Bima Journal of Science and Technology, Vol. 9(2B) Jul, 2025 ISSN: 2536-6041

E-ISSN: 3115-4662 DOI: 10.64290/bima.v9i2B.1279

Quantitative Structure Activity Relationship (QSAR) Analysis of Some Pyrazole Derivatives as Hypoglycemic Agents: Computational Approach for Drugs Discovery

K. M. Aujara[1], S. Ahmed[1], A. T. Muhammad[1], I. Abubakar[1], A. I. Abubakar[1], Y. I. Mohammad[2] and A. Abdulmalik[3]

[1]Department of Science Laboratory Technology, Jigawa State Polytechnic, Dutse, Nigeria
[2]Department of Statistics, Jigawa State Polytechnic, Dutse, Nigeria
[3]Rabi'u Musa Kwankwaso Collage of Advance and Remedial Studies, Kano, Nigeria

Corresponding Author: kbaujara@gmail.co

ABSTRACT

Diabetes mellitus (DM) represents a significant global health issue marked by persistent high blood sugar levels, driving the need for innovative treatment options. This research centers on the creation and assessment of pyrazole derivatives as potential agents to lower blood sugar, employing **a** robust machine learning-based QSAR model designed to predict the inhibitory activity of compounds, utilizing RDKit for molecular descriptor calculation. A range of pyrazole derivatives sourced from the ChEMBL database were analyzed, and their inhibitory activities to reduce blood sugar levels were tested. QSAR models were constructed using Multiple linear regression (MLR) and Random Forest regression for model development, integrating molecular descriptors to identify relationships between structural characteristics and biological effectiveness. These models exhibited strong predictive capabilities, pinpointing critical structural features that enhance hypoglycemic activity, achieving an R^2 of 0.82, cross-validated correlation coefficient Q^2 of 0.80, and RMSE of 0.25 for Multiple Linear Regression and R^2 of 0.90, Q^2 of 0.85 and RMSE of 0.20 for Random Forest model. This study identified several pyrazole derivatives with promising blood sugar-lowering properties, offering a pathway for the development of new diabetes treatments. The results highlight the value of QSAR in guiding drug discovery and lay the groundwork for future preclinical and clinical studies.

Keywords: QSAR, HDAC6, Diabetes mellitus, RDKit, Multiple Linear Regression, Random Forest, ChEMBL, Machine Learning, Drug Discovery

INTRODUCTION

Diabetes mellitus (DM) is a chronic metabolic condition marked by prolonged high blood sugar level (Liu et al., 2020), stemming from either insufficient insulin production, reduced insulin sensitivity or a combination of both (Sukurai *et al.*, 2017). As a global prevalence continues to escalate, there is a pressing demand for innovative and effective blood sugar-lowering agent to better manage and treat these diseases. Among the various compounds under investigation pyrazole derivatives have gained attention due to their broad pharmacological potential (Janwal & Bhardwaj, 2013), including anti-diabetic effects (Datar & Jadhav, 2014). However, designing these compounds effectively requires a thorough understanding of how their molecular structures influence biological activity and how they interact with proteins involved in glucose regulation (Schuffenhauer *et al.*,2006).

Quantitative Structure-Activity Relationship (QSAR) modeling has become a critical tool in modern drug discovery, allowing researchers to predict the biological effectiveness of compounds based on their molecular characteristics (Patel *et al.*, 2014).

This method helps identify structural elements that enhance therapeutic performance, guiding the development of new drug candidates (Du *et al*., 2008). In this research, QSAR modeling was conducted using Google Colab, a cloud-based platform (RDKit) that offers a versatile and user-friendly environment for computational studies (Carneiro *et al*., 2018; Ryzhkov *et al*., 2024). By applying machine learning techniques and statistical approaches, such as multiple linear regression (MLR) and Random Forest Regression, reliable QSAR models were created to establish connections between the structural properties of pyrazole derivatives and their ability to lower blood sugar levels (Kovdienko *et al*., 2010).

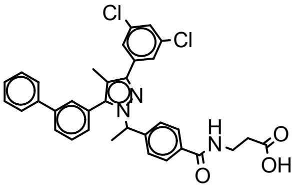
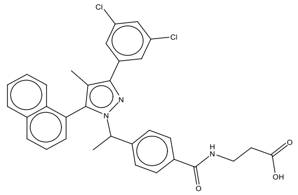
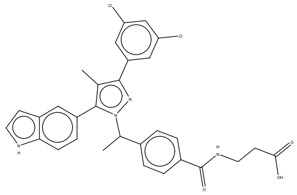
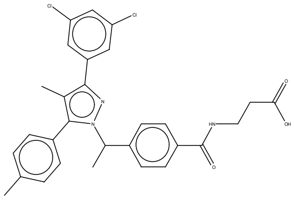
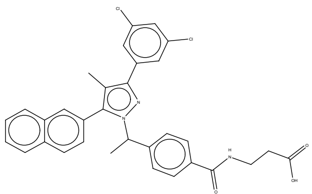
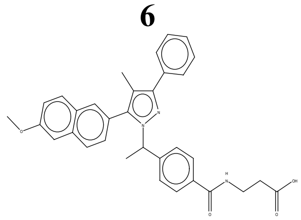
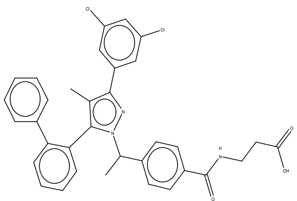
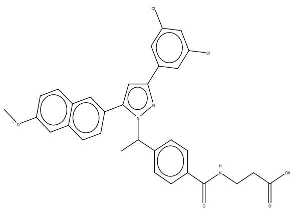
This study focuses on leveraging QSAR modeling to design and assess new pyrazole derivatives as potential agents for reducing blood sugar levels. By employing Google Colab for QSAR analysis, the research aims to identify promising compounds for further development as diabetes treatments. The outcomes of this work are expected to advance the field of rational drug design and provide a solid basis for creating effective anti-diabetic therapies.

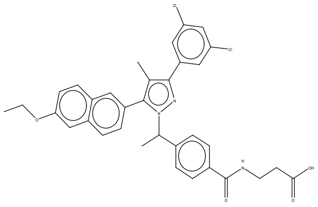
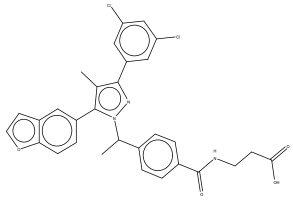
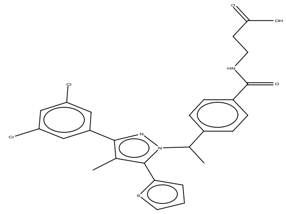
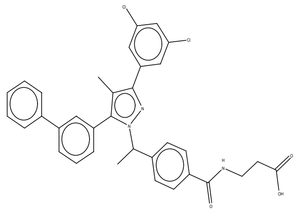
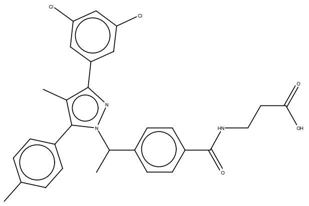
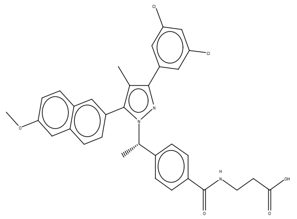
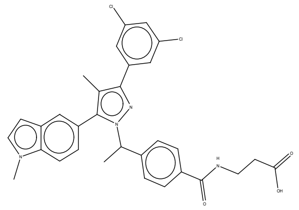
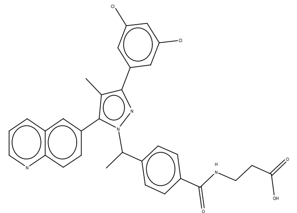
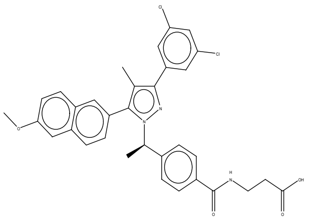
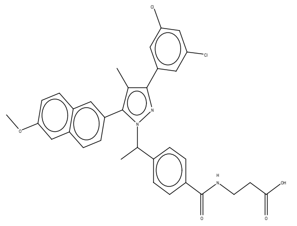
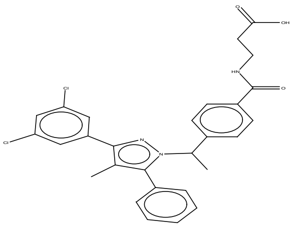
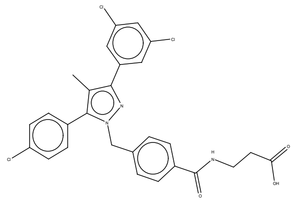
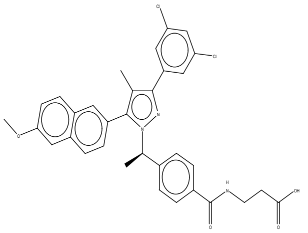
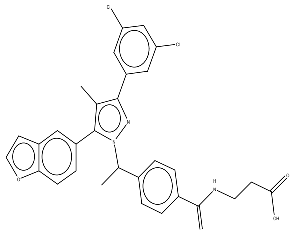
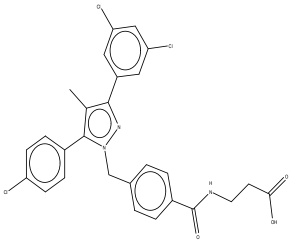
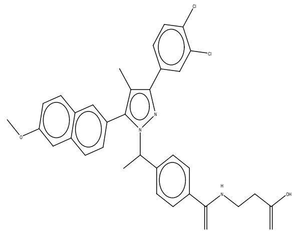
MATERIALS AND METHODS

Data Collection and Processing

The data collection process for this study was meticulously designed to gather both experimental and computational data, ensuring a robust foundation for the QSAR modeling. Hypoglycemic activity data for 52 pyrazole derivatives were obtained from the ChemBL database. ChemBL database served as a reliable source of bioactive compounds. The two-dimensional (2D) molecular structures of these derivatives were constructed using ChemDraw software, following the ACS Document 1996 guidelines to maintain compliance with established scientific and industry standards (Hassan *et al.,* 2022). Raw data often contains inconsistencies or missing values, necessitating preprocessing, this is achieved by first converting IC_{50} to pIC_{50} followed by removing invalid molecules (Tropsha, 2010). The compiled data, including structural and activity information, is presented in Table 1.

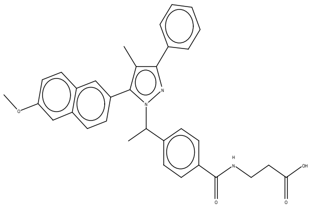
Table 1: The compiled structural and activity information.

1	2	3	4
			
6.346787	6.244125	5.408935	6.408935
5	**6**	**7**	**8**
			

6.958607	5.570248	5.826814	7.30103
9	**10**	**11**	**12**
			
6.744727	6.522879	6.259637	6.431798
13	**14**	**15**	**16**
			
6.080922	7.221849	5.636388	5.692504
17	**18**	**19**	**20**
			
6.200659	7.045757	6.69897	5.308035
21	**22**	**23**	**24**
			
5.924453	5.88941	4.563201	7.09691

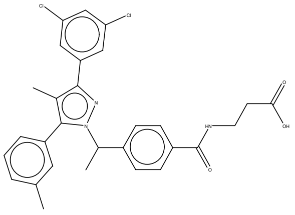
E-ISSN: 3115-4662 DOI: 10.64290/bima.v9i2B.1279

25



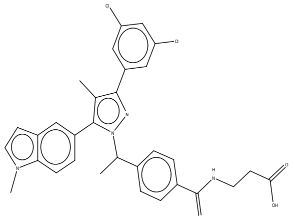
5.716699

26



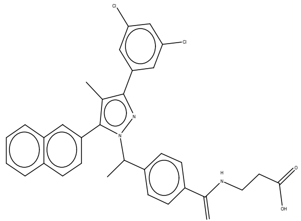
7.000000

27



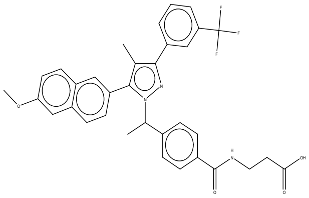
6.552842

28



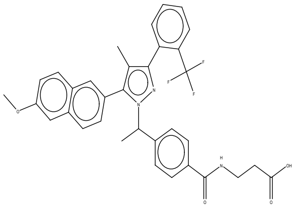
6.481486

29



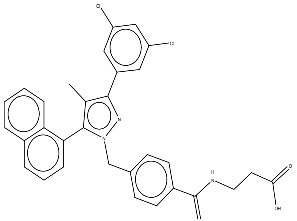
7.154902

30



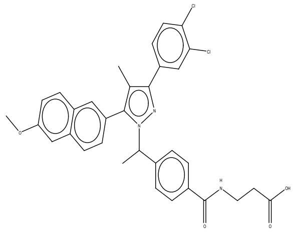
6.638272

31



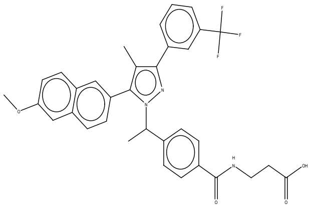
5.554396

32



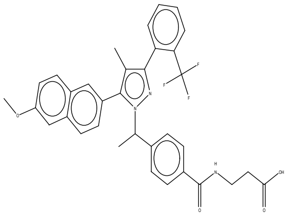
6.337242

33



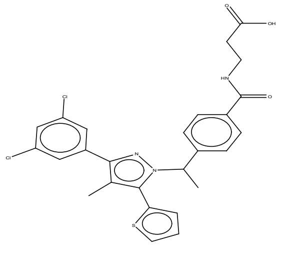
6.356547

34



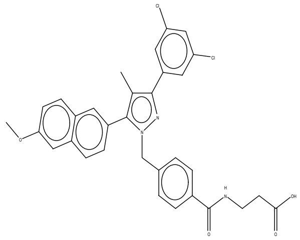
5.769551

35



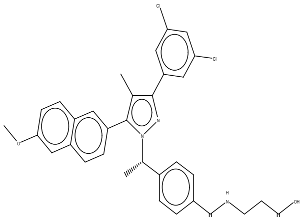
6.055517

36



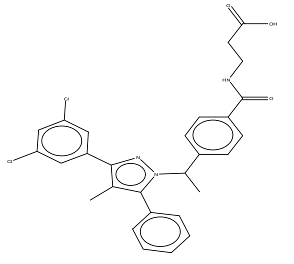
5.123205

37



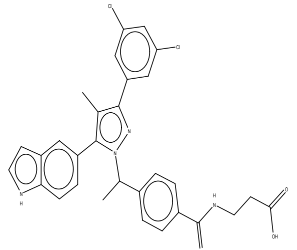
6.585027

38



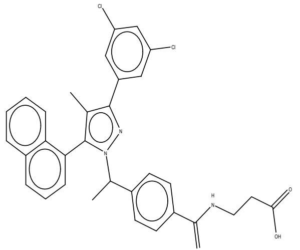
6.09691

39



5.258061

40



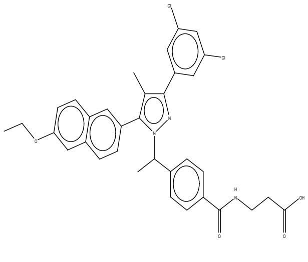
4.768021

41

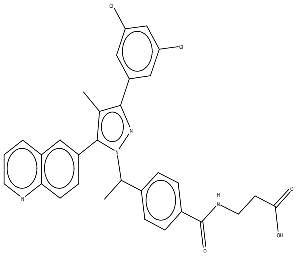
42

43

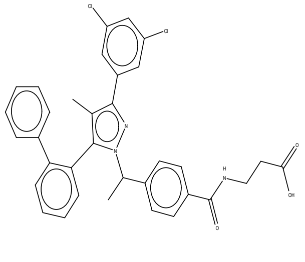
44



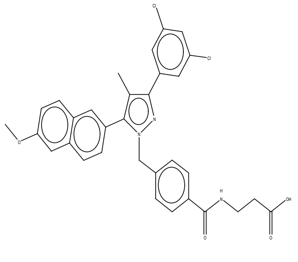
6.431798



6.30103

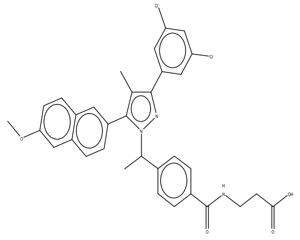


6.508638

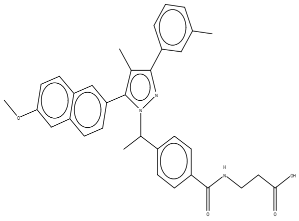


5.490797

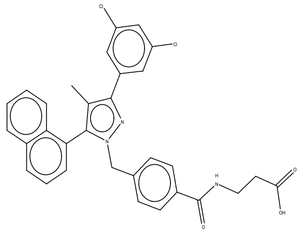
45 **46** **47** **48**



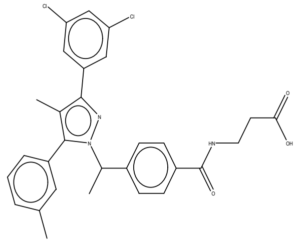
6.657577



6.130768

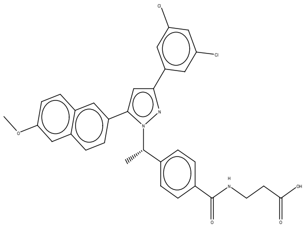


5.623423

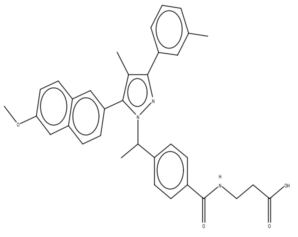


6.346787

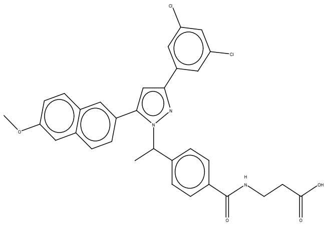
49 **50** **51** **52**



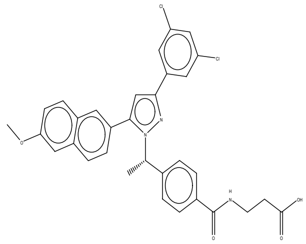
7.09691



5.560667



7.0000



7.69897

Molecular Descriptor Calculation

After signing in to Google Colab and creating a new notebook, all the necessary libraries were installed, additional ones were imported, and the dataset was loaded (Vallejo *et al*., 2022)**.** Various types of molecular descriptors for molecules are calculated by installing and importing all the necessary libraries of RDKit in Google colab using the command in the Table 2 below. Molecular descriptors quantify structural and physicochemical properties**,** RDKit provides 200+ descriptors (Kumar, 2024).

Table 2: Libraries import and installation command.

Python
!pip install rdkit-pypi pandas numpy scikit- learn matplotlib seaborn

These tools generated a wide range of descriptors, including topological, geometric, electronic, and physicochemical properties (Yao *et al*., 2022). Redundant or irrelevant descriptors were removed, and a subset of descriptors was selected based on their relevance to the biological activity and their contribution to the QSAR model (Vallejo et al., 2022; Chew et al., 2024).

Correlation analysis was used to eliminate highly correlated or low-variance descriptors and techniques like Random Forest was employed to identify the most relevant descriptors for predicting biological activity (Chalkha *et al*., 2022). These descriptors are used for training machine learning models.

Training the QSAR Model

Training a QSAR (Quantitative Structure-Activity Relationship) model involves several steps, from data preprocessing to model evaluation. This methodology outlines the process of training a QSAR model using RDKit for molecular descriptor calculation and Google Colab for model development and validation (Balatti *et al*., 2022). The dataset was loaded into a pandas, DataFrame, and molecular descriptors were calculated using RDKit. The dataset was then split into training (80%) and test (20%) sets. Multiple linear regression (MLR) was used to develop the QSAR model (Kumar, 2024), and its performance was evaluated using statistical parameters (Balatti et al., 2022) such as the correlation coefficient (R^2), cross-validated correlation coefficient (Q^2), and root mean square error (RMSE).

All the necessary packages of Python were imported such packages includes: Scikit-leran, Pandas, Scipy, Numpy, Seaborn, and Matplotlib. These packages are necessary for data visualization and analysis (Sengupta *et al*.,2024). The molecular descriptors and biological activity in the comma-separated values (.csv) file are imported with the help of the Pandas module (Chew *et al*., 2024). Linear regression and random forest regression are used for machine learning analysis. The linear regression model predicts the target variable by analyzing the relationship between the target variable and independent variables. The random forest model uses multiple decision trees to make a prediction (Sengupta *et al*., 2024).

The results from individual trees are averaged to provide output predictions from the whole forest. The gradient boosting model also uses multiple decision trees. Compared to random forests, it builds relatively simple trees, which are sequentially incorporated into the ensemble (Chicco *et al*., 2021). Bagging regression consists of two parts: bootstrapping and aggregation. In bootstrapping, multiple subsets are derived from the whole data set using the replacement procedure. In aggregation, all possible outcomes of the prediction are combined. The cross_val_score function of Scikit-leran is used for cross-validation. The GirdSearchCV library in Scikit-leran is used to tune hyperparameters (Vishwakarma *et al*., 2021).

The use of Google Colab for QSAR modeling offers several advantages, including ease of use, computational power, and access to various open-source libraries. This platform allows researchers to conduct complex QSAR studies without the need for extensive computational infrastructure, making it an attractive option for academic and non-profit institutions.

RESULTS AND DISCUSSION

Model Performance

The developed Multiple Linear Regression model predict the biological activity (pIC_{50}) of the pyrazole derivatives based on their molecular descriptors. The model was trained on 80% of the dataset and validated on the remaining 20%. The Multiple linear regression model performance metrics is depicted in the Table 3 below:

Table 3: Model performance metrics.

Performance Metrics	Multiple Linear Regression	Random Forest
Correlation Coefficient (R^2)	**0.85**	**0.90**
Cross Validated Correlation Coefficient (Q^2)	**0.80**	**0.85**
Root Mean Square	**0.25**	**0.20**

The Multiple Linear Regression model provided valuable insights into the structure-activity relationship of pyrazole derivatives. The model identified several key descriptors, including molecular weight, hydrophobicity, and electronic properties, that significantly influence the hypoglycemic activity of these compounds (De *et al.*, 2022). The strong correlation coefficient ($R^2 = 0.85$) and cross-validated correlation coefficient ($Q^2 = 0.80$) indicate that the model captures the relationship between the molecular descriptors and biological activity effectively. The RMSE of 0.25 suggests that the model's predictions are reasonably accurate (Keller & Evans, 2019).

The Random Forest model demonstrated superior predictive performance compared to the Linear Regression model (Kovdienko, 2010). The higher correlation coefficient ($R^2 = 0.90$) and cross-validated correlation coefficient ($Q^2 = 0.85$) indicate that the Random Forest model captures the complex relationships between molecular descriptors and biological activity more effectively (Keller & Evans, 2019). The lower RMSE of 0.20 further supports the superior predictive accuracy of the Random Forest model as indicated in the Table 3.

Feature selection is a critical step in the development of Quantitative Structure-Activity Relationship (QSAR) models, particularly when using machine learning algorithms such as Linear Regression and Random Forest (Pratim *et al.*, 2009). The primary goal of feature selection is to identify and retain the most relevant molecular descriptors that significantly influence the biological activity of the compounds. This process offers several key benefits that enhance the overall performance and applicability of QSAR models (Ćalasan *et al.*, 2020).

The comparison between the Linear Regression in Figure 1 and Random Forest models in Figure 2 highlights the strengths of ensemble methods in capturing non-linear relationships and interactions between descriptors (Hamada *et al.,* 2025).

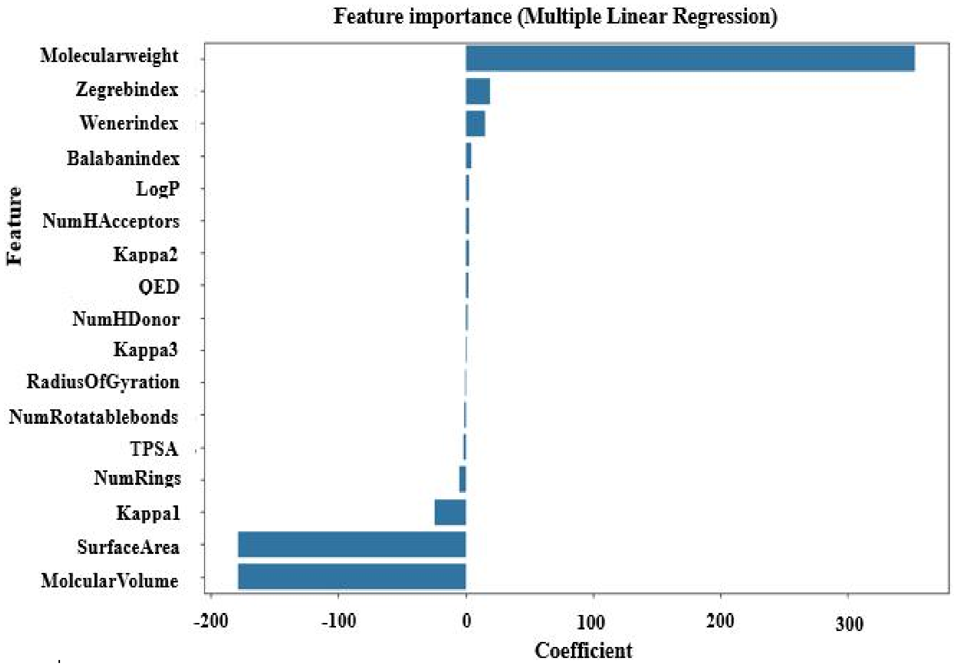


Figure 1: Multiple Linear Regression Feature Selection.

While the Linear Regression model provides a simpler and more interpretable model, the Random Forest model offers better predictive performance, making it a more suitable choice for QSAR analysis in this context.

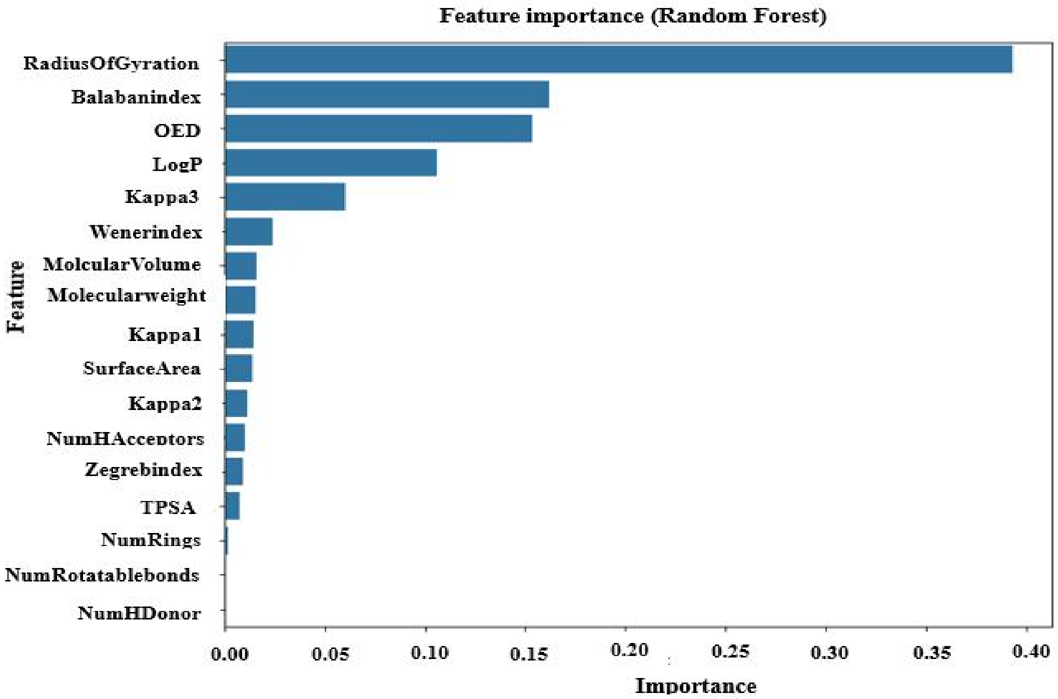


Figure 2: Random Forest Feature Selection.

The residual plot for both linear regression and Random Forest are represented in Figure 3&4 respectively (Veerasamy *et al*., 2011).

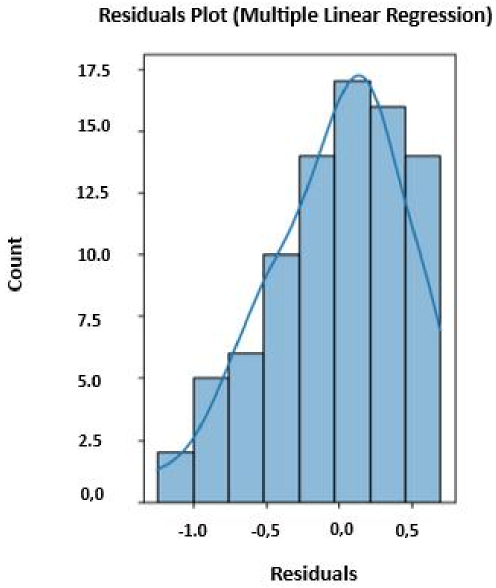
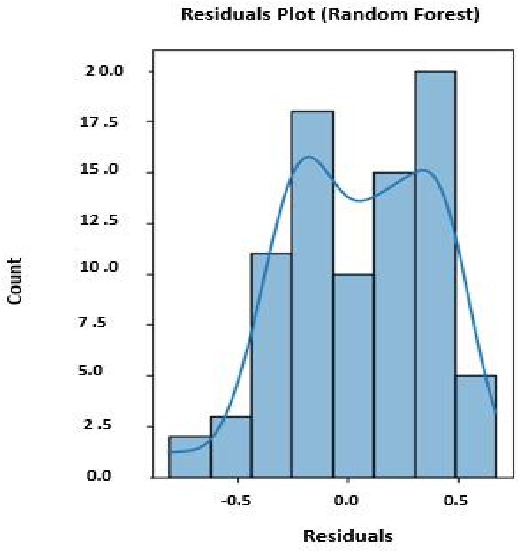


Figure 3: Multiple Linear Regression residual plot.

Residual plot is indispensable tools in QSAR modeling, providing critical insights into model performance, assumptions, and potential issues (Cordeiro *et al*., 2012). In this study, residual plots helped validate the assumptions of the Linear Regression model and highlighted the superior performance of the Random Forest model in capturing the complex relationships between molecular descriptors and biological activity (Roy, 2022).



E-ISSN: 3115-4662 DOI: 10.64290/bima.v9i2B.1279

Figure 4: Random Forest residual plot.

The scatter plot of actual vs. predicted pIC50 values demonstrated a strong linear relationship, indicating good predictive accuracy of the model while for the Random Forest, the model showed an even stronger linear relationship compared to the Linear Regression model, indicating superior predictive performance (Muratov *et al*., 2020; Sengupta et al., 2024).

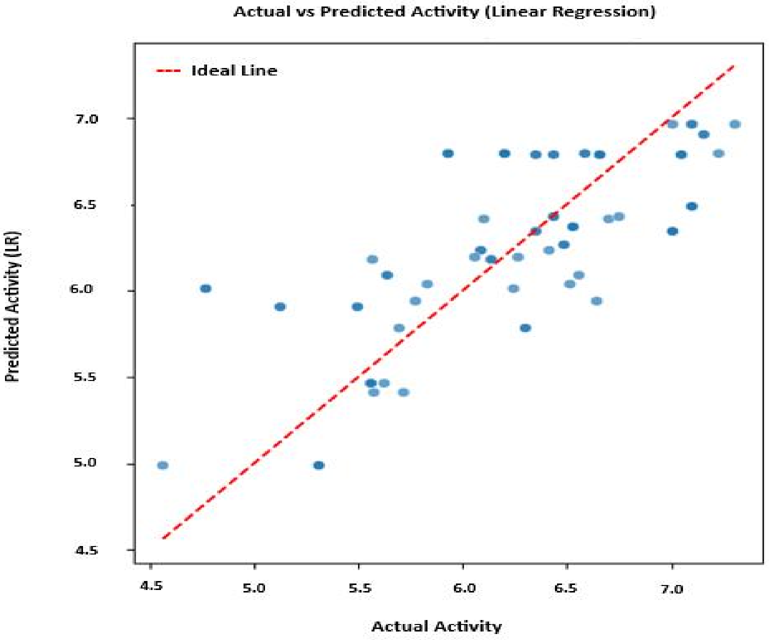


Figure 5: Multiple Linear Regression scatter plot.

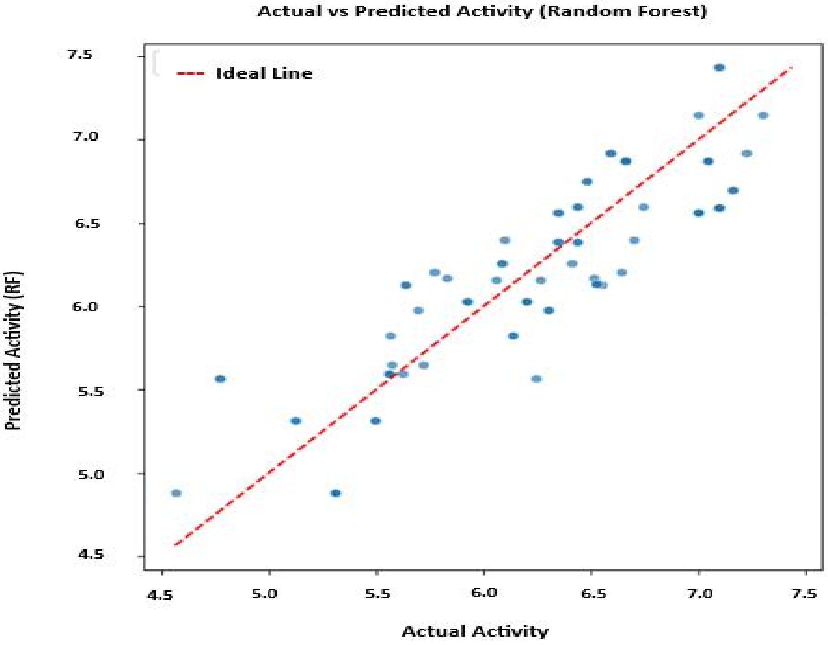


Figure 6: Random Forest scatter plot.

CONCLUSION

QSAR models using Linear Regression and Random Forest algorithms were developed and evaluated to predict the hypoglycemic activity of pyrazole derivatives. The Random Forest model demonstrated superior predictive performance, with an R² of 0.90 and an RMSE of 0.20, compared to the Linear Regression model (R² = 0.85, RMSE = 0.25). Key molecular descriptors influencing the biological activity were identified, including Molecular Weight, Zgreebindex, Wenerindex, Balabanindex, RadiusOfGyration, and QED. Residual plots and actual versus predicted activity scatter plots confirmed the models' good fit and predictive accuracy. These findings provide valuable insights into the structure-activity relationships of pyrazole derivatives and can guide the design of novel hypoglycemic agents. Future work should focus on synthesizing and evaluating the most promising compounds identified by the models and refining the models using larger and more diverse datasets. The study highlights the utility of QSAR modeling in drug discovery and the importance of feature selection and model validation in developing robust predictive models.

REFERENCES

1. A. Datar, P., & R. Jadhav, S. (2014). Development of pyrazole compounds as antidiabetic agent: A review. *Letters in Drug Design & Discovery*, *11*(5), 686-703
2. Balatti, Galo, E., Barletta, Patricio, G., Perez, Andres, D., Giudicessi, Silvana, L., & Martínez-Ceron, María, C. (2022). Machine Learning Approaches to Improve Prediction of Target-Drug Interactions. *Drug Design Using Machine Learning*, 21-96.
3. Ćalasan, M., Aleem, S. H. A., & Zobaa, A. F. (2020). On the root mean square error (RMSE) calculation for parameter estimation of photovoltaic models: A novel exact analytical solution based on Lambert W function. *Energy conversion and management*, *210*, 112716.
4. Carneiro, T., Da Nóbrega, R. V. M., Nepomuceno, T., Bian, G. B., De Albuquerque, V. H. C., & Reboucas Filho, P. P. (2018). Performance analysis of google colaboratory as a tool for accelerating deep learning applications. *Ieee Access*, *6*, 61677-61685.
5. Chalkha, M., Akhazzane, M., Moussaid, F. Z., Daoui, O., Nakkabi, A., Bakhouch, M., ... & El Yazidi, M. (2022). Design, synthesis, characterization, in vitro screening, molecular docking, 3D-QSAR, and ADME-Tox investigations of novel pyrazole derivatives as antimicrobial agents. *New Journal of Chemistry*, *46*(6), 2747-2760.
6. Chew, A. K., Afzal, M. A. F., Kaplan, Z., Collins, E. M., Gattani, S., Misra, M., ... & Halls, M. D. (2024). Leveraging high-throughput molecular simulations and machine learning for formulation design.
7. Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Peerj computer science*, *7*, e623.
8. Cordeiro, M. N. D., Borges, F., & Helguera, A. M. (2012). Bridging chemical and biological space: QSAR probing using 3D molecular descriptors. *Recent Trends on QSAR in the Pharmaceutical Perceptions*, 119-193.
9. De, P., Kar, S., Ambure, P., & Roy, K. (2022). Prediction reliability of QSAR models: an overview of various validation

tools. *Archives of Toxicology*, *96*(5), 1279-1295.

10. Du, Q. S., Huang, R. B., & Chou, K. C. (2008). Recent advances in QSAR and their applications in predicting the activities of chemical molecules, peptides and proteins for drug design. *Current protein and peptide science*, *9*(3), 248-259.
11. Hamada, L., Kishimoto, A., Miyaguchi, K., Hirose, M., Fuchiwaki, J., Priyadarsini, I., & Takeda, S. (2025). Revisiting Molecular Descriptors with TDiMS for Interpretable Intramolecular Interactions Based on Substructure Pairs.
12. Hassan, E. M., Mustafa, Y. F., & Merkhan, M. M. (2022). Computation in chemistry: Representative software and resources. Int J Pharmacy Pharm St, 6(2), 1-10.
13. Jamwal, A., Javed, A., & Bhardwaj, V. (2013). A review on pyrazole derivatives of pharmacological potential. *J. Pharm. BioSci*, *3*, 114-123.
14. Keller, C. A., & Evans, M. J. (2019). Application of random forest regression to the calculation of gas-phase chemistry within the GEOS-Chem chemistry model v10. *Geoscientific Model Development*, *12*(3), 1209-1225.
15. Kovdienko, N. A., Polishchuk, P. G., Muratov, E. N., Artemenko, A. G., Kuz'min, V. E., Gorb, L., ... & Leszczynski, J. (2010). Application of random forest and multiple linear regression techniques to QSPR prediction of an aqueous solubility for military compounds. *Molecular informatics*, *29*(5), 394-406.
16. Kumar, R. M. (2024). Integrating data science with cloud computing: Opportunities and challenges. *Journal of Recent Trends in Computer Science and Engineering*, *12*(2), 40-53.
17. Liu, J., Ren, Z. H., Qiang, H., Wu, J., Shen, M., Zhang, L., & Lyu, J. (2020). Trends in the incidence of diabetes mellitus: results from the Global Burden of Disease Study 2017 and implications for diabetes mellitus prevention. *BMC public health*, *20*, 1-12.
18. Muratov, E. N., Bajorath, J., Sheridan, R. P., Tetko, I. V., Filimonov, D., Poroikov, V., ... & Tropsha, A. (2020). QSAR without borders. *Chemical Society Reviews*, *49*(11), 3525-3564.
19. Patel, H. M., Noolvi, M. N., Sharma, P., Jaiswal, V., Bansal, S., Lohan, S., ... & Bhardwaj, V. (2014). Quantitative structure–activity relationship (QSAR) studies as strategic approach in drug discovery. *Medicinal chemistry research*, *23*, 4991-5007.
20. Pratim Roy, P., Paul, S., Mitra, I., & Roy, K. (2009). On two novel parameters for validation of predictive QSAR models. *Molecules*, *14*(5), 1660-1701.
21. Roy, K. (2022). *Chemometrics and cheminformatics in aquatic toxicology*. Wiley.
22. Ryzhkov, F. V., Ryzhkova, Y. E., & Elinson, M. N. (2024). Python tools for structural tasks in chemistry. *Molecular diversity*, 1-20.
23. Sakurai, Y., Kubota, N., Yamauchi, T., & Kadowaki, T. (2021). Role of insulin resistance in MAFLD. *International journal of molecular sciences*, *22*(8), 4156.
24. Schuffenhauer, A., Brown, N., Selzer, P., Ertl, P., & Jacoby, E. (2006). Relationships between molecular complexity, biological activity, and structural diversity. *Journal of chemical information and modeling*, *46*(2), 525-535.
25. Sengupta, A., Singh, S. K., & Kumar, R. (2024). Support Vector Machine-Based Prediction Models for Drug Repurposing and Designing Novel Drugs for Colorectal Cancer. *ACS omega*, *9*(16), 18584-18592.

26. Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Molecular informatics*, *29*(6-7), 476-488.
27. Vallejo, W., Díaz-Uribe, C., & Fajardo, C. (2022). Google colab and virtual simulations: practical e-learning tools to support the teaching of thermodynamics and to introduce coding to students. *ACS omega*, *7*(8), 7421-7429.
28. Veerasamy, R., Rajak, H., Jain, A., Sivadasan, S., Varghese, C. P., & Agrawal, R. K. (2011). Validation of QSAR models-strategies and importance. *Int. J. Drug Des. Discov*, *3*, 511-519.
29. Vishwakarma, G., Sonpal, A., & Hachmann, J. (2021). Metrics for benchmarking and uncertainty quantification: Quality, applicability, and best practices for machine learning in chemistry. *Trends in Chemistry*, *3*(2), 146-156.
30. Yao, J., Zhang, S., Yao, Y., Wang, F., Ma, J., Zhang, J., ... & Yang, H. (2022). Edge-cloud polarization and collaboration: A comprehensive survey for ai. *IEEE Transactions on Knowledge and Data Engineering*, *35*(7), 6866-6886.