Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AISM-REGBF: An Interpretable Regression-Based Framework for Predicting Small- Molecule Potency

Subathra Selvam
SRM Institute of Science and Technology
Priya Dharshini Balaji
SRM Institute of Science and Technology
Thirumurthy Madhavan
thiru.murthyunom@gmail.com

SRM Institute of Science and Technology

Research Article

Keywords: AISM-REGBF, regression, anti-inflammatory, machine learning, small molecules

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11464167/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

AISM-REGBF: An Interpretable Regression-Based Framework for Predicting Small-Molecule Potency

Subathra Selvam [1]| Priya Dharshini Balaji [1] | Thirumurthy Madhavan[1]*

1. Computational Biology Laboratory, Department of Genetic Engineering, School of Bioengineering, SRM Institute of Science and Technology, Kattankulathur-603203, Tamil Nadu, India

***Corresponding author**

Dr.Thirumurthy Madhavan

Associate Professor, Computational Biology Lab,

Department of Genetic Engineering,

School of Bioengineering,

SRM Institute of Science and Technology, SRM Nagar, Kattankulathur,

Chennai 603203, Tamil Nadu, India.

Mail: thiru.murthyunom@gmail.com

Phone number: +919944572918

Abstract

Inflammation is a primary immune response categorized into acute and chronic types, with chronic inflammation linked to various diseases such as autoimmune disorders and cancer. Conventional therapies often face limitations due to side effects and therapeutic resistance, prompting the exploration of small-molecule compounds as safer alternatives. In this study we emphasize the integration of machine learning (ML), particularly through structure-activity relationship (SAR) analysis and quantitative structure-activity relationship (QSAR) modeling, to enhance drug discovery. A series of ML-based regression models, including LightGBM, were developed and systematically evaluated. Various feature selection method, were employed to effectively reduce the input dimensionality and improve model performance. The LightGBM model, based on combined Morgan and substructure count fingerprints, achieved strong predictive performance with an R^2 of 0.75 on the independent test dataset. External validation with FDA-approved compounds affirmed the model's applicability, highlighting key structural features identified through explainable artificial intelligence (XAI) techniques, specifically SHAP and LIME. This study concludes that the AISM-REGBF framework provides a reliable computational tool for predicting the anti-inflammatory potency (pIC_{50}) of small molecules.

Keywords: AISM-REGBF, regression, anti-inflammatory, machine learning, small molecules

1. Introduction

Inflammation is a vital immunological response that protects the body from pathogens, tissue injury, and other cellular stress. Inflammation is often categorized into two types: chronic

and acute, while acute inflammation is self-limiting and essential for host defence. Chronic inflammation is an uncontrolled or persistent condition that significantly contributes to the development of numerous deliberating disorders, including autoimmune disorders, neurodegenerative diseases, cardiovascular diseases, cancer, asthma, psoriasis, rheumatoid arthritis, multiple sclerosis, and diabetes [1][2]. Conventional anti-inflammatory therapeutics such as immunosuppressants, corticosteroids, and non-steroidal anti-inflammatory drugs are widely utilized [3]. Although clinically effective, these agents are frequently associated with adverse side effects and therapeutic resistance. Target specific therapeutic options have expanded, however, their high cost, limited oral bioavailability, and immunogenic potential restrict broader applicability. These limitations continue to drive the search for structurally diverse, potent, and safer small-molecule anti-inflammatory agents [4][5][6]. Recent studies have shown that small-molecule based drugs are safer and more effective inflammatory disease treatments. They are ideal clinical and research therapeutic agents due to their stability and inexpensive production costs [7]. These compounds are more effective than classic anti-inflammatory drugs and have less adverse effects, allowing for more tailored treatment [8][9].

The rapid expansion of biomedical and chemical data has accelerated the integration of machine learning (ML) approaches into drug discovery. Within this framework, structure-activity relationship (SAR) analysis provides a systematic approach to correlate molecular structure with biological activity based on experimental data [10][11]. Quantitative structure-activity relationship (QSAR) modeling further extends this concept by mathematically linking molecular descriptors to pharmacological potency, thereby enabling predictive modeling for lead identification and optimization [12][13]. In recent years, the use of artificial intelligence (AI) and machine learning (ML) methodologies into QSAR modeling has significantly enhanced the drug discovery process, facilitating the creation of more precise, resilient, and scalable predictive models [14].

In this study, we propose a computational framework named AISM-REGBF for predicting the potency of anti-inflammatory small molecules, developed using a well-curated dataset derived from our previous research work. The overall workflow was designed in accordance with our earlier studies, including InflamPred, AISMPred, and MLASM [8][9][15], and the same data pre-processing and optimization strategies were consistently applied. Four distinct categories of molecular features were generated, and feature selection was performed independently for each category using four different selection methods. Subsequently, six ML-based regression models were constructed and evaluated. Based on a comprehensive comparative analysis, the LightGBM regression model trained on the combined Morgan and

substructure count fingerprints demonstrated the most robust predictive performance and was therefore selected as the optimal model. Finally, the overall model behavior and prediction reliability were examined using explainable artificial intelligence (XAI) techniques.

2. Materials and Methods:

The proposed AISM-REGBF architecture depicted in **Figure 1** consist of several subsequent steps: (1) data collection, (2) feature generation, (3) feature selection, (4) development of machine learning models, (5) model evaluation using performance metrics, (6) Murcko scaffold analysis, (7) external validation, and (8) Interpretability analysis.

2.1 Construction of dataset

A comprehensive dataset employed for the construction of AISM-REGBF was retrieved from our previously reported AISMPred research, which was originally curated from the PubChem database (https://pubchem.ncbi.nlm.nih.gov/). A total of 4300 anti-inflammatory small molecules, documented with experimentally validated IC_{50} values for inflammation-related biological targets. The chemical structures were represented in canonical SMILES (simplified molecular-input line-entry system) format. To ensure data quality, duplicates and noisy entries, such as incomplete structures, missing IC_{50} values entities were removed, yielding a refined dataset [16]. The IC_{50} values were converted into pIC_{50} values by a negative logarithmic transformation, to standardize modeling robustness and diminish activity bias. The pIC_{50} values obtained varied from 3.0 to 10.0, encompassing compounds with weak to strong anti-inflammatory potency and accelerating effective regression-based learning. The entire dataset was randomly split into a training set and a test set (80:20), maintaining the data distribution with the reduced selection bias. **Figure 2a** represents the boxplot of pIC50 values of the biological activity distribution of 4300 anti-inflammatory compounds, and **Figure 2b** boxplot illustrates the distribution of relative molecular weight (MW) for these compounds, primarily ranging from 200 to 600 g/mol. The dataset structural diversity was further investigated using plain ring system and Murcko scaffold analysis [17][18]. Across 4,300 anti-inflammatory small compounds, 258 unique simple ring systems and 2,397 distinct Murcko scaffolds were found, underscoring evidential chemical diversity. Among these, 82 plain ring systems were identified with a frequency exceeding 10, signifying frequently repeating structural patterns. **Figure 3** indicates the typical structures of the most commonly seen plain ring systems, whereas **Table 1** presents the distribution of predominant Murcko scaffolds with their frequency count, highlighting the structural diversity within the dataset.

2.2 Feature generation

Initially, a diverse set of descriptors and features were calculated using the RDKit package in Python, and PaDel software was employed to capture the significant chemical and physical properties relevant to identifying anti-inflammatory activity [19][20]. These descriptors are critical for constructing QSAR models, which aim to predict the biological activities of unknown compounds based on their structural characteristics [21][22][8]. The SMILES representation of each compound acted as an input, while different types of molecular characteristics were evaluated, namely 2D descriptor features that encapsulate fundamental physicochemical and structural characteristics of the compounds. And molecular fingerprints that represent substructure patterns significant to biological activity. Initially, the 2D descriptors are calculated using RDKit, and fingerprints are generated using PaDel software. For this study, three types of fingerprints, such as Morgan, substructure count, and CDK-based extended fingerprints, were chosen [23].

2.3 Feature Selection method

A meticulous feature reduction method was carried out to finalize the most significant features [24]. As the dataset comprises four distinct types of features, resulting in the application of four feature selection methods. The dataset initially had 217 2D descriptors, 2048 Morgan fingerprint bits, 317 substructure count descriptors, and 1024 CDK extended fingerprint characteristics. Further application of the selection method reduced the feature space to 86 2D descriptors, 947 Morgan fingerprints, 92 substructure count features, and 539 CDK extended fingerprints. Ensuring the most relevant features have been recruited for model building assuring the importance to anti-inflammatory activity.

2.4 Model building and hyperparameter optimization

In our framework, four conventional and advanced kernel and ensemble-based machine learning algorithms were employed to construct regression models, including Support Vector Regression (SVR)[25], Light Gradient Boosting Machine (LightGBM)[26], Extreme Gradient Boosting (XGBoost)[27], Random Forest (RF)[28]. All models were implemented using the scikit-learn library [29]. A 5- fold cross validation was employed[30], the training set was randomly divided into 5 subsets with equal size. Four subsets were used to build the models, and the other one was used to verify the data and investigate how effectively the models performed. Five iterations of this cross-validation procedure were conducted to ensure that every subgroup had been validated once. And also, extensive hyperparameter optimization was conducted throughout model construction to refine each model performance, generalization ability and to mitigate model underfitting or overfitting.

2.5 Performance evaluation of regression models

The performance of the developed regression models was evaluated using statistical data like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2) [31][32][33].

$$RMSE=\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_{pred}^{i}-y_{exp}^{i}\right)^2}$$

$$MAE=\frac{1}{m}\sum_{i=1}^{m}\left|y_{pred}^{i}-y_{exp}^{i}\right|$$

$$R^2=1-\frac{\sum_{i=1}^{m}\left(y_{pred}^{i}-y_{exp}^{i}\right)^2}{\sum_{i=1}^{m}\left(y_{pred}^{i}-\bar{y}\right)^2}$$

Here, *Yexp* and *Ypred* refer to the experimental and predicted values, respectively, *Y* represents the mean of the experimental observations, and m denotes the sample size [34][35].

2.6 Model interpretation and feature importance

To enhance the interpretability of the proposed predictive architecture and to evaluate the individual features influence, two explainable artificial intelligence (XAI) techniques, SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) were employed [36][37][9][38]. SHAP describes the outline feature roles for all instances and a specific instance locally and globally. While LIME explains how the model operates locally for a single instance. Together SHAP and LIME analyzes collectively validated the significance of the chosen criteria and enhanced the dependability of the proposed model in precisely finding the potency of anti-inflammatory small compounds.

3. Results and Discussion

3.1 Feature Optimization and Dimensionality Reduction

In the development of the AISM-REGBF framework, accurate representation of molecular features plays a critical role in constructing a robust regression-based predictive model. Four distinct categories of molecular descriptors, 2D physicochemical descriptors, Morgan fingerprints, substructure count fingerprints, and CDK extended fingerprints along with their respective combinations, were initially considered as model inputs. A total of 15 different

feature combinations were generated and systematically evaluated, and the retained features for each combination are summarized in **Table 2.**

The feature optimization strategy for 2D molecular descriptors employed a mutual information-based feature selection method to discover the most informative variables. Morgan fingerprints were enhanced by a variance threshold method to remove low-information components. Feature selection for the substructure count fingerprints was conducted with XGBoost-based feature priority ranking, whereas Elastic Net regularization was applied to identify important features from the CDK extended fingerprints.

As a result of this feature selection process, the original feature space was substantially reduced to 86 for 2D descriptors, 947 for Morgan fingerprints, 92 for substructure count features, and 539 for CDK extended fingerprints. A number of feature combinations were employed to construct regression predicting models from the 1,664 filtered features. For model building and evaluation, the final dataset was randomly split into 80:20 ratio for training and independent test sets.

3.2 Regression Model Development and Performance Evaluation

A suite of machine learning algorithms, Random Forest (RF), Support Vector Regression (SVR), LightGBM (LGBM), and Extreme Gradient Boosting (XGBoost), was applied to identify the most predictive and interpretable model. The prediction efficacy of all 15 feature combinations from 6 distinct model methods was built and evaluated using cross-validation on the training set and subsequently confirmed on an independent test set. The performance metrics of all developed models utilizing different feature sets are detailed in Supplementary **Table 1.** The results indicate that both the individual fingerprint types and selected combinations demonstrate strong predictive capability. Among the combined feature sets, models integrating Morgan and substructure count fingerprints, as well as those combining Morgan, substructure count, and CDK extended fingerprints, exhibited superior performance. This improvement can be attributed to the enhanced ability of combined fingerprints to capture complementary structural information and better represent the relationship between molecular features and molecular bioactivity.

In the preliminary evaluation, the Morgan, substructure count, and CDK extended combined fingerprint set exhibited robust predictive efficacy, attaining a training performance of R^2 of 0.685, MAE of 0.485, and RMSE of 0.677, whereas the independent test set results in R^2 of 0.742, MAE of 0.446, and RMSE of 0.616. These findings demonstrate that hybrid feature representations can proficiently encapsulate complementing structural information.

Nevertheless, taking consideration of both predictive accuracy and computational efficiency, especially the reduced number of input features, the combination of Morgan and substructure count fingerprints (Morgan + Sub) was selected as the most ideal input for the final model. **Table 3** summarizes the performance metrics acquired throughout the training phase and the independent test validation. Feature sets incorporating Morgan and substructure fingerprints tend to outperform other individual and hybrid combinations, indicating that these descriptors offer a more informative and discriminative depiction of structural patterns associated with anti-inflammatory potency. **Figure 4** illustrates the performance of the LightGBM model constructed with the Morgan + Sub feature combination. The model achieved a R^2 of 0.69, MAE of 0.48, and RMSE of 0.67 on the training set, while the independent test set exhibited R^2 of 0.75, MAE of 0.44, and RMSE of 0.61. These results indicate that the AISM-REGBF architecture has high prediction accuracy, generalization, and durability.

3.3 External Validation of the AISM-REGBF Regression Model

An external validation was carried out to evaluate the robustness of the proposed LightGBM regression model, employing a dataset of 1300 FDA-approved drugs obtained from the ZINC database. Validation on a larger, independent dataset enables a more reliable evaluation of the model's relevance in real-world contexts. To ensure consistency in the training workflow, molecular fingerprints for the external compounds have been generated using the same computational approach utilized during model construction. Only the features previously identified by the variance threshold approach for Morgan fingerprinting and XGBoost-based feature selection for substructure count fingerprints were preserved, and these two feature sets have combined as the final input for prediction.

The projected pIC_{50} values derived from the AISM-REGBF framework varied from 3 to 8, indicating a progressive shift from diminished to enhanced anti-inflammatory action. Numerous FDA-approved medicines, such as JAK inhibitors, tyrosine kinase inhibitors, and traditional anti-inflammatory medications, have significant expected efficacy. Many of these chemicals are recognized for their ability to regulate critical inflammatory signaling pathways, hence affirming the biological significance of the predictions. **Table 4** represents the FDA-approved drugs with the highest anticipated activity.

3.4 Model interpretability and feature contribution analysis using XAI

Figure 4a illustrates the ten most influential molecular features contributing to the prediction of anti-inflammatory activity using SHAP analysis. Among these, SubFPC307, corresponding to the presence of a chiral center, emerges as the most significant feature, followed closely by

SubFPC300, which represents 1,3-tautomerizable functionality. Additional key contributors include SubFPC302 (rotatable bonds), SubFPC295 (CONS bond patterns), SubFPC2 (secondary carbon), SubFPC1 (primary carbon), SubFPC275 (heterocyclic structures), SubFPC274 (aromatic rings), and SubFPC3 (tertiary carbon).

These findings highlight the crucial role of structural and substructural characteristics in governing anti-inflammatory potency. Chiral centers can significantly influence molecular orientation and stereochemical compatibility with biological targets, thereby affecting activity. The presence of 1,3-tautomerizable groups reflects the ability of molecules to interconvert between structural forms, which may alter hydrogen bonding patterns and enhance binding interactions. Rotatable bonds contribute to molecular flexibility, enabling compounds to adopt conformations favorable for target engagement. CONS bond patterns are indicative of stable connectivity frameworks that can support specific molecular recognition.

Furthermore, different carbon atom environments primary, secondary, and tertiary play a vital role in defining the molecular scaffold and steric architecture. Aromatic rings and heterocyclic moieties are particularly important, as they frequently participate in π–π interactions and hydrogen bonding, which are essential for stabilizing ligand-target interactions. Collectively, these features underscore the structural determinants underlying anti-inflammatory activity and validate the relevance of the descriptors identified by the SHAP analysis.

Figure 4b presents the LIME-based local explanation for a representative compound predicted by the LightGBM regression model, highlighting the most influential features contributing to the predicted pIC_{50} value. Several Morgan fingerprint bits (MFP) and substructure count features (SubFPC) were identified as key contributors, consistent with the global SHAP analysis. Notably, features such as MFP_FP_226, MFP_FP_672, MFP_FP_1601, and MFP_FP_1434 exhibit strong negative contributions, indicating the absence of specific structural motifs that may reduce anti-inflammatory potency. In contrast, MFP_FP_1509, MFP_FP_331, MFP_FP_1810, and SubFPC55 show positive contributions, suggesting that the presence of these molecular fragments enhances predicted activity.

Substructure-based features identified by LIME, including SubFPC291, SubFPC62, SubFPC197, and SubFPC55, correspond to chemically meaningful motifs such as specific carbon environments, connectivity patterns, and functional group arrangements. These features are closely related to aromatic frameworks, carbon hybridization states, and scaffold rigidity, which are known to influence molecular stability and interaction with inflammatory targets. The consistency between LIME and SHAP explanations confirms that both global and local interpretations emphasize similar structural determinants, reinforcing the reliability and

interpretability of the proposed AISM-REGBF regression model for predicting anti-inflammatory activity.

4. Conclusion

In this study, we developed a robust and efficient predictive framework for predicitng the pIC50 values of anti-inflammatory small molecules (AISMs). Emphasizing the importance of feature selection and optimal feature combinations, our modeling strategy employed a hybrid representation integrating Morgan fingerprints and substructure count features, which substantially enhanced predictive performance. Among the evaluated models, the LightGBM regression model demonstrated strong consistency and stability, achieving a superior coefficient of determination (R^2) of approximately 0.75 on the independent test dataset. External validation further confirmed the applicability of the proposed framework. Overall, the AISM-REGBF framework provides a reliable and efficient tool for predicting the anti-inflammatory potency (pIC_{50}) of small molecules, facilitating data-driven prioritization in drug discovery.

5. Data availability

All the datasets generated for this study are available at https://github.com/Subathra15/AISMPred/tree/main

Funding Declaration

The authors received no specific funding for this research.

Ethics Declaration

Not applicable.

Orchid

Subathra Selvam https://orcid.org/0009-0004-0850-7997

Priya Dharshini Balaji https://orcid.org/0000-0002-7213-6084

Thirumurthy Madhavan http://orcid.org/0000-0001-5625-1210

Reference

[1] R. Medzhitov, “Origin and physiological roles of inflammation.,” *Nature*, vol. 454, no. 7203, pp. 428–435, Jul. 2008, doi: 10.1038/nature07201.

[2] D. Furman *et al.*, “Chronic inflammation in the etiology of disease across the life span.,” *Nat. Med.*, vol. 25, no. 12, pp. 1822–1832, Dec. 2019, doi: 10.1038/s41591-019-0675-0.

[3] J. Lu *et al.*, “Machine learning risk prediction model for acute coronary syndrome and death from use of non-steroidal anti-inflammatory drugs in administrative data.,” *Sci. Rep.*, vol. 11, no. 1, p. 18314, Sep. 2021, doi: 10.1038/s41598-021-97643-3.

[4] J. Guan *et al.*, “Predicting Anti-inflammatory Peptides by Ensemble Machine Learning and Deep Learning.,” *J. Chem. Inf. Model.*, vol. 63, no. 24, pp. 7886–7898, Dec. 2023, doi: 10.1021/acs.jcim.3c01602.

[5] R. Zhang *et al.*, “InflamNat: web-based database and predictor of anti-inflammatory natural products.,” *J. Cheminform.*, vol. 14, no. 1, p. 30, Jun. 2022, doi: 10.1186/s13321-022-00608-5.

[6] A. Raza, J. Uddin, A. Almuhaimeed, S. Akbar, Q. Zou, and A. Ahmad, “AIPs-SnTCN: Predicting Anti-Inflammatory Peptides Using fastText and Transformer Encoder-Based Hybrid Word Embedding with Self-Normalized Temporal Convolutional Networks.,” *J. Chem. Inf. Model.*, vol. 63, no. 21, pp. 6537–6554, Nov. 2023, doi: 10.1021/acs.jcim.3c01563.

[7] S. Lei, X. Chen, J. Wu, X. Duan, and K. Men, “Small molecules in the treatment of COVID-19,” *Signal Transduct. Target. Ther.*, vol. 7, no. 1, p. 387, 2022, doi: 10.1038/s41392-022-01249-8.

[8] S. Selvam, P. D. Balaji, H. Sohn, and T. Madhavan, “AISMPred: A Machine Learning Approach for Predicting Anti-Inflammatory Small Molecules.,” *Pharmaceuticals (Basel).*, vol. 17, no. 12, Dec. 2024, doi: 10.3390/ph17121693.

[9] S. Selvam, P. D. Balaji, R. A. Uthra, C. G. Anupama, H. Sohn, and T. Madhavan, “InFlamPred: a machine learning framework for anti-inflammatory small molecule prediction,” *Chem. Pap.*, vol. 79, no. 9, pp. 5957–5969, 2025, doi: 10.1007/s11696-025-04165-7.

[10] N. A. Meanwell, “Improving drug candidates by design: a focus on physicochemical properties as a means of improving compound disposition and safety.,” *Chem. Res. Toxicol.*, vol. 24, no. 9, pp. 1420–1456, Sep. 2011, doi: 10.1021/tx200211v.

[11] A. Cherkasov *et al.*, “QSAR modeling: where have you been? Where are you going to?,” *J. Med. Chem.*, vol. 57, no. 12, pp. 4977–5010, Jun. 2014, doi: 10.1021/jm4004285.

[12] P. Mishra, S. Vipsita, and T. K. Sahoo, “CLPCA-enhanced QSAR modelling for high-

throughput prediction of SARS bioactivity inhibitors," *Netw. Model. Anal. Heal. Informatics Bioinforma.*, vol. 14, no. 1, p. 128, 2025, doi: 10.1007/s13721-025-00617-y.

[13] M. Koirala, L. Yan, Z. Mohamed, and M. DiPaola, "AI-Integrated QSAR Modeling for Enhanced Drug Discovery: From Classical Approaches to Deep Learning and Structural Insight.," *Int. J. Mol. Sci.*, vol. 26, no. 19, Sep. 2025, doi: 10.3390/ijms26199384.

[14] P.-Y. Wu *et al.*, "Development of machine learning-based quantitative structure-activity relationship models for predicting plasma half-lives of drugs in six common food animal species.," *Toxicol. Sci.*, vol. 203, no. 1, pp. 52–66, Jan. 2025, doi: 10.1093/toxsci/kfae125.

[15] P. D. Balaji, S. Selvam, H. Sohn, and T. Madhavan, "MLASM: Machine learning based prediction of anticancer small molecules.," *Mol. Divers.*, vol. 28, no. 4, pp. 2153–2161, Aug. 2024, doi: 10.1007/s11030-024-10823-x.

[16] L.-Q. Bao *et al.*, "Development of Activity Rules and Chemical Fragment Design for In Silico Discovery of AChE and BACE1 Dual Inhibitors against Alzheimer's Disease.," *Molecules*, vol. 28, no. 8, Apr. 2023, doi: 10.3390/molecules28083588.

[17] Y. Zhao, Q. Wan, and X. He, "Construction of IRAK4 inhibitor activity prediction model based on machine learning.," *Mol. Divers.*, vol. 28, no. 4, pp. 2289–2300, Aug. 2024, doi: 10.1007/s11030-024-10926-5.

[18] Y. Zhao, Y. Chen, X. Tao, Y. Wang, and F. Tang, "An integrated approach for novel PTP1B inhibitor screening: combining machine learning models, molecular docking, molecular and dynamics simulations.," *Mol. Divers.*, Jul. 2025, doi: 10.1007/s11030-025-11292-6.

[19] C. W. Yap, "PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints," *J. Comput. Chem.*, vol. 32, no. 7, pp. 1466 – 1474, 2011, doi: 10.1002/jcc.21707.

[20] H. Kaneko, "Molecular Descriptors, Structure Generation, and Inverse QSAR/QSPR Based on SELFIES.," *ACS omega*, vol. 8, no. 24, pp. 21781–21786, Jun. 2023, doi: 10.1021/acsomega.3c01332.

[21] Y. Li, D. Ma, D. Chen, and Y. Chen, "ACP-GBDT: An improved anticancer peptide identification method with gradient boosting decision tree.," *Front. Genet.*, vol. 14, p. 1165765, 2023, doi: 10.3389/fgene.2023.1165765.

[22] N. Bajiya, S. Choudhury, A. Dhall, and G. P. S. Raghava, "AntiBP3: A Method for

Predicting Antibacterial Peptides against Gram-Positive/Negative/Variable Bacteria," *Antibiotics*, vol. 13, no. 2, 2024, doi: 10.3390/antibiotics13020168.

[23] E. L. Willighagen *et al.*, "The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching.," *J. Cheminform.*, vol. 9, no. 1, p. 33, Jun. 2017, doi: 10.1186/s13321-017-0220-4.

[24] E. Odhiambo Omuya, G. Onyango Okeyo, and M. Waema Kimwele, "Feature Selection for Classification using Principal Component Analysis and Information Gain," *Expert Syst. Appl.*, vol. 174, p. 114765, 2021, doi: https://doi.org/10.1016/j.eswa.2021.114765.

[25] R. Rodríguez-Pérez and J. Bajorath, "Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery.," *J. Comput. Aided. Mol. Des.*, vol. 36, no. 5, pp. 355–362, May 2022, doi: 10.1007/s10822-022-00442-9.

[26] G. Ke *et al.*, "LightGBM: a highly efficient gradient boosting decision tree," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 3149–3157.

[27] M. Klug *et al.*, "A Gradient Boosting Machine Learning Model for Predicting Early Mortality in the Emergency Department Triage: Devising a Nine-Point Triage Score.," *J. Gen. Intern. Med.*, vol. 35, no. 1, pp. 220–227, Jan. 2020, doi: 10.1007/s11606-019-05512-7.

[28] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, 2006, doi: 10.1007/s10994-006-6226-1.

[29] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, Jan. 2012.

[30] J. Allgaier and R. Pryss, "Cross-Validation Visualized: A Narrative Guide to Advanced Methods," *Mach. Learn. Knowl. Extr.*, vol. 6, no. 2, pp. 1378–1388, 2024, doi: 10.3390/make6020065.

[31] J. Gill *et al.*, "Evaluating the performance of machine-learning regression models for pharmacokinetic drug-drug interactions.," *CPT pharmacometrics Syst. Pharmacol.*, vol. 12, no. 1, pp. 122–134, Jan. 2023, doi: 10.1002/psp4.12884.

[32] H. Khoshvaght, R. R. Permala, A. Razmjou, and M. Khiadani, "A critical review on selecting performance evaluation metrics for supervised machine learning models in wastewater quality prediction," *J. Environ. Chem. Eng.*, vol. 13, no. 6, p. 119675, 2025, doi: https://doi.org/10.1016/j.jece.2025.119675.

[33] I. B. Suleiman, “Performance of linear regression in predicting academic outcome,” *J. Electr. Syst. Inf. Technol.*, vol. 13, no. 1, p. 24, 2026, doi: 10.1186/s43067-026-00325-1.

[34] S. Bitam, M. Hamadache, and S. Hanini, “Discovery of novel natural product-derived EGFR inhibitors using multiple linear regression, stacked ensemble regression, and fingerprinting approaches.,” *J. Comput. Aided. Mol. Des.*, vol. 40, no. 1, p. 18, Dec. 2025, doi: 10.1007/s10822-025-00731-z.

[35] M. Chotaliya and S. S. Agrawal, “Toxigraphnet: a graph neural network framework for precise toxicity prediction of drug molecules.,” *J. Comput. Aided. Mol. Des.*, vol. 39, no. 2, p. 101, Oct. 2025, doi: 10.1007/s10822-025-00683-4.

[36] A. M. Salih *et al.*, “A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME,” *Adv. Intell. Syst.*, vol. 7, no. 1, p. 2400304, 2025, doi: https://doi.org/10.1002/aisy.202400304.

[37] A. Schindele *et al.*, “Interpretable machine learning for thyroid cancer recurrence predicton: Leveraging XGBoost and SHAP analysis,” *Eur. J. Radiol.*, vol. 186, p. 112049, 2025, doi: https://doi.org/10.1016/j.ejrad.2025.112049.

[38] W. Wolak, A. Plichta, and H. Orlicki, “Interpretable ensemble learning for tumor-type prediction with a SHAP-based evaluation of CatBoost and voting classifiers,” *Sci. Rep.*, vol. 16, no. 1, p. 1401, 2025, doi: 10.1038/s41598-025-31079-x.

Figure 1. Schematic representation of the AISM-REGNET workflow, including data pre-processing, dataset split, feature encoding, regression modeling, chemical space analysis, and independent test validation.

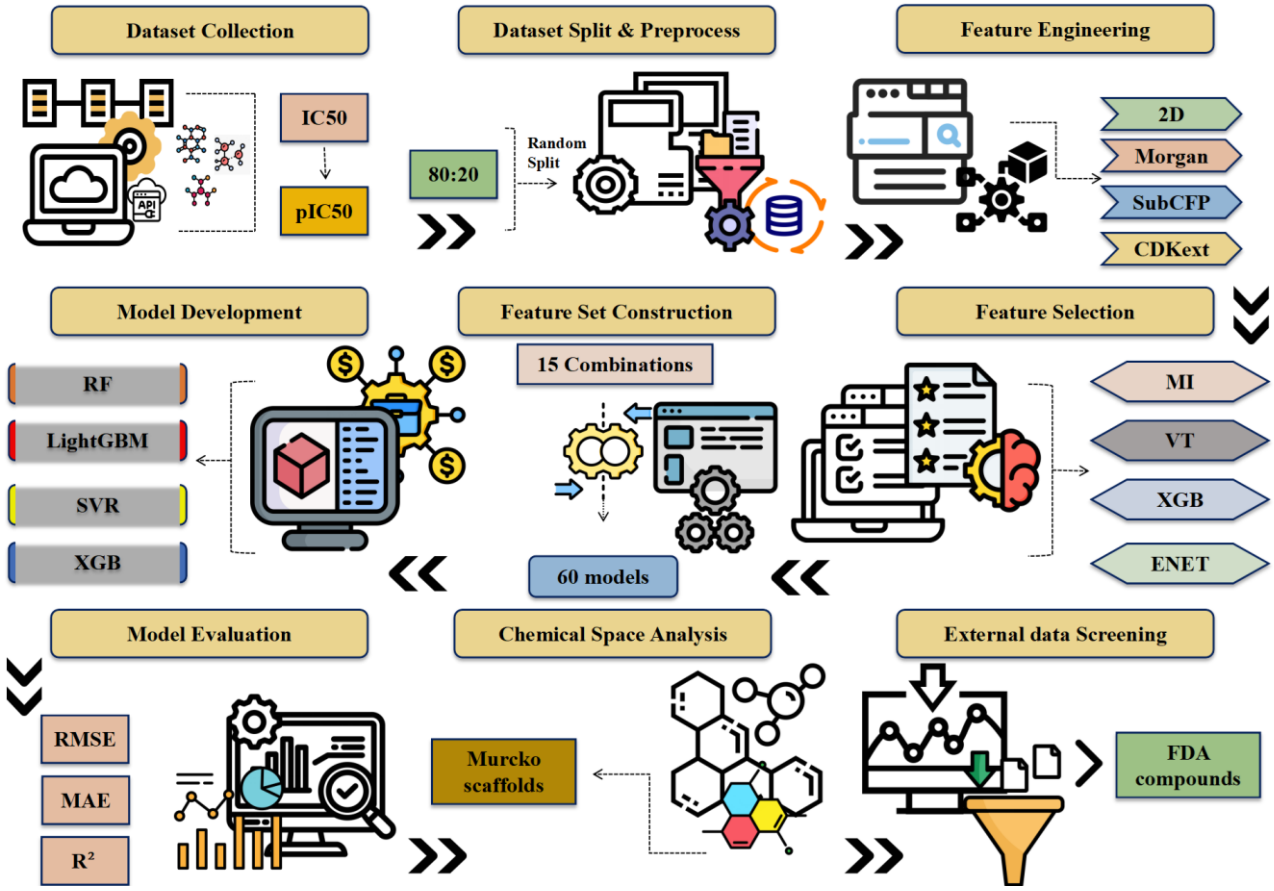


Figure 2. Statistical distribution of activity and physicochemical properties visualized using boxplots. A) Boxplot of pIC_{50} values demonstrating median activity, interquartile spread, and

potential outliers within the dataset. B) Boxplot of molecular weight (MW) values showing variability and structural diversity of compounds employed in model construction.

A) **B)**

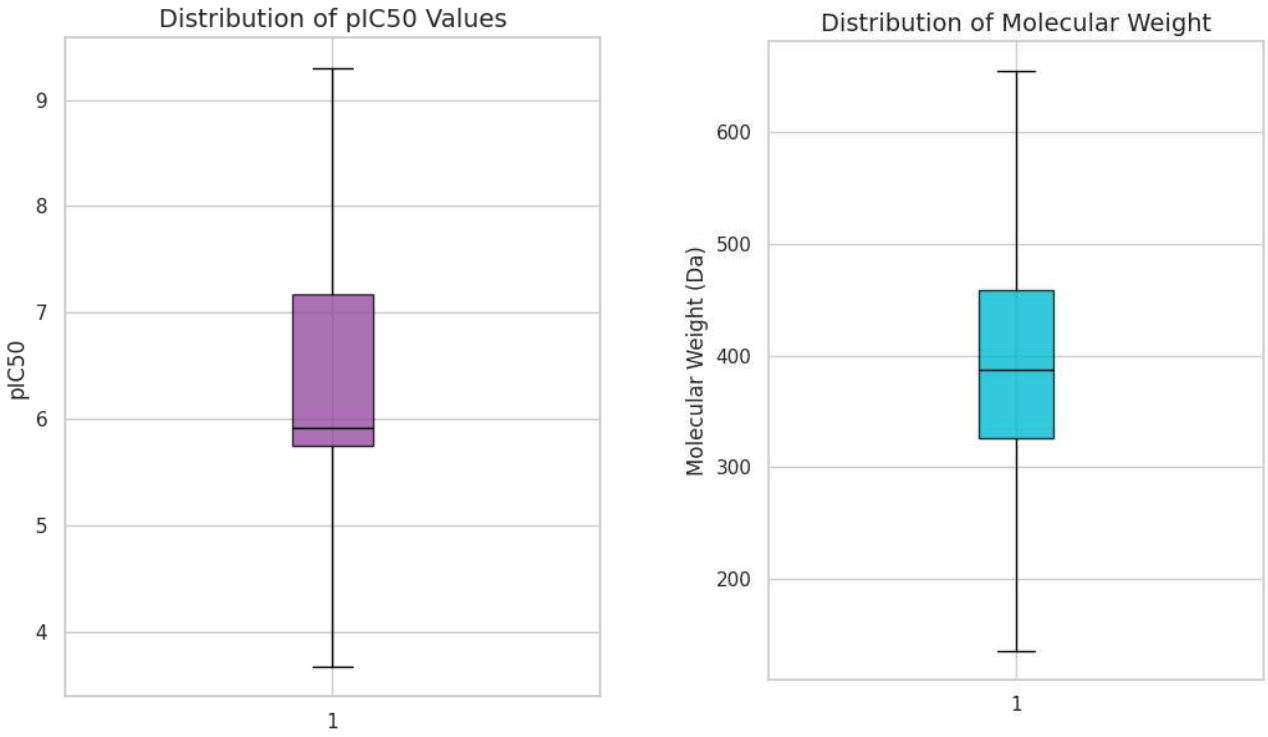


Figure 3. The top most frequently occurring plain ring systems identified through RDKit-based ring decomposition of the dataset. The structures represent dominant core motifs within the anti-inflammatory chemical space.

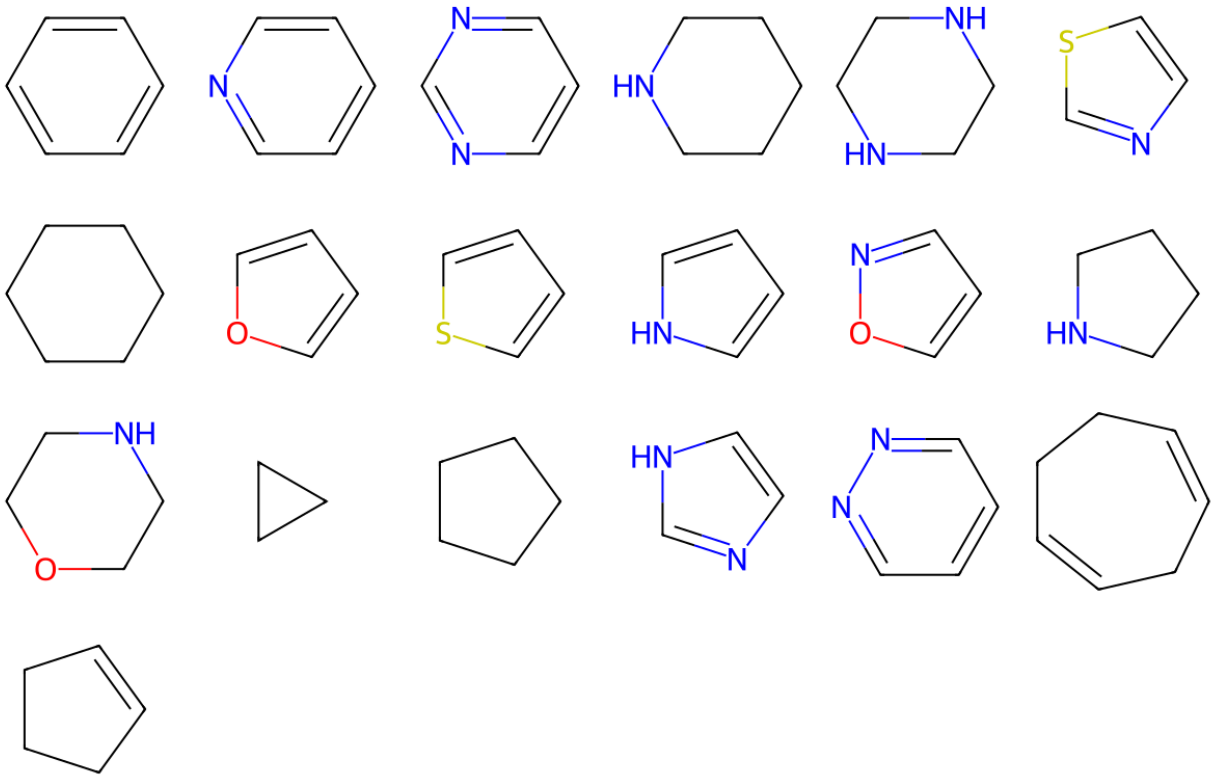


Figure 4. Regression performance visualization of the selected LightGBM model using Morgan and substructure count fingerprint features. Scatter plots depict observed versus predicted pIC_{50} values for the training and independent test sets, highlighting model robustness and generalization capability.

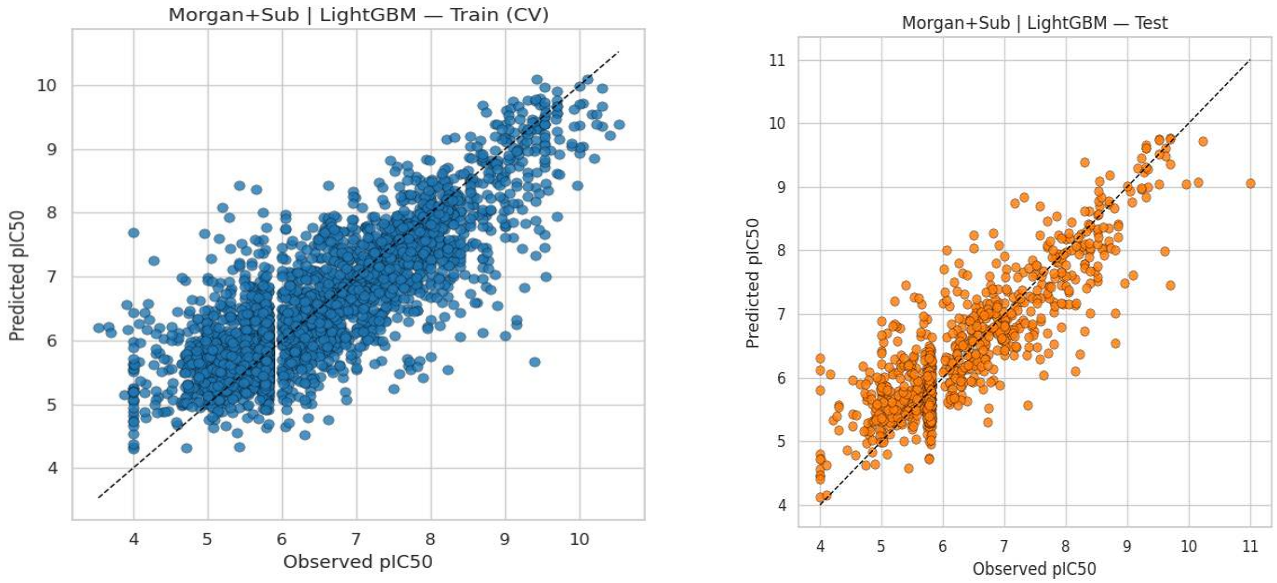
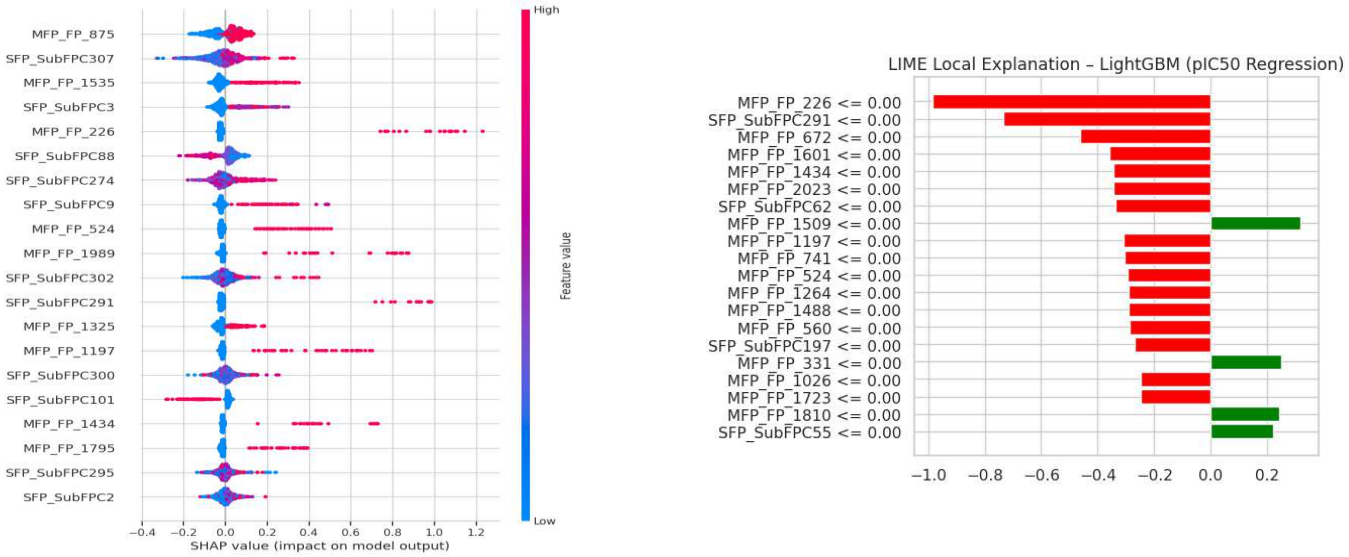


Figure 5. Model interpretability analysis of the selected LightGBM regression model. A) SHAP summary plot illustrating the global impact of molecular features on predicted pIC_{50} values. Each point represents the contribution of a feature to the model output across the dataset. B) LIME feature contribution plot showing the top 20 features influencing the predicted pIC_{50} value for an individual compound. Positive and negative contributions indicate features driving the prediction toward higher or lower activity values, respectively.

A) **B)**



Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableAISMR.docx