Materials Today Communications 35 (2023) 106402

Contents lists available at ScienceDirect

Materials Today Communications

journal homepage: www.elsevier.com/locate/mtcomm

A combination of machine learning model and density functional theory method to predict corrosion inhibition performance of new diazine derivative compounds

Muhamad Akrom [a,e], Supriadi Rustad [e,*], Adhitya Gandaryus Saputro [b,c], Aditianto Ramelan [d], Fadjar Fathurrahman [b,c], Hermawan Kresno Dipojono [b,c,**]

[a] Doctoral Program of Engineering Physics, Faculty of Industrial Technology, Bandung Institute of Technology, Bandung 40132, Indonesia
[b] Advanced Functional Materials Research Group, Bandung Institute of Technology, Bandung 40132, Indonesia
[c] Research Center for Nanosciences and Nanotechnology, Bandung Institute of Technology, Bandung 40132, Indonesia
[d] Materials Science and Engineering Research Group, Faculty of Mechanical and Aerospace Engineering, Bandung Institute of Technology, Bandung 40132, Indonesia
[e] Research Center for Materials Informatics, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Indonesia

ARTICLE INFO

Keywords:
ML
QSPR
DFT
Diazine
Corrosion inhibition

ABSTRACT

This study proposes a novel approach that combines machine learning (ML) and density functional theory (DFT) methods to construct a quantitative structure-properties relationship (QSPR) model for diazine derivatives as anti-corrosion inhibitors. A dataset is constructed by combining three existing diazine isomer datasets to represent diazine compounds. Thirty-two different ML algorithms were implemented on the dataset, and the gradient boosting regressor (GBR) model was identified as the best predictive model for diazine and each isomer dataset based on the coefficient of determination (R^2) and root mean square error (RMSE) metric values. This consistency was also observed when the GBR model was implemented on four other diazine derivatives, resulting in high corrosion inhibition efficiency (CIE) values ranging from 85.02 % to 94.99 %. The DFT calculations for these derivatives also showed strong adsorption energies ranging from − 4.41 to − 6.09 eV, in line with the CIE trend obtained from the ML prediction. This novel approach can provide insights into the properties of prospective organic corrosion inhibitors prior to experimental investigations, which could accelerate the development of new and effective organic corrosion inhibitors.

1. Introduction

The use of organic compounds as corrosion inhibitors for industrial metals has been the focus of numerous studies due to their non-toxic, environmentally friendly, and cost-effective properties, as well as their high corrosion inhibition capabilities. Efficient organic inhibitors typically contain heteroatomic groups (such as N, S, O, or P) and aromatic rings in their structure. Among the organic compounds, diazine isomers, such as pyridazine (1,2-diazine), pyrimidine (1,3-diazine), and pyrazine (1,4-diazine) families [1], have been extensively investigated for their effectiveness and efficiency in inhibiting metal corrosion due to their aromatic rings and heteroatom groups (O and N). Experimental investigations have utilized weight loss methods, electrochemical impedance spectroscopy, and potentiodynamic polarization to evaluate the performance of diazine isomers as inhibitors [2–4]. These studies have reported excellent CIE for these isomers. However, the experimental evaluation of a wide range of inhibitor candidates is a time-consuming and resource-intensive process.

Since the diazine isomers had been experimentally proven to have excellent CIEs, it is interesting to investigate whether the high CIEs are also true for other diazine derivatives. Some of them are 1-Phenyl-1 H-pyrazolo[3,4-*d*]-pyrimidine-4-ol; 1-(3-methyl-1-phenyl-1 H-pyrazole-4-

Abbreviations: DFT, density functional theory; GBR, gradient boosting regression; CIE, corrosion inhibition efficiency; ML, machine learning; QSPR, quantitative structure-property relationship.

* Corresponding author.
** Correspondence to: Advanced Functional Materials Research Group and Research Center for Nanosciences and Nanotechnology, Bandung Institute of Technology, Bandung 40132, Indonesia.
E-mail addresses: srustad@dsn.dinus.ac.id (S. Rustad), dipojono@tf.itb.ac.id (H.K. Dipojono).

https://doi.org/10.1016/j.mtcomm.2023.106402
Received 20 February 2023; Received in revised form 23 May 2023; Accepted 6 June 2023
Available online 8 June 2023

Table 1
A diazine dataset, constructed from three diazine isomer datasets.

Dataset	Number of molecules	Ref.
Pyridazine (1,2-diazine)	18	[13,14]
Pyrimidine (1,3-diazine)	94	[15,16]
Pyrazine (1,4-diazine)	8	[17]

yl)-ethenone; 1-Phenyl-1 h-pyrazolo[3,4-*d*]-pyrimidine-4-amine; and 1-Phenyl-1 h-pyrazolo[3,4-*d*]-pyrimidine-4-carbonitrile. These four diazine derivatives are organic compounds that have not been tested as candidates for corrosion inhibitors but have been reported in the literature as anticoagulant [5], antipsychotic [6], antimitotic [7], analgesic [8], anti-inflammatory [9], glucosamine reductases [10], and antimicrobial [11,12]. They also have aromatic rings and heteroatom groups (N and O) previously recognized as good inhibitors' specific properties.

Recently, a combination of ML and DFT methods has revealed a QSPR. Chemical properties derived from DFT calculations are input features, while CIE from the experimental study is treated as a target of a variety of accurate QSPR models. The correlation between molecular structure characteristics and their chemical properties makes the QSPR a cheap, fast, and reliable technique in the exploration of new inhibitors. Various ML algorithms have been widely used in the development of QSPR models to evaluate the performance of inhibitors. Some were implemented and compared in a search for the best model for a certain dataset. Quadri et al. [13] used multilinear regression (MLR) and artificial neural network (ANN) to develop a QSPR model for designing pyridazine corrosion inhibitors and found that between the two, the ANN model produces the best prediction performance. Pyridazine was also studied by Assiri et al. [14] comparing three ML algorithms namely partial least square (PLS), ANN, and principal component regression (PCR), and the result was that PCR produced the best predictive ability among the tested algorithms. Alamri et al. [15] investigated another diazine isomer called pyrimidine using linear and non-linear algorithms, namely PLS and random forest (RF). They reported that the RF performed the best of the two. For the pyrimidine dataset, with the same algorithms, Quadri et al. [16] also found that the non-linear ANN won over the linear MLR. In the case of the pyrazine dataset, MLR was implemented by Obot et al. [17] to produce the CIE prediction of new pyrazine derivatives. The pyridazine, pyrimidine, and pyrazine belong to the diazine compounds, therefore it is hypothesized that they should be associated with the same best QSPR model that is suitable for the diazine as well as each isomer dataset. The search for the best model for diazine compounds is crucial, especially to investigate the potential of diazine derivatives as inhibitors using ML.

The development of a systematic and efficient approach for identifying the best predictive model for diazine compounds is of great importance in the search for effective corrosion inhibitors using ML. To the best of our knowledge, this study represents the first effort to combine ML and DFT methods to construct a QSPR model for diazine compounds to identify potential corrosion inhibitors. The proposed

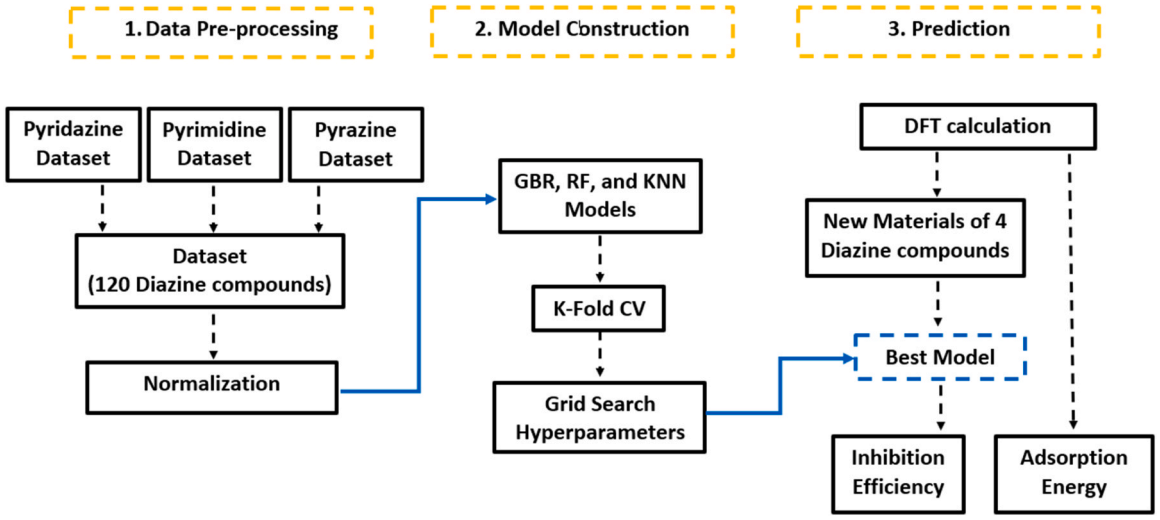


Fig. 1. (a) Model development of QSPR and (b) K-Fold CV model.

Table 2
The hyperparameters of GBR, RF, and KNN, and their range and optimal values.

Model	Hyperparameter	Dimension	Range value	Optimal value
GBR	n-Estimators	1 – infinity	50, 100, 500, 1000	100
	Max Depth	1 – infinity	3, 5, 7, 9	3
	Learning Rate	0 – infinity	0.01, 0.05, 0.1, 0.2	0.1
RF	n-Estimators	10 – 100	20, 50, 80, 100	50
	Max Depth	Integer	3, 5, 7, 9	3
	Min Sample Leaf	Integer	2, 4, 5, 6	5
KNN	n-Neighbors	Integer	2, 3, 4, 5	3
	Weights	Uniform, Distance	uniform, distance	distance
	Leaf Size	Integer	10, 20, 30, 40	10

method has been shown to be highly effective in identifying the best predictive model for diazine compounds, and it is anticipated that it will be applicable to the search for other organic corrosion inhibitors beyond diazine compounds. In this work, we construct a dataset of 120-diazine compounds by integrating its isomer datasets. To search for the best model suitable for the diazine compounds, as many as 15 linear and 17 non-linear different ML algorithms are tested on this diazine dataset, and the best model is then used to make the CIE prediction of another four diazine derivatives that have not been used as corrosion inhibitors. The DFT calculation is used to provide input features of the chemical properties of these derivatives to make a prediction of their CIEs. It is also used to calculate the adsorption energies of these diazine derivatives to evaluate their interactions with the iron surface. This paper starts with an introduction followed by materials and methods in Section 2, results and discussions in Section 3, and enclosed by a conclusion in Section 4.

2. Materials and methods

2.1. Dataset and chemical properties

A dataset of 120 diazine-derived compounds (pyridazine, pyrimidine, and pyrazine) constructed from published literature (Table 1), is used to statistically validate the QSPR model to consider and analyse a corrosion inhibition design. A total of 11 chemical properties calculated by DFT are used as independent variables to evaluate their correlation to the CIE as the dependent variable of inhibitor compounds. The chemical properties include HOMO, LUMO, energy gap (ΔE), ionization potential (I), electron affinity (A), electronegativity (χ), global hardness (η), global softness (σ), dipole moment (μ), electrophilicity (ω), and a fraction of electron transferred (ΔN). Corrosion inhibition is highly dependent on these chemical properties [18,19].

2.2. Data preprocessing

After the data is collected and structured, it is normalized using the standard scaler normalize method to avoid model sensitivity problems for very large or too small data (Fig. 1a). This preprocessing step is important in reducing the inaccuracy of the model in making predictions [20,21]. The Min-Max scaler normalization obeys Eq. (1),

$$X_{scaled} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

where X_{min} and X_{max} are minimum and maximum values of any feature X. The most extreme value for each component is changed to 1, the base estimation of that element is changed to 0, and every other value is altered to a decimal between 0 and 1.

2.3. Model development

The QSPR model is developed through a comparison of 15 linear and 17 non-linear QSPR models using a supervised ML approach. After an initial trial of the diazine dataset, all linear models show high statistical errors and poor input-output correlation for all features, while non-linear models produce better predictive performance than their linear counterparts do. In their works of comparison between linear and non-linear models, Ser et al. [22], Quadri[a] et al. [23], and Quadri[b] et al. [24], reported that the latter tends to be immune to multicollinearity in the feature sets as well as generate lower statistical errors. Among non-linear algorithms implemented on the diazine dataset, the GBR, RF, and K-nearest neighbors (KNN) emerge as the best three models based on their R^2 and root mean square error (RMSE) RMSE values (Fig. 1a). These models are then used to evaluate the corrosion inhibition performance of diazine-derived compounds based on the relationship between dependent (chemical properties) and independent CIE variables. All parameters of each ML algorithm are set by default as per the release of the sci-kit-learn 1.2.0 module.

Internal validation is applied using the k-fold cross-validation (k-fold CV) method to determine a stable and optimum ML model through repeated training (iteration) until the smallest statistical error value is obtained. To overcome the problem of variance and bias in ML, the k-fold CV data is divided into 10 folds (k = 10), where in each training iteration, one fold is used as a testing set, and the other folds are used as a training set (Fig. 1b). The choice of fold depends on the size of the data set, k = 5 or 10 is generally used [25].

2.4. Hyperparameter optimization

The Grid Search hyperparameter method is applied to search for parameters that match the characteristics of the model, thereby increasing the prediction accuracy. This approach is the easiest to find ideal parameters because it evaluates every possible combination in the provided discrete parameter space [26,27]. Since there is no certainty that the model is the best for the problem at hand, then model improvement is needed. Even though it plays an important role in determining the most suitable parameter for the best model, the hyperparameter is different for different algorithms.

The hyperparameters of each model and, their range and optimal values are presented in Table 2. The GBR and RF share the first two parameters, namely n-Estimators and Max Depth, and differ in the third parameter, Learning Rate for the former and Min Sample Leaf for the latter. The hyperparameters of the KNN are totally different from those of previous algorithms. They are n-Neighbors, Weights, and Leaf Size. The dimension spaces, used range values, and optimal values of the hyperparameters of each algorithm are listed in columns 3, 4, and 5 of Table 2. As mentioned in Section 2.1, as many as 11 features are extracted from the DFT calculation, and due to their positive correlation to the target, all of them are used in the hyperparameter optimization process for each model.

2.5. Assessment metrics

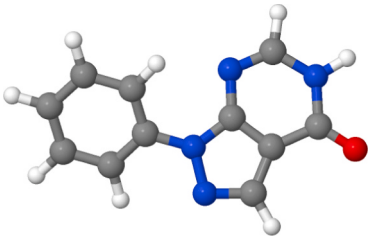
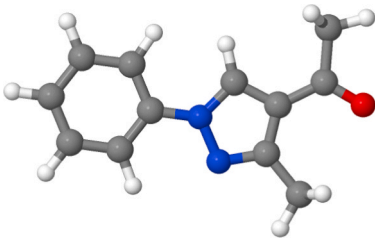
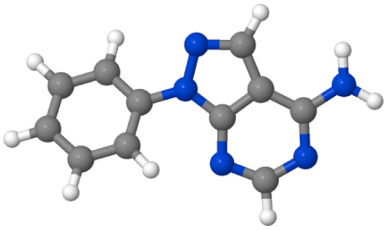
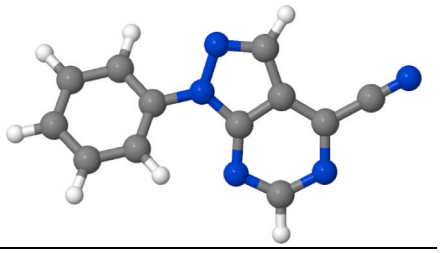
The QSPR model is tested using two strong metrics, namely R^2 and RMSE. These metrics are expressed in Eqs. (2) to (3), respectively,

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

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_i' - Y_i)^2} \tag{3}$$

where Y_i, $\overline{Y}_i$, and Y_i' are actual, mean of actual, and predicted values,

Table 3
Structure of novel diazine molecules (gray, white, blue, and red represent C, H, N, and O atom colors, respectively).

Inhibitor code	Molecule	Structure
P1	1-Phenyl-1 H-pyrazolo [3,4-*d*]-pyrimidin-4-ol	
P2	1-(3-methyl-1-phenyl-1 H-pyrazol-4-yl)-ethenone	
P3	1-Phenyl-1 h-pyrazolo [3,4-*d*]-pyrimidin-4-amine	
P4	1-Phenyl-1 h-pyrazolo [3,4-*d*]-pyrimidine-4-carbonitrile	

respectively. The R^2 measures the suitability of the model, when its value is close to one, it indicates a good fit. The RMSE measures the deviation between the predicted and the actual values [28,29]. The smaller the RMSE, the smaller the prediction error. Such statistical error metrics are used to evaluate the accuracy of the model, the lower the statistical error the better the predictability of the ML model.

2.6. Feature importance

To determine the importance of features in the ML model for predicting CIE, we utilized a feature importance technique that provides insight into the contribution of each feature to the model. However, it is worth noting that the technique does not reveal how a feature directly impacts the dependent (target) variable [30]. Specifically, we employed a permutation technique that involves randomly shuffling or changing feature values to evaluate their impact on model performance. This technique is useful for determining the contribution of each feature to the model's overall performance. Notably, features with a substantial impact on model performance will lead to a higher drop in performance metrics, while those with a lower impact will result in a smaller drop. The feature importance score (FIS) for each feature was calculated by comparing the model's performance on the original dataset to its performance on the permuted random dataset and calculating the difference between the two. The higher the difference, the more significant the feature is.

2.7. Molecular design and prediction

The inhibition performance of a series of 4 new diazine-derived compounds as corrosion inhibitor candidates is predicted using the best model of the ML algorithm through statistical comparison and investigation. The structure of the 4 new diazine-derived compounds is presented in Table 3. The inhibitor compounds are modeled using atomic simulation environment (ASE) software and visualized using Jmol software. The DFT simulation is implemented to calculate the chemical properties using Quantum Espresso software and the Koopmans method. The Koopmans method is generally used to determine molecular properties related to the reactivity and selectivity of a compound. Following Alamri et al. [15], Ser et al. [22], and Geerlings et al. [31], this work uses the DFT and Koopmans method to calculate chemical properties, and the results are presented in Table 4.

To examine the interaction behavior between inhibitor molecules and iron surface, the adsorption energy is also calculated by the DFT using Eq. (4) as.

$$E_{ads} = E_{mol/surf} - (E_{mol} + E_{surf}) \quad (4)$$

where $E_{mol/surf}$, E_{mol}, and E_{surf} correspond to the total energy of the combined system of a molecule and a surface configuration, an isolated molecule, and a surface configuration, respectively.

3. Results and discussions

3.1. Inhibition capability of inhibitors

The current study presents an analysis of the predictive performance of GBR, RF, and KNN models in making CIE predictions for the diazine dataset, as evidenced by Table 5 and Fig. 2. Furthermore, a comparison of prediction performance between the GBR model and other works (PCR, RF, and MLR) on the same dataset is presented in Table 6 and Fig. 3. In addition, we provide a comparison of the current work with relevant literature such as ANN, genetic algorithm (GA)-ANN, autoregressive with exogenous (ARX), multi-layer perceptron neural network (MLPNN), MLR, PLS, and support vector machine (SVM) in Table 7.

Table 5
Calculated metrics of GBR, RF, and KNN algorithms for diazine dataset.

Model	Training		Testing	
	R^2	RMSE	R^2	RMSE
GBR	0.98	1.04	0.97	1.64
RF	0.96	2.17	0.93	2.56
KNN	0.92	2.79	0.89	2.92

Table 4
Chemical properties for novel diazine molecules by DFT calculation.

Inhibitor Code	HOMO	LUMO	**ΔE**	I	A	X	H	Σ	μ	Ω	**ΔN**
P1	-6.29	-1.52	4.77	-3.91	6.29	1.52	3.91	2.39	0.42	3.20	0.19
P2	-6.07	-1.46	4.61	-3.77	6.07	1.46	3.77	2.31	0.43	3.07	0.23
P3	-5.92	-1.33	4.59	-3.63	5.92	1.33	3.63	2.30	0.44	2.86	0.26
P4	-5.81	-1.29	4.52	-3.55	5.81	1.29	3.55	2.26	0.44	2.79	0.28

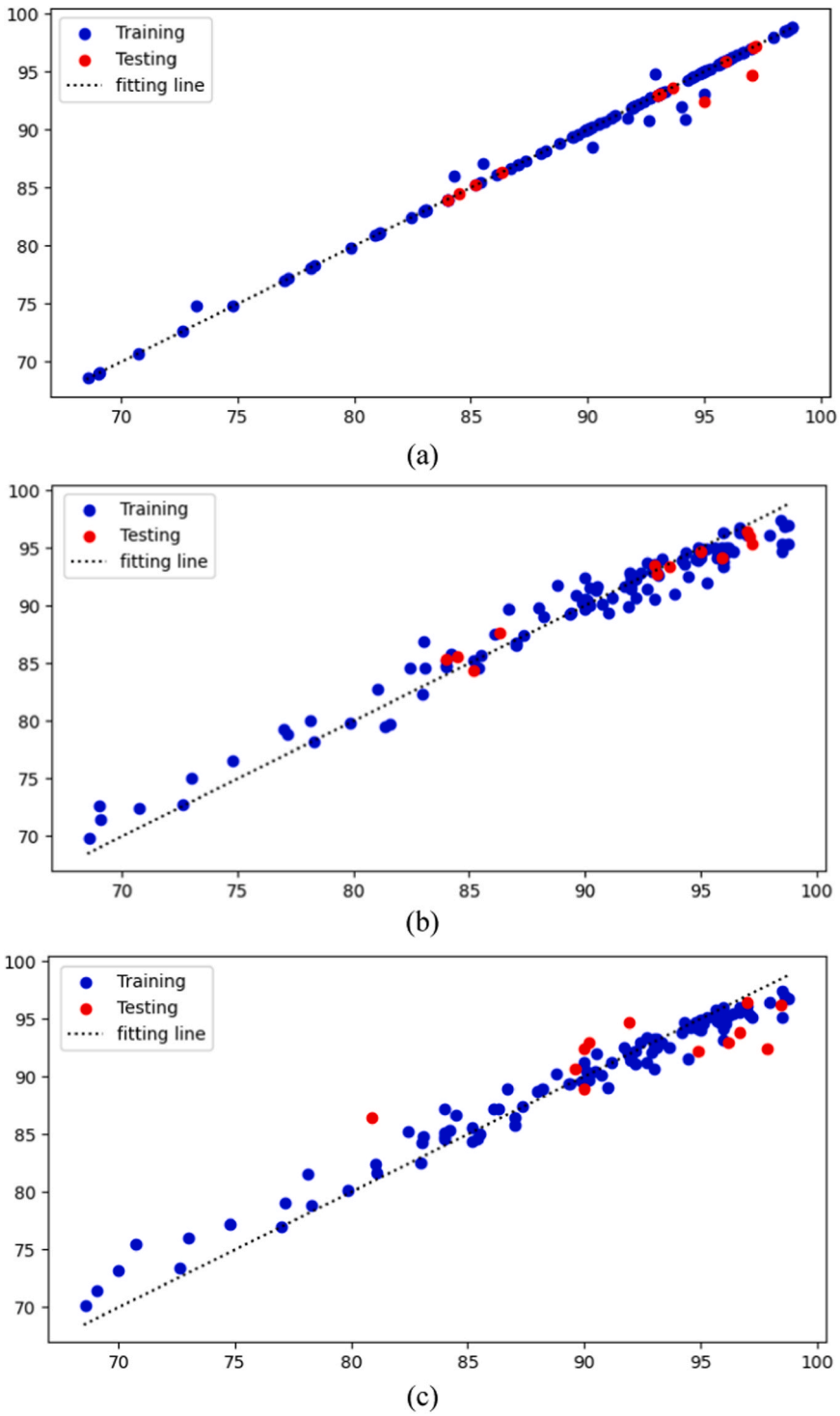


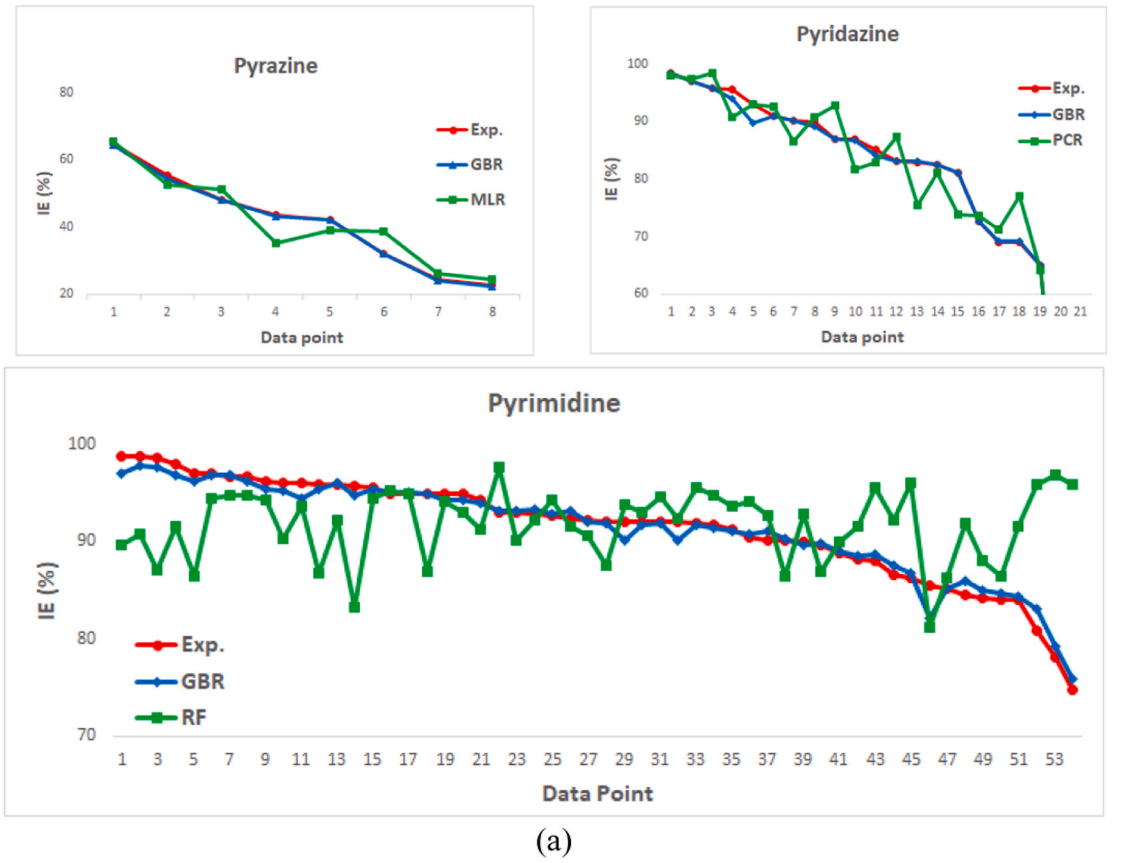
Fig. 2. Scatter plots of prediction data produced by the (a) GBR, (b) RF, and (c) KNN models.

Table 6
Comparison between metrics of the present work (GBR) and those of other works (PCR, RF, and MLR), each for the same dataset.

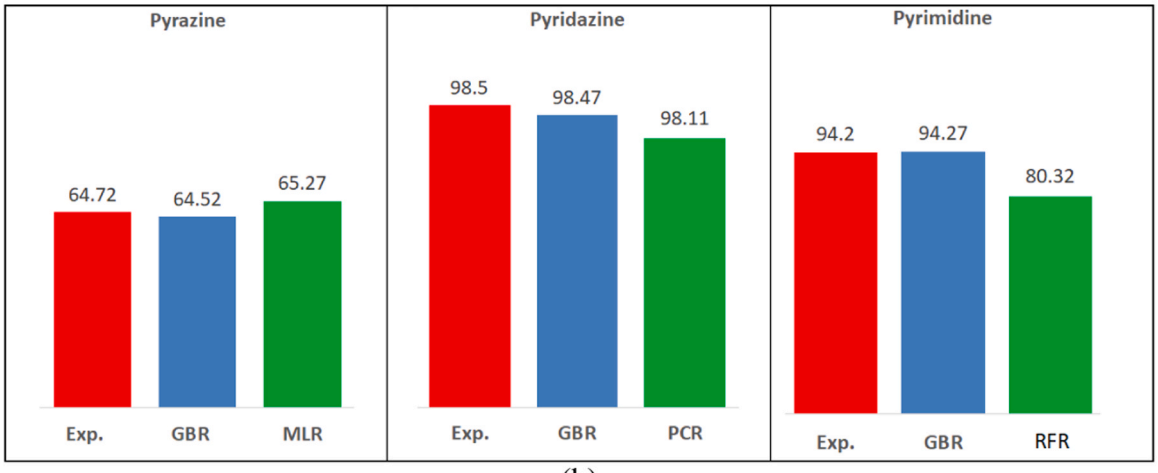
Dataset		Other work			Present work	
Compound	Number	Model	R^2	MSE	R^2	MSE
Pyridazine [14]	21	PCR	0.920	NA	0.989	1.23
Pyrimidine [15]	54	RF	NA	32.60	0.971	0.98
Pyrazine [17]	8	MLR	0.903	NA	0.990	1.21

Notably, Table 8 presents the CIE predicted value of GBR, as well as the adsorption energy of DFT calculation for diazine derivatives. Finally, the optimized molecule adsorption structures of molecules on the iron surface are presented in Fig. 4.

Based on our results, the GBR algorithm exhibits superior predictive performance, as demonstrated by the highest R^2 values and the smallest RMSE during both the training and testing stages (Table 5) for the diazine dataset. These findings suggest that the GBR algorithm provides a good fit and that the selection of $k = 10$ for 120 data points effectively reduces bias in training and variance in testing. As noted by Yuan et al.



(a)



(b)

Fig. 3. Plots of (a) experimental and predicted CIEs for Pyrazine, Pyridazine, and Pyrimidine datasets, and (b) the best CIE GBR-MLR, GBR-PCR, and GBR-RF prediction comparisons relative to the experimental value.

[25], the optimal choice of k values is contingent upon dataset size; for their dataset of 90 data points, they found that k = 5 was optimal. The small error values are further substantiated by the distribution of predicted data points generated by the GBR, RF, and KNN models along the fitting line (Fig. 2). Notably, the GBR-generated predictions are in closer proximity to the fitting line (actual data) compared to those of the RF and KNN models. Therefore, we assert that the GBR algorithm demonstrates the most favorable predictive performance, and thus presents the optimal model for evaluating new diazine compounds as corrosion inhibitors.

It is interesting to check whether the GBR is also the best model for each isomer dataset. Table 6 shows the prediction performance of the GBR versus PCR for pyridazine, RF for pyrimidine, and MLR for pyrazine datasets. Based on the R^2 or MSE values, the GBR is proven to be the best among the previous models. This result is consistent with the prediction data and the best CIE values that are the closest to the experimental values presented in Fig. 3a and Fig. 3b, respectively [14,15,17]. The finding that the GBR is the best model to investigate new diazine derivatives, is supported by its superiority in predicting the CIE of diazine as well as each isomer dataset.

The literature shows that various ML algorithms such as ANN, GA-ANN, ARX, MLPNN, MLR, PLS, and SVM, have been implemented to investigate various compounds as corrosion inhibitors [13,16,22–24,32,33–36]. Table 7 presents the kind and number of inhibitors, and

Table 7
The RMSE of the present work among those of similar works in the literature.

Model	Dataset	Number of compounds	RMSE	Ref.
ANN	Pyridazine	20	10.56	[13]
GA-ANN	Pyridine-Quinoline	41	8.83	[22]
ANN	Pyrimidine	40	2.91	[16]
MLPNN	Ionic liquid	30	5.47	[23]
ANN	Quinoxaline	40	5.42	[24]
ANN	Amino acid	17	8.00	[32]
ARX	Drug	250	7.03	[33]
MLR	Benzimidazole-Imidazole	34	4.58	[34]
PLS	Thiadiazol	24	2.51	[35]
SVM	Amino acid	19	1.48	[36]
GBR	Diazine	120	1.04	Present work

Table 8
The CIE of novel diazine molecules by GBR and its adsorption energy calculation by DFT.

Inhibitor Code	CIE (%)	Adsorption Energy (eV)
P1	86.87	-4.41
P2	85.02	-4.24
P3	90.67	-4.47
P4	94.99	-6.09

associated RMSE values of various algorithms and compounds found in the literature, accompanied by the result of the present work. The table shows that the present study on the modeling of diazine derivative compounds for corrosion inhibition using GBR has yielded highly promising results.

Due to its convincing performance, the GBR model is implemented to investigate four new diazine derivatives as corrosion inhibitor candidates based on their chemical properties. The approximate protective capacity of the new inhibitor compounds is predicted in the range from 85.02 % to 94.99 % (Table 8). The results suggest that these compounds are theoretically predicted to be excellent corrosion inhibitors. As mentioned above, since corrosion inhibition is related to the interaction of the inhibitor with the metal surface, the adsorption energy of the molecules is determined using DFT calculations. Table 8 shows that the inhibitor molecules are adsorbed with quite strong adsorption energy on the iron surface. The calculated adsorption energies range from − 4.41 eV to − 6.09 eV with the P2 molecule having the smallest and the P4 molecule having the highest adsorption energy and CIE. It is good to note that the result of adsorption energy calculations shows conformity with the GBR prediction of CIEs, obeying the theory that corrosion inhibition is related to the ability to interact between inhibitor molecules and metal surfaces. Inhibitors have the potential to adsorb onto metal surfaces, thereby impeding the corrosive attack on the metal surfaces. This mechanism ultimately inhibits the oxygen reduction reaction, which is a key process involved in the corrosion of metals [37,38]. The inhibition capacity of an inhibitor is directly proportional to the strength of its adsorption onto the metal surface, as reported in previous studies [39,40]. Furthermore, our analysis of important molecular features also supports this notion. For instance, we observed that the P4 molecule exhibited the lowest energy gap (as shown in Table 4), suggesting that it has a greater tendency to form strong bonds with the metal surface, thereby improving its inhibition ability.

From the top and side views of the optimized adsorption geometries of molecules (P1, P2, P3, and P4) on the iron surface related to the calculation of the adsorption energy, it is observed that all four molecular structures stick to the iron surface (Fig. 4). It exhibits strong chemisorption and seems to form an adsorbed layer on the iron surface. All the tested diazine derivatives are adsorbed on the iron surface via the aromatic rings and heteroatom groups (N and O). Several organic compounds containing N, P, O, and S atoms have already been tested in various environmental conditions for corrosion protection [41]. Kozlica et al. also mentioned that organic inhibitors possess a functional group with heteroatoms (N, P, O, or S) containing lone pair electrons, and/or delocalized π electrons due to the presence of multiple bonds or aromatic rings [42]. The presence of aromatic rings and heteroatom groups in the molecular structure of the compound makes this inhibitor molecule effective as a corrosion inhibitor. This configuration allows the diazine-derived compounds to donate electrons to the vacant d-orbitals on the iron surface to form a stable bond. Therefore, the studied diazine derivatives tend to be chemically adsorbed on the iron surface. The effectiveness of inhibitor molecules depends on their ability to form an adsorbed layer on the metal surface to prevent charge and mass transfers, thus protecting the metal from corrosive environments [43,44].

3.2. Feature importance analysis

The performance of corrosion inhibition can be affected by many complex factors, including the chemical properties of the inhibiting molecules, therefore, the relationship of chemical properties to CIE is considered. Based on the feature importance analysis (Fig. 5), the energy gap has a large absolute value of the FIS > 0.7. This shows that the energy gap is the chemical property that has the highest correlation with CIE. Nonetheless, other descriptors also show a positive correlation with CIE, therefore feature selection is not needed in this study.

The fact that the energy gap has the highest correlation values, is in line with a theory regarding the CIE of inhibitor molecules. The energy gap reflects the ability of the inhibitor molecule to bind to the metal surface, where a lower energy gap indicates that the molecule tends to easily release electrons to interact with the metal surface [45]. The diazine derivative molecules have characteristics and tendencies as electron donors in interacting with metal surfaces. These characteristics support the formation of adsorption bonds on metal surfaces. The formation of the adsorbed layer can be facilitated by the donor-acceptor ability of the lone electron pair of the heteroatom group of the inhibitor molecule with the empty d-orbitals of the metal surface atoms and/or the interaction between the π-electrons of the aromatic ring of the inhibitor molecule and the empty d-orbitals of the metal surface atoms [46]. The active group or functional group (such as heteroatoms) in the inhibitor molecule donates free electrons to the empty d-orbital on the metal surface, while the π-orbital on the aromatic ring of the inhibitor molecule accepts π-electrons from the empty d-orbital of the metal [47].

4. Conclusions

This study successfully employs a combination of ML and DFT methods to develop a QSPR predictive model for the effective design of diazine corrosion inhibitors. A dataset comprising three diazine isomers (pyridazine, pyrimidine, and pyrazine) is defined to represent the diazine compounds. Our search for the best predictive model for the diazine dataset reveals that out of the 15 linear and 17 non-linear algorithms tested, the GBR model emerges as the most accurate predictor for diazine compounds, based on the R^2 and RMSE metrics. This superiority of the GBR model is also observed when applied to each diazine isomer dataset.

Consequently, the GBR model is used to investigate the potential of four additional diazine derivatives (1-Phenyl-1 H-pyrazolo[3,4-*d*]-pyrimidine-4-ol, 1-(3-methyl-1-phenyl-1 H-pyrazole-4-yl)-ethenone, 1-Phenyl-1 h-pyrazolo[3,4-*d*]-pyrimidine-4-amine, and 1-Phenyl-1 h-pyrazolo[3,4-*d*]-pyrimidine-4-carbonitrile) as corrosion inhibitors. The results show that these derivatives are well-predicted inhibitors, with high CIEs ranging from 85.02 % to 94.99 %. DFT calculations further reveal that these derivatives have strong adsorption energies (ranging from −4.41 eV to −6.09 eV), which aligns with the trend observed in the CIE

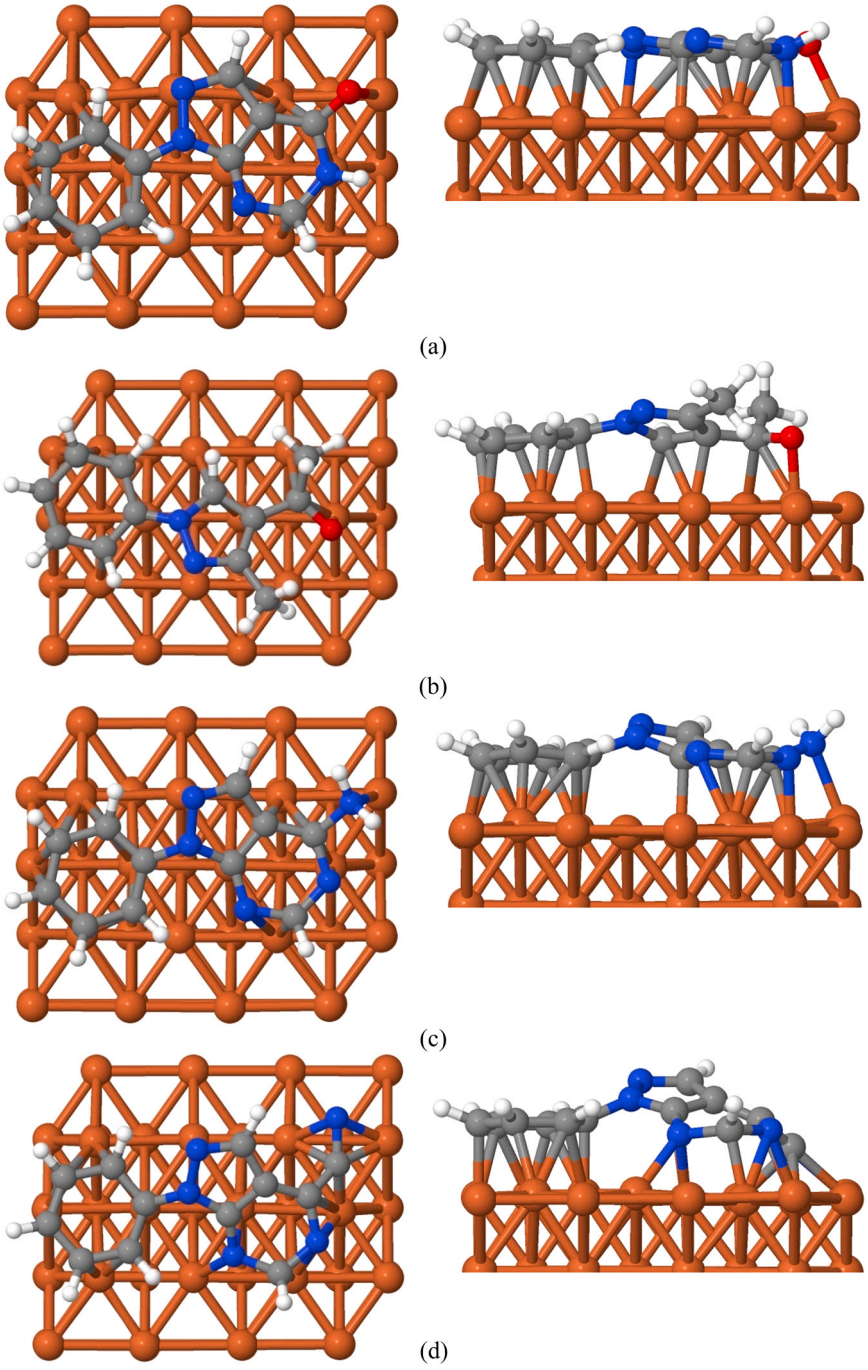


Fig. 4. Adsorption geometry (top and side view) of (a) P1, (b) P2, (c) P3, and (d) P4 molecules on the iron surface.

predictions made using the ML model. Both the ML and DFT methods support the conclusion that the 1-Phenyl-1 h-pyrazolo[3,4-*d*]-pyrimidine-4-carbonitrile molecule shows the best inhibition performance, and the adsorption energy strongly correlates with the CIE of diazine compounds.

The combination of ML and DFT methods provides a complementary approach for future work, particularly in updating the dataset of compounds. As new compounds are discovered through ML, DFT calculations can help update their dataset. This approach can also be extended to other compound classes beyond diazine derivatives.

CRediT authorship contribution statement

Muhamad Akrom: Writing – original draft, Performed DFT and machine learning investigation. **Supriadi Rustad:** Conceptualization, Writing – review & editing, Supervision. **Adhitya Gandaryus Saputro:** Review, Supervision. **Aditianto Ramelan:** Supervision. **Fadjar Fathurrahman:** Review. **Hermawan Kresno Dipojono:** Review, Supervision.

Declaration of Competing Interest

All 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.

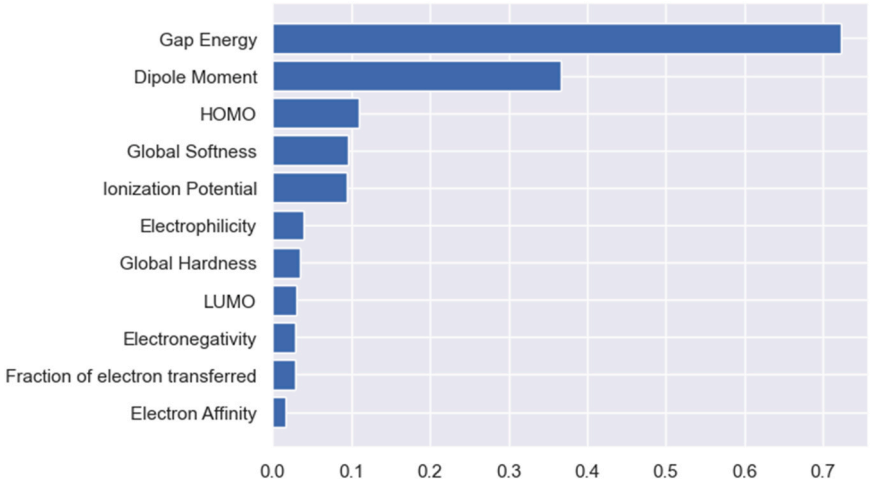


Fig. 5. The FIS of diazine derivatives by GBR.

Data availability

Data and code will be made available on request.

Acknowledgements

All calculations were performed using Computation Facility at the Research Center for Materials Informatics, Universitas Dian Nuswantoro, and the High-Performance Computer facility at the Research Center for Nanosciences and Nanotechnology, Bandung Institute of Technology. This work is funded by BRIN Indonesia through the "RIIM G3″ program. HKD acknowledges funding support by DIKTIRISTEK through the "PFR 2023" program. MA acknowledges scholarship support from LPDP Indonesia.

References

[1] D. Lin-Vien, N.B. Colthup, W.G. Fateley, J.G. Grasselli, Aromatic and heteroaromatic rings. in: The Handbook of Infrared and Raman Characteristic Frequencies of Organic Molecules, Elsevier, 1991, pp. 277–306, https://doi.org/10.1016/b978-0-08-057116-4.50023-7.

[2] A. Hassan, M.S. Numin, K. Jumbri, K.E. Kee, N. Borhan, Review on the recent development of fatty hydrazide as corrosion inhibitor in acidic medium: experimental and theoretical approaches, Metals 12 (7) (2022), https://doi.org/10.3390/met12071058.

[3] K. Rasheeda, V.D.P. Alva, P.A. Krishnaprasad, S. Samshuddin, Pyrimidine derivatives as potential corrosion inhibitors for steel in acid medium – an overview, Int. J. Corros. Scale Inhib. 7 (1) (2018) 48–61, https://doi.org/10.17675/2305-6894-2018-7-1-5.

[4] S.A. Umoren, M.T. Abdullahi, M.M. Solomon, An overview on the use of corrosion inhibitors for the corrosion control of Mg and its alloys in diverse media, J. Mater. Res. Technol. 20 (2022) 2060–2093, https://doi.org/10.1016/j.jmrt.2022.08.021.

[5] P. Hilgard , R.D. Thornes+, Anticoagulants in the Treatment of Cancer, Pergamon Press, 1976.

[6] L.D. Wise et al., A Series of Novel Potential Antipsychotic Agents, 1987.

[7] J.B. Jiang, D.P. Hesson, B.A. Dusak, D.L. Dexter, G.J. Kang, E. Hamel, Synthesis and biological evaluation of 2-styrylquinazolin-4(3H)-ones, A New Class of Antimitotic Anticancer Agents Which Inhibit Tubulin Polymerization, 1990.

[8] A. Gürsoy, S. Qeref Demirayak, G. Çapan, K. Erol, K. Vural, Synthesis and Preliminary Evaluation of New 5-Pyrazolinone Derivatives as Analgesic Agents, 2000.

[9] E.-S.A.M. Badaweya , I.M. El-Ashmaweyb, Nonsteroidal antiinflammatory agents-Part 1: Antiinflammatory, Analgesic and Antipyretic Activity of Some New 1-(pyrimidin-2-yl)-3-Pyrazolin-5ones and 2-(pyrimidin-2-yl)-l,2,4,5,6,7-Hexahydro-3H-indazol-3-ones, 1998.

[10] A.M. Gilbert, et al., Pyrazolidine-3,5-diones and 5-hydroxy-1H-pyrazol-3(2H)-ones, inhibitors of UDP-N-acetylenolpyruvyl glucosamine reductases, J. Med Chem. 49 (20) (2006) 6027–6036, https://doi.org/10.1021/jm060499t.

[11] G. Daidone et al., Antimicrobial anD Antineoplastic Activities of New 4-diazopyrazole Derivatives, 1998.

[12] N.J. Thumar, M.P. Patel, Synthesis, characterization, and antimicrobial evaluation of carbostyril derivatives of 1H-pyrazole, Saudi Pharm. J. 19 (2) (2011) 75–83, https://doi.org/10.1016/j.jsps.2011.01.005.

[13] T.W. Quadri, et al., Development of QSAR-based (MLR/ANN) predictive models for effective design of pyridazine corrosion inhibitors, Mater. Today Commun. 30 (2022), https://doi.org/10.1016/j.mtcomm.2022.103163.

[14] E.H. El Assiri, et al., Development and validation of QSPR models for corrosion inhibition of carbon steel by some pyridazine derivatives in acidic medium, Heliyon 6 (10) (2020), https://doi.org/10.1016/j.heliyon.2020.e05067.

[15] A.H. Alamri, N. Alhazmi, Development of data driven machine learning models for the prediction and design of pyrimidine corrosion inhibitors, J. Saudi Chem. Soc. 26 (6) (2022), 101536, https://doi.org/10.1016/j.jscs.2022.101536.

[16] T.W. Quadri, et al., Predicting protection capacities of pyrimidine-based corrosion inhibitors for mild steel/HCl interface using linear and nonlinear QSPR models, J. Mol. Model 28 (9) (2022), https://doi.org/10.1007/s00894-022-05245-1.

[17] I.B. Obot, S.A. Umoren, Experimental, DFT and QSAR models for the discovery of new pyrazines corrosion inhibitors for steel in oilfield acidizing environment, Int J. Electrochem Sci. 15 (9) (2020) 9066–9080, https://doi.org/10.20964/2020.09.72.

[18] R.L. Camacho-Mendoza, L. Feria, L.Á. Zárate-Hernández, J.G. Alvarado-Rodríguez, J. Cruz-Borbolla, New QSPR model for prediction of corrosion inhibition using conceptual density functional theory, J. Mol. Model 28 (8) (2022), https://doi.org/10.1007/s00894-022-05240-6.

[19] C. Beltran-Perez, et al., A general use QSAR-ARX model to predict the corrosion inhibition efficiency of drugs in terms of quantum mechanical descriptors and experimental comparison for lidocaine, Int J. Mol. Sci. 23 (9) (2022), https://doi.org/10.3390/ijms23095086.

[20] M. Ahsan, M. Mahmud, P. Saha, K. Gupta, Z. Siddique, Effect oF Data Scaling Methods on Machine Learning Algorithms and Model Performance, Technolgies 9 (3) (2021) 52, https://doi.org/10.3390/technologies9030052.

[21] T. Sutojo, S. Rustad, M. Akrom, A. Syukur, G.F. Shidik, H.K. Dipojono, A machine learning approach for corrosion small datasets, Npj Mater. Degrad. 7 (1) (2023), https://doi.org/10.1038/s41529-023-00336-7.

[22] C.T. Ser, P. Žuvela, M.W. Wong, Prediction of corrosion inhibition efficiency of pyridines and quinolines on an iron surface using machine learning-powered quantitative structure-property relationships, Appl. Surf. Sci. 512 (2020), https://doi.org/10.1016/j.apsusc.2020.145612.

[23] T.W. Quadri, et al., Multilayer perceptron neural network-based QSAR models for the assessment and prediction of corrosion inhibition performances of ionic liquids, Comput. Mater. Sci. 214 (2022), https://doi.org/10.1016/j.commatsci.2022.111753.

[24] T.W. Quadri, et al., Computational insights into quinoxaline-based corrosion inhibitors of steel in HCl: quantum chemical analysis and QSPR-ANN studies, Arab. J. Chem. 15 (7) (2022), https://doi.org/10.1016/j.arabjc.2022.103870.

[25] X. Yuan, et al., Applied machine learning for prediction of CO2 adsorption on biomass waste-derived porous carbons, Environ. Sci. Technol. 55 (17) (2021) 11925–11936, https://doi.org/10.1021/acs.est.1c01849.

[26] L. Yang, A. Shami, On hyperparameter optimization of machine learning algorithms: theory and practice, Neurocomputing 415 (2020) 295–316, https://doi.org/10.1016/j.neucom.2020.07.061.

[27] Z.M. Alhakeem, Y.M. Jebur, S.N. Henedy, H. Imran, L.F.A. Bernardo, H.M. Hussein, Prediction of ecofriendly concrete compressive strength using gradient boosting regression tree combined with GridSearchCV hyperparameter-optimization techniques, Materials 15 (21) (2022), https://doi.org/10.3390/ma15217432.

[28] Y. jun Lv, et al., Steel corrosion prediction based on support vector machines, Chaos Solitons Fractals 136 (2020), https://doi.org/10.1016/j.chaos.2020.109807.

[29] T.H. Nguyen, T.L. Chau, T. Hoang, T. Nguyen, Developing artificial neural network models to predict corrosion of reinforcement in mechanically stabilized earth walls, Neural Comput. Appl. (2022), https://doi.org/10.1007/s00521-022-08043-1.

[30] A. Altmann, L. Toloşi, O. Sander, T. Lengauer, Permutation importance: a corrected feature importance measure, Bioinformatics 26 (10) (2010) 1340–1347, https://doi.org/10.1093/bioinformatics/btq134.

[31] P. Geerlings, F. De Proft, W. Langenaeker, Conceptual density functional theory, Chem. Rev. 103 (5) (2003) 1793–1873, https://doi.org/10.1021/cr990029p.
[32] B. El Ibrahimi, A. Jmiai, L. Bazzi, S. El Issami, Amino acids and their derivatives as corrosion inhibitors for metals and alloys, Arab. J. Chem. 13 (1) (2020) 740–771, https://doi.org/10.1016/j.arabjc.2017.07.013.
[33] C. Beltran-Perez, et al., A general use QSAR-ARX model to predict the corrosion inhibition efficiency of drugs in terms of quantum mechanical descriptors and experimental comparison for lidocaine, Int. J. Mol. Sci. 23 (9) (2022), https://doi.org/10.3390/ijms23095086.
[34] M.H. Keshavarz, K. Esmaeilpour, A.N. Golikand, Z. Shirazi, Simple approach to predict corrosion inhibition efficiency of imidazole and benzimidazole derivatives as well as linear organic compounds containing several polar functional groups, Z. Anorg. Allg. Chem. 642 (16) (2016) 906–913, https://doi.org/10.1002/zaac.201600230.
[35] K. Sadik, S. Byadi, M.E. Hachim, N. El Hamdani, Podlipnik, A. Aboulmouhajir, Multi-QSAR approaches for investigating the relationship between chemical structure descriptors of Thiadiazole derivatives and their corrosion inhibition performance, J. Mol. Struct. 1240 (2021), https://doi.org/10.1016/j.molstruc.2021.130571.
[36] H. Zhao, X. Zhang, L. Ji, H. Hu, Q. Li, Quantitative structure-activity relationship model for amino acids as corrosion inhibitors based on the support vector machine and molecular design, Corros. Sci. 83 (2014) 261–271, https://doi.org/10.1016/j.corsci.2014.02.023.
[37] M. Akrom, et al., DFT and microkinetic investigation of oxygen reduction reaction on corrosion inhibition mechanism of iron surface by Syzygium aromaticum extract, Appl. Surf. Sci. 615 (2023), https://doi.org/10.1016/j.apsusc.2022.156319.
[38] N. Arrousse, et al., The inhibition behavior of two pyrimidine-pyrazole derivatives against corrosion in hydrochloric solution: experimental, surface analysis and in silico approach studies, Arab. J. Chem. 13 (7) (2020) 5949–5965, https://doi.org/10.1016/j.arabjc.2020.04.030.
[39] M.S.S. Carranza, Y.I.A. Reyes, E.C. Gonzales, D.P. Arcon, F.C. Franco, Electrochemical and quantum mechanical investigation of various small molecule organic compounds as corrosion inhibitors in mild steel, Heliyon 7 (9) (2021), https://doi.org/10.1016/j.heliyon.2021.e07952.
[40] A. Kokalj, Corrosion inhibitors: physisorbed or chemisorbed? Corros. Sci. 196 (2022) https://doi.org/10.1016/j.corsci.2021.109939.
[41] D. Kumar, V. Jain, B. Rai, Capturing the synergistic effects between corrosion inhibitor molecules using density functional theory and ReaxFF simulations - a case for benzyl azide and butyn-1-ol on Cu surface, Corros. Sci. 195 (2022), https://doi.org/10.1016/j.corsci.2021.109960.
[42] D.K. Kozlica, A. Kokalj, I. Milošev, Synergistic effect of 2-mercaptobenzimidazole and octylphosphonic acid as corrosion inhibitors for copper and aluminium – an electrochemical, XPS, FTIR and DFT study, Corros. Sci. 182 (2021), 109082, https://doi.org/10.1016/J.CORSCI.2020.109082.
[43] A. Dehghani, A.H. Mostafatabar, G. Bahlakeh, B. Ramezanzadeh, A detailed study on the synergistic corrosion inhibition impact of the Quercetin molecules and trivalent europium salt on mild steel; electrochemical/surface studies, DFT modeling, and MC/MD computer simulation, J. Mol. Liq. 316 (2020), https://doi.org/10.1016/j.molliq.2020.113914.
[44] A. Thakur, S. Kaya, A.S. Abousalem, A. Kumar, Experimental, DFT and MC simulation analysis of Vicia sativa weed aerial extract as sustainable and eco-benign corrosion inhibitor for mild steel in acidic environment, Sustain Chem. Pharm. 29 (2022), https://doi.org/10.1016/j.scp.2022.100785.
[45] T. Le Minh Pham, T. Khoa Phung, H. Viet Thang, DFT insights into the adsorption mechanism of five-membered aromatic heterocycles containing N, O, or S on Fe(1 1 0) surface, Appl. Surf. Sci. 583 (2022), https://doi.org/10.1016/j.apsusc.2022.152524.
[46] S. Kamal, et al., Synthesis, characterization and DFT studies of water stable Cd(II) metal–organic clusters with better adsorption property towards the organic pollutant in waste water, Inorg. Chim. Acta 512 (2020), https://doi.org/10.1016/j.ica.2020.119872.
[47] E. Ech-chihbi, et al., Computational, MD simulation, SEM/EDX and experimental studies for understanding adsorption of benzimidazole derivatives as corrosion inhibitors in 1.0 M HCl solution, J. Alloy. Compd. 844 (2020), https://doi.org/10.1016/j.jallcom.2020.155842.