JOTCS A
Journal of the Turkish Chemical Society
Section: A
Chemistry
TURKISH CHEMICAL SOCIETY

QSER Modeling of Half-Wave Oxidation Potential of Indolizines by Theoretical Descriptors

Nabil Bouarra[1,2*], Nawel Nadji[1,2], Soumaya Kherouf[3], Loubna Nouri[1,4], Amel Boudjemaa[1], Khaldoun Bachari[1], Djelloul Messadi[3]

[1]Scientific and Technical Research Center in Physico-Chemical Analysis, Industrial Zone, Tipaza, 42004, Algeria.

[2]Badji Mokhtar University, Laboratory of Environmental Engineering, Annaba, 23000, Algeria.

[3]Badji Mokhtar University, Department of Chemistry, Annaba, 23000, Algeria.

[4]University of Sciences and Technology Houari – Boumediene, Laboratory of Reaction Engineering, Algiers, 16111, Algeria.

Abstract: Indolizine derivatives hold essential biological functions and have been researched for hypoglycemic, antibacterial, anti-inflammatory, analgesic, and anti-tumor actions. Indolizine scaffold has intrigued conjecture and continuous attention and has become an effective parent system for generating powerful novel medication candidates. This research focused on applying the quantitative structure-electrochemistry relationship (QSER) approach to the half-wave potential ($E_{1/2}$) for Indolizine derivatives using theoretical molecular descriptors. After calculating the descriptors and splitting the data into both sets, training and prediction. The QSER model was constructed using the Genetic Algorithm/Multiple Linear Regression (GA/MLR) technique, which was used to choose the optimal descriptors for the model. A four-parameter model has been established. Many assessment procedures, including cross-validation, external validation, and Y-scrambling testing, were used to assess the model's performance. Furthermore, the applicability domain (AD) was investigated using the Williams and Insubria graphs to assess the correctness of the established model's predictions. The constructed model exhibits great goodness-of-fit to experimental data, as well as high stability (R^2=0.893, Q^2_{LOO}= 0.851, Q^2_{LMO}=0.843 $RMSE_{tr}$= 0.052, s= 0.056). Prediction results show a good agreement with the experimental data of $E_{1/2}$ (R^2ext= 0.912, Q^2_{F1}= 0.883, Q^2_{F2}= 0.883, Q^2_{F3}= 0.919, CCC_{ext}= 0.942, $RMSE_{ext}$=0.045).

Keywords: QSER, cyclic voltammetry, indolizines, molecular descriptors, MLR.

Submitted: February 07, 2022. **Accepted:** April 11, 2022.

Cite this: Bouarra N, Nadji N, Kherouf S, Nouri L, Boudjemaa A, Bachari, et al. QSER Modeling of Half-Wave Oxidation Potential of Indolizines by Theoretical Descriptors. JOTCSA. 2022;9(3):709–20.

DOI: https://doi.org/10.18596/jotcsa.1065043.

***Corresponding author. E-mail:** bouarranabil@yahoo.com.

INTRODUCTION

Indolizine is a heteroaromatic molecule composed of two condensed rings (five and six members) and a bridging nitrogen atom (1). Indolizine has been referred to by various names in the literature, including pyrindole, pyrrodine, pyrrolo[1,2-a]pyridine, and pyrrocoline (2). Indolizines (indolizine derivatives) are heterocyclic compounds comprised of indolizine heterocyclic nuclei. Indolizidines are widely dispersed in nature, particularly in plants, but aromatic indolizines are uncommon (2). Heterocycles possessing indolizine cores play an essential role in pharmaceutical and

materials chemistry. Several high-performance materials, dyes, and medicines are intrinsically heterocyclic (3). Numerous pharmacological activities have been reported for indolizines, like anti-inflammatory activity (4), antiviral activity (5), aromatase inhibitory activity (6), analgesic activity (7), and anticancer activity (8,9). Indolizine scaffold has intrigued conjecture and continuous attention and has become a significant parent system for the generation of novel medication candidates (9).

Electrochemical techniques are helpful tools for studying electron-transfer processes and may also provide important information that can contribute to understanding various biological phenomena (10). The oxidation reaction is the most commonly seen route during the beginning phase of drug biotransformation. Because of this, electrochemistry is often used as a simulation technique in drug metabolism investigations (11). Organic compounds' half-wave oxidation potential ($E_{1/2}$) is a significant electrochemical feature, which is a constant that defines an oxidation-reduction system according to the definition. The $E_{1/2}$ may be used to predict the electrochemical properties of other organic molecules as well (12). A typical electrochemical technique used in studies of electro-oxidation systems is voltammetry (13). Since the synthesis and the evaluation of novel drugs based on indolizines and their investigation by voltammetric methods are restricted in time and cost (14), the construction of theoretical models to predict the features of these compounds is essential and required.

The quantitative structure-property relationship (QSPR) approach, referred to as the quantitative structure-electrochemistry relationship (QSER), allows for the prediction and interpretation of $E_{1/2}$ of drugs and organic compounds, based on the relationship between both their $E_{1/2}$ and structural molecular descriptors. These descriptors contain chemical information related to the molecule's physicochemical features (15). QSER models are suitable because they minimize the number of experiments by saving time and money while measuring physicochemical or bio-activities. Several studies on the use of QSPR in electrochemistry have been conducted (16-20). Hemmateenejad and Shamsipur used PCR and PC-ANN to determine the $E_{1/2}$ of 69 organic compounds. They developed an ideal PC-ANN model that can explain 96% of the $E_{1/2}$ data variances (16). Nesmerak *et al*. relate Hammet substituent constants and HOMO orbital energy to $E_{1/2}$ of 40 benzoxazines. They discovered a significant relationship between HOMO and E1/2 oxidation. (17). Fatemi et al. constructed a QSPR model based on multiple linear regression to predict $E_{1/2}$ values of 15 substituted nitrobenzenes (18). Hemmateenejad and Yazdani used MLR and PCR to investigate the half-wave reduction potential ($E_{1/2}$) of 40 steroids (19). Goudarzi *et al*. (20) used a genetic algorithm-partial least squares (GA-PLS) and stepwise regression-partial least squares (SR-PLS) approach to estimate the half-wave reduction potentials of 21 chlorinated organic compounds.

In this work, we have attempted to develop a new QSER model by predicting the half-wave oxidation potential of different sets of indolizines. Our purposes are:

1) To investigate the relationship between the half-wave oxidation potentials of indolizines and their molecular structures;

2) To build a precise and stable model with great predictive potential using a rapid and straightforward method of regression;

3) To predict $E_{1/2}$ values for different indolizines without experimental data using the established model.

MATERIAL AND METHODS

Dataset

Fifty-two structurally diverse indolizines were selected for data; their molecular structures are described in Table S1 in supplementary materials. Experimental $E_{1/2}$ values were obtained from the literature (21). Recorded values vary from 0.362 to 0.966 Volts. Table 1 summarizes the data obtained for indolizine derivatives by cyclic voltammetry (CV). The cyclic voltammograms were recorded according to these experimental conditions: A platinum disk electrode (d=1.0 mm). acetonitrile solutions (1mM) of the substrate containing 0.1 M TBATFB as the supporting electrolyte, and all measurements were performed at 20 °C and at 1 V/s scan rate (21).

Table 1: Cyclic voltammetry data and descriptors values for the studied compounds.

Nº.	$E_{1/2}$ (V) b	T(O..O)	SIC4	R8m	TPSA(NO)
1	0.386	0	0.869	0.223	37.12
2	0.429	0	0.869	0.378	37.12
3	0.461	0	0.869	0.503	37.12
4	0.385	0	0.83	0.251	37.12
5	0.362	26	0.802	0.274	55.58
6	0.435	0	0.873	0.379	37.12
7	0.446	6	0.878	0.399	46.35

Nº.	$E_{1/2}$ (V) b	T(O..O)	SIC4	R8m	TPSA(NO)
8	0.647	5	0.903	0.383	46.35
9	0.671	4	0.903	0.418	46.35
10	0.443	0	0.882	0.458	37.12
11	0.522	0	0.882	0.442	37.12
12	0.45	0	0.91	0.474	37.12
13	0.436	0	0.875	0.346	37.12
14	0.391	0	0.856	0.421	37.12
15	0.492	0	0.878	0.403	46.35
16	0.443	20	0.819	0.343	64.81
17	0.676	8	0.896	0.462	63.42
18	0.688	10	0.907	0.394	63.42
19	0.679	12	0.882	0.386	63.42
20	0.692	0	0.877	0.347	54.19
21	0.77	0	0.875	0.473	71.26
22	0.807	0	0.875	0.594	71.26
23	0.825	0	0.875	0.758	71.26
24	0.74	0	0.816	0.475	71.26
25	0.68	64	0.848	0.486	89.72
26	0.792	0	0.872	0.423	71.26
27	0.773	0	0.879	0.393	71.26
28	0.966	0	0.872	0.729	71.26
29	0.743	0	0.842	0.49	74.5
30	0.776	0	0.875	0.486	71.26
31	0.815	0	0.882	0.525	71.26
32	0.804	18	0.878	0.473	80.49
33	0.477	0	0.872	0.35	47.47
34	0.688	22	0.872	0.485	64.54
35	0.711	22	0.878	0.5	64.54
36	0.411	0	0.816	0.464	47.47
37	0.686	0	0.889	0.248	71.26
38	0.683	0	0.858	0.356	71.26
39	0.791	36	0.875	0.454	89.72
40	0.772	32	0.88	0.485	89.72
41	0.754	34	0.88	0.481	89.72
42	0.782	0	0.901	0.514	71.26
43	0.788	0	0.879	0.457	71.26
44	0.78	0	0.847	0.491	71.26
45	0.809	0	0.903	0.623	71.26
46	0.79	0	0.903	0.615	71.26
47	0.722	0	0.869	0.391	60.36
48	0.669	0	0.869	0.38	60.36
49	0.606	0	0.88	0.429	48.12
50	0.671	0	0.889	0.566	48.12

Nº.	$E_{1/2}$ (V) b	T(O..O)	SIC4	R8m	TPSA(NO)
51	0.698	0	0.889	0.691	48.12
52	0.601	0	0.857	0.419	48.12

Generation of Descriptors

ChemDraw 7.0 software (22) was used to sketch the chemical structures of all molecules.

The three-dimensional geometries were optimized using the semi-empirical PM7 method (23) and the MOPAC software (24) to reach the low-energy conformation for each chemical compound. After the geometric optimization, the Dragon software (V.5.5) was used to generate more than 3000 descriptors (25) from different families, including topological descriptors, molecular counts, connection indices, information indices, 2D autocorrelations, edge adjacency indices, topological charge indices, and eigenvalues-based indices, among the molecular descriptors generated. Constant or almost constant descriptor values and descriptors that were found highly correlated (r >0.95) (26) were omitted to minimize repetitive and unnecessary information.

GA-MLR procedure

The obtained descriptors and experimental $E_{1/2}$ values were analyzed using a genetic algorithm-multivariate linear regression (GA-MLR). GA (27,28) is done to explore the feature space and choose the main descriptors related to the compounds' activities or properties ($E_{1/2}$ in this study). Briefly, the GA is built up of the following fundamental phases: 1) a vector (chromosome) comprising zeros and ones (genes) is produced with the size corresponding to the number of factors; 2) a population of chromosomes is randomly generated; 3) the value of fitness function is examined for every new created chromosomes(The fitness function here is the cross-validation coefficient (Q^2_{LOO})); 4) the chromosomes with the better predictions (according to their fitness function value) are then used to generate new populations by operations including selection, crossover and mutation. These phases of evolution continue until the halting criteria are fulfilled. After that, the MLR is used to associate the descriptors chosen by GA with the values of $E_{1/2}$. The MLR provides an equation relating the structural descriptors to the $E_{1/2}$:

$$E_{1/2} = b_0 + b_1y_1 +...+ b_ny_n \quad (1)$$

Where the intercept (b_0) and the regression coefficients of the descriptors (b_i) are calculated using the least-squares method. y_i is the independent variable or descriptor.

Validation of QSER model

Following the Organization for Economic Cooperation and Development (OECD) guidelines, a quantitative structure-activity relationship (QSAR) model should give acceptable metrics of quality, robustness, and reliability. Whereas a training set provides the model's internal performance, reliability is evaluated using a suitable test set (29).

The following statistical metrics (R^2 and Q^2_{LOO}) were calculated to verify the model's accuracy. R^2 evaluates the model's fit to the observed data in the training set. In other words, R^2 governs the fit of the build model. The cross-validation coefficient (Q^2_{LOO}), one of the most frequent internal validation procedures, was calculated for the quantitative assessment of model robustness. This procedure was repeated for the full training set by eliminating one molecule and developing and verifying each molecule's model (29, 30).

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (2)$$

$$Q^2_{LOO} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_{i/i})^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \quad (3)$$

Where y_i is the experimental $E_{1/2}$, $\hat{y}_i$ is the value of $E_{1/2}$ calculated by the model equation, $\bar{y}$ is the average value of $E_{1/2}$ for the whole set, n is the total compounds in the training set, and $\hat{y}_{i/i}$ is the value of $E_{1/2}$ predicted by the generated model according to the LOO method.

Internal validation using leave-many-out (LMO) is an effective method. In theory, LMO model validation employs fewer training sets than the LOO procedure. The LOO (leave-one-out) procedure employs n training sets of n-1 objects in and predicts each excluded object in the test set, which may be performed several times owing to the possibility of more combinations leaving several compounds out of the training set. It can reasonably be inferred that the model obtained is stable if there is an excellent average QSPR model in Q^2_{LOO} validation (31). In this work, 30% of the compounds were separated from the training set randomly.

External validation was used to assess the developed model's prediction performance based on a series of coefficients: R^2_{ext} (which describes the correlation inside the validation set between both predicted and experimental values), Q^2_{F1}(32), Q^2_{F2}(33), Q^2_{F3} (34,35) and the Concordance Correlation Coefficient (CCC) (36-38). This last one verifies the tiniest variation in predictions between experimental and external data. Moreover, the root means squared error (RMSE), which recapitulates the total error of the developed model, measures and compares the reliability of predictions in both the training ($RMSE_{tr}$) and the prediction set ($RMSE_{ext}$), defined as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (4)$$

One of the most commonly used strategies for ensuring the accuracy and robustness of the created model is Y-scrambling. It is not unusual for a model with good statistical results for training to have fortuitous correlations but descriptors that do not necessarily relate to the modeled property. The Y-scrambling procedure detects these random models. The experimental property of the training set is randomly mixed, and the learning algorithm is retrained to obtain a model using the same descriptors. Typically, the resulting models should have poor efficiency (30).

RESULTS AND DISCUSSION

Development of the Model

The experimental data of the $E_{1/2}$ were randomly split into two subsets, namely training set (70%) and prediction set (30%). The QSARINS software used the hybrid Genetic Algorithm-multiple linear regression (GA-MLR) approach on the training set to build numerous linear models (39). The set of parameters were used in QSARINS, including the population size of 1000, the generation per size of 1000, the number of models per size of 100, the mutation rate of 80, the crossover rate of 0.6, and the QUIK rule of 0.05. As a result, different models of many sizes have been generated based on the statistics on the cross-validation (Q^2_{LOO}), multiple correlation coefficients (R^2), and standard error (s). However, some of them may be over-fitted.

Figure 1 depicts the effects of the number of descriptors on R^2 and Q^2_{LOO} statistics. As seen in fig.1, models containing 5 and 6 descriptors do not significantly improve model statistics.

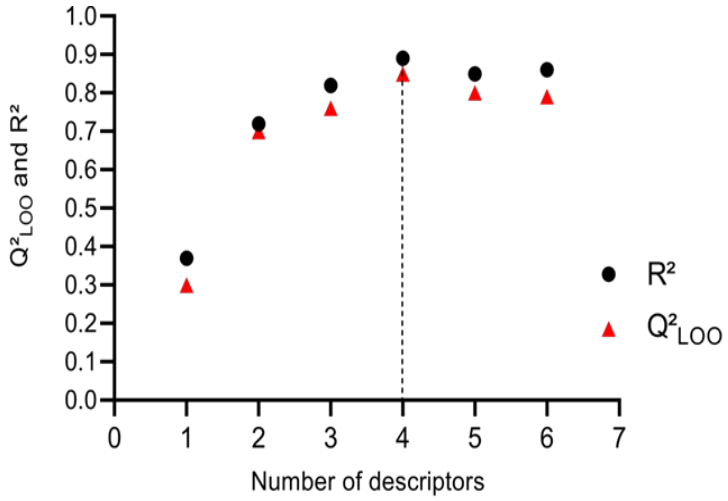

Figure 1: The plot of Q^2_{LOO} and R^2 for the obtained models *v*ersus the number of descriptors.

Based on figure 1, the model of 4-parameters should be chosen as the best model defined by the following equation:

$$\boldsymbol{E_{1/2}}\ (V) = -1.28 - 0.003\ \boldsymbol{T\ (O..O)} + 1.46\ \boldsymbol{SIC4} + 0.327\ \boldsymbol{R8m} + 0.008\ \boldsymbol{TPSA\ (NO)} \quad (5)$$

R^2= 0.893, Q^2_{LOO}= 0.851, Q^2_{LMO}= 0.847, $RMSE_{cv}$= 0.061, $RMSE_{tr}$= 0.052, CCC_{tr}= 0.943, $RMSE_{ext}$= 0.045, R^2_{ext}= 0.912, Q^2_{F1}= 0.883, Q^2_{F2}= 0.883, Q^2_{F3}= 0.919, CCC_{ext}= 0.942, s = 0.0581, F= 66.727.

The statistics prove the constructed model's stability, robustness, and predictive ability (Equation 5). Therefore, the model was accepted with values of R^2 and CCC_{tr} above 0.7 and 0.85, respectively. Moreover, this model has the smallest values for $RMSE_{tr}$ and the highest values for CCC_{tr}, indicating that this model has the lowest error, i.e., the minor differences from the predicted data. In addition, the model's Q^2_{LOO} and Q^2_{LMO} values are more significant than 0.6 and close to R^2. Furthermore, the established model has the lowest RMSEcv values, proving its efficiency. The built model's external validation results showed a high predictive ability because the R^2_{ext} and the CCC_{ext} values are more significant than 0.7 and 0.85, respectively. Other external validation metrics (Q^2_{F1}, Q^2_{F2}, and Q^2_{F3}) show that they accepted the model based on the literature-recommended criteria (38).

The correlation matrix presented in table 2 was examined to ensure that these descriptors are independent. It is clear from table 2 that none of the descriptor pairs has a significant correlation. In addition, to check the multicollinearity of the descriptors concerned, the Variance Inflation Factor (*VIF*) was calculated (40). From table 2, the *VIF* values are less than 5.0, indicating that the selected descriptors are not collinear. As a result, the model that has been established is reliable.

Table 2: The Correlation matrix.

Descriptors	Definition	T(O..O)	SIC4	R8m	TPSA(NO)	VIF
T(O..O)	Sum of topological distances between O..O	1				**1.653**
SIC4	Structural Information Content index (neighborhood symmetry of 4-order)	-0.101	1			**1.065**
R8m	R autocorrelation of lag 8 / weighted by mass	0.051	0.217	1		**1.310**
TPSA(NO)	Topological polar surface area using N, O, S and P polar contributions	0.589	0.05	0.389	1	**1.936**

Predicted data versus experimental ones and radar plots for training and prediction sets are given in figure 2. The experimental and predicted values are fairly similar, as illustrated in Fig.2 (left). This model matches the experimental data well (R^2 = 0.883, $RMSE_{tr}$ = 0.052, for the training set and R^2_{ext} = 0.912, $RMSE_{ext}$ = 0.045, for the prediction set). The difference between the experimental and predicted $E_{1/2}$ in training and prediction sets may be explained by the degree of overlap between the experimental and predicted $E_{1/2}$ lines in the radar plot (Fig. 2. Right). The radar plot shows a good overlap between experimental and predicted data.

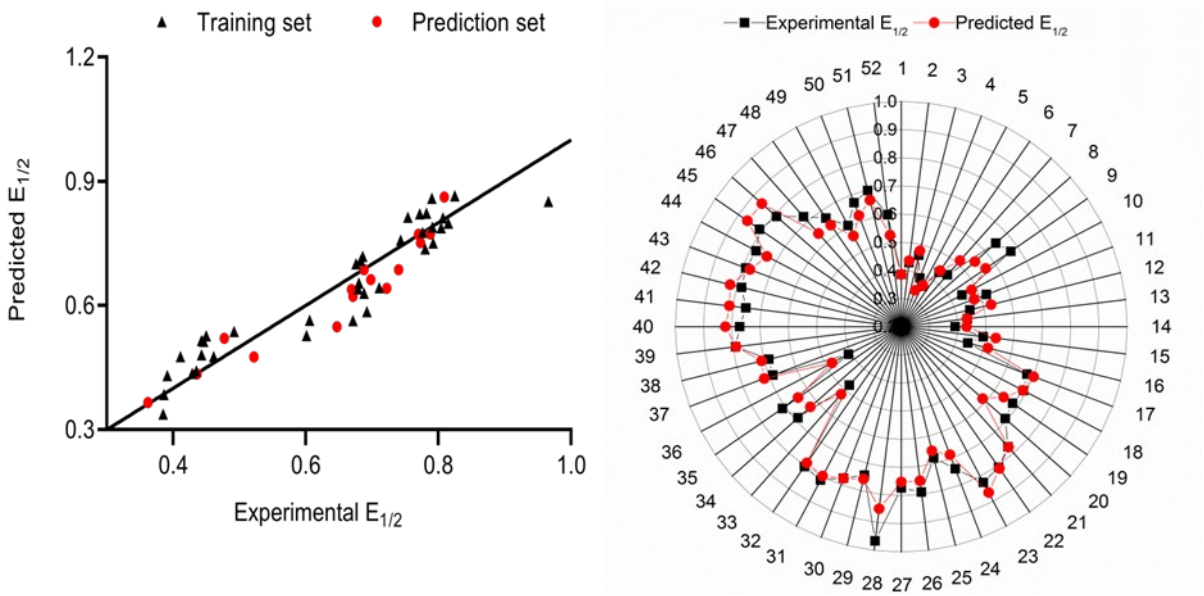

Figure 2: Predicted versus experimental values of E1/2 **(Left)**. The radar plot of QSER model **(Right)**.

The residuals of the training and prediction data sets are represented in Figure 3. As seen in Figure 3, all residuals are scattered consistently and randomly on both sides of the zero line. Consequently, the developed model has no systematic errors.

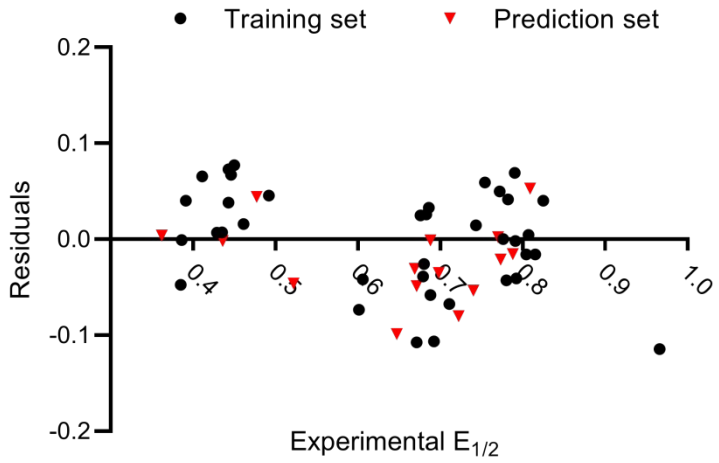

Figure 3: The residuals *vs.* training and prediction values.

The randomization test was applied to avoid correlations by chance and to validate the developed model. By generating two hundred models, the results (R^2Ysc and Q^2Ysc as a function of Kxy) are reproduced in Figure 4, where K_{xy} is the overall correlation in the model descriptors (including $E_{1/2}$). The low R^2Ysc and Q^2Ysc (0.4) values indicate that the favorable results of the created model were not related to random correlations.

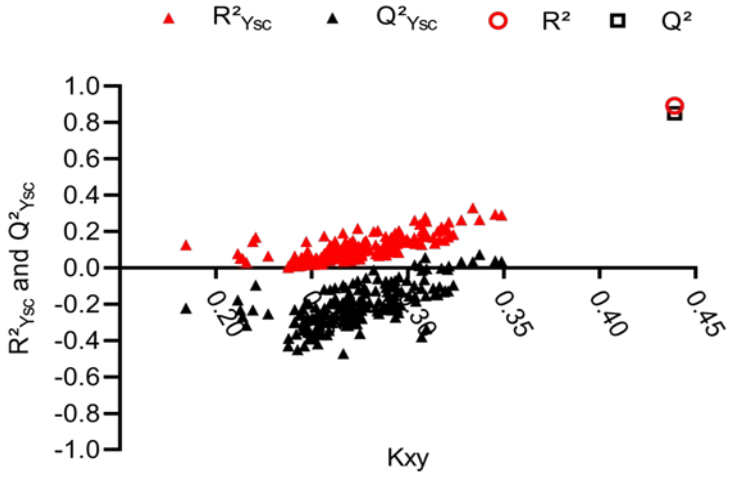

Figure 4: Randomization test.

The applicability domain (AD) belongs to the model validation technique, also known as the model prediction space. The AD was checked to: 1) determine the correct zone for model predictions such that predictions in this zone are reliable and 2) use this AD for making predictions of new compounds. The Williams plot is a standard graphic representation of the standardized residuals vs. leverages values (h_i). The details of this concept are defined in the literature (30,41). Suppose a compound has high leverage (> h^*), this compound will be out of AD. In general, h^* equals 3(p+1)/n, where p is the model's size and n is the number of training compounds. The high standardized residuals (>|3SD| units) (26) are another criterion that places a chemical out of AD.

Figure 5 presents the standardized prediction errors as a function of the values of the leverages (h_i). As shown in figure 4, the presence of an influential point (Compound #25) of the training set, h^* equal 0.405. Thus, this compound can have a positive leverage effect on the stability of the model and make it more accurate. We also note that all residuals are in the range (± 3 SD) (horizontal lines); this denotes that the developed model has an excellent predictive capacity.

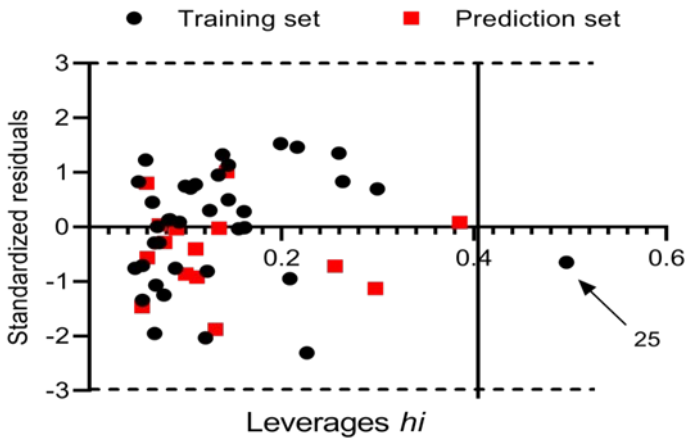

Figure 5: Williams plot of the developed model.

Furthermore, the applicability domain was checked using the Insubria graph (40), which plots leverage values vs. predicted values for compounds with no experimental data. This is useful for visualizing the proposed model's interpolated and extrapolated predictions for novel chemicals without experimental data (51 compounds with no data in this work). The minimum and maximum values of the experimental $E_{1/2}$ of the training set are always presented in the graph, a zone of higher reliability, where predictive ability is good, for both structures less than h^* and $E_{1/2}$ predictions placed within Y_{min} and Y_{max}. In addition, chemical predictions are extrapolated and may be less accurate if their leverages are $h_i > h^*$ (outside the structural domain of the training set).

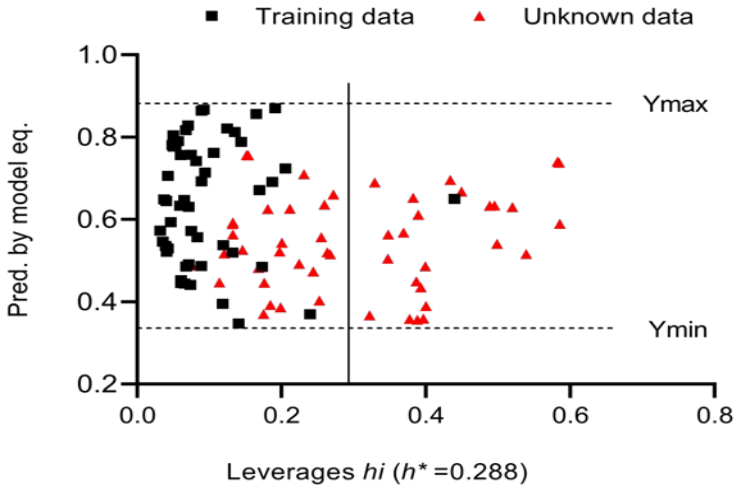

Figure 6: Insubria graph for the developed model: Leverages h_i *vs*. predicted $E_{1/2}$.

The Insubria graph of the developed model is reported in Figure 6. As can be seen in this figure, the predictions for 53% of indolizines from the prediction set were located within the model's AD, suggesting that this model reliably interpolated the $E_{1/2}$ predictions. Otherwise, 47% of these compounds are outside the AD ($h_i > h^*$), meaning that the predictions obtained are less reliable since they are extrapolations from the structural domain of the model. It demonstrates that, with a few exceptions, the model developed in this study can make reliable predictions for structurally similar indolizines.

Interpretation of descriptors

The following is an order of decreasing descriptor significance in the model:
TPSA(NO) (39.7013%) >T(O..O) (22.9569%) > R8m (19.6070%) > SIC4 (17.7348%).

TPS(NO) (Topological Polar Surface Area defined by nitrogen and oxygen contributions) is the most crucial descriptor in the constructed model. It belongs to the molecular properties descriptor family, expressed as a polar part of the molecule linked to oxygen, nitrogen, sulfur atoms, and hydrogen connected to these heteroatoms and specific charge interactions (15,43). Based on the summation of the tabular polar fragment surface contributions, the topological polar surface area may

be calculated directly (i.e., atoms regarding their bonding pattern). TPSA(NO) has demonstrated a strong correlation (0.826) with half-wave potential in the developed model. The positive TPSA(NO) coefficient suggests that as the topological polar surface area increases, the $E_{1/2}$ increases as well.

The second important descriptor is T(O..O) (Sum of the topological distance between (O..O), which belongs to the topological descriptors (15,43). The topological distance is the length (i.e., the number of involved bonds) of the shortest route between two atoms. The equivalent row sum of the distance matrix, which is the sum of the topological distance between the atom and every other atom, is the atom's distance degree. The sum of topological distances between (O..O) is obtained by the sum of topological distances between all pairs of (O..O) (15,43). The negative coefficient of T(O..O) in equation 5 demonstrates an inverse relationship with $E_{1/2}$, implying that decreasing the descriptor value increases $E_{1/2}$.

The third important descriptor was R8m (the R autocorrelation of lag 8/weighted by atomic masses) (15,43), which belongs to GETAWAY descriptors. The GETAWAY descriptors are built on the leverage matrix, the most widely used for regression diagnostics and the same one determined in statistics (42). Via the use of the molecular influence matrix, atomic connections by molecular topology, and chemical information, these molecular descriptors attempt to match 3D Molecular Geometry through the use of distinct atomic weights and the use of the molecular influence matrix (atomic mass, polarizability, van der Waals volume and electronegativity, etc.) (15,43). The SIC4 descriptor has a positive sign, indicating that the $E_{1/2}$ is associated with this descriptor.

The structural information content (neighborhood of symmetry of 4-order) (SIC4) (15,43) was the last important descriptor. This descriptor is among the information indices determined based on neighbor degrees and edge multiplicity for the H-included molecular graph (15,43). The information indices can be used as a quantitative indicator of structural homogeneity or graph diversity. Therefore, it is related to the symmetry associated with the structure. The positive coefficient of SIC4 implies that the $E_{1/2}$ may grow as SIC4 increases.

According to our findings, $E_{1/2}$ of indolizines is mainly determined by combining the descriptors mentioned above: molecular properties, topological properties, GETAWAY property descriptors, and information index descriptors.

CONCLUSION

The available evidence in the literature suggests that indolizines comprise a substantial class of heterocycles that possess various intriguing biological actions, including hypoglycemic, antibacterial, analgesic, and anti-inflammatory activities. The half-wave potential ($E_{1/2}$) is an essential electrochemical characteristic. It is a valuable metric for determining the antioxidant activity of organic compounds. The primary goal of this work is to build a quantitative structure-electrochemistry relationship model that could be used to predict the oxidation half-wave potential ($E_{1/2}$) of a series of indolizines using the GA-MLR approach. The developed model has four descriptors derived from the structures of the chemical compounds. The proposed model has a high statistical significance. The $E_{1/2}$ predictions have a good match with the experimental values. Furthermore, the leverage approach assessed the model's applicability domain. The developed model can correctly predict the $E_{1/2}$ for novel compounds that are structurally similar to indolizines, as well as other existing indolizines with undefined $E_{1/2}$ experimental values.

CONFLICT OF 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.

ACKNOWLEDGMENTS

We thank Prof. Paola Gramatica for the free license of QSARINS. We are thankful to the Algerian Directorate-General for Scientific Research and Technological Development (DGRSDT) for providing financial assistance for this research.

REFERENCES

1. Georgescu E, Dumitrascu F, Georgescu F, Draghici C, Barbu L. A Novel Approach for the Synthesis of 5-Pyridylindolizine Derivatives via 2-(2-Pyridyl) pyridinium Ylides. Journal of Heterocyclic Chemistry. 2013;50(1):78-82. <DOI>.

2. Borrows E, Holland D. The Chemistry of the Pyrrocolines and the Octahydropyrrocolines. Chemical reviews. 1948;42(3):611-43. <DOI>.

3. Katritzky A R, Rees C W, Scriven E F V, Lohray B B, Bhushan V., Comprehensive Heterocyclic Chemistry II. Pergamon Press;1996 .11628 p. ISBN: 0-08-042072-9.

4. Kitadokoro K, Hagishita S, Sato T, Ohtani M, Miki K. Crystal structure of human secretory phospholipase A2-IIA complex with the potent indolizine inhibitor 120–1032. The Journal of Biochemistry. 1998;123(4):619-23. <DOI>.

5. De Bolle L, Andrei G, Snoeck R, Zhang Y, Van Lommel A, Otto M, et al. Potent, selective and cell-mediated inhibition of human herpesvirus 6 at an

early stage of viral replication by the non-nucleoside compound CMV423. Biochemical pharmacology. 2004;67(2):325-36. <DOI>.

6. Sonnet P, Dallemagne P, Guillon J, Engueard C, Stiebing S, Tangue J, Bureau B, Rault S, Auvray P, Moslemi S, Sourdaine P, Séralini G E, New aromatase inhibitors. Synthesis and biological activity of aryl-substituted pyrrolizine and indolizine derivatives, Bioorg Med Chem. 2000;8 (5):945-955. <DOI>.

7. Campagna F, Carotti A, Casini G, Macripo M. Synthesis of new heterocyclic ring systems: indeno [2, 1-b]-benzo [g] indolizine and indeno [1', 2': 5, 4] pyrrolo [2, 1-a] phthalazine. Heterocycles (Sendai). 1990;31(1):97-107. <DOI>.

8. Lillelund VH, Jensen HH, Liang X, Bols M. Recent developments of transition-state analogue glycosidase inhibitors of non-natural product origin. Chemical reviews. 2002;102(2):515-54. <DOI>.

9. Das A, Banik BK. Chapter 5 - Microwave-assisted synthesis of N-heterocycles. In: Das A, Banik B, editors. Microwaves in Chemistry Applications: Elsevier; 2021. p. 143-98. <DOI >.

10. Keyzer H, Eckert GM, Gutmann F. Electropharmacology. CRC Press; 1990. 432 p. ISBN:978-0-8493-5409-0.

11. Eberson L. Electron-Transfer Reactions in Organic Chemistry. In: Gold V, Bethell D, éditeurs. Advances in Physical Organic Chemistry [Internet]. Academic Press; 1982. p. 79-185. <DOI>.

12. Guengerich FP, Willard RJ, Shea JP, Richards LE, Macdonald TL. Mechanism-based inactivation of cytochrome P-450 by heteroatom-substituted cyclopropanes and formation of ring-opened products. Journal of the American Chemical Society. 1984;106(21):6446-7. <DOI>.

13. Scholz F. Electroanalytical Methods: Guide to Experiments and Applications. Springer Science & Business Media; 2009. 366 p. ISBN:978-3-642-02915-8.

14. Macchiarulo A, Costantino G, Fringuelli D, Vecchiarelli A, Schiaffella F, Fringuelli R. 1, 4-Benzothiazine and 1, 4-benzoxazine imidazole derivatives with antifungal activity: a docking study. Bioorganic & medicinal chemistry. 2002;10(11):3415-23. <DOI>.

15. Todeschini R, Consonni V. Handbook of Molecular Descriptors. John Wiley & Sons; 2000. 692 p. <DOI>. ISBN: 9783527613106.

16. Hemmateenejad B, Shamsipur M. Quantitative structure-electrochemistry relationship study of some organic compounds using PC-ANN and PCR. Internet Electronic Journal of Molecular Design. 2004;3(6):316-34. <URL>.

17. Nesmerak K, Nemec I, Sticha M, Waisser K, Palat K. Quantitative structure–property relationships of new benzoxazines and their electrooxidation as a model of metabolic degradation. Electrochimica acta. 2005;50(6):1431-7. <DOI>.

18. Fatemi MH, Hadjmohammadi MR, Kamel K, Biparva P. Quantitative structure–property relationship prediction of the half-wave potential for substituted nitrobenzenes in five nonaqueous solvents. Bulletin of the Chemical Society of Japan. 2007;80(2):303-6. <DOI>.

19. Hemmateenejad B, Yazdani M. QSPR models for half-wave reduction potential of steroids: A comparative study between feature selection and feature extraction from subsets of or entire set of descriptors. Analytica Chimica Acta. 2009;634(1):27-35. <DOI>.

20. Goudarzi N, Goodarzi M, Hosseini MM, Nekooei M. QSPR models for prediction of half wave potentials of some chlorinated organic compounds using SR-PLS and GA-PLS methods. Molecular Physics. 2009;107(17):1739-44. <DOI>.

21. Teklu S, Gundersen L-L, Rise F, Tilset M. Electrochemical studies of biologically active indolizines. Tetrahedron. 2005;61(19):4643-56. <DOI>.

22. ChemDraw Utra “Ultra-chemical structure drawing standard”. Version 7. 2002. Copyright Cambridge Soft Corporation.

23. Stewart JJ. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. Journal of molecular modeling. 2013;19(1):1-32. <DOI>.

24. MOPAC2016, Stewart James J P, Stewart Computational Chemistry, Colorado Springs, CO, USA, <URL> (2016).

25. Todeschini R, Consonni V, Mauri A, Pavan M, DRAGON Software – version 5.4-TALETE srl, (2005).

26. Liu H, Gramatica P. QSAR study of selective ligands for the thyroid hormone receptor β. Bioorganic & medicinal chemistry. 2007;15(15):5251-61. <DOI>.

27. Karakaplan M, Avcu FM. A parallel and non-parallel genetic algorithm for deconvolution of NMR

spectra peaks. Chemometrics and Intelligent Laboratory Systems. 2013;125:147-52. <DOI>.

28. Avcu FM, Karakaplan M. Finding exact number of peaks in broadband UV-Vis spectra using curve fitting method based on evolutionary computing. Journal of the Turkish Chemical Society Section A: Chemistry. 2020;7(1):117-24. <DOI>.

29. Organisation for Economic Co-operation and Development, Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] Models,ENV/JM/MONO (2007) 2, OECD Publishing, Paris. <URL>.

30. Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR & Combinatorial Science. 2003;22(1):69-77. <DOI>.

31. De Lima Ribeiro FA, Ferreira MMC. QSPR models of boiling point, octanol–water partition coefficient and retention time index of polycyclic aromatic hydrocarbons. Journal of Molecular Structure: THEOCHEM. 2003;663(1-3):109-26. <DOI>.

32. Gramatica P. External evaluation of QSAR models, in addition to cross-validation: verification of predictive capability on totally new chemicals. Molecular informatics. 2014;33(4):311-4. <DOI>.

33. Schüürmann G, Ebert R-U, Chen J, Wang B, Kühne R. External validation and prediction employing the predictive squared correlation coefficient-Test set activity mean vs training set activity mean. Journal of Chemical Information and Modeling. 2008;48(11):2140-5. <DOI>.

34. Consonni V, Ballabio D, Todeschini R. Comments on the definition of the Q^2 parameter for QSAR validation. Journal of chemical information and modeling. 2009;49(7):1669-78. <DOI>.

35. Consonni V, Ballabio D, Todeschini R. Evaluation of model predictive ability by external validation techniques. Journal of chemometrics. 2010;24(3-4):194-201. <DOI>.

36. Chirico N, Gramatica P. Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. Journal of chemical information and modeling. 2011;51(9):2320-35. <DOI>.

37. Lawrence I, Lin K. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989:255-68. <DOI>.

38. Chirico N, Gramatica P. Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. Journal of Chemical Information and Modeling. 2012;52(8):2044-58. <DOI>.

39. Gramatica P, Chirico N, Papa E, Cassani S, Kovarich S, QSARINS, Software for the Development and validation of QSAR MLR Models, available on request in <URL>.

40. Kherouf S, Bouarra N, Bouakkadia A, Messadi D. Modeling of linear and nonlinear quantitative structure property relationships of the aqueous solubility of phenol derivatives. Journal of the Serbian Chemical Society. 2019;84(6):575-90. <DOI>.

41. Bouarra N, Nadji N, Nouri L, Boudjemaa A, Bachari K, Messadi D. Predicting retention indices of PAHs in reversed-phase liquid chromatography: A quantitative structure retention relationship approach. Journal of the Serbian Chemical Society. 2021;86(1):63-75. <DOI>.

42. Gramatica P, Cassani S, Roy PP, Kovarich S, Yap CW, Papa E. QSAR modeling is not “push a button and find a correlation”: a case study of toxicity of (benzo-) triazoles on algae. Molecular Informatics. 2012;31(11-12):817-35. <DOI>.

43. Todeschini R, Consonni V. Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references: John Wiley & Sons; 2009. ISBN: 3527628770.

44. Consonni V, Todeschini R, Pavan M. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. Journal of chemical information and computer sciences. 2002;42(3):682-92. <DOI>.