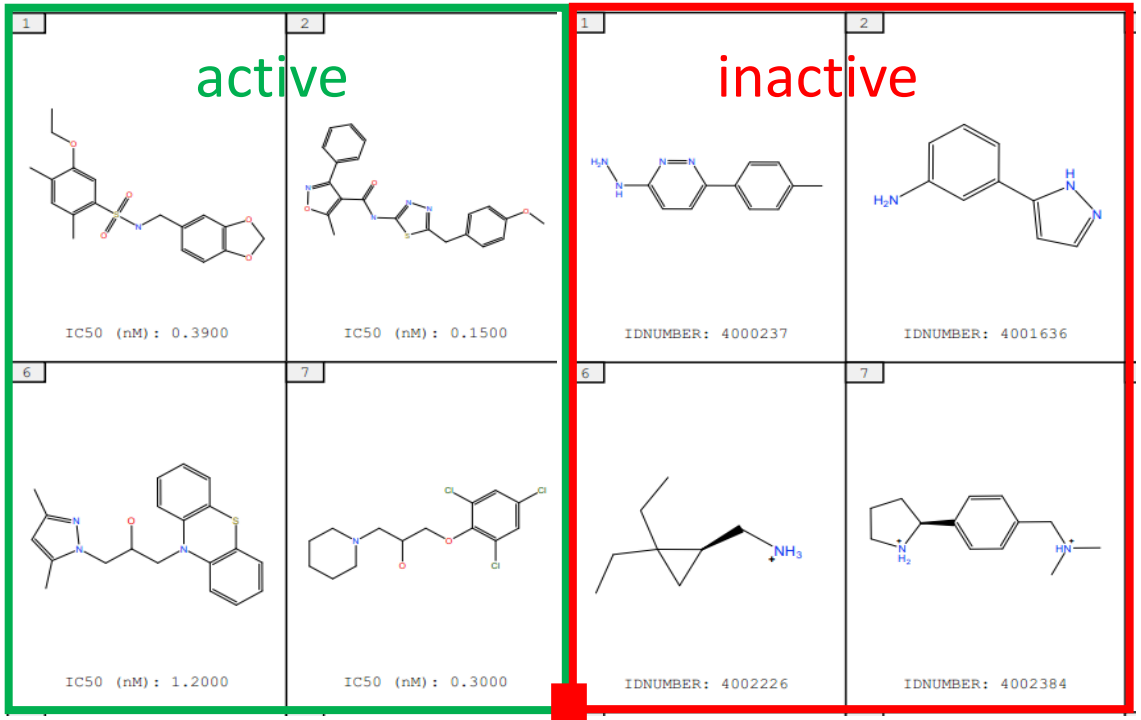
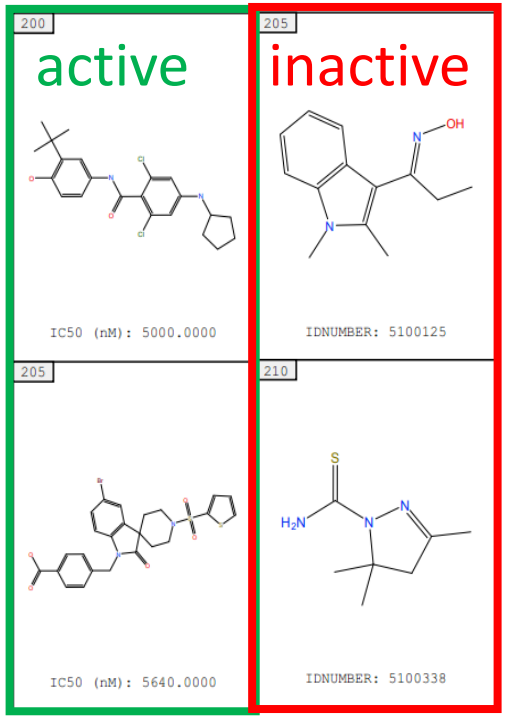
基于机器学习的判别模型构建

判别模型建模一般流程

- 已知活性数据收集
- 数据集预处理（正样本/负样本、训练集/测试集准备等）
- 分子描述属性计算（传统分子描述符、分子指纹等）
- 模型构建（机器学习算法）
- 外部数据集检测
- 未知化合物类别预测

以朴素贝叶斯为例对FXR活性剂与非活性剂进行ML判别模型构建

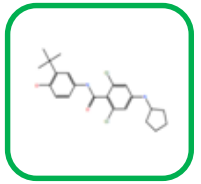




标签值	描述符Xa	描述符Xb	描述符Xc	……
active	Xa1	Xb1	Xc1	……
active	Xa2	Xb2	Xc2	……
inactive	Xa3	Xb3	Xc3	……
……	……	……	……	……

Bayes

描述符Xa	描述符Xb	描述符Xc	……
Xa1	Xb1	Xc1	……
Xa2	Xb2	Xc2	……
Xa3	Xb3	Xc3	……

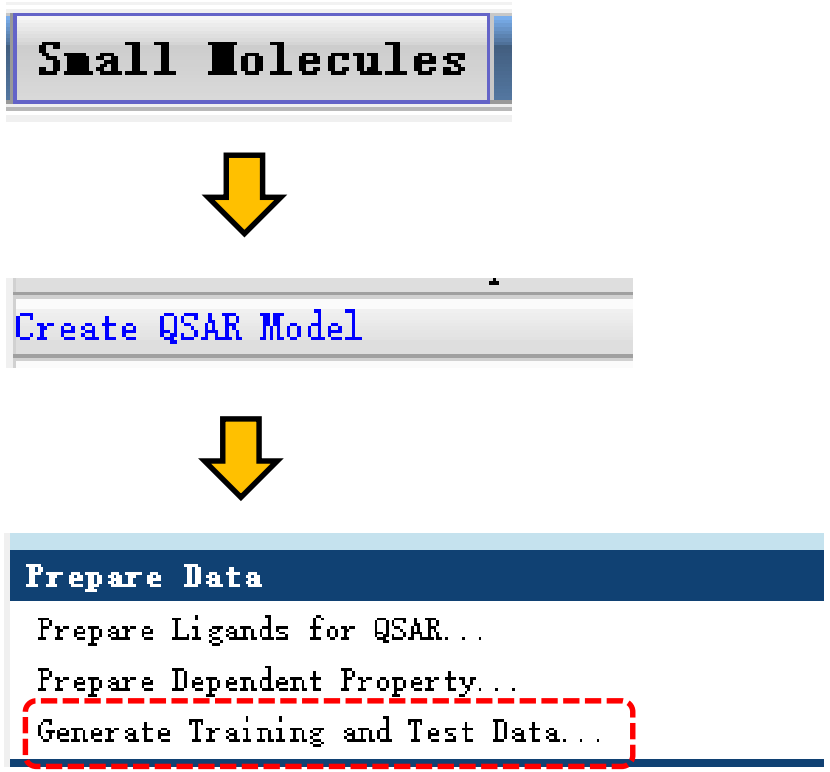


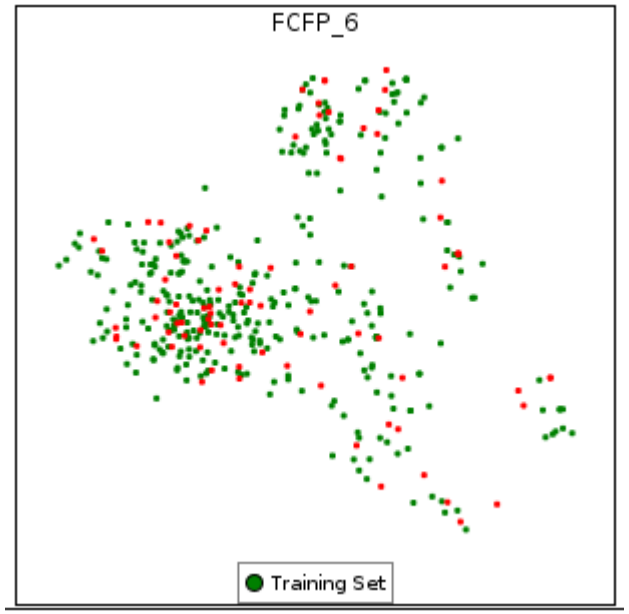
active, 80%

inactive, 20%

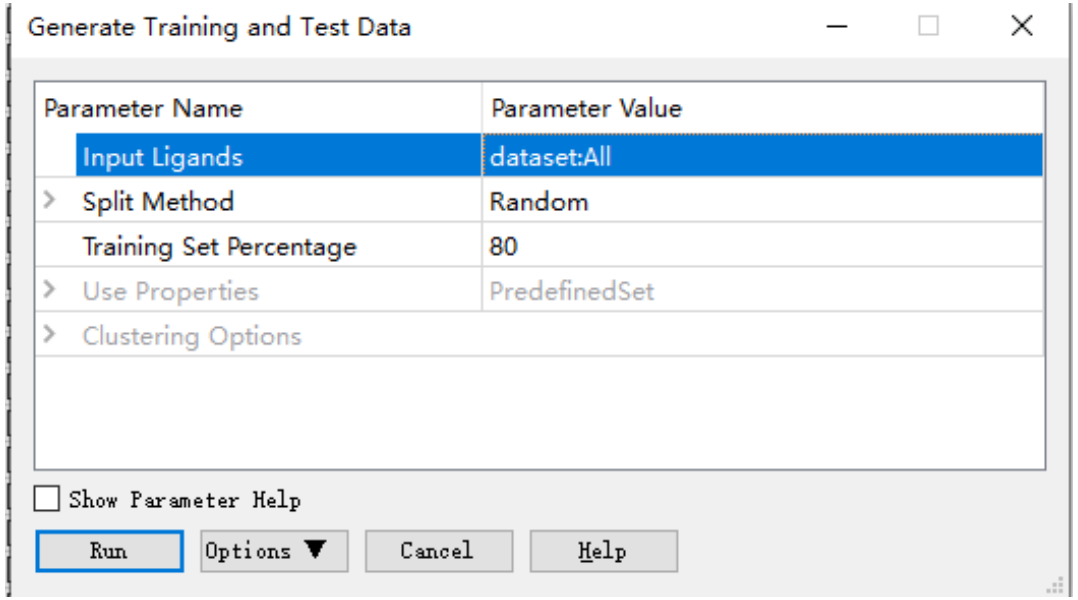
√

数据集准备

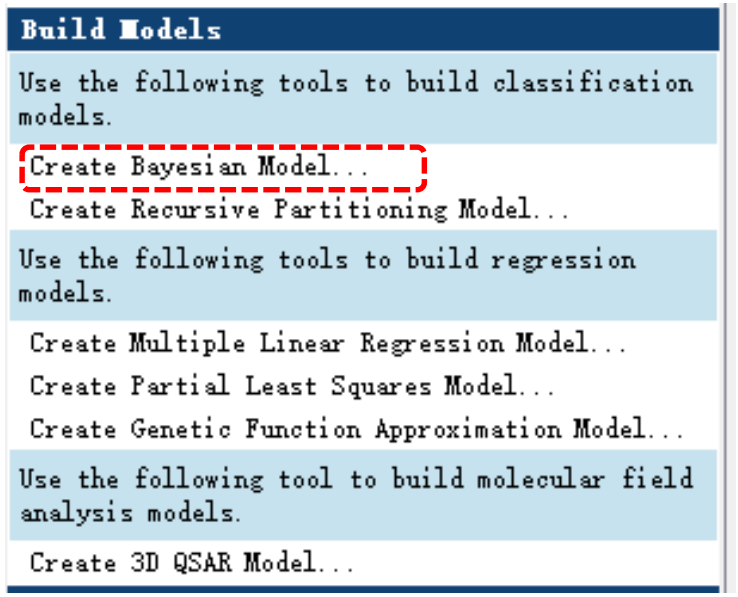


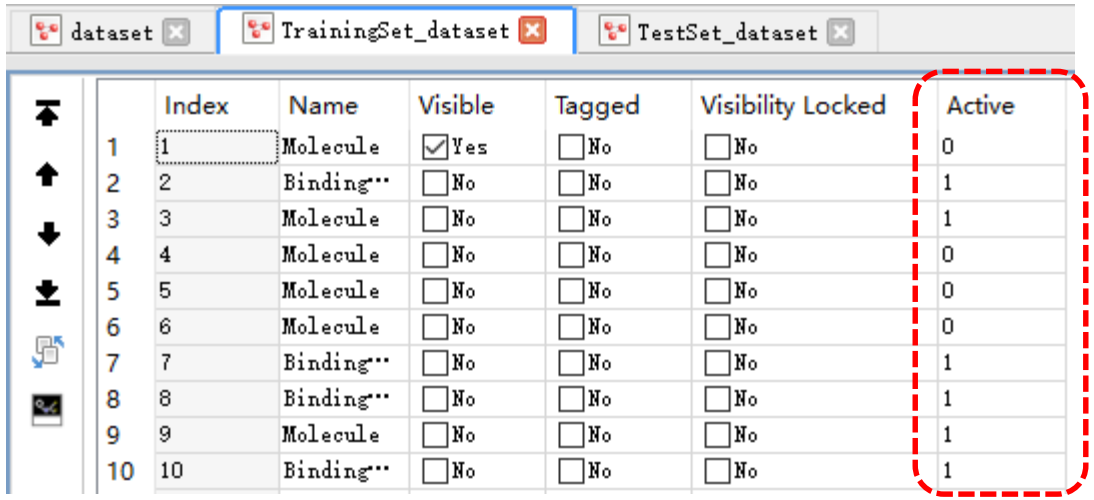


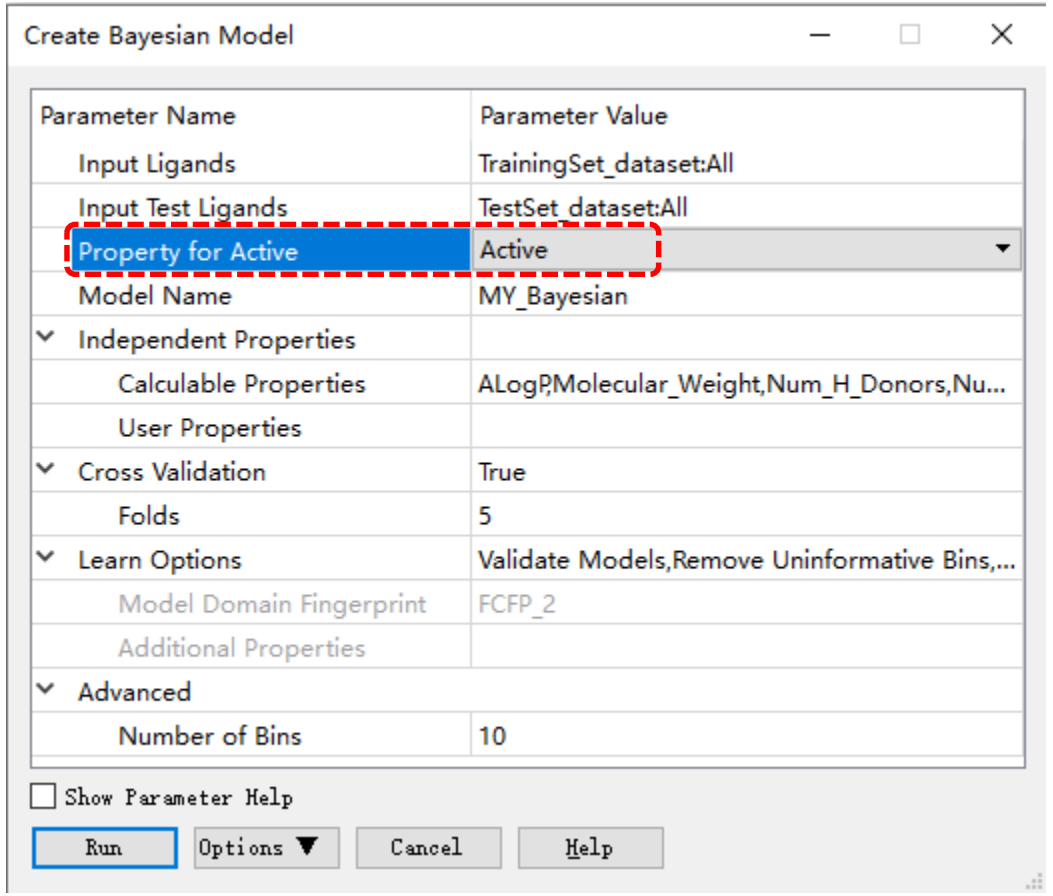
- TrainingSet
- TestSet



模型的参数设置与构建


Build Models
Use the following tools to build classification models.
Create Bayesian Model...
Create Recursive Partitioning Model...
Use the following tools to build regression models.
Create Multiple Linear Regression Model...
Create Partial Least Squares Model...
Create Genetic Function Approximation Model...
Use the following tool to build molecular field analysis models.
Create 3D QSAR Model...

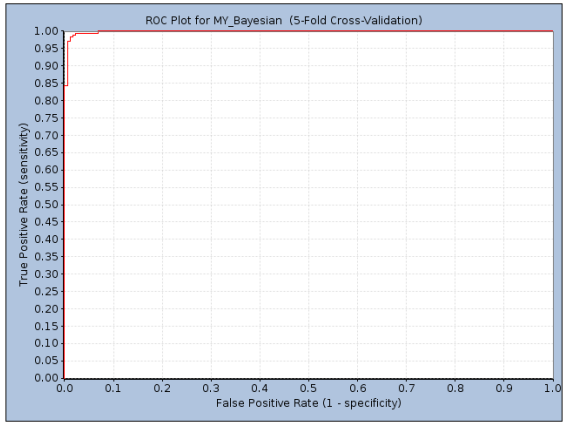

dataset
TrainingSet_dataset
TestSet_dataset
Index Name Visible Tagged Visibility Locked Active
1 1 Molecule Yes No No No 0
2 2 Binding··· No No No No 1
3 3 Molecule No No No No 1
4 4 Molecule No No No No 0
5 5 Molecule No No No No 0
6 6 Molecule No No No No 0
7 7 Binding··· No No No No 1
8 8 Binding··· No No No No 1
9 9 Molecule No No No No 1
10 10 Binding··· No No No No 1


Create Bayesian Model
Parameter Name
Parameter Value
Input Ligands
TrainingSet_dataset:All
Input Test Ligands
TestSet_dataset:All
Property for Active
Active
Model Name
MY_Bayesian
Independent Properties
Calculable Properties
ALogP,Molecular_Weight,Num_H_Donors,Nu...
User Properties
Cross Validation
True
Folds
5
Learn Options
Validate Models,Remove Uninformative Bins,...
Model Domain Fingerprint
FCFP_2
Additional Properties
Advanced
Number of Bins
10
Show Parameter Help
Run
Options
Cancel
Help

结果分析

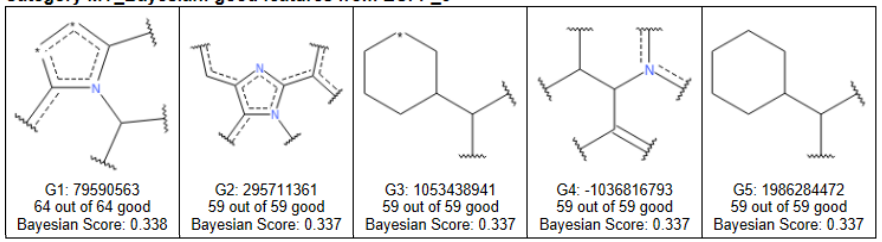
5-Fold Cross-Validation Result									
Model Name	ROC Score	ROC Rating	True Positive	False Negative	False Positive	True Negative	Sensitivity	Specificity	Concordance
MY_Bayesian	0.998	Excellent	171	1	0	174	0.994	1.000	0.997

GA



	Index	Name	Visible	Tagged	Visibility Locked	Active	MY_Bayesian	MY_Bayesian#Prediction
40	30	Molecule	No	No	No	1	14.2845	true
41	31	Molecule	No	No	No	1	11.8302	true
42	32	Molecule	No	No	No	1	11.6981	true
43	33	Molecule	No	No	No	1	9.0923	true
44	36	Molecule	No	No	No	1	3.7689	true
45	48	Molecule	No	No	No	0	-14.5087	false
46	78	Molecule	No	No	No	0	-25.5528	false
47	45	Molecule	No	No	No	0	-12.6683	false
48	46	Molecule	No	No	No	0	-12.7889	false
49	47	Molecule	No	No	No	0	-14.358	false

Category MY_Bayesian: good features from ECFP_6



G1: 79590563
64 out of 64 good
Bayesian Score: 0.338

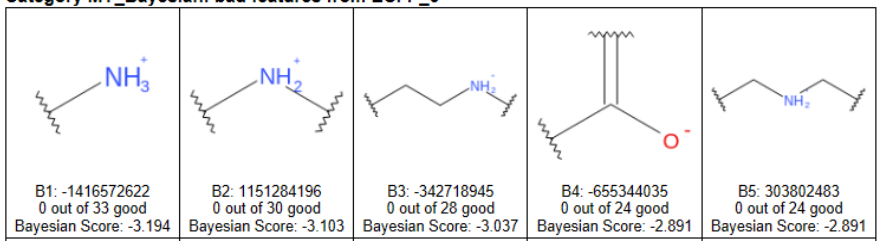
G2: 295711361
59 out of 59 good
Bayesian Score: 0.337

G3: 1053438941
59 out of 59 good
Bayesian Score: 0.337

G4: -1036816793
59 out of 59 good
Bayesian Score: 0.337

G5: 1986284472
59 out of 59 good
Bayesian Score: 0.337

Category MY_Bayesian: bad features from ECFP_6



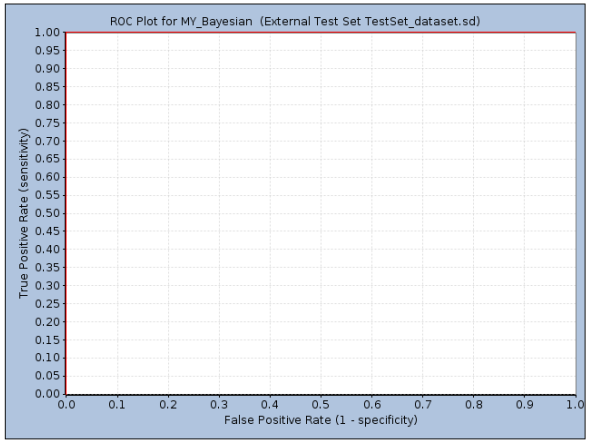
B1: -1416572622
0 out of 33 good
Bayesian Score: -3.194

B2: 1151284196
0 out of 30 good
Bayesian Score: -3.103

B3: -342718945
0 out of 28 good
Bayesian Score: -3.037

B4: -655344035
0 out of 24 good
Bayesian Score: -2.891

B5: 303802483
0 out of 24 good
Bayesian Score: -2.891



Validation Result Using External Test Set TestSet_dataset.sd									
Model Name	ROC Score	ROC Rating	True Positive	False Negative	False Positive	True Negative	Sensitivity	Specificity	Concordance
MY_Bayesian	1.000	Excellent	43	1	0	42	0.977	1.000	0.988

未知化合物活性预测

Calculate Molecular Properties

Property Calculation

Basic Arithmetic...

Calculate RMSD...

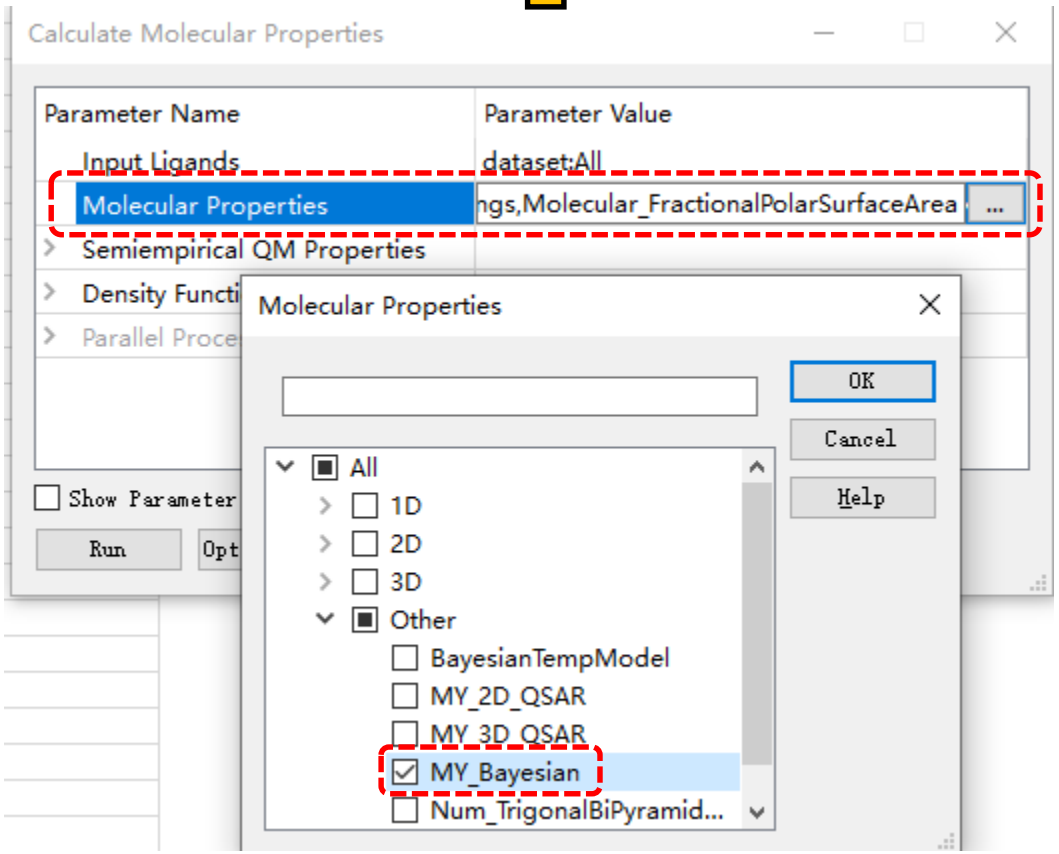
Calculate Molecular Properties...

Calculate Ligand Efficiency...

dataset

	Index	Name	Visible	Tagged	Visibility Locked	index
1	1	Molecule	Yes	No	No	1
2	2	Molecule	No	No	No	2
3	3	Molecule	No	No	No	3
4	4	Molecule	No	No	No	4
5	5	Molecule	No	No	No	5
6	6	Molecule	No	No	No	6
7	7	Molecule	No	No	No	7
8	8	Molecule	No	No	No	8
9	9	Molecule	No	No	No	9
10	10	Molecule	No	No	No	10

	Index	Name	Visible	Tagged	Visibility Locked	MY_Bayesian	MY_Bayesian#Prediction
1	1	1288666	Yes	No	No	-3.20069	true
2	2	1320138	No	No	No	4.79479	true
3	3	1333937	No	No	No	2.05382	true
4	4	2197808	No	No	No	9.37912	true
5	5	2260458	No	No	No	9.71728	true
6	6	2897692	No	No	No	4.33741	true
7	7	2914648	No	No	No	5.33722	true
8	8	2954950	No	No	No	6.9356	true
9	9	3333	No	No	No	5.41247	true
10	10	3567787	No	No	No	-8.91515	false



任务：

（1）了解数据集中正负样本处理方法

（2）模型构建与结果分析

（3）未知化合物活性预测

实验报告：

（1）实验目的

FXR活性剂判别模型的构建

（2）操作流程

正负样本、训练/测试集准备，分子描述符选择，贝叶斯模型构建，模型分析，未知化合物活性预测

（3）结果与讨论

模型精度的影响因素、化合物结构分析等