pISSN 1859-1388
eISSN 2615-9678

QSPR-based calculation model for stability constants of new metal-thiosemicarbazone complexes with hybrid techniques

Nguyen Hoang Minh, Nguyen Minh Quang*

Faculty of Chemical Engineering, Industrial University of Ho Chi Minh city, Vietnam

* Correspondence to Nguyen Minh Quang <nguyenminhquang@iuh.edu.vn>

(*Received: 12 September 2023; Revised: 05 January 2024; Accepted: 09 December 2024*)

Abstract. In the present study, we calculated the logβ12 stability constant of twenty new ML_2 complexes between thiosemicarbazone and metal ions based on the modelling techniques of the quantitative structure and property relationship (QSPR). The QSPR models were developed by combining the genetic algorithm (GA) with multivariate linear regression techniques ($QSPR_{GA-MLR}$), support vector regression ($QSPR_{GA-SVR}$), and artificial neural network ($QSPR_{GA-ANN}$). The descriptive parameters were calculated from semi-empirical quantum computation with the new version PM7 and PM7/sparkle. The resulting $QSPR_{GA-MLR}$ models had three variables, and the $QSPR_{GA-SVR}$ and $QSPR_{GA-ANN}$ models were developed from the variables of the $QSPR_{GA-MLR}$ model. The results show that the best $QSPR_{GA-SVR}$ model had the following optimal parameters: $C = 10.0$; $\gamma = 0.333$; $\varepsilon = 0.10$ with 51 support vectors, and a $QSPR_{GA-ANN}$ model with the network architecture I(3)-HL(10)-O(1) was successfully developed. Furthermore, the quality of QSPR models conformed to statistical values according to OECD principles and Tropsha's criteria.

Keywords: ANN, stability constants logβ12, MLR, QSPR, SVR, thiosemicarbazone

1 Introduction

In recent years, the field of theoretical study has attracted interests in a technique for discovering new chemical compounds. Because of the integration of new mathematical knowledge, cutting-edge computer software, and anecdotes, the establishment of this innovation is based on quantum computing, matched with factual instruments and machine learning strategies. This novel approach is applied in various important industries, including the discovery of new compounds based on characteristics or the development of medications based on activity and toxicity [1]. These methods frequently have several expressions since they are based on the quantitative relationship between structure and properties or activities. Actually, the method is widely used in published papers and applications [2, 3].

Meanwhile, thiosemicarbazones are derivatives that represent a critical group of Schiff compounds carrying sulfur and nitrogen conjugates [4]. This structural characteristic suggests the structural diversity and peculiarity of these derivatives, which leads to their diverse applications. Indeed, in the mid-20th century, thiosemicarbazone derivatives were synthetisised and demonstrated their important application capabilities as drugs against dangerous diseases, such as tuberculosis and herpes [5]. At this stage, the first cancer-preventive activity of thiosemicarbazones was also discovered. At the same time, an extended antibacterial, antifungal, anti-malaria, cancer cell movement, anti-inflammatory and antiviral exercises have also been examined and effectively applied [5].

DOI: 10.26459/hueunijns.v133i1D.7306

Furthermore, the presence of authoritative clusters, such as sulfur and nitrogen donors inside the structure, suggests that thiosemicarbazone derivatives also have complex configurations. These complexes are comparable with the thiosemicarbazone derivatives. This has been illustrated through distributed exploratory work and brings vast applications in explanatory chemistry, which utilizes ligands as potential tests using cheap and easy-to-use UV-VIS procedures to analyze overwhelming metal particles that are destructive to people [6]. In addition, the complex arrangement of ions and atoms within the water environment depends on numerous components and is judged by the complex of solidity and steady esteem. Therefore, chemists have always strived to improve underutilized thiosemicarbazone subordinates with prevailing preferences and multiple applications through stable values.

This work approaches the techniques of quantitative structure-property relationship modeling (QSPR) by using multi-variable linear regression (MLR), supper vector regression (SVR), and artificial neural networks (ANN) in combination with genetic algorithm (GA) to develop corresponding regression models such as $QSPR_{GA\text{-}MLR}$, $QSPR_{GA\text{-}SVR}$ and $QSPR_{GA\text{-}ANN}$. In addition, a semi-experimental quantum calculation method with the new version PM7 is used to optimize the structure of complexes in the study [7]. The molecular depiction parameters and quantum parameters were calculated from the complexes after their structure optimization. The development of these models complies with the rules of the OECD [8] and the pointers of Tropsha [9]. Along these lines, we planned an arrangement of modern ligands and complexes by consolidating synthesized auxiliary outlines joined to the structure of thiosemicarbazone. At this point, the unique complexes were rigorously screened according to the demonstration's criteria, and the consistency was determined by using the model findings. Unused substances within the applicability domain (AD) were subjected to stability constant calculation, whereas the remaining substances outside the application field (outliers) were expelled [1, 8].

2 Computational methods

2.1 Data mining technique

The development of QSPR models involves a series of steps. Among them, data mining is the primary step considered [8].

This study deals with the complexes between thiosemicarbazone (L) and metal ions (M). A typical structure of the ligand and complex is depicted in Figure 1. Generally, a complex is formed as follows:

$$p\text{M} + q\text{L} \rightleftharpoons \text{M}_p\text{L}_q \qquad (1)$$

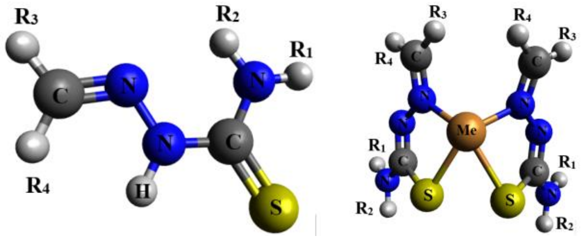
here, the stability constant is utilized as an output value for QSPR models and calculated based on reaction (1) according to equation (2).

$$\beta_{pq} = \frac{\left[M_pL_q\right]}{\left[M\right]^p\left[L\right]^q} \qquad (2)$$

We utilize thiosemicarbazone as a bidentate ligand within the arrangement of an ML_2 complex with a metal ion M (p = 1) joining two ligands (q = 2). Therefore, the stability constant takes the following form.

$$\beta_{12} = \frac{\left[ML_2\right]}{\left[M\right]\left[L\right]^2} \qquad (3)$$

pISSN 1859-1388
eISSN 2615-9678



(a) (b)

Fig. 1. General structure of thiosemicarbazone (a) and its complex (b)

Data mining was carried out in the sequence of secondary collection techniques. In the primary step, a huge dataset comprising stability constant values of ML_2-form test complexes between thiosemicarbazone and metal ions was collected from the works published in prestigious journals. At this point, we utilized the k-means data clustering method to isolate the clusters with little information. This led to the spatial information objects called Voronoi cells. This calculation was presented by McQueen in 1967 [9], taken after other comparable contentions created by Forgey in 1965 [11] and Friedman in 1967 [12]. This process resulted in a data set of 86 stability constant values of the complexes, utilized as a preparation set for building the QSPR models. The detailed information of the data set is depicted in Table 1.

Table 1. Eighty six stability constant values of experimental complexes with maximum values ($\log\beta_{12,max}$) and minimum values ($\log\beta_{12,min}$) in the study

No.	Thiosemicarbazone				Metal ion	Numbers of complexes, *n*	$\log\beta_{12,min}$	$\log\beta_{12,max}$	Ref.
	R_1	R_2	R_3	R_4					
1	H	H	$-C_6H_5$	$-CCH_3{=}\ N{-}OH$	Cu^{2+}	2	7.9164	7.9165	[13,14]
2	H	H	$-C_6H_5$	$-CCH_3{=}\ N{-}OH$	Ni^{2+}	1	9.7118	9.7118	[13]
3	H	H	H	$-C_9H_8N$	Cu^{2+}	16	8.5773	8.6946	[15]
4	H	H	H	$-C_9H_8N$	Ag^{+}	16	10.0713	10.7835	[15]
5	H	H	$-CH_3$	$-CH{=}N{-}NHC_6H_5$	Cu^{2+}	3	22.0200	22.6200	[16]
6	H	H	$-CH_3$	$-CH{=}N{-}NHC_6H_5$	Ni^{2+}	3	20.6300	21.2000	[16]
7	H	H	$-CH_3$	$-CH{=}N{-}NHC_6H_5$	Co^{2+}	3	19.4000	19.9500	[16]
8	H	H	$-CH_3$	$-CH{=}N{-}NHC_6H_5$	Mn^{2+}	3	18.6500	19.1800	[16]
9	H	$-CH_3$	$-CH_3$	$-CH{=}N{-}NHC_6H_5$	Cu^{2+}	3	22.8200	23.4400	[16]
10	H	$-CH_3$	$-CH_3$	$-CH{=}N{-}NHC_6H_5$	Ni^{2+}	3	20.8200	21.4000	[16]
11	H	$-CH_3$	$-CH_3$	$-CH{=}N{-}NHC_6H_5$	Co^{2+}	3	19.6500	20.2100	[16]
12	H	$-CH_3$	$-CH_3$	$-CH{=}N{-}NHC_6H_5$	Mn^{2+}	3	18.9000	19.4500	[16]
13	H	H	$-CH_3$	$-C_{10}H_{12}NO$	Hg^{2+}	16	10.3025	10.7401	[17]
14	H	H	H	$-C_6H_3(OH)OCH_3$	Cu^{2+}	4	17.0500	18.0500	[18]
15	H	H	-	$-C_9H_7NO$	Cu^{2+}	7	14.8530	16.6390	[19]

2.2 Descriptors

A QSPR model relates molecule properties to descriptors in a mathematical equation [1]. These molecular descriptors correlate with properties through empirical values, and these correlations are developed through chemometrics to build meaningful QSPR models. Molecular descriptors describe the characteristics of molecules, providing quantitative parameters for chemical properties [20]. The QSPR model explains the conditions for molecular structure and chemical response, as shown in equation (4).

$$\log\beta_{pq} = f(\text{descriptors}) \quad (4)$$

The descriptors used for modeling in this work include both quantum and molecular descriptors (0D–3D). While the variables of the models are characteristics that describe the structure and properties of molecules in the model equation and are generated via the QSARIS tool, quantum descriptors are calculated from optimized complexes [21]. As a result, each complex will have an experimental stability constant value ($\log\beta_{12,exp}$) corresponding to the molecular descriptions, and the QSPR models are created using the data set made up of stability constants and descriptive parameters.

2.3 Multivariate regression model development

Genetic algorithm

As presented, this work applies the genetic algorithm to detect the best variables for the model. This technique is based on searching the groups of the best or most appropriate descriptive parameters by optimizing the correlation factor and using the Friedman Fitness optimization function [22]. Genetic algorithm is a technique developed and applied by software science researchers in a variety of fields to find appropriate solutions for combined optimization techniques. This logic derives from the evolutionary logic that uses the principles of evolution, such as genetics, mutation, natural selection, and cross-exchange in biology. In the development of the models, GA uses random mutation and genetic reconstruction, also known as a cross-hybrid process.

MLR model

The multivariate linear regression (MLR) method determines the degree of correlation between dependent variables (predicted values) and one or more independent variables. The values of descriptive parameters are independent variables in this study, whereas the stability constant values ($\log\beta_{12}$) of the complex between the studied ligand and the metal ions are dependent variables. The following equation serves as the model for this approach: [23]

$$Y = \beta_0 + \sum_i^k \beta_i . X_i + \varepsilon \quad (5)$$

here, Y is the dependent variable; β_0 and β_i are the regression parameters of the equation; X_i is the i^{th} variables with i being from 1 to k), and ε is the random error.

SVR model

Support vector regression (SVR) is a machine learning method (ML) and is commonly used in statistical mathematics and computer science to develop predictive models with or without monitoring data analytic algorithms, such as data classification, regression analysis, and prediction. Support vector (SV) algorithms are nonlinear calculations first discovered in 1963 in Russia by Vapnik and Lerner. The SV algorithm is a solid foundation in the theory of statistical learning that has been continuously developed over the years by Vapnik and Chervonenkis (1974) and Vapník (1995) [24]. This theory describes machine learning properties and allows to generalize data that are not visible [24]. In 1992, Vapnik and colleagues [25] suggested using the kernel tricks methods to create regression models. The analogy allows undirectional vector multiplication to be replaced by a non-linear kernel function ($K(\vec{x}_i,\vec{x}_j)=\phi(\vec{x}_i).\phi(\vec{x}_j)$) that allows the algorithmic analogy to match the maximization of the super plane margin in the conversion space, and the result shows that the analogy works well.

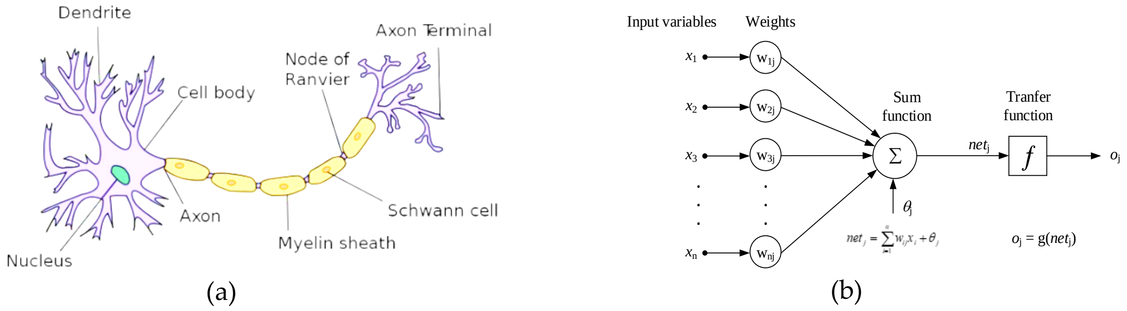
The SVR's effectiveness depends on the selection of the kernel function, the kernel parameters (γ, ε, and *C*), and the number of support vectors. Each combination of the parameter selection is typically tested via cross-assessment, and the kernel parameters with the best accuracy are selected [24]. The modeling process across the entire dataset optimizes the parameters as stated with a corresponding kernel function. There are many kernel functions used in regression model training. In searching for optimum parameters, we used the Gaussian basis function, or radial basis function (RBF) as follows [24]:

$$K(\vec{x}_i,\vec{x}_j)=\exp\{-\gamma\|\vec{x}_i.\vec{x}_j+r\|^2\},\ \gamma>0 \qquad (6)$$

here, γ is the nuclear parameter and *r* is the constant.

ANN model

A technique for replicating the action of the biological nervous system that creates animal brains is called an artificial neural network (ANN). ANN comprises a set of interconnected nodes or units called artificial neurons and establishes a flexible model between animal brain neurons. Biological neurological activity and simulation of ANN information processing is presented in Figure 2 [26]:



(a)

(b)

Fig. 2. (a) Structure and transmission of biological neurons; (b) Information process of a biologically simulated neuron

Multi-layer perceptron (MLP) networks and back-propagation algorithms [27] are commonly used to prepare nonlinear ANN models. This interpretation was initiated by Rumelhart and colleagues in 1986 [27]. An input layer (I), an output layer (O), and one or more hidden layers (HL) make up this kind of MLP network. The back-propagation algorithm in this case operates in three stages: forward propagation, backward propagation, and updating the appropriate weights and deviations. When the target function's value drops low enough, the interpretation ends.

During model practice, a transfer function (*f*) connects the total function (*net*$_j$) to the output (o_j). The transfer function transmits the output to the network from the total function's result along with a required ANN output threshold value (θ_j). In this work, two transfer functions, log-sigmoid and hyperbolic sigmoid, are employed as follows [26]:

$$f(z)=\log sig(z)=\frac{1}{1+e^{-z}} \qquad (7)$$

$$f(z)=\tan sig(z)=\frac{1-e^{-z}}{1+e^{-z}} \qquad (8)$$

2.4 Model validation

Evaluating the model is an essential step in determining whether the QSPR model is predictable. Model evaluations typically need to represent both internal and external evaluations on two independent datasets. The training data set serves as the basis for the internal evaluation process, while the other data set is used for the external evaluation. In this work, the internal evaluation is conducted by using R^2_{train} statistics in conjunction with cross-validation (CV) and the statistical $Q^2_{\text{CV-LOO}}$ index and is based on the original data set of 86 experimental values (Table 1). The experimental stability constant has 18 values in the external validation set (EV) with the evaluation indicator Q^2_{EV}. Formula (9) is used to calculate all these values in the data sets [23]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \qquad (9)$$

here, Y_i, $\hat{Y}_i$, and $\bar{Y}$ are the experimental, predictive, and average stability constants, respectively.

We also used the adjustment coefficient R^2_{adj} to adjust R^2_{train} when putting the variables into equation (5). The index is calculated according to equation (10) [23]:

$$R^2_{adj} = R^2 - \frac{k-1}{N-1}\left(1 - R^2\right) \qquad (10)$$

The mean square error (MSE) is the deviation of the residue, and this quantity plays an important role in determining the data. It contributes to evaluate the predictability of the model and means that it evaluates the difference between the experimental and calculated stability constants from the model. Meanwhile, the root mean square error (RMSE) is the square root of the mean squared error, and this quantity is used in this study. The RMSE is calculated according to equation (11):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\mathrm{Y_i} - \hat{\mathrm{Y}}_\mathrm{i})^2}{N-k-1}} \qquad (11)$$

here, N is the number of variables in the training, and k is the number of variables in the model.

Similarly, the SVR and ANN models use statistical parameters, such as R^2_{train}, Q^2_{test}, and Q^2_{CV}, to evaluate the model [1, 25, 26]. During the model preparation, SVR identifies the optimal kernel parameters. MSE_{ANN} is trained concurrently by altering the nodes (m) in the hidden layer HL(m) of the MLP network with the architecture of I(n)-HL(m)-O(k) until the ultra-minimum value between the values of the output (o) and the real value (t) is reached [28]. Thus, the following description is given for this quantity [26]:

$$MSE_{\text{ANN}} = \frac{1}{n}\sum_{1}^{n}\left(t_i - o_i\right)^2 \qquad (12)$$

Additionally, this study compares the predictability of the models (MLR, SVR, and ANN) by using the mean absolute value of the relative error MARE(%). In principle, the lower this value is, the better is the model of the predictive quality. This means that the predictive values are closer to the experimental values. The MARE(%) value is calculated according to equation (13) [29]:

$$MARE,\% = \frac{1}{n}\frac{\left|\log\beta_{12,\text{exp}} - \log\beta_{12,\text{pred}}\right|}{\log\beta_{12,\text{exp}}}.100 \qquad (13)$$

here, n is the sample size, $\beta_{12,\text{exp}}$ and $\beta_{12,\text{pred}}$ are the corresponding experimental and predictive stability constants.

3 Results and discussion

3.1 Constructing QSPR$_{\text{GA-MLR}}$ models

The stepwise regression approach was used to train the QSPR$_{\text{GA-MLR}}$ model, and the genetic algorithm, which is managed by the R^2_{adj} index on the QSARIS system [21], is used to add one-on-one variables into the model equations. The initial experimental data set (Table 1) was randomly split during model training into a training set accounting for around 80% and a CV set accounting for 20% of the total data. Because of the short data set utilizing the $Q^2_{\text{CV-LOO}}$ statistical measure, the CV of the model in the operational training phase is based on the leave-one-out (LOO) elimination strategy. Additionally, statistical aggregations such as R^2_{train}, RMSE, PRESS, and F_{stat} (Fischer's value) were used to evaluate the conformance of the models [1, 8, 9, 23]. Twelve QSPR$_{\text{GA-MLR}}$ models produced findings that complied with the criteria listed in Table 2. According to the findings based on the information in Table 2, all twelve models received three descriptive variables and met the statistical requirements. However, the use of external assessment methods is required in order to choose a full model. Eighteen experimental compounds were employed in the investigation, and their stability constant values (log$\beta_{12,\text{exp}}$) are shown in Table 4. The best-selected QSPR$_{\text{GA-MLR}}$ model will have the $Q^2_{\text{EV- MLR}}$ values that meet the statistical requirements (>0.6) [9], the larger the better, and the MARE(%) value is the smaller the better [29].

Two QSPR$_{\text{GA-MLR}}$ models, QSPR$_{\text{GA-MLR3}}$ and QSPR$_{\text{GA-MLR12}}$, may be shown as satisfying the statistical values of $Q^2_{\text{EV-MLR}}$ based on the experimental data and the predictive findings of the stability constant (log$\beta_{12,\text{pred}}$) in Figure 4. However, the QSPR$_{\text{GA-MLR12}}$ model receives the $Q^2_{\text{EV- MLR}}$ value of 0.9227, and the QSPR$_{\text{GA–MLR3}}$ model receives the $Q^2_{\text{EV–MLR3}}$ value of 0.7967, which means that the $Q^2_{\text{EV-MLR}}$ value of the QSPR$_{\text{GA-MLR 12}}$ model is significantly larger than that of QSPR$_{\text{GA- MLR3}}$. By contrast, the MARE(%) value of these two models had nearly equal values, 10.5596 and 11.6106, respectively. Thus, it is possible to assert that the QSPR$_{\text{GA-MLR12}}$ model predicts better than the QSPR$_{\text{GA–MLR3}}$ model. As a consequence, the QSPR$_{\text{GA-MLR12}}$ model was chosen to look for ML models of SVR and ANN models as well as develop new complexes.

Table 2. Statistical values for twelve QSPR$_{\text{GA-MLR}}$ models (k = 3)

Notation	The models of QSPR$_{\text{GA-MLR}}$
GA-MLR1	logβ_{12} = 40.233 + 0.005×MW – 12.485×$xc3$ + 0.744×$dipole$ R^2 = 0.921; R^2_{adj} = 0.918; $Q^2_{\text{CV-LOO}}$ = 0.915; MSE = 2.060; $PRESS$ = 181.367; F_{stat} = 317.231
GA-MLR2	logβ_{12} = -18.406 + 1.023×$SdsCH$ + 0.605×$SssNH$ + 35.458×$Hmin$ R^2 = 0.926; R^2_{adj} = 0.923; $Q^2_{\text{CV-LOO}}$ = 0.914; $RMSE$ = 1.929; $PRESS$ = 182.136; F_{stat} = 340.680
GA-MLR3	logβ_{12} = 46.440 + 1.014×$SssNH$ – 0.013×ΔH_f – 10.335×$SssS$ R^2 = 0.881; R^2_{adj} = 0.876; $Q^2_{\text{CV-LOO}}$ = 0.861; $RMSE$ = 3.102; $PRESS$ = 296.166; F_{stat} = 201.514
GA-MLR4	logβ_{12} = 15.492 + 1.093×$SssNH$ + 2.117×$logP$ – 2.603×$SsCH3$ R^2 = 0.916; R^2_{adj} = 0.913; $Q^2_{\text{CV-LOO}}$ = 0.905; $RMSE$ = 2.191; $PRESS$ = 202.764; F_{stat} = 296.586
GA-MLR5	logβ_{12} = -2,493 + 1,694×$SdsCH$ – 3,717×$LUMO$ + 0,580×$Hother$ R^2 = 0.890; R^2_{adj} = 0.885; $Q^2_{\text{CV-LOO}}$ = 0.873; $RMSE$ = 2.869; $PRESS$ = 269.696; F_{stat} = 220.060
GA-MLR6	logβ_{12} = -56.902 + 1.304×$SdsCH$ + 43.601×$Ovality$ – 2.282×$SssCH2$ R^2 = 0.927; R^2_{adj} = 0.924; $Q^2_{\text{CV-LOO}}$ = 0.921; $RMSE$ = 1.905; $PRESS$ = 169.285; F_{stat} = 345.367
GA-MLR7	logβ_{12} = 39.440 – 10.418×$nelem$ + 2.775×$numHBa$ + 0.239×SdO R^2 = 0.953; R^2_{adj} = 0.951; $Q^2_{\text{CV-LOO}}$ = 0.948; $RMSE$ = 1.227; $PRESS$ = 110.572; F_{stat} = 551.198

Notation	The models of QSPR$_{GA-MLR}$
GA-MLR8	$\log\beta_{12} = 17.881 - 2.220\times HsOH - 3.185\times SaasC + 0.452\times SaaCH$ $R^2 = 0.933$; $R^2_{adj} = 0.930$; $Q^2_{CV-LOO} = 0.928$; $RMSE = 1.745$; $PRESS = 154.123$; $F_{stat} = 379.470$
GA-MLR9	$\log\beta_{12} = 12.404 + 2.446\times numHBa + 0.869\times HssNH + 1.167\times x0$ $R^2 = 0.902$; $R^2_{adj} = 0.899$; $Q^2_{CV-LOO} = 0.891$; $RMSE = 2.540$; $PRESS = 233.159$; $F_{stat} = 252.159$
GA-MLR10	$\log\beta_{12} = 17.881 - 3.185\times SaasC + 0.451\times SaaCH - 0.635\times SsOH$ $R^2 = 0.933$; $R^2_{adj} = 0.930$; $Q^2_{CV-LOO} = 0.928$; $RMSE = 1.745$; $PRESS = 154.123$; $F_{stat} = 379.470$
GA-MLR11	$\log\beta_{12} = 47.833 - 13.757\times xc3 + 0.234\times SdO - 1.219\times HsOH$ $R^2 = 0.917$; $R^2_{adj} = 0.914$; $Q^2_{CV-LOO} = 0.907$; $RMSE = 2.153$; $PRESS = 198.253$; $F_{stat} = 302.361$
GA-MLR12	$\log\beta_{12} = 75.148 + 1.098\times SssNH - 14.719\times SssS - 7.935\times Hmax$ $R^2 = 0.933$; $R^2_{adj} = 0.930$; $Q^2_{CV-LOO} = 0.924$; $RMSE = 1.747$; $PRESS = 161.466$; $F_{stat} = 379.466$

3.2 Constructing QSPR$_{GA-SVR}$ models

The QSPR$_{GA-SVR}$ model was developed from variables such as SssNH, SssS, and Hmax of the QSPR$_{GA-MLR12}$ model built in the above section. This non-linear regression technique is one of the ML methods along with the ANN network that has been widely used in developing effective regression models.

In this study, the RBF function was used to search for the optimal values of kernel parameters, such as C, γ, and ε, with support vector numbers (n) [24]. These optimal values are found when successfully setting up the SVR model with the statistical values of the training data set, with suitable values of R^2_{train}, Q^2_{test}, and Q^2_{CV}. At the same time, the preparation of the SVR model is combined with the external assessment technique on the data set in Table 4 controlled by statistical values Q^2_{EV-SVR} and MARE(%).

The QSPR$_{GA-SVR}$ model preparation process was carried out with the Weka tool [30] with the change of the C core parameter at 0.010, 0.1, 1.0, 10, and 100; the γ parameter was randomly selected from 0.01 to 10.0. The QSPR$_{GA-SVR}$ model result was successfully constructed with parameters such as C = 10.0; γ = 0.333; ε = 0.100, and the support vector number n = 51. The statistical values of the model are as follows: R^2_{train} = 0.9810; Q^2_{test} = 0.9892, Q^2_{CV} = 0.9867, and RMSE = 0.980. Meanwhile, the external evaluation process received very good results: Q^2_{EV-SVR} = 0.9397 (Figure 4) and MARE(%) = 6.3199.

3.3 Constructing QSPR$_{GA-ANN}$ models

As presented, the QSPRGA-ANN model was also developed based on the three descriptive variables of the QSPR$_{GA-MLR12}$ model, namely SssNH, SssS, and Hmax. The construction was based on the MLP network type and the back-propagation algorithm. Therefore, the network architecture used in this case is I(3)-HL(m)-O(1), in which the input are three descriptive variables, the output is log$_{\beta12}$, and the hidden layer nodes are m. Statistical parameters R^2_{train}, Q^2_{test}, and Q^2_{CV} are used to control the network, and the results must respond to Tropsha standards [9].

The search for the ideal ANN model involves two steps: first, proceeding through MLP networks with various topologies using the initial training data set to identify the hidden nodes of HL(m). Table 3 displays the findings for the networks that met the statistical criteria; then, using the MARE(%) value and the Q^2_{EV-ANN} index between the predicted and experimental values from the external evaluation data set in Table 4 to identify the best network. The ANN model of the

pISSN 1859-1388
eISSN 2615-9678

architecture I(3)-HL(10)-O(1) network with the statistical parameters denoted in bold in Table 3 was found, and the external evaluation results in the value $Q^2_{\text{EV-ANN}}$ of 0.9407 (Figure 3b) and the MARE(%) of 5.5795 % (Figure 4). To compare the anticipated and experimental values on the external evaluation data set of these three models, we performed a one-way ANOVA analysis. The findings demonstrated that there was no statistically significant difference between the three predicted models (F = 0.2490 $F_{0.05}$ = 2.7395).

Table 3. Initial survey results of QSPR$_{\text{GA-ANN}}$ models with the architecture of I(3)-HL(m)-O(1)

QSPR$_{\text{GA-ANN}}$	R^2_{train}	Q^2_{test}	Q^2_{CV}	Training error	Test Error	Validation Error	Transfer Function
I(3)-HL(6)-O(1)	0.993	0.993	0.994	0.178	0.133	0.207	Eq. (8)
I(3)-HL(9)-O(1)	0.998	0.992	0.994	0.044	0.147	0.156	Eq. (8)
I(3)-HL(4)-O(1)	0.994	0.990	0.994	0.163	0.166	0.149	Eq. (8)
I(3)-HL(10)-O(1)	**0.997**	**0.992**	**0.994**	**0.080**	**0.157**	**0.187**	**Eq. (7)**
I(3)-HL(7)-O(1)	0.995	0.993	0.995	0.134	0.145	0.203	Eq. (7)

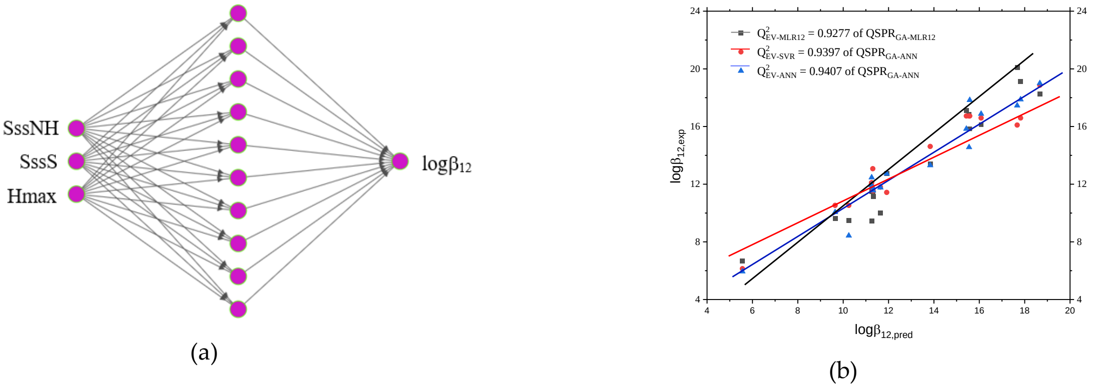


Fig. 3. (a) ANN model with I(3)-HL(10)-O(1) architecture; (b) Correlation in external-validation data set of three QSPR models

3.4 External validation of QSPR models

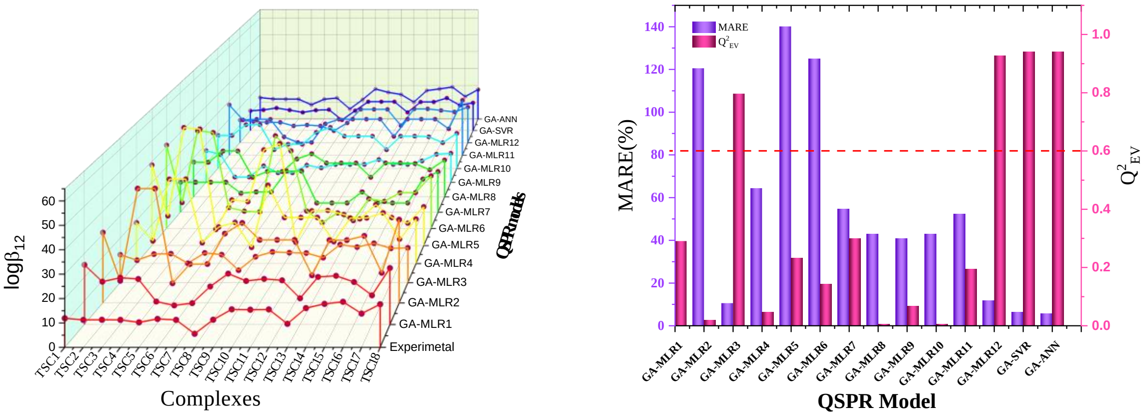
The model must be subjected to external validation before it is finished. It needs to be assessed using a different data set. This work uses an external data set with eighteen experimental complexes (Table 4). The assessment process was based on the Q^2_{EV} and MARE(%) values of the models. Figure 4 illustrates the results from the MLR, SVR, and ANN models.

Table 4. External validation of 18 experimental stability constant values

Notation	Thiosemicarbazone				Metal ions	log$\beta_{12,\text{exp}}$	ref.
	R$_1$	R$_2$	R$_3$	R$_4$			
TSC1	H	$-CH_3$	$-CH_3$	$-C_5H_4N$	Ni^{2+}	11.9191	[31]
TSC2	H	H	$-CH_3$	$- C_5H_4N$	Ni^{2+}	11.2410	[32]
TSC3	H	$-C_6H_5$	$-C_6H_5$	$-C_5H_4N$	Cu^{2+}	11.3050	[33]
TSC4	H	$-C_6H_5$	$-C_6H_5$	$-C_5H_4N$	Zn^{2+}	11.2610	[33]

DOI: 10.26459/hueunijns.v133i1D.7306

Notation	Thiosemicarbazone				Metal ions	log$\beta_{12,exp}$	ref.
	R$_1$	R$_2$	R$_3$	R$_4$			
TSC5	H	H	H	– C_9H_8N	Hg^{2+}	10.2493	[15]
TSC6	H	$–C_6H_5$	$–CH_3$	– $C_{10}H_{12}NO$	Cu^{2+}	11.6434	[34]
TSC7	H	$–C_6H_5$	$–CH_3$	– $C_{10}H_{12}NO$	Hg^{2+}	11.2569	[34]
TSC8	$–CH_3$	$–CH_3$	$–C_5H_4N$	$–C_5H_4N$	Zn^{2+}	5.5600	[35]
TSC9	H	C_6H_5	$–C_5H_4N$	$–C_5H_4N$	Ni^{2+}	11.3200	[36]
TSC10	H	H	H	$–C_{10}H_6OH$	Sm^{3+}	15.5800	[37]
TSC11	H	H	H	$–C_{10}H_6OH$	Eu^{3+}	15.4300	[37]
TSC12	H	H	H	$–C_{10}H_6OH$	Gd^{3+}	15.5500	[37]
TSC13	H	H	$–CH_3$	– $C_{10}H_{12}NO$	Ag^+	9.6457	[17]
TSC14	H	H	H	$–C_{10}H_6OH$	Co^{2+}	16.0800	[38]
TSC15	H	H	H	$–C_{10}H_6OH$	Ni^{2+}	17.8200	[38]
TSC16	H	H	H	$–C_{10}H_6OH$	Cu^{2+}	18.6700	[38]
TSC17	H	H	–	$–C_9H_7NO$	Cd^{2+}	13.8370	[19]
TSC18	H	H	H	$–C_6H_4NO_2$	Pb^{2+}	17.6700	[39]



(a) (b)

Fig. 4. (a) The predicted stability constant values (log$\beta_{12,pred}$) from QSPR models in the external validation data set; (b) The MARE(%) and Q^2_{EV} values from QSPR models

The acceptance of the QSPR$_{GA-MLR12}$ model was analyzed in the above section. Thus, according to the results from Figure 4, the QSPR$_{GA-SVR}$ and QSPR$_{GA-ANN}$ I(3)-HL(10)-O(1) models received the Q^2_{EV} values of 0.9397 and 0.9407, respectively, and the MARE(%) values of 6.3199 and 5.5795, respectively. This confirms that the predicted results of the two machine learning models are very good, and the ANN model gives the best-predicting results, then come the SVR model and the MLR model. Furthermore, the predicted values log$\beta_{12,pred}$ of the ANN network are very close to the published log$\beta_{12,exp}$ values.

pISSN 1859-1388
eISSN 2615-9678

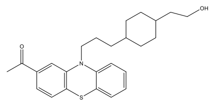
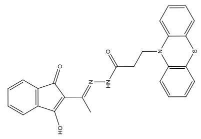
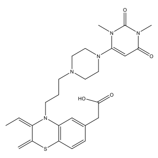
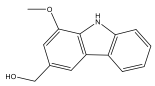
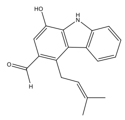
3.5 Development of new thiosemicarbazone and complexes

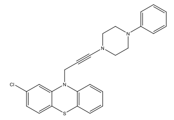
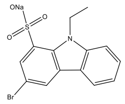
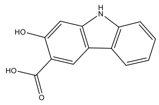
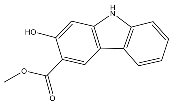
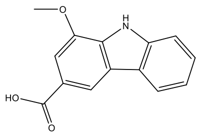
We chose to create a novel thiosemicarbazone using the carbazole and phenothiazine derivatives by putting these groups in the R_4 site of the thiosemicarbazone structure (Figure 1a), with hydrogen atoms in different positions (R_1, R_2, and R_3). The complexes consist of a few common metal ions: Ag^+, Cu^{2+}, Cd^{2+}, Ni^{2+}, and Zn^{2+} between novel ligands. The original derivatives benefit their particular antibacterial and antiviral activities and this led to the selection of the derivatives [40, 41]. The effect of thiosemicarbazone derivatives is the focus of this long-term research project. Additionally, these derivatives were synthesized in the real world and have made comprehensive announcements of applications in a variety of sectors [40, 41].

Twenty-two thiazole derivatives and 22 carbazole compounds were used to design a successful investigation producing 220 complexes with the five metal ions described above [42]. The application domain (AD) was tested by carefully screening and embedding the new complexes into the training data space [1, 8]. The findings were 20 novel complexes with 10 ligands and metal ions in the application domain AD through the D-cook value (<2.0) [9]; the stability constants of the new complexes were anticipated from three constructed QSPR models. Table 5 provides the expected values of these new compounds ($\log\beta_{12,new}$).

Furthermore, the single factor ANOVA method was also used to re-test the $\log\beta_{12,new}$ predicted values from the $QSPR_{GA-MLR12}$, $QSPR_{GA-SVR}$, and $QSPR_{GA-ANN}$ models. The results show that the stability constant value ($\log\beta_{12,new}$) of the models is correlative, meaning the difference is insignificant ($F = 0.4006 < F_{0.05} = 3.1588$).

Table 5. Twenty new developed complexes with the $\log\beta_{12,new}$ calculated values from QSPR models

R_4 site	Metal ions	$\log\beta_{12,new}$		
		$QSPR_{GA-MLR12}$	$QSPR_{GA-SVR}$	$QSPR_{GA-ANN}$
	Ag^+	7.1144	7.9174	8.2359
	Ag^+	8.9684	7.9190	8.8966
	Ag^+	6.4773	7.9199	8.8533
	Cu^{2+}	8.2846	7.9203	8.2977
	Cd^{2+}	7.7944	7.9196	8.7863

R$_4$ site	**Metal ions**	**log$\beta_{12,new}$**		
		QSPR$_{\text{GA-MLR12}}$	**QSPR$_{\text{GA-SVR}}$**	**QSPR$_{\text{GA-ANN}}$**
	Ag+	8.4101	7.9174	8.2331
	Cu^{2+}	12.3953	11.0277	11.1617
	Ni^{2+}	12.2136	12.3616	12.1617
	Zn^{2+}	13.8441	11.9620	12.1368
	Cd^{2+}	9.1190	7.9191	8.8685
	Ni^{2+}	9.4532	8.8563	8.8770
	Zn^{2+}	9.4532	9.2365	9.6530
	Cu^{2+}	8.7180	8.9193	8.8214
	Ni^{2+}	8.8265	9.1353	9.1214
	Zn^{2+}	8.9563	9.5639	9.8363
	Cd^{2+}	8.3839	7.9194	8.8100
	Cu^{2+}	8.6728	8.7569	8.8450
	Ni^{2+}	8.8695	8.9266	8.9653
	Zn^{2+}	9.0232	9.3251	10.0232
	Cd^{2+}	8.3386	7.9193	8.8354

4 Conclusion

We successfully developed three quantitative structure-property relationship models using combined techniques: multivariate linear regression (QSPR$_{\text{GA-MLR}}$), support vector regression (QSPR$_{\text{GA-SVR}}$), and artificial neural network (QSPR$_{\text{GA-ANN}}$) with variable analysis by using genetic algorithm. The structure of the experimental complexes between thiosemicarbazone and metal ions was optimized via semi-empirical quantum computation with the new versions PM7 and PM7/sparkle. The construction process as well as the training of models were fully evaluated through internal and external evaluation carefully based on statistical indicators such as R^2_{train}, R^2_{adj}, $Q^2_{\text{CV-LOO}}$, Q^2_{EV}, RMSE, and MARE(%) and in combination with the one-factor ANOVA method to compare the correlation in the prediction of the models. The results were obtained from three models, including the QSPR$_{\text{GA-MLR12}}$ model, the QSPR$_{\text{GA-SVR}}$ model with parameters C = 10.0; γ = 0.333, and ε = 0.10 with the support vector numbers equal to 51, and the QSPR$_{\text{GA-ANN}}$ I(3)-HL(10)-O(1) architecture. All of the models met the requirements for predictability. Among them, the ANN model has the best predictive ability, followed by the SVR model, and finally the MLR model. The results of this study enable the design of new thiosemicarbazone derivatives and the prediction of their ability to complex with metal ions, thus opening up new directions for applying these ligands and complexes to multiple applications in the fields of analytical chemistry, pharmaceutical drug development, and environmental impact assessment.

pISSN 1859-1388
eISSN 2615-9678

References

1. Roy K, Kar S, Das RN. A Primer on QSAR/QSPR Modeling. Fundamental Concepts. New York (USA). Springer; 2015.
2. Jiang J, Duan W, Wei Q, Zhao X, Ni L, Pan Y, Shu CM. Development of quantitative structure-property relationship (QSPR) models for predicting the thermal hazard of ionic liquids: A review of methods and models. J Mol Liq. 2020;301:112471.
3. Abramenko N, Kustov L, Metelytsia L, Kovalishyn V, Tetko I, Peijnenburg W. A review of recent advances towards the development of QSAR models for toxicity assessment of ionic liquids. J Hazar Mater. 2020;384.
4. Kumar S, Dhar DN, Saxena PN. Applications of metal complexes of Schiff bases-A review. J Sci Ind Res. 2009;68.
5. Khan T, Ahmad R, Joshi S, Khan AR. Anticancer potential of metal thiosemicarbazone complexes: a review. Chem Sin. 2015;6.
6. Patel NC, Patel BA. Spectrophotometric Method for determination of Copper (II) using p-Chlorobenzaldehyde -4-(2′-carboxy-5-sulphophenyl)-3-thiosemicarbazone [pCBCST]. Res J Chem Sci. 2014;4(2):1-6.
7. Stewart JJP. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. J Mol Model. 2013;19:1-32.
8. OECD, Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships Models, Organisation for Economic Co-operation and Development (France); 2007.
9. Golbraikh A, Tropsha A. Beware of q2!. J Mol Graph Model. 2002;20(4):269-76.
10. MacQueen JB. Some methods for classification and analysis of multivariate observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press (USA). 1967;281-297.
11. Forgy EW. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics. 1965;21:768-769.
12. Friedman HP, Rubin J. On some invariant criteria for grouping data. Journal of the American Statistical Association. 1967;62:1159-1178.
13. Reddy KH, Prasad NBL, Reddy TS. Analytical properties of 1-phenyl-1,2-propanedione-2-oxime thiosemicarbazone: simultaneous spectrophotometric determination of copper(II) and nickel(II) in edible oils and seeds. Talanta. 2003; 59:425-433.
14. Reddy KH, Prasad NBL. Spectrophotometric determination of copper (II) in edible oils and seed using novel oxime-thiosemicarbazones. India J Chem. 2004;43A:111-114.
15. Reddy SAN, Reddy KJ, Duk LK, Reddy AV. Evaluation of 2,6-diacetylpyridinebis-4-phenyl-3-thiosemicarbazone as complexing reagent for zinc in food and environmental samples. J Saudi Chem Soc. 2012;16:1-9.
16. El-Karim ATA, El-Sherif AA. Potentiometric, equilibrium studies and thermodynamics of novel thiosemicarbazones and their bivalent transition metal(II) complexes. J Mol Liq. 2016;219:914-922.
17. Atalay T, Ozkan E. Thermodynamic studies of some complexes of 4′-morpholinoacetophenone thiosemicarbazone. Thermochimica Acta. 1994; 237:369-374.
18. Garg BS, Jain VK. Determination of thermodynamic parameters and stability constants of complexes of biologically active o-vanillinthiosemicarbazone with bivalent metal ions. Thermochimica Acta. 1989;146:375-379.
19. Sarkar K, Garg BS. Determination of thermodynamic parameters and stability constants of the complexes of p-MITSC with transition metal ions. Thermochimica Acta. 1987;113:7-14.
20. Guha R, Willighagen E. A survey of quantitative descriptions of molecular structure. Curr Top Med Chem. 2012;12:1946-1956.
21. QSARIS 1.1. Statistical Solutions Ltd (USA); 2001.
22. Holland JH. Genetic algorithms. Sci Am. 1992; 267:44-50.
23. Steppan DD, Werner J, Yeater PR. Essential Regression and Experimental Design for Chemists and Engineers; 1998.
24. Alex JS, Bernhard S. A tutorial on support vector regression. Statistics and Computing. 2004;14:199-222.
25. Cortes C, Vapnik VN. Support-vector networks. Machine Learning. 1995;20:251-261.
26. Gasteiger J, Zupan J. Neural networks in chemistry. Chiw Inr Ed EngI. 1993;32:503-521.

27. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533-536.

28. Matlab R2016a 9.0.0.341360, USA. MathWorks; 2016.

29. Quang NM, Mau TX, Nhung TTA, An TNM, Tat PV. Novel QSPR modeling of stability constants of metal-thiosemicarbazone complexes by hybrid multivariate technique: GA-MLR, GA-SVR and GA-ANN. J Mol Struct. 2019;1195:95-109.

30. Waikato Environment for Knowledge Analysis-Version 3.9.3. The University of Waikato Hamilton. New Zealand; 1999-2018.

31. Reddy DN, Reddy KV, Tegegne BM, Reddy VK. Development of a Highly Sensitive Extractive Spectrophotometric Method for the Determination of Nickel(II) from Environmental Matrices Using 2-Acetylpyridine-4-methyl-3-thiosemicarbazone. American J Anal Chem. 2012;3:719-726.

32. Reddy SL, Sekhar KBC. Analytical applications of 3-acetylpyridine thiosemicarbazone (3-APT): Simple and sensitive spectrophotometric determination of nickel (II) in soil and alloy samples. Int J Bas App Chem Sci. 2013;3(4):62-68.

33. Atalay T, Akgemci EG. Thermodynamic Studies of Some Complexes of 2-benzoylpyridine 4-phenyl-3-thiosemicarbazone. India J Chem. 1998;22:123-127.

34. Atalay T, Ozkan E. Thermodynamic stabilities: thermodynamic parameters of some complexes of 4′-morpholinoacetophenone 4-phenyl-3-thiosemicarbazone. Thermochimica Acta. 1994;246:193-197.

35. Gaál A, Orgován G, Polgári Z, Réti A, Mihucz VG, Bősze S, Szoboszlai N, Streli C. Complex forming competition and in-vitro toxicity studies on the applicability of di-2-pyridylketone-4,4,-dimethyl-3-thiosemicarbazone (Dp44mT) as a metal chelator. J Inorg Biochem. 2014;130:52-58.

36. Bernhardt PV, Sharpe PC, Islam M, Lovejoy DB, Kalinowski DS, Richardson DR. Iron chelators of the dipyridylketone thiosemicarbazone class: Precomplexation and transmetalation effects on anticancer activity. J Med Chem. 2009;52(2):407-415.

37. Sahadev S, Sharma RK, Sindhwani SK. Potentiometric studies on the complexation equilibria between Some trivalent lanthanide metal ions and biologically active 2-Hydroxy-1-naphthaldehyde thiosemicarbazone (HNATS). Monatshefte fur Chemie. 1992;123:883-889.

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

39. Sawhney SS, Sati RM. pH-metric studies on Cd(II)-, Pb(II)-, AI(III)-, Cr(III)- AND Fe(III)-p-nitrobenzaldehyde thiosemicarbazone systems. Thermochimica Acta. 1983;66:351-355.

40. Al-Busaidi, Haque, Al Rasbi, Khan M. Phenothiazine-based derivatives for optoelectronic applications: A review. Synthetic Metals. 2019;257:116-189.

41. Huang L, Feng ZL, Wang YT, Lin LG. Anticancer carbazole Alkaloids and coumarins from Clausena plants: A review. Chin J Nat Med. 2017;15(12):881-888.

42. Quang NM. Design, screening and synthesis of thiosemicarbazone derivatives and metal-thiosemicarbazone complexes using quantum chemistry calculation and QSPR modeling methods. Theoretical chemistry and physical chemistry, Hue City; 2020.