Scientific Research Publishing

Computational Chemistry, 2026, 14(4), 105-130
https://www.scirp.org/journal/cc
ISSN Online: 2332-5984
ISSN Print: 2332-5968

Prediction of the Hydrophobic Character of Aromatic Amines by Quantum Chemistry and QSPR Methods: The Case of Aniline and Its Derivatives

Fatogoma Diarrassouba[1,2*], Kafoumba Bamba[1], Nahossé Ziao[1]

[1]Laboratoire de Thermodynamique et Physico-Chimie du Milieu (LTPCM), Unité de Formation et de Recherche des Sciences Fondamentales et Appliquées (UFR-SFA), Université Nangui ABROGOUA, Abidjan, Côte d'Ivoire
[2]Unité de Formation et de Recherche Sciences et Technologies (UFR-ST), Université Alassane OUATTARA, Bouaké, Côte d'Ivoire
Email: *fatogomadiarras89@gmail.com

How to cite this paper: Diarrassouba, F., Bamba, K. and Ziao, N. (2026) Prediction of the Hydrophobic Character of Aromatic Amines by Quantum Chemistry and QSPR Methods: The Case of Aniline and Its Derivatives. *Computational Chemistry*, **14**, 105-130.
https://doi.org/10.4236/cc.2026.144007

Received: August 7, 2026
Accepted: September 15, 2026
Published: September 18, 2026

Abstract

This study established a relationship between the water/octanol partition coefficient (Log*P*) and molecular descriptors from quantum chemistry using a QSPR approach. To this end, a QSPR model based on three molecular descriptors, namely ionization energy (*IE*), total electronic energy (E_T), and dipole moment (μ), was developed. To evaluate the quality, robustness, and predictive power of the resulting model, several statistical and validation parameters were calculated ($R^2 = 0.9501$; $R^2_{adjusted} = 0.9386$; $s = 0.2000$; $F = 82.5446$; $Q^2_{LOO} = 0.9149$; $\overline{r_m^2(LOO)} = 0.6651$; $\Delta r_m^2(LOO) = 0.0000$; $Q^2_{ext} = 0.9825$; $\overline{r_m^2(test)} = 0.9013$; $\Delta r_m^2(test) = 0.0787$). The values obtained for these different indicators reveal that the developed QSPR model is valid, robust, and efficient for predicting the water/octanol partition coefficient of the studied anilines. This model can therefore be applied to estimate the hydrophobic character of other compounds belonging to the same chemical family, provided they fall within its range of applicability. Consequently, the design of new anilines with specific water/octanol partition coefficients could be facilitated by modulating the three molecular descriptors included in the developed QSPR model.

Keywords

Quantum Chemistry, QSPR Model, Statistical Data Analysis, Aniline

1. Introduction

Aniline is an important basic chemical, with an annual global production of approximately 10 billion pounds, and is widely used as a feedstock for polyurethane, rubber processing chemicals, herbicides, and dyes and pigments [1]. For example, aniline is a key feedstock for polymeric methylenediphenyldiamine (MDA)/methylenediphenyldiisocyanate (PMDI). Nearly 80% of aniline consumption in the United States is used for the production of PMDI, an intermediate product used in the manufacture of a wide variety of commercial polyurethane products. Aniline is also an important feedstock for the production of rubber processing chemicals, dyes and pigments, specialty fibers, pesticides, and various other chemicals, including pharmaceuticals [2]. A study of global aniline production allows us to assess its competitiveness, not only in terms of market share but also in terms of available technologies. PMDI production accounted for 73% of global aniline consumption and has been the primary driver of global aniline demand growth since 1982. Consequently, nitrobenzene/aniline/PMDI consumption largely follows the trends of major global economies and is heavily dependent on construction/renovation activities and automotive production. Current aniline manufacturing relies on a three-step process: 1) nitric acid production, 2) nitration of benzene with an HNO_3/H_2SO_4 mixture, and 3) hydrogenation of nitrobenzene with metal catalysts (Raney Ni or Pt/Pd on carbon) [1]. Although this indirect method of aniline production has a long history, it has several drawbacks, including corrosivity, environmental concerns, and high raw material and investment costs [3]. Furthermore, the current process is energy-intensive and generates a large amount of acidic waste and phenolic byproducts. This poses serious environmental problems and runs counter to the ecological trends of the global chemical industry [4]. Knowledge of aniline's lipophilicity is essential for predicting its behavior in the environment and in living organisms. It allows us to anticipate its ability to cross cell membranes and its risk of accumulation. Lipophilicity helps estimate whether the product will accumulate in the fats of living organisms (bioaccumulation). Combined with its half-life, it helps calculate its risk to aquatic ecosystems. Knowing its lipophilicity allows us to design liquid-liquid extraction, separation, and purification processes during its synthesis. In this context, the use of methods other than experimentation is essential. Among these, Quantitative Structure-Property/Activity Relationships (QSPR/QSAR) are attracting increasing interest and are now recommended by the most recent regulations [5] [6]. These approaches make it possible to establish mathematical relationships (models) linking the physicochemical properties or biological activities to the molecular structure of compounds. They also contribute to a better understanding of the mechanisms underlying these properties or activities and offer the possibility of predicting them for molecules lacking experimental data. In this study, the main objective is to model the water/octanol partition coefficient ($\log P$) using QSPR methods. More specifically, the aim is to establish a correlation between this lipophilicity

parameter and certain molecular descriptors from quantum chemistry, based on an experimental database composed of twenty-four (24) aniline derivatives.

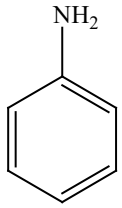
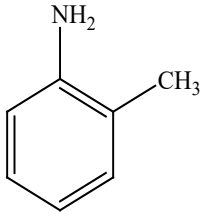
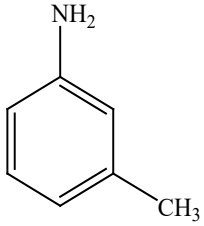
2. Materials and Methods

2.1. Materials

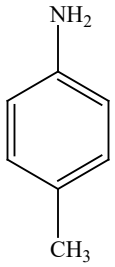
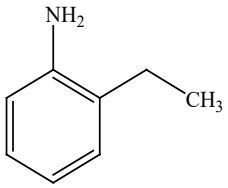
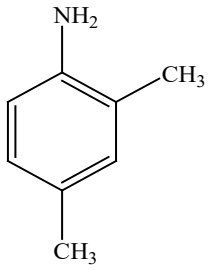
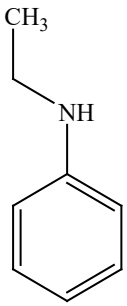
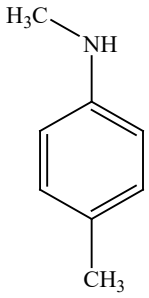
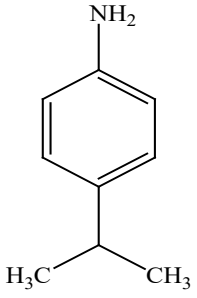
2.1.1. Experimental Database

An experimental database of twenty-four (24) molecules belonging to the aniline family was used in this study. These compounds were derived from the work of James Sangster published in 1989 [7] and have experimental values for the water/octanol partition coefficient (Log*P*). To develop and validate the QSPR model, this database was divided into two subsets: a training set and a test set. Several approaches are proposed in the literature for this allocation [8]. Among them, the commonly adopted rule is to allocate approximately 75% of the compounds to the training set and the remaining 25% to the test set [9]. However, the most widely used method in QSPR/QSAR modeling remains the random selection of compounds. In this context, it is generally recommended that the test set represent at least 20% of the total dataset [10] [11]. In the present study, random selection was used. Thus, seventeen (17) molecules were assigned to the training set, while the remaining seven (7) molecules constituted the test set. Following **Table 1** presents the two-dimensional structures of these different compounds as well as their experimental water/octanol partition coefficient values.

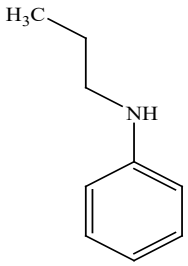
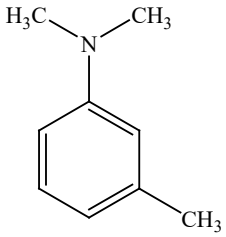
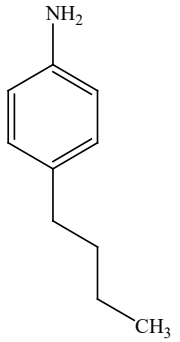
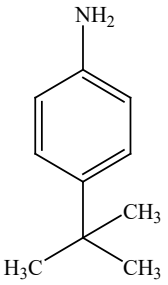
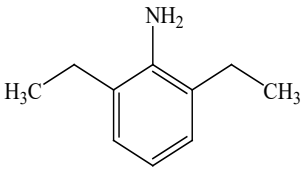
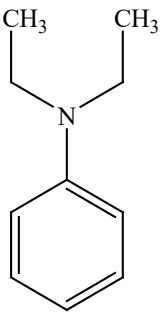
Table 1. Series of aniline studied (experimental database) [7].

Training set			
Compound	**Name**	**2D structure**	**LogP_{exp}**
AM_1	Aniline		**0.90 ± 0.05**
AM_2	2-Methylaniline		**1.32 ± 0.10**
AM_3	3-Methylaniline		**1.40 ± 0.05**

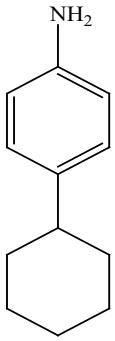
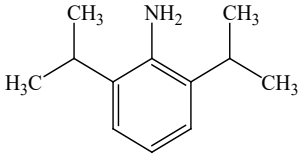
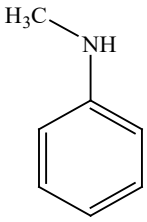
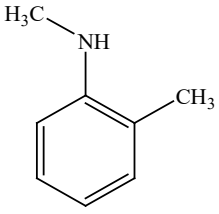
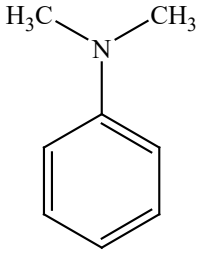
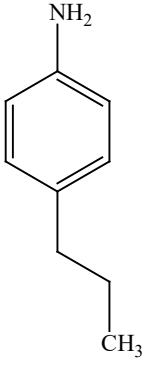
Continued

AM_4	4-Methylaniline		**1.39 ± 0.05**
AM_5	2-Ethylaniline		**1.74 ± 0.25**
AM_6	2,4-Dimethylaniline		**1.68 ± 0.20**
AM_7	N-Ethylaniline		**2.16 ± 0.20**
AM_8	4,N-Dimethylaniline		**2.15 ± 0.20**
AM_9	4-Isopropylaniline		**2.34 ± 0.25**

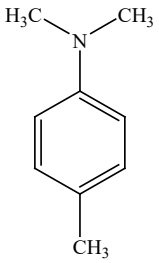
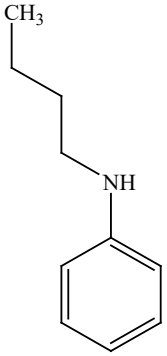
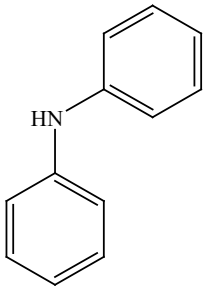
Continued

AM_10	N-Propylaniline		**2.45 ± 0.25**
AM_11	3,N,N-Trimethylaniline		**2.80 ± 0.25**
AM_12	4-n-Butylaniline		**3.05 ± 0.30**
AM_13	4-tert-Butylaniline		**2.47 ± 0.25**
AM_14	2,6-Dithylaniline		**3.10 ± 0.30**
AM_15	N,N-Diethylaniline		**3.31 ± 0.20**

Continued

AM_16	4-Cyclohexylaniline		**3.65 ± 0.30**
AM_17	2,6-Diisopropylaniline		**3.18 ± 0.30**
		Test set	
AM_18	N-Methylaniline		**1.66 ± 0.15**
AM_19	2,N-Dimethylaniline		**2.16 ± 0.20**
AM_20	N,N-Dimethylaniline		**2.31 ± 0.10**
AM_21	4-n-Propylaniline		**2.40 ± 0.25**

Continued

AM_22	4,N,N-Trimethylaniline		**2.81 ± 0.30**
AM_23	N-Butylaniline		**3.10 ± 0.20**
AM_24	N-Benzylaniline		**3.13 ± 0.30**

2.1.2. Used Theory Level and Software

The three-dimensional (3D) structures of the studied molecules were generated and visualized using GaussView 5.0 [12]. Subsequently, geometry optimization and frequency calculations were performed with Gaussian 09 [13] at a temperature of 298.15 K and a pressure of 1 atm in the gas phase. All quantum chemical calculations were carried out at the B3LYP/6-311G(d, p) level of theory. The two-dimensional (2D) molecular structures were drawn using ChemSketch[1]. Graphical representations were produced with Microsoft Excel[2], while XLSTAT[3] was employed for statistical analyses and model development. The calculation of leverage values was performed using Minitab 18[4].

3. Methods

3.1. Statistical Analysis

To develop a Quantitative Structure-Property Relationship (QSPR) model, an

[1](2015) ACDLABS 10. Advanced Chemistry Development Inc.
[2]Microsoft Corporation (2010) Microsoft Excel 2010 [Computer Software].
[3]Addinsoft (2014) XLSTAT (Version 2014.5.03) [Computer Software]
[4]Minitab, LLC. (2017) Minitab Statistical Software (Version 18) [Computer Software].

appropriate statistical analysis method is required to establish and quantify the relationship between the property under investigation and the molecular descriptors representing the chemical structure. Several modeling and statistical analysis techniques are available for this purpose. In the present study, Multiple Linear Regression (MLR), which involves the use of several explanatory variables, was employed to construct the QSPR model.

The general form of the Multiple Linear Regression (MLR) equation [14] is written as follows:

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_n X_n \quad (1)$$

where Y represents the dependent variable (or response), $X_1, X_2, \cdots, X_n$ correspond to the independent variables or descriptors included in the model, $a_1, a_2, \cdots, a_n$ are the regression coefficients associated with these descriptors, and a_0 denotes the constant term of the model. For given values of the explanatory variables $X_1, X_2, \cdots, X_n$, the response variable Y is assumed to follow a normal distribution. Furthermore, the descriptors included in a regression model should exhibit low correlation with each other to avoid multicollinearity problems. To ensure the statistical robustness of the model, the maximum number of descriptors should generally not exceed one-fifth of the total number of components in the training set. A regression model that is well-fitted to the experimental data results in a scatter plot where the points are distributed close to the regression line, thus demonstrating the quality of the fit.

3.2. Molecular Descriptors

In a QSAR study, two molecular descriptors are considered independent when they show no statistical correlation between them. In other words, knowing the value of one does not allow us to predict or explain the value of the other. Thus, when a set of descriptors is available, it is essential to perform a bivariate analysis of the data. This analysis consists of calculating the linear correlation coefficient r between all pairs of descriptors in order to assess their degree of intercorrelation. This approach, called objective analysis, allows us to select the most relevant descriptors while reducing their number, without taking into account the dependent variable (biological activity or property being studied). Depending on the value of the correlation coefficient *r*, the relationships between two descriptors can be interpreted as follows [15]:

- ✓ $r > 0.97$, the descriptors are strongly correlated;
- ✓ $r > 0.90$, the descriptors are moderately correlated;
- ✓ $r < 0.90$, the descriptors are weakly correlated;
- ✓ $r < 0.50$, the descriptors are considered uncorrelated.

In the context of constructing a QSAR/QSPR model, two strongly or slightly correlated descriptors cannot be simultaneously included in the same regression equation due to the multicollinearity issues they may cause. Furthermore, since the correlation can be either positive or negative, the analysis performed in this work relies on the absolute value of the correlation coefficient. The various

molecular descriptors calculated and used in this study are presented below:

- **Electronic energy (E_T)**

The total electronic energy corresponds to the overall energy of the molecule. It is obtained using the Gaussian 09 software after geometric and frequency optimization calculations are performed on the studied molecules.

- **Ionization energy (IE)**

According to the Koopmans approximation [16], the ionization energy is defined as the negative of the energy of the highest occupied molecular orbital (HOMO). It is expressed by the following relation:

$$IE = -E_{\text{HOMO}} \tag{2}$$

- **Dipole moment (μ)**

The dipole moment is a parameter directly related to the non-uniform distribution of electrical charges within a molecule. It thus constitutes a measure of the molecule's polar or nonpolar character. This descriptor is widely used in Quantitative Structure-Activity/Property Relationship (QSAR/QSPR) studies due to its influence on molecular interactions. The dipole moment values were determined from geometric and frequency optimization calculations performed on the studied molecules.

3.3. Statistical and Validation Parameters

Several statistical and validation parameters were computed using Equations (3)-(6).

$$\text{ESS} = \sum \left(Y_{i,\text{cal}} - \overline{Y}_{\text{exp}} \right)^2 \tag{3}$$

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

$$\text{RSS} = \sum \left(Y_{i,\text{exp}} - Y_{i,\text{cal}} \right)^2 \tag{5}$$

$$\text{TSS} = \text{ESS} + \text{RSS} \tag{6}$$

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

- **Determination coefficient (R^2)** [17]

The coefficient of determination is obtained from the following relationship:

$$R^2 = 1 - \frac{\sum \left(Y_{i,\text{exp}} - Y_{i,\text{cal}} \right)^2}{\sum \left(Y_{i,\text{exp}} - \overline{Y}_{\text{exp}} \right)^2} = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{7}$$

with

$$R = \pm \sqrt{\frac{\sum \left(Y_{i,\text{cal}} - \overline{Y}_{\text{exp}} \right)^2}{\sum \left(Y_{i,\text{exp}} - \overline{Y}_{\text{exp}} \right)^2}} = \pm \sqrt{\frac{\text{ESS}}{\text{TSS}}} \tag{8}$$

- **Standard deviation (s)** [18]

This parameter reflects the degree of data dispersion around the mean. A value approaching zero indicates a better model fit and, consequently, higher reliability

of the predicted results.

$$s = \sqrt{\frac{\sum\left(Y_{i,\exp} - Y_{i,\text{cal}}\right)^2}{n-p-1}} = \sqrt{\frac{\text{RSS}}{n-p-1}} \tag{9}$$

- **Adjusted determination coefficient (R^2_{adjusted})** [19]

Unlike the coefficient of determination R^2, this parameter provides a measure of the model's robustness. It is particularly useful in multiple regression analyses because it accounts for the number of parameters (descriptors) included in the model.

$$R^2_{\text{adjusted}} = 1 - \frac{(n - \text{Intercept})}{n-p-1} \cdot \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{(n - \text{Intercept})}{n-p-1} \cdot \left(1 - R^2\right) \tag{10}$$

- **Fisher-Snedecor coefficient (*F*)** [20]

The Fisher-Snedecor test is used to evaluate the overall significance of a linear regression model. A regression equation is considered globally significant when it includes at least one explanatory variable that significantly contributes to the prediction of the dependent variable. The Fisher-Snedecor statistic is related to the coefficient of determination through the following relationship:

$$F = \frac{n-p-1}{p} \cdot \frac{\text{ESS}}{\text{RSS}} = \frac{n-p-1}{p} \cdot \frac{R^2}{1-R^2} \tag{11}$$

- **Cross-validation coefficient (Q^2_{LOO})** [21]

It is used to measure the predictive performance of the model based on the training set data.

$$Q^2_{\text{LOO}} = 1 - \frac{\sum\left(y_{i,\exp} - y_{i,\text{pred}}\right)^2}{\sum\left(y_{i,\exp} - \overline{y}_{\exp}\right)^2} = 1 - \frac{\text{PRESS}}{\text{TSS}} \tag{12}$$

- **Cross-validation criteria (PRESS)** [21]

The sum of squared prediction errors, known as PRESS (Prediction Sum of Squares), is defined by the following relationship:

$$\text{PRESS} = \sum\left(y_{i,\exp} - y_{i,\text{pred}}\right)^2 \tag{13}$$

This criterion is used to select models with strong predictive performance; the preferred model is the one that yields the lowest PRESS value. The standard deviation of the prediction error (SDEP) is then derived from PRESS.

$$\text{SDEP} = \sqrt{\frac{\sum\left(y_{i,\exp} - y_{i,\text{pred}}\right)^2}{n}} = \sqrt{\frac{\text{PRESS}}{n}} \tag{14}$$

In these expressions, n is the number of molecules in the training set, p is the number of explanatory variables. $y_{i,\exp}$ and $y_{i,\text{pred}}$ are respectively the experimental and predicted values of the property for molecule i, and $\overline{y}_{\exp}$ is the average value of the property for the training set.

- **Todeschini's parameter ($^cR^2_P$)** [22]

$^cR^2_P$ is the corrected form of P.P. Roy's parameter noted R^2_P [23]. It allows one to know if the model is due to chance correlations or not. If this parameter is

greater than 0.50, the model is not due to chance correlations. It is defined as:

$$^{c}R_P^2 = R\sqrt{R^2 - R_r^2} \tag{15}$$

With R_r^2, the average value of R_{ri}^2 of the models was obtained with the randomized property.

- **External validation coefficient (Q_{ext}^2)** [24]

It measures the accuracy of the predictions on the test set data.

$$Q_{\text{ext}}^2 = 1 - \frac{n}{n_{\text{ext}}} \frac{\text{PRESS}(\text{test})}{\text{TSS}} \tag{16}$$

Here, n_{ext} refers to the number of test set compounds.

- **Parameter (RMSEP)** [24]

The external predictive ability of the Quantitative Structure-Property Relationship (QSPR) model may further be determined by the root mean square error in prediction, given by:

$$\text{RMSEP} = \sqrt{\frac{\sum\left(y_{\text{exp(test)}} - y_{\text{pred(test)}}\right)^2}{n_{\text{ext}}}} \tag{17}$$

- **Roy K. *et al.* [25] parameters ($\overline{r_m^2}$ and Δr_m^2)**

For an acceptable prediction, the value of Δr_m^2 should preferably be lower than 0.20 when the value of $\overline{r_m^2}$ is more than 0.50.

$$\overline{r_m^2} = \frac{\left(r_m^2 + r_m'^2\right)}{2} \tag{18}$$

$$\Delta r_m^2 = \left|r_m^2 - r_m'^2\right| \tag{19}$$

Here

$$r_m^2 = r^2\left(1 - \sqrt{r^2 - r_0^2}\right) \tag{20}$$

and

$$r_m'^2 = r^2\left(1 - \sqrt{r^2 - r_0'^2}\right) \tag{21}$$

The parameters r^2 and r_0^2 are the coefficients of determination between the observed and predicted values of the compounds (training set or test set) with and without intercept, respectively. The parameter $r_0'^2$ bears the same meaning but uses the reversed axes.

- **External validation criteria or "Tropsha's criteria"** [21] [26].

There are five such criteria:

✓ Criterion 1: $R_{\text{ext}}^2 > 0.70$

✓ Criterion 2: $Q_{\text{ext}}^2 > 0.60$

✓ Criterion 3: $\dfrac{\left|R_{\text{ext}}^2 - R_0^2\right|}{R_{\text{ext}}^2} < 0.1$ and $0.85 < k < 1.15$

✓ Criterion 4: $\dfrac{\left|R_{\text{ext}}^2 - R_0'^2\right|}{R_{\text{ext}}^2} < 0.1$ and $0.85 < k' < 1.15$

✓ Criterion 5: $\left|R_{\text{ext}}^2 - R_0^2\right| < 0.3$

where, R^2_{ext} stands for the determination coefficient of molecules for the test set; R^2_0 represents the determination coefficient of the regression between predicted and experimental values for the test set without intercept; R'^2_0 is the determination coefficient of the regression between experimental and predicted values for the test set without intercept. k stands for the slope of the correlation line (values predicted according to the experimental values with intercept = 0), and k' is the slope of the correlation line (experimental values according to the predicted values with intercept = 0).

- **Domain of Applicability**

The usefulness of a QSPR model essentially lies in its ability to predict the properties of new chemical compounds. Thus, after building the model, it is essential to define its domain of applicability (DA). A model is considered valid only within this domain, and any prediction made outside of it must be interpreted as an extrapolation, and therefore unreliable. The most commonly used method for determining the DA relies on calculating the leverage value of each compound [27]. The Williams diagram, which represents the standardized (or studentized) residuals R as a function of the leverage values h_{ii}, is used in this study to graphically evaluate the domain of applicability of the developed QSPR models. Leverage corresponds to a measure of the distance of an observation from the centroid of the data in the space of explanatory variables. It thus allows the identification of observations that are structurally distant from others. For an observation i, the leverage is defined by the following expression:

$$h_{ii} = x_i \left(X^{\mathrm{T}} X \right)^{-1} x_i^{\mathrm{T}} \tag{22}$$

where x_i is the descriptor vector of the considered composite, X is the descriptor matrix derived from the descriptor values of the training set, and the threshold is defined as follows:

$$h^* = \frac{3(p+1)}{n} \tag{23}$$

where n is the number of compounds in the training set, and p is the number of descriptors in the proposed model. It should be noted that [28]:

- If the h_{ii} value of a compound in the training set is greater than h^*, its structure strongly contributes to the model construction.
- If all points have values within the intervals $0 \le h_{ii} \le h^*$ and $-3\sigma \le R \le 3\sigma$, the model is considered statistically acceptable and valid.
- A compound with a standardized (or studentized) residual greater than three standard deviations ($R > 3\sigma$) and a leverage value below the threshold ($h_{ii} < h^*$) is considered aberrant in terms of response (Y aberrant).
- Conversely, a compound with a leverage value above the threshold ($h_{ii} > h^*$) but a residual below 3σ is considered structurally influential (X aberrant). When it belongs to the training set, it can improve the robustness and accuracy of the model and is then referred to as a beneficial influential point.
- Compounds with $h_{ii} > h^*$ strengthen the model when they belong to the training set, but otherwise will have questionable predicted values, though not

necessarily outliers, as the residuals may be low.

3.4. Statistical Tests

- **Shapiro-Wilk Test**

Widely used, the Shapiro-Wilk test [29] is based on the W statistic. Compared to other tests, it stands out for its high power with small samples ($n \leq 50$). The test statistic is given by:

$$W = \frac{\left[\sum_{i=1}^{k=\left[\frac{n}{2}\right]} a_i \left(y_{n-i+1} - y_i\right)\right]^2}{\sum_i \left(y_i - \bar{y}\right)^2} \quad \text{with} \quad -a_i = a_{n-i+1} \tag{24}$$

y_i corresponds to the series of sorted data.

$\left[\frac{n}{2}\right]$ is the entire part of the report $\frac{n}{2}$.

a_i These are constants generated from the mean and variance-covariance matrix of the quantiles of a sample of size n following a normal distribution. The constants a_{n-i+1} are provided in specific tables. The W statistic can therefore be interpreted as the coefficient of determination (the square of the correlation coefficient) between the series of quantiles generated from the normal distribution and the empirical quantiles obtained from the data. The null and alternative hypotheses of the Shapiro-Wilk test are as follows:

H_0: The sample of size n follows a normal distribution.

H_1: The sample of size n does not follow a normal distribution.

The value of the W statistic is higher, the more credible the agreement with the normal distribution.

✓ If $W < W_{\text{critical}}$, reject H_0 (25)

✓ If $W > W_{\text{critical}}$, do not reject H_0 (26)

- **Durbin-Watson test**

Developed by J. Durbin and G. Watson (1950-1951), the Durbin-Watson test [30]-[32] is used to detect autocorrelation among the residuals of a linear regression. In practice, error terms are often autocorrelated, which can lead to poor parameter estimation. It is assumed that the residuals ε_i are stationary and normally distributed with mean 0. The null and alternative hypotheses of the Durbin-Watson test are:

H_0: The residuals are not autocorrelated ($\rho = 0$).

H_1: The residuals are distributed according to a first-order autoregressive process (AR 1) ($\rho > 0$).

The test statistic d is written as:

$$d = \frac{\sum_{i=2}^{n} \left(\varepsilon_i - \varepsilon_{i-1}\right)^2}{\sum_{i=1}^{n} \varepsilon_i^2} \tag{27}$$

where $\varepsilon_i = y_i - \hat{y}_i$ and y_i and $\hat{y}_i$ are, respectively, the observed and predicted values of the response (dependent variable) for compound i. The upper and lower

critical values, d_U (Upper) and d_L (Lower), were tabulated (Durbin-Watson table) for different values of k (number of explanatory variables) and n (sample size).

✓ If $d < d_L$, reject H_0: $\rho = 0$; (28)

✓ If $d > d_U$, do not reject H_0: $\rho = 0$; (29)

✓ If $d_L < d < d_U$, the test is inconclusive. (30)

4. Results and Discussion

1) Values of calculated molecular descriptors

In this study, three molecular descriptors were calculated: ionization energy (*IE*), electronic energy (E_T), and dipole moment (μ). The correlation between these descriptors, taken two at a time, was also studied using the correlation matrix. Table 2 presents the values of these different descriptors.

Table 2. Values of calculated molecular descriptors.

Training set				
Compound	**LogP_{exp}**	***IE* (eV)**	**E_T (u.a)**	**μ (D)**
AM_1	**0.90**	4.3807	−287.6815	1.7146
AM_2	**1.32**	4.3291	−327.0090	1.7493
AM_3	**1.40**	4.3209	−327.0090	1.5618
AM_4	**1.39**	4.2428	−327.0081	1.4627
AM_5	**1.74**	4.3170	−366.3311	1.7673
AM_6	**1.68**	4.2010	−366.3356	1.5285
AM_7	**2.16**	4.2067	−366.3237	1.9053
AM_8	**2.15**	4.0903	−366.3232	1.5713
AM_9	**2.34**	4.2627	−405.6558	1.5801
AM_10	**2.45**	4.1898	−405.6479	2.0112
AM_11	**2.80**	4.0425	−405.6369	1.7900
AM_12	**3.05**	4.2479	−444.9810	1.5175
AM_13	**2.47**	4.2625	−444.9766	1.5800
AM_14	**3.10**	4.2294	−444.9800	1.8200
AM_15	**3.31**	4.0211	−444.9602	1.9844
AM_16	**3.65**	4.2458	−522.4199	1.5216
AM_17	**3.18**	4.2568	−523.6248	1.6367
Test set				
AM_18	**1.66**	4.2054	−326.9966	1.8748
AM_19	**2.16**	4.1704	−366.3239	1.9257
AM_20	**2.31**	4.0886	−366.3095	1.9741
AM_21	**2.40**	4.2532	−405.6569	1.5213
AM_22	**2.81**	3.9852	−405.6362	1.6192
AM_23	**3.10**	4.1857	−444.9720	2.0401
AM_24	**3.13**	4.1715	−518.7821	0.8627

2) Normality test of experimental water/octanol partition coefficients

Before any QSPR modeling, it is essential to test the normality distribution of the experimental data to be modeled. The Shapiro-Wilk test allows us to verify the normality distribution of the data. A summary of the test parameters is listed in Table 3 below.

Table 3. Shapiro-Wilk test parameter values.

W	$W_{critical}$	p-value	1 - α
0.9658	0.916	0.5651	0.05

Analysis of the data in Table 3 shows that the calculated p-value is greater than 1 - $\alpha = 0.05$ (5% significance level). Regarding $W_{critical}$, the value is lower than that of the calculated W. Under these conditions, the theoretical values of the water/octanol partition coefficient follow a normal distribution. This normal distribution is confirmed by the distribution of the scatter plot along the first bisector (the line with the equation $y = x$) (Figure 1).

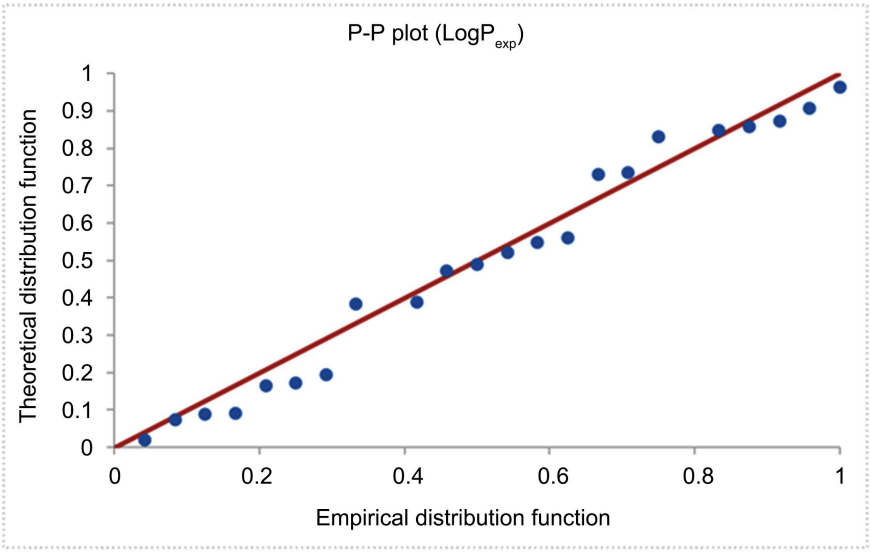


Figure 1. Graph P-P plot (LogP_{exp}) of the model.

3) Study of the intercorrelation between molecular descriptors

The correlation between pairwise descriptors must be zero or low to avoid information redundancy. Table 4 below contains the correlation coefficient values between the different calculated molecular descriptors.

Table 4. Correlation matrix of descriptors.

Variables	IE	E_T	μ
EI	**1.0000**		
E_T	0.2990	**1.0000**	
μ	−0.3356	0.0037	**1.0000**

The correlation matrix shows that the absolute values of the correlation coefficients for the pairwise descriptors are less than 0.50 ($|r| < 0.50$). These values, all below 0.50, indicate the absence of correlation between these pairwise descriptors. Since the correlations between the different pairs are zero, these three descriptors can therefore coexist in the same QSPR model. A QSPR model will thus be established based on these three parameters.

4) Development of the QSPR model: LogP_{theo} = f(*IE*, E_T, and μ)

A multiparametric QSPR model dependent on ionization energy (*IE*), electronic energy (E_T), and dipole moment (μ) was developed using XLSTAT software.

- **QSPR model regression equation**

The XLSAT software allowed us to determine the coefficients of the regression equation for the developed QSPR model. The regression coefficients are summarized in Table 5.

Table 5. Regression coefficient values for the model.

Source	Value	Standard Error	t	Pr > |t|
Constant	5.4544	2.7675	1.9709	0.0704 (·)
IE	−1.8954	0.5725	−3.3110	0.0056 (**)
E_T	−0.0105	0.0008	−13.3267	<0.0001 (***)
μ	0.3945	0.3129	1.2608	0.2295 ()

$$|t_{critical}| = 2.16 .$$

The regression coefficients in Table 5 assigned to the different explanatory variables (molecular descriptors) lead to the following regression equation:

$$\text{Log}P_{theo} = 5.4544 - 1.8954IE - 0.0105E_T + 0.3945\mu .$$

The regression equation indicates that the coefficients for ionization energy (*IE*) and electronic energy (E_T) are negative, while that of the dipole moment (μ) has the opposite sign. Under these conditions, the water/octanol partition coefficient increases along with the dipole moment. Indeed, an increase in this independent variable also leads to an increase in the water/octanol partition coefficient. Conversely, an increase in ionization energy leads to a decrease in the partition coefficient. As for electronic energy, its decrease leads to an increase in the property being studied. It is also noted that the p-value of the initial constant belongs to the interval [0.01; 0.1]. This means that the initial constant has a nearly significant contribution to the prediction of the partition coefficient. Regarding ionization energy, its p-value is within the interval [0.001; 0.01]. Consequently, the contribution of ionization energy is highly significant () in predicting the water/octanol partition coefficient. For the explanatory variable E_T, its p-value falls within the range [0; 0.001[, indicating a highly significant (*) influence on the water/octanol partition coefficient. For the dipole moment, the p-value is between 0.1 and 1. This clearly shows that this variable is not significant () in predicting

the property under study. Therefore, we can limit ourselves to determining the first two variables to evaluate the water/octanol partition coefficient for this family of compounds. The absolute values of the t-test and the contribution values reveal that electronic energy still provides the strongest contribution (67.07%), indicating that it is the main predictive descriptor of the partition coefficient for the studied aniline family. The ionization energy contributes 16.66% while the dipole moment contributes 6.35%. The ranking of the descriptors in descending order of priority in predicting the water/octanol partition coefficient is as follows: $E_T > IE > \mu$.

- **Analysis of Variance (ANOVA) table for the model**

Table 6 allows for the easy calculation of various statistical parameters of the model.

Table 6. ANOVA table of the model.

Source	DF	Sum of squares	Average of squares	*F*	Pr > *F*
Model	3	9.9013	3.3004	82.5446	<0.0001 (***)
Error	13	0.5198	0.0400		
Corrected total	16	10.4211			

(***) means: "Highly significant".

In Table 6 (ANOVA table), the p-value falls within the range [0; 0.001[, indicating that the model's regression equation is highly significant (***) for predicting the water/octanol partition coefficient of the studied molecules. This significance is confirmed by the very high Fisher's exact test (F = 82.5446), which is much greater than the significance threshold (F_{limit} = 3.41). More precisely, at least one of the explanatory variables is relevant for explaining the dependent variable (lipophilicity parameter). A summary of the parameters determined using the ANOVA table is provided in Table 7.

- **Statistical parameters of the model**

Various parameters related to the developed QSPR model were calculated to assess its quality. These parameters are summarized in Table 7 below:

Table 7. Statistical parameters of the model.

n	*R*	R^2	R^2_{adjusted}	*s*	*F*
17	0.9747	0.9501	0.9386	0.2000	82.5446

Table 7 shows that the correlation coefficient is very high (R = 0.9747). This indicates that the hydrophobicity parameter is strongly correlated with the three selected molecular descriptors. The adjusted coefficient of determination R^2_{adjusted} provides greater accuracy than the standard coefficient of determination R^2. The value of R^2_{adjusted} is 0.9386. This value reveals that 93.86% of the experimental variance of the water/octanol partition coefficient is explained by the model descriptors. Furthermore, the low standard deviation (s = 0.2000) demonstrates a good fit and high predictive reliability. The statistical parameters yielded satisfactory

results, thus demonstrating the good quality of the developed QSPR model. However, validation tests are necessary to avoid overestimating the model's predictive capacity with respect to actual external molecules.

5) Internal validation of the model

Internal validation of a QSPR model applies only to the training set. The validation methods used are Leave-One-Out (LOO) cross-validation and property randomization testing.

- **Leave-One-Out cross-validation of the model**

The Y-randomization parameters are summarized in Table 8.

Table 8. Statistical parameters of the LOO cross-validation of the model.

n	Q^2_{LOO}	$\overline{r^2_m(\text{LOO})}$	$\Delta r^2_m(\text{LOO})$	PRESS	SDEP
17	0.9149	0.6651	0.0000	0.9166	0.0079

Remarkably, the cross-validation coefficient LOO of the model is 0.9149 and is greater than 0.90 ($Q^2_{\text{LOO}} > 0.90$). This demonstrates the model's excellent prediction of the partition coefficient for the studied family of molecules, according to Eriksson *et al.* [33]. Furthermore, out of 100 molecules in the training set, approximately 91 have their water/octanol partition coefficients predicted by the model. The model therefore exhibits high predictive power for the molecules in the training set. This result shows that our developed QSPR model is very insensitive to the Leave-One-Out operation, as the cross-validation coefficient Q^2_{LOO} is very high, as is the R^2 coefficient of determination. This justifies its robustness. Regarding the coefficient $\overline{r^2_m(\text{LOO})}$, its value is 0.6651 and is greater than 0.50, while that of $\Delta r^2_m(\text{LOO})$ is less than 0.2. Consequently, the developed model is acceptable for predicting the water/octanol partition coefficient of the training set. To determine whether the established QSPR model is randomized, a randomization test of the property under study was performed.

- **Y-Randomization Test of the Model**

For randomization, a circular permutation (16 iterations) was performed. A summary of the mean values of the randomization parameters is given in Table 9.

Table 9. Mean values of model randomization parameters.

Randomized parameter	R^2_r	s_r	F_r	$^cR^2_p$
Average value	0.2564	0.7691	1.6903	0.8118

The mean value of the randomized coefficient of determination R^2_r is very low ($R^2_r = 0.2564$), indicating that the regression line equation accounts for only 25.64% of the point distribution (water/octanol partition coefficient). Furthermore, there is significant dispersion of the point cloud around the regression line, confirmed by a high randomized standard deviation ($s_r = 0.7691$). The very low value of the statistic ($F_r = 1.6903$) of the randomized model shows that the randomized model equation is not significant. As for the Todeschini corrected

parameter ${}^{c}R_P^2$, its value is 0.8118 and is well above 0.50 (${}^{c}R_P^2 > 0.50$). Since this parameter is greater than 0.50, it is clear that the developed QSPR model is not due to chance correlations.

6) External validation of the model

External validation applies only to the test set. For this validation, parameters were determined. Furthermore, the Tropsha criteria were verified.

- **Statistical parameters of the external validation of the model**

Table 10. Statistical parameters of the external validation of the model.

n_{ext}	R_{ext}^2	Q_{ext}^2	$\overline{r_m^2(\text{test})}$	$\Delta r_m^2(\text{test})$	PRESS (test)	RMSEP
7	0.9637	0.9825	0.9013	0.0787	0.0751	0.1035

From the analysis of the data in Table 10, it can be indicated that the model has very high predictive power due to the high value of the external validation coefficient ($Q_{ext}^2 = 0.9825$). This means that out of 100 molecules in the test set, approximately 98 have their water/octanol partition coefficients predicted by the model. Similarly, 96.37% of the experimental variance of the water/octanol partition coefficient is explained by the model's descriptors. Regarding $\overline{r_m^2(\text{test})}$, the value is greater than 0.50, while that of $\Delta r_m^2(\text{test})$ is less than 0.2. Thus, the developed model is acceptable for predicting the properties of the molecules in the test set. Furthermore, the five (5) Tropsha criteria have been met.

- **Verification of the Tropsha criteria for the model**

✓ **Criterion 1:** $R_{ext}^2 = 0.9637 > 0.70$

✓ **Criterion 2:** $Q_{ext}^2 = 0.9825 > 0.60$

✓ **Criterion 3:** $\dfrac{\left|R_{ext}^2 - R_0^2\right|}{R_{ext}^2} = 0.0021 < 0.1$ and $k = 1.0009$ with $0.85 < k < 1.15$

✓ **Criterion 4:** $\dfrac{\left|R_{ext}^2 - R_0'^2\right|}{R_{ext}^2} = 0.0075 < 0.1$ and $k' = 0.9975$ with $0.85 < k' < 1.15$

✓ **Criterion 5:** $\left|R_{ext}^2 - R_0^2\right| = 0.002 < 0.3$

We observe that all five (05) Tropsha criteria are met. Consequently, the model performs very well in predicting the water/octanol partition coefficient of the molecules in the test set of the experimental database.

- **Ratio $\tau = \text{Log}P_{exp}/\text{Log}P_{theo}$ of the test set**

The model's performance was also studied by analyzing the ratio $\tau = \text{Log}P_{exp}/\text{Log}P_{theo}$ of the test series. The values obtained are summarized in Table 11.

Analysis of Table 11 shows that the values of the ratio $\tau = \text{Log}P_{exp}/\text{Log}P_{theo}$ are approximately equal to one. This demonstrates that the model performs very well in predicting the water/octanol partition coefficient of the studied series of molecules and can be used to predict the partition coefficient of other molecules in the same family.

Table 11. Values of the ratio $\tau = \text{Log}P_{exp}/\text{Log}P_{theo}$.

$\text{Log}P_{exp}$	$\text{Log}P_{theo}$	$\tau = \text{Log}P_{exp}/\text{Log}P_{theo}$
1.6600	1.6595	1.0003
2.1600	2.1592	1.0004
2.3100	2.3334	0.9900
2.4000	2.2563	1.0637
2.8100	2.8025	1.0027
3.1000	3.0020	1.0327
3.1300	3.3402	0.9371

- **Comparison between model-predicted and experimental values of the water/octanol partition coefficient in the test set**

The theoretical values of the lipophilicity parameter were compared to the experimental values based on the regression constants of the graphs $\text{Log}P_{theo} = f(\text{Log}P_{exp})$ of Figure 2 and $\text{Log}P_{exp} = f(\text{Log}P_{theo})$ of Figure 3.

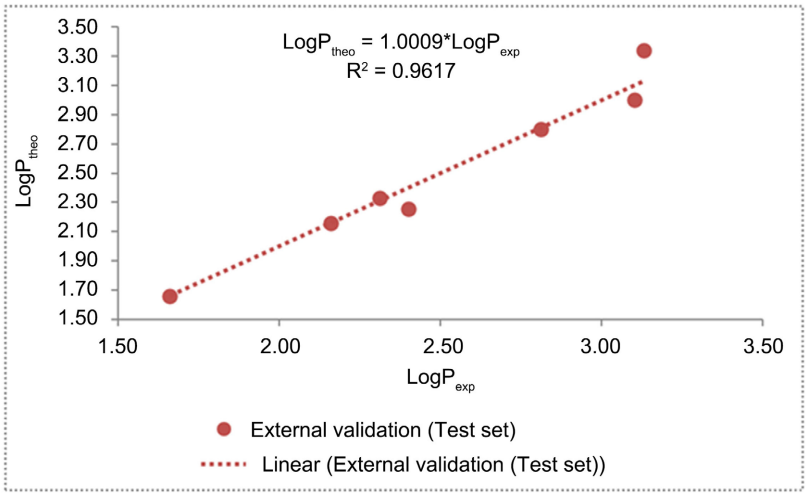


Figure 2. $\text{Log}P_{theo} = f(\text{Log}P_{exp})$ graph of the model's test series (intercept = 0).

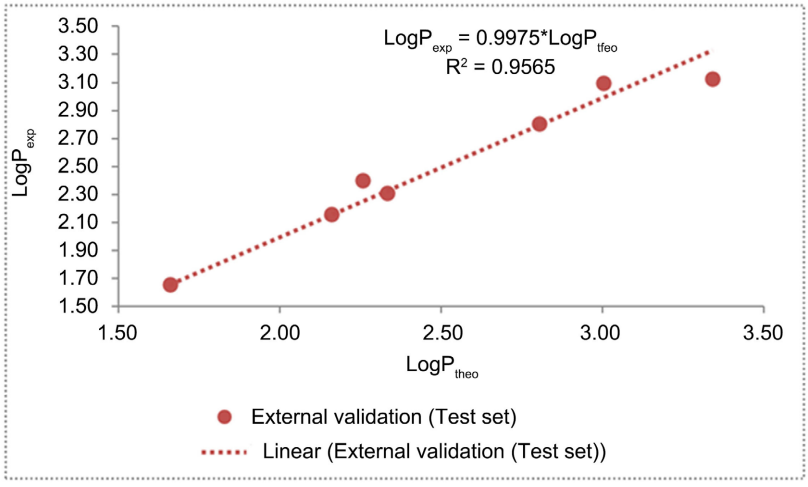


Figure 3. $\text{Log}P_{exp} = f(\text{Log}P_{theo})$ graph of the model's test series (intercept = 0).

According to Figure 2 and Figure 3, the values of the slopes $k = 1.0009$ and $k' = 0.9975$ are very close to unity. This means that the calculated water/octanol partition coefficient is very close to the experimental value ($\text{Log}P_{\text{theo}} = f(\text{Log}P_{\text{exp}})$). This leads to a ratio closer to unity. The results of the external validation further demonstrate the model's high performance in predicting the partition coefficient for the series of studied molecules. It can be used effectively for predicting the lipophilicity parameter of new aniline within its range of applicability.

- **Correlation between model-predicted and experimental values of the water/octanol partition coefficient**

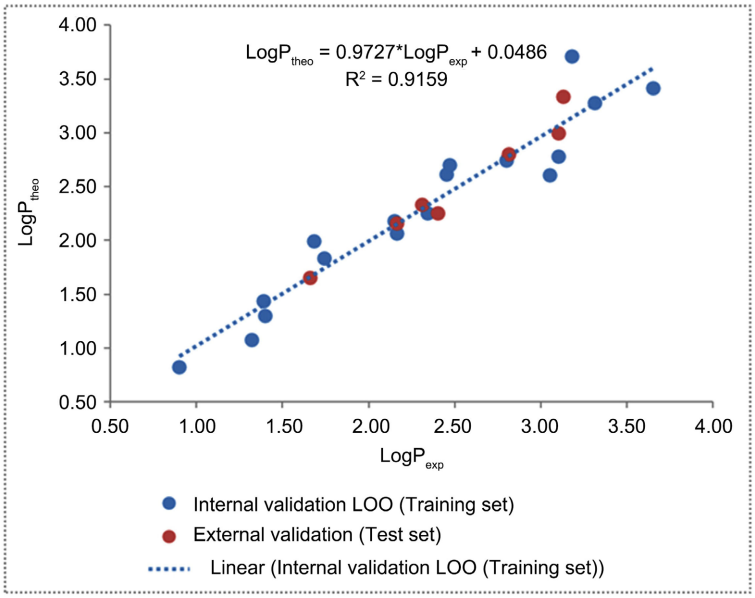


Figure 4. $\text{Log}P_{\text{theo}}$-$\text{Log}P_{\text{exp}}$ scatter plot of the model.

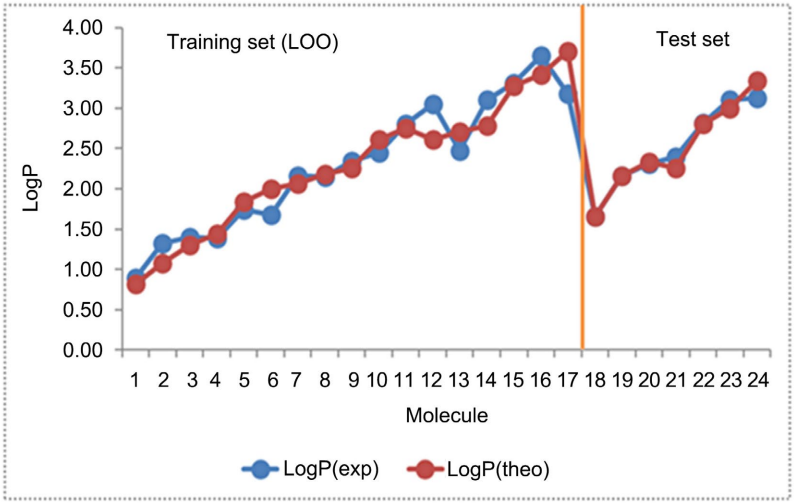


Figure 5. Similarity between values predicted by the model and experimental values.

Across the scatter plot distribution around the regression line in Figure 4, a strong linear correlation is observed between the model-predicted and experimental

values of the water/octanol partition coefficient. Figure 5 shows a similarity between the curves of model-predicted values and experimental values, particularly for the test set. Consequently, these graphs confirm that the model is validated and accurate in predicting the water/octanol partition coefficient. This also reflects the suitability of the theoretical framework used to develop this QSPR model.

7) Statistical Tests

- **Normality Test of the Model Data**

The test performed to verify the normality of the model data is the Shapiro-Wilk test. The test parameters are summarized in Table 12.

Table 12. Shapiro-Wilk test parameter values for the model.

W	$W_{critique}$	p-value	$1 - \alpha$
0.9755	0.916	0.8003	0.05

Analysis of the values in Table 12 shows that the calculated p-value is greater than $1 - \alpha = 0.05$ (5% significance level). Regarding $W_{critical}$, its value is lower than that of the calculated W. Consequently, the assumption of normality is consistent with our data. This normal distribution is confirmed by the distribution of the scatter plot along the first bisector (the line with equation $y = x$) (Figure 6).

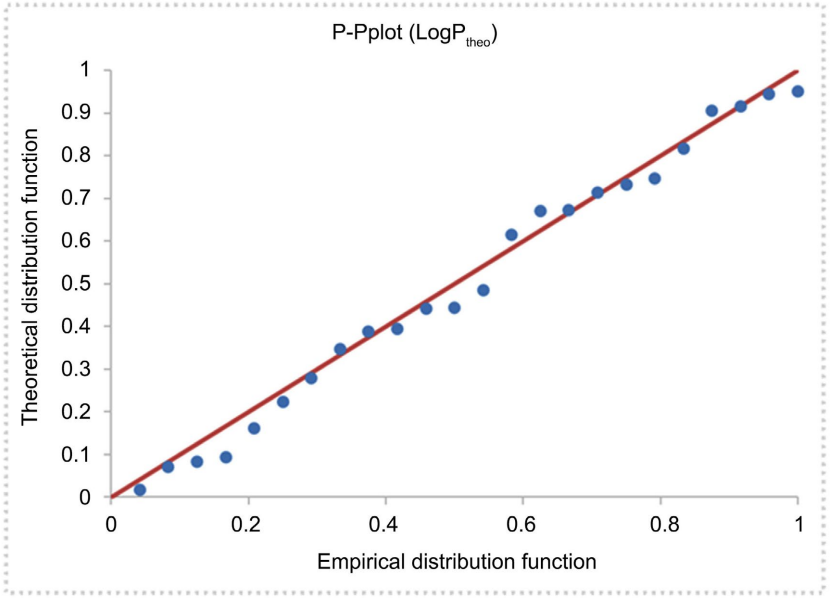


Figure 6. P-P plot ($\text{Log}P_{theo}$) graph of the model.

- **Durbin-Watson test of the model**

Table 13. Parameter values of the model Durbin-Watson test.

d	d_L	d_U	p-value	$1 - \alpha$
2.6278	1.10	1.66	0.9118	0.05

The data in Table 13 show that the calculated statistical test d is greater than the maximum critical value ($d_U = 1.66$). Also, the calculated p-value is greater than $1 - \alpha = 0.05$ (significance level of 5%). It is therefore clear that the residuals are not autocorrelated. These residuals contain no information that could influence the prediction of the water/octanol partition coefficient by the model.

8) Model applicability domain

The applicability domain of the QSPR model was defined using the Williams diagram, which is simply the graphical representation of the studentized residuals as a function of the leverage. Figure 7 below shows the plot of this diagram.

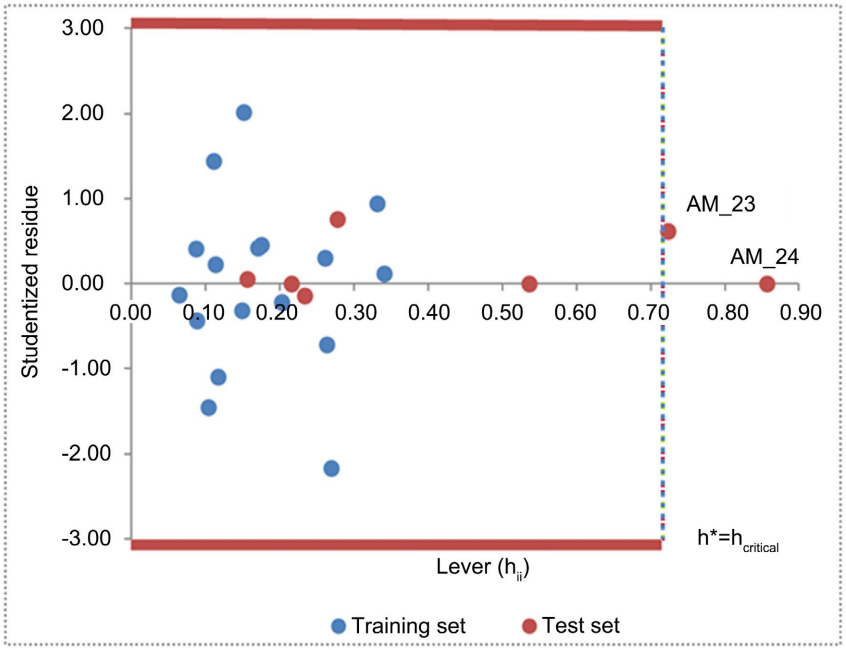


Figure 7. Williams diagram of the model.

On the Williams diagram, all observations in the training set have studentized residuals within ±3 standard deviations. The same is true for the test set. Furthermore, apart from observations AM_23 and AM_24 in the test set, the leverages of the others are all below the threshold value $h^* = 0.7059$. This indicates that no observation in either set is an outlier. The results of the external validation reveal that the model can be used to predict the water/octanol partition coefficient of future anilines belonging to the same family within its domain of applicability.

5. Conclusion

This work aimed to linearly correlate the water/octanol partition coefficient of a series of twenty-four (24) aniline molecules with molecular descriptors from quantum chemistry. To this end, a multiparametric QSPR model dependent on ionization energy (EI), electronic energy (E_T), and dipole moment (μ) was developed. This model exhibits several highly satisfactory statistical and validation parameters: $R^2 = 0.9501$; $R^2_{\text{adjusted}} = 0.9386$; $s = 0.2000$; $F = 82.5446$; $Q^2_{\text{LOO}} = 0.9149$;

$\overline{r_m^2(\text{LOO})} = 0.6651$; $\Delta r_m^2(\text{LOO}) = 0.0000$; $Q_{\text{ext}}^2 = 0.9825$; $\overline{r_m^2(\text{test})} = 0.9013$; $\Delta r_m^2(\text{test}) = 0.0787$. These various parameters reveal that the developed QSPR model is validated and performs well in predicting the water/octanol partition coefficient. It is acceptable as a predictive model. Consequently, it can now be used to predict future aniline water/octanol partition coefficients of the same family within its scope of applicability. Therefore, we plan to use this model to design new aniline compounds with desired lipophilicity parameters.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Nexant ChemSystems (2003) PERP Report: Nitrobenzene/Aniline. Nexant ChemSystems.

[2] Schmerling, L. (1960) Preparation of Aromatic Amines. U.S. Patent 2948755A. U.S. Patent and Trademark Office.

[3] Wiley (2006) Ullmann's Encyclopedia of Industrial Chemistry. 7th Edition, Wiley.

[4] Baker, R.T. and Tumas, W. (1999) Toward Greener Chemistry. *Science*, **284**, 1477-1479. https://doi.org/10.1126/science.284.5419.1477

[5] European Parliament and Council of the European Union (2006) Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the registration, evaluation, authorisation and restriction of chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. *Official Journal of the European Union*, **L 396**, 1-849. https://eurlex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32006R1907

[6] Margossian, N. (2008) Le règlement REACH—La règlementation européenne sur les produits chimiques. Dunod.

[7] Sangster, J. (1989) Octanol-water Partition Coefficients of Simple Organic Compounds. *Journal of Physical and Chemical Reference Data*, **18**, 1111-1229. https://doi.org/10.1063/1.555833

[8] Roy, K., Kar, S. and Das, R.N. (2015) Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment. Academic Press.

[9] Cornillon, P.A. and Matzner-Løber, E. (2011) Régression avec R. Springer.

[10] Gramatica, P., Giani, E. and Papa, E. (2007) Statistical External Validation and Consensus Modeling: A QSPR Case Study for Koc Prediction. *Journal of Molecular Graphics and Modelling*, **25**, 755-766. https://doi.org/10.1016/j.jmgm.2006.06.005

[11] Kahn, I., Fara, D., Karelson, M., Maran, U. and Andersson, P.L. (2004) QSPR Treatment of the Soil Sorption Coefficients of Organic Pollutants. *Journal of Chemical Information and Modeling*, **45**, 94-105. https://doi.org/10.1021/ci0498766

[12] Dennington, R., Keith, T. and Millam, J. (2009) GaussView Version 5. Semichem Inc.

[13] Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., *et al.* (2009) Gaussian 09, Revision A.02. Gaussian, Inc.

[14] Snedecor, G.W. and Cochran, W.G. (1991) Statistical Methods. Wiley-Blackwell.

[15] Trinajstič, N., Nikolič, S., Basak, S.C. and Lukovits, I. (2001) Distance Indices and Their Hyper-Counterparts: Intercorrelation and Use in the Structure-Property Modeling. *SAR and QSAR in Environmental Research*, **12**, 31-54. https://doi.org/10.1080/10629360108035370

[16] Koopmans, T. (1934) Über die Zuordnung von Wellenfunktionen und Eigenwerten zu den Einzelnen Elektronen Eines Atoms. *Physica*, **1**, 104-113. https://doi.org/10.1016/s0031-8914(34)90011-2

[17] Chatterje, S. and Hadi, A.S. (2006) Regression Analysis by Example. 4th Edition, John Wiley & Sons, 366 p.

[18] Siegel, A.F. (1997) Practical Business Statistics. 3rd Edition, IRWIN.

[19] Besse, P. (2003) Pratique de la modélisation statistique. Publications du laboratoire de statistique et Probabilité.

[20] Cook, R.D. and Weisberg, S. (1994). An Introduction to Regression Graphics. Wiley, 265 p. https://doi.org/10.1002/9780470316863

[21] Golbraikh, A. and Tropsha, A. (2002) Beware of Q2! *Journal of Molecular Graphics and Modelling*, **20**, 269-276. https://doi.org/10.1016/s1093-3263(01)00123-1

[22] Mitra, I., Saha, A. and Roy, K. (2010) Exploring Quantitative Structure-Activity Relationship Studies of Antioxidant Phenolic Compounds Obtained from Traditional Chinese Medicinal Plants. *Molecular Simulation*, **36**, 1067-1079. https://doi.org/10.1080/08927022.2010.503326

[23] Pratim Roy, P., Paul, S., Mitra, I. and Roy, K. (2009) On Two Novel Parameters for Validation of Predictive QSAR Models. *Molecules*, **14**, 1660-1701. https://doi.org/10.3390/molecules14051660

[24] Consonni, V., Ballabio, D. and Todeschini, R. (2010) Evaluation of Model Predictive Ability by External Validation Techniques. *Journal of Chemometrics*, **24**, 194-201. https://doi.org/10.1002/cem.1290

[25] Roy, K., Mitra, I., Kar, S., Ojha, P.K., Das, R.N. and Kabir, H. (2012) Comparative Studies on Some Metrics for External Validation of QSPR Models. *Journal of Chemical Information and Modeling*, **52**, 396-408. https://doi.org/10.1021/ci200520g

[26] 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

[27] Gramatica, P. (2007) Principles of QSAR Models Validation: Internal and External. *QSAR & Combinatorial Science*, **26**, 694-701. https://doi.org/10.1002/qsar.200610151

[28] Lu, H.X., Liu, W.Q., Yang, F., Zhou, H., Liu, F.P., Yuan, H., *et al.* (2020) Thermal Conductivity Estimation of Diverse Liquid Aliphatic Oxygen-Containing Organic Compounds Using the Quantitative Structure-Property Relationship Method. *ACS Omega*, **5**, 8534-8542. https://doi.org/10.1021/acsomega.9b04190

[29] Shapiro, S.S. and Wilk, M.B. (1965) An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, **52**, 591-611. https://doi.org/10.1093/biomet/52.3-4.591

[30] Durbin, J. and Watson, G.S. (1951) Testing for Serial Correlation in Least Squares Regression. II. *Biometrika*, **38**, 159-179. https://doi.org/10.2307/2332325

[31] MacKinnon, J.G. (2008) Durbin-Watson Statistic. In: MacKinnon, J.G., Ed., *The New Palgrave Dictionary of Economics*, Palgrave Macmillan UK, 1-3.

https://doi.org/10.1057/978-1-349-95121-5_2200-1

[32] Chattefuee, S. and Hadi, A.S. (2006) Regression Analysis by Example. 4th Edition, Wiley, 366 p.

[33] Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T.D., McDowell, R.M. and Gramatica, P. (2003) Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs. *Environmental Health Perspectives*, **111**, 1361-1375. https://doi.org/10.1289/ehp.5758