
X <sub>15</sub>	[C]([C][C] = [C])
X <sub>16</sub>	[C]([C][C][C][H])

X <sub>27</sub>	[F]([C])
X <sub>28</sub>	[Br]([C])
X <sub>29</sub>	[I]([C])
X <sub>30</sub>	[Cl]([C])
X <sub>31</sub>	[H]([C])
X <sub>32</sub>	[H]([O])

### Construction of constraint equations:

The key feature of signature is its ability to develop a set of Diophantine (polynomial equation in which only [integer](#) solutions are allowed.) constraint equations [26]. These equations are divided in two groups: consistency equations and graphicality equation. Equations (1, 3, 5, 7, 8, 9, 10 and 11) are referred to as the consistency equations, and are created from the height 1 atomic signatures. The purpose of the consistency equations is to account for the fact that a bond in one atomic signature must appear in another atomic signature, but in the reverse order.

The second type of constraint equation is known as the graphicality equation. It is developed from the height 0 atomic signatures, and is a necessary condition for a connected graph [27]. The conditions for satisfying the graphicality equation are that the sum of vertex (root atom) degrees must be even, and the number of vertices of an odd degree must be even. The graphicality equation for the volatile organic compounds dataset is provided in equations (2, 4, 6 and 12).

$$-X_2 - X_4 - 3X_5 - 2X_6 - X_7 - 3X_8 - 2X_9 - X_{10} - 2X_{11} - 2X_{12} - 2X_{13} - X_{16} - X_{17} - 2X_{18} - 3X_{20} - 2X_{21} - X_{22} + 2X_{31} = 0 \quad \text{Eq. 1}$$

$$\text{Mod}(X_3 + X_5 + X_6 + 2X_7 + X_9 + 2X_{10} + X_{11} + X_{12} + 2X_{13} + 2X_{14} + 2X_{15} + 3X_{16} + X_{17} + X_{19} + X_{21} + 2X_{22} + 3X_{23}, 2) = 0 \quad \text{Eq. 2}$$

$$-X_{19} - X_{20} - X_{21} - X_{22} - X_{23} + 2X_{24} + X_{25} = 0 \quad \text{Eq. 3}$$

$$\text{Mod}(X_{15} + X_{17} + X_{18}, 2) = 0 \quad \text{Eq. 4}$$

$$-X_2 - X_{14} - X_{19} + 2X_{26} = 0 \quad \text{Eq. 5}$$

$$\text{Mod}(2X_1 + 2X_3 + 2X_4, 2) = 0 \quad \text{Eq. 6}$$

$$-X_1 - X_6 + X_{30} = 0 \quad \text{Eq. 7}$$

$$-X_7 - X_8 - X_9 + X_{29} = 0 \quad \text{Eq. 8}$$

$$-X_{10} - X_{12} - X_{19} + X_{28} = 0 \quad \text{Eq. 9}$$

$$-X_{25} + 2X_{32} = 0 \quad \text{Eq. 10}$$

$$-X_{11} + 2X_{27} = 0 \quad \text{Eq. 11}$$

$$\text{Mod}(X_1 + X_2 + X_3 + X_4 + 2X_5 + 2X_6 + 2X_7 + 2X_8 + 2X_9 + 2X_{10} + 2X_{11} + 2X_{12} + 2X_{13} + X_{14} + X_{15} + 2X_{16} + X_{17} + X_{18} + X_{19} + 2X_{20} + 2X_{21} + 2X_{22} + 2X_{23} - X_{26} - X_{27} - X_{28} - X_{29} - X_{30} - X_{31} - 2X_{32}, 2) = 0 \quad \text{Eq. 12}$$

Any solution which satisfies all the constraint equations is evaluated for fitness by means of a QSPR generated on the training set. The molecular descriptors (molecular signatures) were calculated using the values of table 2 and some calculation examples are registered on table 3.

**Table 2: Constraint equations solutions values**

The atoms (X <sub>i</sub> )	First values	Second values
X <sub>1</sub>	2	1
X <sub>2</sub>	3	2
X <sub>3</sub>	3	2
X <sub>4</sub>	1	2
X <sub>5</sub>	3	1
X <sub>6</sub>	1	2
X <sub>7</sub>	2	1
X <sub>8</sub>	2	1
X <sub>9</sub>	3	1
X <sub>10</sub>	1	1
X <sub>11</sub>	6	6

The atoms (X <sub>i</sub> )	First values	Second values
X <sub>17</sub>	1	2
X <sub>18</sub>	2	1
X <sub>19</sub>	1	1
X <sub>20</sub>	3	2
X <sub>21</sub>	3	3
X <sub>22</sub>	1	1
X <sub>23</sub>	2	1
X <sub>24</sub>	1	2
X <sub>25</sub>	8	4
X <sub>26</sub>	3	2
X <sub>27</sub>	3	3

X <sub>12</sub>	2	2	X <sub>28</sub>	4	4
X <sub>13</sub>	2	2	X <sub>29</sub>	7	3
X <sub>14</sub>	2	1	X <sub>30</sub>	3	3
X <sub>15</sub>	3	3	X <sub>31</sub>	36	28
X <sub>16</sub>	1	1	X <sub>32</sub>	4	2

**Table 3: molecular signatures calculation examples**

Molecules	Molecular Signature	Calculation
n-Propane	$\sigma^1(C) = [C]([C][C][H][H]) + 2[C]([C][H][H][H]) + 8[H]([C])$	$\sigma^1(C) = X_{13} + 2X_5 + 8X_{31}$
n-Butane	$\sigma^1(C) = 2[C]([C][C][H][H]) + 2[C]([C][H][H][H]) + 10[H]([C])$	$\sigma^1(C) = 2X_{13} + 2X_5 + 10X_{31}$
Methanol	$\sigma^1(C) = [C]([O][H][H][H]) + [O]([C][H]) + [H]([O]) + 3[H]([C])$	$\sigma^1(C) = X_{20} + X_{25} + X_{32} + 3X_{31}$
Ethanol	$\sigma^1(C) = [C]([C][O][H][H]) + [O]([C][H]) + [C]([C][H][H][H]) + 5[H]([C]) + [H]([O])$	$\sigma^1(C) = X_{21} + X_{25} + X_{32} + 5X_{31}$
Acetaldehyde	$\sigma^1(C) = [C]([C][H][H][H]) + [C]([C] = [O][H]) + [O]([C]) + 4[H]([C])$	$\sigma^1(C) = X_5 + X_2 + X_{26} + 4X_{31}$
Propanal	$\sigma^1(C) = [C]([C][H][H]) + [C]([C] = [C][H]) + [O]([C]) + 4[H]([C]) + [C]([C] = [O][H])$	$\sigma^1(C) = X_{18} + X_{17} + X_{26} + 4X_{31} + X_2$
Acetic acid	$\sigma^1(C) = [O]([C]) + 4[H]([C]) + [C]([C][O] = [O]) + [C]([C][H][H][H]) + [O]([C][H] + [H]([O]))$	$\sigma^1(C) = X_{26} + 4X_{31} + X_{19} + X_5 + X_{25} + X_{32}$
2- Butanone	$\sigma^1(C) = [C]([C][C] = [O]) + 2[C]([C][H][H][H]) + [C]([C][C][H][H]) + [O]([C]) + 8[H]([C])$	$\sigma^1(C) = X_{14} + 2X_5 + X_{13} + X_{26} + 8X_{31}$
2-Methyl Butyl Acetate	$\sigma^1(C) = [C]([C][O][H][H]) + 3[C]([C][H][H][H]) + [C]([C][O] = [O]) + [O]([C][C]) + 14[H]([C]) + [O]([C]) + [C]([C][C][H][H]) + [C]([C][C][C][H])$	$\sigma^1(C) = X_{21} + 3X_5 + X_{19} + X_{24} + X_{16} + X_{13} + X_{26} + 14X_{31}$
Bromooctane	$\sigma^1(C) = 5[C]([C][C][H][H]) + [C]([C][C][H][Br]) + 2[C]([C][H][H][H]) + [Br]([C]) + 17[H]([C])$	$\sigma^1(C) = 5X_{13} + X_{10} + 17X_{31} + 2X_5 + X_{28}$

### Multiple Linear Regressions (MLR)

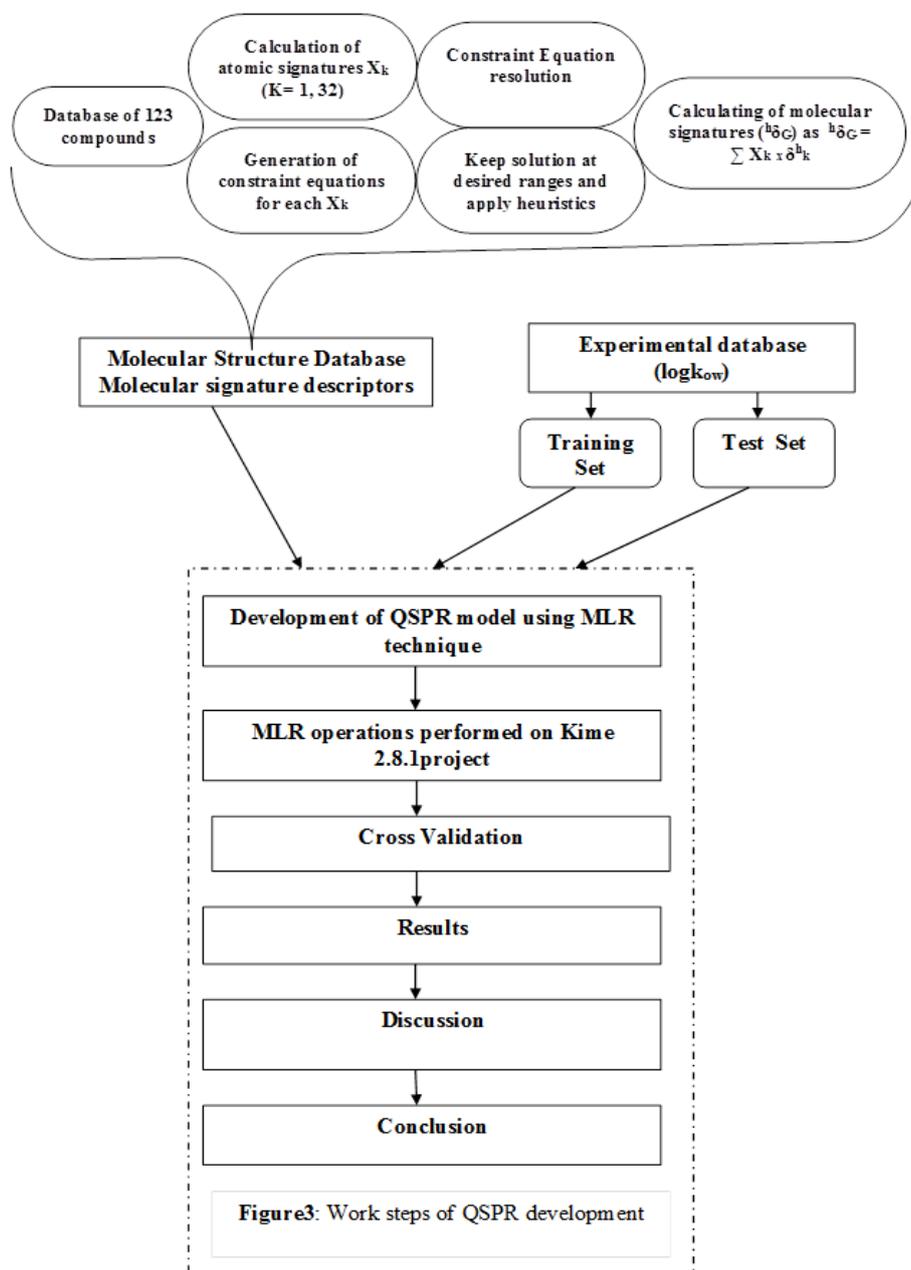
The multiple linear regression statistic technique is used to study the relationship between one dependent variable and several independent variables. It is a mathematic technique that minimizes differences between actual and predicted values. The multiple linear regression model (MLR) was generated using KNIME software which shown on Figure 2 and 3, to predict  $\log k_{ow}$  of 123 volatile organic compounds belonging to 7 different classes: (hydrocarbons (32), alcohols(27), aldehydes (11), ketones(16), carboxylic acids(15), esters(13) and halogen compounds(9)).

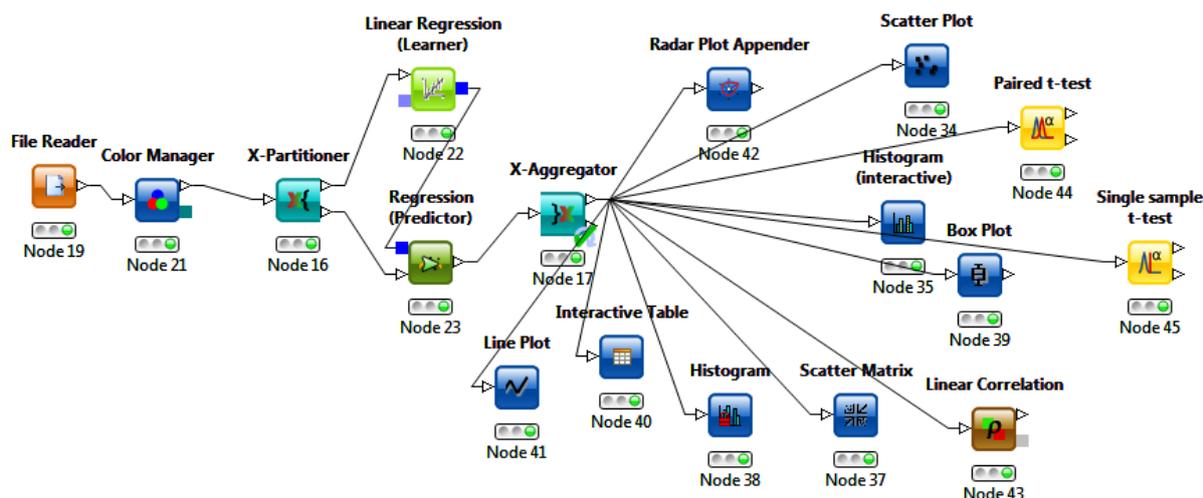
Before applying multiple linear regression (MLR), there is a preparatory phase shown in the work diagram include the following: selection of the data set from [4], calculation of the unique height1 atomic signatures  $X_k$  from the molecular graph table1, construction of constraint equations, constraint equations solutions ( The atoms ( $X_k$ ) calculation values table 2, for the calculation of molecular signatures  $^h\delta_G$  table 3.

Multiple Linear Regression was applied using KNIME software. The data base contains the values of molecular signatures as independent variables and the experimental values of  $\log k_{ow}$  as dependent variable. The calculated model for predicting  $\log k_{ow}$  using MLR was validated with random sampling cross validation. The MLR model is represented by following equation:

$$\text{Log}K_{ow} = b + \sum_{i,j=1}^{\infty} a_i x_j \quad \text{Eq.13}$$

Where  $a_i$  is the regression coefficients and  $b$  is the regression constant.





**Figure 2:** Multi-linear regression Knime diagram of all modelling process

## RESULTS AND DISCUSSION

Some of molecular descriptor, which encodes the topological features of molecules, was calculated to describe their molecular structure. Forward stepwise regression routine is used to develop the linear model for the prediction of  $\log_{\text{ow}}$  using calculated molecular descriptors (molecular signatures) with random sampling cross validation method. After calculation for each class then for the totality of chemical classes the final QSPR models were developed and given in Table 4. The statistical parameters used in this work are: Correlation coefficient ( $R^2$ ), Standard Deviation (SD) and Standard Error Mean (SEM), Cross-validated coefficient ( $Q^2$ ), in the case of Eq. 13, for each class and all classes of compounds.

**Table 4:** statistical parameters ( $R^2$ ,  $Q^2$ , SD, SEM) in the case of eq. 13.

Class	N°	Training			Validation		
		$R^2$	SD	SEM	$Q^2$	SD	SEM
Alcohols	27	0.95	0.472	0.090	0.95	0.610	0.298
Hydrocarbones	32	0.99	0.215	0.038	0.99	0.631	0.0024
Carboxylic acides	15	0.99	0.186	0.040	0.99	0.717	0.0127
Aldehydes	11	0.84	0.559	0.168	0.88	1.954	0.2969
Esters	13	0.79	0.716	0.198	0.87	1.255	0.1041
Ketones	16	0.97	0.366	0.009	0.98	1.418	0.0412
Halogen compounds	9	0.57	1.426	0.475	0.81	2.630	1.0980
All compounds	123	0.97	0.256	0.00071	0.99	0.147	0.00087

Based on the above calculated statistical parameters, we observe that there is a difference between the correlation values of these chemicals classes. The compounds which have the nearest  $R^2$  values to the unity show a strong correlation between measured and predicted  $\log_{\text{kow}}$ s. But those having the lowest values my present weak or eventually no correlation. In fact, carboxylic acids and hydrocarbons have the highest  $R^2$  values, but for halogen compounds and esters have the lowest ones. This difference is due to the chemical structure of the compounds for each class, especially by the presence of the COO- group and dipole-dipole interactions in esters. As for the halogen compounds the high electronegativity of the atoms due to their high effective nuclear charge and to the polarity of the C-X bond induces heterolytic mechanisms.

**Model Equation:**

$$\log K_{ow} = b - a_1 \times X_{32} + a_2 \times X_{31} + a_3 \times X_{30} + a_4 \times X_{29} + a_5 \times X_{28} - a_6 \times X_{26} - a_7 \times X_{24} - a_8 \times X_{23} - a_9 \times X_{22} - a_{10} \times X_{21} - a_{11} \times X_{20} - a_{12} \times X_{18} + a_{13} \times X_{17} - a_{14} \times X_{16} - a_{15} \times X_{14} - a_{16} \times X_{13} + a_{17} \times X_5 + a_{18} \times X_4 + a_{19} \times X_3 + a_{20} \times X_2$$

Where  $\log k_{ow}$  is the predicted or expected value of the dependent variable,  $X_j$  are distinct independent or predictor variables,  $b$  is the value of dependent variable when all of the independent variables  $X_j$  are equal to zero, and  $a_1$  through  $a_{20}$  are the estimated regression coefficients. Each regression coefficient represents the change in the dependent variable relative to a one unit change in the respective independent variable. The estimated regression coefficients and their values are registered on Table 5.

**Table 5: the estimated regression coefficients and their values**

Eq parameters	Values
$a_{10}$	-0.564
$a_{11}$	-0.508
$a_{12}$	0.114
$a_{13}$	0.555
$a_{14}$	0.105
$a_{15}$	-0.211
$a_{16}$	0.219
$a_{17}$	0.543
$a_{18}$	0.397
$a_{19}$	-0.034
$a_{20}$	0.031
Eq parameters	Values
$b$	0.332
$a_1$	0.007
$a_2$	0.001
$a_3$	0.101
$a_4$	0.108
$a_5$	0.097
$a_6$	-0.316
$a_7$	1.517
$a_8$	-1.962
$a_9$	-1.790

The parameters (predictor variables) of QSPR equation reflect quantitatively the well known fact that the  $\log k_{ow}$  of a compound depends on these variables together. The calculated value for the  $\log k_{ow}$  are in good agreement with those of the experimental values. The predicted values for  $\log k_{ow}$  for the compounds in the training and test sets using equation  $\log K_{ow}$  were plotted against the experimental  $\log k_{ow}$  values in Figure 4.

The histogram shown in Figure 5 displays all rows of the incoming data, the x axis of the histogram view shows ten selected binning columns (bins) of the predicted  $\log k_{ow}$  values. Whilst, the y axis indicates the aggregation values which represent the affiliation frequency of some compounds in the selected predicted  $\log k_{ow}$  bins. According to the range in the histogram, the chemical compounds that have  $\log k_{ow}$  values between 0.988 and 3.625 are the most common, accounting for 70.7 % of the total studied compounds.

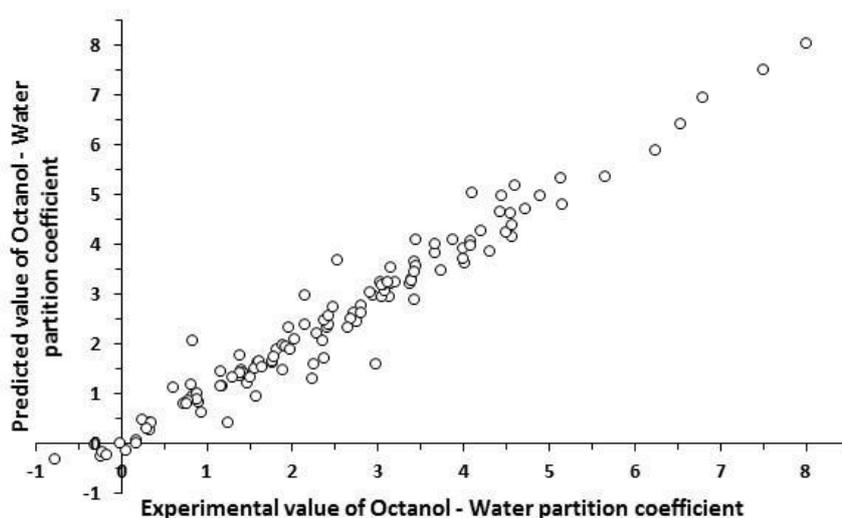


Figure 4: Shows the scatter plot of  $\log_{ow}(\text{predi})/\log_{ow}$

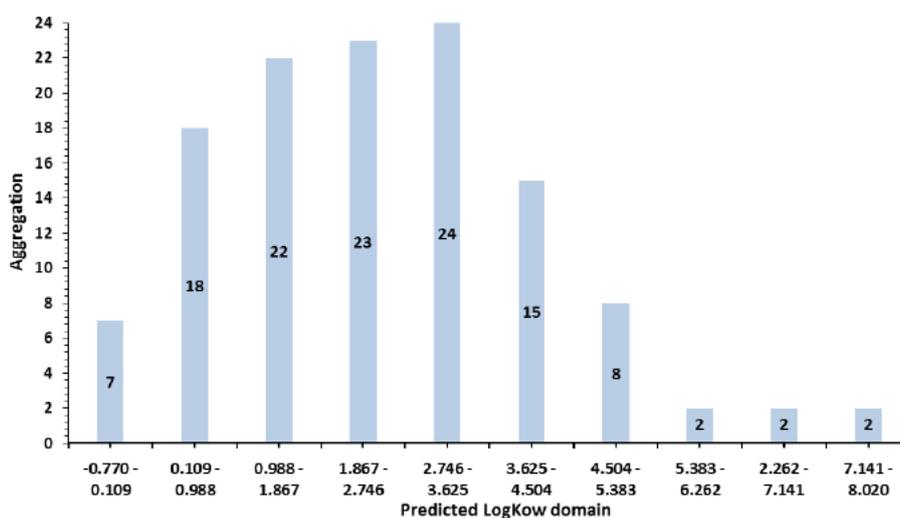


Figure 5: Histogram relating aggregation values to the predicted logkow bins

## CONCLUSION

In this work we presented the concept of atomic signature, utility of signature in QSPR study and how signature notation can be used to denote the molecular graph of a compound. We tested the height-1 signatures to use it as descriptors in a QSPR by correlating the  $\log_{ow}$  values with molecular descriptors (molecular signatures). We found a well linear correlation. The main advantage of signature versus other descriptors is its readiness for molecular structure representation.

New multi-linear QSPR model was developed to study the possibility of estimation (prediction) for the octanol-water partition coefficients of 123 volatile organic compounds belonging to seven different chemical classes using QSPR method and the concept of atomic signature. The linear relationship between  $\log_{ow}$  and the molecular descriptors (molecular signatures) with ( $R^2 = 0.97$ ,  $SD = 0.25$ ) produced a well mathematical relationship  $\log_{ow} = b + \sum(a_i \times x_j)$ . All molecular descriptors of this model can be easily calculated from the chemical structure of a molecule. The QSPR model developed in this study can provide a useful tool to predict the  $\log_{ow}$  of new compounds.

## REFERENCES

- [1] P. Isnard et S. Lambert, Estimating bioconcentration factors from octanol-water partition coefficient and aqueous solubility, 17, 1(1988) 21-34.
- [2] C.Hansch et A.Leo, Substituent Constants for Correlation Analysis in Chemistry and Biology, Wiley, New York, NY, 1979.
- [3] W.J. Lyman, W.F. Reehl et D.H. Rosenblatt, Handbook of Chemical Property Estimation Methods: Environmental Behavior of Organic Compounds, American Chemical Society, D.C. Washington, USA, 1990.
- [4] E. S. Souza, L. Zaramello , C. A. Kuhnen , B. S. Junkes , R. A.Yunes et V. E. F. Heinzen, Estimating the Octanol/Water Partition Coefficient for Aliphatic Organic Compounds Using Semi-Empirical Electrotopological Index ,Int. J. Mol. Sci. 12 (2011), 7250-7264.
- [5] H. Kubinyi, QSAR: Hansch Analysis and Related Approaches, VCH, Weinheim, 1993.
- [6] F. A. L. Ribeiro et M. M. C. Ferreira, QSPR models of boiling point, octanol–water partition coefficient and retention time index of polycyclic aromatic hydrocarbons, Journal of Molecular Structure (Theochem) 663 (2003) 109–126.
- [7] F. Spafiu, A. Mischie, P. Ionita , A. Beteringhe, T. Constantinescu et A. T. Balaban , New alternatives for estimating the octanol/water partition coefficient and water solubility for volatile organic compounds using GLC data (Kovàts retention indices) 2009 (x) 174-194 .
- [8] R. P. Schwarzenbach, P.M. Gschwend et D. M. Imboden, 2nd. Environmental organic Chemistry, John Wiley & Sons, Inc: New York, 2003.
- [9] J .Sangster, Wiley Series in Solution Chemistry, John Wiley & Sons: New York. 2 (1997).
- [10] A. C. Kokossis et Yang, Future system challenges in the design of renewable bio-energy systems and the systems of sustainable biorefineries. Design for Energy and Environment, 2009.
- [11] N. Chemmangattuvalappil, A Systematic Property Based Approach for Molecular Synthesis Using Higher Order Molecular Groups and Molecular Descriptors, Alabama, December 13, 2010.
- [12] L. B. Kier et L. H. Hall, Chemometrics Series, 9: Molecular Connectivity in Structure-Activity Analysis. New york: John Wiley & Sons, 1986.
- [13] Fengping Liu, Chenzhong Cao et Bin Cheng, A Quantitative Structure-Property Relationship (QSPR) Study of Aliphatic Alcohols by the Method of Dividing the Molecular Structure into Substructure, Int. J. Mol. Sci. 12 (2011) 2448-2462.
- [14] A.R. Katritzky, R. Petrukhin et D. Tatham, Interpretation of quantitative structure-property and activity relationships, J. Chem. Inf. Comput. Sci. 41 (2001) 679–685.
- [15] A.R. Katritzky, D.A. Dobchev, S. Slavov et M. Karelson, Legitimate utilization of large descriptor pools for QSPR/QSAR models, J. Chem. Inf. Model. 48 (2008) 2207–2213.
- [16] E.J. Delgado, J.B. Alderete et A.J. Gonzalo, A simple QSPR model for predicting soil sorption coefficients of polar and nonpolar organic compounds from molecular formula, J. Chem. Inf. Comput. Sci. 43(2003) 1928–1932.
- [17] D.H. Laura, S.P. David, N. Florian et B.O. John, A study of QSPR models of solubility, melting point, and Log P, J. Chem. Inf. Model. 48(2008) 220–232.
- [18] D. A. Winkler, The role of quantitative structure - activity relationships (QSAR) in biomolecular discovery. Briefings in Bioinformatics, 3 (2002) 73-86.
- [19] M. T. D. Cronin, A. P. Worth, (Q) SARs for Predicting Effects Relating to Reproductive Toxicity. QSAR & Combinatorial Science, 27(2008) 91-100.
- [20] T. I. Netzeva, M. Pavan et A. P. Worth, Review of (Quantitative) Structure–Activity Relationships for Acute Aquatic Toxicity. QSAR & Combinatorial Science, 27 (2008) 77-90.
- [21] M. Grover, B. Singh, M. Bakshi et S. Singh, Quantitative structure-property relationships in pharmaceutical research - Part 1. Pharmaceutical Science & Technology Today, 3(2000a) 28-35.
- [22] M. Grover, B. Singh, M. Bakshi et S. Singh, Quantitative structure-property relationships in pharmaceutical research - Part 2. Pharmaceutical Science & Technology Today, 3(2000b) 50-57.
- [23] J. Dearden et A.Worth, In Silico Prediction of Physicochemical Properties: European Commission, Joint Research Centre. (2007).
- [24] A. R. Katritzky, M. Kuanar, S. Slavov, C. D. Hall, M. Karelson et I. Kahn, Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction. Chemical Reviews, 110 (2010) 5714-5789.



- [25] J. F. Faulon, D. P. Visco, Jr. et R. S. Pophale, The Signature Molecular Descriptor. 1. Extended Valence Sequences vs. Topological Indices in QSAR and QSPR studies, *J. Chem. Inf. Comput. Sci.*, 43(2003) 707 – 720.
- [26] D. Weis, J. L. Faulon, R. Le Bone et D. Visco, The Signature Molecular Descriptor. 5. The Design of Hydrofluoroether Foam Blowing Agents Using Inverse- QSAR, *Ind. Eng. Chem. Res.*, 44 (2005) 8883-8891.
- [27] P. Donald, Jr Visco, R.S. Pophale, M. D. Rintoul et J. L. Faulon, Developing a methodology for an inverse quantitative structure-activity relationship using the signature molecular descriptor, *Journal of Molecular Graphics and Modelling* 20 (2002) 429–438.