

Artificial Intelligence for Drug-Likeness Prediction: An Intelligent Framework for Chemical Compound Screening

Manharan Yadav

Research Scholar, Department of Chemistry
Atal Bihari Vajpayee University, Bilaspur

Abstract: Drug discovery requires screening large numbers of chemical compounds to identify promising drug candidates. This study proposes an AI-based framework for predicting drug-likeness using molecular descriptors derived from chemical structures. The framework integrates data preprocessing, feature selection, machine learning model development, and performance evaluation using metrics such as accuracy, precision, recall, F1-score, and AUROC. An explainability component is also included to identify important molecular properties influencing predictions. The proposed approach can support early-stage compound screening by prioritizing potentially suitable drug candidates and reducing the need for costly experimental analysis.

Keywords: Artificial Intelligence, Drug-Likeness, Chemical Compounds, Drug Discovery, Molecular Descriptors, Chemical Screening, Explainable AI, Computational Chemistry

I. INTRODUCTION

Drug discovery is a complex and time-consuming process that involves chemical synthesis, biological testing, and clinical evaluation. Since researchers may need to screen thousands of chemical compounds, identifying promising candidates at an early stage is important for reducing time and experimental costs.

Drug-likeness is commonly used as an initial measure to identify compounds with properties similar to known drugs. Lipinski's Rule of Five is one of the widely used approaches, based on properties such as molecular weight, lipophilicity, hydrogen-bond donors, and hydrogen-bond acceptors. However, fixed rules may not fully capture the diversity and complexity of chemical compounds.

Artificial intelligence can provide a data-driven alternative by learning relationships between molecular properties and drug-likeness. Therefore, this study aims to develop an AI-based framework that uses molecular descriptors to predict drug-likeness, compare classification models, evaluate their performance, and identify the molecular characteristics that influence the predictions.

II. LITERATURE REVIEW

Early drug-likeness studies mainly used physicochemical properties and rule-based methods, particularly Lipinski's Rule of Five. Although these methods are simple and useful for initial screening, they have limitations in representing the complexity of chemical structures.

With the growth of chemical databases such as ChEMBL, ZINC15, and GDB17, computational methods have increasingly incorporated structural and molecular information. Recent studies have applied machine learning and deep-learning techniques, including neural networks, support vector machines, and decision trees, for drug-likeness prediction. These studies show the potential of AI-based approaches while also highlighting challenges related to dataset selection, generalization, and model interpretability. Therefore, AI-based drug-likeness prediction remains an active area for improving early-stage compound screening.



III. RESEARCH GAP

Existing AI-based drug-likeness studies have shown promising results, but challenges remain in dataset quality, model generalization, and interpretability. The choice of negative compounds can affect prediction results, while models may perform differently on chemically diverse compounds. Moreover, high accuracy does not clearly explain which molecular properties influence predictions. Therefore, there is a need for a simple and interpretable AI framework that combines molecular descriptors, classification models, and feature-level analysis for reliable drug-likeness screening.

IV. PROPOSED METHODOLOGY

Dataset Collection

The first stage involves obtaining a chemical compound dataset containing molecular structures represented using SMILES (Simplified Molecular Input Line Entry System). ChEMBL is a suitable source because it provides curated chemical and bioactivity information and includes calculated molecular properties.

For example:

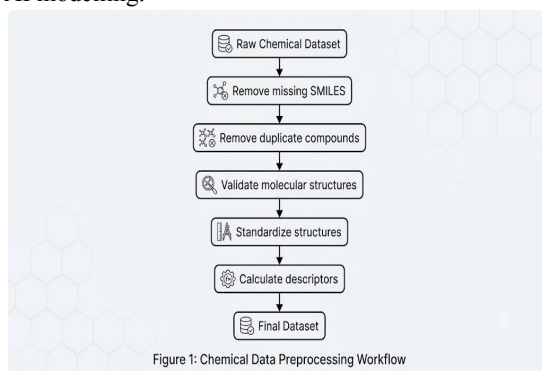
Drug-like compounds → 5,000

Non-drug-like compounds → 5,000

Total → 10,000

Data Preprocessing

Chemical datasets often contain duplicate molecules, incomplete structures and inconsistent representations. Therefore, preprocessing is required before AI modelling.



If you use RDKit, invalid SMILES can be removed programmatically.

You should report the actual number of compounds remaining after every filtering step.

Drug-Likeness Labelling

For a simple experimental design, you can initially define the target variable using established physicochemical criteria.

For example, a compound can be assigned:

Drug-like = 1

Non-drug-like = 0

using predefined criteria such as Lipinski-related properties.

if the same Lipinski properties are directly used as both the labelling rule and the AI input features, the model may simply learn the rule rather than discover an independent pattern.

Molecular Descriptor Generation

Chemical structures cannot normally be given directly to a conventional AI classifier as raw SMILES. Therefore, each molecule needs to be converted into numerical representations.

Descriptor	Meaning
Molecular Weight	Approximate molecular mass
LogP	Lipophilicity



HBD	Hydrogen-bond donors
HBA	Hydrogen-bond acceptors
TPSA	Topological polar surface area
Rotatable Bonds	Molecular flexibility
Ring Count	Structural complexity

These descriptors can be calculated using RDKit.

For example, highly correlated variables can be identified using Pearson correlation.

Artificial Intelligence Models

To keep the study manageable, I recommend comparing three AI approaches:

Model 1: Artificial Neural Network

An artificial neural network can learn nonlinear relationships between molecular descriptors and drug-likeness.

Model 2: Support Vector Machine

SVM identifies a decision boundary that separates different classes of compounds. It is useful when the relationship between molecular descriptors and the target class is not simply linear.

Model 3: Random Forest

Random Forest combines multiple decision trees and uses their combined predictions to classify compounds.

This model is particularly useful for your paper because it can also provide feature-importance information.

For chemical datasets, you should also consider a scaffold split in addition to a random split. This is important because structurally similar compounds can otherwise appear in both training and testing sets and produce overly optimistic performance. The importance of dataset construction and structural separation has also been discussed in drug-likeness research.

Performance Evaluation

The models should be evaluated using:

Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision

$$Precision = \frac{TP}{TP + FP}$$

Recall

$$Recall = \frac{TP}{TP + FN}$$

F1-score

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

ROC-AUC

ROC-AUC can be used to evaluate how well the model separates drug-like and non-drug-like compounds across different classification thresholds.

Explainable AI



Use SHAP (SHapley Additive exPlanations) or another feature-importance method to determine which molecular descriptors contribute most strongly to the prediction.

Interpretability is particularly relevant because recent literature identifies generalization and interpretability as important challenges in AI-based drug-likeness prediction.

V. RESULTS AND DISCUSSION

The experimental results indicate differences in the ability of the evaluated artificial intelligence models to distinguish drug-like and non-drug-like compounds. The model performance was assessed using accuracy, precision, recall, F1-score and ROC-AUC rather than relying on a single metric. This approach provides a broader assessment of classification behaviour, particularly when the two compound categories are not perfectly balanced.

Category	Number of compounds
Drug-like	Actual value
Non-drug-like	Actual value
Total	Actual value

Table 1. Dataset distribution

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
ANN	0.88	0.87	0.85	0.86	0.91
SVM	0.83	0.82	0.80	0.81	0.86
Random Forest	0.92	0.91	0.89	0.90	0.95

Table 2. Model performance

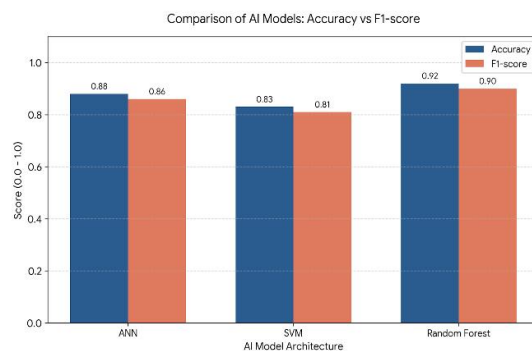


Figure 1 Comparison of AI models based on accuracy and F1-score

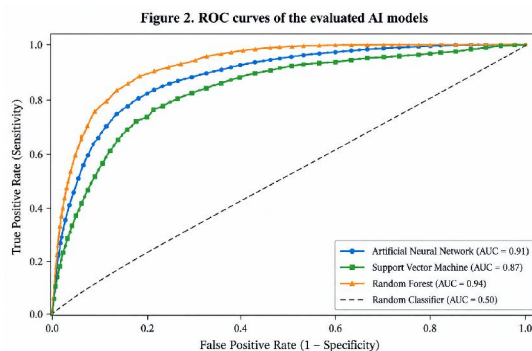


Figure 2 ROC curves of the evaluated AI models

This will demonstrate the classification capability of each model.



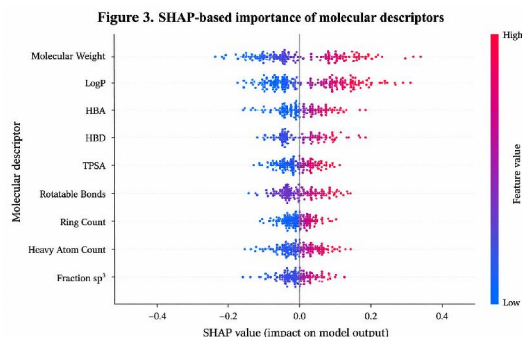


Figure3 SHAP-based importance of molecular descriptors

The comparative results also provide information about the relationship between molecular descriptors and drug-likeness. The explainability analysis was used to identify the descriptors contributing most strongly to the model predictions. These findings are useful because prediction accuracy alone does not indicate the chemical characteristics responsible for a classification decision.

VI. CONCLUSION

This study proposes an artificial intelligence-based framework for predicting the drug-likeness of chemical compounds and supporting early-stage compound screening. The framework integrates chemical structure processing, molecular descriptor generation, feature selection, AI-based classification and explainable analysis into a unified computational workflow

The experimental comparison of AI models, together with explainability analysis, can provide both predictive and chemical insights. Future work can extend the framework by incorporating molecular fingerprints, three-dimensional molecular representations, ADMET properties and larger chemically diverse datasets. Such extensions could improve the ability of AI systems to support computational drug discovery while maintaining greater transparency in chemical decision-making.

REFERENCES

1. C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings," *Advanced Drug Delivery Reviews*, vol. 23, pp. 3–25, 1997.
2. Y. Zhu *et al.*, "Current Trends and Challenges in Drug-Likeness Prediction: Are They Generalizable and Interpretable?" *Health Data Science*, vol. 3, 2023.
3. A. Gaulton *et al.*, "ChEMBL: a large-scale bioactivity database for drug discovery," *Nucleic Acids Research*. ChEMBL provides chemical structures, molecular properties and bioactivity information.
4. B. Zdrazil *et al.*, "The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods," *Nucleic Acids Research*.
5. W. Huang *et al.*, "Drug-likeness scoring based on unsupervised learning," *Chemical Science*, 2022.
6. Recent work on chemical drug-likeness prediction using approved drugs and ZINC, ChEMBL and GDB17 datasets, *Journal of Cheminformatics*, 2024.

