Computational Chemistry, 2017, 5, 38-50
http://www.scirp.org/journal/cc
ISSN Online: 2332-5984
ISSN Print: 2332-5968

Quantum Chemistry Prediction of Molecular Lipophilicity Using Semi-Empirical AM1 and *Ab Initio* HF/6-311++G Levels

Ouanlo Ouattara, Nahossé Ziao*

Laboratoire de Thermodynamique et de Physico-Chimie du Milieu, UFR SFA, Université NanguiAbrogoua,
Abidjan, Côte d'Ivoire
Email: *nahosse_ziao@yahoo.fr

How to cite this paper: Ouattara, O. and Ziao, N. (2017) Quantum Chemistry Prediction of Molecular Lipophilicity Using Semi-Empirical AM1 and *Ab Initio* HF/6-311++G Levels. *Computational Chemistry*, **5**, 38-50.
http://dx.doi.org/10.4236/cc.2017.51004

Received: September 19, 2016
Accepted: January 15, 2017
Published: January 19, 2017

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 eleven quantum various descriptors calculated by means of quantum chemistry methods. According to Quantitative Structure Property Relationship (QSPR) methods, 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 at AM1 and HF/6-311++G theories levels have led to establish a QSPR relation able to predict molecular lipophilicity with over 95% confidence.

Keywords

Molecular Lipophilicity, Molecular Descriptors, Quantum Chemistry, Statistical Analysis

1. Introduction

The informations contained in molecular structure can be accessed and described by the mean of various physicochemical quantities named descriptors. For decades, many studies have been conducted to determine empirically or compute these descriptors and it is well known that they actually can describe molecular structures [1] [2] [3]. In quantum chemistry, the computed descriptors, obviously, will be favoured. The aim of our work is to determine the molecular descriptors that can reliably predict the molecular lipophilicity by quantum chemistry methods. The suitable descriptors will be selected from an initial set of eleven, 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 methods, a performant QSPR model.

2. Computational Details

2.1. 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's coefficient. The training set corresponds to fourteen molecules and test set, nine molecules (**Table 1**). All molecules are codified CA*i*, the *i* running from 1 to 23.

2.2. Computational Theories Levels and Softwares

All molecules have been fully optimized using GAUSSIAN 03 [5] software at semi-empirical AM1 method and *ab initio* HF/6-311++G method. The basis set 6-311++G is sufficient, especially, the use of both polarization and diffuse functions is not necessary since we are not in a case of intermolecular study. Two other softwares have been used, according their specificities, to do statistical analysing of the results and to plot graphics, *i.e.* XLSTAT [6] and EXCEL [7].

2.3. Statistical Analysing

QSPR study needs a statistic analysis all along the validation process. In this work, we used the multiple linear regression analysis method [8] [9], corresponding to the below general equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Y: Property studied; $X_1, X_2, \cdots, X_p$: explanatory variables (descriptors) of the studied property; $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$: model regression coefficients. Excel 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 set, according Vessereau [10]. The first criterion requires that there must be a linear dependency between the property studied and the descriptors. For each descriptor, one must have $|R| \geq 0.50$ where R is the linear correlation coefficient. The second criterion indicates that the descriptors must be independent each from other, so we must have $a_{ij} < 0.70$ where 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:

$$R = \frac{\operatorname{cov}(X,Y)}{S_X \cdot S_Y}; \quad a_{ij} = \frac{\operatorname{cov}(X_i, X_j)}{\operatorname{var}(X_i)}$$

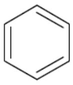
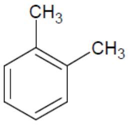
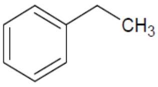
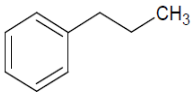
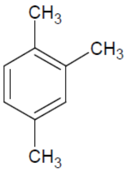
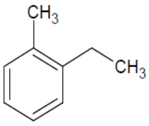
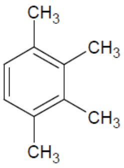
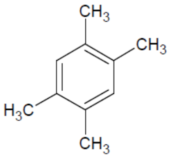
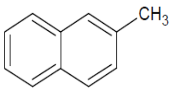
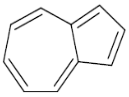
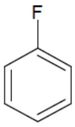
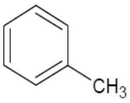
The determination coefficient R^2 [12] is given by the following equation:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \text{ and } R = \sqrt{R^2}$$

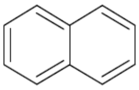
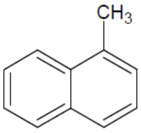
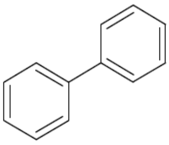
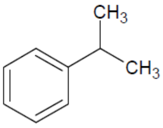
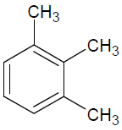
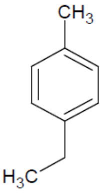
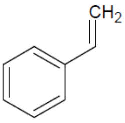
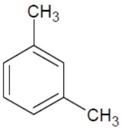
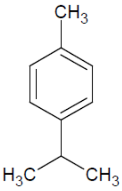
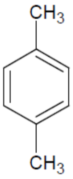
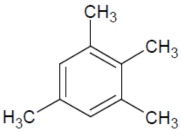
TSS: Total Sum of Squares; *ESS*: Extended Sum of Squares; *RSS*: Residual Sum of Squares. A linear regression equation significancy is drawn from Fisher's coefficient (*F*) [13]. The higher this coefficient is, the better the linear regression equation is.

$$F = \frac{n-p-1}{p} \cdot \frac{ESS}{RSS}$$

Table 1. Training setand test set samples molecules and theirli pophilicities.

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
	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

Continued

	CA13	3.35 ± 0.10
	CA14	3.87 ± 0.20
	CA15	3.98 ± 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

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

The predicting power of a model can be obtained from five Tropsha's criteria [14] [15]. If at least three of the criteria are satisfied, then the model will be considered efficient in predicting the property studied. These criteria are:

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

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

2.4. Molecular Descriptors Selection

There are thousands of molecular descriptors from the literature and quantum chemical calculations. For our study, we considered eleven quantum descriptors (**Table 2**).

Table 3 and **Table 4** give the values of the quantum descriptors at AM1 and HF/6-311++G levels respectively. These values were used to calculate correlation linear coefficient *R*, the partial coefficient correlation a_{ij} and to establish regression models. According to **Table 5**, the rejected descriptors have a correlation coefficient value less than 0.50 and those selected have a coefficient greater than 0.50. We hold the following results. At semi-empirical level, AM1, the selected descriptors are $\varepsilon_{\text{HOMO}}, \varepsilon_B, \chi$ and Q. At *ab initio* level HF/6-311++G, the selected descriptors are $\varepsilon_{\text{HOMO}}, \varepsilon_B, \chi, \eta, S, q_-$ and Q. The last step is to verify the criterion 2 (**Table 6** and **Table 7**). According to **Table 6**, the descriptors $\varepsilon_{\text{HOMO}}$ and χ are dependent. This leads us to consider two groups of descriptors at AM1 level. In the group **1**, the selected de-

Table 2. List of eleven quantum descriptors.

Quantum descriptors	Notation	Expression
Dipolar moment	μ	
Energy of the HOMO	$\varepsilon_{\text{HOMO}}$	
Energy of the LUMO	$\varepsilon_{\text{LUMO}}$	
Acidity by hydrogen bonding [16]	ε_A	$\varepsilon_A = 0.01\cdot\left[\varepsilon_{\text{LUMO}} - \varepsilon_{\text{HOMO}}\left(\text{H}_{2\text{O}}\right)\right]$
Basicity by hydrogen bonding [16]	ε_B	$\varepsilon_B = 0.01\cdot\left[\varepsilon_{\text{LUMO}(\text{H}_{2\text{O}})} - \varepsilon_{\text{HOMO}}\right]$
Chemical elecrtonegativity [17]	χ	$\chi = \frac{\varepsilon_{\text{HOMO}} - \varepsilon_{\text{LUMO}}}{2}$
Chemical hardness [17]	η	$\eta = \frac{\varepsilon_{\text{LUMO}} - \varepsilon_{\text{HOMO}}}{2}$
Chemical softness [17]	S	$S = \eta^{-1}$
Smallestnegative charge of the molecule	q_-	
Larger positive charge of the hydrogenatoms of the molecule	q_+	
Sum of absolutes values of net electrical charges of Mulliken	Q	

Table 3. Values of the training set quantum descriptors at AM1 level.

CODE	μ	$\varepsilon_{\text{HOMO}}$	$\varepsilon_{\text{LUMO}}$	ε_A	ε_B	χ	η	S	q_-	q_+	Q
CA1	0.0009	−0.3547	0.0204	0.0048	0.0052	−0.1672	0.1876	5.3319	−0.1301	0.1301	1.5614
CA2	0.4692	−0.3375	0.0192	0.0048	0.0050	−0.1592	0.1784	5.6070	−0.1775	0.1296	2.0264
CA3	0.2453	−0.3444	0.0194	0.0048	0.0051	−0.1625	0.1819	5.4975	−0.2072	0.1304	2.0942
CA4	0.2589	−0.3441	0.0192	0.0048	0.0051	−0.1625	0.1817	5.5051	−0.2119	0.1306	2.4142
CA5	0.2977	−0.3298	0.0186	0.0048	0.0049	−0.1556	0.1742	5.7405	−0.1777	0.1296	2.2557
CA6	0.4228	−0.3382	0.0195	0.0048	0.0050	−0.1594	0.1789	5.5913	−0.2067	0.1297	2.3258
CA7	0.4717	−0.3277	0.0199	0.0048	0.0049	−0.1539	0.1738	5.7537	−0.1796	0.1289	2.4884
CA8	0.0000	−0.3246	0.0182	0.0048	0.0049	−0.1532	0.1714	5.8343	−0.1760	0.1292	2.4782
CA9	0.3123	−0.3173	−0.0086	0.0045	0.0048	−0.1630	0.1544	6.4788	−0.1795	0.1321	2.3463
CA10	1.5121	−0.2948	−0.0319	0.0043	0.0046	−0.1634	0.1315	7.6075	−0.1880	0.1410	2.0976
CA11	1.5754	−0.3508	0.0060	0.0046	0.0051	−0.1724	0.1784	5.6054	−0.1657	0.1479	1.5913
CA12	0.2652	−0.3429	0.0191	0.0048	0.0051	−0.1619	0.1810	5.5249	−0.1792	0.1301	1.7976
CA13	0.0003	−0.3201	−0.0098	0.0045	0.0048	−0.1650	0.1552	6.4454	−0.1278	0.1321	2.1093
CA14	0.2741	−0.3155	−0.0098	0.0045	0.0048	−0.1627	0.1529	6.5424	−0.1811	0.1325	2.3500

Table 4. Values of the test set quantum descriptors at HF/6-311++G level.

CODE	μ	$\varepsilon_{\text{HOMO}}$	$\varepsilon_{\text{LUMO}}$	ε_A	ε_B	χ	η	S	q_-	q_+	Q
CA1	0.0000	−0.3409	0.0424	0.0055	0.0038	−0.1493	0.1917	5.2178	−0.3387	0.3387	4.0650
CA2	0.6870	−0.3202	0.0394	0.0055	0.0036	−0.1404	0.1798	5.5617	−1.6167	0.3402	13.3364
CA3	0.4144	−0.3273	0.0387	0.0055	0.0037	−0.1443	0.1830	5.4645	−1.1851	0.3616	7.9439
CA4	0.4319	−0.3266	0.0391	0.0055	0.0037	−0.1438	0.1829	5.4690	−1.1932	0.3730	9.4211
CA5	0.4248	−0.3104	0.0397	0.0055	0.0035	−0.1354	0.1751	5.7127	−1.8709	0.3375	18.3793
CA6	0.7839	−0.3187	0.0395	0.0055	0.0036	−0.1396	0.1791	5.5835	−1.6439	0.3713	15.3392
CA7	0.6708	−0.3076	0.0396	0.0055	0.0035	−0.1340	0.1736	5.7604	−1.8444	0.3554	21.3145
CA8	0.0000	−0.3026	0.0401	0.0055	0.0034	−0.1313	0.1714	5.8360	−2.8671	0.3183	24.3437
CA9	0.5046	−0.2909	0.0392	0.0055	0.0033	−0.1259	0.1651	6.0588	−1.5820	0.3619	10.2537
CA10	1.7852	−0.2624	0.0366	0.0055	0.0030	−0.1129	0.1495	6.6890	−0.5776	0.3718	6.2841
CA11	2.5200	−0.3536	0.0383	0.0055	0.0040	−0.1577	0.1960	5.1033	−0.5641	0.3546	3.6156
CA12	0.4218	−0.3274	0.0397	0.0055	0.0037	−0.1439	0.1836	5.4481	−1.3335	0.3521	8.7112
CA13	0.0000	−0.2948	0.0387	0.0055	0.0034	−0.1281	0.1668	5.9970	−0.4731	0.3394	5.4112
CA14	0.3839	−0.2894	0.0384	0.0055	0.0033	−0.1255	0.1639	6.1013	−1.7341	0.3853	10.2685

scriptors are Energy of the HOMO ($\varepsilon_{\text{HOMO}}$), Basicity by hydrogen bonding (ε_B) and Sum of absolutes values of net electrical charges of Mulliken (Q). For the group **2**, the selected descriptors are Basicity by hydrogen bonding (ε_B), Chemical electronegativity (χ) and Sum of absolutes values of net electrical charges of Mulliken (Q).

According to Table 7, the descriptors $\varepsilon_{\text{HOMO}}$ and χ are dependent. This leads us to consider two groups of descriptors for the level calculation HF/6-311++G. So, we can

Table 5. Selection of quantum descriptors according criterion 1 [10] at AM1 and HF/6-311++G levels.

Equation	**Niveau AM1**		**Niveau HF/6-311++G**	
	Correlation coefficient $\lvert R \rvert$	Rejected $\lvert R \rvert < 0.50$	Correlation coefficient $\lvert R \rvert$	Rejected $\lvert R \rvert < 0.50$
$\log P_{\exp} = f(\mu)$	0.3173	Rejected	0.3551	Rejected
$\log P_{\exp} = f(\varepsilon_{\text{HOMO}})$	0.5727	Selected	0.6186	Selected
$\log P_{\exp} = f(\varepsilon_{\text{LUMO}})$	0.1127	Rejected	0.2126	Rejected
$\log P_{\exp} = f(\varepsilon_{A})$	0.0600	Rejected	0.2126	Rejected
$\log P_{\exp} = f(\varepsilon_{B})$	0.5641	Selected	0.6186	Selected
$\log P_{\exp} = f(\chi)$	0.7228	Selected	0.6241	Selected
$\log P_{\exp} = f(\eta)$	0.3572	Rejected	0.6122	Selected
$\log P_{\exp} = f(S)$	0.2980	Rejected	0.5522	Selected
$\log P_{\exp} = f(q_{-})$	0.4134	Rejected	0.7340	Selected
$\log P_{\exp} = f(q_{+})$	0.4414	Rejected	0.1300	Rejected
$\log P_{\exp} = f(Q)$	0.9818	Selected	0.7060	Selected

Table 6. Selection of quantum descriptors according criterion 2 [10] at AM1 level.

Correlation between	AM1 level	
	Coefficient a_{ij}	Criterion 2: Independent descriptors if $a_{ij} < 0.70$
$\varepsilon_{\text{HOMO}}$ and ε_{B}	−97.3800	Independent
$\varepsilon_{\text{HOMO}}$ and χ	0.8450	Dependent
$\varepsilon_{\text{HOMO}}$ and Q	0.0269	Independent
ε_{B} and χ	−0.0078	Independent
ε_{B} and Q	−0.0003	Independent
χ and Q	0.0124	Independent

settled two groups. For the first group **3**, descriptors selected are Energy of the HOMO ($\varepsilon_{\text{HOMO}}$), Basicity by hydrogen bonding (ε_B), Chemical hardness (η), Chemical softness (S), Smallest negative charge of the molecule (q_-), Sum of absolutes values of net electrical charges of Mulliken (Q). For the last group **4**, the selected descriptors are Basicity by hydrogen bonding (ε_B), Chemical electronegativity (χ), Chemical hardness (η), Chemical softness (S), Smallest negative charge of the molecule (q_-) and Sum of absolutes values of net electrical charges of Mulliken (Q).

3. Results and Discussion

3.1. Prediction of Lipophilicity at Semi-Empirical Level AM1 (Model 1)

Figure 1 shows that the group 2 quantum descriptors retained are linearly dependent on molecular lipophilicity. The actual plot on Figure 1 is $\text{Descriptors} = f(\log P_{\exp})$. Indeed, there are several descriptors corresponding to a single value of $\log P_{\exp}$, and it has

Table 7. Selection of quantum descriptors according criterion 2 at HF/6-311++G level.

Correlation between	HF/6-311++G level	
	Coefficient a_{ij}	Criterion 2 Independent descriptorsif $a_{ij} < 0.70$
$\varepsilon_{\text{HOMO}}$ and ε_B	−100.00	Independent
$\varepsilon_{\text{HOMO}}$ and χ	2.0533	Dependent
$\varepsilon_{\text{HOMO}}$ and η	−1.9416	Independent
$\varepsilon_{\text{HOMO}}$ and S	0.0572	Independent
$\varepsilon_{\text{HOMO}}$ and q_-	−0.0065	Independent
$\varepsilon_{\text{HOMO}}$ and Q	0.0007	Independent
ε_B and χ	−0.0205	Independent
ε_B and η	0.0194	Independent
ε_B and S	−0.0006	Independent
ε_B and q_-	0.00006	Independent
ε_B and Q	−0.000007	Independent
χ and η	−0.9416	Independent
χ and S	0.0277	Independent
χ and q_-	−0.0034	Independent
χ and Q	0.0004	Independent
η and S	−0.0295	Independent
η and q_-	0.0031	Independent
η and Q	−0.0003	Independent
S and q_-	−0.0676	Independent
S and Q	0.0078	Independent
q_- and Q	−0.0985	Independent

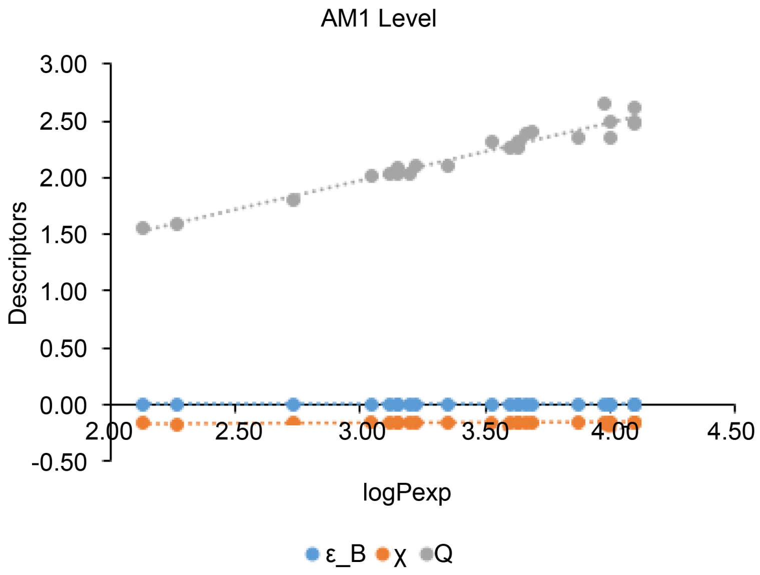


Figure 1. Graphs $\text{Descriptors} = f\left(\log P_{\exp}\right)$ at semi-empirical AM1 level.

been impossible with the software Excel to plot on a same graph $\log P_{\exp} = f\left(\text{Descriptors}\right)$.

The quantum descriptors of group **2** were used for the establishment of **Model 1** because they give a more significant regression equation in the sense of Fisher than group **1**.

Model 1:

$$\log P = 1.9891 - 417.8917 \cdot \varepsilon_B + 3.2938 \cdot \chi + 1.8490 \cdot Q$$

$$n = 14;\ R^2 = 0.9729;\ R = 0.9863;\ s = 0.1171;\ F = 119.4556;\ FIT = 1.2422$$

According to the statistical *t*_test, the importance of quantum descriptors in **Model 1** is in the following descending order: $Q > \varepsilon_B > \chi$. In Table 8 are various statistical parameters for **Model 1** validation. Table 8 shows that the **Model 1** has a very high predictive capability, since up to 95.60%, of the test molecules have their game lipophilicities predicted. This means that **Model 1** can be used to reliably predict the aromatic compounds unavailable lipophilicities.

Verification of Tropsha criteria for Model 1.

1) $R^2_{\text{ext}} = 0.9900 > 0.70$; 2) $Q^2_{\text{ext}} = 0.9560 > 0.60$; 3) $R^2_{\text{ext}} - R^2_0 / R^2_{\text{ext}} = 0.0515 < 0.10$

4) $\left|R^2_{\text{ext}} - R^2_0\right| = 0.0510 \le 0.30$; 5) $k = 1.1059$ and $0.85 < k < 1.15$

All values satisfy Tropsha's criteria. **Model 1** is retained as predictive model of molecular lipophilicity. Statistical parameters are gathered in Table 8.

3.2. Prediction of Lipophilicity at *Ab Initio* Level HF/6-311++G (Model 2)

Figure 2 shows that there is indeed a linear dependence between the quantum descriptors of group 4 and the molecular lipophilicity. The quantum descriptors of group 4 were used for the establishment of **Model 2** as they give a more significant regression equation in the sense of Fisher than group 3.

Model 2:

$$\log P = 93.8066 - 98.5843 \cdot \chi - 361.2443 \cdot \eta - 7.1577 \cdot S - 0.1749 \cdot q_- + 0.0217 \cdot Q$$

$$n = 14;\ R = 0.9340;\ R^2 = 0.8724;\ s = 0.2839;\ F = 10.9402;\ FIT = 0.1367$$

According to the statistical *t*_test, the importance of quantum descriptors in **Model 2** is in the following descending order: $\eta > S > \chi > Q > q_-$. Table 9 shows the various statistical parameters for validating the **Model 2**. Table 9 shows that the **Model 2** has a low predictive ability ($Q^2_{ext} < 0.60$), since only 59.71%, of the test molecules have their game lipophilicities predicted. This means that the **Model 2** cannot be used to reliably

Table 8. Statistical parameters of the Model 1 (Semi-empirical level AM1).

Model 1parameters		Internal validation LOO (Training set)		External validation (Test set)	
n	14	n	14	n	9
R^2	0.9729 (97.29%)	PRESS	0.3716	R^2_{ext}	0.9900 (99%)
R^2_{ajust}	0.9647	Q^2_{LOO}	0.9265 (92.65%)	PRESS	0.1429
F	119.4556			Q^2_{ext}	0.9560 (95.60%)
s	0.1171	s_{press}	0.1928	s_{press}	0.1691

predict the aromatic compounds unavailable lipophilicities.

Verification of Tropshacriteria for Model 2.

1) $R^2_{\text{ext}} = 0.4006 < 0.70$; 2) $Q^2_{ext} = 0.5971 < 0.60$;

3) $R^2_{\text{ext}} - R^2_0 / R^2_{\text{ext}} = 0.5300 > 0.10$

4) $\left| R^2_{\text{ext}} - R^2_0 \right| = 0.2123 \leq 0.30$; 5) $k = 0.3741$ and $k < 0.85$

All Tropsha criteria, excepted criterion 4, are not satisfied. **Model 2** established at HF/6-311++G level is validated, since $R^2 = 0.8724 > 0.70$, but is not efficient in predicting the lipophilicity. He is dismissed as a model for lipophilicity prediction. This unsuitable prediction of lipophilicity is certainly due to the use of an extended basic function, taking into account the diffuse functions on all atoms. The use of diffuse functions seems unefficient when calculating lipophilicity. Statistical parameters are gathered in Table 9.

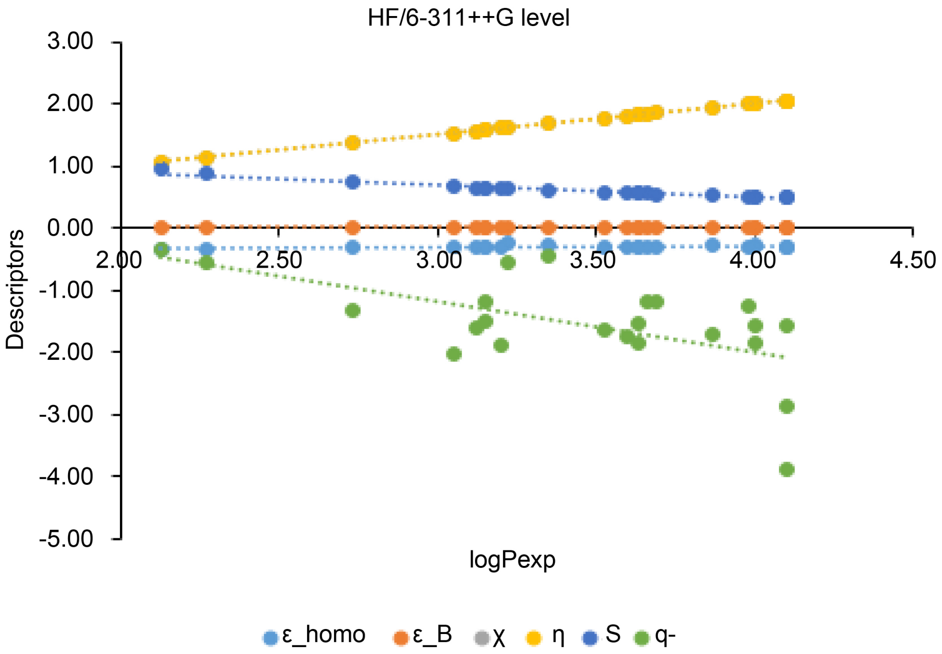


Figure 2. Graphs $\text{Descriptors} = f\left(\log P_{\text{exp}}\right)$ at *ab initio* HF/6-311++G level.

Table 9. Statistical parameters of the Model 2 (*ab initio* level HF/6-311++G).

Model 2 parameters		Internal validation LOO (Training set)		Validation externe (Test set)	
n	14	*n*	14	*n*	9
R^2	0.8724 (87.24%)	PRESS	2.5848	R^2_{ext}	0.4006 (40.06%)
R^2_{ajust}	0.6677	Q^2_{LOO}	0.4884 (48.84%)	PRESS	1.3086
F	10.9402			Q^2_{ext}	0.5971 (59.71%)
s	0.2839	s_{press}	0.5684	s_{press}	0.6605

3.3. Correlation between the Predicted and Experimental Values of Lipophilicity

Figure 3 and Figure 4 represent the following graphs $\log P_{\text{pred}}$ depending $\log P_{\text{exp}}$ for internal validation (LOO) and external of our models.

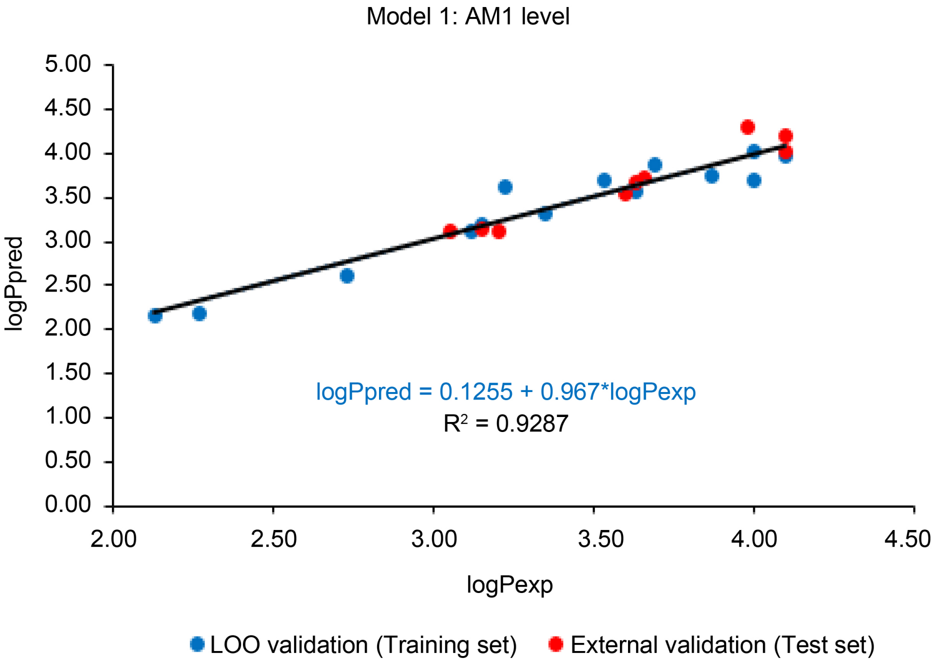


Figure 3. Graph $\log P_{\text{pred}} = f\left(\log P_{\text{exp}}\right)$ of **Model 1.**

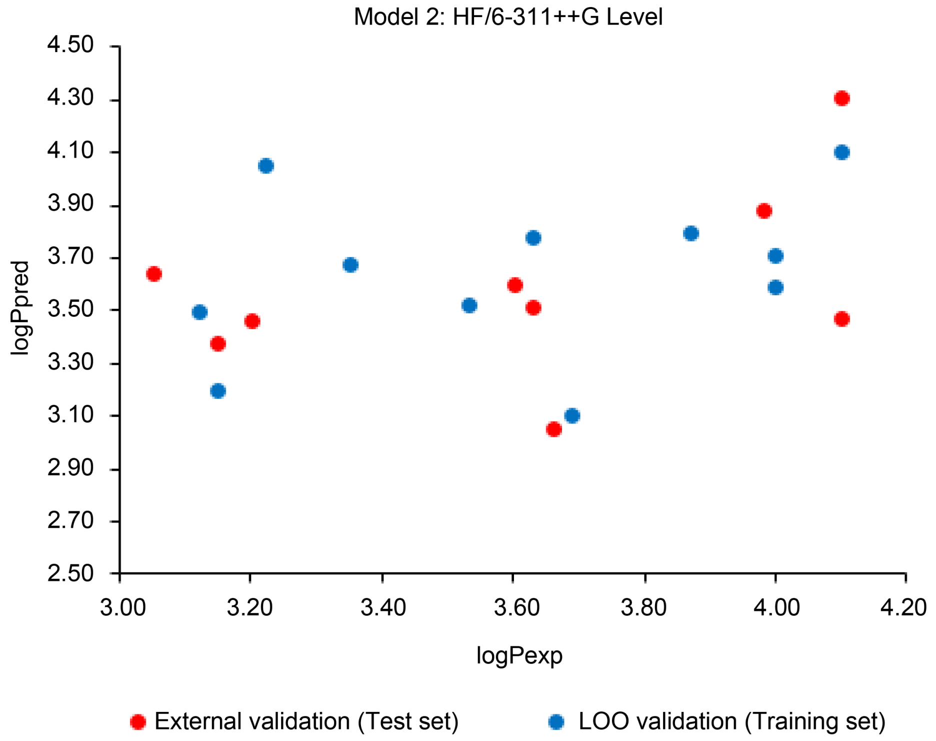


Figure 4. Graph $\log P_{\text{pred}} = f\left(\log P_{\text{exp}}\right)$ of **Model 2.**

Figure 3 shows that there is, indeed, a strong correlation between the predicted and the experimental lipophilicity according **Model 1**. The contrary is observed at Figure 4, for **Model 2**. In the latter case, it can be seen a large dispersion of the points cloud and no linear plot could be obtained. Here is the confirmation that **Model 1** is highly performant, but not **Model 2**.

4. Conclusion

QSPR methodology and quantum chemical methods were used to establish predictive models of molecular lipophilicity. In this work, we identified four groups of quantum descriptors according to the basic criteria usually used for descriptors selection. The results showed that many descriptors strongly correlate lipophilicity. From these descriptors, we have established two lipophilicity prediction models. The statistical analysis led us to select only the semi-empirical (AM1) based model. On the other hand, *ab initio* (HF/6-311++G) based model was rejected because of its low predictive power. Furthermore, the main descriptors that strongly influence the lipophilicity are, from of the selected model, the Basicity by hydrogen bonding (ε_B), Chemical electonegativity (χ) and the Sum of absolutes values of net electrical charges of Mulliken (Q). The *ab initio* based model unefficiency could be due to the use of high theory level, and tends to indicate that high theory levels, and specifically diffuse functions addition, are not suitable for molecular lipophilicity calculation. The performance of the semi-empirical based model could indicate that lipophilicity property is not strongly linked to electronic effect in molecules.

References

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

[2] Todeschini, R. and Consonni, V. (2000) Handbook of Molecular Descriptors. Wiley, Hoboken. https://doi.org/10.1002/9783527613106

[3] Karelson, M., Lobanov, V.S. and Katritzky, A.R. (1996) Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chemical Reviews*, **96**, 1027-1044. https://doi.org/10.1021/cr950202r

[4] Sangster Research Laboratories (1989) Suite M-3, 1270 Sherbrooke ST. West, Montreal, Quebec, Canada H3G 1H7. Received July 21, Revised Manuscript January 30.

[5] Gaussian 03, Revision C.01, Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Montgomery, Jr., J.A., Vreven, T., Kudin, K.N., Burant, J.C., Millam, J.M., Iyengar, S.S., Tomasi, J., Barone, V., Mennucci, B., Cossi, M., Scalmani, G., Rega, N., Petersson, G.A., Nakatsuji, H., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Klene, M., Li, X., Knox, J.E., Hratchian, H.P., Cross, J.B., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R.E., Yazyev, O., Austin, A.J., Cammi, R., Pomelli, C., Ochterski, J.W., Ayala, P.Y., Morokuma, K., Voth, G.A., Salvador, P., Dannenberg, J.J., Zakrzewski, V.G., Dapprich, S., Daniels, A.D., Strain, M.C., Farkas, O., Malick, D.K., Rabuck, A.D., Raghavachari, K., Foresman, J.B., Ortiz, J.V., Cui, Q., Baboul, A.G., Clifford, S., Cioslowski, J., Stefanov, B.B., Liu, G., Liashenko, A., Piskorz, P., Komaromi, I., Martin, R.L., Fox, D.J., Keith, T., Al-Laham, M.A., Peng, C.Y., Nanayakkara, A., Challacombe, M., Gill, P.M.W., Johnson, B., Chen, W., Wong, M.W., Gonzalez, C. and Pople, J.A. (2004) Gaussian, Inc., Wallingford.

[6] XLSTAT Version 2014.5.03 Copyright Addinsoft 1995-2014 (2014) XLSTAT and Addinsoft are Registered Trademarks of Addinsoft. https://www.xlstat.com

[7] Microsoft ® Excel ® 2013 (15.0.4420.1017) MSO (15.0.4420.1017) 64 Bits (2013) Partie de Microsoft Office Professionnel Plus.

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

[9] Rencher, A.C. and Schaalje, G.B. (2008) LinearModels in Statistics. 2nd Edition, John Wiley & Sonc, Inc., Hoboken.

[10] Vessereau, A. (1988) Méthodes statistiques en biologie et en agronomie. Lavoisier (Tec & Doc). Paris, 538 p.

[11] Weisberg, S. (2005) Applied Linear Regression. 3th Edition, John & Sonc, Inc., Hoboken.

[12] Chatterje, S. and Hadi, A.S. (2006) Regression Analysis by Example. 4th Edition, John Wiley & Sonc, Inc., Hoboken. https://doi.org/10.1002/0470055464

[13] Depiereux, E., Vincke, G. and Dehertogh, B. (2005) Biostatistics.

[14] Golbraikh, A. and Tropsha, A. (2002) Beware of q^2!. *Journal of Molecular Graphics and Modelling*, **20**, 269-276. https://doi.org/10.1016/S1093-3263(01)00123-1

[15] Tropsha, A., Gramatica, P. and Gombar, V.K. (2003) The Importance of Being Earnest: Validation Is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR & Combinatorial Science*, **22**, 69-77. https://doi.org/10.1002/qsar.200390007

[16] Abraham, M.H. (1993) Scales of Solute Hydrogen-Bonding: Their Construction and Application to Physicochemical and Biochemical Processes. *Chemical Society Reviews*, **22**, 73-83. https://doi.org/10.1039/cs9932200073

[17] Cardenas-Jiron, G.I., Gutierrez-Oliva, S., Melin, J. and Toro-Labbe, A. (1997) Relations betweenPotential Energy, ElectronicChemicalPotential, and Hardness Profiles. *The Journal of Physical Chemistry A*, **101**, 4621-4627.