FDA Drug Prediction

Shih-Hung Wei
0717034

1. Introduction

I'm a student double majored in Biological Technology, and I'm doing research in a the lab about computational biology, therefore I introduce some knowledge I learned into this final project and try to apply what I have learned in this semester. And I'll briefly introduce the project and go through some materials I apply:

- Motivation

  Drug development is important but time-costing(12∼15 years, 800M USD in average), therefore building a model that can predict drug that is possibly be approved is helpful.

- Kinase

  Kinase is a group of protein, highly related to many diseases such as cancer, cardiovascular disease, autoimmune disease, inflammatory disease, neurological disease, etc. Therefore, kinase inhibitor drugs have always been a popular direction in the field of drug design and development.

- Kinase inhibitor

  Compounds that can inhibit functions of specific kinase. If a compound has been approved by FDA, it's a drug.

- Using SVM/XGBoost/Random Forest model

  End-to-end model doesn't perform well therefore I choose to build a feature-based model, using functional groups(moieties) as features.

2. Related Work

2.1. Features

- Checkmol
- PubChem
- Rings in drugs
- ECFP
- MACCS

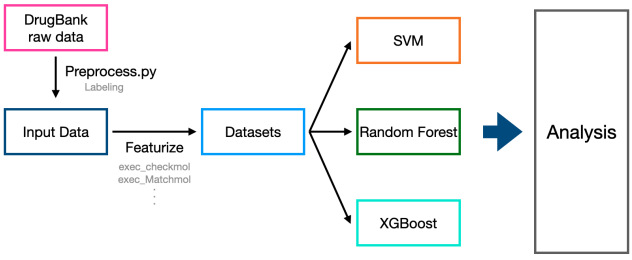
2.2. Baseline

MolecularGNN

Some existed researches use end-to-end methods to build the model. Take molecular GNN as example(Masashi Tsubaki,Bioinformatics, 2018), it directly uses SMILES code(a string format that can represent 3D structure of compounds) as input, and does not extract features. However, these kind of model performs badly, and the output is not interpretable.

3. Methodology

Workflow

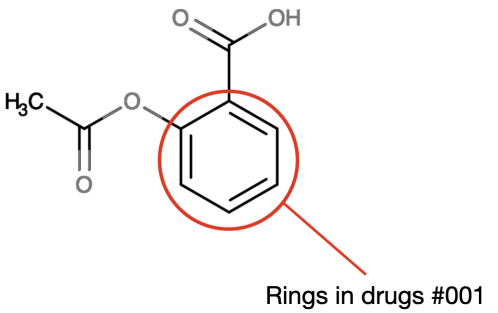


3.1. Datasets

I use the open database: Drugbank as the raw data, and take the approved drugs as positive data, all the other groups(experimental, investigational etc.) of compounds as negative.

3.2. Features

We use the predefined small sub-structures and functional groups in the compounds(we call it moiety) as features. There are many different system of defined moieties, and I choose Checkmol, Pubchem, Rings in drugs, ECFP and MACCs. The image is an example; the sub-structure in the red circle is an moiety(rings in drugs number 001).



Rings in drugs #001

3.3. Models

I choose three classical classifier: SVM, Random Forest and XGBoost to build my model. The input data will be a table of 1744 features x compounds, and output will be a True/False value.

3.4. Evaluation

I use AUC score to evaluate the performance of the model.

4. Result and Analysis

Datapoints: 11141, 7:3 train-test split

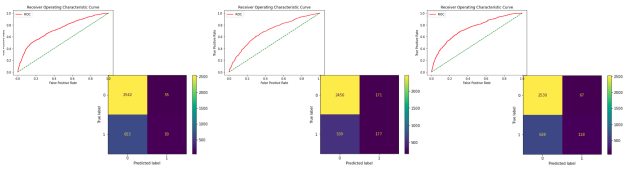
4.1. Baseline

In end-to-end model molecularGNN, AUC score is only 0.68 in training, even only 0.5 in testing.

4.2. Performance

The overall performance of feature based models are around 0.78, which is acceptable though it can absolutely be better.

- SVM
  Accuracy: 0.788
  AUC: 0.70
- Random Forest
  Accuracy: 0.779
  AUC: 0.70
- XGBoost
  Accuracy: 0.782
  AUC: 0.74



Model	SVM	Random Forest	XGBoost
Accuracy	0.788	0.779	0.782
AUC	0.7	0.7	0.74

4.3. Analysis

Although the performance is not good enough, another value of feature based models is that we can interpret the output. I tried to list the feature importance the model learned in Random Forest model, and found that it is interesting and maybe can be developed more deeper in the future. Maybe there actually are some common features in all the approved drugs or compounds that cannot pass the approval.

fea id	pos(%)	neg(%)	fea imp	fea imp rank	
898	60.6%	61.5%	0.02178	1	Drug Moiety #26: butane
903	40.3%	27.1%	0.01671	2	Drug Moiety #32: isobutane
202	62.0%	63.8%	0.01633	3	Checkmol fingerprint #202: heterocyclic compound
55	25.4%	10.3%	0.01617	4	Checkmol fingerprint #055: tert. amine
199	16.8%	11.6%	0.01561	5	Checkmol fingerprint #199: alkene

fea id	pos(%)	neg(%)	pos-neg(%)	fea imp	
55	25.4%	10.3%	15.2%	0.01617	Checkmol fingerprint #055: tert. amine
56	21.8%	7.8%	14.0%	0.01409	Checkmol fingerprint #056: tert. aliphat. amine
903	40.3%	27.1%	13.3%	0.01671	Drug Moiety #32: isobutane
48	14.7%	22.6%	7.9%	0.00831	Checkmol fingerprint #048: prim. amine
904	15.8%	8.0%	7.8%	0.00846	Drug Moiety #33: neopentane

5. Conclusion and Future work

Although the performance is not well enough, at least it runnable and can provide some suggestions(an prediction that is 0.7 accurate). What's more, the strongest advantage of feature based model is that it is interpretable, we don't just simply get the answer from the model, we can also explain the importance of the features and maybe try to find out the reason that cause specific compounds can or cannot be approved. I think if I am going to improve the model in the future, I will analyze all the moiety systems carefully and maybe give them some weight, since not every substructure is equally important in Chemistry. And of course, I will also improve my model without changing the dataset and the method I extract features.