AI-Powered Drug Discovery: Machine Learning Algorithms for Novel Compound Identification

Final report proposal

Hanmo Shi, Yingying He, Yuxiang Tian

November 9, 2024

Abstract:
With the help of artificial intelligence,we will use potential algorithms,SVM, GBM, RF, Logistic Algorithm, and DNN to help people identify and discover new drugs

Contents

1 Task

In this project, the problem we aim to address is AI-driven drug discovery: using machine learning to identify novel compounds. In the field of novel compound research, we face complex challenges. Traditional research methods have shortcomings such as complex reaction pathway design, low experimental efficiency, limited innovation, and resource waste. Therefore, if we can use machine learning to drive drug discovery, it would offer the following advantages:

1. Rapid screening of compounds, optimizing the drug development process, including drug target discovery, drug screening, drug optimization, etc., reducing development costs while improving success rates.
2. Accurate prediction of compound properties to improve drug safety.
3. Mining potential patterns and relationships within data to discover novel compounds and drug targets that traditional methods are unable to identify. This helps break through existing knowledge frameworks and mental models, promoting innovative development in the field of drug research and development.

2 Data

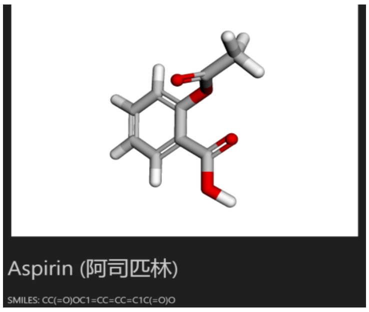
We will use data from PubChem, a high-throughput small molecule database. PubChem is a comprehensive chemical information database that includes multiple resources, such as Compounds, Substances, and BioAssays.

Specifically, our research will focus on the Compounds database, which includes 4,596 target-based drug data, 1,307 drug classification data, and 5,094 drug group data. It provides a comprehensive view of all known information for a single chemical structure, including chemical structure, identifiers, chemical and physical properties, bioactivity, patent information, etc.

The chemical structure of compounds will be used to construct molecular features for drugs, which are key to machine learning models identifying drugs. Bioactivity data will serve as labels or target variables to train the model to predict the bioactivity of new drugs. Chemical and physical properties, patent information, etc., will serve as auxiliary features to help improve the model’s prediction performance.

However, potential obstacles include the fact that although PubChem provides rich chemical information, some drug data may be incomplete or have missing values, which could affect the training effectiveness of the model. Additionally, as the data in PubChem comes from different research institutions and experimental conditions, there may be data bias or inconsistencies, which will require careful screening and correction during data preprocessing.

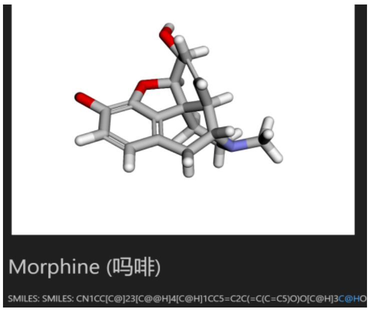
Below is a sample of our dataset, including 3D visualizations of chemical structures generated using Jupyter.



Aspirin (阿司匹林)

SMILES: CC(=O)OC1=CC=CC=C1C(=O)O

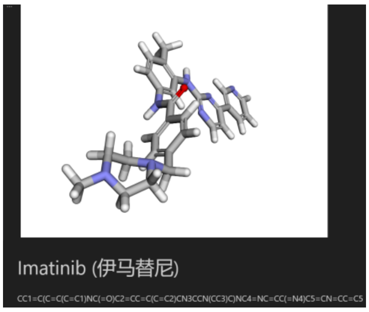
(a) Aspirin



Morphine (吗啡)

SMILES: SMILES: CN1CC[C@]23[C@@H]4[C@H]1CC5=C2C(=C(C=C5)O)O[C@H]3C@HO

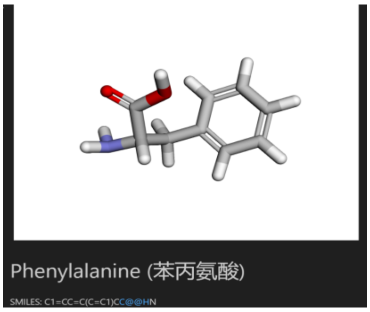
(b) Mosrphine



Imatinib (伊马替尼)

CC1=C(C=C(C=C1)NC(=O)C2=CC=C(C=C2)CN3CCN(CC3)C)NC4=NC=CC(=N4)C5=CN=CC=C5

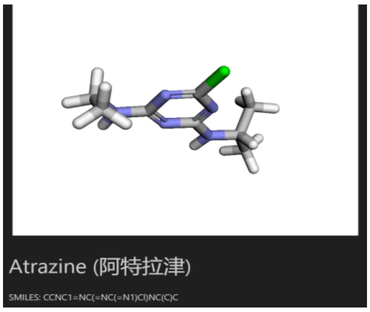
(c) Imatinib



Phenylalanine (苯丙氨酸)

SMILES: C1=CC=C(C=C1)CC@@HN

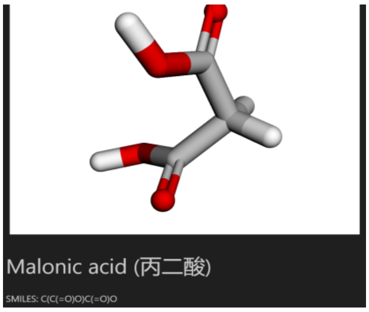
(d) Phenylalanine



Atrazine (阿特拉津)

SMILES: CCNC1=NC(=NC(=N1)Cl)NC(C)C

(e) Atrazine



Malonic acid (丙二酸)

SMILES: C(C(=O)O)C(=O)O

(f) Malonic acid

3 Methods

We will experiment with five algorithms: SVM, GBM, RF, Logistic Regression, and DNN.

Since Logistic Regression is simple, fast, and interpretable, and performs well on small datasets, it will serve as an initial screening model to quickly identify key features related to the target variable. Random Forest (RF) is an ensemble learning algorithm that constructs and combines multiple decision trees to improve prediction accuracy and stability. It is suitable for handling high-dimensional data and can automatically perform feature selection, so it will be used to further screen and validate key features.

Support Vector Machines (SVM) perform well with nonlinear problems, high-dimensional data, and sparse data, and they exhibit strong robustness and generalization ability. Therefore, after initial screening and feature selection, SVM will be used for precise prediction of compound structure and activity.

Gradient Boosting Machines (GBM) improve model performance by combining multiple weak prediction models. It is suitable for handling complex prediction problems. In the precision prediction stage, we will use GBM to further optimize and enhance SVM predictions.

For complex compound structure and activity prediction problems, deep neural networks (such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN)) may perform better. Deep neural networks can automatically learn deeper feature representations of data and improve prediction accuracy through multiple layers of nonlinear transformations. In this project, deep neural networks can be combined with other machine learning algorithms (such as SVM, GBM) to form a hybrid model. For example, deep neural networks can extract deep features of compounds, and then SVM or GBM can be used for further analysis and prediction.

---

4 evaluate

We will evaluate our results using methods such as Balanced F-Score, ANOVA analysis, and t-tests.

The F1 score is calculated as:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{1}$$

Where Precision is:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

And Recall is:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

In this context, Precision refers to the proportion of true positives (TP) among the samples predicted as positive by the model, while Recall refers to the proportion of true positives among all actual positive samples.

We expect the results to include 3D visualizations of novel compounds and their chemical data (including but not limited to molecular formula, molecular weight, melting point, boiling point, solubility, stability, functional group types, etc.).

We hypothesize that the new compound structures and their chemical data predicted by machine learning models will perform better in terms of accuracy and reliability compared to baseline methods (such as traditional chemical synthesis methods and rule-based generation methods). Specifically, we expect machine learning models to predict more compounds with novel structures and high activity potential, while providing more detailed and accurate chemical information. Experimental validation will involve synthesizing and testing the newly predicted compounds to verify their structure and properties. We will compare the machine learning model' s predictions with those of baseline methods and analyze the differences to assess the model's effectiveness.

5 Repository

if you want to konw the progress and results in our project,you can click here to konw the latest information

Besides,please do not forget to click a star for our Repository.Ciallo～(∠·ω<)⌒☆

References

[1] Anuraj Nayarisseri, Ravina Khandelwal, Poonam Tanwar, Maddala Madhavi, Diksha Sharma, Garima Thakur, Alejandro Speck-Planche, and Sanjeev Kumar Singh, *Artificial Intelligence, Big Data and Machine Learning Approaches in Precision Medicine & Drug Discovery*, Current Drug Targets, 2021.

[2] Kit-Kay Mak, Yi-Hang Wong, and Mallikarjuna Rao Pichika, *Artificial Intelligence in Drug Discovery and Development*, Springer, 2024.