



Data-driven investigation to model the corrosion inhibition efficiency of Pyrimidine-Pyrazole hybrid corrosion inhibitors

Muhamad Akrom^{a,b,*}, Supriadi Rustad^{b,*}, Adhitya Gandaryus Saputro^c,
Hermawan Kresno Dipojono^{c,*}

^a Study Program in Informatics Engineering, Faculty of Computer Science, Dian Nuswantoro University, Semarang 50131, Indonesia

^b Research Center for Materials Informatics, Faculty of Computer Science, Dian Nuswantoro University, Semarang 50131, Indonesia

^c Advanced Functional Materials Research Group, Bandung Institute of Technology, Bandung 40132, Indonesia

ARTICLE INFO

Keywords:

Machine learning
Corrosion inhibitor
N-heterocyclic
Pyrimidine-pyrazole hybrid

ABSTRACT

This paper proposes a quantitative structure–property relationship model (QSPR) based on machine learning (ML) for a pyrimidine-pyrazole hybrid as a corrosion inhibitor. Based on the metric values of the coefficient of determination (R^2) and root mean square error (RMSE), the extreme gradient boosting (XGBoost) model was found to be the best predictive model for the N-heterocyclic dataset and its respective non-aggregated dataset. When the XGBoost model was applied to three additional pyrimidine-pyrazole hybrid derivatives, this consistency was also seen, and high corrosion inhibition efficiency (CIE) values were obtained ranging from 82.09% to 95.26%. According to the CIE trends found from the ML predictions, DFT calculations for these derivatives also reveal a strong and suitable adsorption energy trend ranging from -1.40 to -1.52 eV. Also supported by the compatibility of the energy gap trend with the CIE trend of the inhibitor molecule. This innovative method can elucidate the characteristics of potential organic corrosion inhibitors before conducting experimental research, which can speed up the preparation of fresh and strong organic corrosion inhibitors.

1. Introduction

To defend against corrosion attacks on metals, corrosion inhibitors based on organic molecules have been intensively researched. They have drawn a lot of interest because of their non-toxicity, environmental friendliness, affordability, ease of production, and high corrosion inhibition efficiency (CIE) [1,2]. Organic compounds containing heteroatoms such as nitrogen, oxygen, sulfur, and phosphorus contained in aromatic rings, as well as double and triple bonds are usually effective for use as corrosion inhibitors [3,4]. Many N-heterocyclic aromatic organic compounds such as pyrazole [5], pyrimidine [6], pyridazine [7], pyrazine [8], pyridine [9,10], pyrrole [11], quinoline [12], quinoxaline [13], and their derivatives have been studied experimentally and published as effective corrosion inhibitors. In an acidic environment, a new pyrimidine-pyrazole hybrid compound, namely 2,5-dihydro-4H-pyrazolo (3,4-d)pyrimidine-4-one (PP1), 2-(((2,3-dihydro-1H-pyrazol-1-yl)methyl)amino) pyrimidine-4,6-diol (PP2), and 2-(bis((1H-pyrazol-1-yl)methyl)amino)pyrimidine-4,6-diol (PP3) have been evaluated experimentally

and demonstrated their effectiveness as corrosion inhibitors in iron [14,15]. The CIE of the iron surface by these three molecules increases with increasing concentration. However, experimental testing is expensive, time-consuming, and requires significant resources [16–19].

Due to the rapid rise of technology and data science via machine learning (ML), notably in the design and discovery of innovative materials, materials informatics has become a priority of study [20]. Because molecular characteristics can be quantified and are directly related to the chemical structure of compounds, ML techniques based on the quantitative structure–property relationship (QSPR) method are frequently used to assess the performance of compounds [21,22]. ML models have frequently used the QSPR. In the field of materials informatics, the QSPR is a quick, dependable, and inexpensive approach [23,24]. Finding the optimal model with high predictive accuracy in the development of ML is very important, especially for testing the potential of new organic compounds as corrosion inhibitors.

Extreme gradient boosting (XGBoost) [25], an ensemble model based on decision tree methods used in artificial intelligence, is growing in

* Corresponding authors at: Research Center for Materials Informatics, Faculty of Computer Science, Dian Nuswantoro University, Semarang 50131, Indonesia (M. Akrom, S. Rustad); Advanced Functional Materials Research Group, Bandung Institute of Technology, Bandung 40132, Indonesia (H.K. Dipojono).

E-mail addresses: m.akrom@dsn.dinus.ac.id (M. Akrom), srustad@dsn.dinus.ac.id (S. Rustad), dipojono@itb.ac.id (H.K. Dipojono).

<https://doi.org/10.1016/j.comptc.2023.114307>

Received 25 June 2023; Received in revised form 25 August 2023; Accepted 1 September 2023

Available online 3 September 2023

2210-271X/© 2023 Elsevier B.V. All rights reserved.

Table 1
QCDs of novel pyrimidine-pyrazole hybrid molecules by DFT and Koopmans calculations.

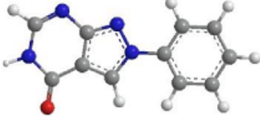
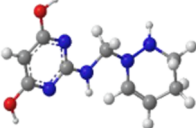

QCDs	Molecule		
	PP1 2,5-dihydro-4H-pyrazolo(3,4-d)pyrimidine-4-one	PP2 2-(((2,3-dihydro-1H-pyrazol-1-yl)methyl)amino) pyrimidine-4,6-diol	PP3 2-(bis((1H-pyrazol-1-yl)methyl) amino)pyrimidine-4,6-diol
			
HOMO	-7.98	-8.45	-8.03
LUMO	-6.76	-5.83	-6.06
ΔE	1.22	2.62	1.97
I	7.98	8.45	8.03
A	6.76	5.83	6.06
χ	4.27	7.03	6.91
η	1.09	1.24	1.19
σ	0.92	0.81	0.84
μ	16.26	13.82	14.18
ΔN	0.81	0.05	0.19

Table 2
Comparison between XGBoost and other models (RF and KNN) for the *N*-heterocyclic constructed dataset.

Model	Training R ²	RMSE	Testing R ²	RMSE
XGBoost	9.998	4.133	0.986	5.060
RF	0.918	6.001	0.897	6.723
KNN	0.745	10.594	0.726	10.961

popularity as the algorithm of choice for machine learning researchers. XGBoost has been used effectively and shows excellent performance in various classification and regression-based studies [26], such as for estimating bridge damage conditions in the Korean bridge management system [27], predicting the evapotranspiration rate of greenhouse tomato plants [28], solar radiation forecasting [29], product sensory characteristics of grape to wine [26], multiparametric MRI texture analysis in patients with invasive breast cancer [30], estimated safety assessment of oil and gas pipelines [31], predicting the internal corrosion rate of oil and gas pipelines [18], and predicting microbiologically induced corrosion of concrete in gutters [32]. XGBoost is an enhanced supervised learning method with excellent performance in solving some complex engineering problems. XGBoost has also shown considerable potential in various data mining investigations in recent years. Among the other advantages of XGboost are: (1) It is flexible, supports objective functions and user-defined evaluation functions, and is scalable and excellent at modeling small and constrained data sets; (2) With the gradient enhancement technique, XGBoost can create a strong classifier based on a weak classifier, and it has been proven that XGBoost successfully classifies tabular data; (3) Parallel processing is supported by XGBoost, so calculations are much faster compared to other algorithms; (4) To reduce model complexity, XGBoost adds a regularization component to reduce model variance; (5) XGBoost's built-in features are designed to handle lost data intelligently and offer ways to prevent overfitting; (6) Support the development of more robust learning models by combining multiple tree models.

The comprehensive implementation of various ML models has not been properly addressed to overcome the accuracy problem in modeling corrosion inhibition. The main challenge in developing ML models is the high accuracy of the model so that the prediction results are relevant to their actual properties. The current study focuses on investigating the corrosion inhibition performance of novel pyrimidine-pyrazole-derived hybrid organic compounds (PP1, PP2, PP3) using the ML-based XGBoost

model. This work helps optimize the performance of inhibitor synthesis prior to experimental analysis to achieve performance effectiveness and efficiency.

2. Method

2.1. Dataset, feature, and target

We constructed a dataset consisting of 218 aromatic *N*-heterocyclic organic compounds compiled from the published literature [16,17,33–38]. A total of 10 quantum chemical descriptors (QCDs), namely HOMO and LUMO energy, energy gap (ΔE), ionization potential (I), electron affinity (A), electronegativity (χ), global hardness (η), global softness (σ), dipole moment (μ), and a fraction of electrons transferred (ΔN) were used as features and CIE as targets. QCDs are very important factors in corrosion inhibition. In ML-based QSPR studies, QCDs are usually calculated using density functional theory (DFT) and Koopmans theory methods, while CIE is obtained from experimental studies.

Among the QCDs used, the features of HOMO and LUMO energies, and dipole moment (μ) are obtained by DFT simulation, while other features such as energy gap (ΔE), ionization potential (I), electron affinity (A), electronegativity (χ), global hardness (η), global softness (σ), and a fraction of electrons transferred (ΔN) are calculated using the Koopmans method [17,39,40] based on the following equation:

$$\Delta E = LUMO - HOMO$$

$$I = -HOMO$$

$$A = -LUMO$$

$$\chi = -\frac{HOMO + LUMO}{2}$$

$$\eta = \frac{LUMO - HOMO}{2}$$

$$\sigma = \frac{1}{\eta}$$

$$\Delta N = \frac{\chi_{Fe} - \chi_{Inh}}{2(\eta_{Fe} + \eta_{Inh})}$$

In this case, the inhibitor molecule has a predicted electronegativity and molecular hardness of χ_{Inh} and η_{Inh} , respectively, while the surface

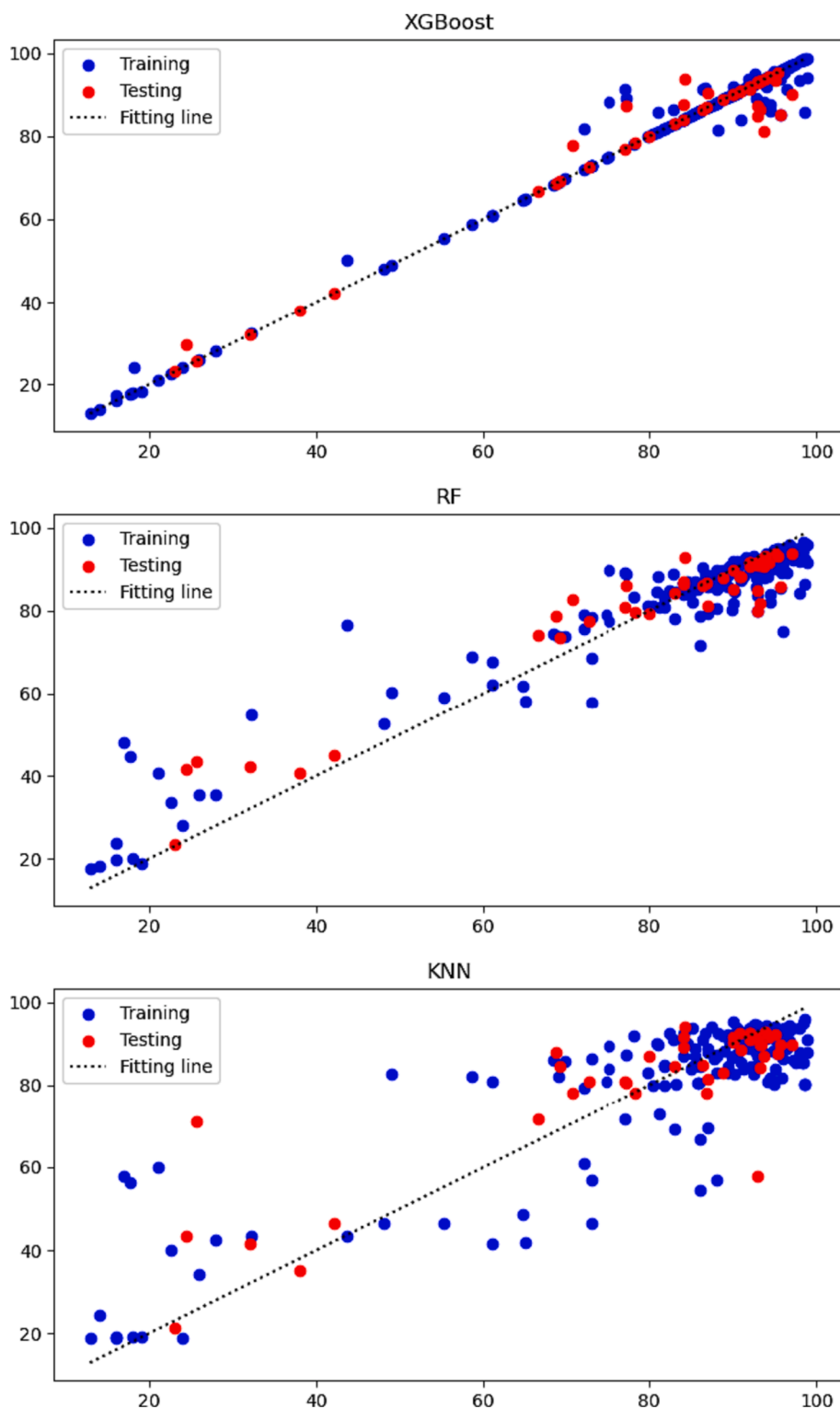


Fig. 1. Scatter plot of prediction data by models.

electronegativity of iron χ_{Fe} is 7 and the molecular hardness of iron η_{Inh} is 0 [17,41].

2.2. Machine learning model development

The *N*-heterocyclic dataset was used to evaluate the predictive performance of the XGBoost model in investigating the CIE of three pyrimidine-pyrazole-derived hybrid organic compounds, they are PP1,

PP2, and PP3. To validate the efficacy of the XGBoost model, we compared it with other models, namely random forest (RF) and k-nearest neighbors (KNN). In addition, validation was also carried out on each dataset that was not combined. To validate the prediction results of XGBoost, the CIE value was also compared with the results of experimental studies.

The robust scaler technique is used as a data normalization tool when working with extremely large or insufficiently sized data to prevent

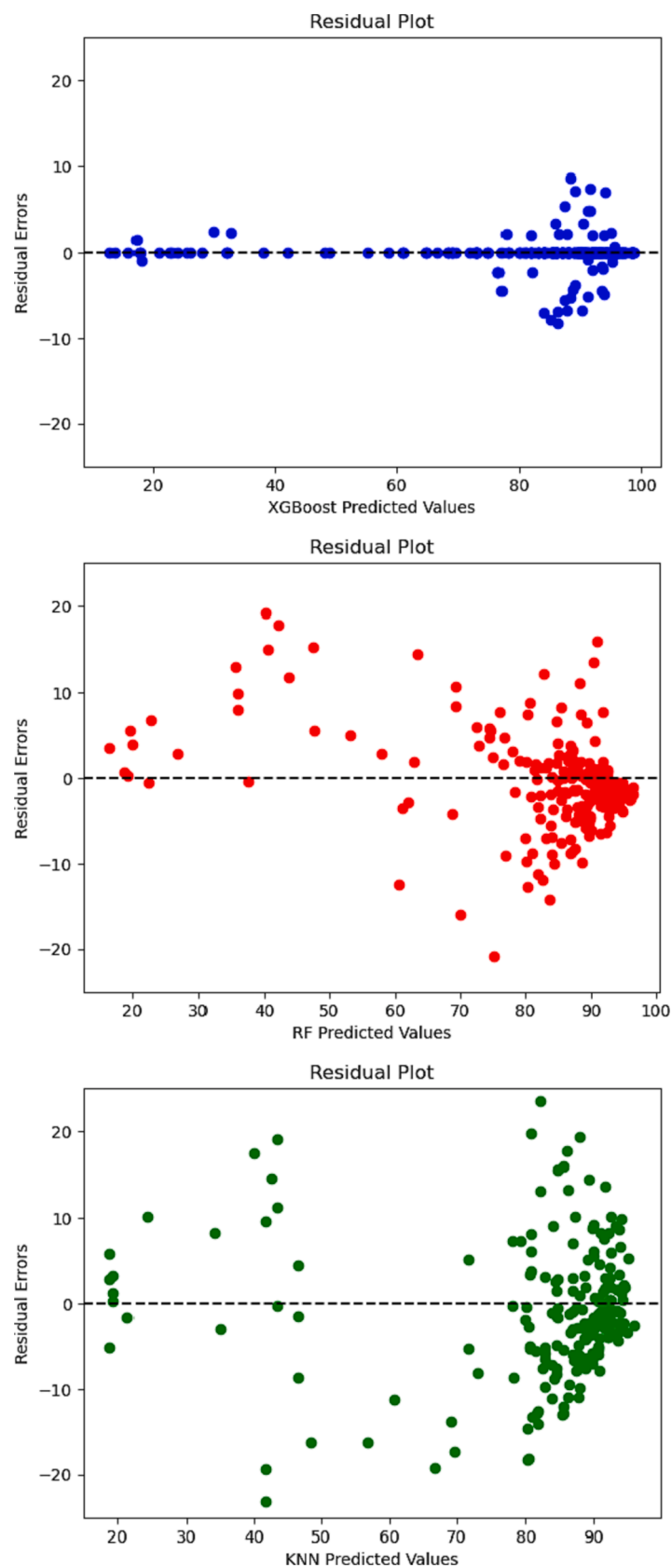


Fig. 2. Residual error plot by models.

Table 3

Comparison between XGBoost and other models for each dataset before merged.

Dataset	XGBoost R ²	RMSE	Other work Model	R ²	RMSE
Pyrimidine	0.991	0.52	ANN [33]	na	2.91
Pyrimidine	0.989	0.49	RF [16]	na	5.71
Pyridazine	0.991	0.88	ANN[34]	na	10.56
Pyridazine	0.991	2.48	ANN [36]	0.920	na
Pyridine-Quinoline	0.745	10.59	ANN [17]	0.726	10.96
Pyridine-Quinoline	0.985	3.20	MLR [37]	0.930	na
Pyrazine	0.993	0.29	MLR [38]	0.903	na
Quinoxaline	0.995	0.76	ANN [35]	na	5.42

Table 4

The CIE comparison of novel pyrimidine-pyrazole hybrid compounds between XGBoost prediction and experimental studies, and its adsorption energy on the iron surface.

Inhibitor	CIE (%) XGBoost	Experimental	E _{ads} (eV)
PP1	95.26	96 [15]	−1.52
PP2	82.09	81 [14]	−1.40
PP3	90.77	92 [14]	−1.48

model sensitivity problems. This preprocessing procedure is essential to lowering model prediction errors [42]. The stability of the model of any ML method is evaluated using the k-fold cross-validation (k-fold CV) approach through repeated training (iteration) until a minimum statistical error value is obtained [43]. Data is divided into 10 folds, with one-fold serving as the test set and the other as the training set in each training iteration to address the variance and bias issues in ML. Depending on the magnitude of the data collection, multiples of k = 5 or 10 are typically utilized [44].

2.2.1. K-Nearest Neighbor (KNN)

The KNN algorithm runs based on the Euclidean distance function (eq. (1)) to determine the labeling of unclassified sample points, determined by the maximum class label at the k nearest neighbor points.

$$d(x, p) = \sqrt{\sum_{i=1}^m (x_i - p_i)^2} \quad (1)$$

$$y = \frac{1}{K} \sum_{i=1}^K y_i \quad (2)$$

with m being the number of descriptors and x and p being the descriptors of the training and testing sets, respectively. To produce the prediction result, the K target values from the training set are averaged (equation (2)), where y is the outcome of the testing prediction and y_i is the ith target.

2.2.2. Random forest (RF)

Random forest uses two stages to get the result. In the first stage, a replacement-based bootstrap technique is used to randomly select several classifier trees using the input features and the training data set to build decision trees. The classifier created in the first stage is used to predict the results in the second stage [45]. So, it is expected that as the forest tree cover increases, the RF accuracy will also increase. The average yield of the forecast results generated from each decision tree is used to calculate the result in the manner described below:

$$y = \frac{1}{T} \sum_{i=1}^T y_i(x') \quad (3)$$

where y is the desired result, y_i represents the ith individual decision tree, T is the total number of trees, and x' stands for unidentified cases.

2.2.3. Xgboost

XGBoost was introduced by Chen and Guestrin [25] within the Gradient Boosting (GB) framework [46], is an improved supervised method. Because of its advantages, mainly great efficiency, and flexibility, the XGBoost approach has recently attracted a lot of attention in handling challenging regression and classification engineering issues. The loss function in XGBoost includes an additional regularization term to the objective function to smooth the final learned weights and prevent overfitting. Additionally, it uses first- and second-order gradient statistics to improve the loss function. To alleviate the overfitting problem, XGBoost also offers row and column sampling. The parallel and distributed computation makes it possible to explore models more quickly. According to Chen and Guestrin [25], the loss function in XGBoost can be expressed as follows:

$$O = \sum_i l(\hat{y}_i, y_i) + \sum_j \varphi(t_j) \quad (4)$$

$$\varphi(t) = \sigma P + \frac{1}{2} \mu \|Q\|^2 \quad (5)$$

A regularization term (φ) is added to the objective function (O) in XGBoost based on a loss function (l) in GB. Here, the predictions and observations are denoted by \hat{y}_i and y_i respectively. A specific tree structure is represented by t. P and Q represent, respectively, the number of leaf nodes and the score at each node. Regularization is controlled by the parameters σ and μ .

2.3. Assessment metrics

Reliable metrics like the coefficient of determination (R²) and root mean square error (RMSE) are used to assess the performance of models. When R², a metric of the model's appropriateness, is close to 1, it implies a good fit. RMSE is used to determine the difference between the actual and anticipated value. The prediction error decreases along with the RMSE. This statistic is used to assess the model's accuracy because a smaller statistical error signifies a more predictable model. [47,48].

$$R^2 = \frac{\sum_{i=1}^n (Y'_i - \bar{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y}_i)^2} \quad (6)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y'_i - Y_i)^2} \quad (7)$$

where Y_i, Y_i', and \bar{Y}_i represent the experimental, predicted, and mean of experimental values, respectively.

2.4. DFT calculation

All models were built using the Atomic Simulation Environment (ASE) [49] and the Chemdraw Ultra [50] software. The Quantum-Espresso software is used to perform spin-polarized calculations [51]. Based on the generalized gradient approximation (GGA) of the Perdew, Burke, and Ernzerhof (PBE) functional, the exchange–correlation functional is employed [52]. Ultrasoft pseudopotentials (Smith) are used to characterize the interactions between the ion core and electron. Grimme's semi-empirical correction is used to take into account van der Waals interactions [53]. A 2 × 2 × 1 k-point mesh is employed as Monkhorst-Pack sampling, with the exception of isolated molecules. For computations involving isolated molecules, a cubic cell measuring 30 × 30 × 30 Å³ is used [54,55]. At 40 Ry and 200 Ry, respectively, the kinetic energy cut-off and electron density cut-off converge. The system's atomic positions are relaxed up to the sum of all forces acting in a given direction is less than 0.05 eV Å^{−1}. The system is optimized until each atom has less than 1 × 10^{−6} eV of total energy. A 15 Å vacuum zone is created on the surface normal direction to stop intercellular

communication. Using bcc unit cells, the optimal geometry for bulk Fe is computed. Our calculated lattice constant, 2.85 Å, agrees well with experimental observations, 2.87 Å, and previous theoretical findings 2.83 Å [56]. The estimated lattice constant builds the Fe(110) surfaces. The 4x4 sheet with 64 Fe atoms was used in four layers with the bottom two layers frozen and the others allowed to be kept free to relax. Because it has the lowest surface energy, the Fe(110) surface was selected [2,57]. Only one molecule for each type of inhibitor is calculated to calculate adsorption energy.

The ability of inhibitor molecules to inhibit corrosion is related to their adsorption behavior on metal surfaces. To evaluate the interaction between the inhibitor molecule and the iron surface, the adsorption energy is also calculated by DFT using the following equation:

$$E_{ads} = E_{Inh-Fe} - (E_{Inh} + E_{Fe}) \quad (8)$$

where E_{Inh-Fe} , E_{Inh} , and E_{Fe} correspond to the total energy of the combined molecule-surface system, isolated molecule, and surface configuration, respectively.

The QCDs of the 3 new hybrid pyrimidine-pyrazole derivatives are presented in Table 1. DFT simulation was implemented to calculate adsorption energy and HOMO-LUMO energy using Quantum Espresso software. The Koopmans method is used to determine QCDs related to the reactivity of a compound.

3. Result and discussion

Table 2, Fig. 1, and Fig. 2 provide evidence that this research analyzes the predictive performance of the XGBoost model validated by other models (RF and KNN) in investigating corrosion inhibition performance for the *N*-heterocyclic compound dataset that has been constructed. In addition, Table 3 offers a comparison of the predictive ability of the XGBoost model with models from other studies for the same dataset before merging. Table 4 presents the CIE values of the three new pyrimidine-pyrazole hybrid compounds predicted by XGBoost and experimental results, as well as their adsorption energy values.

According to our findings, XGBoost has the best predictive performance, as evidenced by the largest R^2 value and lowest RMSE during the training and testing phases on the *N*-heterocyclic data set (Table 2). According to these results, XGBoost offered a good fit, and the choice of $k = 5$ for 218 data points succeeded in reducing bias in training and variation in testing. The predicted distribution of data points (Fig. 1) generated by the XGBoost model along the fit line increasingly favors minimal error values. Compared to the predictions of the RF and KNN models, the estimates generated by XGBoost are closer to fitting lines (actual data points). In addition, the residual error plot (Fig. 2) also shows superior predictive accuracy from XGBoost where the residual error is the smallest when compared to the other two models. If the residual error is minimal, the model can accurately predict the true value of the data; conversely, if the residual error is high, the model cannot do it. We therefore argue that the GBR method offers the best predictive performance and is therefore an ideal model to assess new diazine chemicals as corrosion inhibitors.

It will be interesting to see if XGBoost is also the best model for each unmerged dataset. Table 3 displays the predictive performance of XGBoost for each of the pyridazine, pyrimidine, pyrazine, pyridine-quinoline, and quinoxaline datasets compared to an artificial neural network (ANN), multilinear regression (MLR), and RF models from other work. XGBoost's superiority in predicting CIE was also found in each of these datasets, this supports the conclusion that XGBoost is an ideal model to discover new pyrimidine-pyrazole hybrid derivatives.

Due to its convincing performance, the XGBoost model was then used to evaluate three new pyridine-pyrazole hybrid derivatives as candidates for corrosion inhibitors. According to Table 4, the estimated CIE of the new inhibitor compounds ranged from 82.09 to 95.26%. These results are identical to the results of experimental studies. The findings

imply that this chemical could theoretically function as a superior corrosion inhibitor.

As previously indicated, DFT simulations are used to evaluate the adsorption energies of molecules because corrosion inhibition is related to the interaction of the inhibitor with the metal surface. Table 4 also shows that the inhibitor molecule has strong adsorption energy when interacting with the iron surface. PP2 molecules have the lowest adsorption energy and CIE, while PP1 molecules have the largest. The calculated adsorption energies range from −1.40 to −1.52 eV. There is an agreement between the adsorption energies on CIE of the tested molecules in this study, in accordance with the results of other authors [58–64], this supports the general notion that corrosion inhibition is related to the molecular capacity inhibitor to interact with the metal surface. This relationship does not necessarily show a proportional relationship between CIE values and adsorption energy, there are many other and complex factors that can affect CIE besides adsorption energy, such as chemical and physical properties of inhibitor molecules, metal surface properties, corrosive environments, and other properties.

In addition, based on Table 1, the CIE trend is also related to the molecular gap energy trend. A lower energy gap means that it is easier for the molecule to release (donor) electrons to interact with the metal surface [65]. The lone electron pairs of the heteroatom groups of the molecule and the empty d-orbitals of the metal surface atoms, as well as the interactions between the electrons of the aromatic ring of the molecule and the empty d-orbitals of the metal surface atoms, can act as donors and acceptors to facilitate the formation of the adsorbed layer, respectively [66]. The ability of the inhibitor to be adsorbed on the metal surface can prevent corrosive attacks on the metal surface because the corrosion reaction on the metal surface can be prevented by this mechanism [67,68]. Many organic compounds with N, P, O, and S heteroatomic groups and aromatic rings have undergone corrosion prevention tests in various environmental settings and have shown satisfactory performance [3,4]. The effectiveness of molecular inhibitors depends on their ability to form an adsorbed layer on the metal surface to prevent charge and mass transfer, thereby protecting the metal from corrosive environments [69,70].

4. Conclusion

This study develops an ML-based QSPR predictive model for the design of pyrimidine-pyrazole hybrid corrosion inhibitor compounds. The XGBoost model emerged as the most accurate predictor of *N*-heterocyclic derivatives, according to the R^2 and RMSE metrics. When applied to any unaggregated dataset, the advantages of the XGBoost model are also apparent. Consequently, the XGBoost model was used to investigate the potential of three new pyrimidine-pyrazole hybrid derivatives (2,5-dihydro-4H-pyrazolo[3,4-d] pyrimidine-4-one (PP1), 2-(((2,3 - dihydro-1H-pyrazol-yl) methyl) amino) (PP2), and 2-(bis((1H-pyrazol-1-yl) amino) pyrimidine-4,6-diol) as corrosion inhibitors. The predicted results showed that these compounds are good corrosion inhibitors, with high CIE ranging from 82.09% to 95.26%. Subsequent investigation of the DFT revealed that this compound has strong adsorption energy (range −1.40 to −1.52 eV). The adsorption energy trend obtained is in accordance with the predicted CIE trend and experimental results from the literature. In addition, the energy gap trend also supports the CIE trend. This study provides a complementary approach for future work, particularly in updating the data set of hybrid pyrimidine-pyrazole derivatives as corrosion inhibitors. When new compounds are discovered via ML, DFT calculations can help update the data set. This approach may also expand to other classes of compounds beyond the pyrimidine-pyrazole hybrid derivatives.

CRedit authorship contribution statement

Muhamad Akrom: Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing. **Supriadi Rustad:**

Supervision, Conceptualization, Methodology.

Declaration of Competing Interest

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

Data availability

Data will be made available on request.

Acknowledgments

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. SR acknowledges funding support by BRIN Indonesia through the "RIIM G3" program. HKD acknowledges funding support by DIKTIRISTEK through the "PFR 2023" program.

References

- [1] Y. Cui, T. Zhang, F. Wang, New understanding on the mechanism of organic inhibitors for magnesium alloy, *Corros Sci* 198 (2022), 110118, <https://doi.org/10.1016/j.corsci.2022.110118>.
- [2] H. Jin, D.J. Blackwood, Y. Wang, M.F. Ng, T.L. Tan, First-principles study of surface orientation dependent corrosion of BCC iron, *Corros Sci* 196 (2022), 110029, <https://doi.org/10.1016/j.corsci.2021.110029>.
- [3] 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>.
- [4] 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>.
- [5] G.H. Sayed, M.E. Azab, K.E. Anwer, M.A. Raouf, N.A. Negm, Pyrazole, pyrazolone and enaminonitrile pyrazole derivatives: Synthesis, characterization and potential in corrosion inhibition and antimicrobial applications, *J Mol Liq* 252 (2018) 329–338, <https://doi.org/10.1016/j.molliq.2017.12.156>.
- [6] T.K. Sarkar, V. Saraswat, R.K. Mitra, I.B. Obot, M. Yadav, Mitigation of corrosion in petroleum oil well/tubing steel using pyrimidines as efficient corrosion inhibitor: Experimental and theoretical investigation, *Mater Today Commun* 26 (2021), 101862, <https://doi.org/10.1016/j.mtcomm.2020.101862>.
- [7] W. Luo, et al., A new pyridazine derivative synthesized as an efficient corrosion inhibitor for copper in sulfuric acid medium: Experimental and theoretical calculation studies, *J Mol Liq* 341 (2021), 117370, <https://doi.org/10.1016/j.molliq.2021.117370>.
- [8] X. Li, S. Deng, H. Fu, Three pyrazine derivatives as corrosion inhibitors for steel in 1.0 M H₂SO₄ solution, *Corros Sci* 53 (10) (2011) 3241–3247, <https://doi.org/10.1016/j.corsci.2011.05.068>.
- [9] S.A. Abd El-Maksoud, A.S. Fouda, Some pyridine derivatives as corrosion inhibitors for carbon steel in acidic medium, *Mater Chem Phys* 93 (1) (2005) 84–90, <https://doi.org/10.1016/j.materchemphys.2005.02.020>.
- [10] C. Verma, K.Y. Rhee, M.A. Quraishi, E.E. Ebenso, Pyridine based N-heterocyclic compounds as aqueous phase corrosion inhibitors: A review, *J Taiwan Inst Chem Eng* 117 (2020) 265–277, <https://doi.org/10.1016/j.jtice.2020.12.011>.
- [11] C.B. Verma, E.E. Ebenso, I. Bahadur, I.B. Obot, M.A. Quraishi, 5-(Phenylthio)-3H-pyrrole-4-carbonitriles as effective corrosion inhibitors for mild steel in 1 M HCl: Experimental and theoretical investigation, *J Mol Liq* 212 (2015) 209–218, <https://doi.org/10.1016/j.molliq.2015.09.009>.
- [12] L. Jiang, Y. Qiang, Z. Lei, J. Wang, Z. Qin, B. Xiang, Excellent corrosion inhibition performance of novel quinoline derivatives on mild steel in HCl media: Experimental and computational investigations, *J Mol Liq* 255 (2018) 53–63, <https://doi.org/10.1016/j.molliq.2018.01.133>.
- [13] D.S. Chauhan, P. Singh, M.A. Quraishi, Quinoxaline derivatives as efficient corrosion inhibitors: Current status, challenges and future perspectives, *J Mol Liq* 320 (2020), 114387, <https://doi.org/10.1016/j.molliq.2020.114387>.
- [14] 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>.
- [15] S. Echihi, et al., Experimental and theoretical investigation to the mild steel's corrosion inhibition using pyrazole pyrimidine derivative, *Chem. Data Collect.* 46 (2023), 101049, <https://doi.org/10.1016/j.cdc.2023.101049>.
- [16] 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), <https://doi.org/10.1016/j.jscs.2022.101536>.
- [17] C.T. Ser, P. Zuvela, 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), 145612, <https://doi.org/10.1016/j.apsusc.2020.145612>.
- [18] M.E.A. Ben Seghier, D. Höche, M. Zheludkevich, Prediction of the internal corrosion rate for oil and gas pipeline: Implementation of ensemble learning techniques, *J Nat Gas Sci Eng* 99 (2022), <https://doi.org/10.1016/j.jngse.2022.104425>.
- [19] 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>.
- [20] A. Agrawal and A. Choudhary, "Deep materials informatics: Applications of deep learning in materials science," *MRS Communications*, vol. 9, no. 3. Cambridge University Press, pp. 779–792, Sep. 01, 2019, doi: 10.1557/mrc.2019.73.
- [21] 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>.
- [22] 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>.
- [23] A. A. Toropov and A. P. Toropova, "QSPR/QSAR: State-of-art, weirdness, the future," *Molecules*, vol. 25, no. 6. MDPI AG, Mar. 02, 2020, doi: 10.3390/molecules25061292.
- [24] M.E. Belghiti, et al., Computational simulation and statistical analysis on the relationship between corrosion inhibition efficiency and molecular structure of some hydrazine derivatives in phosphoric acid on mild steel surface, *Appl Surf Sci* 491 (2019) 707–722, <https://doi.org/10.1016/j.apsusc.2019.04.125>.
- [25] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
- [26] C. E. J. Armstrong, J. Niimi, P. K. Boss, V. Pagay, and D. W. Jeffery, "Use of Machine Learning with Fused Spectral Data for Prediction of Product Sensory Characteristics: The Case of Grape to Wine," *Foods*, vol. 12, no. 4, Feb. 2023, doi: 10.3390/foods12040757.
- [27] S. Lim, S. Chi, Xgboost application on bridge management systems for proactive damage estimation, *Adv. Eng. Inf.* 41 (2019), <https://doi.org/10.1016/j.aei.2019.100922>.
- [28] J. Ge, et al., Prediction of greenhouse tomato crop evapotranspiration using XGBoost machine learning model, *Plants* 11 (15) (2022), <https://doi.org/10.3390/plants11151923>.
- [29] X. Li, et al., Probabilistic solar irradiance forecasting based on XGBoost, *Energy Rep.* 8 (2022) 1087–1095, <https://doi.org/10.1016/j.egyrs.2022.02.251>.
- [30] A. Syed, R. Adam, T. Ren, J. Lu, T. Maldjian, T.Q. Duong, Machine learning with textural analysis of longitudinal multiparametric MRI and molecular subtypes accurately predicts pathologic complete response in patients with invasive breast cancer, *PLoS One* vol. 18 (2023), <https://doi.org/10.1371/journal.pone.0280320>.
- [31] W. Liu, Z. Chen, Y. Hu, XGBoost algorithm-based prediction of safety assessment for pipelines, *Int. J. Press. Vessel. Pip.* 197 (2022), <https://doi.org/10.1016/j.ijpvp.2022.104655>.
- [32] Y. Wang, F. Su, Y. Guo, H. Yang, Z. Ye, L. Wang, Predicting the microbiologically induced concrete corrosion in sewer based on XGBoost algorithm, *Case Stud. Constr. Mater.* 17 (2022), <https://doi.org/10.1016/j.cscm.2022.e01649>.
- [33] 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) Sep, <https://doi.org/10.1007/s00894-022-05245-1>.
- [34] 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>.
- [35] T. W. Quadri et al., "Computational insights into quinoxaline-based corrosion inhibitors of steel in HCl: Quantum chemical analysis and QSPR-ANN studies," *Arabian Journal of Chemistry*, vol. 15, no. 7 (2022) doi: 10.1016/j.arabjc.2022.103870.
- [36] 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>.
- [37] R.L. Camacho-Mendoza, L. Feria, L.A. 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>.
- [38] 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>.
- [39] 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), 103870, <https://doi.org/10.1016/j.arabjc.2022.103870>.
- [40] T.L. Yusuf, T.W. Quadri, G.F. Tolufashe, L.O. Olasunkanmi, E.E. Ebenso, W.E. Van Zyl, Synthesis and structures of divalent Co, Ni, Zn and Cd complexes of mixed dichalcogen and dipnictogen ligands with corrosion inhibition properties:

- Experimental and computational studies, *RSC Adv* 10 (69) (2020) 41967–41982, <https://doi.org/10.1039/d0ra07770d>.
- [41] A. Kokalj, On the HSAB based estimate of charge transfer between adsorbates and metal surfaces, *Chem Phys* 393 (1) (2012) 1–12, <https://doi.org/10.1016/j.chemphys.2011.10.021>.
- [42] M. Ahsan, M. Mahmud, P. Saha, K. Gupta, Z. Siddique, Effect of data scaling methods on machine learning algorithms and model performance, *Technologies (Basel)* 9 (3) (2021) 52, <https://doi.org/10.3390/technologies9030052>.
- [43] A. Botchkarev, A new typology design of performance metrics to measure errors in machine learning regression algorithms, *Interdiscip. J. Inf. Knowl. Manag.* 14 (2019) 45–76, <https://doi.org/10.28945/4184>.
- [44] X. Yuan, Z. Ge, Z. Song, Soft sensor model development in multiphase/multimode processes based on Gaussian mixture regression, *Chemometrics and Intelligent Laboratory Systems* 138 (Nov. 2014) 97–109, <https://doi.org/10.1016/j.chemolab.2014.07.013>.
- [45] L. Breiman, "Random Forests" (2001).
- [46] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Front Neurobot.* vol. 7, no. DEC, 2013, doi: 10.3389/fnbot.2013.00021.
- [47] S. Bafandeh, I. And, and M. Bolandraftar, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background." [Online]. Available: www.ijera.com.
- [48] P.D. Pately, M.R. Pately, N. Kaushik-Basu, T.T. Talele, 3D QSAR and molecular docking studies of benzimidazole derivatives as hepatitis C virus NS5B polymerase inhibitors, *J Chem Inf Model* 48 (1) (2008) 42–55, <https://doi.org/10.1021/ci700266z>.
- [49] S.R. Bahn, K.W. Jacobsen, An object-oriented scripting interface to a legacy electronic structure code, *Comput Sci Eng* 4 (3) (May 2002) 56–66, <https://doi.org/10.1109/5992.998641>.
- [50] K.R. Cousins, Computer review of chemdraw ultra 12.0, *J Am Chem Soc* 133 (21) (2011) 8388, <https://doi.org/10.1021/ja204075s>.
- [51] P. Giannozzi, et al., "QUANTUM ESPRESSO: A modular and open-source software project for quantum simulations of materials," *Journal of Physics Condensed Matter*. vol. 21(39) (2009) doi: 10.1088/0953-8984/21/39/395502.
- [52] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized Gradient Approximation Made Simple," 1996.
- [53] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, "A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu," *Journal of Chemical Physics*, vol. 132, no. 15, Apr. 2010, doi: 10.1063/1.3382344.
- [54] A.G. Saputro, et al., Dissociative Oxygen Reduction Reaction Mechanism on the Neighboring Active Sites of a Boron-Doped Pyrolyzed Fe-N-C Catalyst, *J. Phys. Chem. C* 124 (21) (2020) 11383–11391, <https://doi.org/10.1021/acs.jpcc.0c00632>.
- [55] H.K. Dipojono, et al., Oxygen reduction reaction mechanism on a phosphorus-doped pyrolyzed graphitic Fe/N/C catalyst, *New J. Chem.* 43 (28) (2019) 11408–11418, <https://doi.org/10.1039/c9nj02118c>.
- [56] T. Ossowski, A. Kiejna, Oxygen adsorption on Fe(110) surface revisited, *Surf Sci* 637–638 (2015) 35–41, <https://doi.org/10.1016/j.susc.2015.03.001>.
- [57] J. Cai, D. Lu, The formation energy and bonding characteristics of small helium-vacancy clusters on the low-index surface of α -Fe by first principles calculations, *Comput Mater Sci* 92 (2014) 387–394, <https://doi.org/10.1016/j.commatsci.2014.05.064>.
- [58] D. Kumar, V. Jain, B. Rai, Imidazole derivatives as corrosion inhibitors for copper: A DFT and reactive force field study, *Corros Sci* 171 (2020), 108724, <https://doi.org/10.1016/J.CORSCI.2020.108724>.
- [59] D. Kumar, N. Jain, V. Jain, B. Rai, Amino acids as copper corrosion inhibitors: A density functional theory approach, *Appl Surf Sci* 514 (2020), 145905, <https://doi.org/10.1016/J.APSUSC.2020.145905>.
- [60] Z. Cao, Y. Tang, H. Cang, J. Xu, G. Lu, W. Jing, Novel benzimidazole derivatives as corrosion inhibitors of mild steel in the acidic media. Part II: Theoretical studies, *Corros Sci* 83 (2014) 292–298, <https://doi.org/10.1016/J.CORSCI.2014.02.025>.
- [61] I.B. Obot, Z.M. Gasem, Theoretical evaluation of corrosion inhibition performance of some pyrazine derivatives, *Corros Sci* 83 (2014) 359–366, <https://doi.org/10.1016/J.CORSCI.2014.03.008>.
- [62] Ş. Erdoğan, Z.S. Safi, S. Kaya, D.Ö. Işın, L. Guo, C. Kaya, A computational study on corrosion inhibition performances of novel quinoline derivatives against the corrosion of iron, *J Mol Struct* 1134 (2017) 751–761, <https://doi.org/10.1016/J.MOLSTRUC.2017.01.037>.
- [63] S.K. Saha, M. Murmu, N.C. Murmu, I.B. Obot, P. Banerjee, Molecular level insights for the corrosion inhibition effectiveness of three amine derivatives on the carbon steel surface in the adverse medium: A combined density functional theory and molecular dynamics simulation study, *Surf. Interfaces* 10 (2018) 65–73, <https://doi.org/10.1016/J.SURFIN.2017.11.007>.
- [64] D. Quy Huong et al., "Pivotal Role of Heteroatoms in Improving the Corrosion Inhibition Ability of Thiourea Derivatives," *ACS Omega*, vol. 5, no. 42, pp. 27655–27666, Oct. 2020, doi: 10.1021/acsomega.0c04241.
- [65] T. Le Minh Pham, T. Khoa Phung, and 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.* vol. 583, (2022) doi: 10.1016/j.apsusc.2022.152524.
- [66] 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, *Inorganica Chim Acta* 512 (2020), <https://doi.org/10.1016/j.ica.2020.119872>.
- [67] M. Akrom, S. Rustad, A.G. Saputro, A. Ramelan, F. Fathurrahman, H.K. Dipojono, A combination of machine learning model and density functional theory method to predict corrosion inhibition performance of new diazine derivative compounds, *Mater Today Commun* 35 (2023), 106402, <https://doi.org/10.1016/J.MTCOMM.2023.106402>.
- [68] 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>.
- [69] 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>.
- [70] 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>.