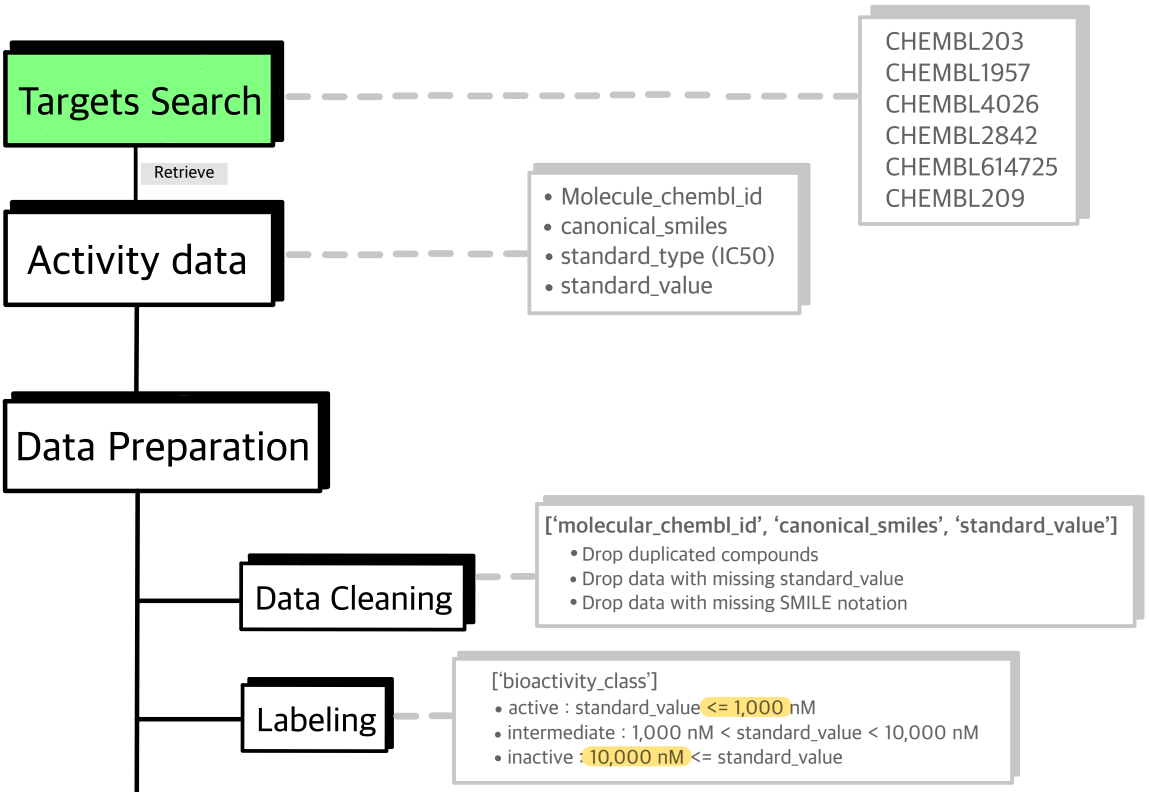
🟩 Finishsed

🟥 In Progress

DRUG TARGET IDENTIFICATION

FLOW DIAGRAM



Targets Search

CHEMBL203
CHEMBL1957
CHEMBL4026
CHEMBL2842
CHEMBL614725
CHEMBL209

Retrieve

Activity data

- Molecule_chembl_id
- canonical_smiles
- standard_type (IC50)
- standard_value

Data Preparation

Data Cleaning

['molecular_chembl_id', 'canonical_smiles', 'standard_value']

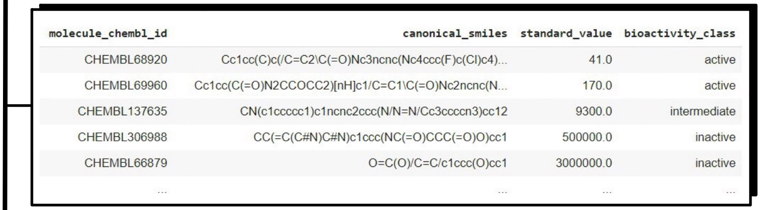
- Drop duplicated compounds
- Drop data with missing standard_value
- Drop data with missing SMILE notation

Labeling

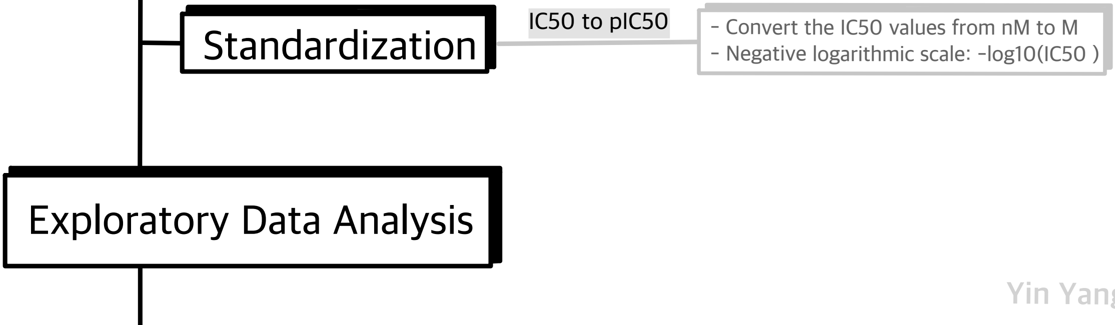
['bioactivity_class']

- active : standard_value <= 1,000 nM
- intermediate : 1,000 nM < standard_value < 10,000 nM
- inactive : 10,000 nM <= standard_value

data to .CSV file



molecule_chembl_id	canonical_smiles	standard_value	bioactivity_class
CHEMBL68920	Cc1cc(C)c(/C=C2\C(=O)Nc3ncnc(Nc4ccc(F)c(Cl)c4)...	41.0	active
CHEMBL69960	Cc1cc(C(=O)N2CCOCC2)[nH]c1/C=C1\C(=O)Nc2ncnc(N...	170.0	active
CHEMBL137635	CN(c1ccccc1)c1ncnc2ccc(N/N=N/Cc3ccccn3)cc12	9300.0	intermediate
CHEMBL306988	CC(=C(C#N)C#N)c1ccc(NC(=O)CCC(=O)O)cc1	500000.0	inactive
CHEMBL66879	O=C(O)/C=C/c1ccc(O)cc1	3000000.0	inactive
...	...	...	...



Standardization

IC50 to pIC50

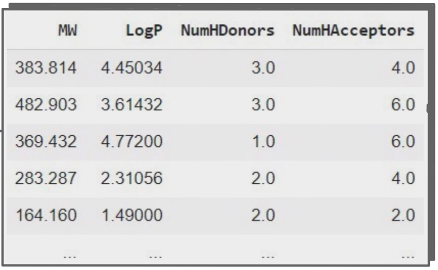
- Convert the IC50 values from nM to M
- Negative logarithmic scale: -log10(IC50)

Exploratory Data Analysis

Yin Yang

Lipinski 5-rule descripters

- Molecular Weight
- log P (solubility)
- # of hydrogen bond doners
- # of hydrogen bond acceptors

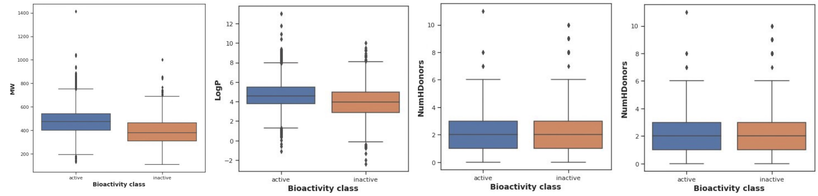


	MW	LogP	NumHDonors	NumHAcceptors
	383.814	4.45034	3.0	4.0
	482.903	3.61432	3.0	6.0
	369.432	4.77200	1.0	6.0
	283.287	2.31056	2.0	4.0
	164.160	1.49000	2.0	2.0
	...	...	...	...

Chemical Space Analysis

Compare active/inactive groups

Box-plots, hypothesis test...



Evaluating the drug-likeness

Active group

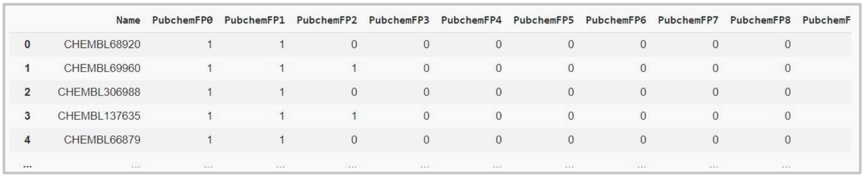
- MW < 500 Dalton
- Octanol-water partition coefficient(LogP) < 5
- Hydrogen bond donor < 5
- Hydrogen bond acceptors < 10

Molecular Descriptor Calculation

PaDEL-Descriptor

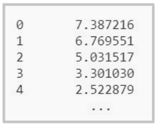
http://pubmed.nobility.nvm.nih.gov/214252294/

Input (PubChem fingerprint)



	Name	PubchemFP0	PubchemFP1	PubchemFP2	PubchemFP3	PubchemFP4	PubchemFP5	PubchemFP6	PubchemFP7	PubchemFP8	PubchemF
0	CHEMBL68920	1	1	0	0	0	0	0	0	0	
1	CHEMBL69960	1	1	1	0	0	0	0	0	0	
2	CHEMBL306988	1	1	0	0	0	0	0	0	0	
3	CHEMBL137635	1	1	1	0	0	0	0	0	0	
4	CHEMBL66879	1	1	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	

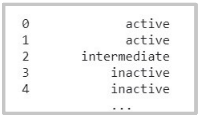
Output (pIC50)



```
0        7.387216
1        6.769551
2        5.031517
3        3.301030
4        2.522879
           ...
```

Regression

Output (active/inactive)



```
0              active
1              active
2        intermediate
3            inactive
4            inactive
              ...
```

Classification

Yin Yang

Modeling

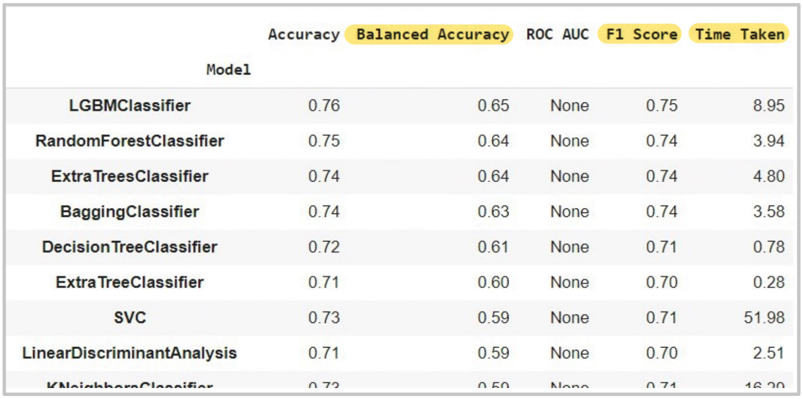
Regression/classification

Remove Low Variance Features

Train/Test Split (80/20)

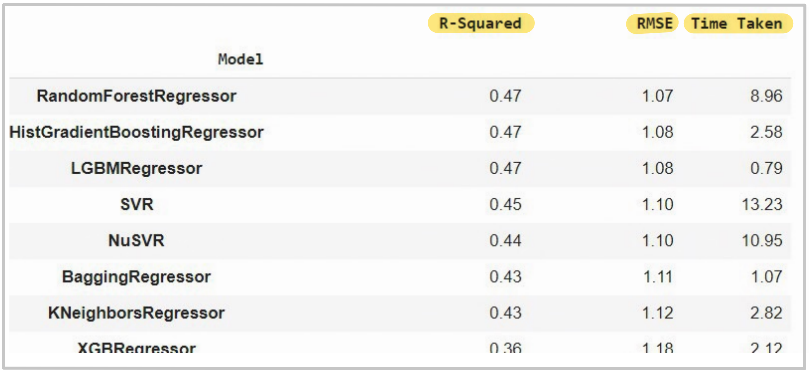
LazyPredict

LazyClasssifier



	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
Model					
LGBMClassifier	0.76	0.65	None	0.75	8.95
RandomForestClassifier	0.75	0.64	None	0.74	3.94
ExtraTreesClassifier	0.74	0.64	None	0.74	4.80
BaggingClassifier	0.74	0.63	None	0.74	3.58
DecisionTreeClassifier	0.72	0.61	None	0.71	0.78
ExtraTreeClassifier	0.71	0.60	None	0.70	0.28
SVC	0.73	0.59	None	0.71	51.98
LinearDiscriminantAnalysis	0.71	0.59	None	0.70	2.51
KNeighborsClassifier	0.73	0.59	None	0.71	16.20

LazyRegressor



	R-Squared	RMSE	Time Taken
Model			
RandomForestRegressor	0.47	1.07	8.96
HistGradientBoostingRegressor	0.47	1.08	2.58
LGBMRegressor	0.47	1.08	0.79
SVR	0.45	1.10	13.23
NuSVR	0.44	1.10	10.95
BaggingRegressor	0.43	1.11	1.07
KNeighborsRegressor	0.43	1.12	2.82
XGBRegressor	0.36	1.18	2.12

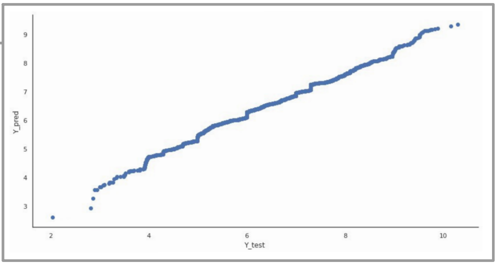
Modeling

Regression Model

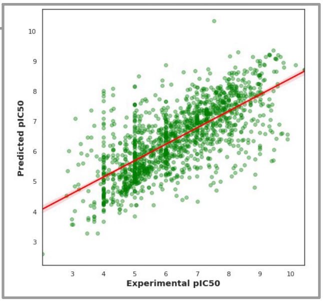
- Decision Tree Regressor
- Extra Trees Regressor

Evaluation

- R-Squared
- Root Mean Squared Error (RMSE)
- Quantile-Quantile Plot



- Fitted-Actual Plot



Classification Model

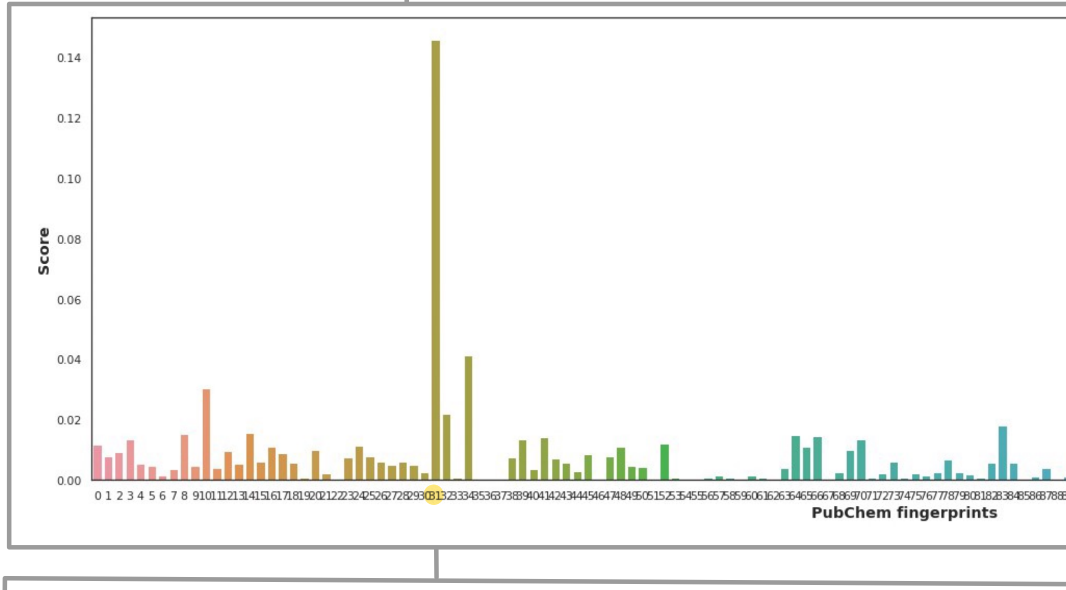
- Decision Tree Classifier
- Extra Trees Classifier

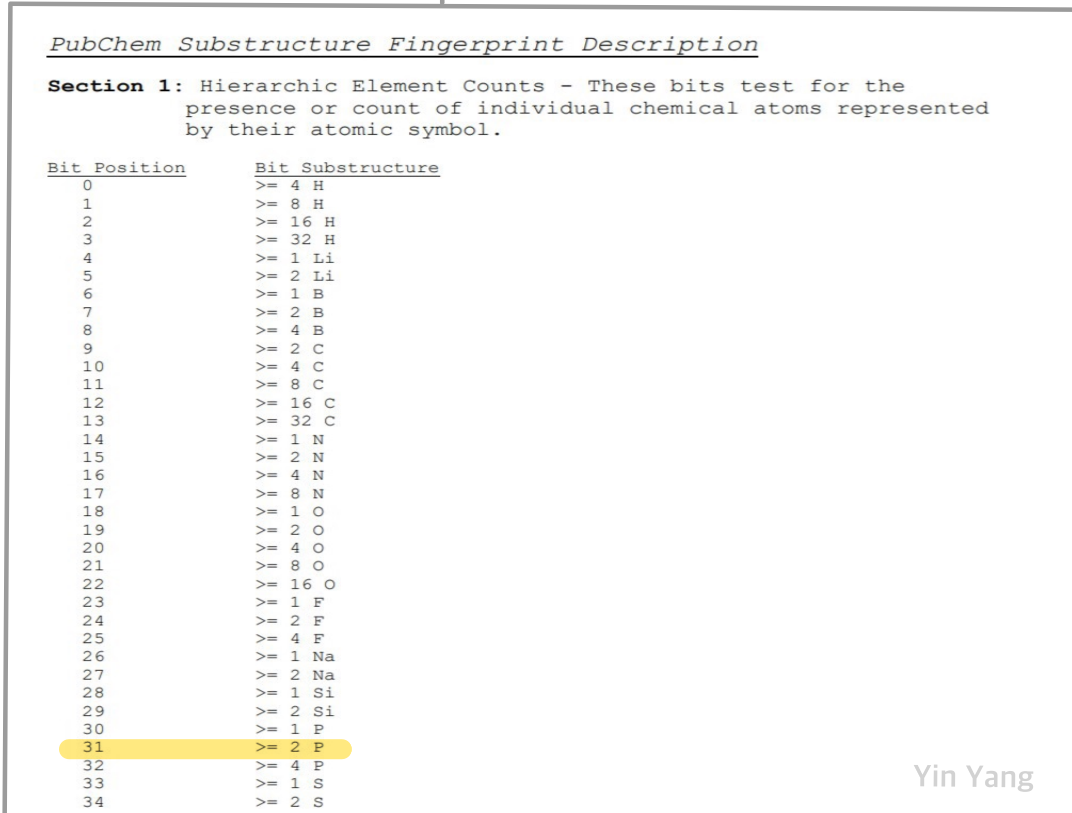
- Prescision
- Recall
- F1-Score

Yin Yang

Feature Importance

For each model





PubChem Substructure Fingerprint Description

Section 1: Hierarchic Element Counts - These bits test for the presence or count of individual chemical atoms represented by their atomic symbol.

Bit Position	Bit Substructure
0	>= 4 H
1	>= 8 H
2	>= 16 H
3	>= 32 H
4	>= 1 Li
5	>= 2 Li
6	>= 1 B
7	>= 2 B
8	>= 4 B
9	>= 2 C
10	>= 4 C
11	>= 8 C
12	>= 16 C
13	>= 32 C
14	>= 1 N
15	>= 2 N
16	>= 4 N
17	>= 8 N
18	>= 1 O
19	>= 2 O
20	>= 4 O
21	>= 8 O
22	>= 16 O
23	>= 1 F
24	>= 2 F
25	>= 4 F
26	>= 1 Na
27	>= 2 Na
28	>= 1 Si
29	>= 2 Si
30	>= 1 P
31	>= 2 P
32	>= 4 P
33	>= 1 S
34	>= 2 S