Heliyon 10 (2024) e36373

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Genetic algorithm multiple linear regression and machine learning-driven QSTR modeling for the acute toxicity of sterol biosynthesis inhibitor fungicides

Mohsen Abbod [a,b], Naser Safaie [b,*], Khodayar Gholivand [c]

[a] Department of Plant Protection, Faculty of Agriculture, Al-Baath University, Homs, Syria
[b] Department of Plant Pathology, Faculty of Agriculture, Tarbiat Modares University, P.O.B. 14115-336, Tehran, Iran
[c] Department of Chemistry, Faculty of Science, Tarbiat Modares University, P.O.B. 14115-175, Tehran, Iran

ARTICLE INFO

Keywords:
Ergosterol
Demethylation inhibitor
Multi-layer perceptron
QSAR
SBIs

ABSTRACT

Sterol Biosynthesis Inhibitors (SBIs) are a major class of fungicides used globally. Their widespread application in agriculture raises concerns about potential harm and toxicity to non-target organisms, including humans. To address these concerns, a quantitative structure-toxicity relationship (QSTR) modeling approach has been developed to assess the acute toxicity of 45 different SBIs. The genetic algorithm (GA) was used to identify key molecular descriptors influencing toxicity. These descriptors were then used to build robust QSTR models using multiple linear regression (MLR), support vector regression (SVR), and artificial neural network (ANN) algorithms. The Cross-validation, Y-randomization test, applicability domain methods, and external validation were carried out to evaluate the accuracy and validity of the generated models. The MLR model exhibited satisfactory predictive performance, with an R^2 of 0.72. The SVR and ANN models obtained R^2 values of 0.7 and 0.8, respectively. ANN model demonstrated superior performance compared to other models, achieving R^2_{cv} and R^2_{test} values of 0.74 and 0.7, respectively. The models passed both internal and external validation, indicating their robustness. These models offer a valuable tool for risk assessment, enabling the evaluation of potential hazards associated with future applications of SBIs.

1. Introduction

SBI fungicides have gained prominence and continue to serve as an important class of fungicides on a global scale [1,2], with total sales surpassing billions of dollars worldwide. SBI fungicides are categorized into four groups (G1-G4) depending on their target site of action within the ergosterol biosynthesis pathway, as delineated by the Fungicide Resistance Action Committee (FRAC) classification. Each group may consist of various chemical classes [3]. Demethylation inhibitors (DMIs) or group G1, are undeniably the main category of SBI fungicides [1–3]. These fungicides have a specific target, namely the Cyp51 enzyme, which effectively impedes the biosynthesis of ergosterol in fungi [4]. With a significant 29.2 % share in fungicide sales, DMIs encompass various chemical classifications such as Triazoles, imidazoles, piperazines, pyrimidines, and pyridines [5]. In addition to DMIs, SBI fungicides consist of other groups known as G2 (Amines), G3 (Keto-reductase inhibitors), and G4 (Squalene-epoxidase inhibitors). Similar to the DMIs, these

* Corresponding author.
E-mail address: nsafaie@modares.ac.ir (N. Safaie).

https://doi.org/10.1016/j.heliyon.2024.e36373
Received 5 May 2024; Received in revised form 12 August 2024; Accepted 14 August 2024
Available online 15 August 2024

groups belong to distinct chemical classes [2,3]. Despite their extensive utilization, numerous studies have revealed the dangers associated with the exposure to SBI fungicides. The exposure to Myclobutanil, Triadimefon [6], and Triazoles [7] has been found to result in reproductive toxicity. Furthermore, the exposure to Triazoles has been linked to hepatotoxic effects [8]. On the other hand, Propiconazole has been observed to exhibit embryotoxicity in mice [9] and also induces liver toxicity [10]. Studies have demonstrated that penconazole can lead to testicular dysfunction and impairment [11], as well as hepatotoxicity [12] in rats.

In the field of toxicological research, the primary approach for categorizing the possible risk associated with pesticides is by assessing their acute oral LD_{50} value [13]. This approach classifies pesticides into five distinct categories. More recently, the acute toxicity hazard classifications outlined in the Globally Harmonized System of Classification and Labelling of Chemicals (GHS) have come into use [13].

Quantitative structure-activity relationship (QSAR) techniques including CoMFA [14], principal components analysis (PCA) [15], and genetic algorithms (GA) [16,17] might offer a valuable instrument to anticipate the biological characteristics of chemical compounds. Early QSAR methods primarily employed linear regression for model building [18]. However, recent developments have seen the integration of machine learning algorithms (ML), including Support Vector Machines (SVM) [19], Artificial Neural Networks (ANNs) [20], and Partial Least Squares (PLS) [21], to explore non-linearity in chemical structure-activity relationships. These algorithms handle big data, enabling researchers to delve deeper into the intricate connections between chemical properties and biological effects [22].

In this work, we have employed the genetic algorithm combined with Multiple Linear Regression (GA-MLR), Support Vector Regression (SVR), and Multi-layer perceptron artificial neural network (MLP-ANN) approaches to propose validated QSTR models to assess the acute toxicity of 45 different SBIs. These models hold potential for the development of future sterol inhibitor fungicides to minimize the risk of toxicity to non-target organisms, especially humans. To our knowledge, this study is the first QSTR analysis specifically focusing on the acute toxicity of SBI fungicides.

2. Materials and methods

2.1. Data set

According to the FRAC classification [3], approximately 50 compounds have been classified as SBIs which are widely used globally in agriculture. The dataset employed in this study included 45 molecules of SBIs fungicides with information on acute oral LD_{50} in rats. These data were collected from the World Health Organization database [13], PubChem [23], and the National Library of Medicine (NLM) [24]. The structure of the compounds was modeled using the MM + force field in HyperChem [25], followed by a geometry optimization using the AM1 semi-empirical method, achieving a root mean square gradient of 0.01 kcal mol^{-1}. The chemical structures of the studied compounds are illustrated in Table S1.

2.2. Descriptors generation

A broad set of 0D-3D molecular descriptors, selected from 1497 calculated using E-Dragon 3.0 software [26], were used to capture structural features for the modeling [26]. A total of 492 descriptors were selected for QSTR modeling after removing those with low variance (standard deviation < 0.001), missing values, and high correlations ($R > 0.9$). The E-Dragon 3.0 provides comprehensive information on these descriptors, including their definitions, calculation procedures, and related references [26].

2.3. Genetic algorithm (GA)

To identify the primary descriptors influencing the toxicity of SBI compounds, the Genetic Algorithm (GA) method was employed. This method, modeled after Darwin's natural selection and evolution principles, has proven effective in variable selection [27]. The optimal models identified five descriptors, and the GA was implemented using MATLAB software (R2022a) [28].

2.4. Dataset splitting using K-means classification

After descriptor selection with the GA, the dataset was split into training and test sets using k-means classification. The dataset was split 80/20 into training and test sets. The test set included 9 randomly selected compounds from each k-means cluster, with the remaining 36 compounds forming the training set [29].

2.5. Multiple linear regression (MLR) model

A Multiple Linear Regression (MLR) analysis was conducted to explore the relationship between physicochemical characteristics and the acute toxicity of SBIs. The LD_{50} values for the 45 compounds were converted to pLD_{50} values ($pLD_{50} = -\log LD_{50}$) and served as the dependent variable. The five molecular descriptors selected previously were used as independent variables. The coefficients for the regression model were calculated using MLR and the least-squares curve fitting method, yielding the following regression equation [30]:

$$Y = a_0 + \sum_{i=1}^{n} a_i X_i \quad (1)$$

In equation (1), Y signifies acute toxicity or pLD_{50}, X_i represents the molecular descriptors, n denotes the number of descriptors, a_0 is the constant of the equation, and a_i are the coefficients associated with each descriptor [30].

2.6. ML-based QSTR models

In addition to GA-MLR, our QSTR modeling incorporated a range of ML techniques, encompassing Support Vector Regression (SVR) and Artificial Neural Networks (ANN).

2.6.1. Support vector regression (SVR) model

Support Vector Regression (SVR) is a form of Support Vector Machine (SVM) algorithms that has been recognized as a potent tool in ML-based QSAR researches [31]. The current investigation utilized Bayesian optimization on the hyperparameters: kernel function, box constraint level, and kernel scale. The molecular descriptors employed in the GA-MLR model were also utilized in the ML-based models. Model accuracy was assessed using two key metrics: the coefficient of determination (R^2) and the mean squared error (MSE). The SVR was carried out using the regression learner tool on MATLAB software (R2022a) [28].

2.6.2. Multi-layer perceptron artificial neural network model (MLP-ANN)

ANNs, as an ML technique, may offer a viable method for investigating complex problems. This particular technique finds its origins in the behavior of biological nervous systems [32]. Among the most frequently employed algorithms for ANNs is the multilayer perceptron (MLP-ANN) [33]. This network might have a single hidden layer or multiple hidden layers, and the number of neurons within each hidden layer is determined by the complexity of the input and output data. This study used MLP-ANN to assess the effectiveness of the molecular descriptors identified by the GA-MLR model. The Levenberg-Marquardt backpropagation algorithm was used for biases and weights optimization due to its fast convergence ability [32,33]. Mean square error (MSE) was used as a Loss Function, mu value was set to be 0.0001 upon completion of the training, maximum number of Epochs and validation checks were 1000 and 6, respectively. Prediction accuracy was estimated by the R^2, and MSE for both train and test sets. The ANN analysis was

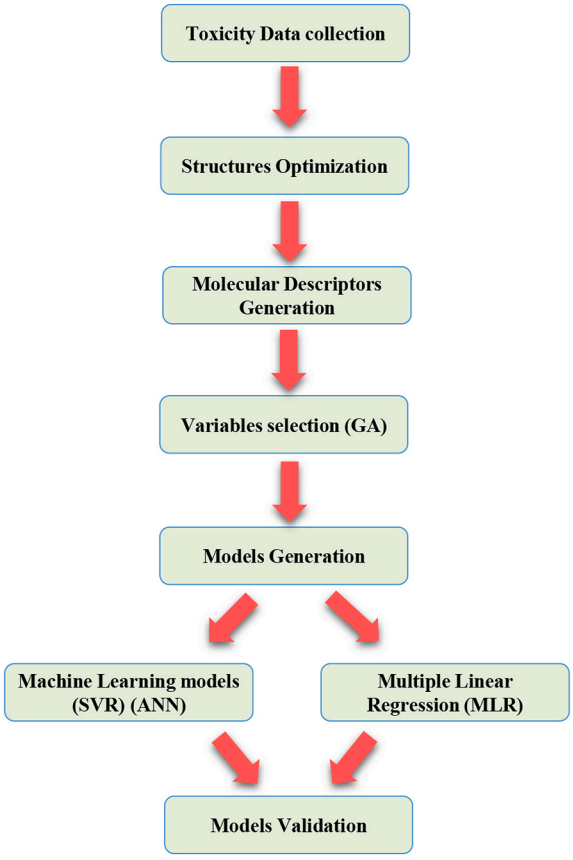

Fig. 1. A flow diagram illustrating the QSTR process used in this study.

achieved using MATLAB software (R2022a) [28].

2.7. Validation of the QSTR models

2.7.1. Cross validation

In our investigation, we explored the examination of the reliability of statistical models through the utilization of cross-validation, a technique that is commonly employed [34,35]. The assessment of the predictive capacity of the models was investigated through Leave-many-out cross-validation (LMO-CV) technique. The LMO-CV employs leave-m-out cross-validation, where m = 20 % of the compounds are used as the validation set, with the remaining 80 % used for training. A model is considered reliable if it gets an R^2_{cv} over 0.5, showing it can make accurate predictions on new data [34,35].

2.7.2. Y-randomization test

Y-randomization helps distinguish real relationships from random chance in an MLR model [36]. The test assesses the significance of an MLR model by scrambling the toxicity activity data (dependent variable) while keeping the molecular descriptors (independent variables) unchanged. A QSTR model is considered valid if it's R^2 and Q^2 values are significantly higher than the average of those obtained from randomized models, suggesting that the relationships found are not due to random chance [36].

2.7.3. Applicability domain (AD)

The OECD principles emphasize the importance of defining a domain of applicability for QSAR models [37]. The evaluation of the domain of applicability aids in determining whether the developed model is suitable for a particular set of molecules. The domain of applicability is characterized as a hypothetical region within the descriptor chemical space and the predicted activity [38]. Within this space, the QSAR model accurately predicts the biological activity of molecules, whereas molecules with inaccurately predicted activities fall outside this space and are considered outliers [38].

2.7.4. External validation

To assess the performance of the generated models, a test set consisting of 9 compounds (20 % of the dataset) was created using the previously described k-means classification method. These compounds were not used during model training. The predictive performance of the model for toxicity was assessed using two metrics: R^2_{test} and MSE_{test}. An R^2_{test} value exceeding 0.5 is considered an acceptable level of predictive performance [39]. The overall process of QSTR modeling is presented schematically in Fig. 1.

3. Results and discussion

3.1. GA-MLR model

GA-MLR approach procedure was carried out to define the main molecular descriptors that affected the acute toxicity of 45 SBIs. The logarithmic values of pLD_{50} (–log LD_{50}) were considered the dependent variable, while molecular structure descriptors were independent variables. Out of 1497 molecular descriptors calculated for each molecule, 492 molecular descriptors were selected and used with pLD_{50} values as input for model development. GA-MLR procedures were performed to develop a 5-variable model based on all types of molecular descriptors (i.e., 0D, 1D, 2D, 3D descriptors). Five molecular descriptors **R3u$^+$**, **ATS6e**, **Mor31u**, **RDF050m**, and **BELv4** (Table 1) [26] were chosen to develop the QSAR models as shown in equation (2):

$$\boldsymbol{pLD_{50}} = -9.04 + 35.48 \times (\mathbf{R3u^+}) + 0.01228 \times (\mathbf{ATS6e}) - 1.21 \times (\mathbf{Mor31u}) - 0.0429 \times (\mathbf{RDF050m}) + 2.415 \times (\mathbf{BELv4}) \quad (2)$$

$$n = 36, R^2_{train} = 0.72, R^2_{adj.} = 0.67, MSE_{train} = 0.04, P < 0.001, \text{VIF values } 1.38 \text{ to } 4.55, R^2_{test=} 0.66, MSE_{test} = 0.062, R^2_{cv} = 0.66.$$

The optimistic statistical quality of the GA-MLR model R^2_{train} (0.72), MSE_{train} (0.04), cross-validated R^2_{cv} (0.66), revealed a good predictive performance on the toxicity of the studied SBIs with maximum of 0.375 log unit deference for Spiroxamine (experimental pLD_{50} of −2.66 vs. predicted value of −3.04) (Fig. 2A and Table 2).: The statistical metrics of the external validation R^2_{test} and MSE_{test} demonstrated a good external performance of the model. The modest variance inflation factor (VIF) values demonstrated by the five descriptors in the MLR model, specifically 2.64, 4.55, 2.81, 1.38, and 2.28 for R3u+, ATS6e, Mor31u, RDF050m, and BELv4, respectively, signify the absence of multicollinearity among these descriptors. Fig. 2A demonstrated that the distribution of observed and predicted pLD_{50} values are significantly correlated because of the low value of MSE for both train and test sets. Eq. (2) revealed a

Table 1
The five selected descriptors by the GA and their meanings.

Descriptors	Chemical meanings	Descriptor group
R3u$^+$	R maximal autocorrelation of lag 3/unweighted	GETAWAY descriptors
ATS6e	Broto-Moreau autocorrelation of lag 6 weighted by Sanderson electronegativity	2D autocorrelations
Mor31u	3D-MoRSE - signal 31/unweighted	3D-MoRSE descriptors
RDF050m	Radial Distribution Function - 050/weighted by mass	RDF descriptors
BELv4	lowest eigenvalue n. 4 of Burden matrix/weighted by Van der Waals volumes	BCUT descriptors

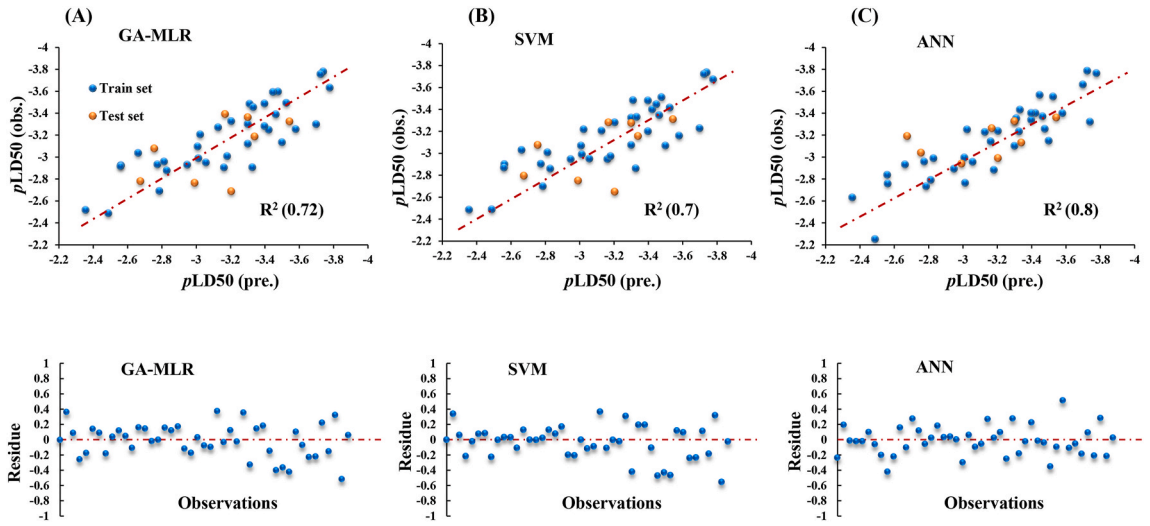

Fig. 2. Experimental *vs* predicted pLD_{50} values and residual plots. (A) GA-MLR, (B) SVM, and (C) ANN.

significant linear correlation between the toxicity of SBI compounds and the five descriptors identified through GA-MLR.

The GETAWAY descriptor **R3u +** [26] with index (+35.48) was the most contributor to the toxicity, followed by **BELv4** BCUT descriptors. The 3D-MoRSE descriptors **Mor31u** [40] and RDF descriptor **RDF050m** [41] affected the toxicity of the compounds in contrast manner with indexes of −1.21 and −0.0429, respectively. The descriptors most frequently chosen by Genetic Algorithm are from the category of 3D molecular descriptors [26], showcasing the superior efficacy of these descriptors in predicting the toxicity of SBIs. Recent QSAR studies have shown that 3D descriptors consistently outperform 2D and quantum-chemical descriptors in terms of predictive accuracy [41,42].

3.2. Y-Randomization test for GA-MLR model

The robustness of the GA-MLR model was evaluated by randomly shuffling the pLD_{50} variable and rebuilding the QSTR models with the same five molecular descriptors. This procedure was repeated 100 times. The resulting models consistently showed lower R^2 and R^2_{CV} values compared to the original GA-MLR model (Fig. 3), indicating that the predictive power of the original model is not solely due to random chance. This finding suggests that the estimated pLD_{50} values, based on the five descriptors outlined in Eq. (2), are not simply due to chance (Fig. 3).

3.3. Applicability domain (AD) analysis of MLR model

The Williams plot for the GA-MLR model is presented in Fig. 4. The AD is defined within a squared region encompassing ±2 standard deviations and a leverage threshold h^* of 0.46. Any compound that demonstrates standardized residuals in prediction above or below 2, or a leverage value above the threshold value ($hi > 0.46$) based on Fig. 4, is classified as an outlier of the AD [43]. As depicted in Fig. 4, the compounds 2 (Fluquinoconazole) and 8 (Prochloraz) from the test and compound 36 (Penconazole) from the train set can be regarded as outliers. Fluquinoconazole exhibits a high leverage value ($hi > 0.46$), indicating its distance from other compounds in the test set and its location outside the AD space. Despite exceeding the residual threshold, Penconazole and Prochloraz remain within the AD due to their leverage values, making their predictions acceptable (Fig. 4). It can be inferred that the developed MLR model exhibits a wide range of applicability and can accurately predict the acute toxicity of the SBIs within the relevant applicability thresholds.

3.4. ML-based QSTR models

In an attempt to enhance the correlation between the predicted toxicities derived from the initial GA-MLR model and the chosen molecular descriptors, another QSTR models were constructed employing SVR, and MLP-ANN. The ML-based QSTR model was built using the five descriptors identified by the GA-MLR model. The optimal hyperparameters for the SVR model, determined by Bayesian optimization, included a Gaussian kernel function, a box constraint level of 197.52, and a kernel scale of 30.55. The MSE_{train} of SVR model was 0.043, while the R^2_{train} and R^2_{cv} were 0.70 and 0.62, respectively. The external validation revealed R^2_{test} and MSE_{test} values of 0.63 and 0.066, respectively (Fig. 2B and Table 3).

In the ANN model, the neuron number in the input and output layers was 5 and 1, respectively. The Tansig transfer function was applied to compute the output of each layer based on its net input [32]. The experiment involved varying the number of neurons in the hidden layer, ranging from 1 to 30. It was found that the network had an architecture of 5/5/1 with 5 neurons in the hidden layer

Table 2
The pLD_{50} values predicted by QSTR models, compared to experimental values.

No.	SBI compound	$pLD_{50,\ (obs.)}$	MLR		SVR		ANN	
			pLD_{50}	Res.	pLD_{50}	Res.	pLD_{50}	Res.
Train test								
1	Azaconazole	−2.489	−2.488	−0.001	−2.489	0.001	−2.254	−0.235
2	Bromuconazole	−2.562	−2.927	0.365	−2.902	0.340	−2.759	0.197
3	Cyproconazole	−3.009	−3.098	0.089	−3.072	0.063	−2.997	−0.011
4	Difenoconazole	−3.162	−2.906	−0.256	−2.947	−0.215	−3.144	−0.02
5	Dodemorph	−3.422	−3.25	−0.173	−3.402	−0.021	−3.403	−0.02
6	Etaconazole	−3.128	−3.272	0.144	−3.207	0.079	−3.229	0.101
7	Fenarimol	−3.398	−3.489	0.091	−3.483	0.085	−3.339	−0.06
8	Fenbuconazole	−3.301	−3.123	−0.178	−3.077	−0.224	−3.103	−0.2
9	Fenhexamid	−3.74	−3.781	0.041	−3.74	−0.001	−3.323	−0.417
10	Fenpropimorph	−3.477	−3.598	0.121	−3.513	0.035	−3.259	−0.218
11	Flusilazole	−2.829	−2.879	0.050	−2.86	0.032	−2.99	0.161
12	Flutriafol	−3.057	−2.952	−0.105	−2.952	−0.105	−2.957	−0.1
13	Imazalil	−2.356	−2.519	0.163	−2.487	0.131	−2.633	0.277
14	Imibenconazole	−3.447	−3.594	0.147	−3.448	0.001	−3.568	0.121
15	Ipconazole	−2.948	−2.931	−0.018	−2.948	0.000	−2.894	−0.054
16	Mefentrifluconazole	−3.301	−3.302	0.001	−3.326	0.025	−3.327	0.026
17	Metconazole	−2.775	−2.934	0.160	−2.905	0.130	−2.959	0.185
18	Myclobutanil	−3.204	−3.328	0.124	−3.283	0.079	−3.239	0.034
19	Nuarimol	−3.311	−3.486	0.175	−3.485	0.174	−3.354	0.043
20	Piperalin	−3.398	−3.283	−0.115	−3.202	−0.196	−3.403	0.005
21	Propiconazole	−3.181	−3.009	−0.172	−2.975	−0.205	−2.885	−0.296
22	Pyributicarb	−3.724	−3.757	0.033	−3.725	0.001	−3.788	0.064
23	Pyrifenox	−3.464	−3.39	−0.074	−3.35	−0.115	−3.373	−0.091
24	Simeconazole	−2.786	−2.691	−0.095	−2.701	−0.086	−2.734	−0.052
25	Spiroxamine	−2.663	−3.038	0.375	−3.031	0.369	−2.934	0.272
26	Tebuconazole	−3.525	−3.495	−0.030	−3.418	−0.108	−3.553	0.027
27	Terbinafine	−3.332	−3.457	0.125	−3.332	0.000	−3.433	0.1
28	Tetraconazole	−3.013	−2.99	−0.023	−2.992	−0.022	−2.765	−0.25
29	Triadimefon	−2.56	−2.917	0.357	−2.871	0.311	−2.84	0.28
30	Triadimenol	−3.58	−3.255	−0.325	−3.162	−0.417	−3.401	−0.18
31	Tridemorph	−2.813	−2.96	0.147	−3.009	0.196	−2.792	−0.021
32	Triflumizole	−3.024	−3.208	0.184	−3.22	0.196	−3.252	0.228
33	Triforine	−3.778	−3.633	−0.145	−3.674	−0.104	−3.766	−0.012
34	Bitertanol	−3.699	−3.301	−0.398	−3.23	−0.469	−3.662	−0.037
35	Epoxiconazole	−3.5	−3.136	−0.363	−3.072	−0.428	−3.15	−0.35
36	Penconazole	−3.327	−2.907	−0.420	−2.864	−0.463	−3.235	−0.092
Test set								
1	Diniconazole	−2.676	−2.781	0.105	−2.797	0.121	−3.192	0.516
2	Fluquinoconazole	−2.049	−1.98	−0.069	−2.147	0.097	−1.943	−0.11
3	Pefurazoate	−2.992	−2.766	−0.225	−2.753	−0.24	−2.941	−0.051
4	Aldimorph	−3.544	−3.326	−0.218	−3.312	−0.23	−3.362	−0.182
5	Fenpropidin	−3.168	−3.393	0.225	−3.283	0.115	−3.263	0.095
6	Hexaconazole	−3.34	−3.189	−0.152	−3.158	−0.182	−3.134	−0.207
7	Oxpoconazole	−2.756	−3.08	0.324	−3.076	0.32	−3.041	0.285
8	Prochloraz	−3.204	−2.689	−0.515	−2.651	−0.553	−2.991	−0.21
9	Triticonazole	−3.301	−3.364	0.063	−3.277	−0.024	−3.329	0.028

having the best performance based on the MSE value. The Levenberg-Marquardt backpropagation algorithm was used to train the network. Mu value was 0.0001 at the end of the 17 epochs. The high R^2_{train} value (0.8) and the low MSE_{train} (0.03) indicated an excellent agreement between predicted and experimental pLD_{50} values (Fig. 2C–Tables 2 and 3). The R^2_{cv} value exhibited an increase of 0.74 in comparison to the other models. The R^2_{test} and MSE_{test} were 0.7 and 0.055 respectively. Table 3 presents the statistical parameters of the generated models.

3.5. Comparison the generated model

Analysis of the statistical metrics obtained from the QSTR models revealed that the ANN model achieved the highest level of accuracy compared to the other models (as shown in Table 3). The higher the R^2, and lower MSE values for both train and test sets, the better accuracy of the model. As shown in Table 3, the MLP-ANN model displayed better statistical parameters than the other models, thus highlighting the superiority of the MLP-ANN model over the GA-MLR and SVR model. External validation further confirmed the superiority of the ANN model, as it achieved higher R^2_{test} and lower MSE_{test} values. However, predictive performance of the MLR model was slightly better than SVR model (Table 3), with revealed the strong linear correlation between the toxicity and the molecular

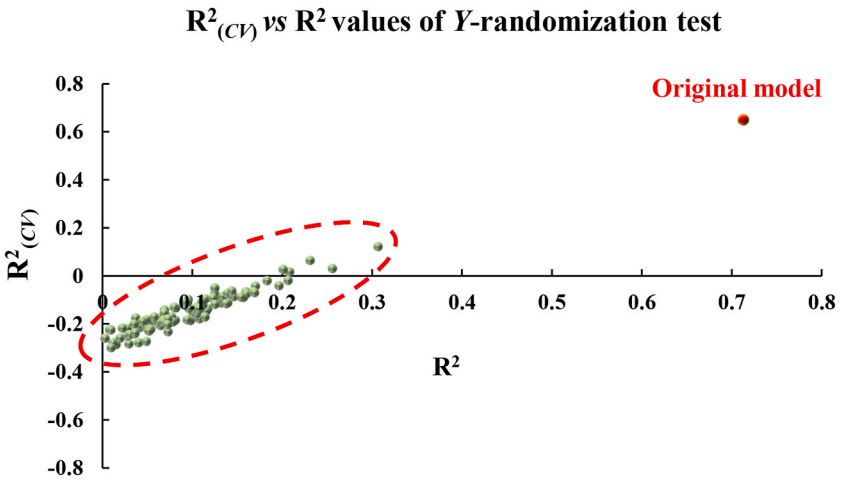

Fig. 3. Y-Randomization test findings for the GA-MLR model.

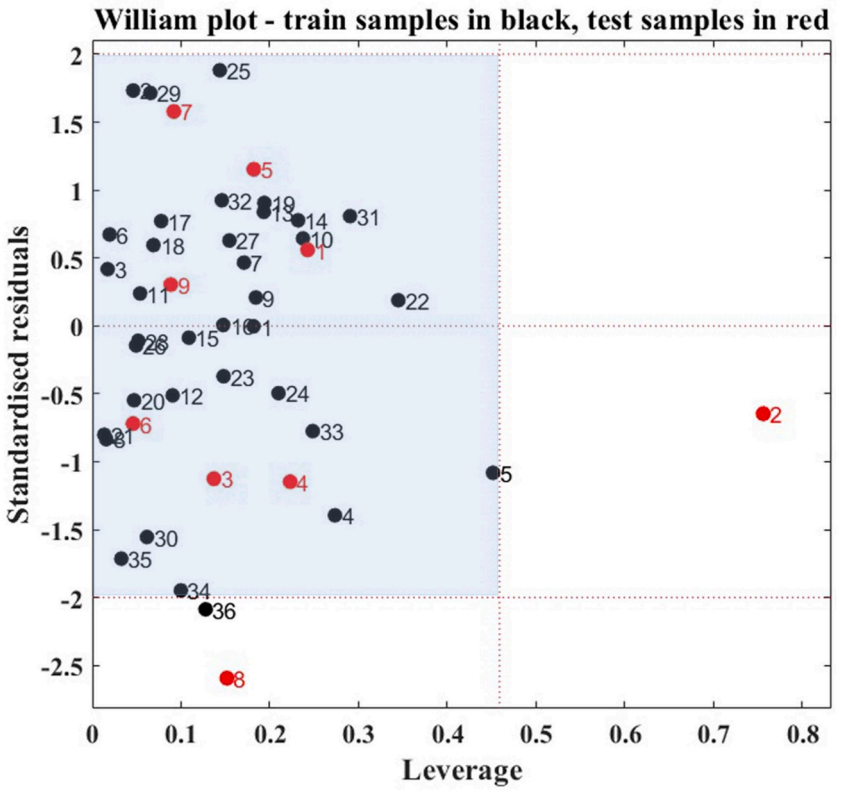

Fig. 4. Williams plots of the GA-MLR model. The training set is marked with black circles, the test set with red circles, and the sky blue area delineates the domain of applicability for the model.

descriptors of the compounds. A comparison of the predicted values for the studied compounds and their residuals for the three models can be found in Table 2 and Fig. 2. It is evident from Fig. 2 that in all models, the residuals were randomly distributed on both sides of zero, and neither model exhibited any proportional or systematic error. This indicates the high level of accuracy of generated models in predicting the toxicity of the studied compound. Several studied demonstrated the useful of ML techniques like SVM [44,45], and ANN [46] in QSAR studies. Support vector regression is widely employed for drug design [47], and toxicity prediction [48]. Herein, SVR model demonstrated good performance in toxicity prediction of the studied compounds but less than the ANN model.

No single algorithm reigns supreme in QSAR model development [49]. ANNs are widely used in QSAR studies and showed high level of accuracy in toxicity prediction [46,50]. The MLP-ANN model achieved the highest predictive performance among the QSTR

Table 3
Comparison of GA-MLR, SVR and ANN generated models.

Model	Train set			Test set	
	R^2_{train}	MSE	R^2_{cv}	R^2_{test}	MSE_{test}
MLR	0.72	0.040	0.66	0.66	0.062
SVR	0.70	0.043	0.62	0.63	0.066
ANN	0.8	0.03	0.74	0.7	0.055

models evaluated in this study, as confirmed by statistical analysis.

SBIs are extensively used in the field of medicine and agriculture for combating fungal pathogens [1,51] The inhibition of Ergosterol biosynthesis has promptly resulted in the emergence of drug resistance, thus compelling researchers and companies to explore new low-risk compounds and formulas falling into this category of antifungal agents [51,52]. The potential environmental and medicinal toxicities are essential factors in designing novel effective and low-risk SBI agents. Within the scope of this study, we have introduced individual QSTR models that have demonstrated a high ability in predicting the adverse effects associated with this particular type of chemical, thereby aiding researchers in the future development of safe compounds.

4. Conclusions

The present study affirms the potential of suggested models for acute toxicity prediction based on the R^2, R^2_{cv}, and MSE metrics. The current research presents four models aimed at predicting the potential toxicity of a significant group of compounds in the fields of agrochemicals and medicine. The generated models exhibited a good performance in predicting the acute toxicity of SBI fungicides. The results indicate that the MLP-ANN model outperformed the other models. The models demonstrated that the following five descriptors had a strong impact on the toxicity of SBIs: R3u+, ATS6e, Mor31u, RDF050m, Mor15u, and BELv4. In terms of risk assessment, these models can serve as a checkpoint to evaluate the potential hazards of proposed SBIs in the future.

Data availability statement

The trained ML models are available on request from the corresponding author.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Mohsen Abbod: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Investigation, Formal analysis, Data curation. **Naser Safaie:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Khodayar Gholivand:** Writing – review & editing, Validation, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e36373.

References

[1] J. Houšť, J. Spížek, V. Havlíček, Antifungal drugs, Metabolites 10 (3) (2020) 106, https://doi.org/10.3390/metabo10030106.
[2] K. Stenzel, J.P. Vors, Sterol biosynthesis inhibitors, Modern Crop Protection Compounds (2) (2019) 797–844, https://doi.org/10.1002/9783527699261.ch19. John Wiley & Sons.
[3] F.M. Fishel, M.M. Dewdney, Fungicide Resistance Action Committee's (FRAC) Classification Scheme of Fungicides According to Mode of Action. *PI94*, University of Florida, 2012, p. 7p.
[4] Z. Li, N. Liu, W. Yang, J. Tu, Y. Huang, W. Wang, C. Sheng, Controlling antifungal activity with light: optical regulation of fungal ergosterol biosynthetic pathway with photo-responsive CYP51 inhibitors, Acta Pharm. Sin. B 13 (7) (2023) 3080–3092, https://doi.org/10.1016/j.apsb.2023.02.008.
[5] S. Saha, Fungicides: the uncharted domain, J. Mycopathol. Res. 61 (2) (2023) 149–155, https://doi.org/10.57023/JMycR.61.2.2023.149.

[6] U.S. Environmental Protection Agency (U.S. EPA), Myclobutanil; Pesticide tolerances for emergency exemptions, Fed. Regist. 70 (2005) 49499–49507.
[7] X. Chen, J. Zheng, J. Zhang, M. Duan, H. Xu, W. Zhao, Y. Yang, C. Wang, Y. Xu, Exposure to difenoconazole induces reproductive toxicity in zebrafish by interfering with gamete maturation and reproductive behavior, Sci. Total Environ. 838 (2022) 155610, https://doi.org/10.1016/j.scitotenv.2022.155610.
[8] L.P. Marciano, L.F. Costa, N.S. Cardoso, J. Freire, F. Feltrim, G.S. Oliveira, F.B. Paula, A.C. Silvério, I. Martins, Biomonitoring and risk assessment of human exposure to triazole fungicides, Regul. Toxicol. Pharmacol. 147 (2024) p105565, https://doi.org/10.1016/j.yrtph.2024.105565.
[9] A.E.F.M. El-Shershaby, F.E.D.M. Lashein, A.A. Seleem, A.A. Ahmed, Developmental neurotoxicity after penconazole exposure at embryo pre-and post-implantation in mice, J. Histotechnol. 43 (3) (2020) 135–146, https://doi.org/10.1016/j.envpol.2020.116116.
[10] F.A.O. Joint, Pesticide residues in food 2017, FAO Plant Production and Protection Paper (2017) 232.
[11] A. Nowrozi, H. Johari, M. Shariati, Investigating the Protective effects of vitamin D3 on the Physiological and histopathological changes of the testis in adult rats treated with Penconazole, Pars J. Med. Sci. 21 (3) (2023) 21–29.
[12] M. Chaâbane, N. Soudani, K. Benjeddou, M. Turki, F. Ayadi Makni, T. Boudawara, N. Zeghal, R. Ellouze Ghorbel, The protective potential of Nitraria retusa on penconazole-induced hepatic injury in adult rats, Toxicol, Environ. Chem. 97 (9) (2015) 1253–1264, https://doi.org/10.1080/02772248.2015.1093633.
[13] G. Who, The WHO Recommended Classification of Pesticides by Hazard and Guidelines to Classification 2009, 2010.
[14] S. Banerjee, S.K. Baidya, N. Adhikari, T. Jha, 3D-QSAR studies: CoMFA, CoMSIA, and topomer CoMFA methods, in: Modeling Inhibitors of Matrix Metalloproteinases, CRC Press, 2023, pp. 32–53.
[15] M. Greenacre, P.J. Groenen, T. Hastie, A.I. d'Enza, A. Markos, E. Tuzhilina, Principal component analysis, Nat. Rev. Methods Primers 2 (1) (2022) 100, https://doi.org/10.1038/s43586-022-00184-w.
[16] M.A. Albadr, S. Tiun, M. Ayob, F. Al-Dhief, Genetic algorithm based on natural selection theory for optimization problems, Symmetry 12 (11) (2020) 1758, https://doi.org/10.3390/sym12111758.
[17] T.R. Noviandy, A. Maulana, G.M. Idroes, N.B. Maulydia, M. Patwekar, R. Suhendra, R. Idroes, Integrating genetic algorithm and LightGBM for QSAR modeling of acetylcholinesterase inhibitors in Alzheimer's disease drug discovery, Malacca Pharm 1 (2) (2023) 48–54, https://doi.org/10.60084/mp.v1i2.60.
[18] C. Hansch, T. Fujita, p-σ-π analysis. A method for the correlation of biological activity and chemical structure, J. Am. Chem. Soc. 86 (8) (1964) 1616–1626, https://doi.org/10.1021/ja01062a035.
[19] F. Khajehgili-Mirabadi, M.R. Keyvanpour, Enhancing QSAR modeling: a fusion of sequential feature selection and support vector machine, in: 2023 14th International Conference on Information and Knowledge Technology *(IKT)*, IEEE, 2023, pp. 44–49, https://doi.org/10.1109/IKT62039.2023.10433035.
[20] M. Gackowski, B. Madriwala, M. Koba, In silico design, docking simulation, and ANN-QSAR model for predicting the anticoagulant activity of thiourea isosteviol compounds as FXa inhibitors, Chem. Pap. 77 (11) (2023) 7027–7044, https://doi.org/10.1007/s11696-023-02994-y.
[21] X. Huo, J. Xu, M. Xu, H. Chen, An improved 3D quantitative structure-activity relationships (QSAR) of molecules with CNN-based partial least squares model, Artif. Intell. Life Sci. 3 (2023) 100065, https://doi.org/10.1016/j.ailsci.2023.100065.
[22] W.M. Neal, P. Pandey, S.I. Khan, I.A. Khan, A.G. Chittiboyina, Machine learning and traditional QSAR modeling methods: a case study of known PXR activators, J. Biomol. Struct. Dyn. 42 (2) (2024) 903–917, https://doi.org/10.1080/07391102.2023.2196701.
[23] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B.A. Shoemaker, P.A. Thiessen, B. Yu, L. Zaslavsky, PubChem in 2021: new data content and improved web interfaces, Nucleic Acids Res. 49 (D1) (2021) D1388–D1395, https://doi.org/10.1093/nar/gkaa971.
[24] National library of medicine (NLM). https://www.nlm.nih.gov/, 2011.
[25] HyperChem, Hypercube, Inc., http://www.hyper.com.
[26] R. Todeschini, V. Consonni, M. Pavan, DRAGON–Software for the calculation of molecular descriptors, rel. 1.12 for Windows, Free download available at: http://www.disat.unimib/chm, 2001.
[27] B. Alhijawi, A. Awajan, Genetic algorithms: theory, genetic operators, solutions, and applications, Evol. Intell. 17 (3) (2024) 1245–1256, https://doi.org/10.1007/s12065-023-00822-6.
[28] V. Matlab, (R2022a), The MathWorks inc., Natick, Massachusetts, 2022.
[29] J.T. Leonard, K. Roy, On selection of training and test sets for the development of predictive QSAR models, QSAR Comb, Sci. 25 (3) (2006) 235–251, https://doi.org/10.1002/qsar.200510161.
[30] L.S. Aiken, S.G. West, R.R. Reno, Multiple Regression: Testing and Interpreting Interactions, sage, 1991, p. 212p.
[31] M. Abbod, A. Mohammad, Combined interaction of fungicides binary mixtures: experimental study and machine learning-driven QSAR modeling, Sci. Rep. 14 (1) (2024) p12700, https://doi.org/10.1038/s41598-024-63708-2.
[32] I. Pantic, J. Paunovic, J. Cumic, S. Valjarevic, G.A. Petroianu, P.R. Corridon, Artificial neural networks in contemporary toxicology research, Chem. Biol. Interact. 369 (2023) p110269, https://doi.org/10.1016/j.cbi.2022.110269.
[33] M.W. Gardner, S.R. Dorling, Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences, Atmos. Environ. 32 (14–15) (1998) 2627–2636, https://doi.org/10.1016/S1352-2310(97)00447-0.
[34] A. Rakhimbekova, T.N. Akhmetshin, G.I. Minibaeva, R.I. Nugmanov, T.R. Gimadiev, T.I. Madzhidov, I.I. Baskin, A. Varnek, Cross-validation strategies in QSPR modelling of chemical reactions, SAR QSAR Environ. Res. 32 (3) (2021) 207–219, https://doi.org/10.1080/1062936X.2021.1883107.
[35] K. Roy, I. Mitra, On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design, Comb. Chem. High. Thro. Screen. 14 (6) (2011) 450–474, https://doi.org/10.2174/138620711795767893.
[36] C. Rücker, G. Rücker, M. Meringer, y-Randomization and its variants in QSPR/QSAR, J. Chem. Inf. Model. 47 (6) (2007) 2345–2357, https://doi.org/10.1021/ci700157b.
[37] OECD, Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models, OECD Series on Testing and Assessment, OECD, 2014. https://doi:10.1787/9789264085442-en.
[38] K. Roy, S. Kar, P. Ambure, On a simple approach for determining applicability domain of QSAR models, Chemometr. Intell. Lab. Syst. 145 (2015) 22–29, https://doi.org/10.1016/j.chemolab.2015.04.013.
[39] A. Golbraikh, A. Tropsha, Beware of q2, J. Mol. Graph. Model. 20 (4) (2002) 269–276, https://doi.org/10.1016/S1093-3263(01)00123-1.
[40] J. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, L. Steinhauer, V. Steinhauer, Chemical information in 3D space, J. Chem. Inf. Comput. Sci. 36 (5) (1996) 1030–1037, https://doi.org/10.1021/ci960343+.
[41] M.C. Hemmer, V. Steinhauer, J. Gasteiger, Deriving the 3D structure of organic molecules from their infrared spectra, Vib. Spectrosc. 19 (1) (1999) 151–164, https://doi.org/10.1016/S0924-2031(99)00014-4.
[42] E. Estrada, E. Molina, 3D connectivity indices in QSPR/QSAR studies, J. Chem. Inf. Comput. Sci. 41 (3) (2001) 791–797, https://doi.org/10.1021/ci000156i.
[43] R. Todeschini, V. Consonni, P. Gramatica, Chemometrics in QSAR, in: Comprehensive Chemometrics, vol. 4, Elsevier, 2009, pp. 129–172, https://doi.org/10.1016/B978-044452701-1.00007-7.
[44] F. Zhang, Z. Wang, W.J. Peijnenburg, M.G. Vijver, Machine learning-driven QSAR models for predicting the mixture toxicity of nanoparticles, Environ. Int. 177 (2023) 108025, https://doi.org/10.1016/j.envint.2023.108025.
[45] B.S. Luka, T.K. Yuguda, M. Adnouni, R. Zakka, I.B. Abdulhamid, B.G. Gargea, Drying temperature-dependent profile of bioactive compounds and prediction of antioxidant capacity of cashew apple pomace using coupled Gaussian Process Regression and Support Vector Regression (GPR–SVR) model, Heliyon 8 (9) (2022) e10461, https://doi.org/10.1016/j.heliyon.2022.e10461.
[46] M. Kianpour, E. Mohammadinasab, T.M. Isfahani, Prediction of oral acute toxicity of organophosphates using QSAR methods, Curr. Comput. Aided Drug Des. 17 (1) (2021) 38–56, https://doi.org/10.2174/1573409916666191227093237.
[47] X. Yao, H. Liu, R. Zhang, M. Liu, Z. Hu, A. Panaye, J.P. Doucet, B. Fan, QSAR and classification study of 1, 4-dihydropyridine calcium channel antagonists based on least squares support vector machines, Mol. Pharm. 2 (5) (2005) 348–356, https://doi.org/10.1021/mp050027v.
[48] C. Jiang, P. Zhao, W. Li, Y. Tang, G. Liu, In silico prediction of chemical neurotoxicity using machine learning, Toxicol. Res. 9 (3) (2020) 164–172, https://doi.org/10.1093/toxres/tfaa016.

[49] Z. Wu, M. Zhu, Y. Kang, E.L.H. Leung, T. Lei, C. Shen, D. Jiang, Z. Wang, D. Cao, T. Hou, Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets, Brief. Bioinform. 22 (4) (2021) bbaa321, https://doi.org/10.1093/bib/bbaa321.

[50] M. Kianpour, E. Mohammadinasab, T.M. Isfahani, Comparison between genetic algorithm-multiple linear regression and back-propagation-artificial neural network methods for predicting the LD50 of organo (phosphate and thiophosphate) compounds, J. Chin. Chem. Soc. 67 (8) (2020) 1356–1366, https://doi.org/10.1002/jccs.201900514.

[51] J. Cui, B. Ren, Y. Tong, H. Dai, L. Zhang, Synergistic combinations of antifungals and anti-virulence agents to fight against Candida albicans, Virulence 6 (4) (2015) 362–371, https://doi.org/10.1080/21505594.2015.1039885.

[52] A. Schoeneberg, M. Hu, Efficacy evaluation of demethylation inhibitors and mixtures against *Colletotrichum spp*. causing strawberry anthracnose, J. Plant Pathol. 104 (4) (2022) 1483–1489, https://doi.org/10.1007/s42161-022-01191-2.