RESEARCH ARTICLE OPEN ACCESS

Can Empirical Descriptors Reliably Predict Molecular Lipophilicity ? A QSPR Study Investigation

Ouanlo OUATTARA, Thomas Sopi AFFI, Mamadou Guy-Richard KONE, Kafoumba BAMBA, Nahossé ZIAO*

Laboratoire de Thermodynamique et de Physico-Chimie du Milieu, UFR SFA, Université Nangui Abrogoua, 02 BP 801 Abidjan 02, République de Côte-d'Ivoire

ABSTRACT

Reliable prediction of lipophilicity in organic compounds involves molecular descriptors determination. In this work, the lipophilicity of a set of twenty-three molecules has been determined using up to seven various empirical descriptors. According to Quantitative Structure Property Relationship (QSPR) method, a first set of fourteen molecules was used as training set whereas a second set of nine molecules was used as test set. Calculations made with empirical descriptors, after a severe statistical analysis, have led to establish a QSPR relation able to predict molecular lipophilicity with over 95% confidence.

***Keywords**:* Lipophilicity, molecular descriptors, QSPR, statistical analysis.

I. INTRODUCTION

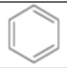
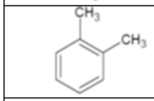
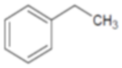
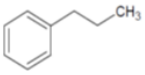
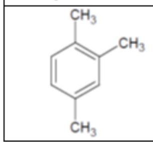
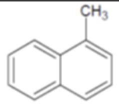
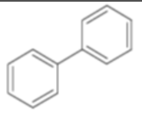
The informations contained in molecular structure can be accessed and described by the using of various physicochemical quantities named descriptors. For decades, many studies have been conducted to determine numerous descriptors of many kind, and it is well known that they actually can describe molecular structures [1-3]. The aim of our work is to determine the molecular descriptors that can reliably predict the molecular lipophilicity by empirical methods. The suitable descriptors will be selected from an initial set of seven empirical descriptors, only taking into account the ones who are highly correlated with the molecular lipophilicity while being independent one from each other in pairs. The whole process will lead to establish and validate by statistical method, a performant QSPR model.

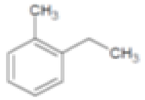
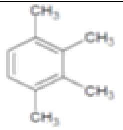
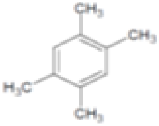
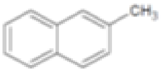
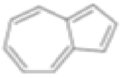
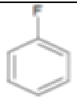
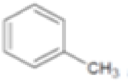
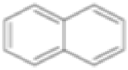
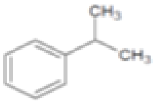
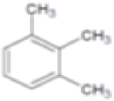
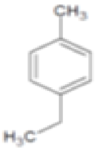
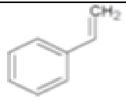
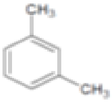
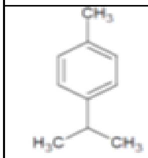
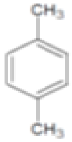
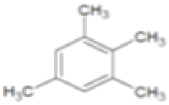
II. COMPUTATIONAL METHODS

1. Training and test sets molecules

Both training and test sets are constituted from a sample of twenty-three aromatic compounds with known experimental values [4] of molecular lipophilicity expressed as $\log P_{Exp}$, where P_{Exp} is the experimental value of octanol-water partition coefficient. The training set corresponds to fourteen molecules and the test set, to nine (Table 1). All molecules are codified CA*i*, the *i* running from 1 to 23.

Table 1 : Training set and test set samples molecules and their lipophilicities.

Training set		
Molecule	**Code**	**$\log P_{Exp}$**
	CA1	2.13 ± 0.10
	CA2	3.12 ± 0.20
	CA3	3.15 ± 0.20
	CA4	3.69 ± 0.15
	CA5	3.63 ± 0.15
	CA14	3.87 ± 0.20
Test set		
Molecule	**Code**	**$\log P_{Exp}$**
	CA15	3.98 ± 0.10

	CA6	3.53 ± 0.30
	CA7	4.00 ± 0.20
	CA8	4.10 ± 0.20
	CA9	4.00 ± 0.20
	CA10	3.22 ± 0.20
	CA11	2.27 ± 0.20
	CA12	2.73 ± 0.10
	CA13	3.35 ± 0.10
	CA16	3.66 ± 0.20
	CA17	3.60 ± 0.20
	CA18	3.63 ± 0.40
	CA19	3.05 ± 0.30
	CA20	3.20 ± 0.20
	CA21	4.10 ± 0.10
	CA22	3.15 ± 0.20
	CA23	4.10 ± 0.20

2. Computation details

Empirical descriptors have been computed using ACD/ChemSketch sofware [5]. Two other sofwares have been used, according their specificities, to perform statistical analysing of the results and to plot graphics, i.e XLSTAT [6] and EXCEL [7].

3. Statistical analysing

To establish predictive models of molecular lipophilicity, we used the method of multiple linear regression analysis [8-9] which is given by the general equation 1:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \qquad (1)$$

Y: Property studied ; $X_1, X_2, \ldots, X_p$: explanatory variables (descriptors) of the studied property ; $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$: model regression coefficients ; ε : model error ; p : number of explanatory variables. XLSTAT software directly provides these linear regression equations with the regression analysis tool. The final choice of predictive descriptors is based on two fundamental criteria for selecting descriptors sets [10]. The first criterion requires that there must be a linear dependency between the property studied, meaning here the lipophilicity, and descriptors such as $|R| \geq 0.50$. The second criterion indicates that the descriptors must be independent one of each other as $a_{ij} < 0.70$. Wherein R is the linear correlation coefficient and a_{ij} is the partial correlation coefficient between descriptors i and j. XLSTAT software directly provides these coefficients. In the case of simple linear regression [11], expressions of R and a_{ij} are given by equations 2 and 3:

$$R = \frac{cov(X,Y)}{S_X \cdot S_Y} \qquad (2) \qquad ;$$

$$a_{ij} = \frac{cov(X_i, X_j)}{var(X_i)} \qquad (3)$$

In the case of multilinear regression, the following relations 4, 5, 6 and 7 are used to calculate the statistical parameters needed to validate a model.

$$TSS = \sum \left(Y_{i,exp} - \bar{Y}_{exp}\right)^2 \qquad (4)$$

$$ESS = \sum \left(Y_{i,cal} - \bar{Y}_{exp}\right)^2 \qquad (5)$$

$$RSS = \sum \left(Y_{i,exp} - \bar{Y}_{i,cal}\right)^2 \qquad (6)$$

$$TSS = ESS + RSS \qquad (7)$$

Where TSS : Total Sum of Squares ; ESS : Extended Sum of Squares ; RSS : Residual Sum of Squares.

The determination coefficient R^2 [12] is given by the following equations 8 and 9 :

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \qquad (8) \ ; \ R = \sqrt{\frac{ESS}{TSS}} \qquad (9)$$

The relations 10 and 11 give respectively the standard deviation (s) and the adjusted coefficient of determination R^2 are :

$$s = \sqrt{\frac{RSS}{n-p-1}} \qquad (10)$$

$$R^2_{ajust} = 1 - \frac{(n - Intercept)}{n-p-1} \cdot (1 - R^2) \qquad (11)$$

To show that a linear regression equation is significant relative to another equation, we compare their Fisher's coefficients F (relation 12) [13]. The higher Fisher's coefficient is, the more significant the regression equation will be.

$$F = \frac{n-p-1}{p} \cdot \frac{ESS}{RSS} \qquad (12)$$

Where n : number of molecules ; p : number of explanatory variables.

To calculate the statistical prediction parameters of a model, we use the following relation 13 :

$$PRESS = \sum \left(Y_{i,exp} - Y_{i,pred}\right)^2 \qquad (13)$$

Where PRESS : Predictive Residual Sum of Squares. The expression of the internal cross-validation coefficient (Q^2_{LOO}) is given by equation 14:

$$Q^2_{LOO} = 1 - \frac{PRESS}{TSS} \qquad (14)$$

The external validation coefficient (Q^2_{ext}) is given by the relation 15:

$$Q^2_{ext} = 1 - \frac{n}{n_{ext}} \cdot \frac{PRESS}{TSS} \qquad (15)$$

Where n_{ext} : Number of molecules in the test set ; LOO : Leave-One-Out (Cross-validation by omission of a molecule). To show that a model is efficient in predicting a given property, we apply the five Tropsha's criteria [14-15] to this model. If at least 3/5 of the criteria are checked, then the model will be considered efficient in predicting the property studied [16]. These criteria are :

Criterion 1 : $R^2_{ext} > 0.70$; Criterion 2 : $Q^2_{ext} > 0.60$; Criterion 3 : $\frac{R^2_{ext} - R^2_0}{R^2_{ext}} < 0.10$ and $0.85 \leq k \leq 1.15$

Criterion 4 : $\frac{R^2_{ext} - R'^2_0}{R^2_{ext}} < 0.10$ and $0.85 \leq k \leq 1.15$; Criterion 5 : $|R^2_{ext} - R^2_0| \leq 0.30$

4. Molecular descriptors selection

There are numerous empirical descriptors from the literature. For our study, we considered seven empirical descriptors (Tables 2). Tables 3 gives the values of the empirical descriptors. These values were used not only to calculate the linear correlation coefficient R and the partial correlation coefficient a_{ij}, but also to establish the regression models.

Table 2 : List of seven empirical descriptors.

Empirical descriptors	Notation	Expression
Molecular volume [17]	V_M	V_M M
Molar refractivity [18]	R_M	R_M $-(n^2-1)\ .\ M$
Molar parachor [19]	P_r	P_r $_{1/4}\ M$
Molar polarizability [20-21]	P_M	P_M $-\underline{(\varepsilon_r-1)}\cdot\underline{M}$
Surface tension [22]	γ	$\gamma = \frac{F}{L}$
Molecular density	d	
Refractive index	n	

Table 3 : Values of the empirical descriptors of the training set.

CODE	$\mathbf{logP_{exp}}$	$\mathbf{V_M}$ (cm^3)	$\mathbf{R_M}$ (cm^3)	$\mathbf{P_r}$ (cm^3)	$\mathbf{P_M}$ $(10^{-24}cm^3)$	γ $(dyne \cdot cm^{-1})$	**d** $(g \cdot cm^{-3})$	**n**
CA1	2.13	89.400	26.250	207.200	10.400	28.800	0.873	1.498
CA2	3.12	121.900	35.900	282.500	14.230	28.700	0.870	1.500
CA3	3.15	122.200	35.800	283.800	14.190	29.000	0.868	1.497
CA4	3.69	138.700	40.430	323.600	16.020	29.600	0.866	1.494
CA5	3.63	138.200	40.720	320.200	16.140	28.700	0.869	1.500
CA6	3.53	138.500	40.620	321.500	16.100	29.000	0.867	1.498
CA7	4.00	154.500	45.550	357.800	18.050	28.700	0.868	1.501
CA8	4.10	154.500	45.550	357.800	18.050	28.700	0.868	1.501
CA9	4.00	139.800	48.920	348.700	19.390	38.700	1.016	1.616
CA10	3.22	123.500	44.090	311.100	17.480	40.200	1.037	1.632
CA11	2.27	93.600	26.240	214.400	10.400	27.400	1.026	1.472
CA12	2.73	105.700	31.070	244.900	12.320	28.800	0.871	1.499
CA13	3.35	123.500	44.090	311.100	17.480	40.200	1.037	1.632
CA14	3.87	139.800	48.920	348.700	19.390	38.700	1.016	1.616

Table 4 : Selection of empirical descriptors according criterion 1 [10].

Equation	**Coefficient of correlation** $\|R\|$	**Rejected descriptor if** $\|R\| < 0.5$
$logP_{exp} = f(V_M)$	0.9815	Selected
$logP_{exp} = f(R_M)$	0.9295	Selected
$logP_{exp} = f(P_r)$	0.9895	Selected
$logP_{exp} = f(P_M)$	0.9292	Selected
$logP_{exp} = f(\gamma)$	0.2943	Rejected
$logP_{exp} = f(d)$	0.0346	Rejected
$logP_{exp} = f(n)$	0.3156	Rejected

Table 5: Selection of empirical descriptors according criterion 2 [10].

Correlation between :	**coefficients** a_{ij}	**Criterion 2 : Independent descriptors if** $a_{ij} < 0{,}70$
V_M and R_M	0.3281	Independent
V_M and P_r	2.4150	Dependent
V_M and P_M	0.1300	Independent
R_M and P_r	6.3571	Dependent
R_M and P_M	0.3964	Independent
P_r and P_M	0.0577	Independent

According to Table 4, the rejected descriptors have a correlation coefficient value less than 0.50 and those selected have a coefficient greater than 0.50. The selected descriptors are V_M , R_M , P_r and P_M. The last step is to verify the criterion 2 (Tables 5). According to Table 5, it is noted that the molar parachor (P_r) depends both molecular volume (V_M) and molar refractivity (R_M), and descriptors which are themselves independent. We can exclude the molar parachor (P_r) from the list of four empirical descriptors selected by the criterion 1. The remain selected empirical descriptors are Molecular volume (V_M), Molar refractivity (R_M) and Molar polarizability (P_M).

III. RESULTS AND DISCUSSION

1. QSPR model

Fig. 1 shows that the empirical descriptors retained are linearly dependent on molecular lipophilicity. The graph of this Fig. 1 is cooresponds to the plot *Descriptors = f (logP$_{exp}$)*. Indeed, there are several descriptors for a single value of $logP_{exp}$, and it was impossible with the Excel software to plot on the same graph *logP$_{exp}$ = f (Descriptors)*. The prediction model of molecular lipophilicity established on empirical descriptors is given below :

$$logP = -0.4547 + 0.0217 \cdot V_M + 0.7689 \cdot R_M - 1.8745 \cdot P_M$$

$$n = 14 \ ; \ R = 0.9925 \ ; \ R^2 = 0.9851$$

$$s = 0.0867 \ ; \ F = 220.9188 \ ; \ FIT = 2.2877$$

According to the statistic t test (relating to the significance of the descriptors), the importance of empirical descriptors in the model is in the following descending order : $V_M > R_M > P_M$. In Table 6, the validation statistical parameters of the model are recorded. This Table 6 shows that the model has a very high predictive capacity, since 97.84% of the molecules in the test set have their lipophilicities predicted. This means that the model can be used to reliably predict the aromatic compounds unavailable lipophilicities.

Verification of Tropsha's criteria

$$(1)\ R^2_{ext} = 0.9953 > 0.70$$

$$(2)\ Q^2_{ext} = 0.9784 > 0.60$$

$$(3)\ R^2_{ext} - R^2_0 / R^2_{ext} = 0.0281 < 0.10$$

$$(4)\ |R^2_{ext} - R^2_0| = 0.028 \leq 0.30 \quad ;$$

$$(5)\ k = 1.1095 \ \text{ and } \ 0.85 < k < 1.15$$

We note that all values satisfy Tropsha's criteria. Therefore, the model is efficient in predicting the molecular lipophilicity.

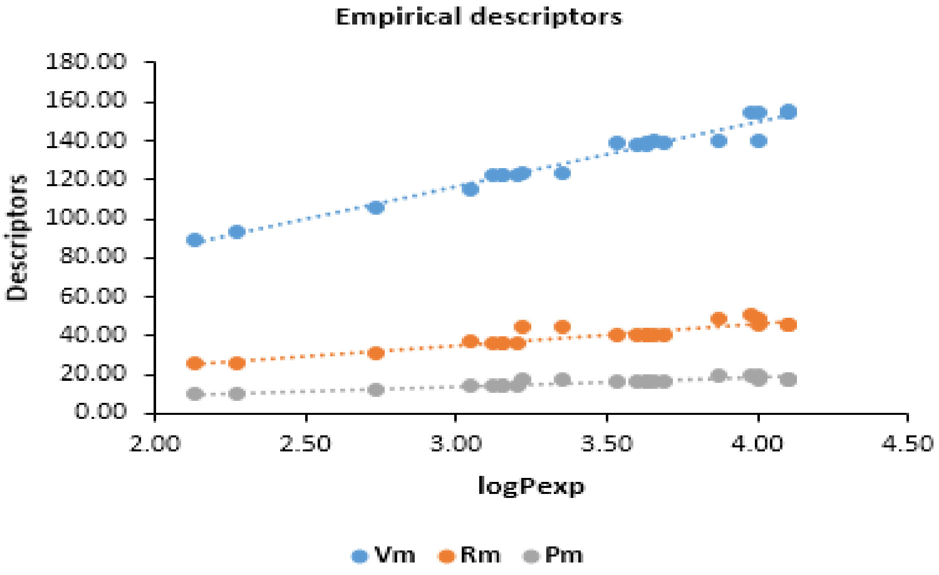


Figure 1 : Graphs $Descriptors = f(logP_{exp})$ of the model.

Table 6 : Statistical parameters of the model

Model parameters		**Internal validation LOO (Training set)**		**External validation (Test set)**	
n	14	n	14	n	9
R^2	0.9851 (98.51%)	$PRESS$	0.1810	R^2_{ext}	0.9953 (99.53%)
R^2_{ajust}	0.9807	Q^2_{LOO}	0.9642 (96.42%)	$PRESS$	0.0703
F	220.9409			Q^2_{ext}	0.9784 (97.84%)
s	0.0867	s_{press}	0.1343	s_{press}	0.1186

2. Correlation between the predicted and experimental values of lipophilicity

Fig. 2 represents the following graphs $logP_{pred}$ depending $logP_{exp}$ for internal validation (LOO) and external validation of the model. Fig. 2 shows that there is, indeed, a strong correlation between the predicted and the experimental lipophilicity according the model, since the correlation coefficient equals the high value of 0.9642. Here is the confirmation the model is highly performant in the prediction of molecular lipophilicity.

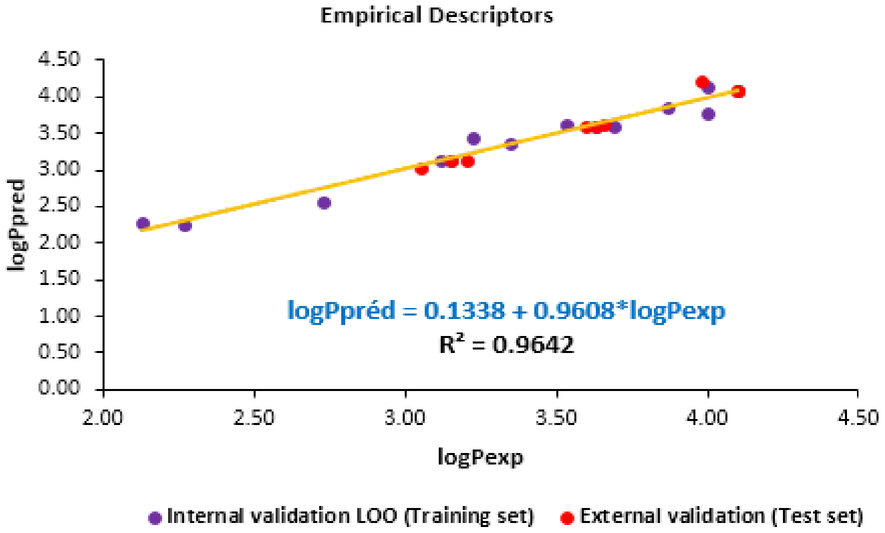


Figure 2 : Graph $logP_{pred} = f(logP_{exp})$ of the model

IV. CONCLUSION

The QSPR method was used to establish a model for molecular lipophilicity prediction. In this work, we first identified the suitable empirical descriptors in lipophilicity prediction, according to the criteria usually used for the selection of descriptors. The results showed that three empirical descriptors, i.e Molecular volume (V_M), Molar refractivity (R_M) and Molar polarizability (P_M) are strongly correlated with molecular lipophilicity. From these three descriptors, we have established a QSPR model for predicting molecular lipophilicity. Statistical parameters analysis has led to a satisfactory conclusion. Indeed, the results obtained have successfully overcome the statistical validation process, and the model has a very high predictive capacity with a coefficient of determination R^2 of 0.9953. Furthermore, the predictive validation coefficient equals 0.9642. According the model, the increasing of Molecular volume (V_M) and/or Molar refractivity (R_M) will also lead to molecular lipohilicity increasing. In the other hand, the increasing of Molar polarizability (P_M) will lead to molecular lipohilicity decreasing The establishment of a highly efficient QSPR model constitutes a noteworthy advance in molecular lipophilicity prediction.

REFERENCES

[1]. M. Karelson, Molecular Descriptors in QSAR/QSPR (Wiley, New York, 2000).

[2]. R. Todeschini, V. Consonni, Handbook of Molecular Descriptors (Wiley, 2000)

[3]. M. Karelson, V. S. Lobanov, A. R. Katritzky, Chem. Rev. 96, 1996, 1027.

[4]. Sangster Research Laboratories, (Sherbrooke ST. West, Montreal, Quebec, Canada H3G 1H7, 1989).

[5]. ACD/LogP, v. 10, Advanced Chemistry Development, Inc. (Toronto, On, Canada, 2007).

[6]. XLSTAT Version 2014.5.03 (Copyright Addinsoft 1995-2014, 2014).

[7]. Microsoft ® Excel ® 2013

[8]. P. A. Cornillon, E. M. Atzner-Lober, Régression théorie et Applications (Springer Verlag, Paris, 2007).

[9]. A. C. Rencher, G. B. Schaalje, Linear Models in Statistics (Second Edition, John Wiley & Sonc, Inc., Hoboken, New Jersey, 2008).

[10]. Vessereau A., Méthodes statistiques en biologie et en agronomie (Lavoisier, Tec & Doc, Paris, 1988).

[11]. S. Weisberg, Applied Linear Regression, (thirth Edition, John & Sonc, Inc., Hoboken, New Jersey, 2005).

[12]. S. Chatterje and A. S. Hadi, Regression Analysis by example, (fourth Edition, a John Wiley & Sonc, Inc., Hoboken, New Jersey, 2006).

[13]. Biostatistiques, E. Depiereux, G. Vincke, B. Dehertogh, (FUNDP, 27 Avril 2005).

[14]. A. Golbraikh, A. Tropsha, Beware of q^2 ! J. Mol. Graph. Model. 20, 2002, 269-276.

[15]. A. Tropsha, P. Gramatica, V. K. Gombar, QSAR Comb. Sci., 22, 2003, 69-77.

[16]. O. Ouattara and N. Ziao, Computational Chemistry, 5, 2017, 38-50.

[17]. Michael L. Connolly, J. Am. Chem. Soc. 107, 1985, 1118-1124.

[18]. M. H. Abraham, G. S. Whiting, R. M. Doherty, W. J. Shuely, J. Chem. Soc. Perkin Trans. 2, 1990, 1451-1460.

[19]. O. Exner,. Nature, vol. 196, 1 December 1962, 890 1.

[20]. R. Clausius, Die mechanische U'grmetheorie. (1879).

[21]. O. F. Mossotti, Mem. Di mathem. E fisica in Moderna (1850).

[22]. B. Bonnel, Tension superficielle et capillarité (Octobre 2006).