Anal. Bioanal. Electrochem., Vol. 17, No. 9, 2025, 771-788
https://www.doi.org/10.22034/abec.2025.738911

Analytical & Bioanalytical Electrochemistry

Full Paper

QSPR-based Machine Learning Model for the Prediction of Corrosion Inhibition of Mild Steel in an Acid Medium using Quinoxaline Derivatives

J. Saranya,[1,*] Y. Jeevan Nagendra Kumar,[2] V. P. Radha,[3] Y. Ram Lohit Pranay,[4] M. Amulya,[2] M. Bhavana,[2] F. Benhibae,[5] A. Zarrouk[6,*]

[1]*Department of Chemistry, PSG College of Arts and Science, Coimbatore, Tamil Nadu, India*
[2]*Department of Information Technology, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana, India*
[3]*Department of Chemistry, Dr. N.G.P. Institute of Technology, Anna University, Coimbatore-641048, Tamil Nadu, India*
[4]*Department of Computer Science and Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India*
[5]*Higher Institute of Nursing Professions and Health Techniques of Agadir, Annex Guelmim, Agadir, Morocco*
[6]*Laboratory of Spectroscopy, Molecular Modeling, Materials, Nanomaterials, Water and Environment LS3MN2E, CERNE2D, Faculty of Sciences, Mohammed V University in Rabat, Morocco*

*Corresponding Author, Tel.: +919894774760 (Jagadeesan Saranya); +212665201397 (Abdelkader Zarrouk)
E-Mail: jcsaranya.chem96@gmail.com (Jagadeesan Saranya); azarrouk@gmail.com (Abdelkader Zarrouk)

Received: 11 July 2025 / Received in revised form: 27 October 2025 / Accepted: 23 November 2025 / Published online: 31 December 2025

Abstract- Corrosion inhibitors are widely used to reduce metal degradation, although toxicity and experimental screening costs remain important concerns. In this study, a quantitative structure–property relationship (QSPR) machine-learning framework was evaluated for predicting the corrosion inhibition efficiency of quinoxaline derivatives for mild steel in 1 N sulfuric acid. The supplied modelling dataset contained 15 compounds and 11 density-functional-theory-derived quantum chemical descriptors. A descriptor-selection procedure based on simulated annealing was used to identify electronegativity (χ), LUMO energy (ELUMO), fraction of electrons transferred (ΔN), and energy gap (ΔE) as the most informative descriptor subset. Seven regression algorithms—linear regression, Lasso, Ridge, Elastic Net, decision tree, random forest, and gradient boosting—were compared. To address predictive validation explicitly, the data were divided reproducibly into 12 training compounds (80%) and

3 independent test compounds (20%) using random state = 42. All preprocessing and model fitting were performed using the training set, and final predictive statistics were calculated only on the held-out test set. The independent test results showed limited generalization because of the very small sample size; the lowest test MAE was obtained by linear regression (MAE = 4.48%, MSE = 21.14, $R^2 = -0.70$). These results indicate that the selected electronic descriptors are mechanistically informative, but a larger chemical dataset is required before the QSPR models can be considered externally validated for routine prediction.

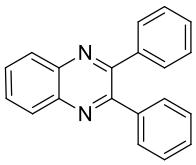
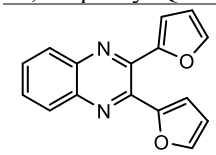
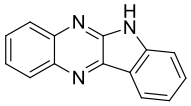
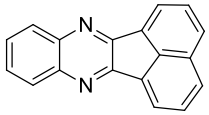
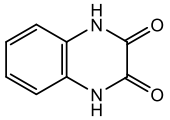
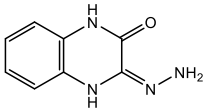
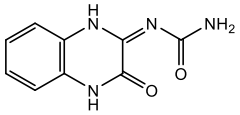
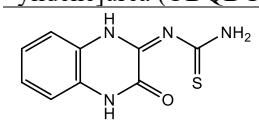
Keywords- Algorithm; Inhibitor; Quantum Chemical Studies; DFT; LR/RF/GB; HOMO/LUMO

1. INTRODUCTION

Mild steel, often referred to as MS, is an iron alloy that is widely used in various sectors, including chemical production, petroleum refining, and thermal power plants, due to its excellent mechanical and electrical properties, as well as its affordability and ease of handling. Additionally, the process of acid pickling, employed to clean metals, has the potential to result in dissolution [1-4]. It is believed that the addition of compounds that limit corrosion is the strategy that has proven to be the most successful in addressing this issue, as this technology is associated with advantages such as high efficiency, practicality, and cost-effectiveness [5-10]. Numerous investigations have been conducted on heterocyclic organic compounds that contain nitrogen, and the findings of these investigations have demonstrated that these compounds are particularly effective in inhibiting corrosion in aggressive media, such as those used in electroplating industries, batteries, and the petroleum sector [11-14]. Most of the literature reports that quinoxaline derivatives were selected as suitable inhibitors for protecting the MS in a solution of sulfuric acid at a concentration of 1 N. These compounds are well-known for their exceptional ability to resist corrosion, as well as their comprehensive range of biological properties, which include antibacterial, antiviral, antifungal, and antioxidant actions [15-19]. Saranya et al. [20] conducted an examination of recent research on quinoxaline derivatives for the prevention of corrosion in various metals. The conclusion was that structural properties, such as nitrogen atoms, aromatic rings, and conjugated double bonds, all contribute to the inhibitory effectiveness of the compound. This conclusion was reached as a result of investigations conducted in both experimental and quantum chemical studies. Several reports have been examined through experimental studies and reported in the scientific literature. These include spontaneous reactions, adsorption mechanisms, types of inhibitors, functional groups, and surface roughness. Quantum chemical studies provide exact information regarding properties such as stability, energy gap, HOMO/LUMO energy levels, electron transport, and Mulliken atomic charges. All of these qualities contribute to determining the effectiveness of the inhibitor [21]. The quantitative structure-property relationship (QSPR) modelling approach helps to scrutinize the suitable chemical compounds. It offers a valuable addition to or substitute for experimental data collection, which is frequently costly and time-consuming [22].

By successfully connecting molecular structure characteristics to their functions, QSPR modeling makes it easier to create novel corrosion inhibitors. Saranya et al. investigated the ability of 13 quinoxaline derivatives to suppress corrosion in mild steel (MS) in 1N sulfuric acid [23–28].

Table 1. Structure and Experimental Inhibition Efficiencies

Structure of the Quinoxaline derivatives	Experimental Inhibition efficiency (%)			Theoretical Efficiency (IE_{Theo} %)
	Weight loss	Polarization	Impedance	
 2,3-diphenyl Quinoxaline (DQ)	85.90	77.73	83.23	75.45
 2,3-di(furan-2-yl)quinoxaline (FQ)	94.62	87.56	92.54	79.23
 6H-indolo[2,3-b]quinoxaline (IQ)	93.63	89.47	91.49	78.69
 Acenaphtho[1,2-b]quinoxaline (AQ)	90.24	85.70	89.78	77.99
 1,4-dihydroquinoxaline-2,3-dione (DQD)	79.97	72.51	78.81	77.46
 (3*E*)-3-hydrazinylidene-3,4-dihydroquinoxalin-2(1*H*)-one (HDQD)	82.50	78.44	81.03	77.23
 1-[(2*E*)-3-oxo-3,4-dihydroquinoxalin-2(1*H*)-ylidene]urea (ODQDU)	86.75	80.44	84.18	77.90
 1-[(2*E*)-3-oxo-3,4-dihydroquinoxalin-2(1*H*)-ylidene]thiourea (ODQDTU)	81.33	81.65	81.63	77.15

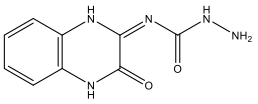
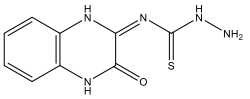
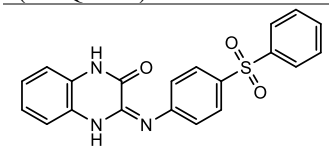
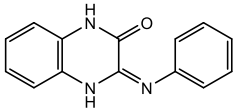
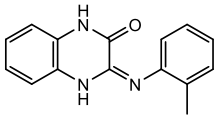
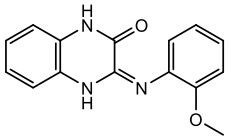
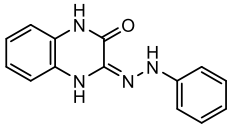
 N-[(2E)-3-oxo-3,4-dihydroquinoxalin-2(1H)-ylidene]hydrazinecarboxamide (ODQDHA)	87.12	82.43	83.95	77.22
 N-[(2E)-3-oxo-3,4-dihydroquinoxalin-2(1H)-ylidene]hydrazinecarbothioamide (ODQDHT)	86.09	83.10	84.94	77.45
 (3*E*)-3-{[4-(phenylsulfonyl)phenyl]imino}3,4-dihydroquinoxalin-2(1*H*)-one (PSDQO)	85.94	82.08	82.13	77.65
 (3*E*)-3-(phenylimino)-3,4-dihydroquinoxalin-2(1*H*)-one (PDQO)	90.91	85.10	84.84	78.13
 (3*E*)-3-[(2-methylphenyl)imino]-3,4-dihydroquinoxalin-2(1*H*)-one (MPDQO)	93.63	87.83	88.62	78.48
 (3*E*)-3-[(2-methoxyphenyl)imino]-3,4-dihydroquinoxalin-(1*H*)-one (MPIDQO)	94.17	90.38	91.51	79.15
 (3*E*)-3-(2-phenylhydrazinylidene)-3,4-dihydroquinoxalin-2(1*H*)-one (PHDQO)	98.69	92.74	92.65	80.15

Table 1 presents the molecular structures of the synthesized quinoxaline compounds and their experimentally determined inhibition efficiencies at a 10 mM concentration, derived from weight loss, impedance, and polarization methods. These compounds can exist in different protonation states, depending on the medium's pH, with fully protonated forms used for computational analysis in highly acidic environments [29,30]. Experimental and quantum chemical studies under different corrosive conditions often demand significant time, labor, and cost resources. To address these drawbacks, machine learning (ML) has emerged as a viable option, leveraging advancements in technology and data science to develop faster, more reliable, and cost-effective solutions [31]. ML is increasingly used in materials discovery and

design, including the identification of potential corrosion inhibitors within materials informatics. Since material properties can be quantitatively associated with their chemical structures, QSPR techniques enable the assessment of material performance while uncovering critical relationships between density functional theory (DFT) parameters and experimental inhibition efficiencies [32,33].

The formulation of QSPR models to evaluate corrosion inhibitor performance has extensively utilized various machine learning algorithms, including artificial neural networks, adaptive neuro-fuzzy inference systems, support vector machines, ordinary least squares regression, multiple linear regression, partial least squares, and autoregressive models with exogenous inputs. Optimization techniques such as genetic algorithms, genetic algorithm-partial least squares, and genetic algorithm-artificial neural networks were also implemented [34]. Ser et al. [35] used GA-PLS and GA-ANN models to predict the effectiveness of pyridine and quinoline derivatives in inhibiting corrosion. The RMSECV and average RMSE values produced by the GA-PLS model were 16.50 and 14.90, respectively. On the other hand, the GA-ANN model provided RMSECV and average RMSE values of 16.70 and 8.80, respectively. Quadri et al. [36] used linear OLS and nonlinear ANN models to predict the performance of 40 quinoxaline derivatives, employing five selected descriptors. The results indicated that the nonlinear ANN model achieved an RMSE of 5.42, demonstrating superior predictive accuracy. Anadebe et al. found that both nonlinear ANN ($R^2 = 0.91$, RMSE = 4.35) and ANFIS models ($R^2 = 1.37$, RMSE = 0.99) exhibited strong performance, with the ANFIS model demonstrating superior predictive capability for the corrosion inhibition efficiencies of 15 expired salbutamol drug molecules [37, 38]. An ARX model was employed to analyze 250 commercial drugs as corrosion inhibitors, resulting in an RMSE of 7.03. Perez et al. highlighted the accuracy of machine learning models in assessing corrosion inhibitors among various chemical compounds [39].

Corrosion studies constitute an important application of machine learning because QSPR models can reveal relationships between molecular descriptors and experimentally measured material properties. The present work therefore develops and critically validates QSPR regression models for quinoxaline corrosion inhibitors. The modelling dataset supplied for this study contains 15 quinoxaline derivatives. Eleven quantum chemical descriptors were considered initially, followed by explicit variable selection and independent test-set validation. Linear regression, Lasso, Ridge, Elastic Net, decision tree, random forest, and gradient boosting were evaluated using the same training and test compounds. The purpose is not only to compare model fit, but also to determine whether the resulting QSPR relationships generalize to compounds that were not used for model fitting.

2. METHODOLOGY

2.1. Quantum chemical parameters (QCP)

Experimental and theoretical information on quinoxaline derivatives in 1 N H2SO4 medium has been reported previously [23–28]. The supplied QSPR modelling dataset contains 15 quinoxaline derivatives and 11 quantum chemical descriptors. The response variable is experimental inhibition efficiency (IE%). The modelling dataset includes DQ, FQ, IQ, AQ, DQD, HDQD, ODQDU, ODQDTU, ODQDHA, ODQDHT, PDQO, MPDQO, MPIDQO, PHDQO, and PSDQO. Table 1 summarizes the compounds for which detailed experimental structure and efficiency information was available in the manuscript source, whereas Table 2 reports the complete 15-compound descriptor matrix used for machine-learning analysis. Figure 1 illustrates the machine-learning workflow.

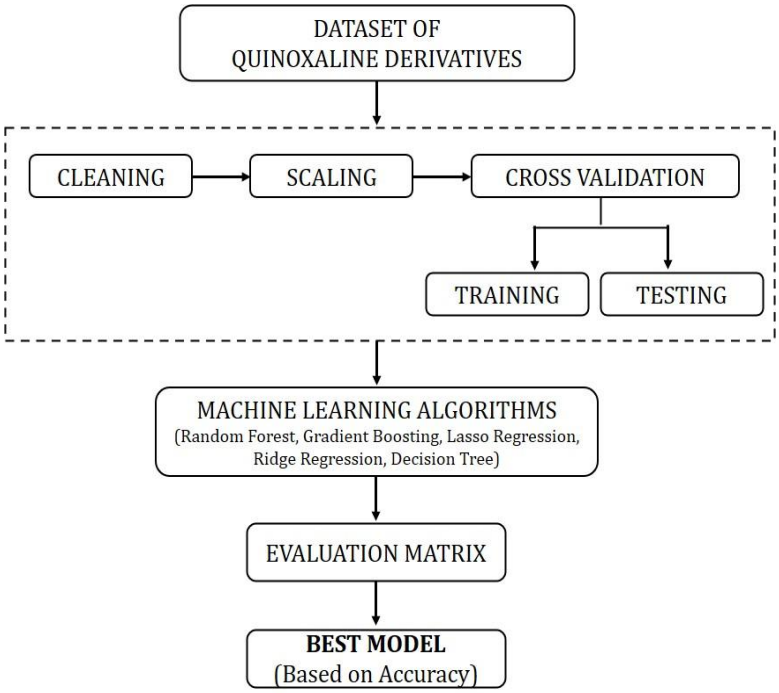


Figure 1. Illustration of Machine learning model

Protonated quinoxaline compounds in an aqueous phase were analyzed through quantum chemical calculations using Gaussian-09 software. The 'B3LYP/6-311G(d,p)' method was employed for these calculations. The solvent effects were considered using the self-consistent reaction field (SCRF) theory, with water as the solvent, and the Tomasi polarised continuum model (PCM) was applied. The key quantum chemical parameters (QCPs) examined are energy gap (ΔE), hardness (η), softness (σ), electronegativity (χ), ionization potential (IP), electron affinity (EA), and the fraction of electrons transferred (ΔN), as reported in the literature [28,33,36]. The output files extracted additional parameters, including E_{HOMO}, E_{LUMO}, dipole moment (μ), and total energy (TE). Table 2 presents the QCPs applied in this study and their respective values.

Table 2. Quantum chemical parameters of the quinoxaline derivatives

QCP	DQ	FQ	IQ	AQ	DQD	HDQD	ODQDU	ODQDTU	ODQDHA	ODQDHT	PDQO	MPDQO	MPIDQO	PHDQO	PSDQO
TE (-)	-880.28	-875.83	-703.36	-801.649	-562.6	-604.05	-717.46	-1040.41	-772.81	-1095.76	-779.8	-894.36	-819.13	-835.13	-1559.55
μ	0.431	1.651	1.91	1.42	6.21	2.05	6.42	7.6	4.28	4.43	4.77	4.25	4.89	3.76	4.11
E_{HOMO}	-6.319	-5.892	-6.057	-6.252	-6.37	-5.03	-6.1	-5.36	-6.16	-6.24	-5.64	-5.26	-5.55	-5.05	-6.24
E_{LUMO}	-2.152	-2.16	-2.181	-2.288	-1.61	-1.23	-1.71	-1.88	-1.91	-2.32	-1.61	-1.5	-1.61	-1.52	-1.78
ΔE	4.167	3.732	3.876	3.963	4.76	3.8	4.4	3.48	4.25	3.92	4.03	3.76	3.94	3.53	4.46
I	6.319	5.892	6.057	6.25	6.37	5.03	6.1	5.36	6.16	6.24	5.64	5.26	5.55	5.05	6.24
A	2.152	2.16	2.18	2.29	1.61	1.23	1.71	1.88	1.91	2.32	1.61	1.5	1.61	1.52	1.78
η	2.083	1.866	1.939	1.98	2.38	1.9	2.195	1.74	2.125	1.96	2.015	1.88	1.97	1.765	2.23
σ	0.48	0.536	0.516	0.505	0.42	0.526	0.456	0.575	0.471	0.51	0.496	0.532	0.508	0.567	0.448
χ	4.235	4.026	4.119	4.27	3.99	3.13	3.905	3.62	4.035	4.28	3.625	3.38	3.58	3.285	4.01
ΔN	0.663	0.797	0.743	0.689	0.632	1.018	0.705	0.971	0.698	0.694	0.837	0.963	0.868	1.052	0.67

2.2. Process of ML model

Seven regression algorithms were compared systematically: linear regression, decision tree, random forest, Lasso, Ridge, Elastic Net, and gradient boosting. The initial candidate matrix contained 11 DFT-derived quantum chemical descriptors. To prevent information leakage, descriptor selection was defined as a pre-modelling step, and the final predictive assessment was performed on compounds excluded from model fitting. The same selected descriptor set and the same 12/3 training–test partition were used for all algorithms so that differences in R^2, MSE, and MAE reflected model behaviour rather than different data partitions.

2.2.1. Variable selection procedure

The complete candidate descriptor set comprised total energy (TE), dipole moment (μ), E_{HOMO}, E_{LUMO}, energy gap (ΔE), ionization potential (I), electron affinity (A), hardness (η), softness (σ), electronegativity (χ), and fraction of electrons transferred (ΔN). Descriptor selection was performed using simulated annealing (SA) as a wrapper-based subset-search method. In the SA procedure, a candidate state represents a subset of the 11 descriptors. A neighbouring state is generated by adding or removing a descriptor, and the candidate subset is evaluated using predictive performance within the training data. Better subsets are accepted, whereas a worse subset can be accepted with probability $P = \exp(-\Delta/T)$, where Δ is the deterioration in the objective function and T is the current temperature. The temperature is gradually reduced so that the search moves from broad exploration to local refinement. The subset repeatedly retained as the most informative consisted of χ, E_{LUMO}, ΔN, and ΔE. These

descriptors were therefore used as inputs to the seven regression models. Importantly, the independent test compounds were not used to calculate the final model coefficients or test statistics. Lasso, Ridge, and Elastic Net additionally apply embedded coefficient shrinkage during fitting; this embedded regularization is distinct from the preceding SA subset search.

2.2.2. Linear regression (LR)

A supervised machine learning algorithm learns from labeled datasets and uses this knowledge to map data points to optimized linear functions. This approach enables the prediction of outcomes for new datasets by identifying the best-fit line equation, which estimates values based on the independent variables. The algorithm examines the correlation between input features (independent variables) and the output (dependent variable), showing this correlation as a linear function optimized to represent a straight line. The primary objective of linear regression is to identify the linear equation that provides the best fit and minimizes the prediction error. The equation is represented as $y = mx + c$, where y represents the dependent variable, x identifies the independent variable, m indicates the slope of the line, which expresses the rate of change of y with respect to x, and c marks the y-intercept, reflecting the value of y when x equals zero. With the use of least squares optimization, the technique helps to lessen the differences between the actual and anticipated values. The approach minimizes the sum of squared residuals, which are the differences between the observed and predicted values, to achieve the best possible compatibility with the data [40].

2.2.3. Decision tree (DT)

DT is a supervised learning algorithm implemented for classification and regression tasks. Classification problems mainly use their ability to manage categorical data and yield interpretable results. A decision tree is constructed as a hierarchical model, with core nodes representing dataset attributes, branches representing decision rules, and leaf nodes expressing outcomes. Decision nodes facilitate decision-making and possess multiple branches, whereas leaf nodes, representing the ultimate outcomes of those decisions, do not branch further. The structure of a decision tree is analogous to an inverted tree, consisting of a root, internal nodes, branches, and leaf nodes [41]. The root node, situated at the top of the tree and serving as the point of departure for data segmentation, represents the entire dataset. At each internal node, which represents the characteristics or attributes of the dataset, there are specific decision rules that govern each split. There is another name for internal nodes: internal nodes. To connect the nodes and outline the pathways connected to the branches, the defined decision criteria are used. In their capacity as terminal nodes, leaf nodes are assigned the responsibility of generating the ultimate prediction or outcome. Unlike internal nodes, leaf nodes do not undergo any further division. They are used to represent the predicted value in regression tasks and the class label in classification tasks. Internal nodes involve further division and are subject to extra division.

2.2.4. Random Forest (RF)

RF is a popular supervised learning method for classification and regression problems, known for its ability to reduce overfitting, handle complex datasets, and make reliable predictions. This ML technique builds numerous decision trees during training and integrates their results to make forecasts more accurate. RF uses a process called bootstrap aggregating, which makes several subsets of the training data by randomly sampling with replacement. For each subgroup, a different decision tree is created. At each node, a random set of attributes is used to split the data, which makes trees diversity more while reducing their mutual dependence. Each tree in the forest is allowed to grow unpruned, enabling it to capture complex patterns within the data. Each tree generates a class prediction for classification tasks, and the random forest identifies the class that appears most frequently across all trees. In regression tasks, each tree produces a numerical prediction, and the Random Forest model computes the average of these predictions to determine the final output [42].

2.2.5. Lasso regression

Lasso Regression, also known as the Least Absolute Shrinkage and Selection Operator, is a linear regression method that involves L1 regularisation to enhance model precision and interpretability. By introducing a penalty based on the absolute value of the magnitude of the coefficients to the loss function, Lasso shrinks the coefficients of less significant features towards zero, resulting in a sparse model that automatically performs feature selection by discarding redundant or irrelevant descriptors. In QSPR modeling of corrosion inhibition prediction, Lasso is instrumental since it, in addition to extracting linear correlations between inhibition efficiency and molecular descriptors, also identifies the most relevant descriptors of the inhibition mechanism, making it perfectly suitable for use when handling high-dimensional chemical data, where not only should prediction be accurate, but also interpretability and reduction of descriptors are equally essential. The trade-off achieved by the Lasso between model simplicity and performance is worthwhile for scientific applications, where understanding feature importance is necessary for experimental verification and molecular design [43].

Cost Function: RSS + λ * $\Sigma|\beta i|$, where RSS is the Residual Sum of Squares, λ is the regularisation parameter, and βi are the coefficients. Shrinks coefficients towards zero and can set them to zero, leading to feature selection.

2.2.6. Ridge Regression

Ridge Regression is a linear regression method that uses L2 regularisation to handle multicollinearity and overfitting in datasets with correlated or multiple features. It alters the ordinary least squares loss function by incorporating a penalty term that is proportional to the square of the absolute value of the coefficients. In contrast to Lasso Regression, Ridge doesn't set coefficients to zero but instead pulls them towards zero, keeping all features in the model,

making Ridge appropriate for cases where all input variables are potentially helpful and should be retained, albeit with diminished influence. In the case of QSPR modeling for corrosion inhibition prediction, Ridge Regression will stabilize the model in the presence of complex DFT descriptors that may have strong intercorrelations[44]. However, its performance degrades in small datasets or when the correspondence between the features and the target variable is highly nonlinear. Still, Ridge offers a linear method with control and interpretability, which is very useful for understanding the combined effect of descriptors on inhibition efficiency. Cost Function: RSS + $\lambda * \Sigma\beta i^2$ is the effect that shrinks coefficients towards zero, but rarely sets them to zero.

2.2.7. Gradient boosting (GB) and Elastic Net

By learning from residuals (the gaps between actual and predicted values), gradient boosting is a supervised ensemble strategy that constructs models one after the other, intending to minimize errors from the preceding model. It works exceptionally well for managing complex datasets and achieving high precision. The core concept of GB is to progressively reduce the loss function by building a model through a series of weak learners, often decision trees [45].

Elastic net linear regression applies the lasso and ridge penalties to regularise regression models. The method combines the Lasso and ridge regression methods by learning from the weaknesses of the former two approaches to enhance statistical model regularisation. The Elastic Net addresses the deficiency of Lasso, particularly in cases where Lasso requires a large number of samples for high-dimensional data. The elastic net algorithm allows for the addition of "n" variables until the point of saturation is reached. If the variables are groups of highly correlated variables, Lasso will select a variable from such sets and omit the rest completely. To address the shortcomings encountered with Lasso, the Elastic Net introduces a quadratic term ($\|\beta\|^2$) to the penalty, which, when applied separately, yields ridge regression. The quadratic term in the penalty pushes the loss function closer to being convex. The elastic net combines the strengths of both worlds – i.e., Lasso and ridge regression.

2.3. Model evaluation

The performance of the models was evaluated using standard metrics, including the coefficient of determination (R^2), mean squared error (MSE), and mean absolute error (MAE). A value of R^2 closer to 1 indicates a better fit of the model to the data [46]. MSE and MAE, on the other hand, measure the difference between the observed and predicted values, providing insight into the model's accuracy. Lower error values suggest a more effective predictive model. The specific formulas for R^2, MSE, and MAE, as provided in Equations 1, 2, and 3, help quantify these measures and guide the optimization of the predictive model.

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

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i' - Y_i)^2 \quad (2)$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |Y_i' - Y_i| \quad (3)$$

where n, Y_i , $\bar{Y}_i$, and Y_i' Represent the number of samples, the actual mean, and predicted values, respectively.

2.4. Model Validation

The modelling dataset contained 15 compounds. A reproducible hold-out validation scheme was used to answer the requirement for an explicit test set. Twelve compounds (80%) were assigned to the training set and three compounds (20%) to the independent test set using a fixed random seed (random_state = 42). The test compounds were ODQDHT, MPDQO, and DQ; the remaining 12 compounds were used for training. Descriptor scaling, where required for the linear regularized models, was fitted on the training data and then applied to the test data. The seven regression models were trained only with the 12 training compounds. R^2, MSE, and MAE reported in Table 3 were then calculated from predictions for the three held-out compounds. Because R^2 is statistically unstable when calculated from only three test observations, MSE and MAE are also emphasized, and the small sample size is explicitly recognized as a limitation. This independent test-set procedure provides a stricter assessment of generalization than reporting training-fit statistics alone.

3. RESULTS AND DISCUSSION

3.1. Selection of Descriptors

According to the results thus far, attempts have been made to correlate specific quantum chemical properties with the observed inhibitory efficiency of the inhibitors. The quantitative structure-property relationship (QSPR) methodology integrates a range of quantum chemical parameters to create a composite index that correlates with experimentally determined inhibition efficiency. The adsorption mechanism significantly influences the choice of the appropriate model equation. The present study demonstrates that the inhibition mechanism predominantly entails the transfer of electrons from the inhibitor to the iron in mild steel. Observations indicate that the inhibitor operates by donating electrons to the metal and accepting electrons from the lone pairs of iron, thereby forming a feedback bond. The linear model for approximating corrosion inhibition potential, as presented by Lukovits et al. [48].

$$IE_{Theo}(\%) = \left(AX_jC_j + B\right) \quad (4)$$

A and B denote the regression coefficients derived from regression analysis, X_j indicates a quantum chemical index associated with molecule j, and C_j represents the experimental concentration of the inhibitor. As shown in Equation 4, the linear model did not exhibit a strong correlation between the experimental and theoretical inhibition efficiencies. A non-linear model (Equation 5), initially proposed by Lukovits et al. [49], was introduced to analyze the interaction between corrosion inhibitors and metal surfaces in acidic environments. The non-linear model, which is derived from the Langmuir adsorption isotherm, can be articulated as follows:

$$IE_{Theo}(\%) = \frac{(AX_j+B)C_j}{1+(AX_j+B)C_j} \times 100 \quad (5)$$

$IE_{Theo}\%$ represents the theoretical inhibition efficiency, A denotes a regression coefficient, B signifies a regression constant, C_j indicates the experimental concentration of the inhibitor, and X_j refers to a quantum chemical index of the molecule. The relationship between quantum chemical parameters (E_{HOMO}, E_{LUMO}, ΔE, and ΔN) and experimental inhibition efficiency for the protonated forms of quinoxaline derivatives has been previously discussed and documented in the literature [24,25]. The QSPR study identified essential molecular descriptors that significantly affect the corrosion inhibition efficacy of quinoxaline derivatives. The simulated annealing algorithm identified the most pertinent global descriptors as follows:

Electronegativity determines the electron-donating or withdrawing capacity of compounds [50]. E_{LUMO} reflects the electron affinity of the compound, influencing its interaction with the metal surface [51]. The fraction of electrons transferred indicates the amount of charge transferred from the inhibitor to the mild steel surface, which is crucial for effective adsorption [52]. The energy gap, ΔE, is an essential component that determines how effectively the inhibitor adheres to the metal surface. As ΔE decreases, the inhibitor becomes more reactive, which enhances its effectiveness. A large ΔE (HOMO-LUMO) gap means that the molecule is more stable, which makes it less effective at stopping other molecules. A lower HOMO energy and a narrower energy gap suggest that the adsorption potential is more substantial and the inhibitor is more effective [53]. The characteristics indicated that electronic properties play a crucial role in preventing corrosion, supporting the idea that molecular interactions, particularly adsorption, are vital. The QSPR analysis reveals the importance of chemical descriptors such as ΔE, E_{HOMO}, E_{LUMO}, and ΔN in determining the effectiveness of an inhibitor. The non-linear Langmuir model fits the experimental data better than the linear model. The findings underscore the importance of adsorption mechanisms and electronic interactions in understanding and enhancing the effectiveness of corrosion inhibitors.

3.2. Model performance

Model performance was assessed exclusively on the independent three-compound test set after fitting each model with the 12 training compounds. Table 3 reports the resulting R^2, MSE,

and MAE values. The test compounds were ODQDHT, MPDQO, and DQ. The results show that training-set fit should not be interpreted as predictive validation in this small dataset. Linear regression produced the lowest test MAE (4.48%) and MSE (21.14), although its test R^2 remained negative (−0.70). The regularized linear models produced similar MAE values of approximately 4.59–4.62%. Decision tree showed the largest test error (MAE = 7.91%), while random forest and gradient boosting produced MAE values of 4.74% and 5.39%, respectively. Thus, none of the evaluated models demonstrated strong independent predictive generalization. The negative test-set R^2 values indicate that the current models do not explain the variation of the held-out compounds better than the test-set mean baseline. This result is consistent with the unfavorable ratio between the number of compounds and the number of candidate descriptors. The descriptor analysis nevertheless remains useful for mechanistic interpretation: χ, E_{LUMO}, ΔN, and ΔE describe electron acceptance/donation and charge-transfer behaviour relevant to adsorption on the mild-steel surface. Predictive claims are therefore restricted to the present chemical domain, and expansion of the dataset is required before a robust externally predictive QSPR model can be established.

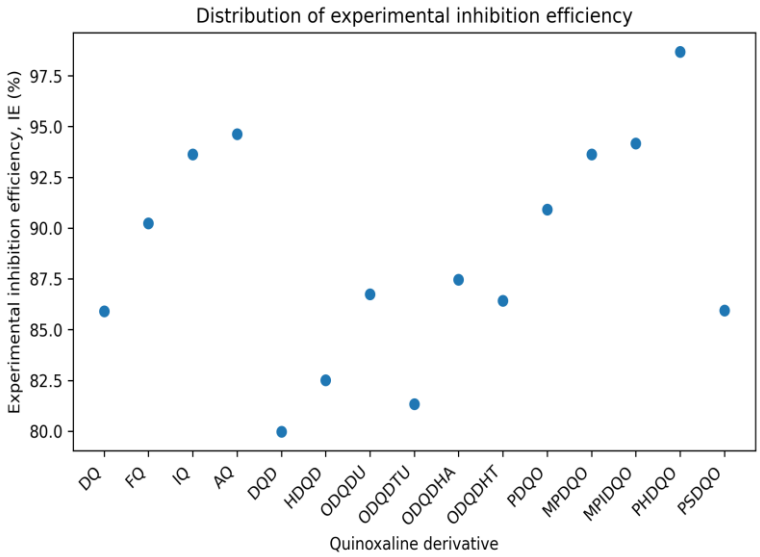


Figure 2. Point distribution of experimental inhibition efficiency for the 15 quinoxaline derivatives used in the QSPR modelling dataset

Figure 2 displays every compound as an individual point so that the distribution of the experimental inhibition-efficiency data can be inspected directly. The response values span approximately 79.97–98.69%, with most compounds concentrated in the mid-to-high efficiency range. This narrow response range and the small number of compounds make independent prediction difficult and also make test-set R^2 sensitive to individual compounds. Accordingly, the model comparison is interpreted using R^2 together with MSE and MAE rather than relying on a single statistic.

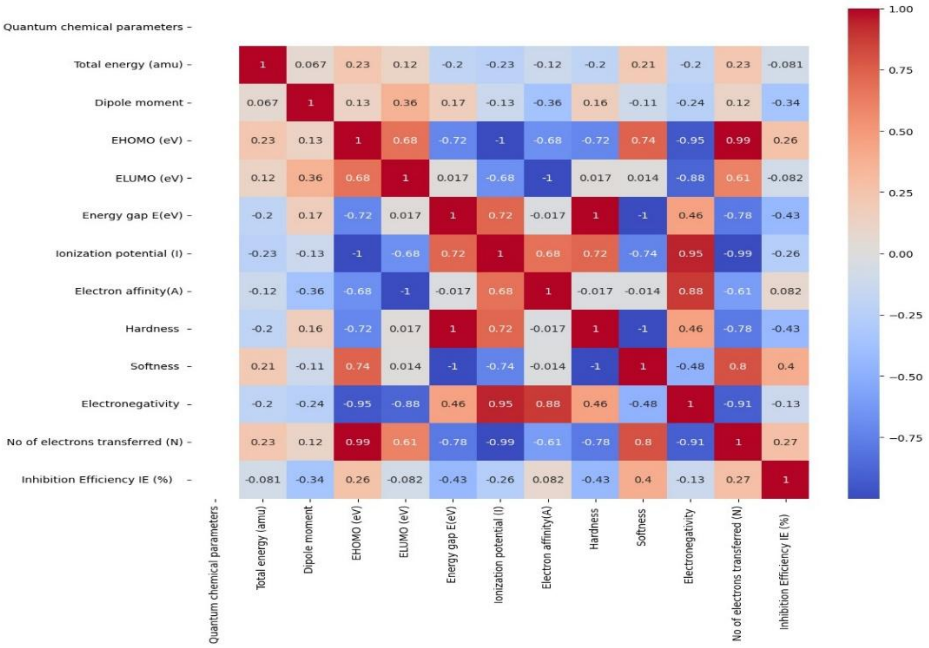


Figure 3. Linear Correlation graph

The independent test-set statistics confirm that none of the seven models is sufficiently validated for routine prediction. Linear regression had the lowest MAE (4.48%), followed closely by Lasso and Elastic Net (approximately 4.60%). Random forest produced an MAE of 4.74%, whereas gradient boosting and decision tree had larger errors. The negative R^2 values for all models must be reported rather than replaced by training-set goodness-of-fit values. These findings demonstrate the importance of held-out validation in small QSPR datasets and prevent overstatement of model performance.

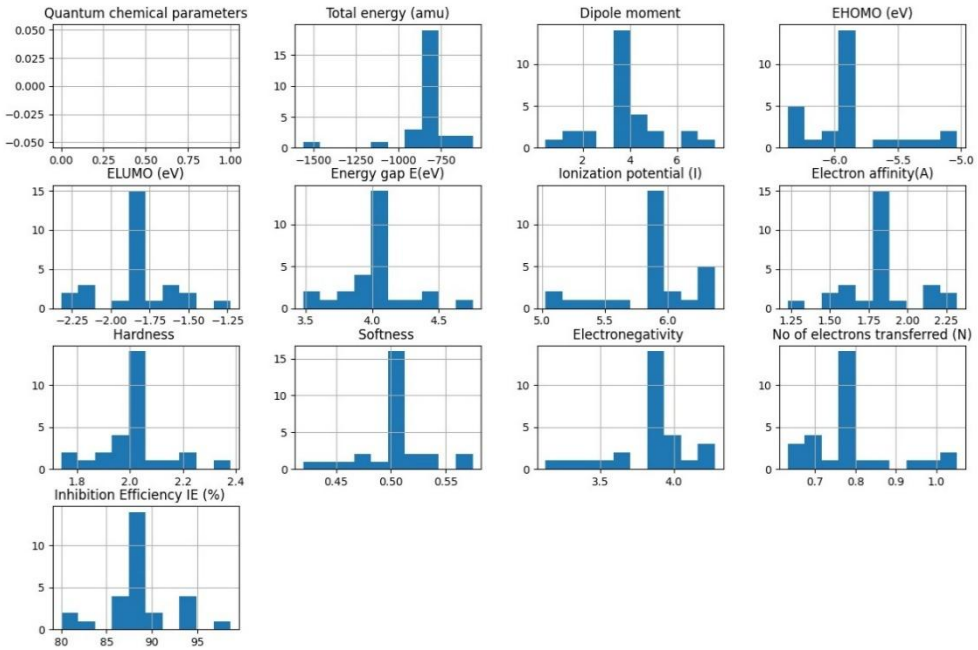


Figure 4. The effect of virtual samples on histograms of feature data

Table 3 summarizes the independent test-set results. The model ranking based on MAE differs from conclusions that would be obtained from training-fit statistics alone. Therefore, the present study does not claim that gradient boosting or random forest establishes a benchmark model. Instead, the test-set analysis shows that the current 15-compound dataset is insufficient for reliable predictive validation. The principal contribution of the modelling analysis is the identification of mechanistically relevant electronic descriptors and the demonstration that larger, chemically diverse datasets are necessary for a stable predictive QSPR model.

Table 3. Model validation

Model	Test R^2	Test MSE	Test MAE (%)
Linear Regression	-0.70	21.14	4.48
Lasso	-0.74	21.58	4.60
Ridge	-0.75	21.73	4.62
ElasticNet	-0.73	21.54	4.59
Decision Tree	-4.77	71.72	7.91
Random Forest	-0.92	23.83	4.74
Gradient Boosting	-1.63	32.70	5.39

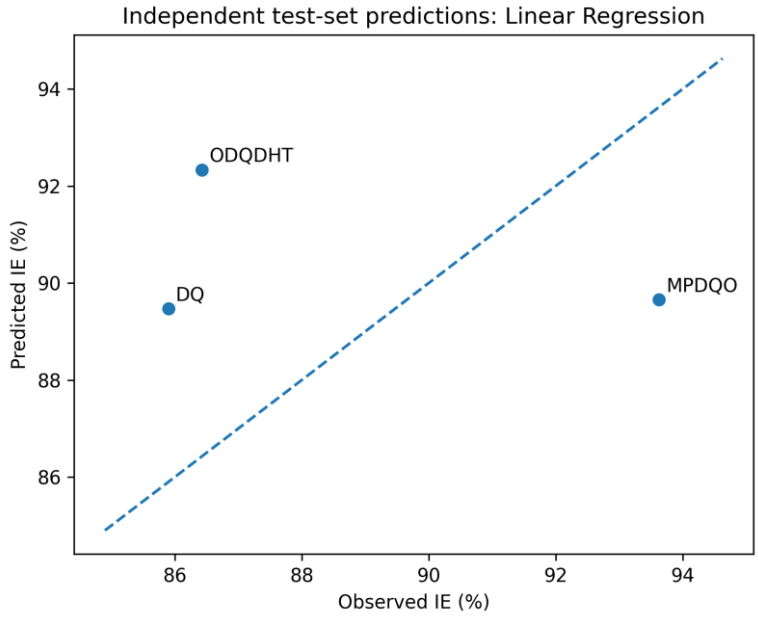


Figure 5. Observed versus predicted inhibition efficiency for the three independent test compounds using Linear Regression, the model with the lowest test MAE

3.3. Practical implications for the Inhibitor design

The selected descriptor–property relationships provide qualitative guidance for inhibitor design, but the independent test-set results require cautious interpretation. The association of

inhibition efficiency with χ, E_{LUMO}, ΔN, and ΔE supports the role of electron transfer and adsorption in the interaction between quinoxaline derivatives and the mild-steel surface. These descriptors may therefore be used to prioritize candidate molecules for further computational and experimental study. However, because none of the seven models achieved strong predictive generalization on the held-out test compounds, the current models should be treated as exploratory QSPR tools rather than validated screening engines. Future studies should expand the number and structural diversity of quinoxaline derivatives and perform external validation on compounds obtained independently from model development.

4. CONCLUSION

This study evaluated QSPR machine-learning models for predicting the corrosion inhibition efficiency of quinoxaline derivatives on mild steel in acidic medium. The supplied modelling dataset contained 15 compounds and 11 DFT-derived quantum chemical descriptors. Simulated-annealing-based descriptor selection identified electronegativity (χ), E_{LUMO}, fraction of electrons transferred (ΔN), and energy gap (ΔE) as the most informative descriptor subset. Seven regression models were compared using an explicit 80:20 hold-out validation design comprising 12 training compounds and 3 independent test compounds. The held-out compounds were ODQDHT, MPDQO, and DQ. On the independent test set, linear regression yielded the lowest MAE (4.48%) and MSE (21.14), but its R^2 was negative (−0.70); the remaining models also produced negative R^2 values. Consequently, none of the evaluated models can yet be regarded as reliably validated for unseen compounds. The electronic descriptors remain useful for interpreting adsorption and electron-transfer effects, but a larger and more diverse dataset, followed by external validation, is required before strong predictive claims can be made.

Declarations of interest

The authors declare no conflict of interest in this reported work.

REFERENCES

[1] D. Winkler, Metals 7 (2017) 553.
[2] B.D.B. Tiu, and R.C. Advincula, React. Funct. Polym. 95 (2015) 25.
[3] S. Muralidharan, K. Phani, S. Pitchumani, S. Ravichandran, and S. Iyer, J. Electrochem. Soc. 142 (1995) 1478.
[4] M. Finšgar, and J. Jackson, Corros. Sci. 86 (2014) 17.
[5] I. Obot, D. Macdonald, and Z. Gasem, Corros. Sci. 99 (2015) 1.
[6] H. Ashassi-Sorkhabi, B. Shaabani, and D. Seifzadeh, Appl. Surf. Sci. 239 (2005) 154.
[7] W. Kohn, and L. J. Sham, Phys. Rev. 140 (1965) A1133.

[8] G. Gece, Corros. Sci. 50 (2008) 2981.
[9] S. Martinez, L. Valek, and I. S. Oslaković, J. Electrochem. Soc. 154 (2007) C671.
[10] A.M. Al-Fakih, Z.Y. Algamal, M.H. Lee, H.H. Abdallah, H. Maarof, and M. Aziz, J. Chemom. 30 (2016) 361.
[11] M. Belghiti, S. Echihi, A. Dafali, Y. Karzazi, M. Bakasse, H. Elalaoui-Elabdallaoui, L. Olasunkanmi, E. Ebenso, and M. Tabyaoui, Appl. Surf. Sci. 491 (2019) 707.
[12] I. Obot, and Z. Gasem, Corros. Sci. 83 (2014) 359.
[13] R. C. Ayers, and N. Hackerman, J. Electrochem. Soc. 110 (1963) 507.
[14] K. Aramaki, and J. Uehara, J. Electrochem. Soc. 136 (1989) 1299.
[15] Y.A. Ammar, M.M.F. Ismail, M.S.A. El-Gaby, and M.A. Zahran, Indian J. Chem. B 41 (2002) 1486.
[16] A.A. Abu-Hashem, M.A. Gouda, and F.A. Badria, Eur. J. Med. Chem. 45 (2010) 1976.
[17] H. Asif, and D. Madhesia, J. Pharm. Res. 4 (2011) 924.
[18] H. Chung, O. Jung, M. J. Chae, S. Hong, K. Chung, S. K. Lee, and C. Ryu, Bioorg. Med. Chem. Lett. 15 (2005) 3380.
[19] M. Suresh, P. Lavanya, D. Sudhakar, K. Vasu, and C. Venkata Rao, J. Chem. Pharm. Res. 2 (2010) 497.
[20] J. Saranya, K. Lavanya, M. H. Kiranmai, R. Subbiah, A. Zarrouk, and S. Chitra, Corros. Rev. 39 (2021) 79.
[21] N. Hackerman, and A. Makrides, Ind. Eng. Chem. 46 (1954) 523.
[22] V. Sastri, and J. Perumareddi, Corrosion 53 (1997) 617.
[23] J. Saranya, P. Sounthari, K. Parameswari, and S. Chitra, J. Mater. Environ. Sci. 8 (2017) 370.
[24] J. Saranya, M. Sowmiya, P. Sounthari, K. Senthil Kumar, K. Parameswari, and S. Chitra, J. Mol. Liq. 216 (2016) 42.
[25] J. Saranya, P. Sounthari, K. Parameswari, and S. Chitra, Measurement 77 (2016) 175.
[26] J. Saranya, P. Sounthari, K. Parameswari, and S. Chitra, Der Pharma Chem. 7 (2015) 187.
[27] J. Saranya, P. Sounthari, K. Parameswari, and S. Chitra, J. Mater. Environ. Sci. 6 (2015) 425.
[28] J. Saranya, P. Sounthari, A. Kiruthika, G. Saranya, S. Yuvarani, K. Parameswari, and S. Chitra, Orient. J. Chem. 30 (2014) 1719.
[29] G. Gece, S. Bilgiç, and Ö. Türksen, Mater. Corros. 61 (2010) 141.
[30] R. Gu, P. Cao, J. Yao, B. Ren, Y. Xie, B. Mao, and Z. Tian, J. Electroanal. Chem. 505 (2001) 95.
[31] N. Obot, and N. Obi-Egbedi, Corros. Sci. 52 (2010) 657.
[32] T. Sutojo, S. Rustad, M. Akrom, A. Syukur, G. F. Shidik, and H. K. Dipojono, npj Mater. Degrad. 7 (2023) 18.
[33] B. El Ibrahimi, et al., J. King Saud Univ. Sci. 32 (2020) 163.

[34] M.F.N. Roy, H. Taufique, H. Khakurel, R. Devanathan, D. D. Johnson, and G. Balasubramanian, npj Mater. Degrad. 6 (2022) 9.
[35] C.T. Ser, P. Žuvela, and M. W. Wong, Appl. Surf. Sci. 512 (2020) 145612.
[36] T.W. Quadri, et al., Arab. J. Chem. 15 (2022) 103870.
[37] V. C. Anadebe, et al., J. Mol. Liq. 349 (2022) 118482.
[38] V. C. Anadebe, V. I. Chukwuike, S. Ramanathan, and R. C. Barik, Process Saf. Environ. Prot. 168 (2022) 499.
[39] C. Beltran-Perez, et al., Int. J. Mol. Sci. 23 (2022) 95086.
[40] X. Wang, and J. Yao, Math. Probl. Eng. 2020 (2020) 1.
[41] T. W. Quadri, et al., Arab. J. Chem. 15 (2022) 103870.
[42] Y. Liu, Y. Guo, W. Wu, Y. Xiong, C. Sun, L. Yuan, and M. Li, Interdiscip. Sci. Comput. Life Sci. 11 (2019) 738.
[43] A.J. Smola, B. Schölkopf and S. Schölkopf, Kluwer Academic Publishers, Dordrecht (2004).
[44] Z. Zhang, Ann. Transl. Med. 4 (2016) atm.2016.03.37.
[45] A. Natekin, and A. Knoll, Front. Neurorobot. 7 (2013) 21.
[46] A. Botchkarev, Interdiscip. J. Inf. Knowl. Manag. 14 (2019) 45.
[47] R. Kohavi, Proc. 14th Int. Joint Conf. Artif. Intell. (1995) 1137.
[48] I. Lukovits, I. Bakó, A. Shaban, and E. Kálmán, Electrochim. Acta 43 (1998) 131.
[49] I. Lukovits, A. Shaban, and E. Kalman, Russ. J. Electrochem. 39 (2003) 177.
[50] M. Akrom, S. Rustad, and H. K. Dipojono, MRS Commun. 14 (2024) 379.
[51] R. G. Pearson, Coord. Chem. Rev. 100 (1990) 403.
[52] N. Islam, and D. Chandra Ghosh, Mol. Phys. 109 (2011) 917.
[53] S. Issaadi, T. Douadi, and S. Chafaa, Appl. Surf. Sci. 316 (2014) 582.
[54] R. G. Pearson, Inorg. Chem. 27 (1988) 734.
[55] H. Nasution, N. Enizan, N. Nurlaili, and J. Syahri, Acta Chim. Asian 3 (2020) 181.
[56] R. Hmamouchi, M. Larif, S. Chtita, A. Adad, M. Bouachrine, and T. Lakhlifi, J. Taibah Univ. Sci. 10 (2016) 451.
[57] J. F. Hair, C. M. Ringle, and M. Sarstedt, Long. Range Plan. 46 (2013) 1.
[58] J. Henseler, C. M. Ringle, and M. Sarstedt, Int. Mark. Rev. 33 (2016) 405.
[59] S. Hadisaputra, A. A. Purwoko, A. Hakim, N. Prasetyo, and S. Hamdiani, ACS Omega 7 (2022) 33054.
[60] D. T. Stanton, Curr. Comput.-Aided Drug Des. 8 (2012) 107.