Physics and Materials Chemistry, 2021, Vol. 7, No. 1, 1-13
Available online at http://pubs.sciepub.com/pmc/7/1/1
Published by Science and Education Publishing
DOI:10.12691/pmc-7-1-1

Predictive Modeling of *Toxoplasma Gondii* Activity of a Series of Substituted Imidazole-Thiosemicarbazides Using Quantum Descriptors

Sopi Thomas Affi[1,2], Bafétigué Ouattara[3], Georges Stéphane Dembélé[1,2], Mamadou Guy-Richard Koné[1,2,*], Nahossé Ziao[1,2]

[1]Laboratoire de Thermodynamique et de Physico-Chimie du Milieu,
UFR SFA, Université NANGUI ABROGOUA, 02 BP 801 Abidjan 02, Côte-d'Ivoire
[2]Groupe Ivoirien de Recherches en Modélisation des Maladies (GIR2M)
[3]Laboratoire de Physique Fondamentale et Appliquée, UFR SFA, Université NANGUI ABROGOUA,
02 BP 801 Abidjan 02, Côte-d'Ivoire
*Corresponding author: guyrichardkone@gmail.com

Received November 01, 2021; Revised December 03, 2021; Accepted December 12, 2021

Abstract Quantitative Structure Activity Relationship (QSAR) study of *Toxoplasma gondii* was done on a series of twenty-five (25) imidazole-thiosemicarbazide molecules. In order to obtain molecular descriptors all these molecules were optimized at B3LYP/LanL2DZ level. This study was performed using the linear multiple regression (MLR), nonlinear regression (MNLR) and artificial neural network (ANN) methods. These statistical methods allow to find three (3) quantitative models. Quantum descriptors which such as energy gap (ΔE), dipole moment (μ), enthalpy of formation (ΔfH), bond length (D(C-S)) and lipophilicity (Logp) were used in models elaboration. Among obtained models, RNA model has much better predictive ability than other models with $\mathbf{R^2}$ = 0.9291 and **RMCE** = 0.00023. A decrease in energy gap (ΔE) which is the main descriptor could significantly improve the *Toxoplasma gondii* $\mathbf{IC_{50}}$ inhibitory concentration of substituted imidazole-thiosemicarbazide analogues. Furthermore, the external validation test $\mathbf{pIC_{50}}$ theo/ $\mathbf{pIC_{50}}$ exp and the applicability domain from Cook's distance were verified.

Keywords: *QSAR, RNA, Energy gap (ΔE), Toxoplasma gondii activity*

Cite This Article: Sopi Thomas Affi, Bafétigué Ouattara, Georges Stéphane Dembélé, Mamadou Guy-Richard Koné, and Nahossé Ziao, "Predictive Modeling of *Toxoplasma Gondii* Activity of a Series of Substituted Imidazole-Thiosemicarbazides Using Quantum Descriptors." *Physics and Materials Chemistry*, vol. 7, no. 1 (2021): 1-13. doi: 10.12691/pmc-7-1-1.

1. Introduction

Toxoplasmosis is a disease which is caused by infection with a parasite called Toxoplasma gondii. It is usually transmitted to humans by domestic animals, especially cats, or by ingestion of meat which is not well cooked. Toxoplasma gondii affects about 30% of population around the world [1]. This parasite causes severe disease in people with HIV/AIDS or in pregnant women creating congenital malformations [2,3]. In this context, Agata et al [4,5] have synthesized and tested the substituted imidazole-thiosemicarbazides to fight against all infections caused by Toxoplasma gondii. Imidazole-thiosemicarbazides are convenient precursors that have been widely used in heterocyclic synthesis. Since a longtime studies on heterocyclic compounds have been an area of interest in medicinal chemistry. Indeed, these compounds and their derivatives have a wide spectrum of biological activities namely antibacterial [6], antifungal [7], anticonvulsant [8], antimicrobial [9], antitumor [10]. $\mathbf{IC_{50}}$ inhibitory activity of imidazole-thiosemicarbazides was used as backbone of our study. The Quantitative Structure-Activity Relationship (QSAR) study, is the method that allows to correlate the molecular structure with a well determined effect such as biological activity or chemical reactivity. It is increasingly used to reduce the excessive number of experiments which are sometimes long, dangerous and costly in terms of time and money [11,12]. The present work is made in the perspective to establish a model which will be able to fight against the infection caused by Toxoplasma gondii. All this contributes to the reduction of drug production costs [13,14] and contributes to the protection of environment. In general, the QSAR model is a function of one fifth (1/5) of the initial database.

The main objective of this work is to develop reliable models to explain and predict the IC50 antibacterial activity (median inhibitory concentration in μg/mL) of a series of twenty-five (25) substituted imidazole-thiosemicarbazide derivatives (Figure 1). These compounds were synthesized and tested by Agata et al [4,5] for their biological activities.

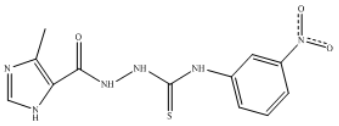
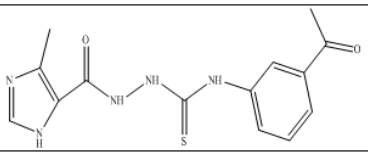
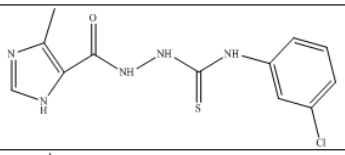
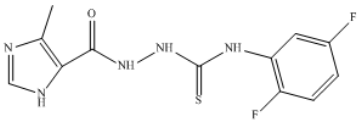
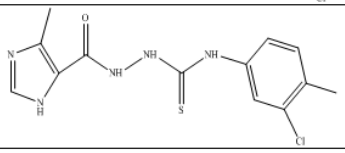
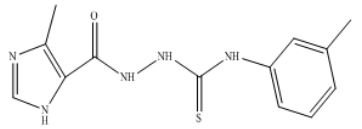
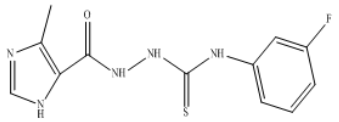
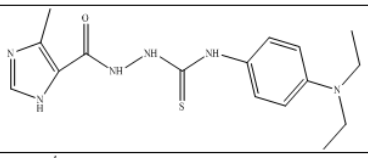
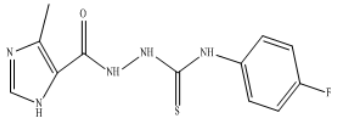
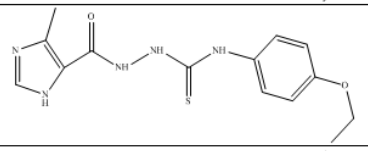
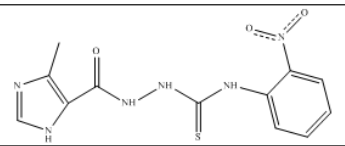
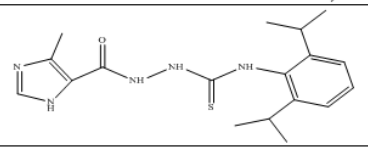
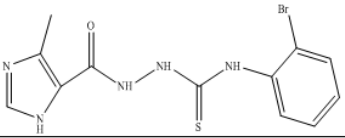
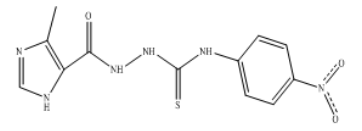
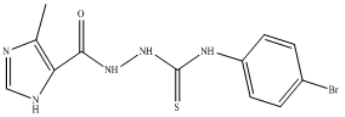
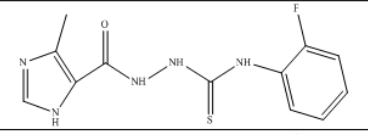
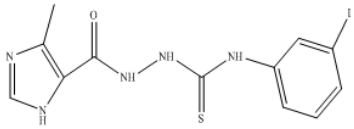
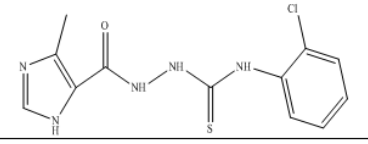
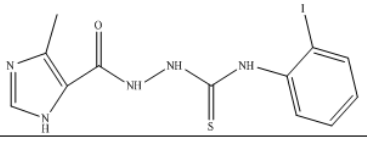
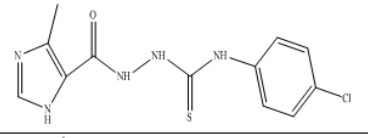
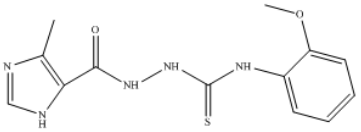
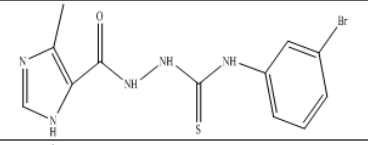
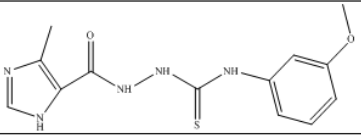
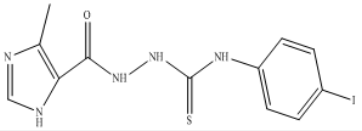
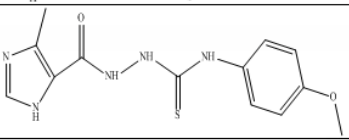
Codes	Molecules	IC50	Codes	Molecules	IC_{50} (µg/mL)
1		14.78	14		95.5
2		25.70	15		53.7
3		18.62	16		37.15
4		113.45	17		25.70
5		110.31	18		95.28
6		43.05	19		97.72
7		27.65	20		21.62
8		14.57	21		68.83
9		10.3	22		35.61
10		20.31	23		73.37
11		33.11	24		15.64
12		79.65	25		22.01
13		113.99			

Figure 1. Molecular structure, code and inhibitory concentration (IC_{50}) of the twenty-five (25) substituted imidazole-thiosemicarbazide derivatives

2. Materials and Methods

2.1. Computational Level of Theory

In order to predict the antibacterial activity of imidazole-thiosemicarbazides quantum chemical calculations were performed using Gaussian 09 software [15]. DFT methods are generally known to generate a variety of molecular properties [16,17,18] in QSAR studies. These increase the predictability of QSAR models while reducing the computational time and implication of cost in new drug conception [19,20]. The B3LYP/LanL 2 DZ level of theory was used to determine the molecular descriptors. The modeling was done using three methods. The first method is multilinear and nonlinear regression (MNLR) regression which are implemented in Excel [21] and XLSTAT [22] spreadsheets. The third method is the artificial neuron method which is included in the JMP Pro software [23].

2.2. Used Molecular Descriptors

In order to develop our QSAR model, some theoretical descriptors were determined. In particular, the energy gap (ΔE), dipole moment (μ), enthalpy of formation (ΔfH), bond length (D(C-S)) and lipophilicity (Logp).

The energy gap represents the energy difference between the boundary orbitals namely HOMO and LUMO, it is a parameter that gives more information about reactivity and stability of a molecule. The interactions are more favored, when the energy gap (ΔE) between the HOMO and LUMO is smaller [24]. They are thus stabilizing for the new chemical edifice formed [25]. The energy gap (ΔE) is calculated from equation (1):

$$\Delta E = E_{LUMO} - E_{HOMO} \quad (1)$$

The theory of boundary molecular orbitals allows us to analyze the reactivity of molecules which is made in terms of the interactions of reactants molecular orbitals [24]. But energies like E_{LUMO} et E_{HOMO} can be used to calculate many other parameters.

Lipophilicity expresses the affinity of a molecule for a lipidic environment such as oil, cell membrane, lipidic solvent [26]. This physico-chemical parameter is commonly measured by the distribution of neutral ou soluble molecule, between water and another immiscible solvent like n-octanol (or octan-1-ol) which is generally used [26,27,28,29]. Lipophilicity is evaluated from the value of logP. LogP. is equal to the logarithm of concentrations ratio of the studied substance in octanol and in water logP = log (Coct/Water). Indeed, positive and very high value of logP expresses the fact that the considered molecule is much more soluble in octanol than in water, which reflects its lipophilic character, Conversely, logP with negative value means that the considered molecule is hydrophilic. When logP is nul, it means that the molecule has the same solubility level in the two solvents. In this work, Chemsketch software [30] was used to determine the different values of logp. In practice we express the lipophilicity by the decimal logarithm of the partition coefficient logP. Thus :

If logP > 0; then P>1, the molecule is lipophilic. It is soluble in the lipidic phase so the molecule is not polar.

If logP < 0 then P<1, the molecule is hydrophilic. It is soluble in water so the molecule is then polar.

Another parameter related to the distribution of charges, is the dipole moment. This parameter is based on the existence of electrostatic dipoles. It is a global distribution of electric charges in a molecular system, so that the barycenter of positive charges does not coincide with that of negative charges. The dipole moment is a vector quantity. For a distribution of charges q_i at distance $\vec{r}_i$ each, the total dipole moment is established as follows:

$$\vec{\mu}_D = \sum q_i \vec{r}_i \quad (2)$$

The dipole moment makes it possible to describe the global polarity as well as the existence of interactions of molecular systems such as Van der Waals forces, and also to predict their solubility in polar solvents. The dipole moment is an important property that gives an idea of the reactivity of the molecule [31]. Furthermore, it indicates the stability of a molecule in water. Thus, a high dipole moment will reflect a low solubility in organic solvents and a high solubility in water [37,38].

The standard enthalpy of formation at temperature T of a chemical compound is the difference in enthalpy involved in the formation of one mole of this compound from simple, pure bodies, taken in the standard state and stable at the temperature considered T. The enthalpy of formation is calculated through the following formulas proposed by Otchersky et al [15]:

$$\Delta H_f^0(M,0K) = \sum_{atoms} x\Delta H_f^0(X,0K) - \sum D_0 \quad (3)$$

$$\Delta H_f^0(M,298K) = \Delta H_f^0(M,0K) + \left(H_M^0(298K) - H_M^0(0K)\right) - \sum_{atoms} x\left(H_X^0(298K) - H_X^0(0K)\right) \quad (4)$$

with:

$$\sum D_0 = \sum x\varepsilon_0 - \varepsilon_0(M) - \varepsilon_{ZPE} \quad (5)$$

$\sum \boldsymbol{D_0}$: Atomization energy;
$\boldsymbol{\varepsilon_0(M)}$: Total energy of the molecule;
$\boldsymbol{\varepsilon_{ZPE}}$: Zero point energy of the molecule;
$\boldsymbol{H_X^0(298K) - H_X^0(0K)}$: Enthalpy corrections of the atomic elements. These values are included in the table of Janaf [32].
$\boldsymbol{H_M^0(298K) - H_M^0(0K) = H_{corr} - \varepsilon_{ZPE}(M)}$: Enthalpy correction of the molecule
H_{corr} : Enthalpy of thermal correction.

The geometric descriptor used is the bond length d(C-S) in Armstrong (A°) (Figure 2). This descriptor is illustrated in the figure below around the imidazole-thiosemicarbazide ring.

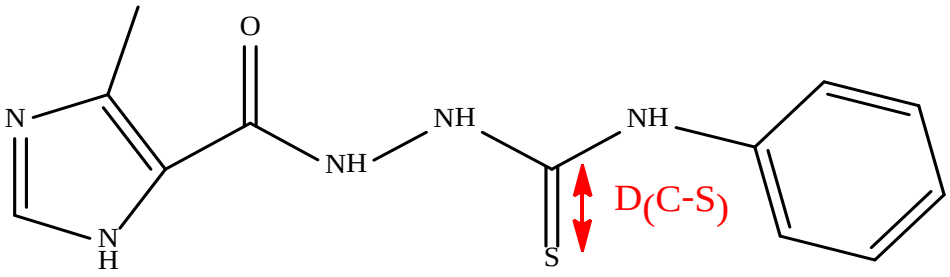


Figure 2. Geometric descriptor of the substituted imidazole-thiosemicarbazide derivatives: the bond length (D(C-S)) in Armstrong (A°)

Table 1. Correlation matrix between the different physico-chemical descriptors

Variables	ΔE_{Gap} (eV)	μ(D)	$\Delta_f H$ (kj/mol)	D(C-S) (A°)	Logp	pIC_{50}(mol/L)
ΔE_{Gap} (eV)	**1.0000**					
μ(D)	-0.1597	**1.0000**				
$\Delta_f H$ (kj/mol)	0.6488	-0.6198	**1.0000**			
D(C-S) (A°)	0.6657	0.3554	0.0602	**1.0000**		
Logp	0.3010	-0.0842	0.0532	0.0245	**1.0000**	
pIC_{50}(mol/L)	-0.2736	0.1086	0.1256	-0.1717	0.1048	**1.0000**

For all the descriptors studied, the analysis of the bivariate data, i.e. the calculation of the partial correlation coefficient between each pair of the set of descriptors is less than 0.70 (aij < 0.70), which means that these different descriptors are independent of each other [33,34]. To better understand the interdependence of the descriptors used, we present the partial correlation coefficient values aij of these descriptors in Table 1.

The partial correlation coefficients aij contained in Table 1 between the descriptor pairs (ΔE_{Gap}, μ), (ΔE_{Gap}, ΔfH), (ΔE_{Gap}, D(C-S)), (ΔE_{Gap}, logp), (μ, ΔfH), (μ, D(C-S)), (μ, logp), (ΔfH, D(C-S)), (ΔfH logp), and (D(C-S), logp) are less than 0.95 (aij < 0.70). This demonstrates the independence of the descriptors used to develop the models.

2.3. Estimation of the Predictive Capacity of a QSAR Model

The quality of a model is determined on the basis of various statistical criteria of analysis, including the coefficient of determination $\mathbf{R^2}$, the standard deviation **RMCE**, the correlation coefficients of the cross-validation $\mathbf{Q^2_{CV}}$ and Fischer **F**. $\mathbf{R^2}$, **S** and **F** relate to the fit of the calculated and experimental values. They describe the predictive ability within the limits of the model, and allow us to estimate the accuracy of the calculated values on the test set [35,36]. As for the cross-validation coefficient $\mathbf{Q^2_{CV}}$, it provides information on the predictive power of the model. This predictive power is called "internal" because it is calculated from the structures used to build the model. The correlation coefficient $\boldsymbol{R^2}$ gives an evaluation of the dispersion of the theoretical values around the experimental values. The quality of the modeling is better when the points are close to the fitting line [37]. The fit of the points to this line can be evaluated by the coefficient of determination.

$$R^2 = 1 - \frac{\sum\left(y_{i,\exp} - \hat{y}_{i,theo}\right)^2}{\sum\left(y_{i,\exp} - \hat{y}_{i,\exp}\right)^2} \qquad (6)$$

Where:

$y_{i,exp}$: Experimental value of anticancer activity

$\hat{y}_{i,theo}$: Theoretical value of anticancer activity and

$\bar{y}_{i,exp}$: Average value of the experimental values of the anticancer activity.

The closer the $\boldsymbol{R^2}$ value will be to 1 the more correlated the theoretical and experimental values are

On the other hand, the variance $\boldsymbol{\sigma^2}$ is determined by relation 7:

$$\sigma^2 = \mathrm{RMCE}^2 = \frac{\sum\left(y_{i,\exp} - y_{i,theo}\right)^2}{n-k-1} \qquad (7)$$

Where k is the number of independent variables (descriptors), n is the number of molecules in the test or training set and n-k-1 is the degree of freedom.

The standard deviation or RMCE is another statistical indicator used. It allows to evaluate the reliability and the precision of a model:

$$\mathrm{RMCE} = \sqrt{\frac{\sum\left(y_{i,\exp} - y_{i,theo}\right)^2}{n-k-1}} \qquad (8)$$

The Fisher **F** test is also used to measure the level of statistical significance of the model, i.e. the quality of the choice of descriptors making up the model.

$$\mathbf{F} = \frac{\sum\left(y_{i,theo} - y_{i,\exp}\right)^2}{\sum\left(y_{i,\exp} - y_{i,theo}\right)^2} * \frac{n-k-1}{k}. \qquad (9)$$

The coefficient of determination of the cross-validation $\mathbf{Q^2_{CV}}$, allows to evaluate the accuracy of the prediction on the test set. It is calculated using the following relation:

$$Q^2_{cv} = \frac{\sum\left(y_{i,theo} - \hat{y}_{i,\exp}\right)^2 - \sum\left(y_{i,theo} - y_{i,\exp}\right)^2}{\sum\left(y_{i,theo} - \hat{y}_{i,\exp}\right)^2} \qquad (10)$$

2.4. Criterion of Acceptance of a Model

The performance of a mathematical model, for Eriksson *et al* [38], is characterized by a value of $Q^2_{cv} > 0.5$ for a satisfactory model when for the excellent model $Q^2_{cv} > 0.9$. According to them, given a test set, a model will perform well if the acceptance criterion $R^2 - Q^2_{cv} < 0.3$ is met.

2.5. Statistical Analysis

2.5.1. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a data analysis tool that allows to explain the structure of correlations or covariances using linear combinations of the original data. Its use allows the interpretation of data in a reduced space [39]. It has been used to assess the relationships between the different variables measured, but above all to access their structure in order to group them by zone. The grouping by zone thus meets the objective of this approach, which is to correlate the classes of physico-chemical descriptors obtained at the sampling stations.

2.5.2. Ascending Hierarchical Classification (AHC)

The purpose of the Ascending Hierarchical Classification (AHC) is to partition a set of individuals into homogeneous classes (an individual is an observation and, in our case, these are samples) [40]. It organizes the individuals, defined by a certain number of variables and modalities, by grouping them hierarchically on a dendrogram. It aggregates those that are most similar to each other using measures of dissimilarity or distance between individuals to form classes. It is performed using data from individuals and variables. AMP allowed for a typology of samples based on energy gap (ΔE), dipole moment (μ), enthalpy of formation (ΔfH), bond length (D(C-S)) and lipophilicity (Logp).

2.5.3. Multiple Linear and Non-Linear Regressions (MLR and NLMR)

The statistical technique of multiple linear regression (MLR) is used to study the relationship between a dependent variable (Property) and several independent variables (descriptors). This statistical method minimizes the differences between the actual and predicted values. It was also used to select the descriptors used as input parameters in the multiple nonlinear regression (MNLR). As for the multiple non-linear regression (MNLR) analysis, it also allows to improve the structure-property relationship in order to quantitatively evaluate the property. It is the most common tool for studying multidimensional data. It is based on the following pre-programmed functions of *XLSTAT*:

$$y = a + (bx_1 + cx_2 + dx_3 + ex_4) + (fx_{12} + gx_{22} + hx_{32} + ix_{42}) \quad (11)$$

Où a, b, c, d,... are the parameters and, x_1, x_2, x_3, x_4,... are the variables.

2.5.4. Artificial Neuron Network (ANN)

Artificial neurons are an inspiration of the human biological neuron. To this end, they are made up of cells or neurons linked together by connections that allow them to send and receive signals from other cells. These neurons are mathematical models made up of several neurons, arranged in different layers. Generally, the network consists of three layers; an input layer, a hidden layer and an output layer, connected through a complex network [41,42]. The most commonly used networks are the Multi-Layer Perceptrons (MLP) whose neurons are generally arranged on layers [43]. In this work, the artificial neural network was obtained using the 5-3-1 multilayer perceptron network, i.e., the network consists of five (5) neurons in the input layer, three (3) neurons in the hidden layer and one (1) neuron in the output layer. The output layer consists of a sigmoid function. The architecture of the applied ANN models is presented in (Figure 3).

2.6. Acceptance Criteria of a QSAR Model

The twenty-five (25) molecules used in this study have Inhibitory Concentration (**IC**) ranging from 10.3 to 113.45 μg/mL. The median inhibitory concentration ($\mathbf{IC_{50}}$) is a measure of the effectiveness of a given compound in inhibiting a specific biological or biochemical function. Biological data are usually expressed as the opposite of the decimal-based logarithm of activity ($-\log_{10}(C)$) to obtain better mathematical values when structures are biologically active [44,45]. The antibacterial activity will be expressed by the antibacterial potential $\mathbf{pIC_{50}}$ defined by equation (12):

$$\mathbf{pIC}_{50} = -\log_{10}\left(\frac{\mathrm{IC}_{50}}{\mathrm{M}} * 10^{-6}\right) \quad (12)$$

Where M is the molecular weight of the compound in g/mol and **IC** is the Inhibitory Concentration in μg/mL.

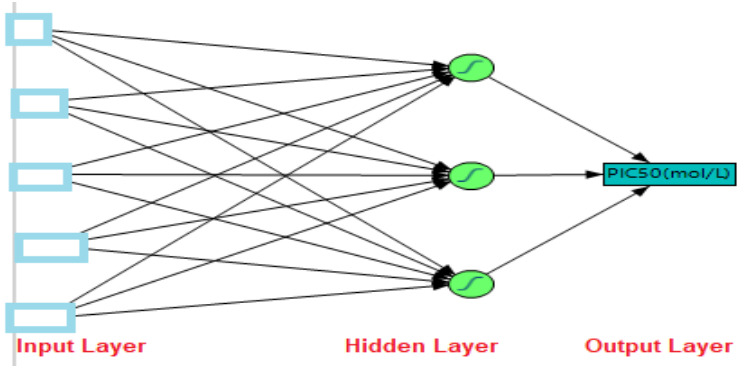


Figure 3. Diagram of the structure of a multilayer perceptron

2.7. Domain of Applicability (DA)

The domain of applicability of a QSAR model is the physicochemical, structural, or biological space in which the model equation is applicable to make predictions for new compounds [46]. It corresponds to the region of chemical space including the compounds in the training set and similar compounds, which are close in the same space [47]. Indeed, the model, which is built on the basis of a limited number of compounds, by relevant descriptors, chosen among many others, cannot be a universal tool to predict the activity of any other molecule with confidence. It appears necessary, even mandatory, to determine the DA of any QSAR model. This is recommended by the Organization for Economic Cooperation and Development (OECD) in the development of a QSAR model [48]. There are several methods for determining the domain of applicability of a model [47]. Of these, the approach used in this work is the Cook distance. The Cook distance measures the effect of deleting a data item. Data with large residuals (outliers) and/or high leverage can distort the outcome and accuracy of a regression. The threshold Cook distance is determined by the following expression [49,50].

$$D_i = 4 / N - k - 1 \quad (13)$$

With N: the number of observations, k: the number of descriptors defined by the model.

3. Results and Discussion

All the values of the physico-chemical descriptors of nineteen (19) compounds of the test set and six (6) other compounds of the validation set are presented in Table 2.

Table 2. Physico-chemical descriptors and experimental pIC_{50}(mol/L) of the test and validation sets

MOLECULES	ΔE(eV)	μ(D)	ΔfH	D(C-S)	Logp	pIC_{50}(mol/L)
Test Set						
1	2.9376	5.8433	-1700.4687	1.7436	1.9000	0.0057
2	4.5487	4.0000	-1398.3878	1.7471	2.5400	0.0051
3	4.5835	4.2725	-1440.2425	1.7483	3.0000	0.0053
4	4.5443	4.0769	-1441.5402	1.7470	2.0000	0.0032
5	4.5911	2.9225	-1441.4235	1.7479	1.9600	0.0033
6	2.8794	3.3570	-1700.9149	1.7392	1.7600	0.0043
7	4.5071	2.6821	-1389.0517	1.7468	2.2000	0.0044
8	4.5443	2.8239	-1389.6516	1.7479	2.6800	0.0052
9	4.5615	3.6710	-1380.4971	1.7476	2.9800	0.0050
10	4.5283	2.8116	-1383.1332	1.7480	2.4600	0.0042
11	4.4619	4.7381	-1525.9971	1.7562	1.4100	0.0048
12	4.5740	5.0431	-1524.2859	1.7508	1.7700	0.0036
13	4.5941	5.1658	-1524.0331	1.7510	1.4600	0.0031
14	3.9525	3.0937	-1557.0649	1.7477	1.3100	0.0032
15	4.3849	2.8569	-1470.5209	1.7505	2.0300	0.0041
16	4.5688	4.1872	-1450.4552	1.7506	1.9700	0.0049
17	4.1893	6.6773	-1749.5335	1.7531	2.7100	0.0046
18	4.5941	5.3001	-1566.0785	1.7511	1.9900	0.0032
19	4.6608	3.8111	-1640.2132	1.7506	4.1900	0.0028
Validation Set						
20	3.2524	5.6033	-1702.2508	1.7423	1.9700	0.0052
21	4.5193	5.1030	-1438.9405	1.7527	1.4800	0.0040
22	4.4940	2.4982	-1398.0734	1.7461	2.0200	0.0047
23	4.5503	2.8481	-1398.6325	1.7474	2.5000	0.0037
24	4.5550	3.8575	-1389.3629	1.7473	2.7200	0.0051
25	4.5174	2.9332	-1383.6879	1.7470	2.9400	0.0041

3.1. Typology of Physicochemical Descriptors of Imidazole-Thiosemicarbazide Derivatives

The correlation circle (Figure 4), the Cartesian diagrams according to F1 and F2 (Figure 5) and the dendrogram of the compounds are shown below.

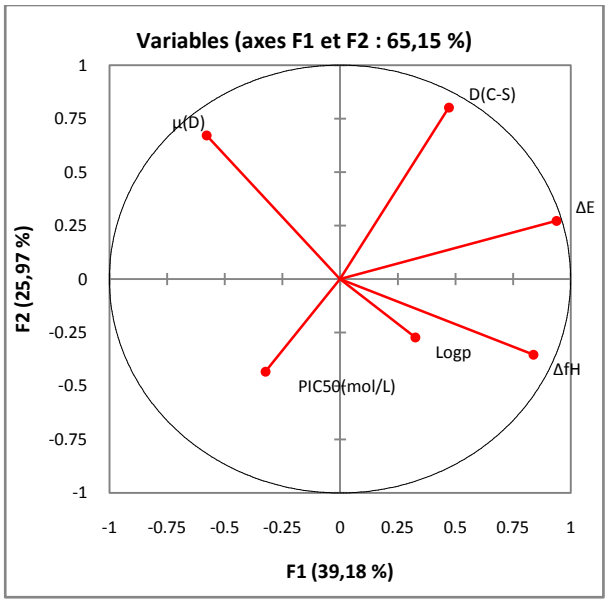


Figure 4. Correlation circle of descriptors descriptors according to F1 x F2

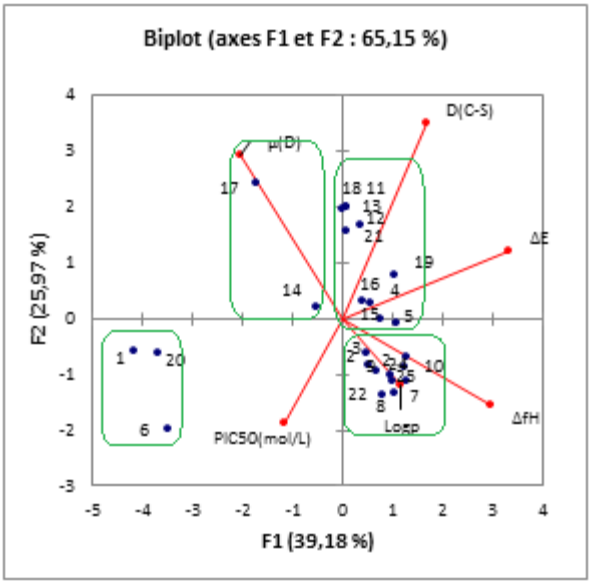


Figure 5. Cartesian diagram according to F1 and F2: correlation between the physicochemical descriptors used and the inhibitory concentration $\mathbf{IC_{50}}$

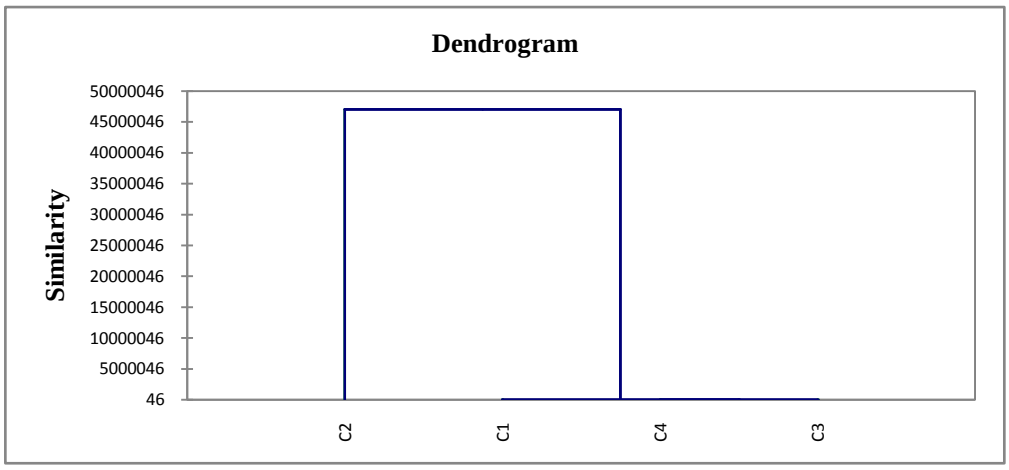


Figure 6. Dendrogram of molecules

PCA data matrix groups the mean values of five (5) variables representing physicochemical descriptors and twenty-five (25) molecules. The resulting matrix provides information on the negative or positive correlation between the variables. The examination of the community circle in Figure 4 associated with the analysis of the factorial structure of the PCA, indicate that two (02) principal components represented by F1 which corresponds to 39.18% of the explained variance and F2 with 25.97% of the explained variance. The two factors F1 and F2 totaling 65.15% of the total variance are sufficient to interpret all the PCA data. Each variable is associated with its factorial weight. Thus, the energy gap (0.88) and the enthalpy of formation (0.70) are positively correlated to **factor 1**, while **factor 2** is positively correlated to the dipole moment (0.45) and the bond length (D(C-S)) (0.64). It appears from this analysis that the F1 factor (39.18%) correlates better with the energy gap and the enthalpy of formation with very high coefficients.

Cartesian diagram in Figure 5 shows that compounds **1**, **20**, and **6** negatively correlated with **factor 2** with an inhibitory potential $\mathbf{pIC_{50}}$. Molecules **14** and **17** are positively correlated with dipole moment μ(D) and **factor 1**. It must be said that compounds **18,13, 11, 12** and **21** are positively correlated with **factor 1** and strongly correlated with distance D(C-S). The molecules **16**, **19** and **4** are correlated to the energy gap ΔE and positively **factor 1**. The same is true for the molecules **15** and **5**. The last group of compounds **(2, 3, 6, 7, 8, 9, 10, 11, 12, 13, 18, 21, 22, 23, 24** and **25**) correlate with logp and enthalpy of formation ΔfH. The latter classification is positively influenced with **factor 2** and negatively with **factor 1**.

The illustration of the distribution of the compounds in the Cartesian diagram is also reflected in the dendrogram of the molecules (Figure 6). In Figure 6, the compounds are grouped into 4 classes (**C1, C2, C3** and **C4**) characterizing the physicochemical descriptors. Thus, the class **C1** (**1**, **6**, and **20**) which consists of three molecules. We have the class **C2** composed of two(2) molecules (**14** and **17**) strongly influenced by the dipole moment μ(D) and the class **C3** constituted by ten(10) molecules (**4, 5, 11, 12, 13, 15, 16, 18, 19** and **21**) influenced by the distance D(C-S). The last class **C4** grouping the remaining ten(10) molecules (**2, 3, 7, 8, 9, 10, 22, 23, 24 and 25**) strongly

correlated with the logp descriptors and the enthalpy of formation ΔfH.

3.2. Multilinear Regression Models (MLR) and Nonlinear Multiple Regression (NLMR)

In a model equation, the negative or positive sign of the coefficient of a descriptor reflects the proportionality effect between the evolution of the inhibitory concentration **IC_{50}** and this physicochemical parameter in the regression equation. Thus, the negative sign indicates that when the value of the descriptor is high, the inhibitory concentration **IC_{50}** decreases, whereas the positive sign reflects the opposite effect. The equations of the best obtained MLR and NLMR models, as well as the statistical indicators will be presented in Table 3 below. These statistical indicators will be followed by Figure 6 and Figure 7 presenting the fit line of the experimental and theoretical data of the **pIC_{50}** inhibitory potentials of the test (blue dots) and validation (red dots) sets of the model. It should be noted that these models were built using the same test and validation sets in Table 1.

Model MLR

$$pIC_{50}^{\exp}(\mathbf{mol/L}) = -0.39454 - 0.00389*\Delta E + 0.00049*\mu(D) + 0.00001*\Delta fH + 0.24785*D(C-S) + 0.00094*logp$$

Model NLMR

$$pIC_{50}^{\exp}(\mathbf{mol/L}) = -23.55095 - 0.00044*\Delta E + 0.00059*\mu(D) + 7.38617E-6*\Delta fH - 27.14561*D(C-S) + 0.00295*Logp - 0.00039*\Delta E^2 - 0.00002*\mu(D)^2 - 1.61252E - 9*\Delta fH^2 + 7.82886*D(C-S)^2 - 0.00038*Logp^2$$

Table 3. Statistical analysis report of the IC_{50} inhibitory potential of substituted imidazole-thiosemicarbazide

Statistical indicators of multilinear regression	Model MLR	Model NLMR
Number of observations **N**	19	19
Coefficient of determination $\mathbf{R^2}$	0.8346	0.8897
Standard deviation **RMCE**	0.00038	0.0004
Fischer's test **F**	85.794	137.07
Cross-validation correlation coefficient Q_{CV}^2	0.8347	0.8897
$\mathbf{R^2}$ - Q_{CV}^2	0.0000	0.0000
Confidence level **α**	> 95%	

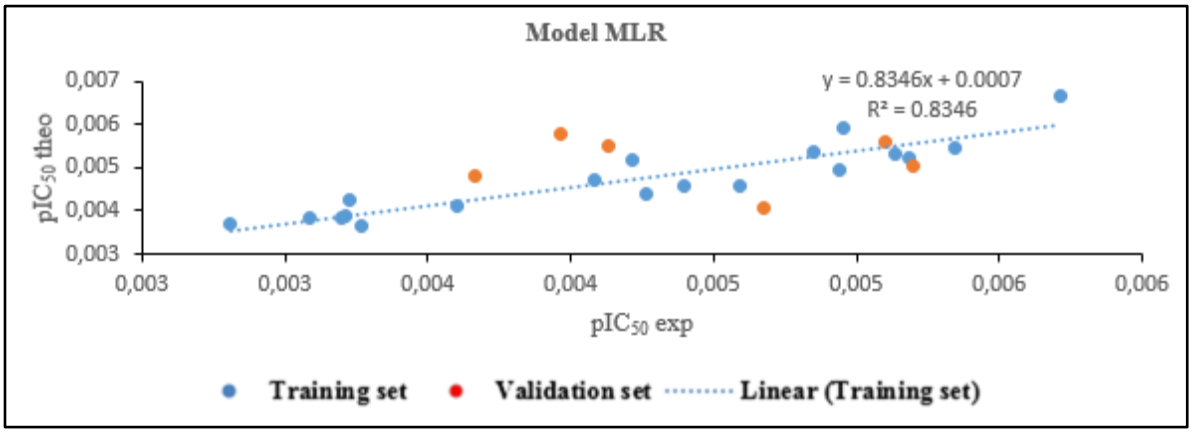


Figure 7. Regression line of the MLR model

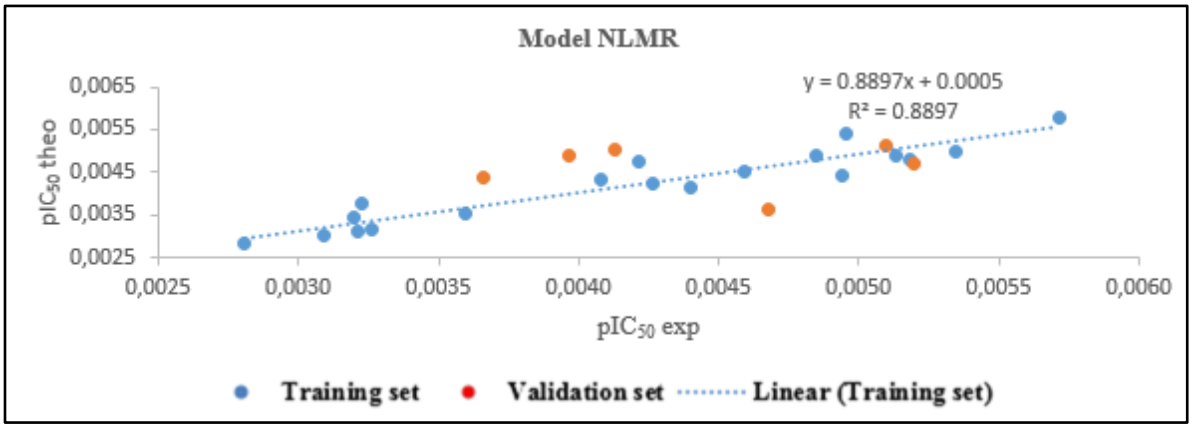


Figure 8. Regression line of the NLMR model

The negative correlation coefficient between the energy gap **ΔE** and the inhibitory potential **pIC_{50}** indicates that these two variables are inversely proportional i.e. decreasing the energy gap favors increasing the inhibitory concentration **IC_{50}** in both models. The significance of these models is shown by the high values of the Fischer **F** coefficient which are 85.794 and 137.07 for the MLR and NLMR models respectively. Moreover, the robustness of these models is reflected by the cross-validation correlation coefficient **Q^2cv** which are also **0.8347** and **0.8897**. These MLR and NLMR models are all acceptable because **R^2 - Q^2cv** =0.000<0.3. The regression lines established by the experimental and theoretical data of pIC_{50} inhibitory potential of the test and validation sets for MLR and NLMR models are shown in Figure 7 and Figure 8.

The external validation test was verified by calculating the **pIC_{50} theo/ pIC_{50} exp** ratio of the MLR and NLMR models. These values are confined in Table 4.

The values of the ratio **pIC_{50}** theo/ **pIC_{50}** exp of the model validation set tend to unity (Table 4) reflecting the good correlation between the theoretical and experimental **pIC_{50}** inhibitory potential of the observations. These models are therefore acceptable for the prediction of the inhibitory concentration IC_{50} of substituted imidazole-thiosemicarbazide derivatives.

Moreover, the low values of the standard error (**RMCE**) which are **0.00038** and **0.0004** for the MLR and NLMR models respectively attest to the good similarity between the predicted and experimental values (Figure 9 and Figure 10). These curves show a similar evolution of the data from these two models for the prediction of the inhibitory concentration **IC_{50}** of substituted imidazole-thiosemicarbazide derivatives despite some deviations recorded.

Table 4. Values of the ratio between theoretical and experimental Transparencies of the validation set

	Validation Set				
	Model MLR pIC_{50} theo	**Model NLMR pIC_{50} theo**	**pIC_{50} exp**	**Model MLR pIC_{50} theo/ pIC_{50} exp**	**Model NLMR T pIC_{50} theo/ pIC_{50} exp**
20	0.005	0.005	0.005	0.865	0.904
21	0.005	0.005	0.004	1.325	1.225
22	0.004	0.004	0.005	0.766	0.766
23	0.004	0.004	0.004	1.162	1.189
24	0.005	0.005	0.005	1.000	1.000
25	0.005	0.005	0.004	1.220	1.244

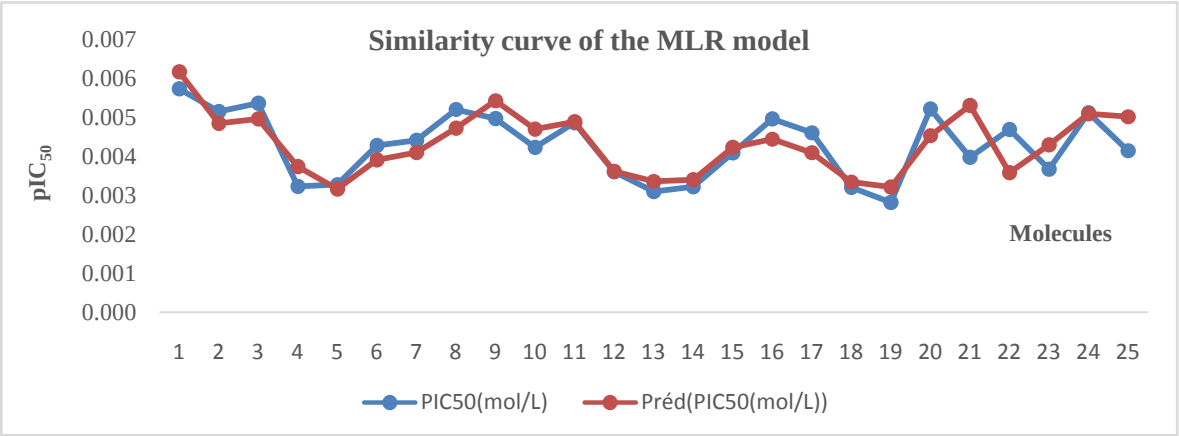


Figure 9. Similarity curve of the experimental and predicted values of the MLR model

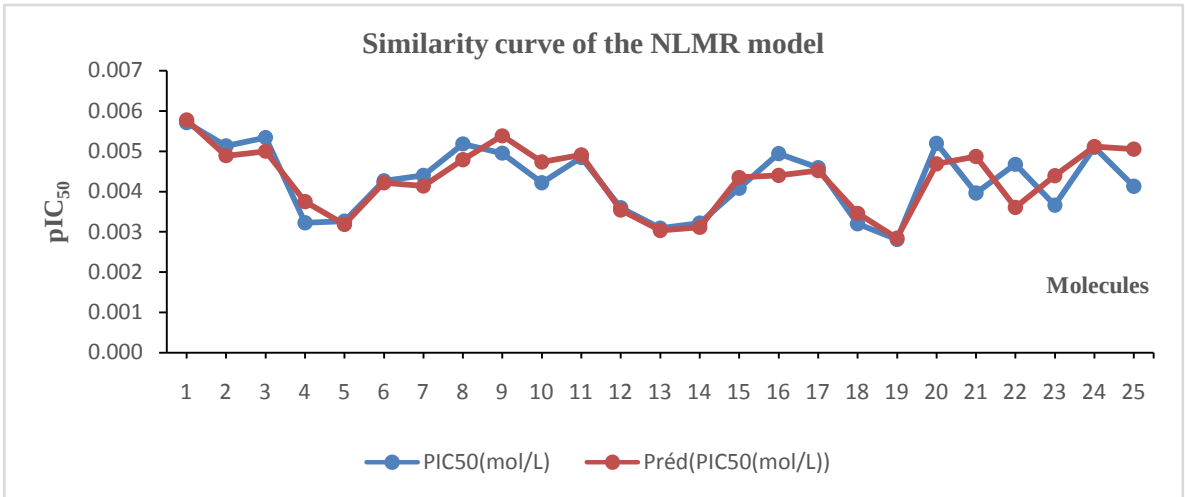


Figure 10. Similarity curve of the experimental and predicted values of the NLMR model

The values of the **pIC_{50} theo/ pIC_{50} exp** ratio of the validation set that tend to unity (Table 3) reflect the good correlation between the theoretical and experimental inhibitory concentrations **IC_{50}** of the observations. Therefore, these models are acceptable for the prediction of the **IC_{50}** inhibitory concentration of substituted imidazole-thiosemicarbazide derivatives.

3.3. Artificial Neural Network (ANN)

The equation of the best model of the Artificial Neural Network obtained, as well as the statistical indicators will be proposed in Table 5 below. This equation will be followed by 3 other equations (X1, X2, X3) obtained from the hidden layer. These three equations from the hidden layer are hyperbolic tangent functions dependent on the molecular descriptors (ΔE, μD, ΔfH, D(C-S, logp).

The statistical indicators of the ANN model will be followed by Figure 11 presenting the fit line of the experimental and theoretical data of the pIC50 inhibitory potentials of the test and validation sets. Here these fitting lines of the test set and the validation set include twenty (20) and five (5) molecules respectively.

$$pIC_{50}^{\exp}\left(\mathbf{mol/L}\right) = 0.0046 + 0.0098 * X_1 - 0.0089 * X_2 + 0.0125 * X_3$$

$$\mathbf{X_1} = \mathrm{TanH}\left(0.5 * \begin{pmatrix} -295.751 - 0.4305 * \Delta E \\ +1.6232 * \mu(D) - 0.0015 * \Delta fH \\ +165.4793 * D(C-S) \\ -0.0494 * Logp \end{pmatrix}\right)$$

$$\mathbf{X_2} = \mathrm{TanH}\left(0.5 * \begin{pmatrix} 111.5418 + 0.3003 * \Delta E \\ +0.3599 * \mu(D) + 0.0048 * \Delta fH \\ -61.2682 * D(C-S) \\ +0.023 * Logp \end{pmatrix}\right)$$

$$\mathbf{X_3} = \mathrm{TanH}\left(0.5 * \begin{pmatrix} 223.1003 + 0.2982 * \Delta E \\ -0.7503 * \mu(D) + 0.0076 * \Delta fH \\ +119.0389 * D(C-S) \\ +0.1874 * Logp \end{pmatrix}\right)$$

Table 5. Statistical analysis report of the IC_{50} inhibitory potential of substituted imidazole-thiosemicarbazide derivatives

Statistical indicators of the Neural Network	Training Set	Validation Set
Number of observations **N**	20	5
Coefficient of determination $\mathbf{R^2}$	0.9291	0.7330
Standard deviation **RMCE**	0.00023	0.00036
Fischer's test **F**	> 95%	

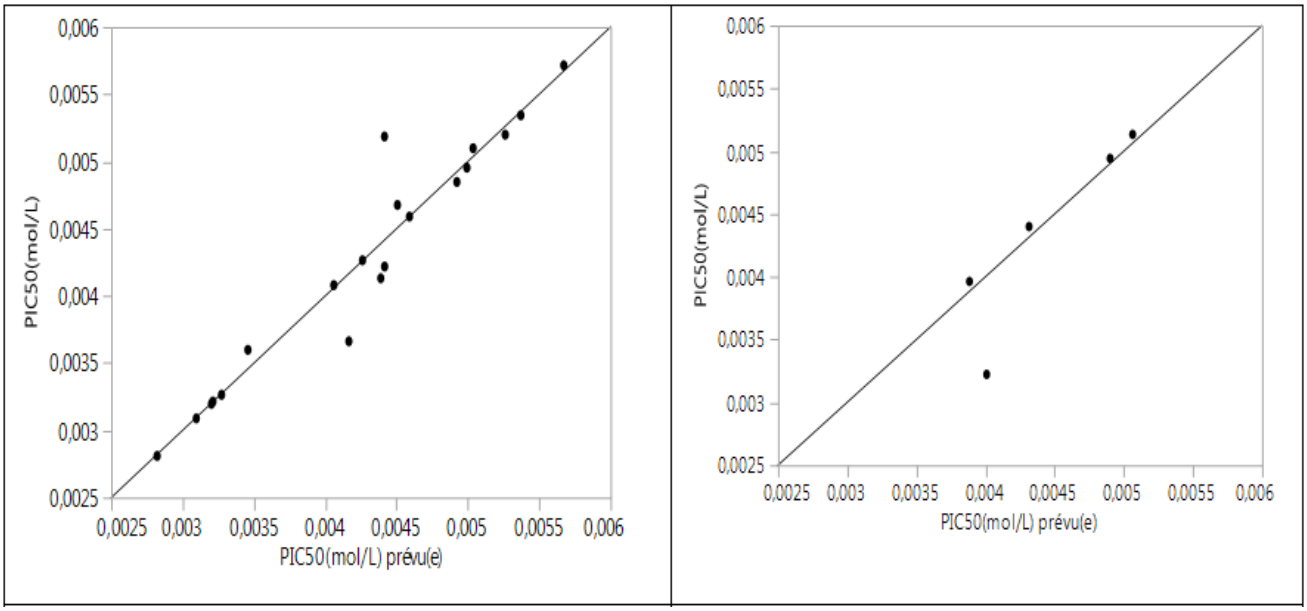


Figure 11. Graphs of observed values versus predicted values of test and validation sets

Of the three models, the model obtained by the statistical **ANN** method has a much better predictive capacity. However, since this model is a function of five physico-chemical descriptors, it is essential to determine the contribution of each one in the prediction of the inhibitory concentration **IC_{50}**. Indeed, the knowledge of this contribution makes it possible to establish the order of priority of the various descriptors and to define the choice of the parameters to be optimized for the good prediction and comprehension of the inhibitory concentration **IC_{50}** of the substituted imidazole-thiosemicarbazide derivatives.

3.4. Analysis of the Contribution of the Descriptors

The contributions of the five physico-chemical descriptors in predicting the inhibitory concentration **IC_{50}** of substituted imidazole-thiosemicarbazide derivatives were illustrated by the normalized coefficients and shown by Figure 12.

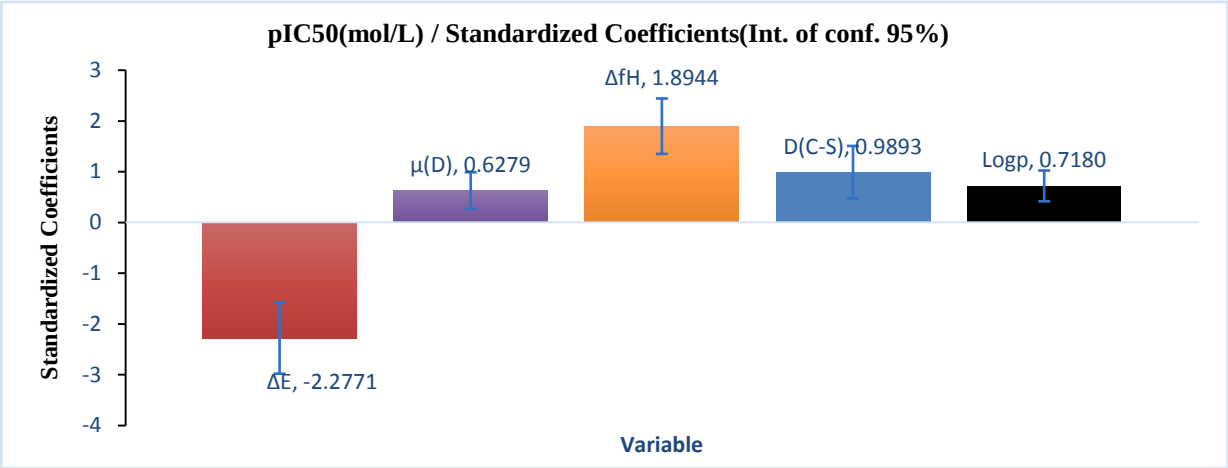


Figure 12. Contribution of descriptors in the models

According to the contribution of these descriptors, the energy gap (ΔE) displays the highest normalized coefficient (-2.2771) followed by the enthalpy of formation (ΔfH) with 1.8944 and the dipole moment (μD) holds the lowest coefficient (0.6279) compared to the other descriptors. It should be noted that the energy gap (ΔE) is the most influential physicochemical descriptor. Thus, to improve the inhibitory concentration $\mathbf{IC_{50}}$ in the synthesis of new substituted imidazole-thiosemicarbazide derivatives, the energy gap (ΔE) must be played to the maximum extent.

3.5. Domain of Applicability of the Model

Here the analysis of the applicability domain in our study being based on the determination of the Cook threshold distance. This parameter was evaluated at the level of the test set.

$$D = 4/N - k - 1 = 4/(19 - 5 - 1) = 0.307 \qquad (14)$$

The cook distances of our molecules are listed in Table 6 and are illustrated in Figure 12.

Here, we observe a large influence of compounds **8**, **16** and **17** on the parameter estimation and predictions. To a lesser extent, the molecules **2**, **3**, **6** and **7**. Compounds **8**, **16** and **17** are considered to be outside the applicability. These erroneous predictions could probably be attributed to erroneous experimental data or the structure of these outliers.

Table 6. Cook distance of the molecules in the test set

Molecules	Cook's distance
Jeu d'essai	
1	-1.0117
2	**0.7085**
3	**0.9312**
4	-1.1759
5	0.2585
6	**0.8453**
7	**0.7266**
8	**1.1027**
9	-1.0596
10	-1.0820
11	-0.0563
12	-0.0083
13	-0.6009
14	-0.4139
15	-0.3188
16	**1.1997**
17	**1.1816**
18	-0.3123
19	-0.9143

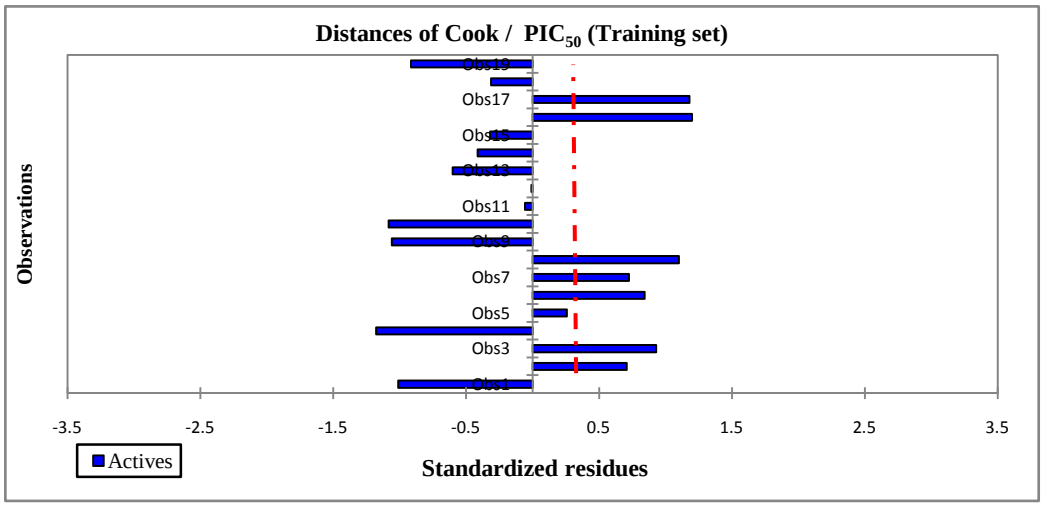


Figure 12. Cook's distance from the test set

4. Conclusion

At the end of this study we were able to show a relationship between the inhibitory concentration **IC_{50}** of *Toxoplasma gondii* and the physicochemical descriptors of imidazole-thiosemicarbazides. The descriptors energy gap (ΔE), dipole moment (μ), enthalpy of formation (ΔfH), bond length (D(C-S)), and lipophilicity (Logp) help explain and predict the antiparasitic activity of imidazole-thiosemicarbazides. Statistical methods such as principal component analysis (**PCA**), Ascending Hierarchical Classification (**AHC**), multilinear and nonlinear regression and artificial neuron method were employed. The study of the robustness of the three (3) models (**MLR**, **NLMR** and **ANN**) constructed shows a good predictive capacity. Moreover, compared to the **MLR** and **NLMR** models, the **ANN** model (**R^2** =0.9291; **RMCE** = 0.00023) is better and is an effective tool for predicting pest activity. The external validation test showed a good correlation between the theoretical and experimental **pIC_{50}** inhibitory potential of the observations. This translates that these models are therefore acceptable for the prediction of the inhibitory concentration **IC_{50}** of substituted imidazole-thiosemicarbazide derivatives. It is apparent from the contribution of the descriptors that a decrease in the energy gap (ΔE) could promote a significant improvement in the *Toxoplasma gondii* **IC_{50}** inhibitory concentration of the substituted imidazole-thiosemicarbazide analogues.

References

[1] J. MCAULEY, K. M. BOYER, D. PATEL, M. METS, C. SWISHER, N. ROIZEN, C. WOLTERS, L. STEIN, M. STEIN et W. SCHEY, "Early and longitudinal evaluations of treated infants and children and untreated historical patients with congenital toxoplasmosis: The Chicago Collaborative Treatment Trial," *Clinical Infectious Diseases* , vol. 19, p. 38-72, 1994.

[2] F. DAFFOS, V. MIRLESSE, P. HOHLFELD, F. JACQUEMARD, P. THULLIEZ et F. FORESTIER, "Toxoplasmosis pregnancy.," *Lancet* , vol. 344, p. 541, 1994.

[3] U. Tendeur, A. Heckeroth et L. Weiss, "Toxoplasma gondii: des animaux aux humains.," *Int. J. Parasitol.*, vol. 30, pp. 1217-1258, 2000.

[4] P. Agata, W. Lidia, B. Adrian, S. Edyta, W. Monika, T. Nazar, H. Anna, H. Miroslaw et D. Katarzyna, "Discovery of Potent and Selective Halogen-Substituted Imidazole-Thiosemicarbazides for Inhibition of Toxoplasma gondii Growth In Vitro via Structure-Based Design," *Molecules*, vol. 24, p. 1618, 2019.

[5] A. Paneth, L. Węglińska, A. Bekier, E. Stefaniszyn, M. Wujec, N. Trotsko, A. Hawrył, M. Hawrył et K. Dzitko, "Discovery of Potent and Selective Halogen-Substituted Imidazole-Thiosemicarbazides for Inhibition of Toxoplasma gondii Growth In Vitro via Structure," *Molecules*, vol. 24, pp. 1-14, 2019.

[6] W. R. Sherman, "5-Nitro-2-furyl-substituted 1,3,4-Oxadiazoles, 1,3,4-Thiazoles, and 1,3,5-Triazines," *The Journal of Organic Chemistry*, vol. 1, pp. 88-95, 1961.

[7] A. Jalilian, S. Sattari, M. Bineshmarvasti, A. Shafiee et M. A. Daneshtalab, *Pharm. Med. Chem.* , vol. 333, p. 347, 2000.

[8] C. B. Chapleo, P. L. Myers, A. C. B. Smith, M. R. Stillings, I. F. Tulloch et D. S. Walter, "Substituted 1,3,4-thiadiazoles with anticonvulsant activity. 4.Amidines," *Journal of medicinal chemistry* , vol. 1, n° %131, pp. 7-11, 1988.

[9] E. F. Da Silva, M. M. Canto-Cavalheiro, V. R. Braz, L. Cysne-Finkelstein et L. L. Leon, "Synthesis, and biological evaluation of new 1,3,4-thiadiazolium-2-phenylamine derivates against Leishmania amazonensis promastigotes and amastigotes," *European Journal of medicinal chemistry*, vol. 12, n° %137, pp. 979-984, 2002.

[10] N. Grynberg, A. Santos et A. Echevarria, "Synthesis and in vivo antitumor activity of new heterocyclic derivatives of the 1,3,4-thiadiazolium-2-aminide class," *Anti-cancer drugs*, vol. 8, n° %11, pp. 88-91, 1997.

[11] T. I. Oprea, "Chemoinformatics in Drug Discovery," *Ed. WILEY-VCH Verlag.*, 2005.

[12] E. A. Rekka et P. N. Kourounakis, "Chemistry and Molecular Aspects of Drug Design and Action," *Ed. Taylor & Francis Group*, 2008.

[13] S. M. Free et J. W. Wilson, "A Mathematical Contribution to Structure-Activity Studies," *J. Med. Chem.*, vol. 7, pp. 395-399, 1964.

[14] C. Hansch et T. Fujita, "ρ – σ – π, analysis: method for correlation of biological activity and chemical structure," *J. Am. Chem. Soc.*, vol. 86, pp. 1616-1626, 1964.

[15] T. Partal et H. K. Cigizoglu, "Estimation and forecasting of daily suspended sediment data using wavelet-neural networks," *Journal of Hydrology*, vol. 358 , n° %134, pp. 317-331, 2008.

[16] K. Hornik, M. Stinchcombe et H. White, "Multilayer feed-forward networks are universal approximators," *Neural Networks computation*, vol. 2, pp. 359-366, 1989.

[17] C. Roussillon, "Prévision de la température par les Réseaux de Neurones Artificiels," Université Victor Hugo Besancon, France, 2004.

[18] JMPPro13, *Statistical Discovery*, Scintilla: SAS institute Inc., 1998-2014.

[19] X. V. 2. C. Addinsoft, *XLSTAT and Addinsoft are Registered Trademarks of Addinsoft.*, 2014, pp. 1995-2014 .

[20] M. Excel, 2016.

[21] M. J. Frisch, G. W. Trucks, H. B. Schlegel et G. E. Scuseria, "Gaussian 09, Revision A.02," Gaussian, Inc., Wallingford CT, 2009.

[22] P. K. Chattaraj, A. Cedillo et R. G. Parr, *J. Phys. Chem.*, vol. 103, p. 7645, 1991.

[23] P. W. Ayers et R. G. Parr, *J. Am Chem. Soc*, vol. 122, p. 2000, 2010.

[24] F. De Proft, J. M. L. Martin, P. Geerlings et :. :, *Chem. Phys Let.*, vol. 250, p. 393, 1996.

[25] C. Hansch, P. G. Sammes et J. B. Taylor, "in:Comprehensive Medicinal Chemistry," *Computers and the medicinal chemist*, vol. 4, pp. 33-58, 1990.

[26] R. Franke, "Theoretical Drug Design Methods," *Elsevier*, 1984.

[27] G. W. Snedecor et W. G. Cochran, "Methods, Statistical," *Oxford and IBH: New Delhi, India*, p. 381, 1967.

[28] N. J.-B. Kangah, M. G.-R. Koné, C. G. Kodjo, B. R. N'guessan, A. L. C. Kablan, S. A. Yéo et N. Ziao, "Antibacterial Activity of Schiff Bases Derived from Ortho Diaminocyclohexane, Meta-Phenylenediamine and 1,6-Diaminohexane: Qsar Study with Quantum Descriptors,," *International Journal of Pharmaceutical Science Invention*, vol. 6, n° %13, pp. 38-43, 2017.

[29] E. X. Esposito, A. J. Hopfinger et J. D. Madura, "Methods for Applying the Quantitative Structure-Activity Relationship Paradigm,," *Methods in Molecular Biology*, vol. 275, pp. 131-213., 2004.

[30] M. Frisch, G. Trucks, H. Schlegel et G. Scuseria, "Revision A.02," chez *Gaussian 09*, Wallingford CT, Gaussian, Inc., 2009.

[31] M. W. Chase, C. A. Davies, J. R. Downey, D. J. Frurip, R. A. McDonald et A. N. Syverud, "JANAF Thermochemical Tables," *J. Phys. Ref.* , vol. 14, n° %11, 1985.

[32] S. Chaltterjee, A. Hadi et B. Price, "Regression Analysis by Examples," *Wiley VCH: New York*, 2000.

[33] H. Phuong, "Synthèse et étude des relations structure/activité quantitatives (QSAR/2D) d'analoguesBenzo[c]phénanthridiniques," France, 2007.

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

[35] J. N'dri, M.-G. Koné, C. KODJO, S. AFFI, A. KABLAN, O. OUATTARA et D. Soro, "Quantitative Activity Structure Relationship (QSAR) of a Series of Azet idinones Derived from Dap-sone by the Method of Density Functional Theory (DFT)," *IRA International Journal of Applied Sciences (ISSN 2455-4499)*, vol. 8, n° %12, pp. 55-62, 2017.

[36] K. R. Clarke et M. Ainsworth, "A method of linking multivariate community structure to environmental variables.," *Marine Ecology*, vol. 92, pp. 205-219, 1993.

[37] B. Escofier et J. Pagès, Analyses factorielles simples et multiples: Objectifs, méthodes et interprétation., vol. 318, Paris: Dunod, 2008.

[38] T. Oprea, Chemoinformatics in drug discovery, Allemagne: Ed. Wiley-VCH Verlag, 2005.

[39] E. Rekka et P. Kourounakis, Chemistry and molecular aspects of drug design and action, Etats Unies: LLC. Ed. Taylor & Francis Group, 2008.

[40] L. Eriksson, J. Jaworska, A. Worth, M. D. Cronin, R. M. Mc Dowell et P. Gramatica, "Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs," *Environmental Health Perspectives*, vol. 111, n° %110, pp. 1361-1375, 2003.

[41] G. Dreyfus, " Réseaux de neurones artificiels," Toulouse, France., 1998.

[42] G. Dreyfus, J. Martinez, M. Samuelides, M. Gordon, F. Badran, S. Thiria et L. Herault, Réseaux de Neurones Artificiels. 2 édition, New York, USA: Groupe Eyrolles, 2002, p. 374 .

[43] I. Rivals, "Modélisation et commande de processus par réseaux de neurones artificiels. Application au pilotage d'un véhicule autonome.," France, 1995.

[44] J. M. Poveda, A. Garcia, P. J. Martin-Alvarez et L. Cabezas, "Application of partial least squares (PLS) regression to predict the ripening time of Manchego cheese," *Food Chemistry,* vol. 84, n° %11, pp. 29-33, 2004.

[45] C. Faur-Brasquet et P. Le Cloirec, "Modelling of the flow behaviour of activated carbon cloths using a neural network approach," *Chemical Engineering and Processing,* vol. 2, n° %142, pp. 645-652, 2003.

[46] V. Labet, "Etude Théorique de Quelques Aspects de la Réactivité des Bases de l'ADN-Définition de nouveaux outils théoriques d'étude de la réactivité chimique. Chemical Sciences.," 2009.

[47] B. Samir, "Etude théorique et expérimentale des réactions de cycloaddition Diels&Alder et 1,3- dipolaire," 2013.

[48] O. Dorosh et Z. Kisiel, "Electric Dipole Moments of Acetone and of Acetic Acid measured in Supersonic Expansion," *Acta Physica Polonica A,* vol. 112 , 2007.

[49] E. Rutkowska, K. Pajak et K. Jozwiak, "Lipophilicity - Methods of Determination and its Role in Medicinal Chemistry," *Acta Poloniae Pharmaceutica - Drug Research,* vol. 70, n° %11, pp. 3-18, 2013.

[50] A. Cozma, V. Zaharia, A. Ignat, S. Gocan et N. Grinberg, "Prediction of the Lipophilicity of Nine New Synthesized Selenazoly and Three Aroyl–Hydrazinoselenazoles Derivatives by Reversed-Phase High Performance Thin-Layer Chromatography," *Journal of Chromatographic Science* , vol. 50, n° %1157, p. 161, 2012.

[51] R. Mannhold, G. I. Poda, C. Ostermann et I. V. Tetko, "Calculation of Molecular Lipophilicity: State-of-the-Art and Comparison of LogP Methods on More Than 96,000 Compounds," *Journal of Pharmaceutical Sciences,* vol. 98, n° %13, pp. 861-893, 2009.

[52] M. A. Bakht, M. F. Alajmi, P. Alam, A. Alam, P. Alam et T. M. Aljarba, "Theoretical and experimental study on lipophilicity and wound healing activity of ginger compounds," *Asian Pacific Journal of Tropical Biomedicine,* vol. 4, n° %14, pp. 329-333, 2014.

[53] J. Kujawski, H. Popielarska, A. Myka, B. Drabińska et M. K. Bernard, "The logP Parameter as a Molecular Descriptor in the Computer-aided Drug Design - an Overview," *Computational Methods In Science And Technology,* vol. 18 , n° %12, pp. 81-88, 2012.

[54] G. Hea, L. Fenga et H. Chena, "International Symposium on Safety Science and Engineering in China," *Proc. Engin,* vol. 43, pp. 204-209, 2012.

[55] K. Roy et e. al, "A Primer on QSAR/QSPR Modeling Chapter 2 Statistical Methods in QSAR/QSPR," *Springer Briefs in Molecular Science,* pp. 37-59, 2015.

[56] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri et V. C. a. R. Todeschini, "Comparison of Different Approaches to Define the Applicability Domain of QSAR Models," *Molecules,* vol. 17, pp. 4791-4810, 2012.

[57] Rakotomalala et Ricco, *Pratique de la Régression Linéaire Multiple Diagnostic et sélection de variables,Version 2.1*.

[58] N. A. Joseph, "Contribution à l'étude de l'activité biologique de composes dérivés du nitrobenzène: étude par diffraction des rayons X – modélisation," 2014.

[59] N. N.-Jeliazkova et J. Jaworska, " An Approach to Determining Applicability Domains for QSAR Group Contribution Models: An Analysis of SRC KOWWIN," *ATLA 33*, p. 461-470, 2005.

[60] acdlabs, *Advanced Chemistry Development/ Chemskecht,* 1994-2010.