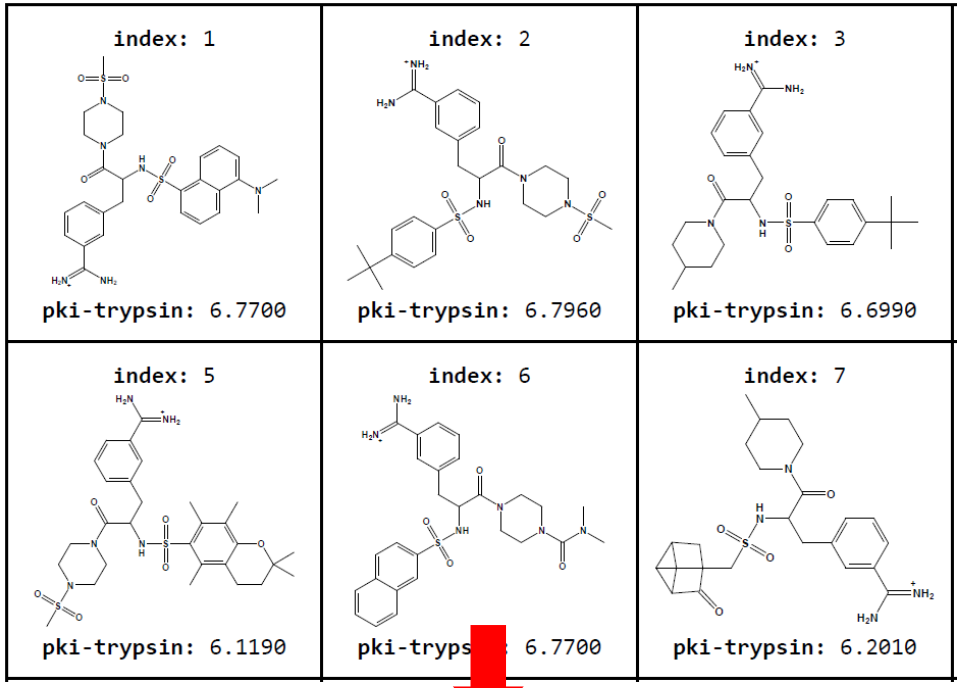
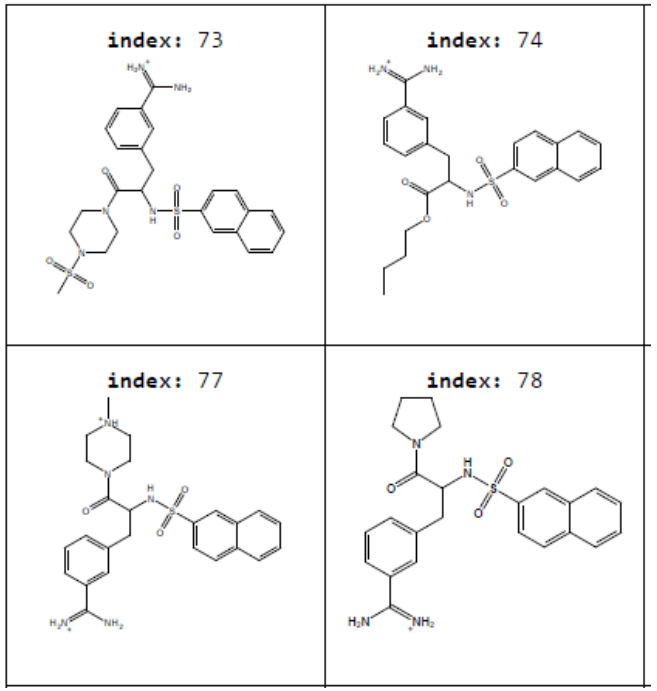
2D-QSAR

QSAR建模一般流程

- 已知活性数据收集
- 数据集准备（训练集与测试集拆分等）
- 分子描述属性计算（传统分子描述符、分子指纹等）
- 模型构建（多元线性回归、偏最小二乘等）
- 外部数据集检测
- 未知活性化合物预测

胰蛋白酶抑制剂QSAR模型的构建



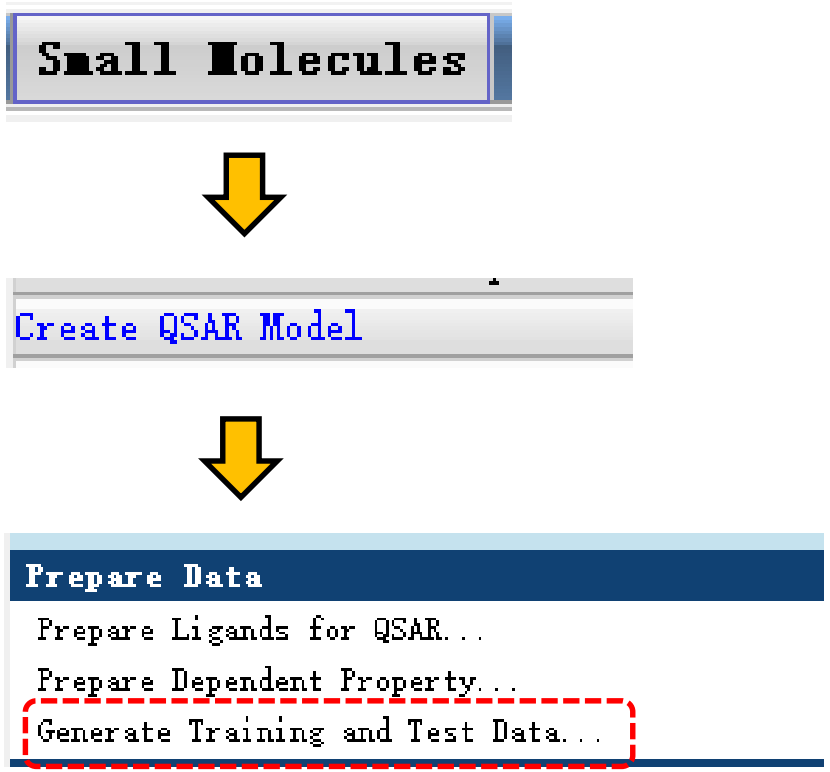


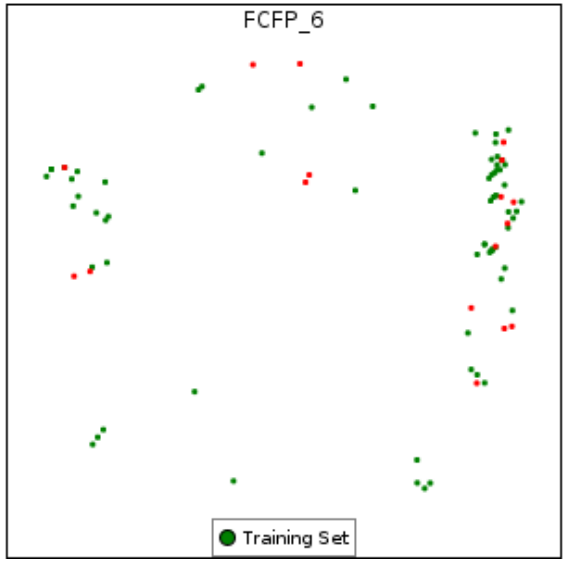
活性值 Y	描述符 Xa	描述符 Xb	描述符 Xc	……
Y1	Xa1	Xb1	Xc1	……
Y2	Xa2	Xb2	Xc2	……
Y3	Xa3	Xb3	Xc3	……
…	……	……	……	……

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

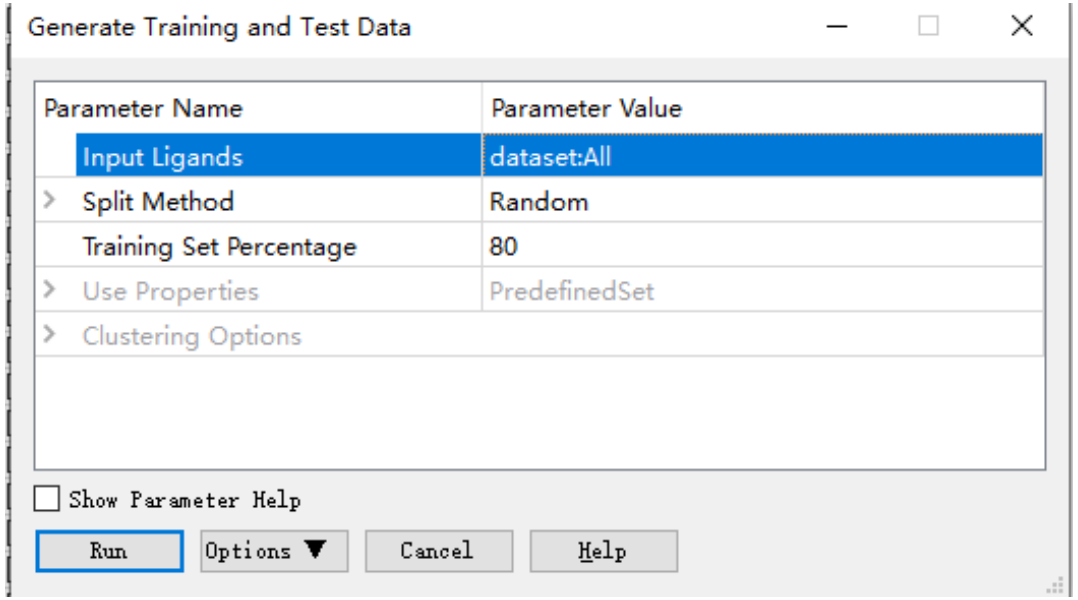
MLR → $Y=a*Xa + b*Xb + c*Xc + d$ → Y_{pred}

数据集准备

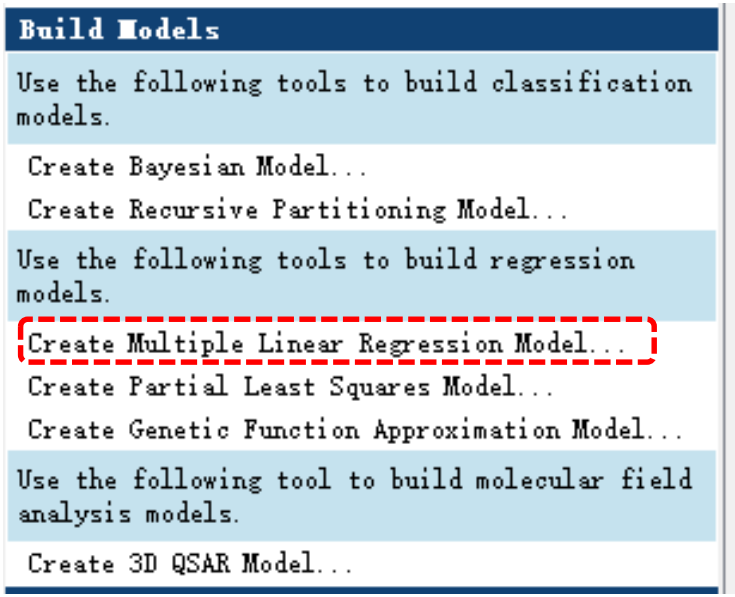


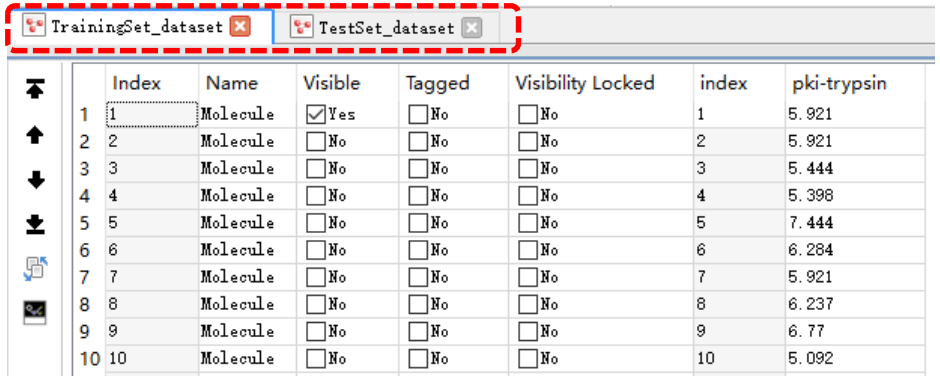


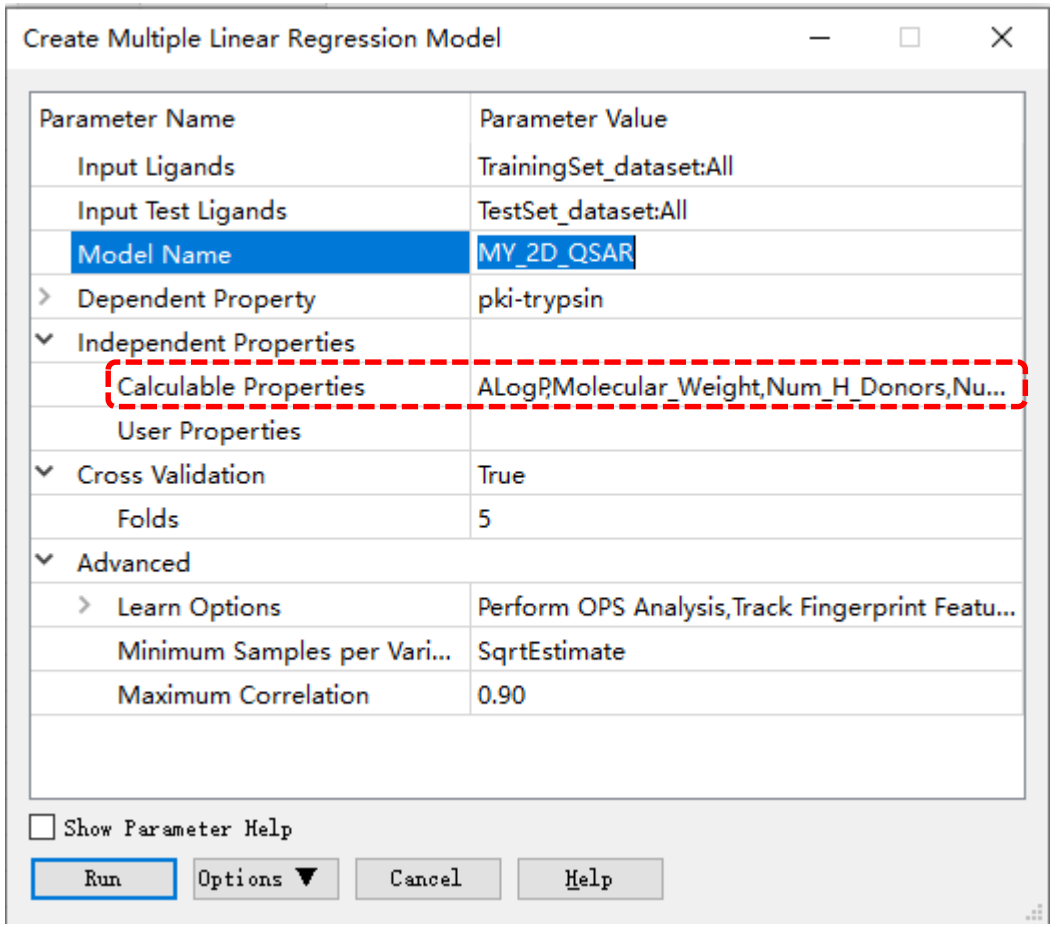
- TrainingSet
- TestSet



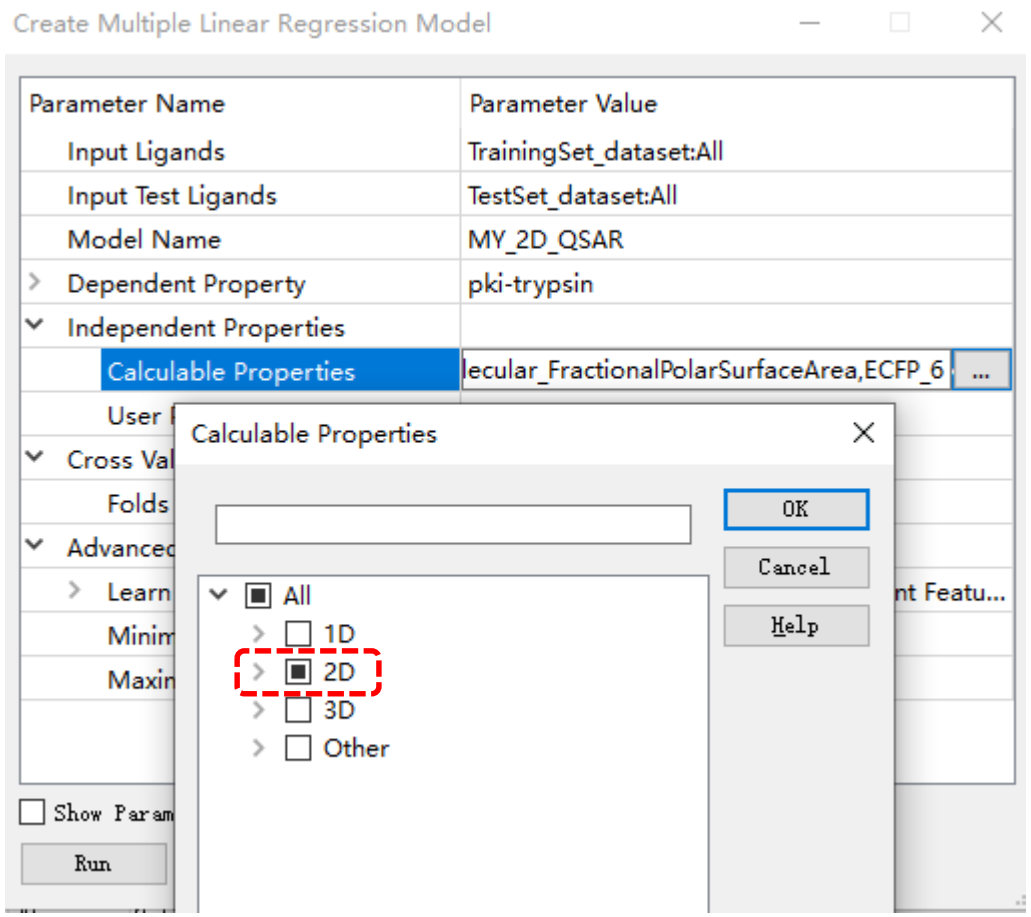
模型的参数设置与构建

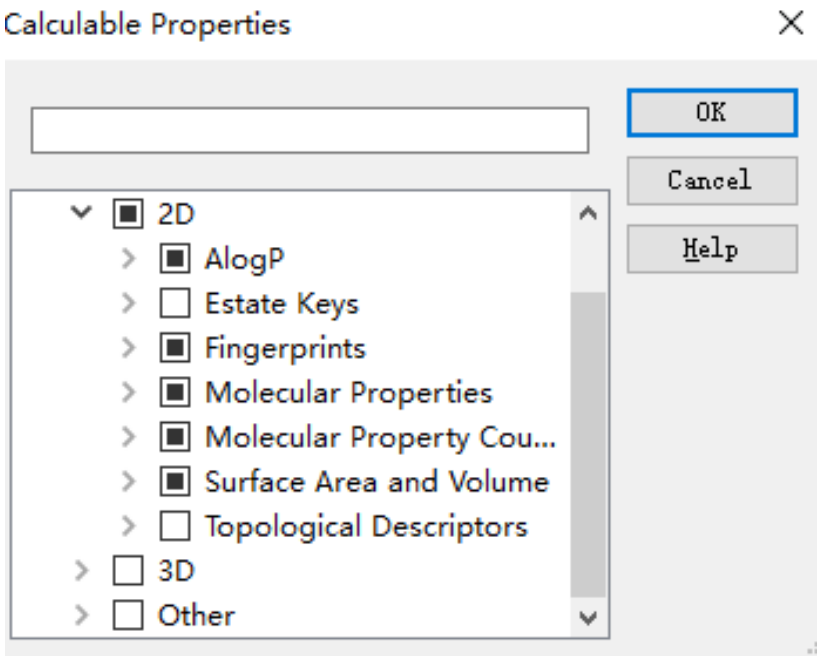

Build Models
Use the following tools to build classification models.
Create Bayesian Model...
Create Recursive Partitioning Model...
Use the following tools to build regression models.
Create Multiple Linear Regression Model...
Create Partial Least Squares Model...
Create Genetic Function Approximation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


TrainingSet_dataset
TestSet_dataset
Index Name Visible Tagged Visibility Locked index pki-trypsin
1 1 Molecule Yes No No 1 5.921
2 2 Molecule No No No 2 5.921
3 3 Molecule No No No 3 5.444
4 4 Molecule No No No 4 5.398
5 5 Molecule No No No 5 7.444
6 6 Molecule No No No 6 6.284
7 7 Molecule No No No 7 5.921
8 8 Molecule No No No 8 6.237
9 9 Molecule No No No 9 6.77
10 10 Molecule No No No 10 5.092


Create Multiple Linear Regression Model
Parameter Name
Parameter Value
Input Ligands
TrainingSet_dataset:All
Input Test Ligands
TestSet_dataset:All
Model Name
MY_2D_QSAR
Dependent Property
pki-trypsin
Independent Properties
Calculable Properties
ALogP,Molecular_Weight,Num_H_Donors,Nu...
User Properties
Cross Validation
True
Folds
5
Advanced
Learn Options
Perform OPS Analysis,Track Fingerprint Featu...
Minimum Samples per Vari...
SqrtEstimate
Maximum Correlation
0.90
Show Parameter Help
Run
Options
Cancel
Help

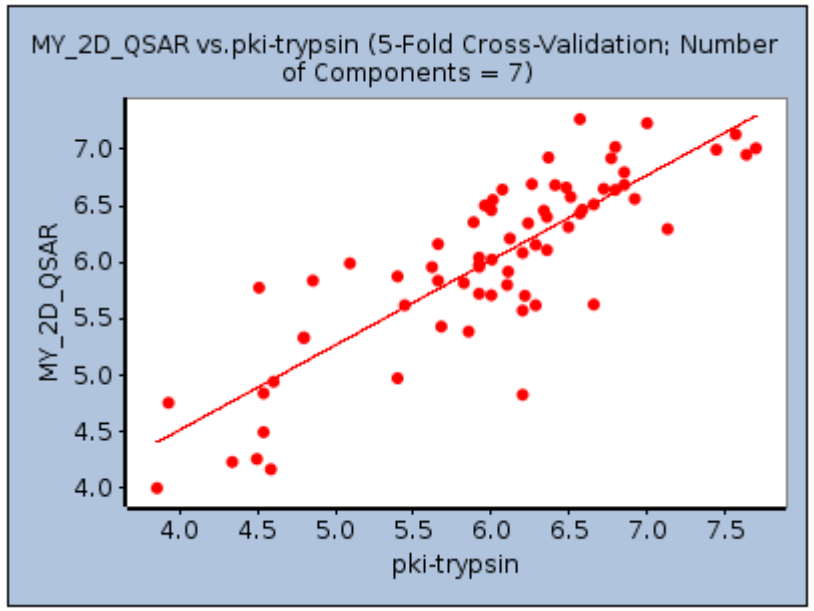
分子描述符选择


Create Multiple Linear Regression Model
Parameter Name
Parameter Value
Input Ligands
TrainingSet_dataset:All
Input Test Ligands
TestSet_dataset:All
Model Name
MY_2D_QSAR
Dependent Property
pki-trypsin
Independent Properties
Calculable Properties
Show Param
Run
Calculable Properties
OK
Cancel
Help
All
1D
2D
3D
Other

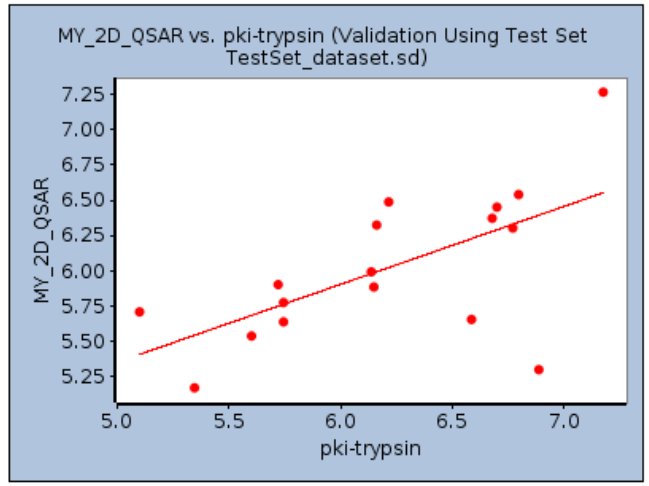

Calculable Properties
OK
Cancel
Help
2D
AlogP
Estate Keys
Fingerprints
Molecular Properties
Molecular Property Cou...
Surface Area and Volume
Topological Descriptors
3D
Other

结果分析

5-Fold Cross Validation Result			
Number of Components	q^2	RMS Error	Mean Absolute Error
1	0.191	0.776	0.644
2	0.504	0.609	0.471
3	0.629	0.527	0.402
4	0.668	0.498	0.392
5	0.698	0.475	0.375
6	0.697	0.476	0.376
7	0.704	0.472	0.366
8	0.695	0.484	0.366
9	0.685	0.494	0.369



	Index	Name	Visible	Tagged	Visibility Locked	index	pki-trypsin	MY_2D_QSAR
1	1	Molecule	☑Yes	☐No	☐No	1	5.745	5.77725
2	2	Molecule	☐No	☐No	☐No	2	6.137	5.99409
3	3	Molecule	☐No	☐No	☐No	3	6.796	6.54063
4	4	Molecule	☐No	☐No	☐No	4	5.602	5.53986
5	5	Molecule	☐No	☐No	☐No	5	6.886	5.30205
6	6	Molecule	☐No	☐No	☐No	6	6.678	6.37342
7	7	Molecule	☐No	☐No	☐No	7	6.699	6.45217
8	8	Molecule	☐No	☐No	☐No	8	6.215	6.48859
9	9	Molecule	☐No	☐No	☐No	9	5.347	5.17323
10	10	Molecule	☐No	☐No	☐No	10	5.745	5.63986
11	11	Molecule	☐No	☐No	☐No	11	6.77	6.30431
12	12	Molecule	☐No	☐No	☐No	12	5.721	5.90382
13	13	Molecule	☐No	☐No	☐No	13	6.149	5.88642
14	14	Molecule	☐No	☐No	☐No	14	6.161	6.32507
15	15	Molecule	☐No	☐No	☐No	15	6.585	5.65766
16	16	Molecule	☐No	☐No	☐No	16	5.102	5.71113
17	17	Molecule	☐No	☐No	☐No	17	7.174	7.26617



Validation Result Using External Test Set TestSet_dataset.sd			
Model Name	q^2	RMS Error	Mean Absolute Error
MY_2D_QSAR	0.388	0.512	0.346

未知化合物活性预测

Calculate Molecular Properties

Property Calculation

Basic Arithmetic...

Calculate RMSD...

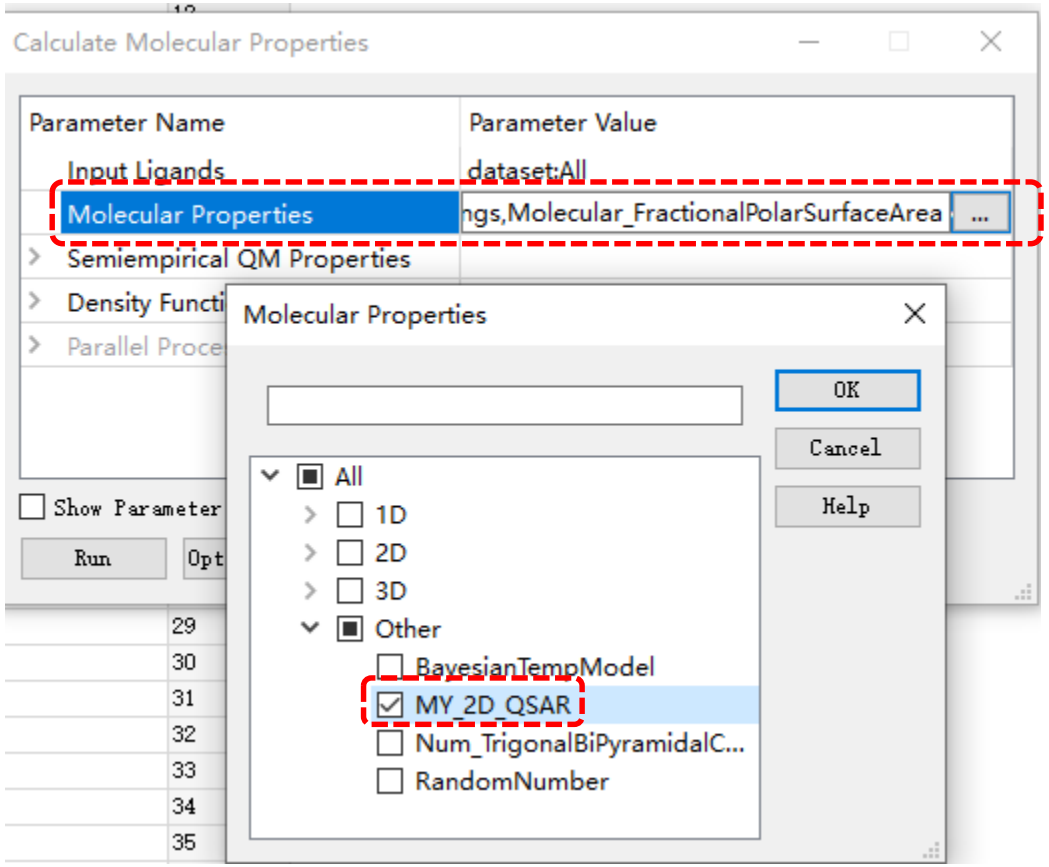
Calculate Molecular Properties...

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	index	MY_2D_QSAR
1	1	Molecule	☑ Yes	☐ No	☐ No	1	6.30431
2	2	Molecule	☐ No	☐ No	☐ No	2	6.54063
3	3	Molecule	☐ No	☐ No	☐ No	3	6.45217
4	4	Molecule	☐ No	☐ No	☐ No	4	6.8804
5	5	Molecule	☐ No	☐ No	☐ No	5	6.08256
6	6	Molecule	☐ No	☐ No	☐ No	6	6.89123
7	7	Molecule	☐ No	☐ No	☐ No	7	6.09373
8	8	Molecule	☐ No	☐ No	☐ No	8	6.24483
9	9	Molecule	☐ No	☐ No	☐ No	9	7.10912
10	10	Molecule	☐ No	☐ No	☐ No	10	5.30205



任务：

（1）了解数据集中训练集与测试集的拆分方法

（2）分子描述符选择与建模方法确定

（3）模型结果分析

（4）未知化合物活性预测

实验报告：

（1）实验目的

胰蛋白酶抑制剂2D-QSAR建模

（2）操作流程

训练集、测试集准备，分子描述符选择，建模方法确认，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