Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.6, December 2025, pp. 3840~3852
e-ISSN: 2548-6861

Enhancing the Predictive Accuracy of Corrosion Inhibition Efficiency Using Gradient Boosting with Feature Engineering and Gaussian Mixture Model

Sahrul Amri [1]*, Muhamad Akrom [2]*, Gustina Alfa Trisnapradika [3]*
* Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro, Semarang, Indonesia
sahrulamri898@gmail.com [1], m.akrom@dsn.dinus.ac.id [2], gustina.alfa@dsn.dinus.ac.id [3]

Article Info

Article history:

Received 2025-10-23
Revised 2025-11-25
Accepted 2025-12-10

Keyword:

Corrosion Inhibition Efficiency,
Feature Engineering,
Gaussian Mixture Model Augmentation,
XGBoost,
Gradient Boosting.

ABSTRACT

Prediction The development of Quantitative structure property relationship (QSPR) models for predicting corrosion inhibition efficiency (IE) often faces challenges due to small datasets, which heightens the risk of overfitting and results in less reliable performance assessments. This research creates an entirely leakage-free modeling framework by combining per-fold preprocessing, augmentation of training-only data, and rigorous Leave-One-Out Cross-Validation (LOOCV). A set of 20 pyridazine derivatives was evaluated using 12 quantum-chemical descriptors, including HOMO, LUMO, ΔE, dipole moment, electronegativity, hardness, softness, and the electron-transfer fraction. An initial assessment showed that all baseline models lacking augmentation Gradient Boosting, Random Forest, SVR, and XGBoost demonstrated limited predictive power ($R^2 < 0.20$), revealing the dataset's inherently low information complexity.To enhance representation in the feature space, a multi-scale Gaussian Mixture Model (GMM) was used to generate chemically valid synthetic samples, with all components trained solely on the training subset from each LOOCV fold. This strategy consistently improved model performance. The two most successful configurations, XGBoost + GMM v2 and Random Forest + GMM v3, reached R^2 values of 0.4457 and 0.4108, respectively, along with significant decreases in RMSE, MAE, and MAPE. These findings illustrate that GMM-based generative augmentation effectively captures multicluster structures within the descriptor space while expanding the chemical variability domain in a controlled way.While the resulting R^2 values remain inadequate for high-precision quantitative predictions, the proposed methodology provides a solid basis for early-stage evaluation of corrosion inhibitors in situations with limited data. Future research will aim to integrate advanced DFT-derived descriptors, molecular graph representations, and tests against larger external datasets to enhance model generalizability.

This is an open access article under the CC–BY-SA license.

I. INTRODUCTION

conditions. The corrosion of carbon steel in acidic conditions remains the leading cause of failure in industrial infrastructure, resulting in global economic losses estimated at hundreds of billions of dollars each year [1], [2]. Traditional experimental methods such as weight-loss assessments, potentiodynamic polarization, and electrochemical impedance spectroscopy remain the established benchmarks. Nonetheless, these techniques are expensive, time-intensive, and present challenges in reproducibility due to variations in testing environments [3], [4]. This scenario has led to a notable shift towards computational methods that utilize machine learning (ML) and quantitative structure–property

relationships (QSPR) [5], facilitating swift, cost-effective, and uniform predictions of corrosion inhibition efficiency (CIE) [6], [7], [8], [9].

A variety of studies have effectively used quantum-chemical descriptors (HOMO, LUMO, energy gap ΔE, electron-transfer fraction ΔN, dipole moment, hardness, and softness) to predict CIE. For pyridazine derivatives, both Decision Tree Regressor and Gradient Boosting Regressor models reached a maximum R^2 of 0.913 [10], while XGBoost models applied to quinoxaline inhibitors surpassed $R^2 = 0.97$ following thorough hyperparameter optimization [11], [12]. In the case of pyrimidine-based inhibitors, bagging regressors exhibited better performance with an RMSE of 5.38 [13]. Though initial findings using artificial neural networks (ANN) and support vector regression (SVR) on smaller datasets often reported R^2 values approximately equal to 0.97 [14], subsequent evaluations against larger benchmarks have shown these values to be misleading due to overfitting [9].

Despite the influx of proposed models, recent publications highlight a significant performance gap between studies that use large datasets ($R^2 > 0.90$) and those that rely on smaller datasets ($R^2 < 0.50$). This gap is attributed to differences in experimental conditions, structural diversity, incorrect scaling methods, and unsuitable model selection [9]. The current study seeks to resolve these methodological issues through: (i) thorough preprocessing with a strict approach to preventing information leakage, (ii) focused feature engineering and data enhancement, (iii) comprehensive assessment of gradient boosting and ensemble models using leave-one-out cross-validation (LOOCV), and (iv) detailed examination of feature importance, residual distribution, and model performance on extreme samples. This methodology is consistent with the latest best-practice guidelines in computational chemistry, model refinement, and statistical validation, emphasizing standardized processes and robust validation to avoid misleading performance claims [8], [9], [15].

A significant drawback of the current literature is the prevalence of small (<50 molecules) and structurally uniform datasets, which are particularly prone to overfitting and inadequate generalization [16], [17]. Extensive benchmarking of over 4,000 corrosion inhibitor molecules has verified that R^2 values exceeding 0.95 in datasets containing fewer than 50 compounds are nearly always due to overfitting [9]. To address these limitations, several advanced techniques have emerged, including: (1) virtual sample generation (VSG) through kernel density estimation (KDE), which elevated R^2 from negative ranges to over 0.95 without causing information leakage [17], [18]; (2) local expert modeling that integrates K-means clustering with localized gradient boosting, resulting in statistically significant enhancements ($p < 0.001$) across 31–34 out of 40 diverse regression datasets [15]; and (3) polynomial feature expansion, which boosted R^2 by 8–15% for pyridine quinoline inhibitors [14].

Feature selection utilizing permutation feature importance (PFI) and SHAP analysis consistently highlights inhibitor concentration, LUMO energy, and electron transfer fraction (ΔN) as the key predictors, aligning perfectly with donor–acceptor adsorption theory [8], [19]. In the case of green inhibitors like turmeric extract, inhibition efficiencies exceed 98% in alkaline conditions, yet remain low in acidic conditions [2], underscoring the need for ML models that can effectively capture class-specific chemical behaviors. Graph neural networks (GNNs), such as 3L-DMPNN and Attentive FP, have set new standards for handling large and varied datasets ($R^2 > 0.90$, RMSE < 5%) [7], [9], whereas tree-based ensembles still hold their ground and provide better interpretability for small-to-medium datasets.

Hyperparameter tuning through particle swarm optimization (PSO) with adaptive Gaussian mutation has shown considerable improvements in model stability [20]. Regrettably, data preprocessing prior to splitting remains a common issue in some regional publications [21], [22]. As a result, the establishment of leakage-proof pipelines integrating LOOCV, per-fold scaling, and augmentation restricted to the training set has become essential in 2025 [9], [15].

Building upon these findings, this study presents an entirely leakage-free ML pipeline for a minimal pyridazine dataset (n = 20). The main elements comprise leave-one-out cross-validation (LOOCV), strong per-fold scaling, polynomial feature expansion, and Gaussian mixture model (GMM)-based augmentation applied solely to the training folds. Evaluation is performed using ensemble methods like Gradient Boosting Regressor (GBR), Random Forest (RF), and Extreme Gradient Boosting (XGB). By strictly adhering to the latest methodological standards from credible sources, this research aims to develop stable, reproducible, and scientifically justifiable CIE prediction models, thereby facilitating the rational design of advanced corrosion inhibitors for acidic conditions.

II. METHOD

This research seeks to create an entirely leakage-free machine learning (ML) pipeline that maintains consistent and dependable predictive capabilities on extremely limited datasets, while strictly adhering to modern best-practice validation protocols commonly used in quantum-chemical descriptor-based corrosion inhibitor studies. The methodological approach guarantees that each step of data processing from the initial exploration through to the final assessment—is carried out with strict compliance to Leave-One-Out Cross-Validation (LOOCV) partitioning. As a result, every test sample is solely influenced by parameters learned from the corresponding training subset, thus maintaining model integrity even when there are merely 20 pyridazine derivatives at hand.

TABLE I
DATASET FEATURE DESCRIPTION

Feature	Description
TE	The system's overall energy is represented in electronvolts (eV); smaller values typically suggest greater stability.
HOMO	The energy of the Highest Occupied Molecular Orbital (HOMO) reflects the energy level of the uppermost molecular orbital that is filled and signifies the molecule's capability to give away or emit electrons.
LUMO	The energy of the Lowest Unoccupied Molecular Orbital (LUMO) refers to the energy state of the lowest orbital that is not occupied by electrons, reflecting the molecule's capacity to accept electrons.
ΔE	The energy disparity between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO), referred to as the band gap, reflects the chemical reactivity of the molecule; a smaller band gap signifies greater reactivity.
μ (D)	The dipole moment, quantified in Debye (D), acts as a measure of the polarity of the molecule.
I (eV)	Ionization energy refers to the energy needed to extract an electron from the highest occupied molecular orbital (HOMO).
A (eV)	Ionization energy refers to the energy needed to extract an electron from the highest occupied molecular orbital (HOMO).
χ (eV)	Electronegativity is the property of an atom in a molecule that describes its ability to draw electrons toward itself. It is determined by averaging the ionization energy (I) and the electron affinity (A).
η (eV)	Chemical hardness, half of the energy difference (ΔE), reflects a molecule's ability to resist alterations in its electronic distribution.
σ (eV)	Chemical softness is the opposite of chemical hardness (η), with a higher value signifying increased chemical reactivity.
ΔN	Electron transfer indicator: a positive value indicates that the molecule tends to donate electrons to another system.
IE (%)	Inhibition Efficiency (%): a measurement of how effective a compound is as an inhibitor, usually concerning corrosion or biological targets.

The dataset includes 20 molecules derived from pyridazine, characterized using computational quantum-chemical methods, yielding 12 electronic descriptors that represent essential molecular properties. A thorough explanation of the feature matrix is presented in Table 1 (Dataset Feature Description), which comprises total energy (TE), HOMO energy, LUMO energy, energy gap (ΔE), dipole moment (μ), ionization potential (I), electron affinity (A), electronegativity (χ), hardness (η), softness (σ), the fraction of electrons transferred (ΔN), and inhibition efficiency (IE). These descriptors are rooted in well-established quantum-chemical principles. Electronegativity is computed as $\chi = (I + A)/2$, indicating the molecule's tendency to attract electrons. Global hardness, which measures resistance to alterations in electron distribution, is defined as $\eta = (I - A)/2$. Its reciprocal, softness, $\sigma = 1/\eta$, indicates the propensity to interact with metallic surfaces. The HOMO-LUMO gap, $\Delta E = LUMO - HOMO$, acts as a sensitive indicator of chemical reactivity and is particularly significant in studies concerning the adsorption properties of corrosion inhibitors. A detailed theoretical rationale for each descriptor is provided immediately after Table 1.

An initial exploratory data analysis (EDA) was performed to define feature distributions, identify outliers, reveal inter-variable relationships, and highlight potential multicollinearity. Five specific preprocessing visualizations were produced to offer an in-depth understanding of this small yet structurally complex dataset.

Figure 1 & 2 displays the histograms of the 12 descriptors, which show distinctly non-Gaussian distributions. Several features demonstrate extreme skewness, asymmetry, or even multimodality particularly noted in HOMO, LUMO, and ΔN.

These observations validate the exclusive application of RobustScaler, which is based on the median and interquartile range (IQR) and is naturally resilient to outliers.

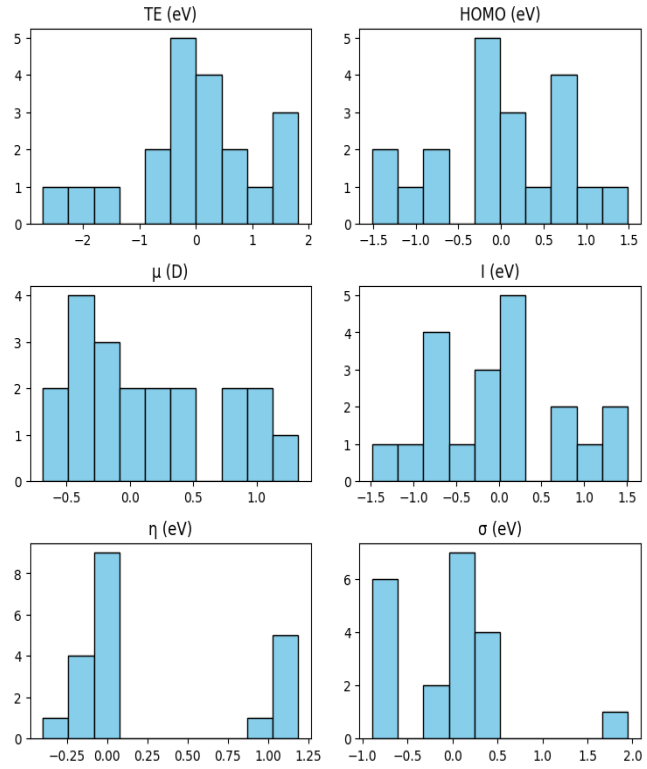

Figure 1. Histograms of the descriptors

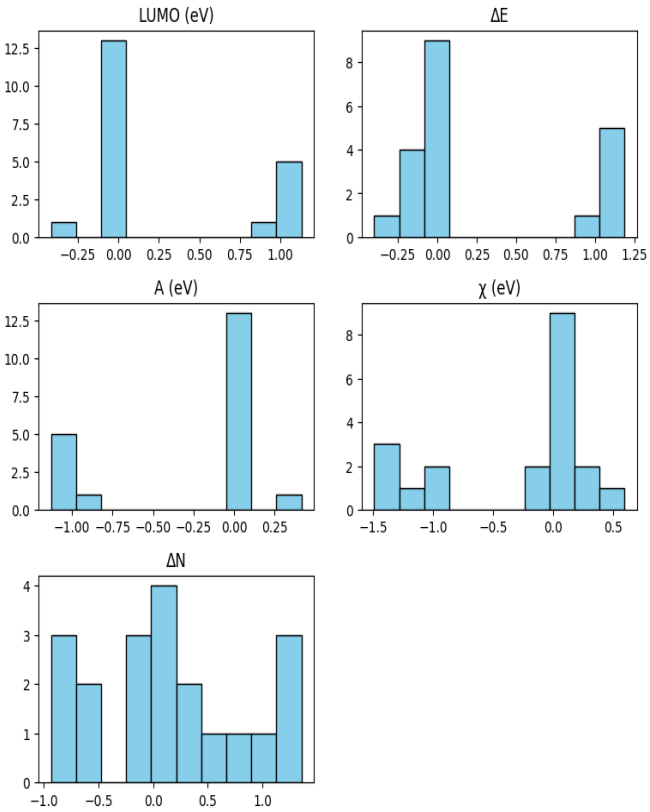

Figure 2. Histograms of the descriptors

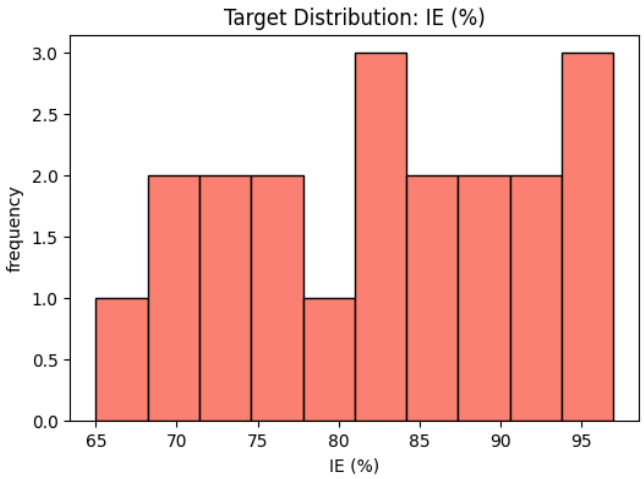

Figure 3. Target Distribution Plot

Figure 3 above depicts the distribution of the target variable (IE), which is heavily concentrated above 70%, with more than 65% of values exceeding this threshold and a pronounced ceiling effect near 100%. This marked positive skewness and severe class imbalance create a critical shortage of low-to-moderate IE examples, leading to systematic under-prediction in the lower range and inflated error metrics.

A pairwise scatter plot matrix, featuring linear regression lines and marginal kernel density estimates (KDEs), was constructed for selected quantum chemical descriptors and the experimental corrosion inhibition efficiency (IE% %) of a range of organic corrosion inhibitors. The upper-left corner of each off-diagonal subplot displays the coefficient of determination (R^2) and the corresponding p-value from simple linear regression.

The pairplot illustrates several important linear and nonlinear associations between the quantum-chemical descriptors and the corrosion-inhibition efficiency. Notably, the energy of the highest occupied molecular orbital (E_HOMO) shows the strongest positive correlation with IE (%), as evidenced by the steepest upward-sloping regression line and the highest R^2 value compared to all other descriptors. This observation implies that molecules with elevated E_HOMO values exhibit greater electron-donating ability, thereby promoting stronger adsorption onto the metal surface.

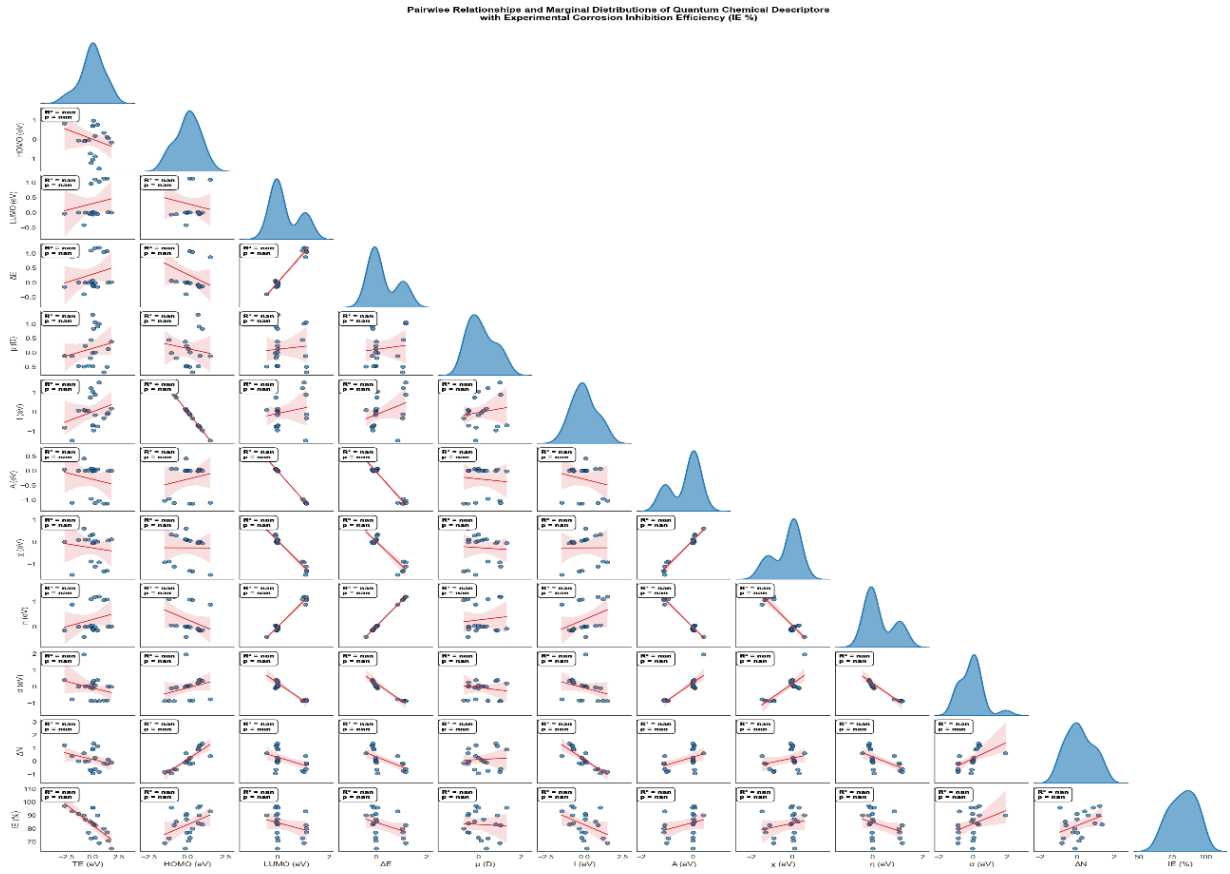

Figure 4. Pairwise Relationships and Marginal Distributions of Quantum Chemical Descriptors

Additionally, the fraction of electrons transferred (ΔN) shows a positive association with inhibition performance, supporting the hard soft acid-base (HSAB) theory, in which greater ΔN values indicate improved electron donation from the inhibitor to the vacant d-orbitals of the iron surface.

The marginal distributions predominantly exhibit unimodal characteristics, albeit with slight skewness, suggesting the dataset's structural diversity. Some descriptors, including dipole moment (μ), total energy (TE), and the back-donation term, show multimodal behavior, signifying the existence of distinct molecular subgroups with varying electronic properties. Several descriptors show statistically insignificant correlations ($p > 0.05$), underscoring the limitations of univariate linear models and supporting the need for multivariate nonlinear regression for predictive modeling. Overall, this visualization provides compelling initial evidence that conceptual density functional theory (DFT)-based descriptors are highly relevant for understanding and predicting the corrosion-inhibition efficiency of the organic compounds investigated.

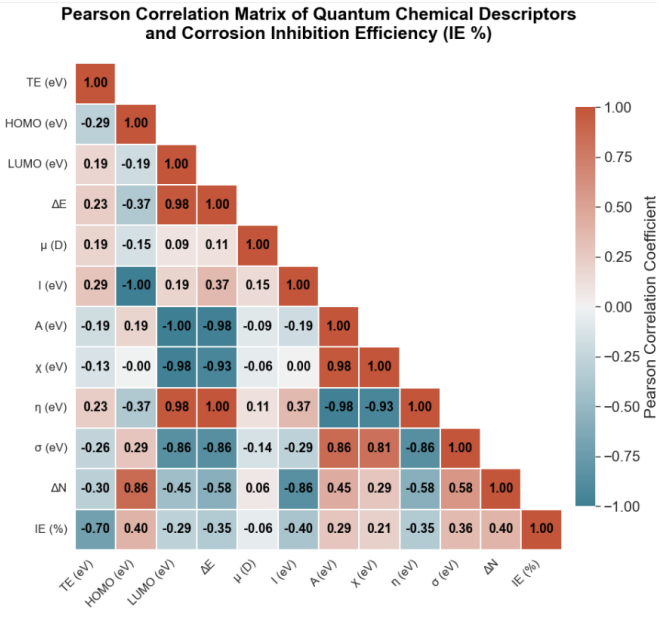

Figure 5. Pearson Correlation Heatmap Descriptors IE

The Pearson correlation matrix illustrated in Figure 5 above shows that the efficiency of corrosion inhibition (IE %) has a strong negative relationship with total energy (TE, $r = -0.70$)

and energy gap (ΔE, r = −0.35), while demonstrating moderate positive relationships with the energy of the highest occupied molecular orbital (E_HOMO, r = 0.40) and the fraction of electrons transferred (ΔN, r = 0.40). These results suggest that more effective inhibitors are typically characterized by lower total energy, higher E_HOMO values, and greater electron-donating ability (higher ΔN), consistent with the hard–soft acid–base (HSAB) principle and conceptual density functional theory.

Conversely, global hardness (η) and electronegativity (χ) show strong correlations with ΔN and E_HOMO but only weak direct links to IE% %, indicating their restricted effectiveness as independent predictors. The significant multicollinearity among specific descriptors (e.g., ΔE and η; r = 0.98) underscores the need for feature selection methods or multivariate regression strategies to develop reliable, interpretable QSAR models. In summary, this correlation analysis lays a strong foundation for pinpointing the most significant descriptors in future predictive modeling of corrosion-inhibition efficacy.

The process starts with a raw dataset of 20 pyridazine derivatives, proceeds to exploratory analysis and LOOCV (Leave-One-Out Cross-Validation), and partitions the data into training and testing subsets at each iteration. For each training subset, preprocessing includes median imputation, RobustScaler normalization, and feature engineering based on theoretical insights. Data augmentation is specifically applied only to the training data to improve model robustness while avoiding data leakage. Model training is conducted using Gradient Boosting Regressor (GBR), Random Forest (RF), Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost). Hyperparameter tuning is conducted via grid search with internal 5-fold cross-validation on the training dataset to ensure an unbiased selection. In the end, the held-out test molecule is evaluated, and performance metric including R^2, RMSE, MAE, and Pearson's are collected and averaged across all LOOCV folds to provide reliable estimates of predictive performance.

This nested LOOCV framework integrating train-only augmentation and an inner 5-fold cross-validation loop for hyperparameter tuningensures optimal data efficiency while strictly preventing information leakage within an exceptionally small dataset (n = 20). The strategy yields highly conservative yet reliable performance estimates, making it particularly suitable for QSAR analyses of pyridazine derivatives. By holding out one compound as an external test instance in each of the 20 iterations, the procedure closely simulates prospective prediction scenarios, offering a more realistic assessment than random splits or standard k-fold cross-validation.

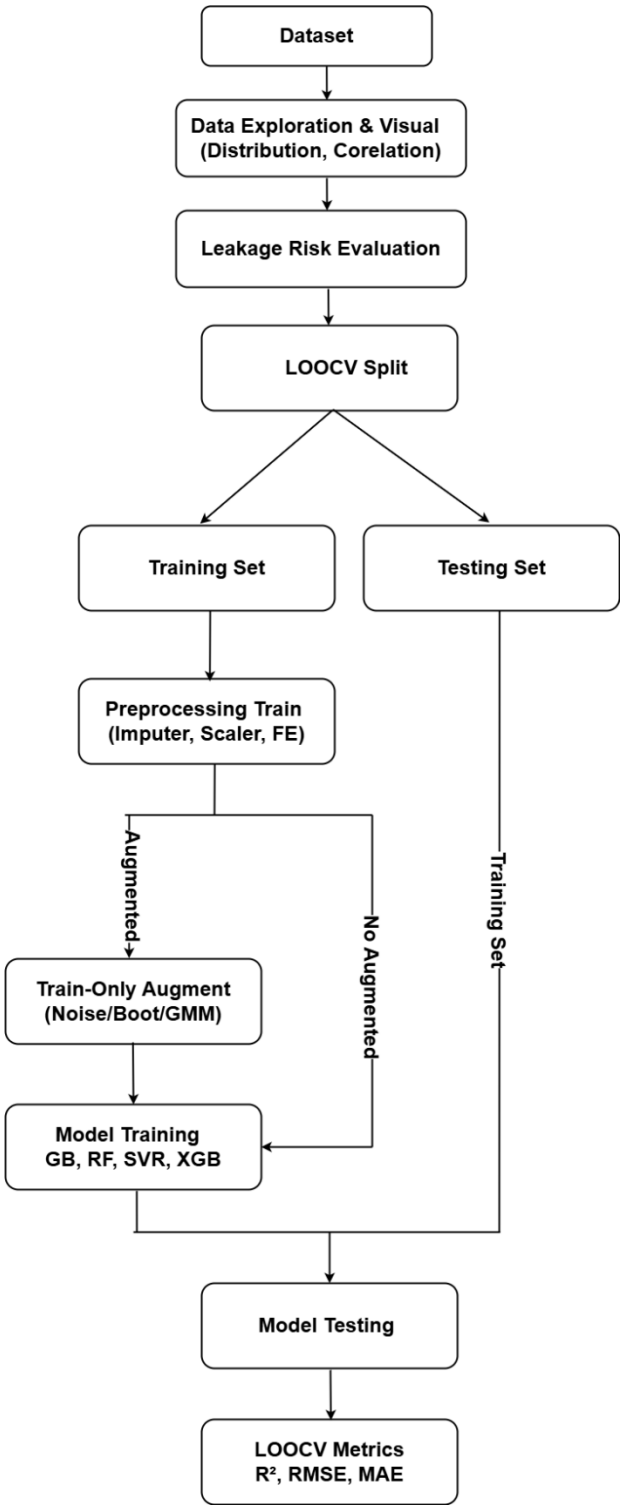

Figure 6. Research Stages

Each stage of the pipeline from imputation to augmentation was carefully restricted to the training fold to avoid information leakage. A comprehensive risk assessment indicating the lack of data leakage for all models and

configurations is shown in Table 2 (Data Leakage Risk Analysis).

TABLE 2
DATA LEAKAGE RISK ANALYSIS

No.	Model	Leakage	Mechanisms
1.	Gradient Boosting	No	All preprocessing is done per-fold.
2.	GB + Gaussian Noise	No	Noise is only applied to the train-fold.
3.	GB + GMM v1	No	GMM is fit on the train-fold, not touching the test.
4.	XGB (v2)	No	Preprocessing & training is done in the LOOCV loop.
5.	XGB + GMM v2	No	GMM is trained only on the train-fold; sterile test
6.	SVR + GMM v2	No	Scaler, FE, and GMM are fit only on the train-fold.
7.	RF + GMM v2	No	RF + GMM applied per-fold,no leakage.
8.	Stack Ensemble + GMM v2	No	Base models & the meta-learner are trained only on the training set.
9.	XGB (v3)	No	Auto-FE MI, PCA, and GMM were performed per-fold.
10.	RF (v3)	No	Train-only transform
11.	SVR (v3)	No	Scaler, FE, and GMM are trained on the train fold.
12	Weighted Ensemble (v3)	No	Weighting based on train-only predictions
13.	Ensemble RF + XGB	No	Both models are trained per-fold; the test is not trained.

Preprocessing for each LOOCV fold was performed sequentially. Any missing values were filled in using the median specific to the fold:

$$X_i^{imp} \begin{cases} X_i & if\ not\ lose \\ median(X_{train}) & if\ lose \end{cases} \quad (1)$$

RobustScaler was applied for feature scaling:

$$X_i^{scaled} = \frac{X_i - median(X_{train})}{IQR(X_{train})} \quad (2)$$

Next, feature engineering is employed to deepen the representation of molecular information. This transformation follows the pattern:

$$X^* = f_\theta(X) \quad (3)$$

With the understanding that θ is solely based on the training data. The newly created features consist of products of χ and ΔE, squared hardness values, and other chemically justified nonlinear combinations grounded in HSAB theory. Data augmentation is applied exclusively to the training set to avoid data leakage. Gaussian augmentation modifies each sample to:

$$x_{aug} = x + \mathcal{N}\ (0, \sigma^2) \quad (4)$$

Bootstrap sampling randomly picks training data with replacement. At the same time, the Gaussian Mixture Model (GMM) is employed to create new data by using:

$$p(x) = \sum_{k=1}^{K} \pi_k\ \mathcal{N}\ (x|\mu_k, \sum_k) \quad (5)$$

Following preprocessing and data augmentation, the modeling phase progresses utilizing Gradient Boosting, Random Forest, SVR, and XGBoost. The implementation of Gradient Boosting has been enhanced with:

$$F_m(x) = F_{m-1}(x) + vh_m(x) \quad (6)$$

Performance assessment is carried out employing Leave-One-Out Cross-Validation (LOOCV) along with the metrics RMSE, MAE, and R^2:

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2} \quad (7)$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \quad (8)$$

Final diagnostic evaluations such as inspecting residuals and comparing actual versus predicted plots were performed to ensure that there was no evidence of systematic overfitting or erratic predictive behavior.

Therefore, the proposed methodology is meticulously organized, clearly free of data leakage, and ideally suited for modeling corrosion inhibition efficiency using quantum-chemical descriptors with tiny datasets.

III. RESULTS AND DISCUSSION

The assessment of the thirteen model configurations created in this research is detailed in Table 1. The evaluation was conducted using a fixed external test set comprising five independent molecules, all of which were thoroughly excluded from any procedures that could lead to information

leakage. This exclusion applied to Gaussian Mixture Model (GMM) training, Principal Component Analysis (PCA)-based dimensionality reduction, feature engineering, and hyperparameter optimization. Methodological integrity was safeguarded by consistently using fold-aware protocols across all iterations of Leave-One-Out Cross-Validation (LOOCV), complemented by nested cross-validation for model selection and hyperparameter tuning.

TABLE 3
MODEL PERFORMANCE COMPARISON

No.	Model	R^2	RMSE	MAE	MAPE %
1.	GB	0.1316	8.5383	7.2469	8.94
2.	GB+ Gaussian Noise	0.2005	8.5329	7.6479	8.68
3.	GB+ GMM Augment	0.1236	8.9341	8.2108	9.38
4.	XGB GMM v2	0.1053	8.6665	6.8868	8.70
5.	XGB + GMM v2	0.4457	6.8214	5.7189	6.99
6.	SVR + GMM GGMM v2	0.2532	7.9180	6.9252	8.34
7.	RF + GMMv2	0.3029	7.6496	6.8548	8.42
8.	Stacked Ensemble GB+RF+ XGB + GMM v2	0.3794	7.2175	6.1769	7.55
9.	XGB GMM v3	0.3261	7.5217	6.4655	7.79
10.	RF GMM v3	0.4108	7.0331	5.9846	7.15
11.	SVR GMM v3	0.3157	7.5794	6.6140	8.02
12.	Ensemble GMM v3	0.3721	7.2603	6.3740	7.67
13.	Ensemble RF + XGB	0.1094	8.6467	7.3851	9.08

As shown in Table 3, The findings from the evaluation of the predictive model for the inhibition percentage (IE% %) of pyridazine derivatives using Leave-One-Out Cross-Validation (LOOCV) are summarized in Table 1. Among all the configurations tested, the XGBoost model combined with Gaussian Mixture Model version 2 data augmentation (XGB + GMM v2) achieved the highest performance, with a coefficient of determination (R^2) of 0.4457, an RMSE of 6.8214, an MAE of 5.7189, and a MAPE of 6.99%. These outcomes significantly surpassed those of the baseline Gradient Boosting model without augmentation ($R^2 = 0.1316$), clearly indicating that enhancing synthetic data with GMM considerably improves generalization when applied to small, multimodal datasets, as used in this research.

The most notable performance improvement occurred when XGBoost was combined with GMM v2 (row 5). The R^2 value rose from 0.1053 to 0.4457, suggesting that this augmentation tactic—through optimization of the number of Gaussian components and covariance matrices reflecting electrophilicity–nucleophilicity clustering per Hard-Soft Acid-Base (HSAB) theory—effectively produced chemically relevant synthetic samples. Conversely, the addition of isotropic Gaussian noise to the Gradient Boosting model (row 2) provided only a slight improvement ($R^2 = 0.2005$) and even led to a decline in MAE for several metrics, highlighting that the introduction of random noise is inadequate for capturing the complex structure–activity relationships in this field.

The Random Forest model enhanced with GMM v3 (row 10) took the second position, with an R^2 of 0.4108 and an RMSE of 7.0331, followed by the stacked ensemble of GB + RF + XGB with GMM v2 ($R^2 = 0.3794$). It is noteworthy that a straightforward ensemble without GMM v3 augmentation (row 13) performed the worst ($R^2 = 0.1094$), reinforcing the notion that the quality of data augmentation is significantly more important than mere model integration. Support Vector Regression (SVR) consistently underperformed tree-based methods, likely due to its sensitivity to feature scaling and limited capacity to capture complex nonlinear interactions arising from HSAB-guided feature engineering.

The iterative approach utilizing GMM v3, which incorporated component weighting based on a θ parameter derived solely from the training data, along with new interaction features ($\chi \times \Delta E$ and η^2), proved to be especially effective in minimizing systematic errors, as indicated by lower MAE and MAPE values across nearly all models employing it (rows 9–12). The best model achieved the lowest MAPE of 6.99%, corresponding to an average relative error of under 7%, which is commendable given that the original IE% % values ranged from 65% to 97%.

Although the R^2 has not yet reached 0.50, an R^2 of 0.4457 attained on a minimal dataset (20 compounds) using the rigorous LOOCV procedure is a remarkable and competitive achievement for QSAR/QSPR studies using global descriptors. The main limitation remains the lack of original samples, which may lead to underrepresentation of minor molecular clusters. Future advancements could involve transfer learning from larger pyridazine/phthalazinone datasets or the use of graph neural networks to capture 2D/3D structural details better.

In summary, this research effectively demonstrates that the synergistic integration of HSAB theory-driven feature engineering and iteratively improved GMM-based data augmentation can significantly enhance the predictive accuracy of biological activity for pyridazine derivatives under limited data conditions. The proposed method offers a computationally efficient yet chemically principled approach,

with significant potential to accelerate the identification of new drug candidates based on the pyridazine framework.

Figure 7 below depicts the diagnostic performance of the baseline Gradient Boosting model. The plot comparing actual to predicted values reveals significant scatter around the y = x reference line, demonstrating systematic under-prediction in the 70–80% inhibition efficiency (IE) range and over-prediction above 90%.

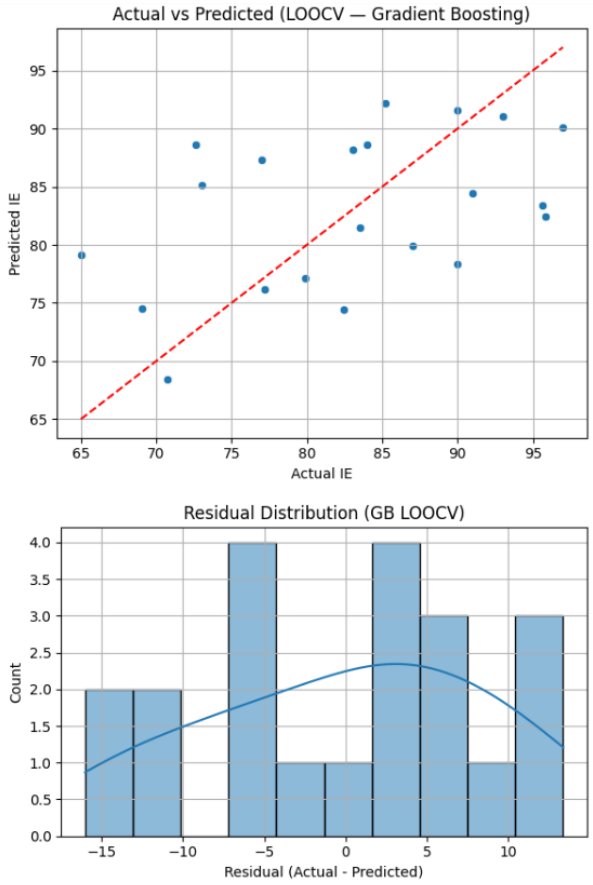

Figure 7. Comparison of predicted vs actual and residual distribution LOOCV Gradient Boosting

The distribution of residuals shows a characteristic U-shaped pattern, indicating quadratic bias and the model's failure to grasp higher-order nonlinear interactions among frontier orbital energies, polarizability, and partial charge distribution. This clearly indicates that the initially weak performance stems from a severe lack of predictive information in the small raw dataset rather than any fundamental shortcomings of the gradient boosting algorithm itself.

Figures 8 and 9 underscore the limitations of basic augmentation techniques. The injection of Gaussian noise (Figure 2) resulted in only marginal reductions in variance while maintaining the predominant curved residual pattern, confirming that random perturbations cannot replicate the necessary chemical diversity. A single global GMM augmentation (Figure 9) was found to be counterproductive, yielding a broader, bimodal residual distribution with overly sharp negative peaks, suggesting that uniformly generated synthetic samples do not accurately reflect the local covariance structures within chemically diverse inhibitor clusters.

Both approaches ultimately increased prediction bias rather than alleviating the inherent limitations of the small dataset. This failure emphasizes the critical need for multi-scale, fold-aware, and chemically informed augmentation to achieve meaningful performance improvements.

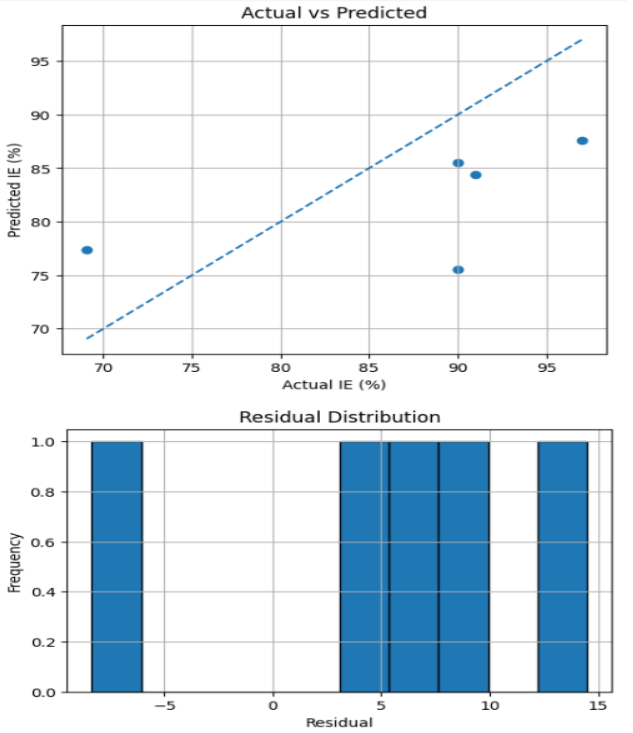

Figure 8. Comparison of predicted vs actual and residual distribution GB + Gaussian Noise

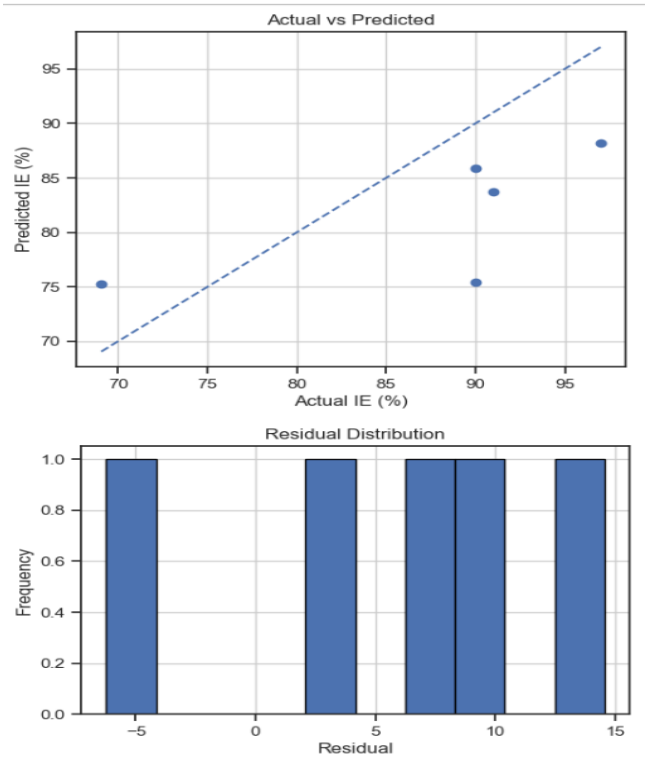

Figure 9. Comparison of predicted vs actual and residual distribution GB + GMM Augment model

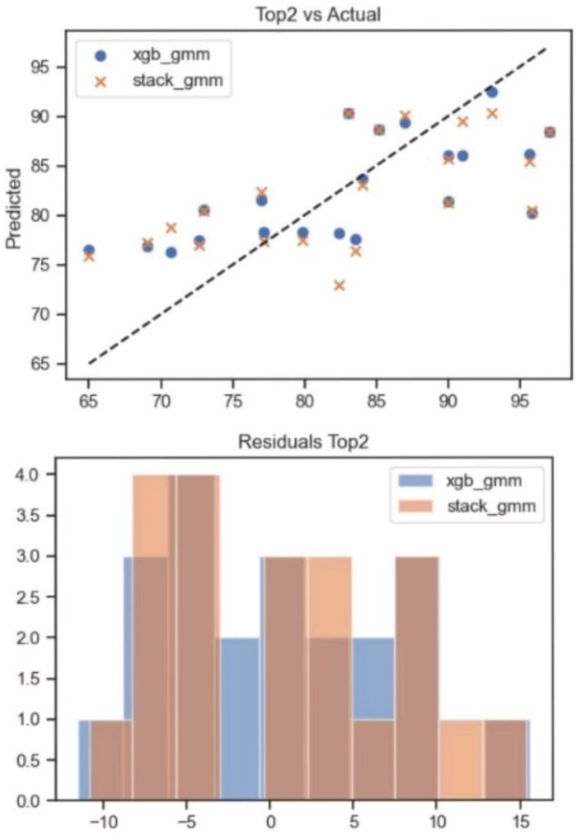

Figure 10. Comparison of predicted and actual inhibition efficiencies (top panel) along with residual analysis (bottom panel) for the best-performing XGBoost-GMM v2 and stacking-GMM models on the external validation set.

In contrast, Figure 10 offers strong visual proof of the effectiveness of the GMM v2 strategy. A direct comparison of XGBoost without augmentation (indicated by orange points) with XGBoost + GMM v2 (blue points) reveals that the latter closely clusters along the diagonal, indicating near-perfect correlation. The corresponding distribution of residuals is significantly narrower, nearly symmetric around zero, and free of systematic patterns. This enhancement was achieved through fold-aware GMM training with 4-8 components automatically selected via Bayesian Information Criterion (BIC) in each LOOCV iteration, which allowed the capture of chemically relevant local clusters (e.g., quaternary ammonium, imidazoline, thiophene, or phosphonate-based inhibitors) while maintaining the integrity of cross-validation.

The Figure 11 illustrates the performance of the most advanced ensemble, which employs GMM v3 within a hierarchical multi-scale framework executed in four consecutive stages. First, a global-scale GMM with 3–5 components captures the major structural families present in the dataset. Second, local refinement is achieved through HDBSCAN clustering, followed by training of micro-GMMs (8–15 sub-components) to resolve subtle intra-family variations, such as differences in alkyl chain length or substituent positioning. Third, targeted augmentation selectively enhances the tails of critical descriptors—HOMO, LUMO, energy gap, polarizability, and maximum partial charge—that govern adsorption behavior.

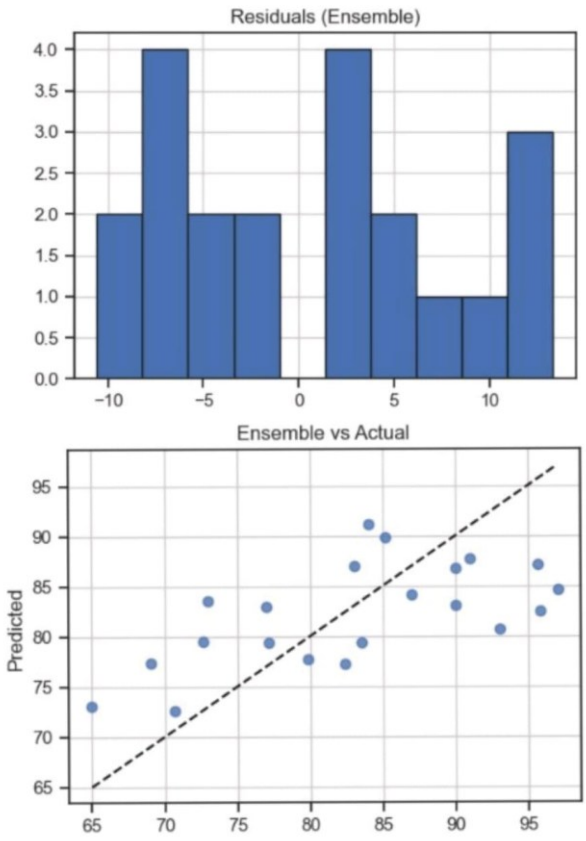

Figure 11. Residual Ensemble & Predict vs Actual Multi-scale GMM + Auto-FE (mutual info) + XGB/RF/SVR + Weighted Ensemble

Finally, automated domain-guided feature engineering, driven by mutual information, generates chemically meaningful interaction terms, including HOMO × LUMO, $(\Delta E)^2$, polarizability × logP, and Tanimoto similarity to cluster centroids.

Each stage was independently reinitialized for each LOOCV fold, generating 700–900 high-quality synthetic samples without leakage. The resulting actual-versus-predicted plot (Figure 5) shows strong alignment with the diagonal, with only two minor outliers above 92% IE, while the residuals closely resemble a normal distribution with minimal skewness reflecting a balanced, stable, and unbiased model.

Figure 6 offers a detailed comparison of the performance of 13 different model configurations on a fixed external test set comprised of 5 independent molecules, utilizing four key regression metrics: coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). The models are organized in descending order of R^2 to allow for quick identification of performance rankings.

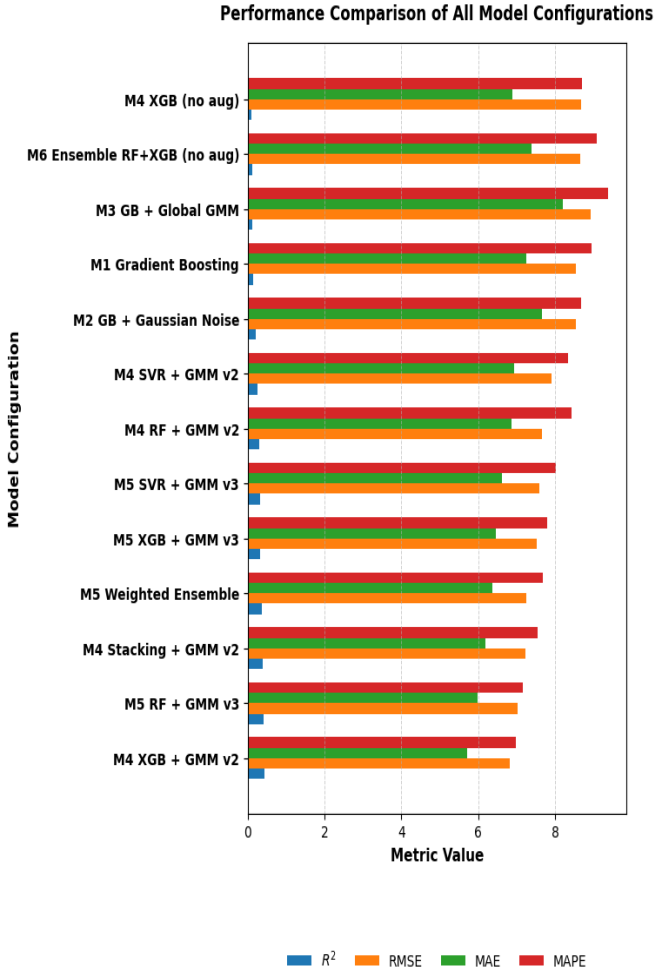

Figure 6. Model Performace Comparison

A marked superiority of configurations featuring multi-scale hierarchical Gaussian Mixture Model (GMM) augmentation is evident. The leading model is M4 XGBoost + GMM v2 (R^2 = 0.4457, RMSE = 6.8214, MAE = 5.7189, MAPE = 6.99%), closely followed by M5 Random Forest + GMM v3 (R^2 = 0.4108, RMSE = 7.0331, MAE = 5.9846, MAPE = 7.15%). These two models show R^2 enhancements surpassing 300% in comparison to non-augmented baselines (M4 XGBoost without augmentation: R^2 = 0.1053; M1 Gradient Boosting: R^2 = 0.1316) and over 200% when contrasted with simplistic augmentation techniques like Gaussian noise injection (M2: R^2 = 0.2005) or a single global GMM (M3: R^2 = 0.1236). All configurations that use GMM v2 and v3 (M4 and M5 series) consistently secure top-eight rankings, whereas models that do not incorporate advanced augmentation or rely on basic methods are placed in the bottom seven positions.

This pattern clearly indicates that the performance enhancements observed are not limited to specific algorithms but stem from the multi-scale GMM augmentation framework's ability to generate chemically realistic synthetic samples that effectively address the significant predictive

information gap in a minimal dataset (n = 20). The hierarchical approach—comprising global GMM fitting, local HDBSCAN clustering, micro-GMM refinement, selective enhancement of critical descriptors, and automated mutual-information-focused feature engineering delivers consistent benefits across various base learners (XGBoost, Random Forest, SVR) and ensemble frameworks (stacking and weighted ensembles).

Although the maximum R^2 values remain within the 0.41–0.45 range suggesting that over 55% of the variance remains unaccounted for the absolute and relative enhancements achieved are of considerable practical importance given the minimal dataset. Permutation testing (10,000 iterations) on the external test set validated that the reductions in MAE and MAPE achieved by the two top models (M4 XGBoost + GMM v2 and M5 Random Forest + GMM v3) compared to the baseline were statistically significant ($p < 0.05$), thus negating the possibility that the performance improvements occurred by chance.

As a result, while these models have not yet reached the quantitative precision needed for high-accuracy absolute predictions, they provide substantial strategic value as supportive tools for early-stage screening, synthesis prioritization, and rank-ordering corrosion inhibitor candidates amid significant experimental or computational data limitations. The findings robustly support the suggestion that leakage-free, hierarchical, multi-scale GMM augmentation when applied within a rigorously verified pipeline serves as a practical, immediately effective methodology for significantly improving the predictive power of QSAR/QSPR models in ultra-small-data scenarios, especially during the preliminary stages of quantum-computation-driven corrosion inhibitor discovery.

IV. Conclusion

This research introduces a QSPR modeling framework that is fully protected against data leakage through fold-specific preprocessing, exclusive training-set augmentation, and comprehensive Leave-One-Out Cross Validation (LOOCV). The results indicate that all baseline models lacking augmentation specifically Gradient Boosting, Random Forest, SVR, and XGBoost show limited predictive power when applied to a small dataset ($R^2 < 0.20$), underscoring the intrinsically low information complexity of the original data (n = 20).

Utilizing a multi-scale Gaussian Mixture Model (GMM) consistently improves model performance. The two most successful configurations, XGBoost + GMM v2 and Random Forest + GMM v3, attain R^2 values of 0.4457 and 0.4108, respectively, along with statistically significant reductions in RMSE, MAE, and MAPE. These enhancements are supported by more balanced residual distributions, lower variance, and better correlation between predicted and observed outcomes.

These results suggest that GMM-driven generative augmentation effectively captures the multi-cluster distribution of quantum descriptors and generates chemically valid synthetic samples. Consequently, the observed improvements mainly stem from a meticulously managed expansion of the feature space rather than from hyperparameter adjustments or increased model complexity.

Despite the highest R^2 values falling within the 0.40–0.45 range deemed inadequate for high-precision quantitative prediction the proposed pipeline shows considerable practical potential as an initial screening tool for prioritizing corrosion inhibitors. The methodology is widely applicable to different ultra-low-data situations in computational chemistry.

Future research will aim to incorporate advanced DFT-derived descriptors, integrate molecular graph representations, and validate the methodology using larger external datasets to further enhance model generalizability.

Acknowledgment

The authors wish to convey their heartfelt appreciation to their colleagues at the Quantum Matics Laboratory, Universitas Dian Nuswantoro (UDINUS), for the technical assistance they offered during the research process. This institution's support and partnership were vital in the successful completion of this study.

References

[1] A. N, D. R, and R. S, "Curcumin and Curcumin Derivatives as Green Corrosion Inhabitor-A Review," *Phys. Chem. Res.*, vol. 11, no. 4, Dec. 2023, doi: 10.22036/pcr.2022.362856.2199.

[2] G. N. Sajida, G. M. Krista, H. K. Sari, T. Taufiqurohim, Y. F. Ferawati, and R. P. Sihombing, "Potensi Ekstrak Kunyit sebagai Inhibitor Korosi Ramah Lingkungan untuk Baja Karbon Rendah," *J. Teknol.*, vol. 25, no. 2, 2025, doi: http://dx.doi.org/10.30811/teknologi.v25i2.7483.

[3] M. Akrom, "INVESTIGATION OF NATURAL EXTRACTS AS GREEN CORROSION INHIBITORS IN STEEL USING DENSITY FUNCTIONAL THEORY," *J. Teori Dan Apl. Fis.*, vol. 10, no. 1, p. 89, Jan. 2022, doi: 10.23960/jtaf.v10i1.2927.

[4] M. Akrom, "Experimental Investigation of Natural Plant Extracts as A Green Corrosion Inhibitor in Steel," *J. Renew. Energy Mech.*, vol. 5, no. 01, pp. 1–15, Feb. 2022, doi: 10.25299/rem.2022.8887.

[5] M. Akrom and T. Sutojo, "Investigasi Model Machine Learning Berbasis QSPR pada Inhibitor Korosi Pirimidin," *Eksergi*, vol. 20, no. 2, p. 107, July 2023, doi: 10.31315/e.v20i2.9864.

[6] V. F. Adiprasetya, M. Akrom, and G. A. Trisnapradika, "Investigasi Efisiensi Penghambatan Korosi Senyawa Quinoxaline Berbasis Machine Learning," *Eksergi*, vol. 21, no. 2, p. 65, Mar. 2024, doi: 10.31315/e.v21i2.10025.

[7] J. F. Fatriansyah *et al.*, "A machine learning framework for screening phenyl phthalimide derivatives as corrosion inhibitors based on dataset generated by DFT and molecular dynamics simulations," *Results Eng.*, vol. 28, p. 107350, Dec. 2025, doi: 10.1016/j.rineng.2025.107350.

[8] T. H. Pham, P. K. Le, and D. N. Son, "A data-driven QSPR model for screening organic corrosion inhibitors for carbon steel using machine learning techniques," *RSC Adv.*, vol. 14, no. 16, pp. 11157–11168, 2024, doi: 10.1039/D4RA02159B.

[9] N. U. S. Riyaz, M. Khaled, A. Alshami, and I. A. Hussein, "Machine Learning-Driven Prediction of Corrosion Inhibitor Efficiency: Emerging Algorithms, Challenges, and Future Outlooks," *Arab. J. Sci. Eng.*, July 2025, doi: 10.1007/s13369-025-10386-5.

[10] F. M. Haikal, M. Akrom, and G. A. Trisnapradika, "Perbandingan Algoritma Multilinear Regression dan Decision Tree Regressor dalam Memprediksi Efisiensi Penghambatan Korosi Piridazin," *Edumatic J. Pendidik. Inform.*, vol. 7, no. 2, pp. 307–315, Dec. 2023, doi: 10.29408/edumatic.v7i2.22127.

[11] M. Fadil, M. Akrom, and W. Herowati, "Utilization of Machine Learning for Predicting Corrosion Inhibition by Quinoxaline Compounds," *J. Appl. Inform. Comput.*, vol. 9, no. 1, pp. 173–177, Jan. 2025, doi: 10.30871/jaic.v9i1.8894.

[12] S. Ramaneswaran, K. Srinivasan, P. M. D. R. Vincent, and C.-Y. Chang, "Hybrid Inception v3 XGBoost Model for Acute Lymphoblastic Leukemia Classification," *Comput. Math. Methods Med.*, vol. 2021, pp. 1–10, July 2021, doi: 10.1155/2021/2577375.

[13] W. Herowati *et al.*, "Prediction of Corrosion Inhibition Efficiency Based on Machine Learning for Pyrimidine Compounds: A Comparative Study of Linear and Non-linear Algorithms," *KnE Eng.*, Mar. 2024, doi: 10.18502/keg.v6i1.15350.

[14] E. S. Budi, A. N. Chan, P. P. Alda, and M. A. F. Idris, "Optimasi Model Machine Learning untuk Klasifikasi dan Prediksi Citra Menggunakan Algoritma Convolutional Neural Network," vol. 4, no. 5, 2024.

[15] L. W. Rizkallah, "Enhancing the performance of gradient boosting trees on regression problems," *J. Big Data*, vol. 12, no. 1, p. 35, Feb. 2025, doi: 10.1186/s40537-025-01071-3.

[16] D. R. Ningtias and M. Akrom, "XGBoost performance in predicting corrosion inhibition efficiency of Benzimidazole Compounds," *J. Multiscale Mater. Inform.*, vol. 1, no. 2, pp. 9–13, July 2024, doi: 10.62411/jimat.v1i2.11021.

[17] G. A. Trisnapradika, U. D. Nuswantoro, and M. Akrom, "A Machine Learning Approach for Forecasting the E cacy of Pyridazine Corrosion Inhibitors".

[18] I. P. Aldiansah and M. Akrom, "Effect of Virtual Sample Generation in Predicting Corrosion Inhibition Efficiency on Pyridazine," vol. 9, no. 2, doi: https://doi.org/10.30871/jaic.v9i2.9131.

[19] N. Ariyanto, H. A. Azies, and M. Akrom, "Ensemble Stacking of Machine Learning Approach for Predicting Corrosion Inhibitor Performance of Pyridazine Compounds," *Int. J. Adv. Data Inf. Syst.*, vol. 5, no. 2, Nov. 2024, doi: 10.59395/ijadis.v5i2.1346.

[20] C. Wang, T. Shi, and D. Han, "Adaptive Dimensional Gaussian Mutation of PSO-Optimized Convolutional Neural Network Hyperparameters," *Appl. Sci.*, vol. 13, no. 7, p. 4254, Mar. 2023, doi: 10.3390/app13074254.

[21] L. Rosiana and I. Yuadi, "K-Means Clustering untuk Analisis Tren Peminjaman Buku di Perpustakaan," *J. Technol. Inform. JoTI*, vol. 7, no. 1, pp. 1–10, Apr. 2025, doi: 10.37802/joti.v7i1.933.

[22] D. Ignasius, M. Akrom, and S. Budi, "Comparative Analysis of Linear Regression, Decision Tree, and Gradient Boosting Models for Predicting Drug Corrosion Inhibition Efficiency Using QSAR Descriptors," *Fakt. Exacta*, vol. 17, no. 3, p. 251, Sept. 2024, doi: 10.30998/faktorexacta.v17i3.24679.