DEVELOPMENT OF NEW METAL-THIOSEMICARBAZONE COMPLEXES BASED ON THE QSPR MODELING USING MLR AND ANN METHODS

Đến tòa soạn 05-03-2021

Nguyen Minh Quang, Tran Nguyen Minh An
Faculty of Chemical Engineering, Industrial University of Ho Chi Minh City, Ho Chi Minh City
Pham Van Tat
Institute of Development and Applied Economics, Hoa Sen University, Ho Chi Minh City

TÓM TẮT

PHÁT TRIỂN CỦA CÁC PHỨC CHẤT MỚI GIỮA ION KIM LOẠI VÀ DẪN XUẤT THIOSEMICARBAZONE DỰA TRÊN SỰ MÔ HÌNH QSPR SỬ DỤNG PHƯƠNG PHÁP MLR VÀ ANN

Trong nghiên cứu này, hằng số bền (logβ11) của 31 phức chất mới giữa một số ion kim loại và dẫn xuất thiosemicarbazone được dự đoán từ kết quả của sự mô hình hóa mối quan hệ định lượng cấu trúc-tính chất (QSPR). Những mô hình QSPR này được phát triển từ 76 giá trị logβ11 của các phức chất thực nghiệm bằng cách sử dụng hai phương pháp phổ biến như hồi quy tuyến tính đa biến (QSPRMLR) và mạng nơ ron nhân tạo (QSPRANN). Bộ mô tả của các phức chất được tính toán từ cấu trúc tối ưu, trong đó các cấu trúc này được tối ưu bằng các tính toán hóa lượng tử bán thực nghiệm với phương pháp mới PM7. Mô hình QSPRMLR tốt nhất tìm được bao gồm năm mô tả: diện tích Cosmo, thể tích Cosmo, ko, SHBa và Gmin. Kết quả nhận được các giá trị thống kê phù hợp (R2train = 0,821; Q2LOO = 0,789; RMSE = 0,745; Fstat = 64,3644 và PRESS = 45,92). Hơn nữa, mô hình mạng nơ ron QSPRANN với kiến trúc I(5)-HL(10)-O(1) được tìm thấy với các giá trị thống kê: R2train = 0,9567, Q2validation = 0,9841 và Q2test = 0,9825. Những mô hình QSPR này đã được kiểm tra chặt chẽ bằng các kỹ thuật đánh giá ngoại và kết quả rất gần với giá trị thực nghiệm. Vì vậy, các kết quả từ các mô hình QSPR có thể được sử dụng để thiết kế các phức chất mới nhằm ứng dụng trong hóa học phân tích.

Keywords: *Artificial neural network, Multivariate linear regression, QSPR, Stability constants logβ$_{11}$, Thiosemicarbazone.*

1. INTRODUCTION

Thiosemicarbazides were first introduced in the literature in the early 19th century [1] and thiosemicarbazones were reported as valuable derivatives for ketones and aldehydes in the early years of the 20th century. Until now, thiosemicarbazone derivatives have been synthesized in practice. Furthermore, the diverse structure of thiosemicarbazone, especially the appearance of potential donors led to their easy complexation with many metal ions. This is the reason why this group of agents has resulted in many useful applications [1]. In the field of chemistry, thiosemicarbazone ligands and their complexes have been receiving more interest in the area of analytical chemistry. Recently, the stability constant of the mentioned complexes has been discovered for related applications like analytical chemistry with the UV/VIS spectrophotometric technique [2].

The QSPR modeling is popularly used in many fields as in silico approach for predicting properties of chemical compounds based on the relationships between the structural characteristics and the properties [3]. According to statistics till 2016, the number of published works related to QSPR models was about 11,000 projects [9]. Nowadays, the QSPR method is widely used and is seen as an effective method for finding new compounds [3]. The QSPR models are developed using mathematical methods, normally, there are two popular approaches to establish QSPR models, that is linear regression (Multivariate linear regression, Partial least square regression, Principal component regression) and machine learning method (Support vector Regression, Artificial neural network) [3].

In this work, we developed the QSPR models with the logarithm of stability constants ($\log\beta_{11}$) of the [ML] complexes between thiosemicarbazone ligands with the metal ions (M = Cu^{2+}, Co^{2+}, Mn^{2+}, Cr^{3+}, Cr^{6+}, Fe^{2+}, Fe^{3+}, Zn^{2+}, Cd^{2+}) in aqueous solution. The $\log\beta_{11}$ values were collected from an experimental published database (Table 1). The 2D and 3D-descriptors of metal-complexes are taken from the results of calculation on the structure optimization of complexes using semi-empirical quantum mechanics [4-5]. The two kinds of QSPR models were used by using the multiple linear regression ($QSPR_{MLR}$) and the artificial neural network ($QSPR_{ANN}$). These QSPR models were internally and externally evaluated on two independent datasets. Besides, a new series of thiosemicarbazone ligands and complexes were designed and calculated the stability constant from the results of the developed QSPR models.

2. COMPUTATIONAL METHODS

There are several steps to construct a QSPR model. The process must comply with The Organisation for Economic Cooperation and Development (OECD) principles [6]. All are presented in the following sections.

2.1. Structure of complex and dataset

The thiosemicarbazone ligands can form several kinds of complexes with metal ions. They are known as monodentate, bidentate and tridentate ligands. This work chooses the monodentate ligand type to form the ML complex that reacted between a metal ion (M) and a thiosemicarbazone ligand (L). The structure of the selected complexes is shown in Figure 1.

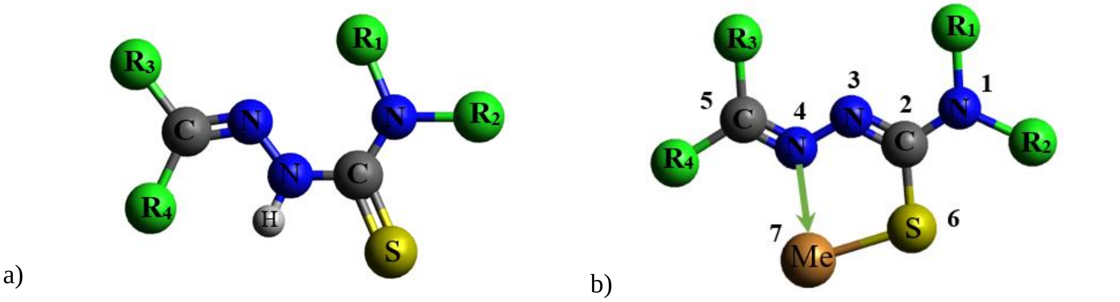


Figure 1. Structure of the thiosemicarbazone ligand (a) and the metal-thiosemicarbazone complex (b)

The properties of complexes is characterized by the stability constant (*β_{11}*) values and the experimental values that used in this research were published in the article. They are target values in this research. So, the data mining is the first step in QSPR modeling research [3]. Firstly, big data was mined from valuable data sources [7-16], then the methods of cluster division as k-means and agglomerative hierarchical clustering (AHC) were used to separate it into several data subsets [3] by using the XLSTAT2016 package [17]. In this study, a dataset including the 76 values log*β_{11}* of metal-thiosemicarbazone complexes was utilized to develop the QSPR models in Table 1.

2.2. Calculation of structural descriptors

The molecular descriptors are recognized as the variables of the descriptive equations in the QSPR modeling. They can be definited as basic numerical characteristics regarding the molecular structures. They have been formed through years of development covering many different theories and until now they are relatively complete [3]. The metal-thiosemicarbazone complexes were sketched the structures by using ChemBioDraw 13.0 [18] and calculated the molecular descriptors from the QSARIS tool [19-20] after they were optimized by using the semi-empirical quantum method with new version PM7 on the MoPac2016 system [6].

2.3. Development of QSPR models

As a mentioned-above matter, the two modeling methods were presented to develop the QSPR regression models in this investigation, which are MLR and ANN methods. Attentively, the artificial neural network models were deeply developed based on the input variables of the $QSPR_{MLR}$ model.

In the $QSPR_{MLR}$ modeling method, the values $\log\beta_{11}$ are confirmed that they are the endpoint values, in this case, they are dependent variables (Y) while the numerical values of structural descriptors (X) are the independent variables in the equation. When the values of X variables correlate well with the values of the Y targets, the equation of the $QSPR_{MLR}$ model is represented as follows: [20-21]

$$Y = b_0 + \sum_{j=1}^{k} b_j X_j \qquad (1)$$

where b_0, is the constant of the model, b_j is the regression coefficients and k is the number of variables in the regression equation.

Table 1. The 76 stability constants of complexes (n) in experimental dataset with minimal ($\log\beta_{11,min}$) and maximal ($\log\beta_{11,max}$) values

No	Thiosemicarbazone ligand				Metal ions	Number of complexes, *n*	$\log\beta_{11,\mathrm{min}}$	$\log\beta_{11,\mathrm{max}}$	Ref.
	R_1	R_2	R_3	R_4					
1	H	H	H	$-C_6H_3BrOH$	Cu^{2+}	1	5.633	5.633	[7]
2	H	H	CH_3	$-CH{=}N{-}NHC_6H_5$	Co^{2+}	4	9.900	10.220	[8-9]
3	H	H	CH_3	$-CH{=}N{-}NHC_6H_5$	Mn^{2+}	3	9.600	9.870	[9]
4	H	H	H	$-C_6H_3(OH)OCH_3$	Cr^{6+}	1	4.842	4.842	[10]
5	H	H	H	$-C_9H_5NOH$	Zn^{2+}	1	6.680	6.680	[11]
6	H	H	H	$-C_6H_3(OH)OCH_3$	Mn^{2+}	3	4.120	5.280	[12]
7	H	H	H	$-C_6H_3(OH)OCH_3$	Pb^{2+}	4	6.530	7.100	[12]
8	H	H	H	$-C_6H_3(OH)OCH_3$	Fe^{2+}	4	7.690	8.150	[12]
9	H	H	H	$-C_6H_3(OH)OCH_3$	Co^{2+}	4	7.860	8.470	[12]
10	H	H	H	$-C_6H_3(OH)OCH_3$	Ni^{2+}	4	8.110	8.650	[12]
11	H	H	$-CH_3$	$-C_6H_4OH$	Cu^{2+}	3	5.810	6.840	[13]
12	H	H	$-CH_3$	$-C_6H_4OH$	Ni^{2+}	2	5.140	5.310	[13]
13	H	H	H	$-C_{10}H_6OH$	Mg^{2+}	2	3.310	3.250	[14]
14	H	H	H	$-C_{10}H_6OH$	Cd^{2+}	4	5.930	6.560	[14]
15	H	H	H	$-C_{10}H_6OH$	Pb^{2+}	1	6.570	6.570	[14]
16	H	H	H	$-C_{10}H_6OH$	Zn^{2+}	1	7.170	7.170	[14]
17	H	H	-	$-C_9H_8NO$	Pb^{2+}	7	7.307	8.109	[15]
18	H	H	-	$-C_9H_8NO$	Zn^{2+}	8	7.039	8.160	[15]
19	H	H	-	$-C_9H_8NO$	Cd^{2+}	7	6.611	7.889	[15]
20	H	H	-	$-C_9H_8NO$	Mn^{2+}	8	5.439	6.041	[15]
21	H	H	H	$-C_6H_4NO_2$	Cr^{3+}	2	10.150	11.250	[16]
22	H	H	H	$-C_6H_4NO_2$	Fe^{3+}	2	11.100	11.630	[16]

In addition, an artificial neural network (ANN) is a non-linear regression method that is known as the deep learning method. Nowadays, this method is used widely in many fields like information technology, drug designs, chemistry and several other fields [22-23]. In this study, we developed the QSPR$_{\text{ANN}}$ models with the multilayer perceptron (MLP) type by using an error back-propagation algorithm [24]. The architecture of the MLP-ANN type has the formation I(*k*)-HL(*m*)-O(*n*). It includes three layers: the input layer (*k*), the hidden layer (*m*) and the output layer (*n*). Therein, the input layer is the variables of the resulted MLR model, the number of hidden neurons is determined by neurons on the input and output layer and the output layer is the stability constant logβ_{11} values. The construction of the ANN model takes place in two stages. Firstly, the *m* values of hidden neurons are prematurely screened by using Neural Designer tools [25], then the best ANN model is excommunicated through external validation on a data set. The second step is run on the Matlab 2016a [26] with Neural Network tool (nntool) toolbox and the process of ANN model training uses two major transfer functions in the neural network research: the hyperbolic sigmoid tangent and log-sigmoid transfer function. Two transfer functions are defined in the following equations: [24-27]

$$a = \tan sig(n) = \frac{2}{\left(1 + e^{-2n}\right)^{-1}} \qquad (2)$$

$$a = \log sig(n) = \frac{1}{1 + e^{-n}} \qquad (3)$$

2.4. Validation of QSPR models

According to OECD principles [6], the QSPR models have to meet the requests of statistics, so it is essential to validate internally and externally on two different datasets and the good models must be received the acceptance criteria of Tropsha's. The indices consist of the values R^2_{train} and Q^2_{LOO} for an internal set or Q^2_{CV} for an external-validation set [20-21]. These standard criteria are calibrated by the same formula (4):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \qquad (4)$$

where Y_i, $\hat{Y}_i$ and $\bar{Y}$ are the experiment, calculated and average values, respectively.

The root means square error (RMSE) determined by the equation 5, is the square root of the mean squared error (MSE). [20-21]. Meanwhile, the ANN models are trained until the mean square error (MSE_{ANN}) is minimized followed by a difference of the output and real values [24-27]. Consequently, MSE$_{\text{ANN}}$ is the average squared error between the network outputs (*o*) and the target outputs (*t*). It is calculated by the equation (6):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\mathrm{Y_i} - \hat{\mathrm{Y}}_\mathrm{i})^2}{N - k - 1}} \qquad (5)$$

$$MSE_{\text{ANN}} = \frac{1}{n}\sum_{1}^{n}\left(t_i - o_i\right)^2 \qquad (6)$$

where N is the number of variables in the training dataset and k is the number of variables in the models.

Furthermore, this work uses the average absolute values of the relative errors MARE (%) to compare the quality of the models. MARE (%) is presented as follows [20]

$$MARE,\% = \frac{\frac{\left|\log\beta_{11,\text{exp}} - \log\beta_{11,\text{cal}}\right|}{\log\beta_{11,\text{exp}}}100}{n} \qquad (7)$$

where *n* is the number of observations; $\beta_{11,exp}$ and $\beta_{11,cal}$ are the experimental and calculated stability constants, respectively,

Finally, $MPx_{k,i}$ (%) quantity is the average contribution percentage [20] which is proposed to find the important variables that have a great influence on the models. It is determined according to formula (8)

$$MPx_{k,i},\% = \frac{1}{N}\sum_{m=1}^{N}\frac{100.\left|b_{k,i}.x_{m,i}\right|}{\sum_{j}^{k}\left|b_{k,j}.x_{m,j}\right|} \qquad (8)$$

where N is the number of observations; m is the number of compounds used to calculate $Px_{k,i}$ value; $b_{k,i}$ are the parameters of the model.

3. RESULTS AND DISCUSSION

3.1. QSPR$_{MLR}$ modeling

The modeling of the multiple linear regression method was operated on MS-EXCEL [28-29] with the Regress system [21] as an add-in program by using the forward and backward elimination regression technique. The method for internally validating the QSPR models in this study is cross-validation (CV) and the process of CV for QSPR models were carried out by the leave-one-out (LOO) method using the statistic Q^2_{LOO}. [20-21].

The full dataset to build the QSPR$_{MLR}$ models including the 76 stability constants values of complexes is separated into the training set and the test set, in which, the test one is randomly selected 20 percent of the original data one. In addition, the criteria of statistical values such as F_{stat} (Fischer's value), RMSE and PRESS are used to evaluate the quality of models [3,6]. The results of QSPR$_{MLR}$ models and the statistical values are indicated in Table 2.

The selection of the best QSPR model is based on the results of Table 2 and Fig 2a. The descriptors of models are chosen according to the changing tendency of the R^2_{train}, Q^2_{LOO}, RMSE and F_{stat} values and the number of variables k that the models reach the goal of the statistical standards. The data from Table 2 presented that when k values increased to 5, the QSPR model met the statistical requirements. Although when the k value is 6, the statistical indexes are better, this variation is negligible and it is not recommended to build the model. So the selection of the model with the k of 5 is the best QSPR model. Besides, the variables from x_1 to x_5 were closely monitored on the basis of the p-value (< 0.05) and t-student characterized the variables [3,6].

The MLR regression model is found as following equation with the statistical values:

$$\log\beta_{11} = -29.585 + 0.310\cdot x_1 - 0.120\cdot x_2 - 0.896\cdot x_3 + 0.249\cdot x_4 - 1.342\cdot x_5$$

$$n = 76;\ R^2_{train} = 0.821;\ Q^2_{LOO} = 0.789;\ RMSE = 0.745;\ F_{stat} = 64.3644 \quad (9)$$

Table 2. Selected models QSPR$_{MLR}$ (k of 1 to 6) and statistical values

k	Variables	*RMSE*	R^2_{train}	Q^2_{LOO}	F_{stat}	*PRESS*
1	x_1	1.576	0.154	0.118	13.4636	191.60
2	x_1/x_2	1.452	0.292	0.248	15.0196	163.24
3	$x_1/x_2/x_3$	1.062	0.626	0.578	40.2067	91.73
4	$x_1/x_2/x_3/x_4$	0.886	0.744	0.708	51.4509	63.41
5	$\boldsymbol{x_1/x_2/x_3/x_4/x_5}$	**0.745**	**0.821**	**0.789**	**64.3644**	**45.92**
6	$x_1/x_2/x_3/x_4/x_5/x_6$	0.709	0.848	0.799	64.5886	43.59

Notation of molecular descriptors

cosmo area	x_1	SHBa	x_4
cosmo volume	x_2	Gmin	x_5
ko	x_3	S6	x_6

As a consequence, the training data set used to develop the MLR is completely qualified; also, the results showed that the predictive ability of the QSPR$_{MLR}$ model is very suitable for this complex group. Therefore, this model can be used to predict new complexes of the same type group based on the Applicability Domain (AD) and Outliers rules [3,6].

On the flip side, $GMPx_i$ values ($GMPx_i$ is the average value of $MPx_{k,i}$) are used to evaluate the affected level of the variables in the models by using three neighboring models. The outcome of data in Table 3 showed that the main contribution of the descriptors in sequential order of cosmo area (x_1) > cosmo volume (x_2) > ko (x_3) with corresponding values of 57.1460, 26.8409 and 11.0105. The cosmo area and cosmo volume parameter are the sums of surface area and volume of molecules that are calculated by COSMO methods [19]. The ko is Kappa zero index or Shannon information index based on atom classes. The Kappa zero index is the information content (IC) index that supplies information as the number of graph vertices, hydride groups, or non-hydrogen atoms [19]. The parameters such as cosmo area, cosmo volume and ko indicate the important role of the structural size of the complexes and the subsistence of the atoms and groups types in the complexes. The two remaining variables (SHBA and Gmin) affect insignificantly the model. These key parameters will be used for the design and search of new ligands and complexes.

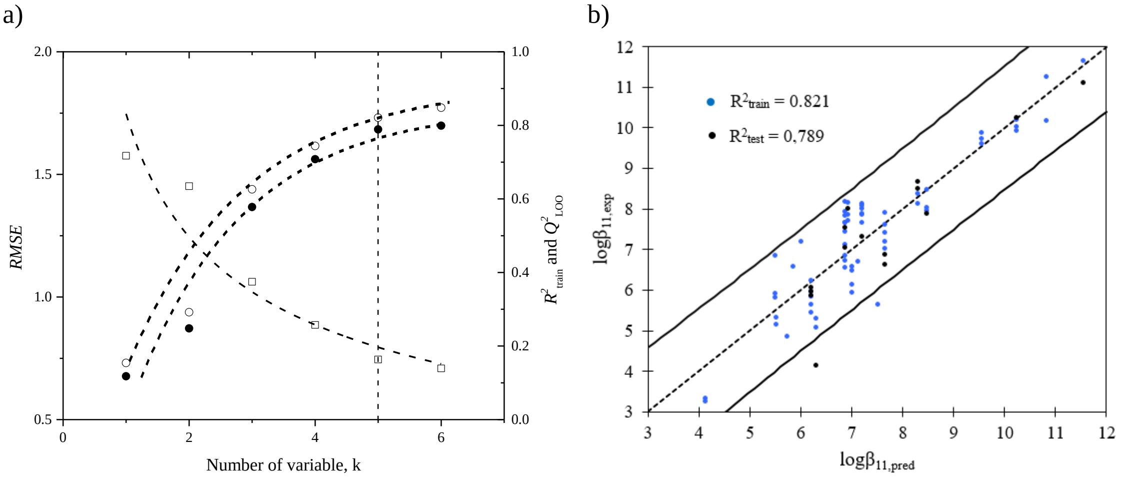


Figure 2. a) Changing tendency of the values RMSE, R^2_{train} and Q^2_{LOO} according to k descriptors; b) Correlation of experimental vs. predicted values $\log\beta_{11}$ of the training compounds using the $QSPR_{MLR}$ model (with k = 5)

Table 3. The full data for calculation of $MPx_{k,i}$ and $GMPx_i$ contribution in models $QSPR_{MLR}$ with k of 4 to 6

Statistical values and variables	$QSPR_{MLR}$			$MPx_{k,i}$, %			$GMPx_i$, %
	$k = 4$	$k = 5$	$k = 6$	$k = 4$	$k = 5$	$k = 6$	
R^2_{train}	0.744	0.821	0.848	–	–	–	–
R^2_{adj}	0.729	0.809	0.827	–	–	–	–
Q^2_{LOO}	0.708	0.789	0.799	–	–	–	–
RMSE	0.886	0.745	0.709	–	–	–	–
constant	-19.4629	-29.5855	-33.1463	–	–	–	–
x_1	0.3066	0.3105	0.3292	55.9339	57.4241	58.0801	57.1460
x_2	-0.1688	-0.1203	-0.1236	33.1196	23.9355	23.4675	26.8409
x_3	-0.5249	-0.8968	-0.9282	7.4385	12.8762	12.7167	11.0105
x_4	0.2172	0.2490	0.2427	3.5081	4.0782	3.7920	3.7928
x_5	–	-1.3426	-1.4183	–	1.6860	1.6993	1.1284
x_6	–	–	-1.0643	–	–	0.2443	0.0814

3.2. $QSPR_{ANN}$ modeling

Firstly, the ANN models are screened to find the architecture of ANN models from the same dataset of the MLR model and the resulting variables of the MLR model. Thereupon, the models are developed upon five-variables

$QSPR_{MLR}$ model and the architecture of the neural network consist of three layers I(5)-HL(*m*)-O(1), in which the input layer I(5) includes five neurons: cosmo area, cosmo volume, ko, SHBa and Gmin; the output layer O(1) includes 1 neuron it is the stability constant logβ_{11} values; the number of the hidden layer (*m*) will be scanned to look for several good models. The results of the *m* neurons are given in Table 5.

In the next step, an external data set is used to train the best ANN model combined with external evaluation for the multivariate linear regression model by the external-validation technique through the Q^2_{CV} index. The results are found the $QSPR_{ANN}$ model with architecture I(5)-HL(10)-O(1) with the best predictability associated with the Q^2_{CV} value of 0.896 as in Figure 3b. Consequently, the hyperbolic tangent function is used for the search of the best network training with the optimum ANN parameters such as the learning rate of 0.01, the momentum constant of 0.05 and the convergent goal of 10^{-7}.

Table 4. The screening of $QSPR_{ANN}$ model I(5)-HL(m)-O(1) with statistical parameters

No	$QSPR_{ANN}$ model	R^2_{train}	Q^2_{test}	$Q^2_{validation}$	Training error	Test Error	Validation Error	Transfer Function
1	I(5)-HL(11)-O(1)	0.9677	0.9753	0.9831	0.0753	0.1214	0.0634	hyperbolic tangent
2	I(5)-HL(8)-O(1)	0.9655	0.9651	0.9823	0.0825	0.1940	0.1418	log-sigmoid
3	I(5)-HL(6)-O(1)	0.9785	0.9768	0.9836	0.0505	0.1303	0.0622	hyperbolic tangent
4	**I(5)-HL(10)-O(1)**	**0.9567**	**0.9823**	**0.9841**	**0.1012**	**0.0820**	**0.0587**	**hyperbolic tangent**
5	I(5)-HL(6)-O(1)	0.9645	0.9795	0.9846	0.0834	0.1000	0.0742	log-sigmoid

In the next step, an external data set is used to train the best ANN model combined with external evaluation for the multivariate linear regression model by the external-validation technique through the Q^2_{CV} index. The results are found the $QSPR_{ANN}$ model with architecture I(5)-HL(10)-O(1) with the best predictability associated with the Q^2_{CV} value of 0.896 as in Figure 3b. Consequently, the hyperbolic tangent function is used for the search of the best network training with the optimum ANN parameters such as the learning rate of 0.01, the momentum constant of 0.05 and the convergent goal of 10^{-7}.

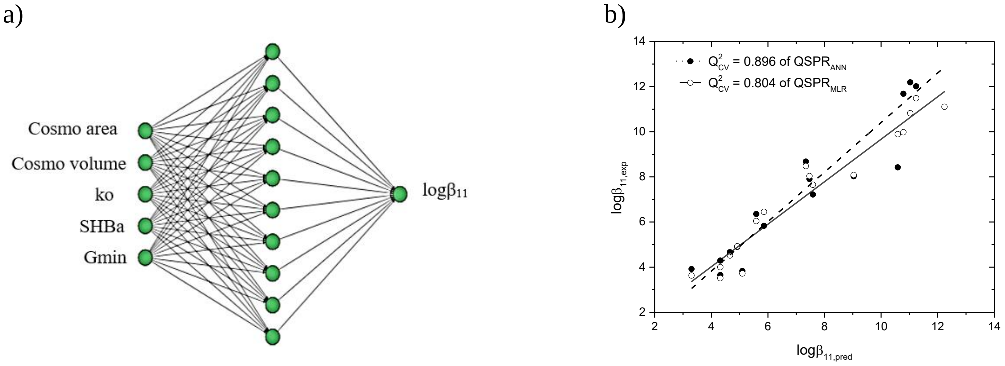


Figure 3. (a) Architecture of neural network I(5)-HL(10)-O(1); (b) The correlation between experimental vs. predicted values of external data set of QSPR models

3.3. The external validation of QSPR models

The external validation is the last step of the model building technique. An external data set including seventeen complexes from the experiment is used for the validation on both $QSPR_{MLR}$ and $QSPR_{ANN}$ models. The results of the evaluation are featured in Table 6.

The calculable data from Table 5 indicated that the *MARE* values of $QSPR_{MLR}$ and $QSPR_{ANN}$ I(5)-HL(10)-O(1) models are 10.478 % and 7.343 %, respectively. As a result, the ANN model with architecture I(5)-HL(10)-O(1) has better predictability than the MLR model and the predicted log$\beta_{11,cal}$ values of the ANN model are close to the experimental log$\beta_{11,\mathrm{exp}}$ values. Based on the results of data analysis in Table 5 and Figure 3b, it can conclude that the prediction of the two models turn out to be in a good agreement with the experimental data [20]. The QSPR$_{\mathrm{ANN}}$ and QSPR$_{\mathrm{MLR}}$ models represent the significant correlation between the predicted values and the experimental values with Q^2_{CV} values of 0.804 and 0.896, respectively [3,6].

Table 5. The experimental log$\beta_{11,exp}$ and external predicted log$\beta_{11,cal}$ values from the QSPR models

Thiosemicarbazone ligand				Metal ions	log$\beta_{11.\mathrm{exp}}$	log$\beta_{11.\mathrm{cal}}$		ref.
R$_1$	R$_2$	R$_3$	R$_4$			QSPR$_{\mathrm{MLR}}$	QSPR$_{\mathrm{ANN}}$	
H	H	H	- C_5H_4N	Mn^{2+}	4.320	3.651	3.512	[30]
H	H	H	- C_6H_4OH	Cu^{2+}	4.920	4.903	4.924	[31]
H	H	H	-$C_{13}H_{16}NO_3$	Fe^{2+}	12.240	14.397	11.107	[32]
H	H	H	-C_4H_3O	Co^{2+}	5.099	3.832	3.718	[33]
H	H	CH_3	-CH=N-NHC_6H_5	Ni^{2+}	10.790	11.684	9.977	[9]
H	CH_3	CH_3	-CH=N-NHC_6H_5	Co^{2+}	10.590	8.422	9.891	[9]
H	CH_3	CH_3	-CH=N-NHC_6H_5	Ni^{2+}	11.030	12.187	10.821	[9]
H	H	H	-$C_{14}H_{12}N$	Cd^{2+}	5.860	5.833	6.450	[34]
H	H	H	-$C_6H_3(OH)OCH_3$	Cd^{2+}	7.340	8.681	8.480	[12]
H	H	H	-$C_6H_3(OH)OCH_3$	Zn^{2+}	7.470	7.903	8.037	[12]
H	H	H	-$C_6H_3(OH)OCH_3$	Cu^{2+}	9.030	8.026	8.083	[12]
H	H	-CH_3	-C_6H_4OH	Mg^{2+}	3.300	3.917	3.628	[13]
H	H	-CH_3	-C_6H_4OH	Mn^{2+}	4.320	4.292	4.001	[13]
H	H	-CH_3	-C_6H_4OH	Cd^{2+}	5.590	6.355	6.043	[13]
H	H	H	-$C_{10}H_6OH$	Mn^{2+}	4.660	4.665	4.517	[14]
H	H	-	-C_9H_8NO	Co^{2+}	7.591	7.218	7.651	[15]
H	H	H	-$C_6H_4NO_2$	Al^{3+}	11.240	12.015	11.482	[16]
				MARE, %		10.478	7.343	

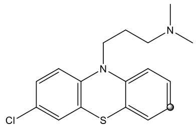
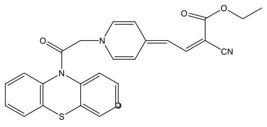
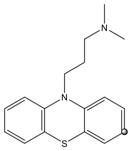
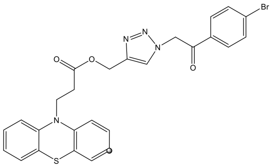
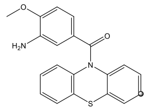
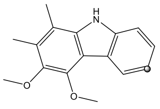
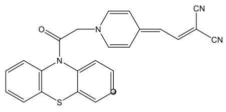
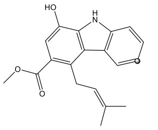
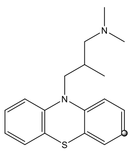
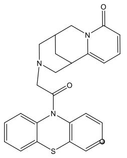
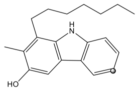
The calculable data from Table 5 indicated that the *MARE* values of QSPR$_{\mathrm{MLR}}$ and QSPR$_{\mathrm{ANN}}$ I(5)-HL(10)-O(1) models are 10.478 % and 7.343 %, respectively. As a result, the ANN model with architecture I(5)-HL(10)-O(1) has better predictability than the MLR model and the predicted log$\beta_{11,\mathrm{cal}}$ values of the ANN model are close to the experimental log$\beta_{11,\mathrm{exp}}$ values. Based on the results of data analysis in Table 5 and Figure 3b, it can conclude that the prediction of the two models turn out to be in a good agreement with the experimental data [20]. The QSPR$_{\mathrm{ANN}}$ and QSPR$_{\mathrm{MLR}}$ models represent the significant correlation between the predicted values and the experimental values with Q^2_{CV} values of 0.804 and 0.896, respectively [3,6].

Furthermore, the one–way ANOVA method is used to evaluate the difference between the experimental and predictive values of both models, accordingly, the differences between the QSPR models are insignificant ($F = 0.0711 < F_{0.05} = 3.1788$).

3.4. Development of new complexes

We selected two kinds of derivatives, namely the phenothiazine and carbazole to design new thiosemicarbazone and the complexes between the new ligands with some metal ions as Ag^+, Ni^{2+}, Cd^{2+}, Cu^{2+} and Zn^{2+}. The derivatives were synthesized and published in the literature [35-38]. The selection is based on the five descriptors: Cosmo area, Cosmo volume, ko, SHBa and Gmin form the results of the QSPR models. The new thiosemicarbazones are designed by means of adding phenothiazine and carbazole groups at the R_4 site on the structure of the thiosemicarbazones and the other positions as R_1, R_2 and R_3 are hydrogen atoms. A series of new complexes are sketched and calculated the structural descriptors. Then they were carefully screened and put into the spatial data of the training set to validate by AD and Outlier [3,6]. As a matter of course, the thirty-one new complexes meet the standards of AD and they are predicted the stability constant from the two built QSP$_{\text{RMLR}}$ and QSPR$_{\text{ANN}}$ models. The prediction of new complexes ($\log\beta_{11,\text{new}}$) is given in Table 6. Using the single-factor ANOVA to compare the predicted $\log\beta_{12,\text{pred}}$ values resulted from the QSPR$_{\text{MLR}}$ and QSPR$_{\text{ANN}}$ models, it indicated that the difference is insignificant between the two models ($F = 0.1335 < F_{0.05} = 4.0012$).

Table 6. The thirty-one new predicted $\log\beta_{11,new}$ values of metal-thiosemicarbazone complexes from the developed QSPR models

R_4 site	metal ions	$\log\beta_{11,\text{pred}}$		R_4 site	metal ions	$\log\beta_{11,\text{new}}$	
		MLR	ANN			MLR	ANN
	Ag^+	7.4741	7.4863		Cd^{2+}	8.7000	9.2852
	Ni^{2+}	10.4742	9.2753		Ni^{2+}	9.8782	9.2888
	Cu^{2+}	10.1601	9.2732		Zn^{2+}	8.5015	9.2764
	Ag^+	9.9575	9.1703		Ag^+	9.8744	9.2879
	Cd^{2+}	9.2782	9.0967		Cd^{2+}	8.9198	9.2876
	Cu^{2+}	7.9016	8.7918		Cu^{2+}	8.6660	9.2874
	Ni^{2+}	8.5959	8.9287		Ni^{2+}	8.6097	9.2875
	Zn^{2+}	7.6084	8.7115		Zn^{2+}	8.9877	9.2874
	Cu^{2+}	10.9610	9.3018		Cd^{2+}	10.2095	9.1353
	Zn^{2+}	10.7333	9.3023		Zn^{2+}	9.4431	8.6861
	Ag^+	10.6217	9.2910		Cd^{2+}	10.9209	9.3612
	Cu^{2+}	8.8434	9.2886		Cu^{2+}	8.3245	9.3022
	Ni^{2+}	10.1772	9.2907		Ni^{2+}	8.7107	9.3080
	Zn^{2+}	8.7257	9.2883		Zn^{2+}	9.7639	9.3655
	Cd^{2+}	7.8791	9.2452		Cd^{2+}	9.3364	9.2880
	Cd^{2+}	7.7389	9.3041				

4. CONCLUSION

In this investigation, two popular methods such as the multivariate linear regression and artificial neural network were used to build successfully the quantitative structure-property relationship (QSPR) models and the QSPR models were developed by using the dataset of structural descriptors and the stability constant values of metal-thiosemicarbazone complexes. The study was a combination of semi-empirical quantum mechanics calculations with new version PM7 and statistics techniques. Moreover, the *in silico* method was studied on big data through design, screening, and mining data techniques. The QSPR models were fully built based on OECD principles and the model acceptance criteria of Golbraikh and Tropsha's as R^2_{train}, Q^2_{LOO}, *MARE*, %, and ANOVA. The results from the new models allowed us to develop thirty-one new complexes with the predicted stability constant values. As a result, the built QSPR models can be useful to explore new complexes.

REFERENCES

1. Casas, J. S.; García-Tasende, M. S.; Sordo, J. *Main group metal complexes of semicarbazones and thiosemicarbazones. A structural review*. Coordination Chemistry Reviews, **209**(1), 197–261 (2000).

2. Singh R. B.; Garg B. S.; Singh. R. P. *Analytical applications of thiosemicarbazones and semicarbazones: A review.* Talanta, **25**(11-12), 619–632 (1978).

3. Kunal, R.; Supratik, K.; Rudra, N. D. *A Primer on QSAR/QSPR Modeling, Fundamental Concepts*. New York: Springer. (2015).

4. Stewart, J. J. P. *MOPAC2016 Version: 17.240W*. Stewart Computational Chemistry, USA (2002).

5. Stewart, J. J. P. *Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters*. J. Mol. Model. **19**, 1-32 (2013).

6. Organisation for Economic Co-operation and Development . *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships Models.* France (2007).

7. Ramanjaneyulu, G.; Reddy, P. R.; Reddy, V. K.; Reddy, T. S. *Direct and Derivative Spectrophotometric Determination of Copper(II) with 5-Bromosalicylaldehyde Thiosemicarbazone*. The Open Anal. Chem. J. **2**, 78-82 (2008).

8. Aljahdali, M.; EL-Sherif, A. A. *Synthesis, characterization, molecular modeling and biological activity of mixed ligand complexes of Cu(II), Ni(II) and Co(II) based on 1,10-phenanthroline and novel thiosemicarbazone*. Inorg. Chimica Acta. **407**, 58-68 (2013)

9. El-Karim, A. T. A.; El-Sherif, A. A. *Potentiometric, equilibrium studies and thermodynamics of novel thiosemicarbazones and their bivalent transition metal(II) complexes*. J. Mol. Liq. **219**, 914-922 (2016).

10. Sreevani, I.; Reddy, P. R.; Reddy, V. K. *A Rapid and Simple Spectrophotometric Determination of Traces of Chromium (VI) in Waste Water Samples and in Soil samples by using 2-Hydroxy, 3-Methoxy Benzaldehyde Thiosemicarbazone (HMBATSC)*. IOSR J. Appl. Phys. **3**(1), 40-45 (2013).

11. Rogolino, D.; Cavazzoni, A.; Gatti, A.; Tegoni, M.; Pelosi; Verdolino, V.; Fumarola, C.; Cretella, D; Petronini, P. G.; Carcelli, M. *Anti-proliferative effects of copper(II) complexes with Hydroxyquinoline-Thiosemicarbazone ligands*. Eu. J. Med. Chem. **128**, 140-153 (2017).

12. Garg, B. S.; Jain, V. K. *Determination of thermodynamic parameters and stability constants of complexes of biologically active o-vanillinthiosemicarbazone with bivalent metal ions*. Thermochimica Acta. **146**, 375-379 (1989).

13. Garg, B. S.; Ghosh, S.; Jain, V. K.; Singh, P. K. *Evaluation of thermodynamic parameters of bivalent metal complexes of 2-hydroxyacetophenonethiosemicarbazone (2-HATS).* Thermochimica Acta. **157**, 365-368 (1990).

14. Sahadev, S; Sharma, R. K.; Sindhwani, S. K. *Thermal studies on the chelation behaviour of biologically active 2-hydroxy-1-naphthaldehyde thiosemicarbazone (HNATS) towards bivalent metal ions: a potentiometric study*. Thermochimica Acta. **202**, 291-299 (1992).

15. Sarkar, K.; Garg, B. S. *Determination of thermodynamic parameters and stability constants of the complexes of p-MITSC with*

transition metal ions. Thermochimica Acta. **113**, 7-14 (1987).

16. Sawhney, S. S.; Sati, R. M. *pH-metric studies on Cd(II)-, Pb(II)-, AI(III)-, Cr(III)-AND Fe(III)-p-nitrobenzaldehyde thiosemicarbazone systems*. Thermochimica Acta. **66**, 351-355 (1983)

17. Addinsoft. *XLSTAT2016 Version 2016.02.28451*. USA (2016).

18. PerkimElmer. *ChemBioDraw Ultra 13.0.0.3015*. CambridgeSoft, England (2012).

19. Statistical Solutions Ltd. *QSARIS 1.1*. USA (2001).

20. Tat, P. V. *Development of QSAR and QSPR*. Ha Noi: Publisher of Natural sciences and Technique (2009).

21. Steppan, D. D.; Werner, J.; Yeater, P. R. *Essential Regression and Experimental Design for Chemists and Engineers*. Free Software Package (1998). http://home.t-online.de/home/jowerner98/indexeng.html.

22. Harish, K. G.; Radha, K. P. *Application of ANN technique to predict the performance of solar collector systems - A review*, Renewable and Sustainable Energy Reviews. **84**, 75-88 (2018).

23. Abdul, A.; Farrokh, J. S.; Alan, S. F.; Kaamran, R. *Artificial neural network (ANN) based model predictive control (MPC) and optimization of HVAC systems: A state of the art review and case study of a residential HVAC system*. Energy and Buildings. **141**, 96-113 (2017).

24. Rojas, R. *Neural Networks*. Springer-Verlag, Berlin (1996).

25. Artelnics. *Neural Designer software*, USA: Artificial Intelligence Techniques, Ltd (2020).

26. MathWorks. *Matlab R2016a 9.0.0.341360*. USA(2016).

27. Gasteiger, J.; Zupan, J. *Neural Networks in Chemistry*. Chiw. Inr. Ed. EngI. **32**, 5-15 (1993).

28. Microsoft. *MS. Excel 2013*, USA (2013).

29. Billo, E. J. *Excel For Scientists and Engineers: Numerical Methods*. USA: John Wiley and Sons, Inc. 03-521 (2007).

30. Kenie, D. N.; Satyanarayana, A. *Protolitic Equilibria and Stability Constants of Mn (II) and Ni (II) Complexes of 3-formylpyridine Thiosemicarbazone in Sodium Dodecyl Sulphate (SDS)- Water Mixture*. Sci. Technol. Arts Res. J. **4**(1), 74-79 (2015).

31. Biswas, R.; Brahman, D.; Sinha, B. *Thermodynamics of the complexation between salicylaldehyde thiosemicarbazone with Cu(II) ions in methanol–1,4-dioxane binary solutions*. J. Serb. Chem. Soc. **79**(5), 565-578 (2014).

32. Milunovic, M. N. M.; Enyedy, E. A.; Nagy, N. V.; Kiss, T.; Trondl, R.; Jakupec, M. A.; Keppler, B. K.; Krachler, R.; Novitchi, G.; Arion, V. B. *L- and D Proline Thiosemicarbazone Conjugates: Coordination Behavior in Solution and the Effect of Copper(II) Coordination on Their Antiproliferative Activity*. Inorg. Chem. **51**, 9309-9321 (2012).

33. Veeranna, V.; Rao, V. S.; Laxmi, V. V.; Varalakshmi, T. R. *Simultaneous Second Order Derivative Spectrophotometric Determination of Cadmium and Cobalt using Furfuraldehyde Thiosemicarbazone (FFTSC)*. Res. J. Pharm. and Tech. **6**(5), 577-584 (2013).

34. Koduru, J. R.; Lee, K. D. *Evaluation of thiosemicarbazone derivative as chelating agent for the simultaneous removal and trace determination of Cd(II) and Pb(II) in food and water samples*. Food Chemistry. **150**, 1-8 (2014).

35. Al-Busaidi, I. J.; Haque, A.; Al Rasbi, N. K.; Khan, M. S. *Phenothiazine-based derivatives for optoelectronic applications: A review*. Synthetic Metals. **257**, 116-139 (2019).

36. Sudeshna, G.; Parimal, K. *Multiple non-psychiatric effects of phenothiazines: A review*. European Journal of Pharmacology. **648**(1-3), 6-14 (2010).

37. Huang, L.; Feng, Z. L.; Wang, Y. T.; Lin, L. G. *Anticancer carbazole Alkaloids and coumarins from Clausena plants: A review*. Chinese Journal of Natural Medicines. **15**(12), 881-888 (2017).

38. Krucaite, G.; Grigalevicius, S. *A review on low-molar-mass carbazole- based derivatives for organic light emitting diodes*. Synthetic Metals **247**, 90-108 (2019).