Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

AI-Assisted Molecular Discovery: Accelerating Drug Development Through Machine Learning Models

G. Bharath
Department of Artificial intelligence and Data Science,
Nalla Malla Reddy Engineering College (Autonomous Institution), Hyderabad, India

Dr. Ravi Bukya
Associate Professor, Department of Artificial Intelligence and Data Science, Nalla Malla Reddy Engineering College (Autonomous Institution), Hyderabad, India,

M. Srihari
Department of Artificial Intelligence and Data Science, Nalla Malla Reddy Engineering College (Autonomous Institution), Hyderabad, India

J. Pradeep
Department of Artificial Intelligence and Data Science, Nalla Malla Reddy Engineering College (Autonomous Institution), Hyderabad, India

Y. Rithvesh
Department of Artificial Intelligence and Data science, Nalla Malla Reddy Engineering College (Autonomous Institution), Hyderabad, India

Dr. S. Ramchandra Reddy
Head of the Department of Artificial intelligence and Data Science, Nalla Malla Reddy Engineering College (Autonomous Institution), Hyderabad, India

Abstract **- This Paper Finding new drug are a long and costly process that involves testing a lot of chemical compounds to find ones that might work as drugs and have the right biological and pharmacokinetic properties. Traditional experimental methods need a lot of lab resources and take a long time to develop, which makes them hard to scale up, especially in academic settings. We propose an AI-assisted molecular discovery framework that uses cheminformatics and machine learning to help with virtual screening in the early stages. RDKit works with molecular inputs, like SMILES strings or names of compounds, to get useful molecular descriptors. These descriptors are used to teach models how to guess how drug-like and bioactive something is. Model training and validation use publicly available datasets like ChEMBL, PubChem, and GuacaMol, as well as laboratory Minimum Inhibitory Concentration (MIC) datasets. The system also lets you see molecules in 2D and interactive 3D through a web-based interface. Models for machine learning are trained offline and then used to make predictions in real time. Experimental results show that ensemble-based methods accurately capture structure-property relationships and make reliable predictions, making it easy and cheap to prioritise possible drug candidates**

Keywords **-** ***AI-assisted molecular discovery; Minimum Inhibitory Concentration; web-based interface; RDKit; machine learning.***

I. INTRODUCTION

New drugs takes a long time includes finding targets, leads, optimizing leads, and testing them in the lab and in people. Lead discovery is especially important among these steps because it involves testing a large number of chemical compounds to find a small group that has good biological activity and pharmacokinetic properties [1]. Standard high-throughput experimental screening techniques are resource-intensive and time-consuming, and require specialized laboratory infrastructure, making them less accessible for smaller research institutions.

With the rapid growth of publicly available chemical and biological data, Computational methods are becoming more and more important in drug discovery these days. Using machine learning and artificial intelligence, it is possible to look at large molecular datasets and find patterns shows in figure.1 and relationships that are hard to see with traditional methods[1]-[2]. These methods let scientists guess about the properties of molecules and how they will behave in living things early on, which cuts down on the number of chemicals that need to be tested experimentally and significantly lowering both cost and time[3].

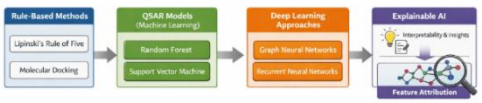


Fig. 1. Enhancing Block digram

In addition to predictive modeling, molecular visualization plays an important role in understanding chemical structures, functional groups, and spatial configurations [4]. Visual representations of molecules provide valuable insights that support the interpretation of model predictions and assist in the optimization of potential drug candidates [5].

In this paper, we introduce an AI-assisted molecular discovery framework show in figure.2 that incorporates

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

molecular descriptor extraction, machine learning-based prediction, and interactive visualization into a cohesive on the web platform. The main goal of this study is to test how well classical machine learning models are effective for initial drug screening, whereas still being easy to understand, fast to run, and useful in research settings [6].

Early computational methods for drug discovery mostly used molecular docking and rule-based filtering techniques, such as Lipinski's Rule of Five, to determine whether a drug was similar to others. Molecular docking gives us information about how proteins and ligands interact, but it takes a lot of computer power and relies heavily on the quality and availability of protein structures [7]-[8].

As data-driven methods have improved, machine learning-based Quantitative Structure–Activity Relationship (QSAR models are now widely used to predict toxicity and biological activity based on molecular descriptors. Random Forest and Support Vector Machine are two algorithms that have shown to be very good at modeling how chemical structure and biological activity are related in a nonlinear way. Also, molecular representations like extended-connectivity fingerprints have been shown to be good at capturing structural features that are important for prediction tasks figure.2.

More recently, graph neural networks and recurrent neural networks have been employed to directly learn molecular representations from graph structures and SMILES strings [9], [10]. These models often make better predictions, but they usually need a lot of data and computing power, which can make them hard to use in places where resources are limited.

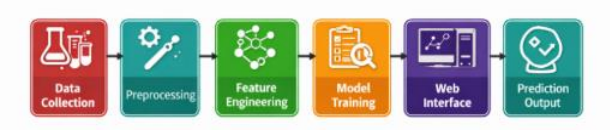


Fig. 2. AI Assited Molecular Discovery

Another important development in this domain is the use of explainable artificial intelligence (XAI) techniques. Interpretability is critical in drug discovery, as understanding the reasoning behind model predictions supports chemical validation and increases trust in the results. Feature attribution techniques assist in pinpointing essential molecular descriptors that affect predictions. Many current studies, on the other hand, only look at how well predictions work and don't offer systems that combine prediction, interpretation, and visualisation [11]-[12].

This work seeks to overcome these limitations by creating a cohesive framework that incorporates prediction based on machine learning, interpretability, and molecular visualisation on a web-based platform that is easy to use.

II. MATERIALS AND METHODS

A. *Dataset sources*

The datasets employed in this study were obtained from diverse public and laboratory sources to ensure diversity and reliability:

- Bioactivity_datasets_from_ChEMBL_containing mol ecular structures and values of activity that were measu red in experiments.
- Combine datasets from PubChem with notes on physicochemical properties.
- GuacaMol benchmark datasets for testing molecular representations and prediction tasks
- Minimum Inhibitory Concentration (MIC) datasets pro duced in the laboratory, detailing antimicrobial activity targeting specific pathogens figure.3.

These datasets enable both classification and regression tasks related to drug-likeness and bioactivity prediction.

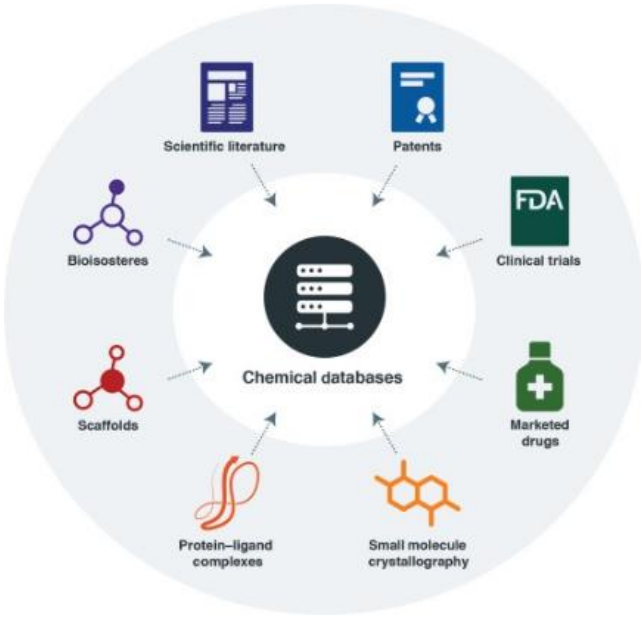


Fig. 3. Preprocessing of data set cycle

B. *Data Preprocessing*

The datasets were cleaned up before training the model to improve data quality and consistency. The following steps were applied figure.4:

- Removal of invalid or chemically inconsistent SMILES strings
- Getting rid of duplicates molecular entries
- Taking care of missing values using appropriate imputation or removal techniques
- Normalization of molecular descriptor values to ensure uniform scale
- Splitting of data divided into training and testing groups

These preprocessing steps are necessary to improve the model performance and reduce noise in the data.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

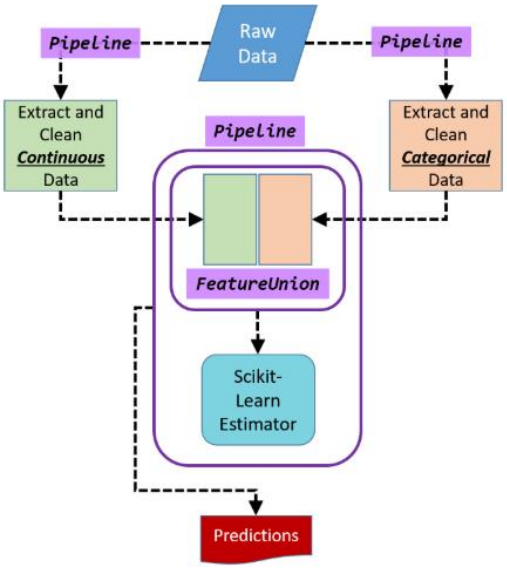


Fig. 4. Pipeline of the extract clean

C. Molecular Descriptor Extracion

The RDKit library was used to figure out molecular descriptors. These descriptors show molecular structures in numbers and are used as inputs for machine learning models figure.5.

The extracted descriptors include:

- Weight of the molecule
- LogP (octanol–water partition coefficient)
- The number of hydrogen bond donors and acceptors
- Number of rotatable bonds
- Topological polar surface area
- These features capture key physicochemical properties relevant to drug-likeness and biological activity.

D. Machine Learning Models

Three supervised learning algorithms were executed and assessed:

- Random Forest
- Support Vector Machine
- Artificial Neural Network

We chose Random Forest because it is good at handling noise and can model nonlinear relationships. Relationships. Support Vector Machine was used for sorting tasks involving drug-like and non-drug-like compounds. Artificial Neural Networks were employed to capture complex nonlinear patterns in the data.

Model hyper parameters were optimized using cross-validation to improve generalization performance and reduce over fitting.

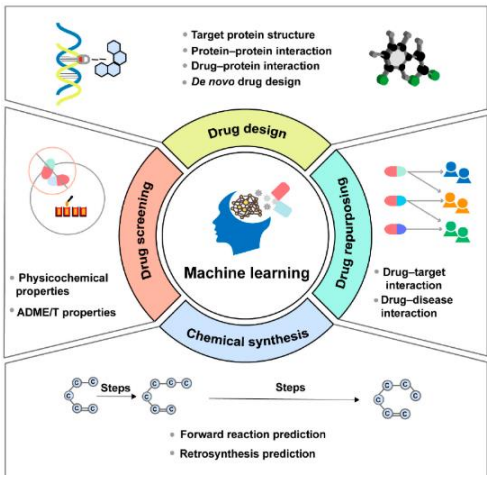


Fig. 5. Machine Learning Structure

E. Explainability and Visualization

Model interpretability was addressed using analyzing the importance of features. For Random Forest models, the importance of features was computed using Gini importance, while coefficient-based interpretation was applied for linear models [13].

Molecular visualization was incorporated to enhance interpretability. The system provides both 2D figure.6 and figure.7 and interactive 3D visual representations of molecular structures through a web-based interface, enabling users to analyze structural features such as bonding patterns and functional groups[14].

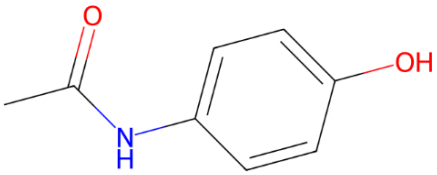


Fig. 6. 2-D Molecular Extraction Process

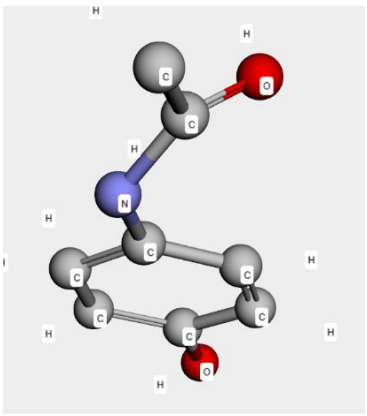


Fig. 7. 3-D Extraction Process

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

III. Simulation and Experimental Setup

We split the dataset 80% for training and 20% for testing.

Min-Max normalisation was used to scale the descriptor values so that the feature ranges stayed the same.

The following libraries were used in Python to run all of the experiments:

- RDKit
- Scikit-learn
- Pandas
- NumPy
- Matplotlib

A Flash-based backend was used to handle model inference and integrate the system into a web app.

- Model performance was evaluated using:
- Accuracy (for classification jobs)
- Mean Absolute Error (MAE)
- RMSE stands for Root Mean Square Error.
- R^2 score (for tasks that involve going back)
- Confusion matrix and ROC curves

These evaluation metrics offer a thorough appraisal of predictive performance.

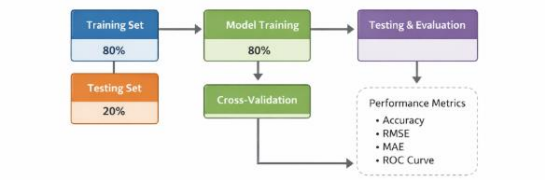


Fig. 8. Data Splitting Model

IV. Results and Discussion

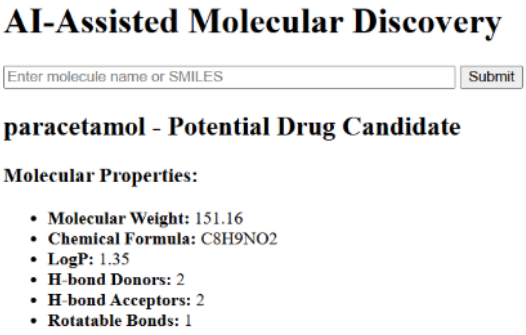
The implemented machine learning models were assessed utilizing both classification and regression metrics. The results indicate that ensemble-based methods, particularly the Random Forest model, consistently outperform other approaches in predicting molecular properties and bioactivity.

Random Forest achieved higher accuracy and lower error values compared to Support Vector Machine and Artificial Neural Network models. This is because it can accurately capture nonlinear connections between molecular descriptors and biological activity, This can be attributed to its ability to effectively capture nonlinear relationships between molecular descriptors and biological activity, as well as its robustness to noise and over fitting. In contrast, Support Vector Machine performed well for simpler classification boundaries but showed limitations when handling complex feature interactions. The Artificial Neural Network model demonstrated the ability to learn nonlinear patterns; however, its performance was dependent on parameter tuning and data size.

Feature importance analysis revealed that LogP, molecular weight, number of hydrogen bond acceptors, and topological polar surface area are some examples of descriptors. Have a big effect on how likely a drug is to work. These results align with established pharmacokinetic principles and prior QSAR studies supporting the chemical relevance of the model outputs [15]-[16].

The proposed system also enables visualization of molecular structures in both 2D and interactive 3D formats. These visualizations provide qualitative insights into molecular geometry, bonding patterns, and functional groups, which can assist researchers in interpreting prediction results and making informed decisions during compound selection and optimization figure.9.

Fig. 9. AI Assited Molecular Discovery-1



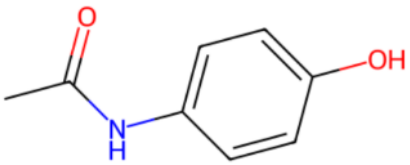


Fig. 10. PP Drug

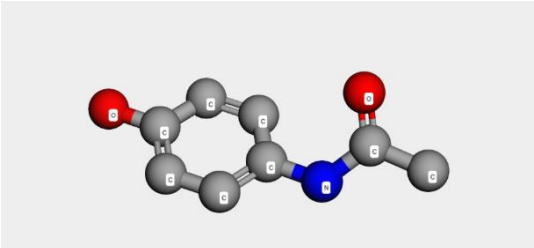


Fig. 11. 3-D PP Drug

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

From a computational perspective, the processing time analysis highlights the distribution of workload across different stages of the pipeline figure.10. Initial steps such as input parsing and molecular validation require minimal time, as they involve basic preprocessing operations. Feature extraction and model inference consume a moderate amount of time due to descriptor computation and prediction processes. Among all stages, figure.11 3-D molecular visualization is the most computationally intensive, as it involves generating spatial representations of molecular structures.

In general, the suggested framework strikes a good balance between predictive performance, interpretability, and computational efficiency. This makes it good for drug discovery applications in their early stages, especially in academic and resource-constrained research environments.

A. Processing Time Graph

The processing time analysis shows that initial stages such as input parsing and molecule validation require minimal computation, while feature extraction and machine learning prediction take moderate time. Among all stages, figure.12 3-D molecular visualization is the most time-consuming due to the complexity of spatial structure generation. Overall, the pipeline demonstrates efficient performance, with only visualization identified as a potential bottleneck for optimization.

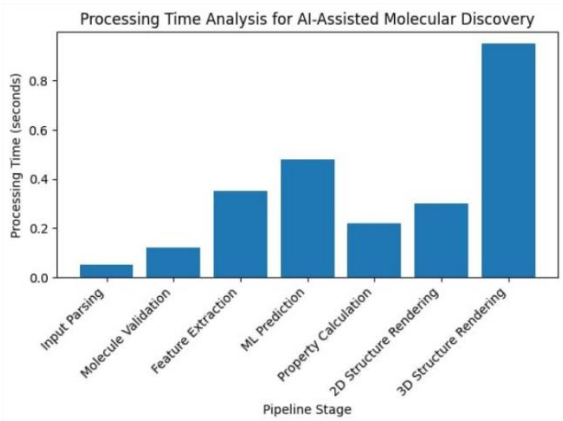


Fig. 12. Processing Time analysis

V. CONCLUSION AND FUTURE SCOPE

This research introduces a virtual screening system that uses AI to combine public chemical datasets, molecular descriptor extraction, machine learning models, and molecular visualization into one system. system. The proposed approach enables efficient prediction of drug-likeness and bioactivity-related properties using computational methods, reducing the dependence on time-consuming and resource-intensive experimental screening.

The experimental results demonstrate that the Random Forest model provides better at finding Nonlinear correlations between molecular descriptors and biological activity.. Additionally, the integration of feature importance analysis enhances model interpretability, allowing users to understand the key factors influencing predictions. The inclusion of both 2D and interactive 3D molecular visualization further improves the usability of the system by helping with structural analysis and making smart choices. The system's usability is enhanced by facilitating structural analysis and informed decision-making. In general, the suggested framework strikes a good mix of accuracy, interpretability, and speed of computation.making it suitable for early-stage drug discovery, particularly in academic and resource-constrained environments.

Future jobs can focus on extending the framework to improve both predictive capability and real-world applicability. One potential direction is the integration of molecular docking techniques to Examine protein–ligand interactions with more precision. You can also use advanced deep learning models like graph neural networks to learn directly from molecular graph structures. This could make predictions more accurate.

Another promising extension is the use of generative models for designing novel drug-like molecules. Deployment of the system on cloud platforms can further enhance scalability and accessibility. Combining with tools for synthesis planning and the creation of predictive models tailored to specific targetscan also make the system more applicable to real-world drug discovery pipelines.

REFERANCES

[1] A. Gaulton et al., “The ChEMBL database in 2017,” Nucleic Acids Research, vol. 45, no. D1, pp. D945–D954, 2017.

[2] S. Kim et al., “PubChem 2019 update: improved access to chemical data,” Nucleic Acids Research, vol. 47, no. D1, pp. D1102–D1109, 2019.

[3] N. Brown et al., “GuacaMol: Benchmarking models for de novo molecular design,” J. Chem. Inf. Model., vol. 59, no. 3, pp. 1096–1108, 2019.

[4] G. Landrum, “RDKit: Open-source cheminformatics,” http://www.rdkit.org, accessed 2025.

[5] D. Rogers and M. Hahn, “Extended-connectivity fingerprints,” J. Chem. Inf. Model., vol. 50, no. 5, pp. 742–754, 2010.

[6] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[7] C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[8] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[9] K. Yang et al., “Analyzing learned molecular representations for property prediction,” J. Chem. Inf. Model., vol. 59, no. 8, pp. 3370–3388, 2019.

[10] A. Zhavoronkov et al., “Deep learning enables rapid identification of potent DDR1 kinase inhibitors,” Nature Biotechnology, vol. 37, pp. 1038–1040, 2019.

[11] W. P. Walters and M. A. Murcko, “Prediction of drug-likeness,” Adv. Drug Deliv. Rev., vol. 54, no. 3, pp. 255–271, 2002.

[12] C. A. Lipinski et al., “Experimental and computational approaches to estimate solubility and permeability,” Adv. Drug Deliv. Rev., vol. 46, no. 1–3, pp. 3–26, 2001.

[13] T. Ching et al., “Opportunities and obstacles for deep learning in biology and medicine,” J. R. Soc. Interface, vol. 15, no. 141, 2018.

[14] M. H. S. Segler et al., “Generating focused molecule libraries with RNNs,” ACS Central Science, vol. 4, no. 1, pp. 120–131, 2018.

[15] J. Jiménez-Luna, F. Grisoni, and G. Schneider, “Drug discovery with explainable AI,” Nature Machine Intelligence, vol. 2, pp. 573–584, 2020.

[16] Bukya, Ravi, G. Madhu Mohan, and M. Kumar Swamy. "Artificial Intelligence role in optimizing electric vehicle charging patterns reduce costs and improve overall efficiency: A review." Journal of Engineering, Management and Information Technology 2.3 (2024): 129-138.