NJmat软件

NJmatML: 材料科学机器学习软件

Python版

NJmatML 0.0.5

pip install NJmatML

- pip install NJmatML
- https://pypi.org/project/NJmatML/
- publicly available
- 用户友好：降低pandas, matplotlib, sklearn, matn
- 功能：文件导入、可视化、机器学习建模、**特征生成**（重要性排名）、特征选择、准确率计算、遗传算法、
- 说明书：https://github.com/Zhang-NJ-Lab/NJmatML/blob/main/2022-11-21/NJ

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号：软著登字第11204908号

软 件 名 称：材料大数据与人工智能软件 [简称：NJmatML] V1.0

著 作 权 人：张磊;毛骏飞;刘柿岑

开发完成日期：2022年12月29日

首次发表日期：未发表

权利取得方式：原始取得

权 利 范 围：全部权利

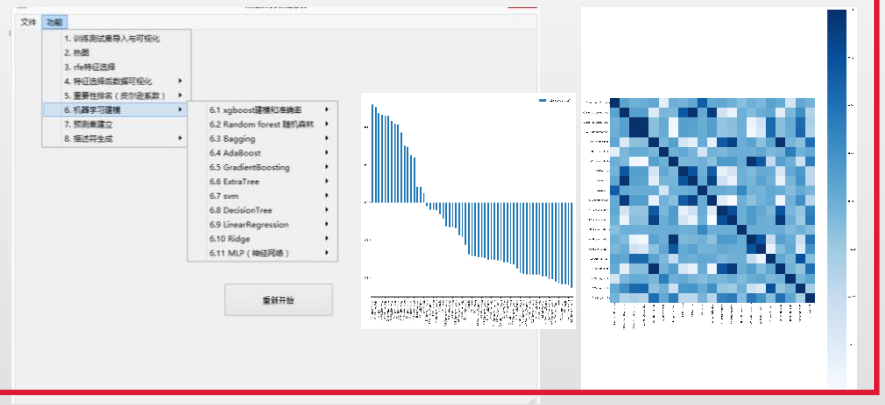
登 记 号：2023SR0617737

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

No.11696395

2023年06月09日

windows版



- **NJmatML的windows版本**
- **下载地址: https://figshare.com/articles/software/NJmatML/24607893**
- **doi: https://doi.org/10.6084/m9.figshare.24607893.v1.exe**

网页版

ZL Lab Nanjing

Machine Learning | Data Import | Feature Engineering | Machine Learning models | Symbolic Regression | Download Report

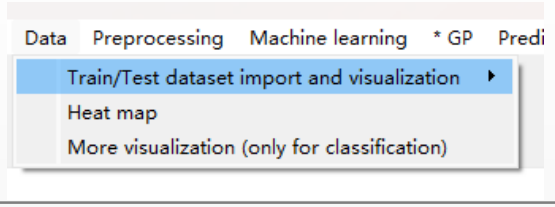
- 网页版 机器学习建模 demo
- https://patrick007.shinyapps.io/zllab_ofi_ml/
- publicly available

实验数据（csv） $\xrightarrow{\text{NJmatML}}$ 材料预测与实验验证

NJmatML: 材料科学机器学习软件

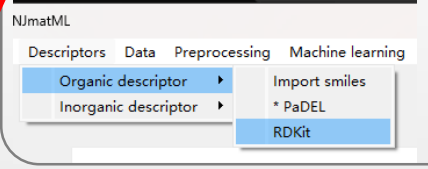
NJmatML主界面

训练测试集导入

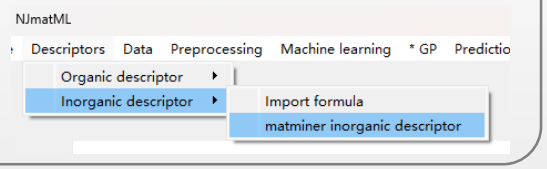


描述符辅助生成：

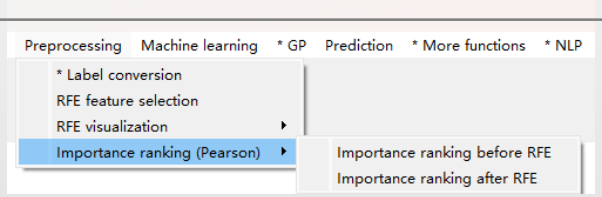
有机材料 (RDKit)



无机材料（Matminer）

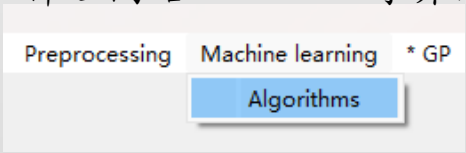


特征选择与可视化

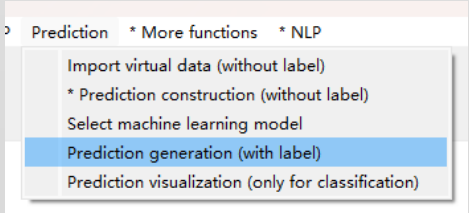


机器学习建模

随机森林/XGBoost/神经网络/Adaboost等算法
回归和分类
遗传算法/Shapley
自主调参

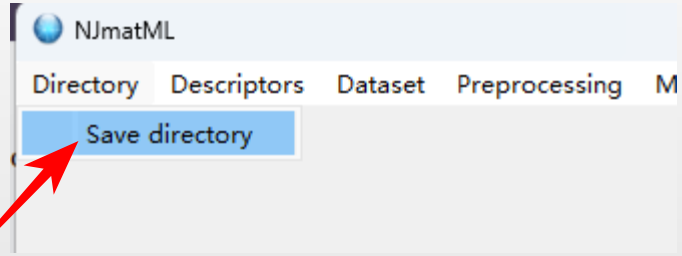


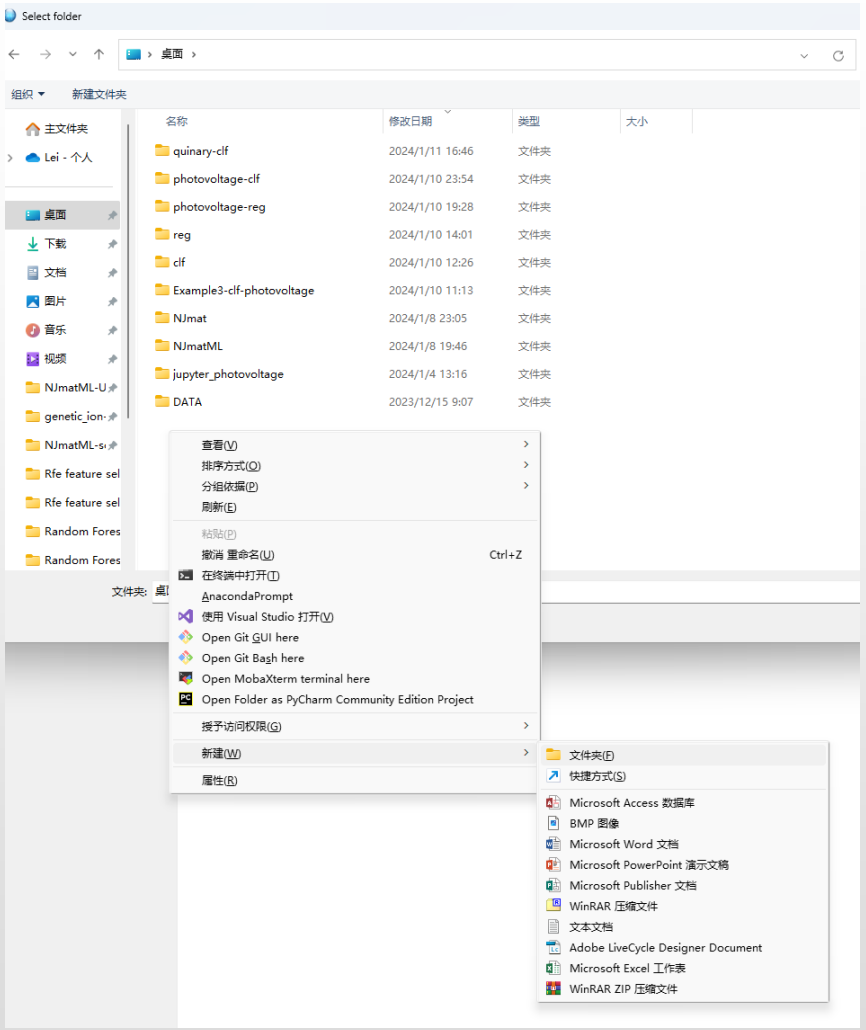
机器学习预测

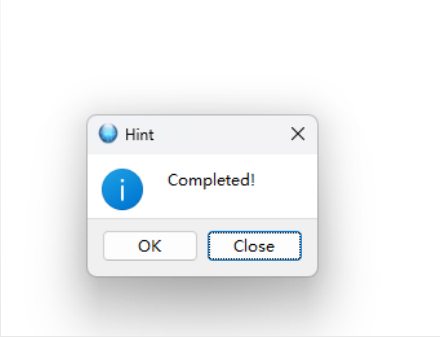


1 选择保存路径

1 选择项目存储文件夹（自定义）







1 选择项目存储文件夹（自定义）

数据集导入

描述符生成

遗传算法建模

传统机器学习建模（分类和回归）

预测虚拟空间

预处理（特征选择等）

Shapley描述符分析

后续在自定义文件夹中生成新的数据文件夹

2 材料特征化 Featurizer

2.1 Featurizer 材料特征化按钮（无机和有机）

2in1:有机无机2合1

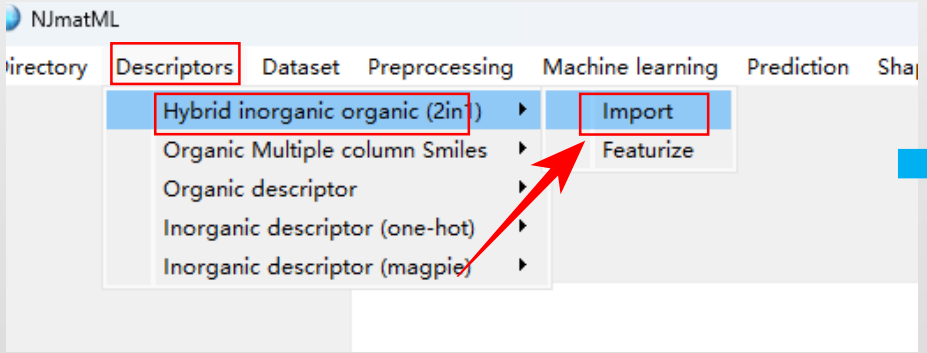
Smiles和化学式都有

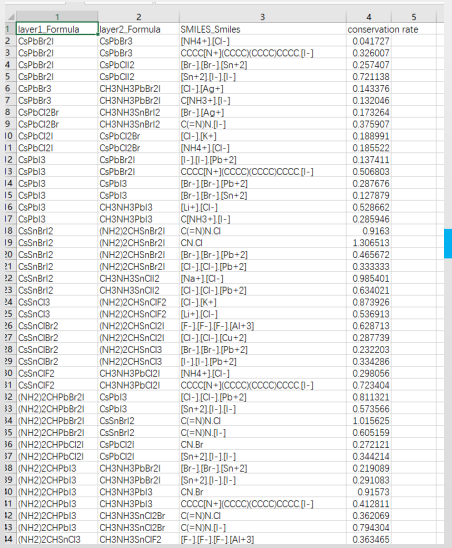
特征化按钮（训练测试集辅助生成）

有机RDKit描述符，无机为Matminer magpie 描述符（元素、原子质量等等）

Descriptors: Hybrid inorganic organic (2in1) → Import → Featurize

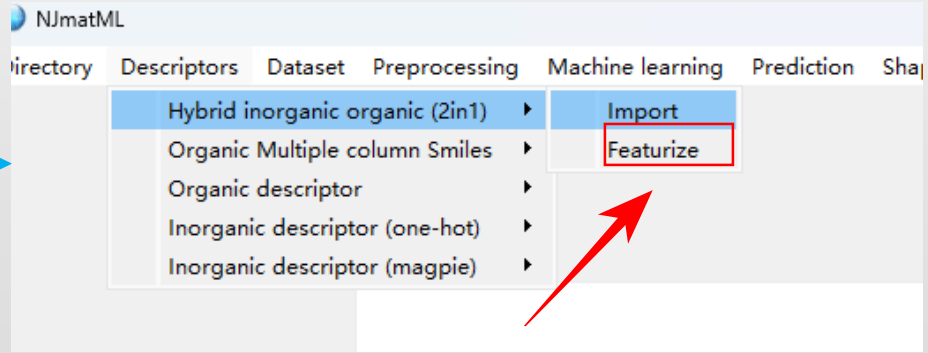
Import raw





	1	2	3	4	5
1	layer1_Formula	layer2_Formula	SMILES_Smiles	conservation rate	
2	CsPbBr2I	CsPbBr3	[NH4+].[Cl-]	0.041727	
3	CsPbBr2I	CsPbBr3	CCCC[N+](CCCC)(CCCC)CCCC.[I-]	0.326007	
4	CsPbBr2I	CsPbCl2	[Br-].[Br-].[Sn+2]	0.257407	
5	CsPbBr2I	CsPbCl2	[Sn+2].[I-].[I-]	0.721138	
6	CsPbBr3	CH3NH3PbBr2I	[Cl-].[Ag+]	0.143376	
7	CsPbBr3	CH3NH3PbBr2I	C[NH3+].[I-]	0.132046	
8	CsPbCl2Br	CH3NH3SnBrI2	[Br-].[Ag+]	0.173264	
9	CsPbCl2Br	CH3NH3SnBrI2	C(=N)N.[I-]	0.375907	
10	CsPbCl2I	CsPbCl2Br	[Cl-].[K+]	0.188991	
11	CsPbCl2I	CsPbCl2Br	[NH4+].[Cl-]	0.185522	
12	CsPbI3	CsPbBr2I	[I-].[I-].[Pb+2]	0.137411	
13	CsPbI3	CsPbBr2I	CCCC[N+](CCCC)(CCCC)CCCC.[I-]	0.506803	
14	CsPbI3	CsPbI3	[Br-].[Br-].[Pb+2]	0.287676	
15	CsPbI3	CsPbI3	[Br-].[Br-].[Sn+2]	0.127879	
16	CsPbI3	CH3NH3PbI3	[Li+].[Cl-]	0.528662	
17	CsPbI3	CH3NH3PbI3	C[NH3+].[I-]	0.285946	
18	CsSnBrI2	(NH2)2CH5SnBr2I	C(=N)N.Cl	0.9163	
19	CsSnBrI2	(NH2)2CH5SnBr2I	CN.Cl	1.306513	
20	CsSnBrI2	(NH2)2CH5SnBr2I	[Br-].[Br-].[Pb+2]	0.465672	
21	CsSnBrI2	(NH2)2CH5SnBr2I	[Cl-].[Cl-].[Pb+2]	0.333333	
22	CsSnBrI2	CH3NH3SnClI2	[Na+].[Cl-]	0.985401	
23	CsSnBrI2	CH3NH3SnClI2	[Cl-].[Cl-].[Pb+2]	0.634021	
24	CsSnCl3	(NH2)2CH5SnClF2	[Cl-].[K+]	0.873926	
25	CsSnCl3	(NH2)2CH5SnClF2	[Li+].[Cl-]	0.536913	
26	CsSnClBr2	(NH2)2CH5SnCl2I	[F-].[F-].[F-].[Al+3]	0.628713	
27	CsSnClBr2	(NH2)2CH5SnCl2I	[Cl-].[Cl-].[Cu+2]	0.287739	
28	CsSnClBr2	(NH2)2CH5SnCl3	[Br-].[Br-].[Pb+2]	0.232203	
29	CsSnClBr2	(NH2)2CH5SnCl3	[I-].[I-].[Pb+2]	0.334286	
30	CsSnClF2	CH3NH3PbCl2I	[NH4+].[Cl-]	0.298056	
31	CsSnClF2	CH3NH3PbCl2I	CCCC[N+](CCCC)(CCCC)CCCC.[I-]	0.723404	
32	(NH2)2CHPbBr2I	CsPbI3	[Cl-].[Cl-].[Pb+2]	0.811321	
33	(NH2)2CHPbBr2I	CsPbI3	[Sn+2].[I-].[I-]	0.573566	
34	(NH2)2CHPbBr2I	CsSnBrI2	C(=N)N.Cl	1.015625	
35	(NH2)2CHPbBr2I	CsSnBrI2	C(=N)N.[I-]	0.605159	
36	(NH2)2CHPbCl2I	CsPbCl2I	CN.Br	0.272121	
37	(NH2)2CHPbCl2I	CsPbCl2I	[Sn+2].[I-].[I-]	0.344214	
38	(NH2)2CHPbI3	CH3NH3PbBr2I	[Br-].[Br-].[Sn+2]	0.219089	
39	(NH2)2CHPbI3	CH3NH3PbBr2I	[Sn+2].[I-].[I-]	0.291083	
40	(NH2)2CHPbI3	CH3NH3PbI3	CN.Br	0.91573	
41	(NH2)2CHPbI3	CH3NH3PbI3	CCCC[N+](CCCC)(CCCC)CCCC.[I-]	0.412811	
42	(NH2)2CHPbI3	CH3NH3SnCl2Br	C(=N)N.Cl	0.362069	
43	(NH2)2CHPbI3	CH3NH3SnCl2Br	C(=N)N.[I-]	0.794304	
44	(NH2)2CHSnCl3	CH3NH3SnClF2	[F-].[F-].[F-].[Al+3]	0.363465	

Featurize



选择起始原始raw数据集（只接受csv格式）

用户需要输入的csv格式：

无机：列名以Formula结尾
有机：列名以Smiles结尾

更准确：
化学式：列名以Formula结尾
Smiles码：列名以Smiles结尾

最右一列为目标数据
其它（左侧）为输入数据

拟合和预测交给软件！

Raw dataset

输入变量　　输出目标

	layer1_Formula	layer2_Formula	SMILES_Smiles	conservation rate
2	CsPbBr2I	CsPbBr3	[NH4+].[Cl-]	0.041727
3	CsPbBr2I	CsPbBr3	CCCC[N+](CCCC)(CCCC)CCCC.[I-]	0.326007
4	CsPbBr2I	CsPbCII2	[Br-].[Br-].[Sn+2]	0.257407
5	CsPbBr2I	CsPbCII2	[Sn+2].[I-].[I-]	0.721138
6	CsPbBr3	CH3NH3PbBr2I	[Cl-].[Ag+]	0.143376
7	CsPbBr3	CH3NH3PbBr2I	C[NH3+].[I-]	0.132046
8	CsPbCl2Br	CH3NH3SnBrI2	[Br-].[Ag+]	0.173264
9	CsPbCl2Br	CH3NH3SnBrI2	C(=N)N.[I-]	0.375907
10	CsPbCl2I	CsPbCl2Br	[Cl-].[K+]	0.188991
11	CsPbCl2I	CsPbCl2Br	[NH4+].[Cl-]	0.185522
12	CsPbI3	CsPbBr2I	[I-].[I-].[Pb+2]	0.137411
13	CsPbI3	CsPbBr2I	CCCC[N+](CCCC)(CCCC)CCCC.[I-]	0.506803
14	CsPbI3	CsPbI3	[Br-].[Br-].[Pb+2]	0.287676
15	CsPbI3	CsPbI3	[Br-].[Br-].[Sn+2]	0.127879
16	CsPbI3	CH3NH3PbI3	[Li+].[Cl-]	0.528662
17	CsPbI3	CH3NH3PbI3	C[NH3+].[I-]	0.285946
18	CsSnBrI2	(NH2)2CHSnBr2I	C(=N)N.Cl	0.9163
19	CsSnBrI2	(NH2)2CHSnBr2I	CN.Cl	1.306513
20	CsSnBrI2	(NH2)2CHSnBr2I	[Br-].[Br-].[Pb+2]	0.465672
21	CsSnBrI2	(NH2)2CHSnBr2I	[Cl-].[Cl-].[Pb+2]	0.333333
22	CsSnBrI2	CH3NH3SnCII2	[Na+].[Cl-]	0.985401
23	CsSnBrI2	CH3NH3SnCII2	[Cl-].[Cl-].[Pb+2]	0.634021
24	CsSnCl3	(NH2)2CHSnClF2	[Cl-].[K+]	0.873926
25	CsSnCl3	(NH2)2CHSnClF2	[Li+].[Cl-]	0.536913
26	CsSnClBr2	(NH2)2CHSnCl2I	[F-].[F-].[F-].[Al+3]	0.628713
27	CsSnClBr2	(NH2)2CHSnCl2I	[Cl-].[Cl-].[Cu+2]	0.287739
28	CsSnClBr2	(NH2)2CHSnCl3	[Br-].[Br-].[Pb+2]	0.232203
29	CsSnClBr2	(NH2)2CHSnCl3	[I-].[I-].[Pb+2]	0.334286
30	CsSnClF2	CH3NH3PbCl2I	[NH4+].[Cl-]	0.298056
31	CsSnClF2	CH3NH3PbCl2I	CCCC[N+](CCCC)(CCCC)CCCC.[I-]	0.723404
32	(NH2)2CHPbBr2I	CsPbI3	[Cl-].[Cl-].[Pb+2]	0.811321
33	(NH2)2CHPbBr2I	CsPbI3	[Sn+2].[I-].[I-]	0.573566
34	(NH2)2CHPbBr2I	CsSnBrI2	C(=N)N.Cl	1.015625
35	(NH2)2CHPbBr2I	CsSnBrI2	C(=N)N.[I-]	0.605159
36	(NH2)2CHPbCl2I	CsPbCl2I	CN.Br	0.272121
37	(NH2)2CHPbCl2I	CsPbCl2I	[Sn+2].[I-].[I-]	0.344214
38	(NH2)2CHPbI3	CH3NH3PbBr2I	[Br-].[Br-].[Sn+2]	0.219089
39	(NH2)2CHPbI3	CH3NH3PbBr2I	[Sn+2].[I-].[I-]	0.291083
40	(NH2)2CHPbI3	CH3NH3PbI3	CN.Br	0.91573
41	(NH2)2CHPbI3	CH3NH3PbI3	CCCC[N+](CCCC)(CCCC)CCCC.[I-]	0.412811
42	(NH2)2CHPbI3	CH3NH3SnCl2Br	C(=N)N.Cl	0.362069
43	(NH2)2CHPbI3	CH3NH3SnCl2Br	C(=N)N.[I-]	0.794304
44	(NH2)2CHSnCl3	CH3NH3SnClF2	[F-].[F-].[F-].[Al+3]	0.363465

需要特征化的变量，例如：
第一列：化学式
第二列：化学式
第三列：Smiles码
....

需要注意：
化学式列需要以Formula结尾
Smiles码列的列名需要以Smiles结尾

第四/五等等列：时间、温度、浓度等其它不需要再次特征化的实验变量

最右一列：输出目标（稳定性、效率、带隙等）

Formula	Formula	Smiles	T, t, C	Output
layer1_Formula	layer2_Formula	SMILES_Smiles		conservation rate
CsPbBr2I	CsPbBr3	[NH4+].[Cl-]		0.041727
CsPbBr2I	CsPbBr3	CCCC[N+](CCCC)(CCCC)CCCC.[I-]		0.326007
CsPbBr2I	CsPbCll2	Br-].[Br-].[Sn+2]		0.257407
CsPbBr2I	CsPbCll2	Sn+2].[I-].[I-]		0.721138
CsPbBr3	CH3NH3PbBr2I	Cl-].[Ag+]		0.143376
CsPbBr3	CH3NH3PbBr2I	C[NH3+].[I-]		0.132046
CsPbCl2Br	CH3NH3SnBrI2	Br-].[Ag+]		0.173264
CsPbCl2Br	CH3NH3SnBrI2	C(=N)N.[I-]		0.375907
CsPbCl2I	CsPbCl2Br	Cl-].[K+]		0.188991
CsPbCl2I	CsPbCl2Br	NH4+].[Cl-]		0.185522
CsPbI3	CsPbBr2I	I-].[I-].[Pb+2]		0.137411
CsPbI3	CsPbBr2I	CCCC[N+](CCCC)(CCCC)CCCC.[I-]		0.506803
CsPbI3	CsPbI3	Br-].[Br-].[Pb+2]		0.287676
CsPbI3	CsPbI3	Br-].[Br-].[Sn+2]		0.127879
CsPbI3	CH3NH3PbI3	Li+].[Cl-]		0.528662
CsPbI3	CH3NH3PbI3	C[NH3+].[I-]		0.285946
CsSnBrI2	(NH2)2CHSnBr2I	C(=N)N.Cl		0.9163
CsSnBrI2	(NH2)2CHSnBr2I	CN.Cl		1.306513
CsSnBrI2	(NH2)2CHSnBr2I	Br-].[Br-].[Pb+2]		0.465672
CsSnBrI2	(NH2)2CHSnBr2I	Cl-].[Cl-].[Pb+2]		0.332833
CsSnBrI2	(NH2)2CHSnCll2	N.[		0.9[illegible]
CsSnBrI2	CH3NH3SnCll2	Cl-].[Cl-].[Pb+2]		0.634021
CsSnCl3	CH3NH3SnClF2	C[NH3+].[K+]		0.87[illegible]
CsSnCl3	(NH2)2CHSnClF2	Li+].[Cl-]		0.536913
CsSnBrI2	(NH2)2CHSnCl2I	F-].[F-].[F-].[Al+3]		0.6[illegible]
CsSnClBr2	(NH2)2CHSnCl2I	Cl-].[Cl-].[Cu+2]		0.287739
CsSnClBr2	(NH2)2CHSnCl3	Br-].[Br-].[Pb+2]		0.2[illegible]
CsSnClBr2	(NH2)2CHSnCl3	I-].[I-].[Pb+2]		0.334286
CsSnClF2	CH3NH3PbCl2I	NH4+].[Cl-]		0.29[illegible]
CsSnClF2	CH3NH3PbCl2I	CCCC[N+](CCCC)(CCCC)CCCC.[I-]		0.723404
(NH2)2CHPbBr2	CsPbI3	Cl-].[Cl-].[Pb+2]		0.8[illegible]
(NH2)2CHPbBr2	CsPbI3	Sn+2].[I-].[I-]		0.573566
(NH2)2CHPbBr2	CsSnBrI2	C(=N)N.Cl		1.01[illegible]
(NH2)2CHPbBr2	CsSnBrI2	C(=N)N.[I-]		0.605159
(NH2)2CHPbCl2	CsPbCl2I	CN.Br		0.272121
(NH2)2CHPbCl2	CsPbCl2I	Sn+2].[I-].[I-]		0.344214
(NH2)2CHPbI3	CH3NH3PbBr2I	Br-].[Br-].[Sn+2]		0.219089
(NH2)2CHPbI3	CH3NH3PbBr2I	Sn+2].[I-].[I-]		0.291083
(NH2)2CHPbI3	CH3NH3PbI3	CN.Br		0.91573
(NH2)2CHPbI3	CH3NH3PbI3	CCCC[N+](CCCC)(CCCC)CCCC.[I-]		0.412811
(NH2)2CHPbI3	CH3NH3SnCl2Br	C(=N)N.Cl		0.362069
(NH2)2CHPbI3	CH3NH3SnCl2Br	C(=N)N.[I-]		0.794304
(NH2)2CHSnCl3	CH3NH3SnClF2	F-].[F-].[F-].[Al+3]		0.363465

变量1
需要
特征化

变量2
需要
特征化

变量3
需要
特征化

其它
变量：
时间
温度
浓度
动作
等等

输出
目标：
稳定性
效率
带隙
性能
等等

时间、温度、浓度

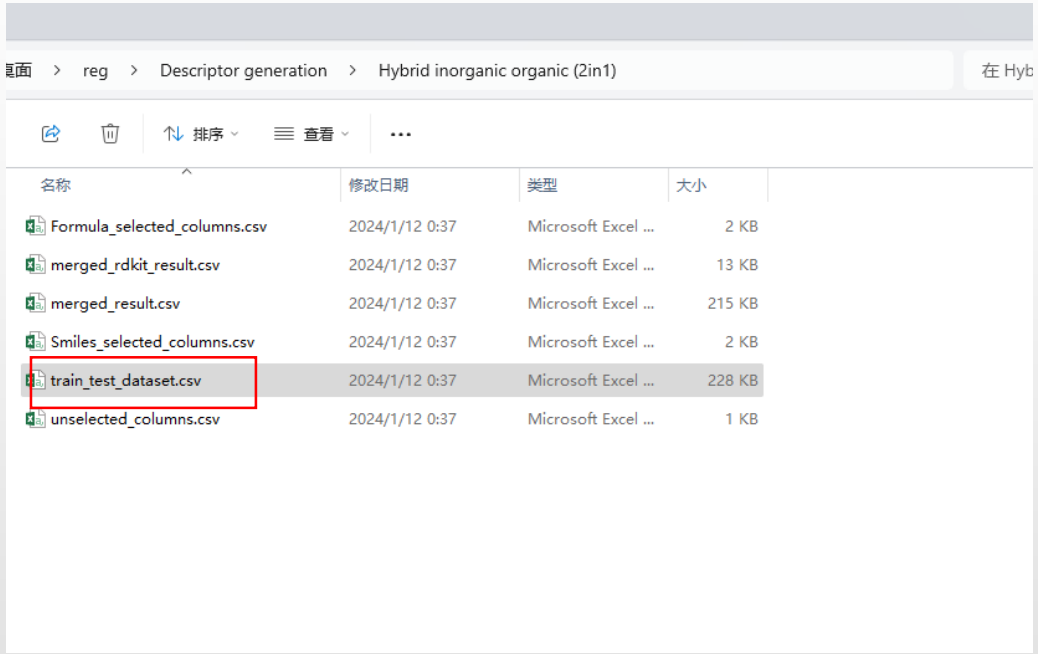
Layer 3 Smiles

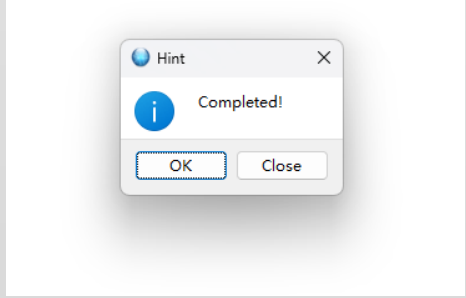
Layer 2 Formula

Layer 1 Formula

NJmatML: 有机-无机复杂材料特征化（2in1）

2in1特征化好的训练测试集train/test dataset保存在：
自定义文件夹→Descriptor generation →Hybrid inorganic organic (2in1) →train_test_dataset.csv





打开**train_test_dataset.csv**
训练测试集
但是需要后面进行特征选择！！！！

	1	2	3	4	5
1	layer1_Formula	layer2_Formula	SMILES_Smiles	conservation rate	
2	CsPbBr2I	CsPbBr3	[NH4+].[Cl-]	0.041727	
3	CsPbBr2I	CsPbBr3	CCCC[N+](CCCC)(CCCC)CCCC.[I-]	0.326007	
4	CsPbBr2I	CsPbCl2I2	[Br-].[Br-].[Sn+2]	0.257407	
5	CsPbBr2I	CsPbCl2I2	[Sn+2].[I-].[I-]	0.721138	
6	CsPbBr3	CH3NH3PbBr2I	[Cl-].[Ag+]	0.143376	
7	CsPbBr3	CH3NH3PbBr2I	C[NH3+].[I-]	0.132046	
8	CsPbCl2Br	CH3NH3SnBrI2	[Br-].[Ag+]	0.173264	
9	CsPbCl2Br	CH3NH3SnBrI2	C(=N)N.[I-]	0.375907	
10	CsPbCl2I	CsPbCl2Br	[Cl-].[K+]	0.188991	
11	CsPbCl2I	CsPbCl2Br	[NH4+].[Cl-]	0.185522	
12	CsPbI3	CsPbBr2I	[I-].[I-].[Pb+2]	0.137411	
13	CsPbI3	CsPbBr2I	CCCC[N+](CCCC)(CCCC)CCCC.[I-]	0.506803	
14	CsPbI3	CsPbI3	[Br-].[Br-].[Pb+2]	0.287676	
15	CsPbI3	CsPbI3	[Br-].[Br-].[Sn+2]	0.127879	
16	CsPbI3	CH3NH3PbI3	[Li+].[Cl-]	0.528662	
17	CsPbI3	CH3NH3PbI3	C[NH3+].[I-]	0.285946	
18	CsSnBrI2	(NH2)2CHSnBrI2I	C(=N)N.Cl	0.9163	
19	CsSnBrI2	(NH2)2CHSnBrI2I	CN.Cl	1.306513	
20	CsSnBrI2	(NH2)2CHSnBrI2I	[Br-].[Br-].[Pb+2]	0.465672	
21	CsSnBrI2	(NH2)2CHSnBrI2I	[Cl-].[Cl-].[Pb+2]	0.333333	
22	CsSnBrI2	CH3NH3SnCl2I2	[Na+].[Cl-]	0.985401	
23	CsSnBrI2	CH3NH3SnCl2I2	[Cl-].[Cl-].[Pb+2]	0.634021	
24	CsSnCl3	(NH2)2CHSnClF2	[Cl-].[K+]	0.873926	
25	CsSnCl3	(NH2)2CHSnClF2	[Li+].[Cl-]	0.536913	
26	CsSnClBr2	(NH2)2CHSnCl2I	[F-].[F-].[F-].[Al+3]	0.628713	
27	CsSnClBr2	(NH2)2CHSnCl2I	[Cl-].[Cl-].[Cu+2]	0.287739	
28	CsSnClBr2	(NH2)2CHSnCl3	[Br-].[Br-].[Pb+2]	0.232203	
29	CsSnClBr2	(NH2)2CHSnCl3	[I-].[I-].[Pb+2]	0.334286	
30	CsSnClF2	CH3NH3PbCl2I	[NH4+].[Cl-]	0.298056	
31	CsSnClF2	CH3NH3PbCl2I	CCCC[N+](CCCC)(CCCC)CCCC.[I-]	0.723404	
32	(NH2)2CHPbBr2I	CsPbI3	[Cl-].[Cl-].[Pb+2]	0.811321	
33	(NH2)2CHPbBr2I	CsPbI3	[Sn+2].[I-].[I-]	0.573566	
34	(NH2)2CHPbBr2I	CsSnBrI2	C(=N)N.Cl	1.015625	
35	(NH2)2CHPbBr2I	CsSnBrI2	C(=N)N.[I-]	0.605159	
36	(NH2)2CHPbCl2I	CsPbCl2I	CN.Br	0.272121	
37	(NH2)2CHPbCl2I	CsPbCl2I	[Sn+2].[I-].[I-]	0.344214	
38	(NH2)2CHPbI3	CH3NH3PbBr2I	[Br-].[Br-].[Sn+2]	0.219089	
39	(NH2)2CHPbI3	CH3NH3PbBr2I	[Sn+2].[I-].[I-]	0.291083	
40	(NH2)2CHPbI3	CH3NH3PbI3	CN.Br	0.91573	
41	(NH2)2CHPbI3	CH3NH3PbI3	CCCC[N+](CCCC)(CCCC)CCCC.[I-]	0.412811	
42	(NH2)2CHPbI3	CH3NH3SnCl2Br	C(=N)N.Cl	0.362069	
43	(NH2)2CHPbI3	CH3NH3SnCl2Br	C(=N)N.[I-]	0.794304	
44	(NH2)2CHSnCl3	CH3NH3SnClF2	[F-].[F-].[F-].[Al+3]	0.363465	

化学式：**Matminer magpie描述符**
Smiles码：RDKit描述符

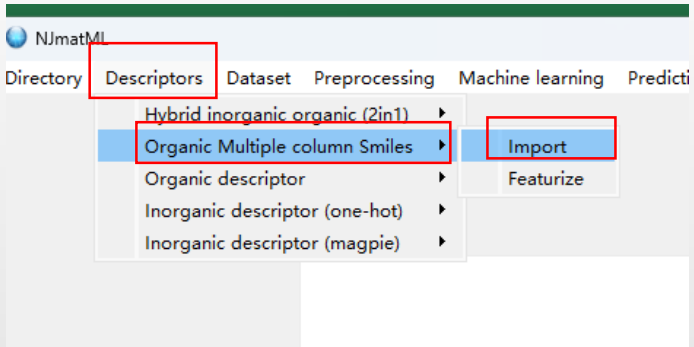
自动特征化

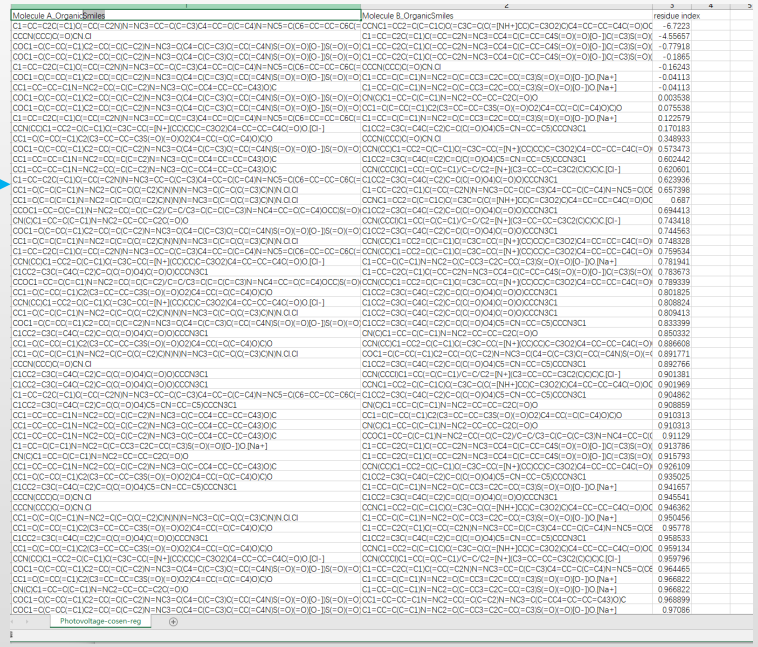
	_formula_1	_formula_1	_formula_1	_formula_1	_formula_1	_formula_1	_formula_1	_formula_1	_formula_1	_formula_1	_formula_1	_formula_1	_formula_1	_formula_1	_formula_1	_formula_1	_formula_1	_formula_1	_formula_1	_formula_1	_formula_1	_formula_1	fo
2	-0.2679	-0.14183	0.126073	52	35	82	47	13.6	35	74.4	5	96	91	27.76	95	125.3636	79.904	207.2	127.296	3
3	-0.2679	-0.14183	0.126073	52	35	82	47	13.6	35	74.4	5	96	91	27.76	95	125.3636	79.904	207.2	127.296	3
4	-0.2679	-0.14183	0.126073	52	35	82	47	13.6	35	74.4	5	96	91	27.76	95	125.3636	79.904	207.2	127.296	3
5	-0.2679	-0.14183	0.126073	52	35	82	47	13.6	35	74.4	5	96	91	27.76	95	125.3636	79.904	207.2	127.296	3
6	-0.29533	-0.14183	0.153503	48.4	35	82	47	16.08	35	74.2	5	95	90	27.68	95	115.9635	79.904	207.2	127.296	4
7	-0.29533	-0.14183	0.153503	48.4	35	82	47	16.08	35	74.2	5	95	90	27.68	95	115.9635	79.904	207.2	127.296	4
8	-0.29533	-0.14183	0.153503	41.2	17	82	65	21.84	17	73.8	5	95	90	27.52	94	98.18309	35.453	207.2	171.747	5
9	-0.29533	-0.14183	0.153503	41.2	17	82	65	21.84	17	73.8	5	95	90	27.52	94	98.18309	35.453	207.2	171.747	5
10	-0.2679	-0.14183	0.126073	44.8	17	82	65	22.24	17	74	5	96	91	27.6	94	107.5832	35.453	207.2	171.747	5
11	-0.2679	-0.14183	0.126073	44.8	17	82	65	22.24	17	74	5	96	91	27.6	94	107.5832	35.453	207.2	171.747	5
12	-0.2679	-0.14183	0.126073	59.2	53	82	29	9.12	53	74.8	5	96	91	27.92	96	144.1638	126.9045	207.2	80.29553	2
13	-0.2679	-0.14183	0.126073	59.2	53	82	29	9.12	53	74.8	5	96	91	27.92	96	144.1638	126.9045	207.2	80.29553	2
14	-0.2679	-0.14183	0.126073	59.2	53	82	29	9.12	53	74.8	5	96	91	27.92	96	144.1638	126.9045	207.2	80.29553	2
15	-0.2679	-0.14183	0.126073	59.2	53	82	29	9.12	53	74.8	5	96	91	27.92	96	144.1638	126.9045	207.2	80.29553	2
16	-0.2679	-0.14183	0.126073	59.2	53	82	29	9.12	53	74.8	5	96	91	27.92	96	144.1638	126.9045	207.2	80.29553	2
17	-0.2679	-0.14183	0.126073	59.2	53	82	29	9.12	53	74.8	5	96	91	27.92	96	144.1638	126.9045	207.2	80.29553	2
18	-0.2679	-0.14445	0.123454	49.2	35	55	20	5.68	53	74.4	5	96	91	27.76	96	117.0657	79.904	132.9055	53.00145	1
19	-0.2679	-0.14445	0.123454	49.2	35	55	20	5.68	53	74.4	5	96	91	27.76	96	117.0657	79.904	132.9055	53.00145	1
20	-0.2679	-0.14445	0.123454	49.2	35	55	20	5.68	53	74.4	5	96	91	27.76	96	117.0657	79.904	132.9055	53.00145	1
21	-0.2679	-0.14445	0.123454	49.2	35	55	20	5.68	53	74.4	5	96	91	27.76	96	117.0657	79.904	132.9055	53.00145	1
22	-0.2679	-0.14445	0.123454	49.2	35	55	20	5.68	53	74.4	5	96	91	27.76	96	117.0657	79.904	132.9055	53.00145	1
23	-0.2679	-0.14445	0.123454	49.2	35	55	20	5.68	53	74.4	5	96	91	27.76	96	117.0657	79.904	132.9055	53.00145	1
24	-0.32038	-0.14445	0.17593	31.2	17	55	38	17.04	17	73.4	5	94	89	27.36	94	71.59489	35.453	132.9055	97.45245	4
25	-0.32038	-0.14445	0.17593	31.2	17	55	38	17.04	17	73.4	5	94	89	27.36	94	71.59489	35.453	132.9055	97.45245	4
26	-0.29533	-0.14445	0.150884	38.4	17	55	38	11.28	35	73.8	5	95	90	27.52	95	89.37529	35.453	132.9055	97.45245	2
27	-0.29533	-0.14445	0.150884	38.4	17	55	38	11.28	35	73.8	5	95	90	27.52	95	89.37529	35.453	132.9055	97.45245	2
28	-0.29533	-0.14445	0.150884	38.4	17	55	38	11.28	35	73.8	5	95	90	27.52	95	89.37529	35.453	132.9055	97.45245	2
29	-0.29533	-0.14445	0.150884	38.4	17	55	38	11.28	35	73.8	5	95	90	27.52	95	89.37529	35.453	132.9055	97.45245	2
30	-0.32038	-0.14445	0.17593	28	9	55	46	19.6	9	73	5	94	89	27.2	93	65.01305	18.9984	132.9055	113.907	4
31	-0.32038	-0.14445	0.17593	28	9	55	46	19.6	9	73	5	94	89	27.2	93	65.01305	18.9984	132.9055	113.907	4
32	-0.2663	-0.23347	0.032826	19.16667	1	82	81	21.38889	1	89	77	96	19	5.666667	92	44.91469	1.00794	207.2	206.1921	5
33	-0.2663	-0.23347	0.032826	19.16667	1	82	81	21.38889	1	89	77	96	19	5.666667	92	44.91469	1.00794	207.2	206.1921	5
34	-0.2663	-0.23347	0.032826	19.16667	1	82	81	21.38889	1	89	77	96	19	5.666667	92	44.91469	1.00794	207.2	206.1921	5

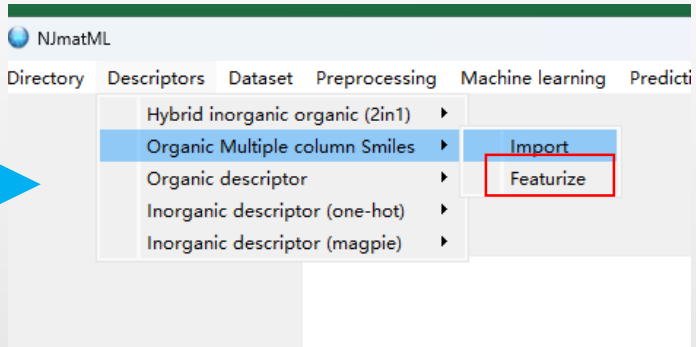
2.2 材料特征化按钮：只有有机部分

特征化按钮（训练测试集辅助生成） 多种有机分子:例如双分子SMILES码 → RDKit描述符

Descriptors: Organic Multiple column Smiles → Import → Featurize







用户需要自己准备的
数据集长啥样?

可以是已经特征化好的数字csv

也可以是最原始的Smiles码或者化学式

如果Smiles码，
请将列名后缀标为Smiles

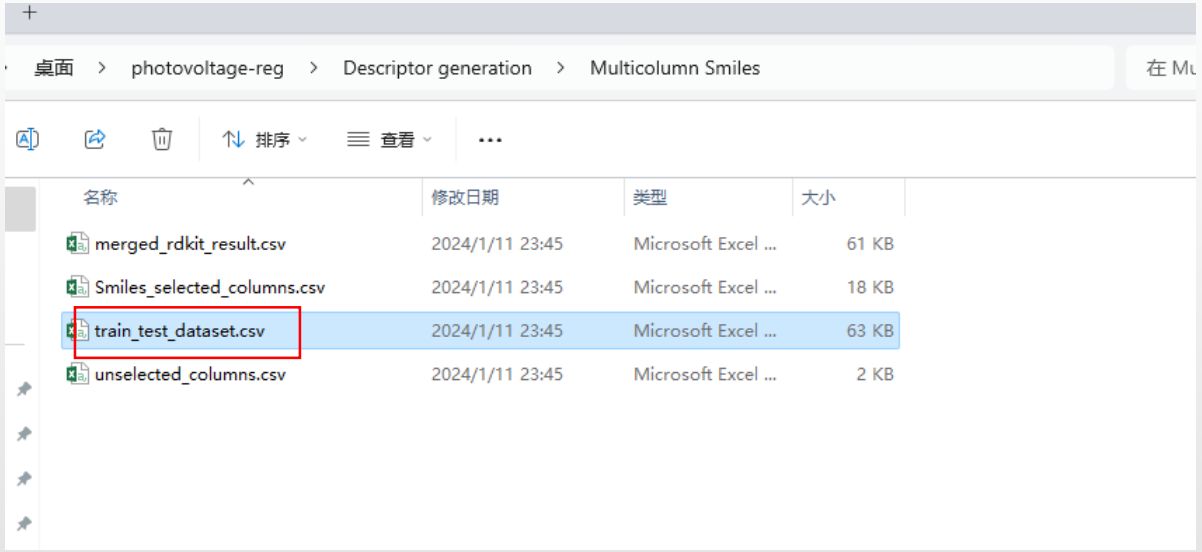
Molecule A_OrganicSmiles	Molecule B_OrganicSmiles	residue index
C1=CC=C2C(=C1)C(=CC(=C2N)N=NC3=CC=C(C=C3)C4=CC=C(C=C4)N=NC5=C(C6=CC=CC=C6C	CCNC1=CC2=C(C=C1C)C(=C3C=C(C(=[NH+]CC)C=C3O2)C)C4=CC=CC=C4C(=O)C	-6.7223
CCCN(CCC)C(=O)CN.Cl	C1=CC=C2C(=C1)C(=CC=C2N=NC3=CC4=C(C=CC=C4S(=O)(=O)[O-])C(=C3)S(=C	-4.55657
COC1=C(C=CC(=C1)C2=CC(=C(C=C2)N=NC3=C(C4=C(C=C3)C(=CC(=C4N)S(=O)(=O)[O-])S(=O)=O	C1=CC=C2C(=C1)C(=CC=C2N=NC3=CC4=C(C=CC=C4S(=O)(=O)[O-])C(=C3)S(=C	-0.77918
COC1=C(C=CC(=C1)C2=CC(=C(C=C2)N=NC3=C(C4=C(C=C3)C(=CC(=C4N)S(=O)(=O)[O-])S(=O)=O	C1=CC=C2C(=C1)C(=CC=C2N=NC3=CC4=C(C=CC=C4S(=O)(=O)[O-])C(=C3)S(=C	-0.1865
C1=CC=C2C(=C1)C(=CC(=C2N)N=NC3=CC=C(C=C3)C4=CC=C(C=C4)N=NC5=C(C6=CC=CC=C6C	CCCN(CCC)C(=O)CN.Cl	-0.16243
COC1=C(C=CC(=C1)C2=CC(=C(C=C2)N=NC3=C(C4=C(C=C3)C(=CC(=C4N)S(=O)(=O)[O-])S(=O)=O	C1=CC=C(C=C1)N=NC2=C(C=CC3=C2C=CC(=C3)S(=O)(=O)[O-])O.[Na+]	-0.04113
CC1=CC=CC=C1N=NC2=CC(=C(C=C2)N=NC3=C(C=CC4=CC=CC=C43)O)C	C1=CC=C(C=C1)N=NC2=C(C=CC3=C2C=CC(=C3)S(=O)(=O)[O-])O.[Na+]	-0.04113
COC1=C(C=CC(=C1)C2=CC(=C(C=C2)N=NC3=C(C4=C(C=C3)C(=CC(=C4N)S(=O)(=O)[O-])S(=O)=O	CN(C)C1=CC=C(C=C1)N=NC2=CC=CC=C2C(=O)O	0.003538
COC1=C(C=CC(=C1)C2=CC(=C(C=C2)N=NC3=C(C4=C(C=C3)C(=CC(=C4N)S(=O)(=O)[O-])S(=O)=O	CC1=C(C=CC(=C1)C2(C3=CC=CC=C3S(=O)(=O)O2)C4=CC(=C(C=C4)O)C)O	0.075538
C1=CC=C2C(=C1)C(=CC(=C2N)N=NC3=CC=C(C=C3)C4=CC=C(C=C4)N=NC5=C(C6=CC=CC=C6C	C1=CC=C(C=C1)N=NC2=C(C=CC3=C2C=CC(=C3)S(=O)(=O)[O-])O.[Na+]	0.122579
CCN(CC)C1=CC2=C(C=C1)C(=C3C=CC(=[N+](CC)CC)C=C3O2)C4=CC=CC=C4C(=O)O.[Cl-]	C1CC2=C3C(=C4C(=C2)C=C(C(=O)O4)C5=CN=CC=C5)CCCN3C1	0.170183
CC1=C(C=CC(=C1)C2(C3=CC=CC=C3S(=O)(=O)O2)C4=CC(=C(C=C4)O)C)O	CCCN(CCC)C(=O)CN.Cl	0.348933
COC1=C(C=CC(=C1)C2=CC(=C(C=C2)N=NC3=C(C4=C(C=C3)C(=CC(=C4N)S(=O)(=O)[O-])S(=O)=O	CCN(CC)C1=CC2=C(C=C1)C(=C3C=CC(=[N+](CC)CC)C=C3O2)C4=CC=CC=C4C(=O	0.573473
CC1=CC=CC=C1N=NC2=CC(=C(C=C2)N=NC3=C(C=CC4=CC=CC=C43)O)C	C1CC2=C3C(=C4C(=C2)C=C(C(=O)O4)C5=CN=CC=C5)CCCN3C1	0.602442
CC1=CC=CC=C1N=NC2=CC(=C(C=C2)N=NC3=C(C=CC4=CC=CC=C43)O)C	CCN(CCCI)C1=CC(=C(C=C1)/C=C/C2=[N+](C3=CC=CC=C3C2(C)C)C)C.[Cl-]	0.620601
C1=CC=C2C(=C1)C(=CC(=C2N)N=NC3=CC=C(C=C3)C4=CC=C(C=C4)N=NC5=C(C6=CC=CC=C6C	C1CC2=C3C(=C4C(=C2)C=C(C(=O)O4)C(=O)O)CCCN3C1	0.623936
CC1=C(C=C(C=C1)N=NC2=C(C=C(C(=C2)C)N)N)N=NC3=C(C=C(C(=C3)C)N)N.Cl.Cl	C1=CC=C2C(=C1)C(=CC(=C2N)N=NC3=CC=C(C=C3)C4=CC=C(C=C4)N=NC5=C(	0.657398
CC1=C(C=C(C=C1)N=NC2=C(C=C(C(=C2)C)N)N)N=NC3=C(C=C(C(=C3)C)N)N.Cl.Cl	CCNC1=CC2=C(C=C1C)C(=C3C=C(C(=[NH+]CC)C=C3O2)C)C4=CC=CC=C4C(=O)	0.687
CCOC1=CC=C(C=C1)N=NC2=CC(=C(C=C2)/C=C/C3=C(C=C(C=C3)N=NC4=CC=C(C=C4)OCC)S(	C1CC2=C3C(=C4C(=C2)C=C(C(=O)O4)C(=O)O)CCCN3C1	0.694413
CN(C)C1=CC=C(C=C1)N=NC2=CC=CC=C2C(=O)O	CCN(CCCI)C1=CC(=C(C=C1)/C=C/C2=[N+](C3=CC=CC=C3C2(C)C)C)C.[Cl-]	0.743418
COC1=C(C=CC(=C1)C2=CC(=C(C=C2)N=NC3=C(C4=C(C=C3)C(=CC(=C4N)S(=O)(=O)[O-])S(=O)=O	C1CC2=C3C(=C4C(=C2)C=C(C(=O)O4)C(=O)O)CCCN3C1	0.744563
CC1=C(C=C(C=C1)N=NC2=C(C=C(C(=C2)C)N)N)N=NC3=C(C=C(C(=C3)C)N)N.Cl.Cl	CCN(CC)C1=CC2=C(C=C1)C(=C3C=CC(=[N+](CC)CC)C=C3O2)C4=CC=CC=C4C(=	0.748328
C1=CC=C2C(=C1)C(=CC(=C2N)N=NC3=CC=C(C=C3)C4=CC=C(C=C4)N=NC5=C(C6=CC=CC=C6C	CCN(CC)C1=CC2=C(C=C1)C(=C3C=CC(=[N+](CC)CC)C=C3O2)C4=CC=CC=C4C(=	0.759534
CCN(CC)C1=CC2=C(C=C1)C(=C3C=CC(=[N+](CC)CC)C=C3O2)C4=CC=CC=C4C(=O)O.[Cl-]	C1=CC=C(C=C1)N=NC2=C(C=CC3=C2C=CC(=C3)S(=O)(=O)[O-])O.[Na+]	0.781941
C1CC2=C3C(=C4C(=C2)C=C(C(=O)O4)C(=O)O)CCCN3C1	C1=CC=C2C(=C1)C(=CC=C2N=NC3=CC4=C(C=CC=C4S(=O)(=O)[O-])C(=C3)S(=C	0.783673
CCOC1=CC=C(C=C1)N=NC2=CC(=C(C=C2)/C=C/C3=C(C=C(C=C3)N=NC4=CC=C(C=C4)OCC)S(	CCN(CC)C1=CC2=C(C=C1)C(=C3C=CC(=[N+](CC)CC)C=C3O2)C4=CC=CC=C4C(=	0.789339
CC1=C(C=CC(=C1)C2(C3=CC=CC=C3S(=O)(=O)O2)C4=CC(=C(C=C4)O)C)O	C1CC2=C3C(=C4C(=C2)C=C(C(=O)O4)C(=O)O)CCCN3C1	[illegible]
CCN(CC)C1=CC2=C(C=C1)C(=C3C=CC(=[N+](CC)CC)C=C3O2)C4=CC=CC=C4C(=O)O.[Cl-]	C1CC2=C3C(=C4C(=C2)C=C(C(=O)O4)C(=O)O)CCCN3C1	[illegible]
CC1=C(C=C(C=C1)N=NC2=C(C=C(C(=C2)C)N)N)N=NC3=C(C=C(C(=C3)C)N)N.Cl.Cl	C1CC2=C3C(=C4C(=C2)C=C(C(=O)O4)C(=O)O)CCCN3C1	[illegible]
COC1=C(C=CC(=C1)C2=CC(=C(C=C2)N=NC3=C(C4=C(C=C3)C(=CC(=C4N)S(=O)(=O)[O-])S(=O)=O	C1CC2=C3C(=C4C(=C2)C=C(C(=O)O4)C5=CN=CC=C5)CCCN3C1	[illegible]
C1CC2=C3C(=C4C(=C2)C=C(C(=O)O4)C(=O)O)CCCN3C1	CN(C)C1=CC=C(C=C1)N=NC2=CC=CC=C2C(=O)O	0.850332
CC1=C(C=CC(=C1)C2(C3=CC=CC=C3S(=O)(=O)O2)C4=CC(=C(C=C4)O)C)O	CCN(CC)C1=CC2=C(C=C1)C(=C3C=CC(=[N+](CC)CC)C=C3O2)C4=CC=CC=C4C(=	[illegible]
CC1=C(C=C(C=C1)N=NC2=C(C=C(C(=C2)C)N)N)N=NC3=C(C=C(C(=C3)C)N)N.Cl.Cl	COC1=C(C=CC(=C1)C2=CC(=C(C=C2)N=NC3=C(C4=C(C=C3)C(=CC(=C4N)S(=O)(	0.891771
CCCN(CCC)C(=O)CN.Cl	C1CC2=C3C(=C4C(=C2)C=C(C(=O)O4)C5=CN=CC=C5)CCCN3C1	[illegible]
C1CC2=C3C(=C4C(=C2)C=C(C(=O)O4)C(=O)O)CCCN3C1	CCN(CCCI)C1=CC(=C(C=C1)/C=C/C2=[N+](C3=CC=CC=C3C2(C)C)C)C.[Cl-]	0.901381
C1CC2=C3C(=C4C(=C2)C=C(C(=O)O4)C(=O)O)CCCN3C1	CCNC1=CC2=C(C=C1C)C(=C3C=C(C(=[NH+]CC)C=C3O2)C)C4=CC=CC=C4C(=O)	[illegible]
C1=CC=C2C(=C1)C(=CC(=C2N)N=NC3=CC=C(C=C3)C4=CC=C(C=C4)N=NC5=C(C6=CC=CC=C6C	C1CC2=C3C(=C4C(=C2)C=C(C(=O)O4)C5=CN=CC=C5)CCCN3C1	0.904862
C1CC2=C3C(=C4C(=C2)C=C(C(=O)O4)C5=CN=CC=C5)CCCN3C1	CN(C)C1=CC=C(C=C1)N=NC2=CC=CC=C2C(=O)O	[illegible]
CC1=CC=CC=C1N=NC2=CC(=C(C=C2)N=NC3=C(C=CC4=CC=CC=C43)O)C	CC1=C(C=CC(=C1)C2(C3=CC=CC=C3S(=O)(=O)O2)C4=CC(=C(C=C4)O)C)O	0.910319
CC1=CC=CC=C1N=NC2=CC(=C(C=C2)N=NC3=C(C=CC4=CC=CC=C43)O)C	CN(C)C1=CC=C(C=C1)N=NC2=CC=CC=C2C(=O)O	0.910313
CC1=CC=CC=C1N=NC2=CC(=C(C=C2)N=NC3=C(C=CC4=CC=CC=C43)O)C	CCOC1=CC=C(C=C1)N=NC2=CC(=C(C=C2)/C=C/C3=C(C=C(C=C3)N=NC4=CC=(	0.91129
C1=CC=C(C=C1)N=NC2=C(C=CC3=C2C=CC(=C3)S(=O)(=O)[O-])O.[Na+]	C1=CC=C2C(=C1)C(=CC=C2N=NC3=CC4=C(C=CC=C4S(=O)(=O)[O-])C(=C3)S(=C	0.913786
CN(C)C1=CC=C(C=C1)N=NC2=CC=CC=C2C(=O)O	C1=CC=C2C(=C1)C(=CC=C2N=NC3=CC4=C(C=CC=C4S(=O)(=O)[O-])C(=C3)S(=C	0.915793
CC1=CC=CC=C1N=NC2=CC(=C(C=C2)N=NC3=C(C=CC4=CC=CC=C43)O)C	CCN(CC)C1=CC2=C(C=C1)C(=C3C=CC(=[N+](CC)CC)C=C3O2)C4=CC=CC=C4C(=	0.926109
CC1=C(C=CC(=C1)C2(C3=CC=CC=C3S(=O)(=O)O2)C4=CC(=C(C=C4)O)C)O	C1CC2=C3C(=C4C(=C2)C=C(C(=O)O4)C5=CN=CC=C5)CCCN3C1	0.935025
C1CC2=C3C(=C4C(=C2)C=C(C(=O)O4)C5=CN=CC=C5)CCCN3C1	C1=CC=C(C=C1)N=NC2=C(C=CC3=C2C=CC(=C3)S(=O)(=O)[O-])O.[Na+]	0.941657
CCCN(CCC)C(=O)CN.Cl	C1CC2=C3C(=C4C(=C2)C=C(C(=O)O4)C(=O)O)CCCN3C1	0.945541
CCCN(CCC)C(=O)CN.Cl	CCNC1=CC2=C(C=C1C)C(=C3C=C(C(=[NH+]CC)C=C3O2)C)C4=CC=CC=C4C(=O)	0.946362
CC1=C(C=C(C=C1)N=NC2=C(C=C(C(=C2)C)N)N)N=NC3=C(C=C(C(=C3)C)N)N.Cl.Cl	C1=CC=C(C=C1)N=NC2=C(C=CC3=C2C=CC(=C3)S(=O)(=O)[O-])O.[Na+]	0.950456
CC1=C(C=CC(=C1)C2(C3=CC=CC=C3S(=O)(=O)O2)C4=CC(=C(C=C4)O)C)O	C1=CC=C2C(=C1)C(=CC(=C2N)N=NC3=CC=C(C=C3)C4=CC=C(C=C4)N=NC5=C(	0.95778
C1CC2=C3C(=C4C(=C2)C=C(C(=O)O4)C(=O)O)CCCN3C1	C1CC2=C3C(=C4C(=C2)C=C(C(=O)O4)C5=CN=CC=C5)CCCN3C1	0.958533
CC1=C(C=CC(=C1)C2(C3=CC=CC=C3S(=O)(=O)O2)C4=CC(=C(C=C4)O)C)O	CCNC1=CC2=C(C=C1C)C(=C3C=C(C(=[NH+]CC)C=C3O2)C)C4=CC=CC=C4C(=O)	0.959134
CCN(CC)C1=CC2=C(C=C1)C(=C3C=CC(=[N+](CC)CC)C=C3O2)C4=CC=CC=C4C(=O)O.[Cl-]	CCN(CCCI)C1=CC(=C(C=C1)/C=C/C2=[N+](C3=CC=CC=C3C2(C)C)C)C.[Cl-]	0.959796
COC1=C(C=CC(=C1)C2=CC(=C(C=C2)N=NC3=C(C4=C(C=C3)C(=CC(=C4N)S(=O)(=O)[O-])S(=O)=O	C1=CC=C2C(=C1)C(=CC(=C2N)N=NC3=CC=C(C=C3)C4=CC=C(C=C4)N=NC5=C(	0.964465
CC1=C(C=CC(=C1)C2(C3=CC=CC=C3S(=O)(=O)O2)C4=CC(=C(C=C4)O)C)O	C1=CC=C(C=C1)N=NC2=C(C=CC3=C2C=CC(=C3)S(=O)(=O)[O-])O.[Na+]	0.966822
CN(C)C1=CC=C(C=C1)N=NC2=CC=CC=C2C(=O)O	C1=CC=C(C=C1)N=NC2=C(C=CC3=C2C=CC(=C3)S(=O)(=O)[O-])O.[Na+]	0.966822
COC1=C(C=CC(=C1)C2=CC(=C(C=C2)N=NC3=C(C4=C(C=C3)C(=CC(=C4N)S(=O)(=O)[O-])S(=O)=O	CC1=CC=CC=C1N=NC2=CC(=C(C=C2)N=NC3=C(C=CC4=CC=CC=C43)O)C	0.968899
COC1=C(C=CC(=C1)C2=CC(=C(C=C2)N=NC3=C(C4=C(C=C3)C(=CC(=C4N)S(=O)(=O)[O-])S(=O)=O	C1=CC=C(C=C1)N=NC2=C(C=CC3=C2C=CC(=C3)S(=O)(=O)[O-])O.[Na+]	0.97086

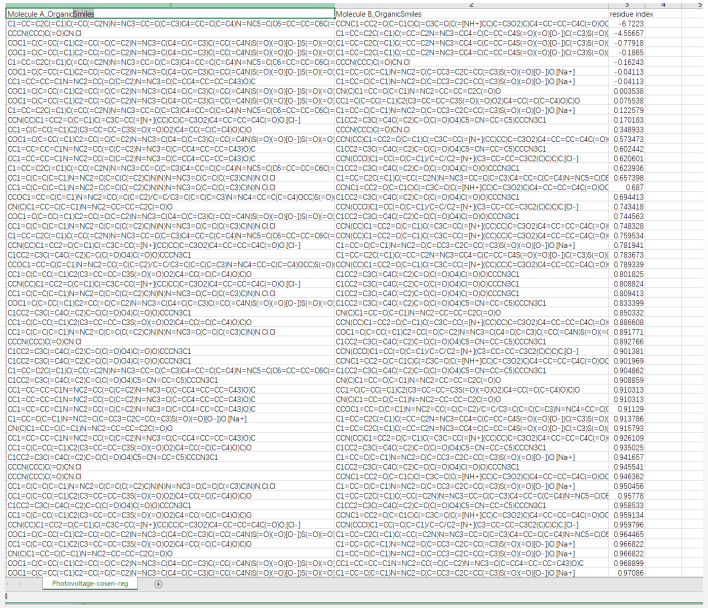
输入变量1（分子A）

输入变量2（分子B）

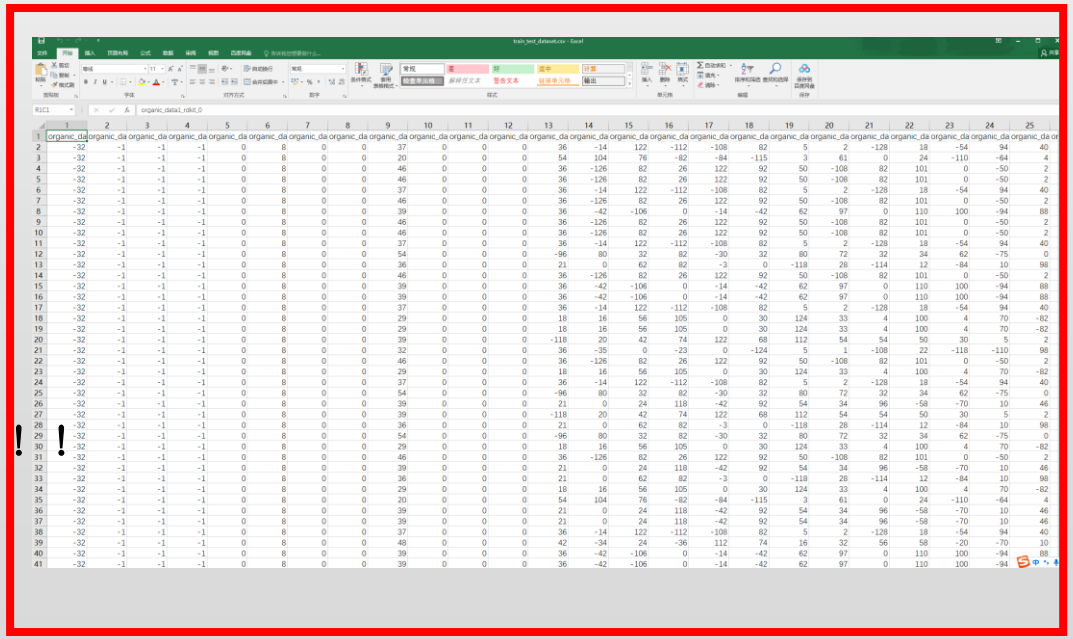
输出列：
稳定性
效率
带隙
性能
等等

NJmatML: 材料特征化（有机部分）





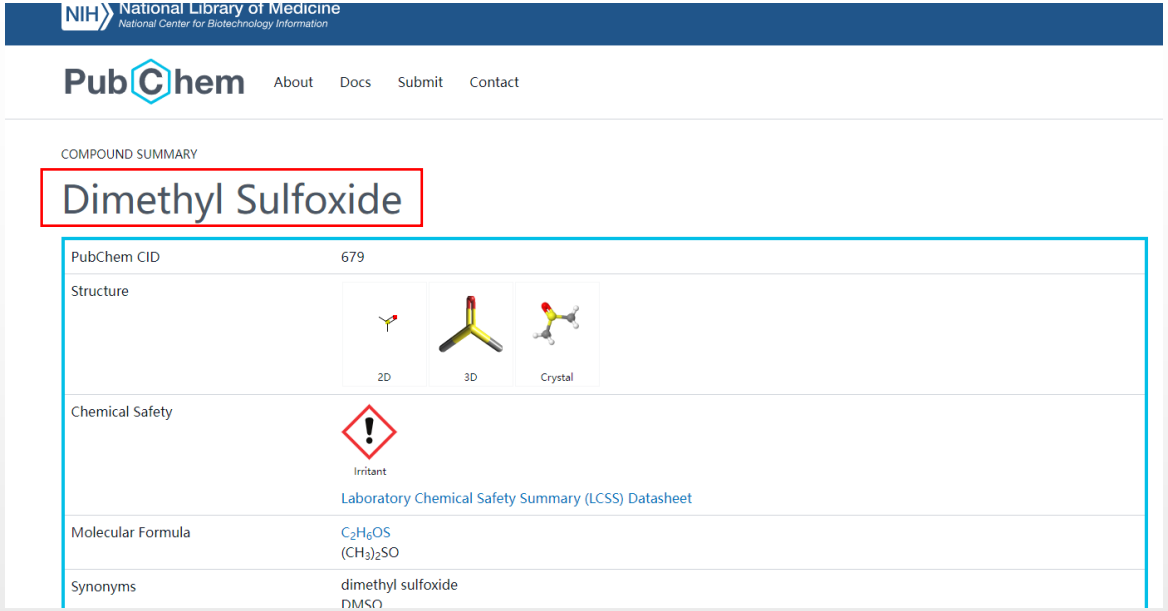
特征化Smiles码

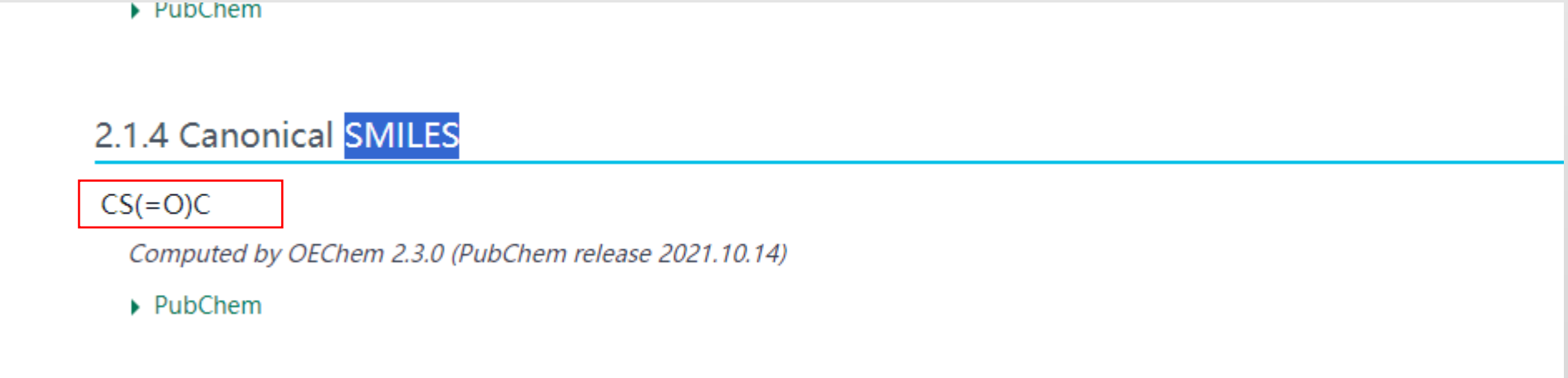


打开train_test_dataset.csv
训练测试集
但是需要后面进行特征选择！！！

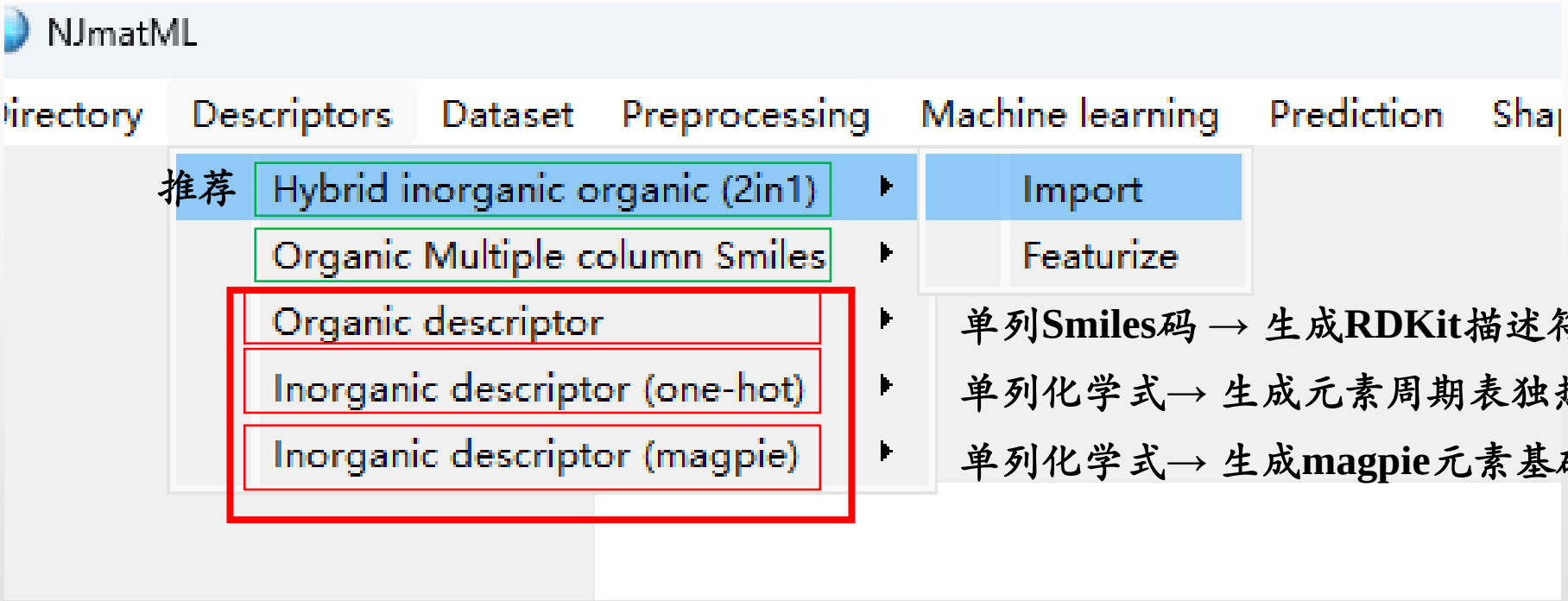
Smiles：可以从Pubchem等网站中查询

Pubchem





2.3其他特征化按钮



单列Smiles码 → 生成RDKit描述符

单列化学式→ 生成元素周期表独热编码

单列化学式→ 生成magpie元素基础描述符

2.3其他化学式特征化按钮：独热编码

	formula
0	(Fe2AgCu2)O3
1	Fe2O3
2	CsPbI3
3	MoS2
4	CuInGaSe
5	Si
6	TiO2

TiO_2:
O: 0.666666为O
Ti: 0.333333为Ti

H	He	Li	Be	B	C	N	O	F	Ne	Na	Mg	Al	Si	P	S	Cl	Ar	K	Ca
0	0	0	0	0	0	0	0.375	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0.6	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.666667	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	
0	0	0	0	0	0	0	0.666667	0	0	0	0	0	0	0	0	0	0	0	

	P	S	Cl	Ar	K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As
0	0	0	0	0	0	0	0	0	0	0	0	0.25	0	0	0.25	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0.4	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0.666667	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.25	0	0.25	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0.333333	0	0	0	0	0	0	0	0	0	0	0

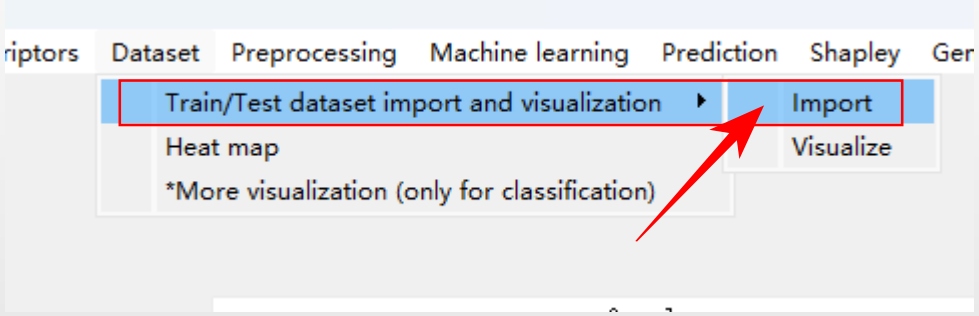
3 导入数据集 Dataset import

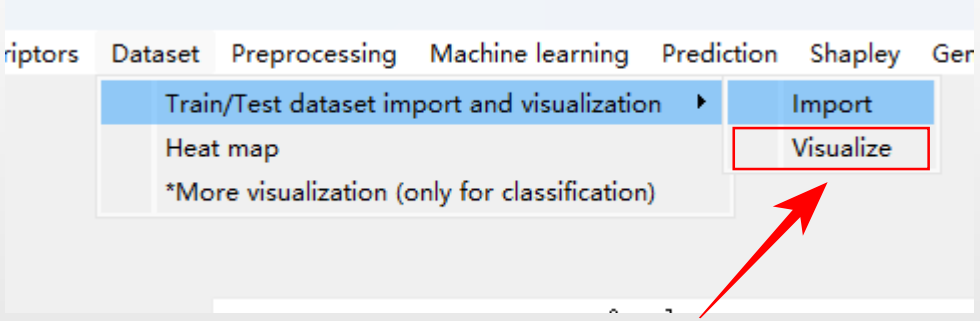
3.1导入数据集

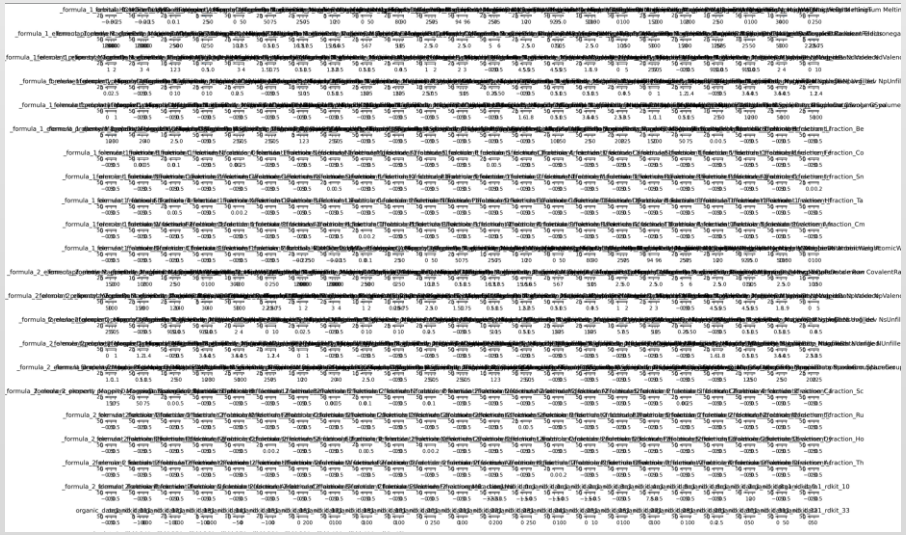
数据集为之前软件自动特征化的train_test_dataset.csv

例如：自定义文件夹→Descriptor generation →Hybrid inorganic organic (2in1) → train_test_dataset.csv

更好的方式：领域知识描述符（customized)

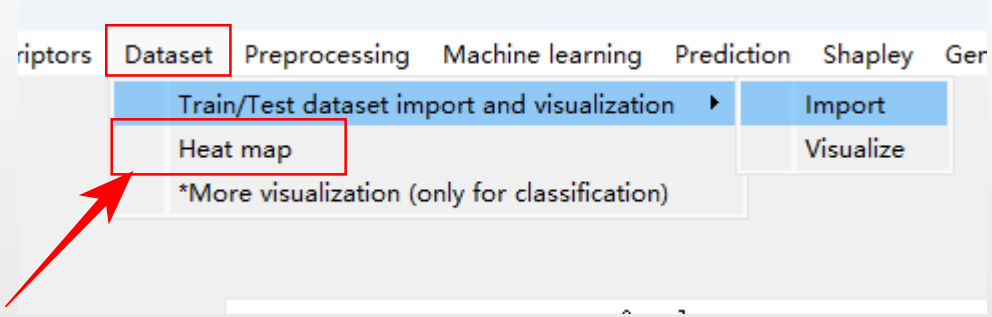


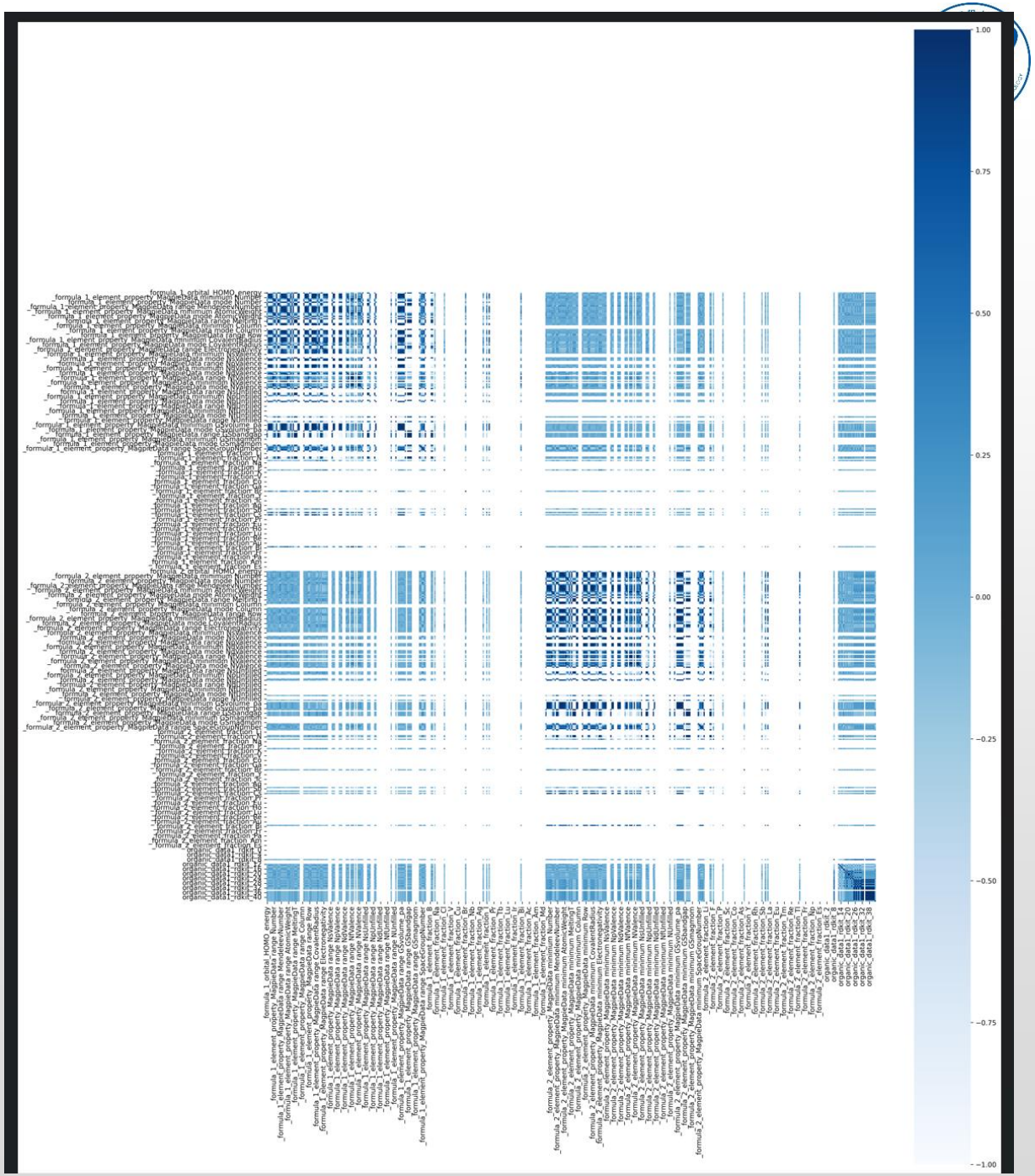




特征选择前的数据可视化

3.2导入数据集后的热图



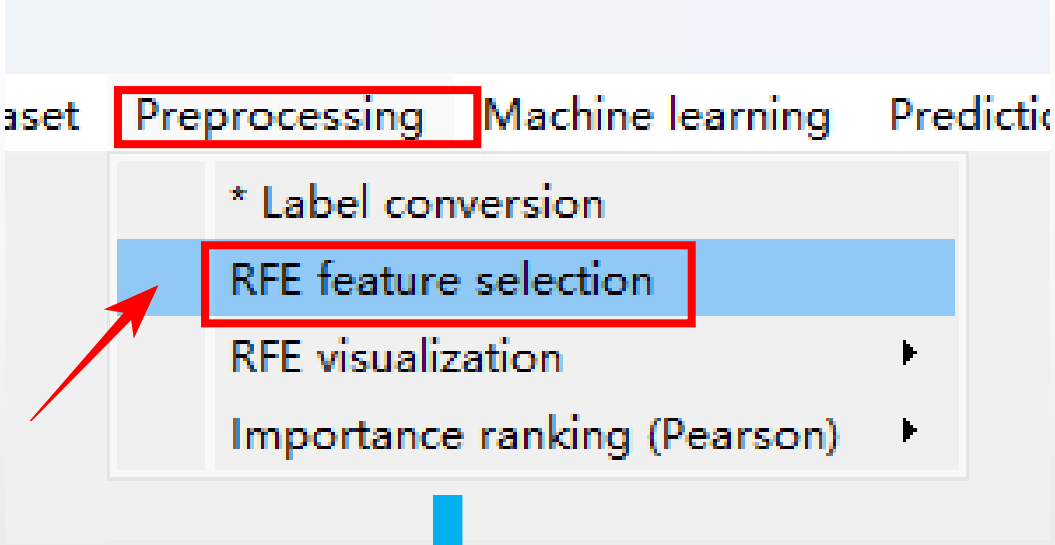


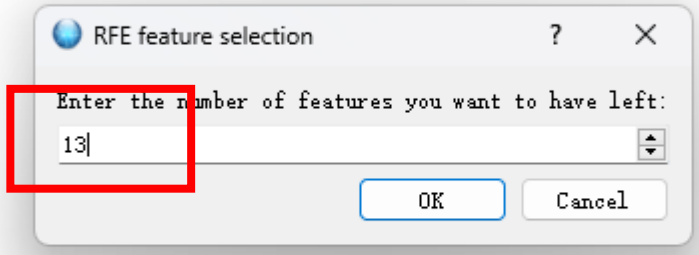
Heat Map

4 Preprocessing

4.1 Preprocessing: RFE特征选择

特征选择：RFE
Recursive feature elimination





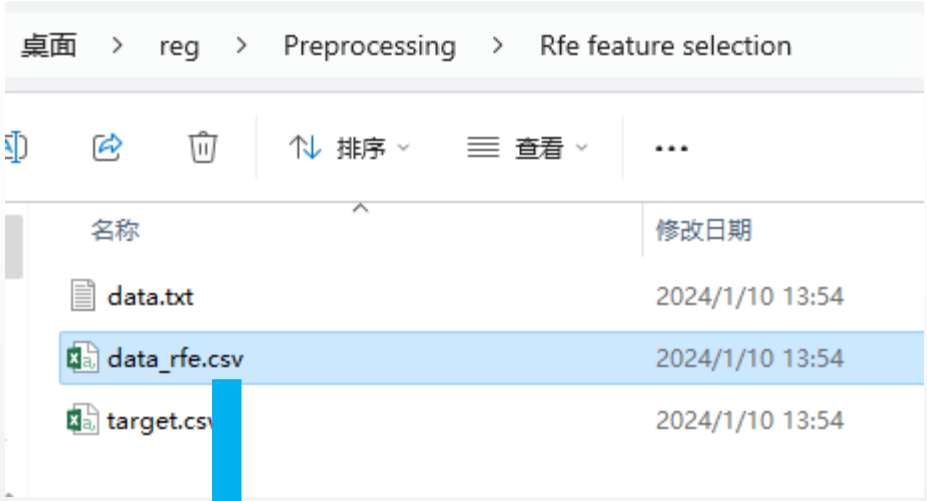
输入想要保留的特征数目
例如 13

4.1 Preprocessing

特征选择后的数据集保存在data_rfe.csv：

自定义folder → Proprocessing → Rfe feature selection → data_rfe.csv

该数据集实质用于机器学习建模

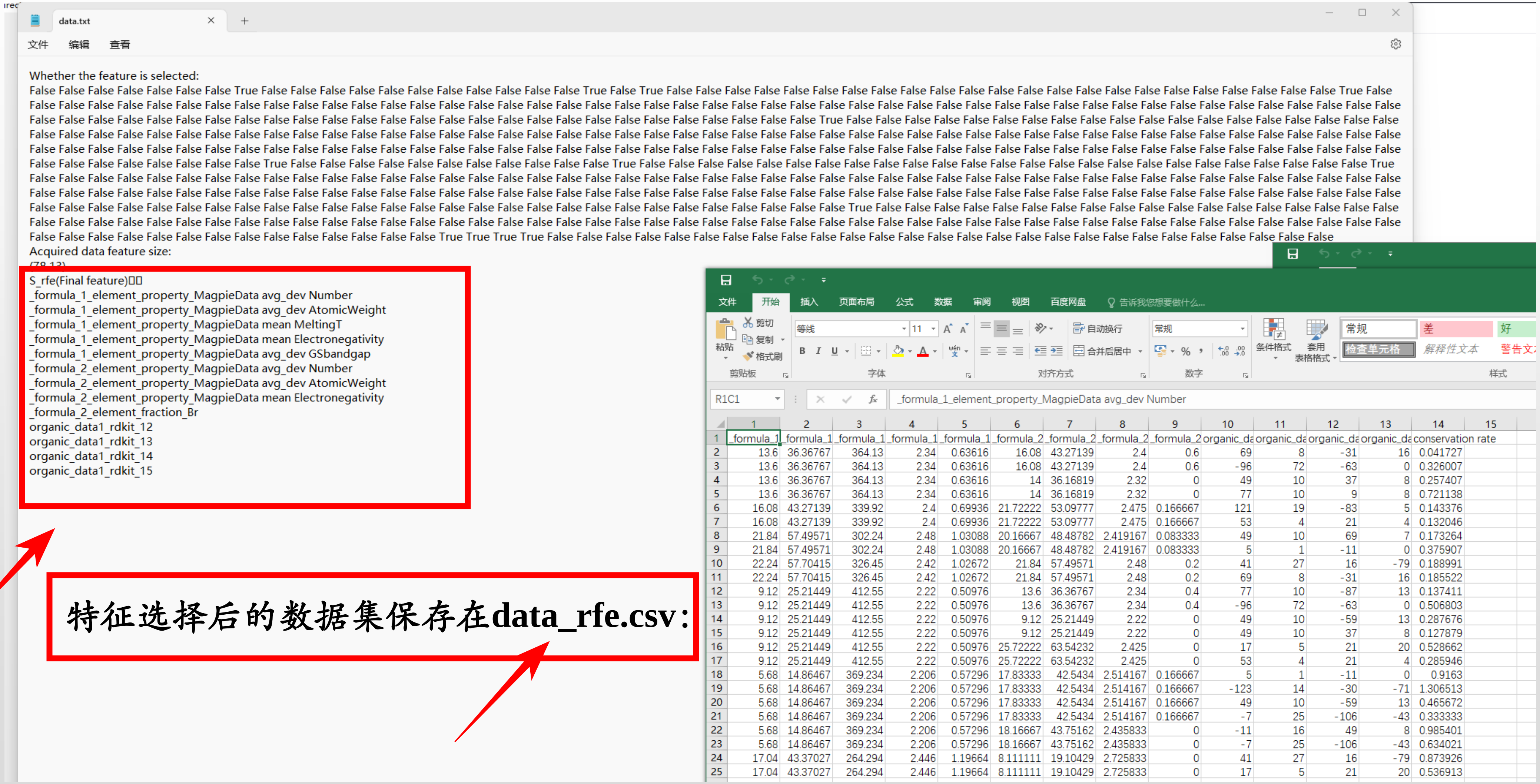


如果对话框填写13，自动生成的data_rfe.csv中的特征为13列，左右一列为输出数据

输入特征　　输出目标

_formula_1	formula_1	formula_1	formula_1	formula_1	formula_2	formula_2	formula_2	formula_2	organic_da	organic_da	organic_da	organic_da	conservation rate
13.6	36.36767	364.13	0.7952	0.63616	16.08	2.4	6	0.6	69	8	-31	16	0.041727
13.6	36.36767	364.13	0.7952	0.63616	16.08	2.4	6	0.6	-96	72	-63	0	0.326007
13.6	36.36767	364.13	0.7952	0.63616	14	2.32	8	0	49	10	37	8	0.257407
13.6	36.36767	364.13	0.7952	0.63616	14	2.32	8	0	77	10	9	8	0.721138
16.08	43.27139	339.92	0.8742	0.69936	21.72222	2.475	7.944444	0.166667	121	19	-83	5	0.143376
16.08	43.27139	339.92	0.8742	0.69936	21.72222	2.475	7.944444	0.166667	53	4	21	4	0.132046
21.84	57.49571	302.24	1.2886	1.03088	20.16667	2.419167	6.388889	0.083333	49	10	69	7	0.173264
21.84	57.49571	302.24	1.2886	1.03088	20.16667	2.419167	6.388889	0.083333	5	1	-11	0	0.375907
22.24	57.70415	326.45	1.2096	1.02672	21.84	2.48	8.4	0.2	41	27	16	-79	0.188991
22.24	57.70415	326.45	1.2096	1.02672	21.84	2.48	8.4	0.2	69	8	-31	16	0.185522
9.12	25.21449	412.55	0.6372	0.50976	13.6	2.34	6	0.4	77	10	-87	13	0.137411
9.12	25.21449	412.55	0.6372	0.50976	13.6	2.34	6	0.4	-96	72	-63	0	0.506803
9.12	25.21449	412.55	0.6372	0.50976	9.12	2.22	6	0	49	10	-59	13	0.287676
9.12	25.21449	412.55	0.6372	0.50976	9.12	2.22	6	0	49	10	37	8	0.127879
9.12	25.21449	412.55	0.6372	0.50976	25.72222	2.425	7.944444	0	17	5	21	20	0.528662
9.12	25.21449	412.55	0.6372	0.50976	25.72222	2.425	7.944444	0	53	4	21	4	0.285946
5.68	14.86467	369.234	0.7162	0.57296	17.83333	2.514167	6.166667	0.166667	5	1	-11	0	0.9163
5.68	14.86467	369.234	0.7162	0.57296	17.83333	2.514167	6.166667	0.166667	-123	14	-30	-71	1.306513
5.68	14.86467	369.234	0.7162	0.57296	17.83333	2.514167	6.166667	0.166667	49	10	-59	13	0.465672

4.1 Preprocessing：特征选择保留特征



```
Whether the feature is selected:
False False False False False False False True False False False False False False False False False False False True False True False False False False False False False False False False False False False False False False False False False False False False False True False
False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False
False False False False False False False False False False False False False False False False False False False False False False False False False False False True False False False False False False False False False False False False False False False False False False False
False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False
False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False
False False False False False False False False True False False False False False False False False False False False True False False False False False False False False False False False False False False False False False False False False False False False False False True
False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False
False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False
False False False False False False False False False False False False False False False False False False False False False False False False False False False False True False False False False False False False False False False False False False False False False False False
False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False False
False False False False False False False False False False False False False False True True True True False False False False False False False False False False False False False False False False False False False False False False False False False False False
Acquired data feature size:
(78,13)
```

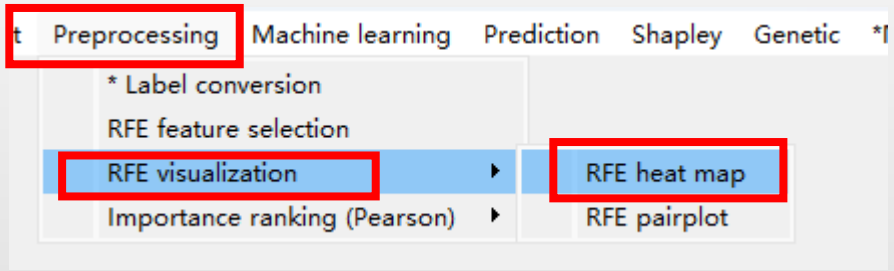
```
S_rfe(Final feature)□□
_formula_1_element_property_MagpieData avg_dev Number
_formula_1_element_property_MagpieData avg_dev AtomicWeight
_formula_1_element_property_MagpieData mean MeltingT
_formula_1_element_property_MagpieData mean Electronegativity
_formula_1_element_property_MagpieData avg_dev GSbandgap
_formula_2_element_property_MagpieData avg_dev Number
_formula_2_element_property_MagpieData avg_dev AtomicWeight
_formula_2_element_property_MagpieData mean Electronegativity
_formula_2_element_fraction_Br
organic_data1_rdkit_12
organic_data1_rdkit_13
organic_data1_rdkit_14
organic_data1_rdkit_15
```

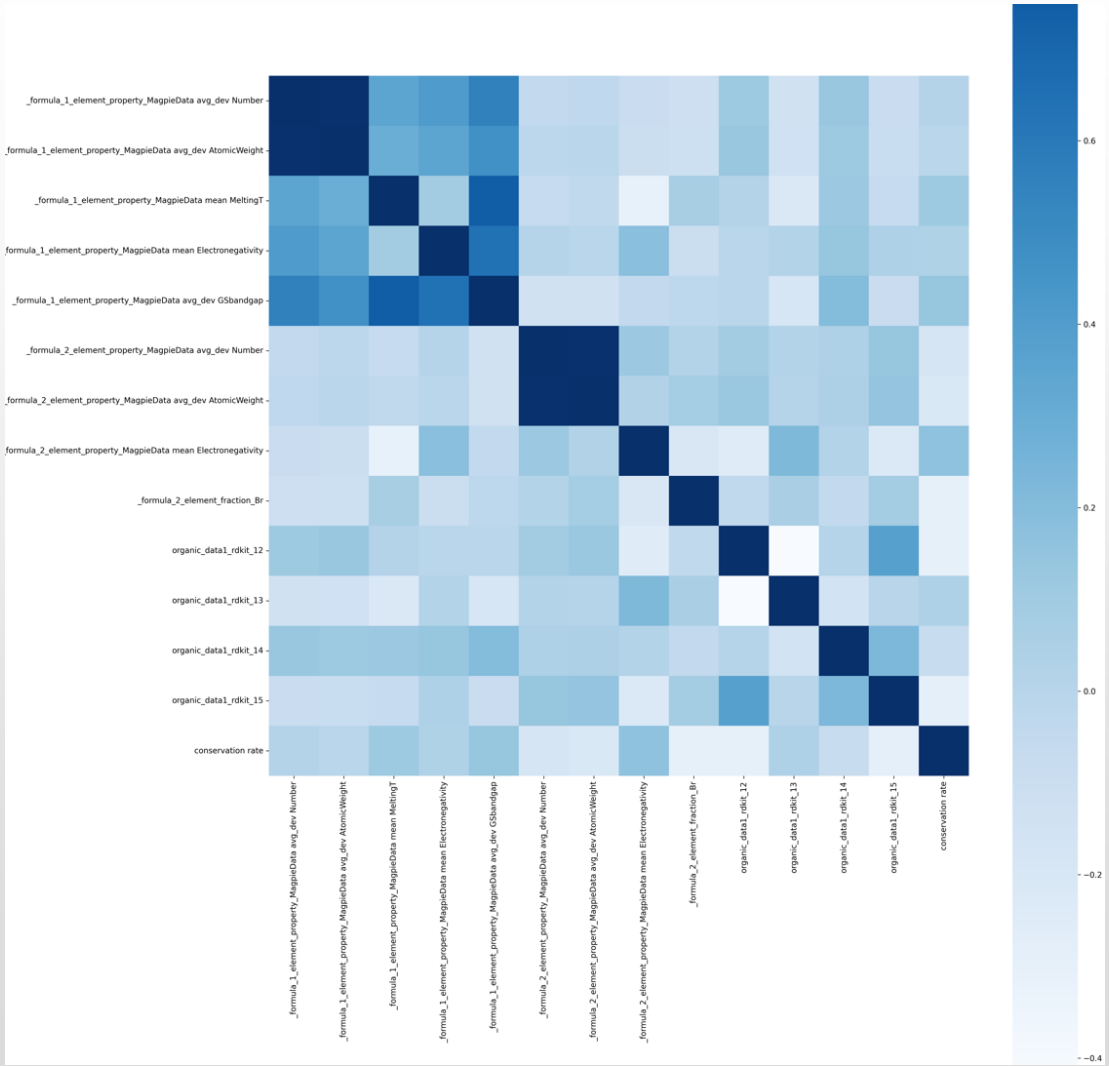
特征选择后的数据集保存在data_rfe.csv：

R1C1 _formula_1_element_property_MagpieData avg_dev Number

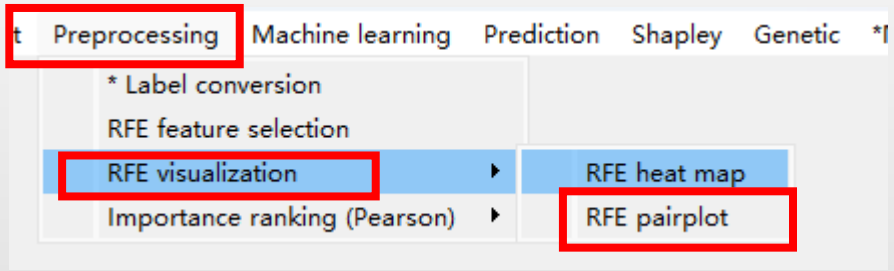
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	_formula_1	_formula_1	_formula_1	_formula_1	_formula_1	_formula_2	_formula_2	_formula_2	_formula_2	organic_da	organic_da	organic_da	organic_da	conservation rate	
2	13.6	36.36767	364.13	2.34	0.63616	16.08	43.27139	2.4	0.6	69	8	-31	16	0.041727	
3	13.6	36.36767	364.13	2.34	0.63616	16.08	43.27139	2.4	0.6	-96	72	-63	0	0.326007	
4	13.6	36.36767	364.13	2.34	0.63616	14	36.16819	2.32	0	49	10	37	8	0.257407	
5	13.6	36.36767	364.13	2.34	0.63616	14	36.16819	2.32	0	77	10	9	8	0.721138	
6	16.08	43.27139	339.92	2.4	0.69936	21.72222	53.09777	2.475	0.166667	121	19	-83	5	0.143376	
7	16.08	43.27139	339.92	2.4	0.69936	21.72222	53.09777	2.475	0.166667	53	4	21	4	0.132046	
8	21.84	57.49571	302.24	2.48	1.03088	20.16667	48.48782	2.419167	0.083333	49	10	69	7	0.173264	
9	21.84	57.49571	302.24	2.48	1.03088	20.16667	48.48782	2.419167	0.083333	5	1	-11	0	0.375907	
10	22.24	57.70415	326.45	2.42	1.02672	21.84	57.49571	2.48	0.2	41	27	16	-79	0.188991	
11	22.24	57.70415	326.45	2.42	1.02672	21.84	57.49571	2.48	0.2	69	8	-31	16	0.185522	
12	9.12	25.21449	412.55	2.22	0.50976	13.6	36.36767	2.34	0.4	77	10	-87	13	0.137411	
13	9.12	25.21449	412.55	2.22	0.50976	13.6	36.36767	2.34	0.4	-96	72	-63	0	0.506803	
14	9.12	25.21449	412.55	2.22	0.50976	9.12	25.21449	2.22	0	49	10	-59	13	0.287676	
15	9.12	25.21449	412.55	2.22	0.50976	9.12	25.21449	2.22	0	49	10	37	8	0.127879	
16	9.12	25.21449	412.55	2.22	0.50976	25.72222	63.54232	2.425	0	17	5	21	20	0.528662	
17	9.12	25.21449	412.55	2.22	0.50976	25.72222	63.54232	2.425	0	53	4	21	4	0.285946	
18	5.68	14.86467	369.234	2.206	0.57296	17.83333	42.5434	2.514167	0.166667	5	1	-11	0	0.9163	
19	5.68	14.86467	369.234	2.206	0.57296	17.83333	42.5434	2.514167	0.166667	-123	14	-30	-71	1.306513	
20	5.68	14.86467	369.234	2.206	0.57296	17.83333	42.5434	2.514167	0.166667	49	10	-59	13	0.465672	
21	5.68	14.86467	369.234	2.206	0.57296	17.83333	42.5434	2.514167	0.166667	-7	25	-106	-43	0.333333	
22	5.68	14.86467	369.234	2.206	0.57296	18.16667	43.75162	2.435833	0	-11	16	49	8	0.985401	
23	5.68	14.86467	369.234	2.206	0.57296	18.16667	43.75162	2.435833	0	-7	25	-106	-43	0.634021	
24	17.04	43.37027	264.294	2.446	1.19664	8.111111	19.10429	2.725833	0	41	27	16	-79	0.873926	
25	17.04	43.37027	264.294	2.446	1.19664	8.111111	19.10429	2.725833	0	17	5	21	20	0.536913	

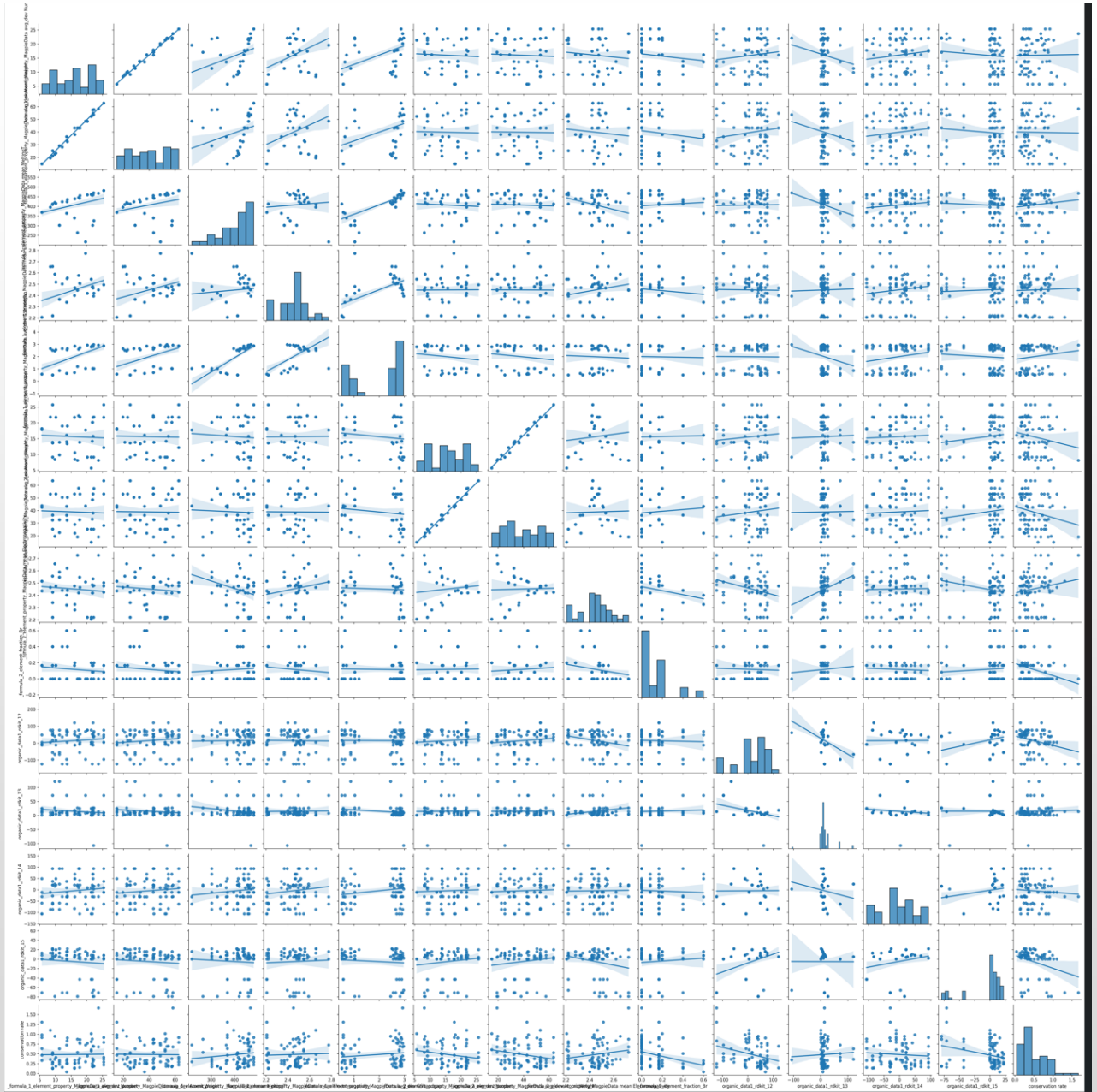
4.2 Preprocessing→特征选择后热图



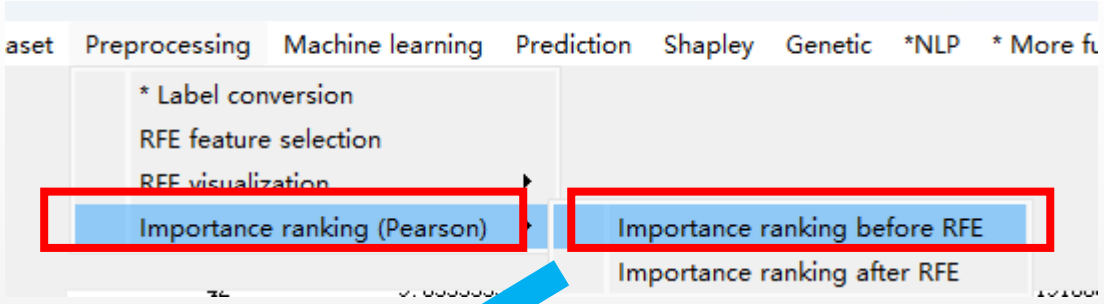


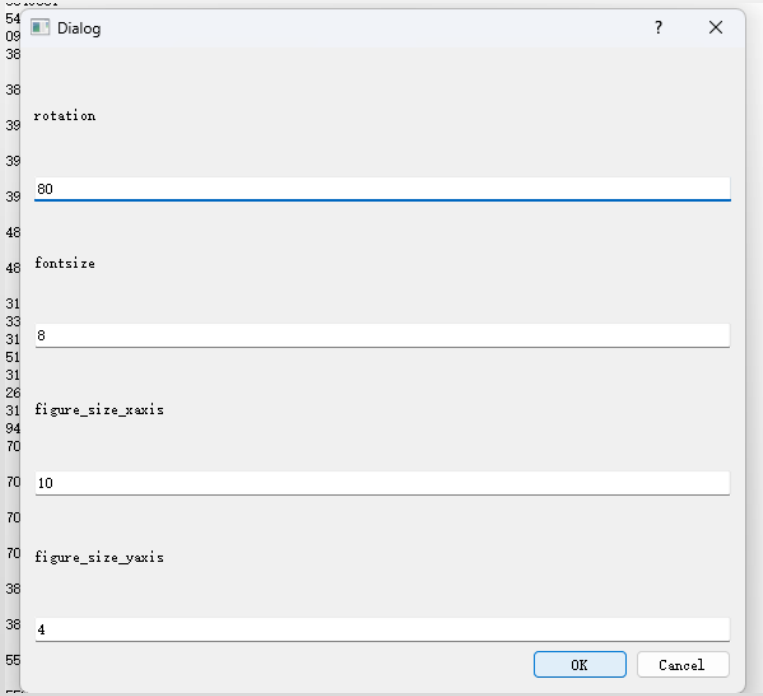
4.2 Preprocessing→特征选择后pairplot



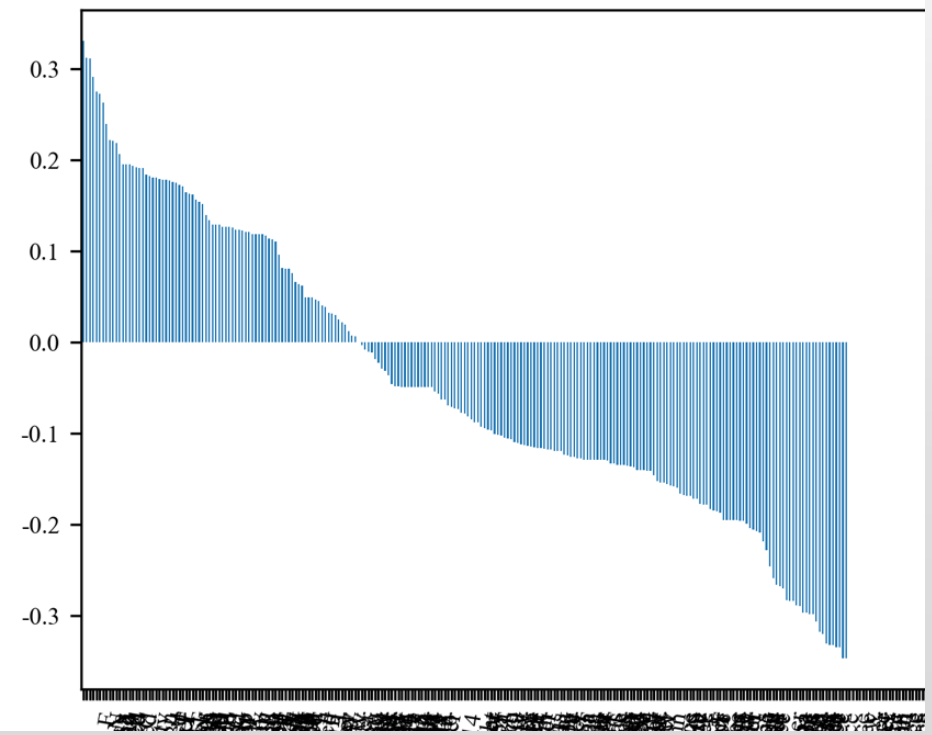


4.3.1 Preprocessing→RFE前的特征重要性排名



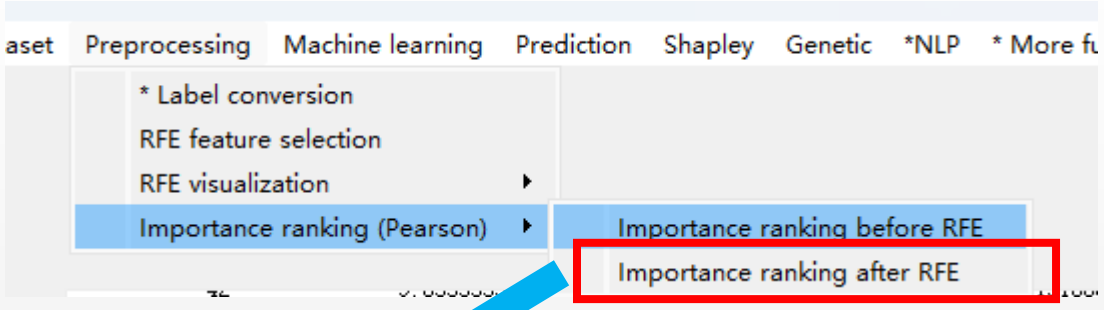


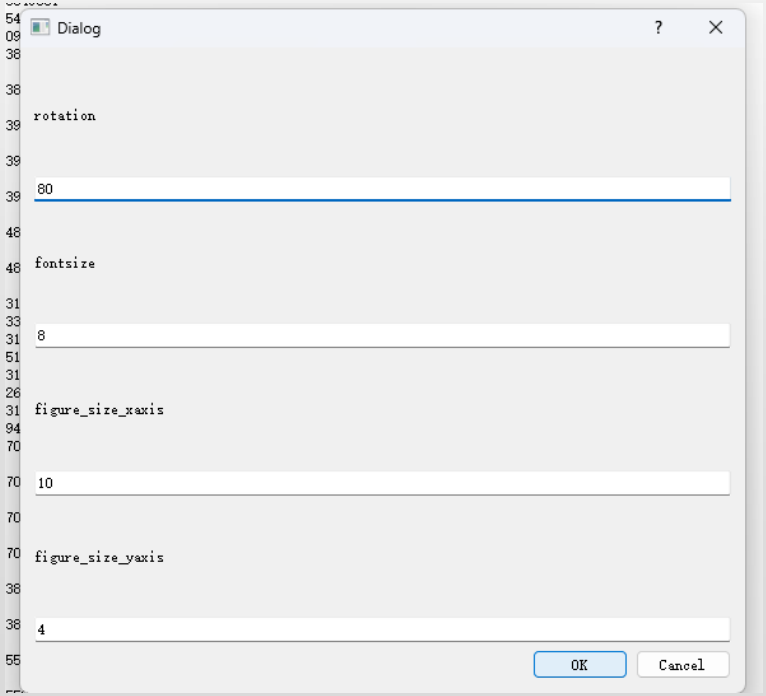
特征重要性排名（Pearson)



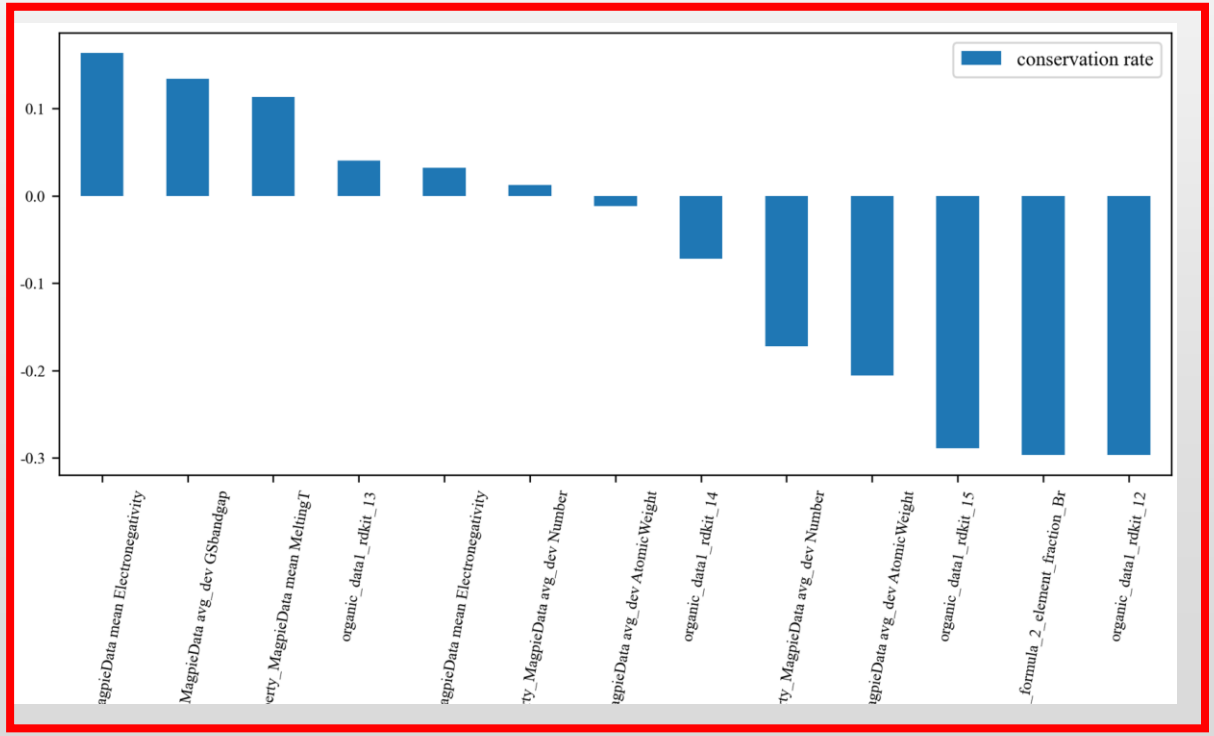
Shap排名更被接受

4.3.2 Preprocessing → RFE后的特征重要性排名





RFE特征选择后特征重要性排名（Pearson)

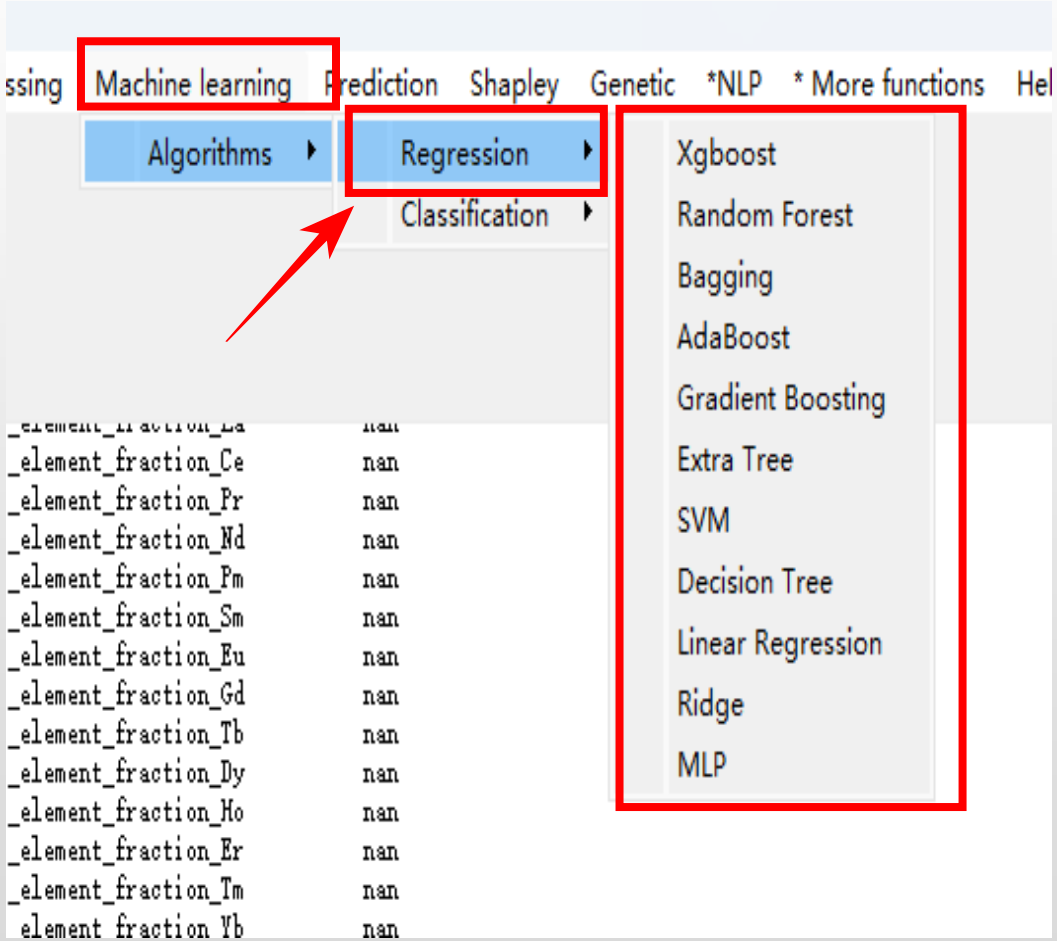


Shap排名更被接受

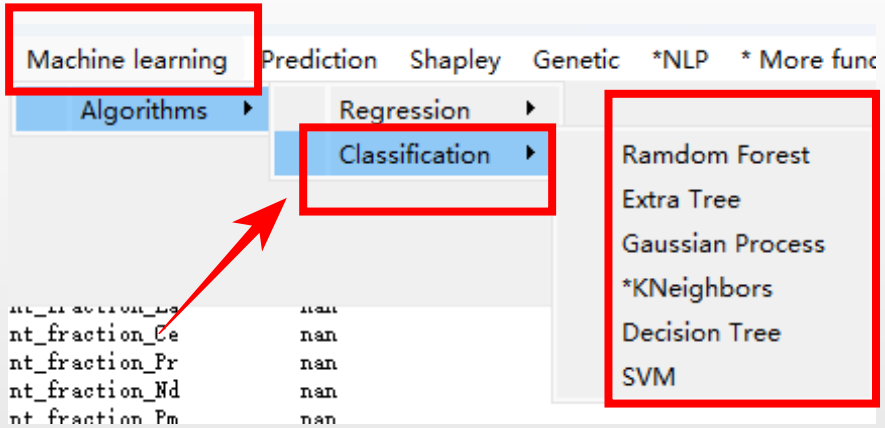
5 机器学习建模

5.1 机器学习建模

机器学习**回归**任务

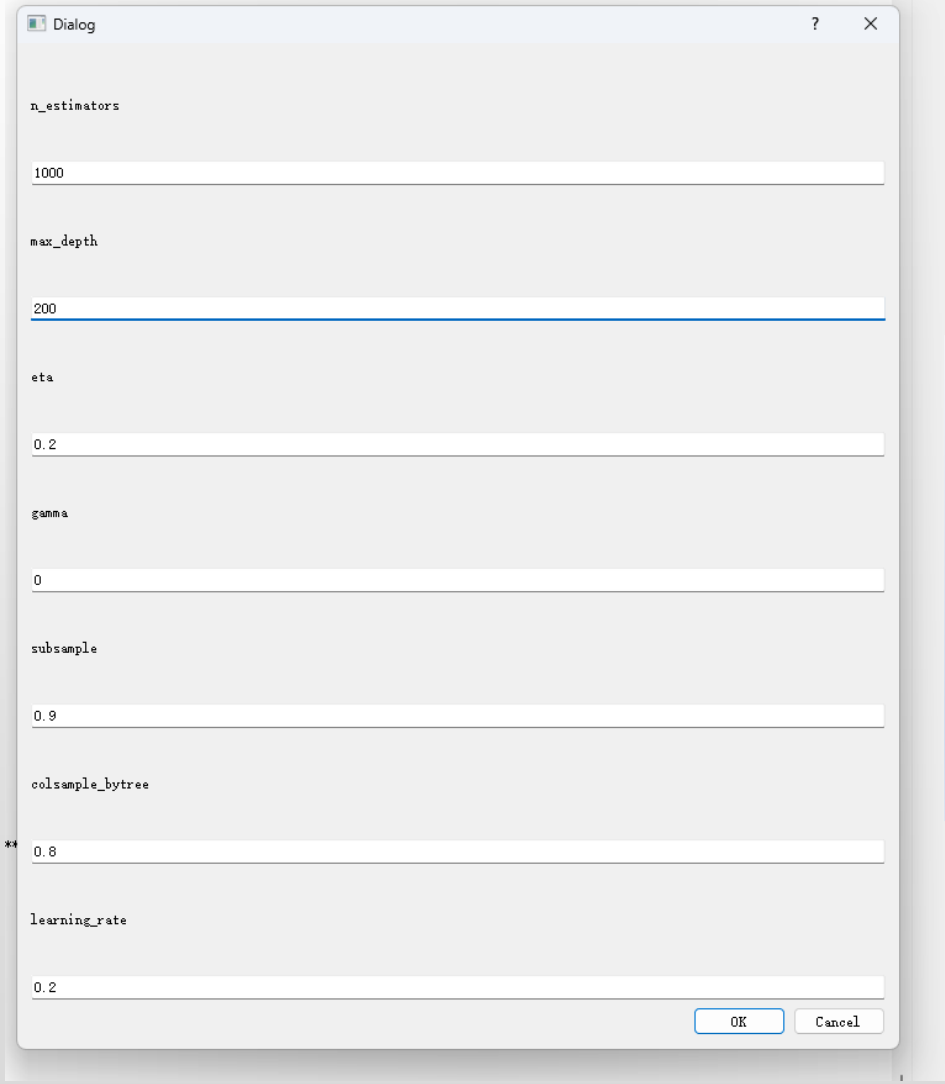


机器学习**分类**任务



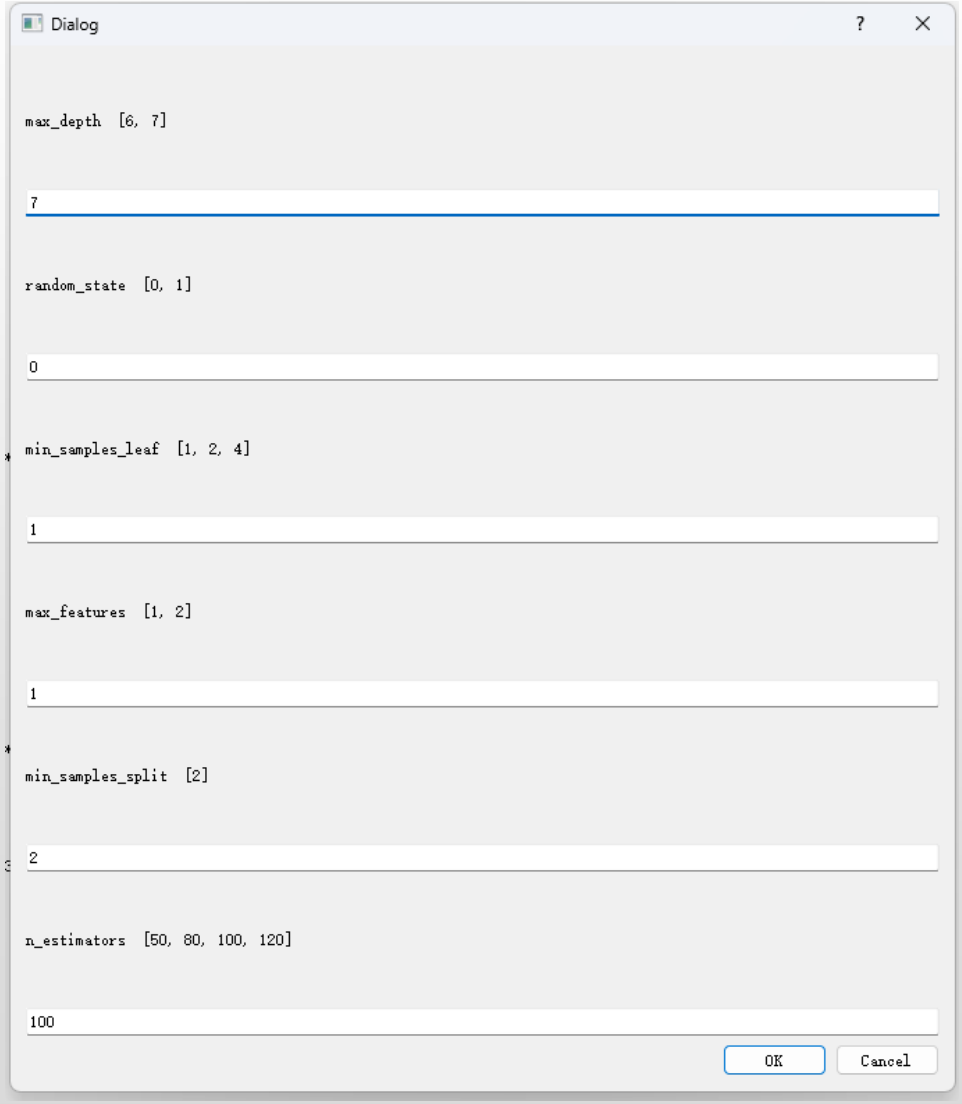
5.2 机器学习:超参数设置（单独对话框）

例：XGBoost 超参数默认，可调参



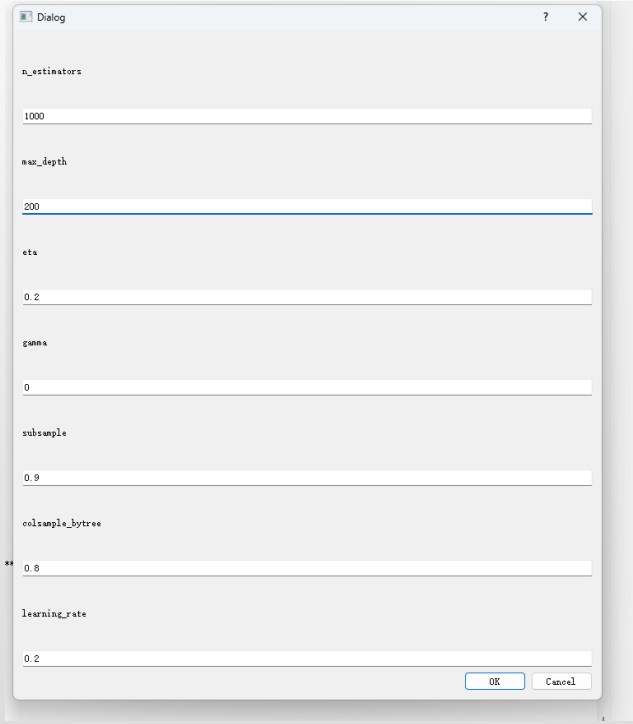
Machine learning 按钮下所有回归和分类机器学习算法皆可自定义超参数

例：随机森林超参数默认，可调参

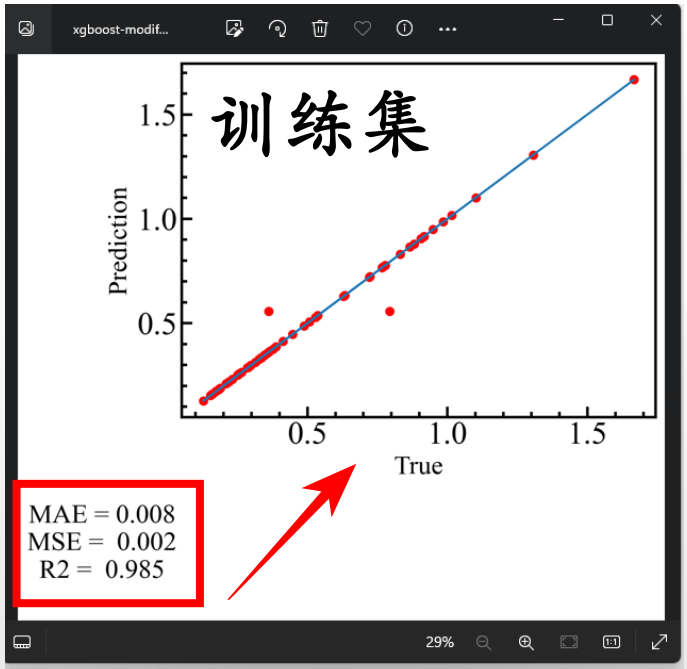


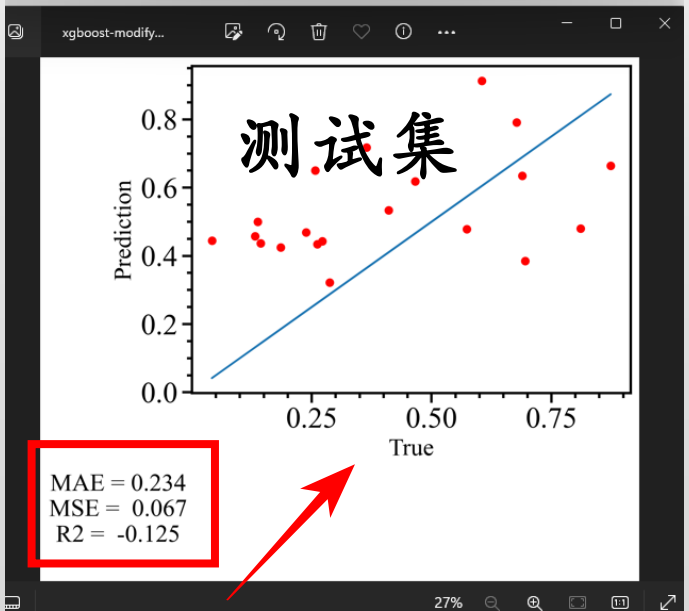
5.3 机器学习:数据拟合、准确率与超参数重新设置

默认超算数

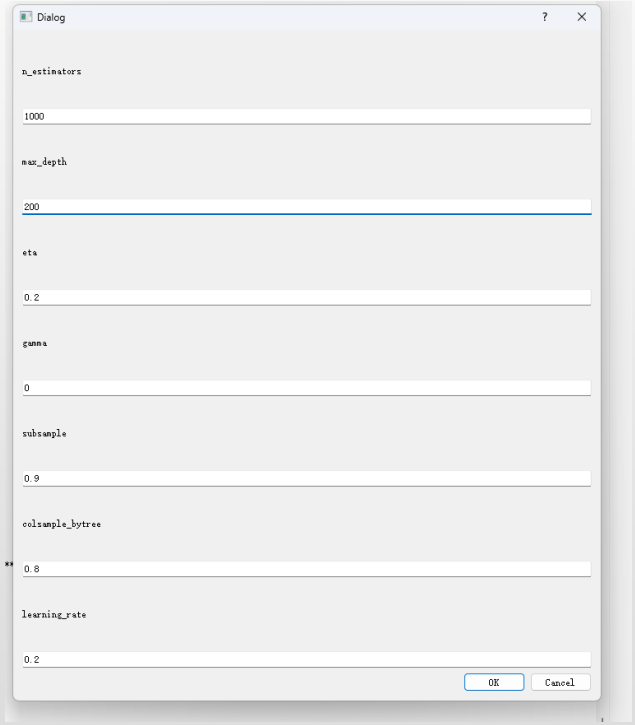


数据拟合





重新调节超算数

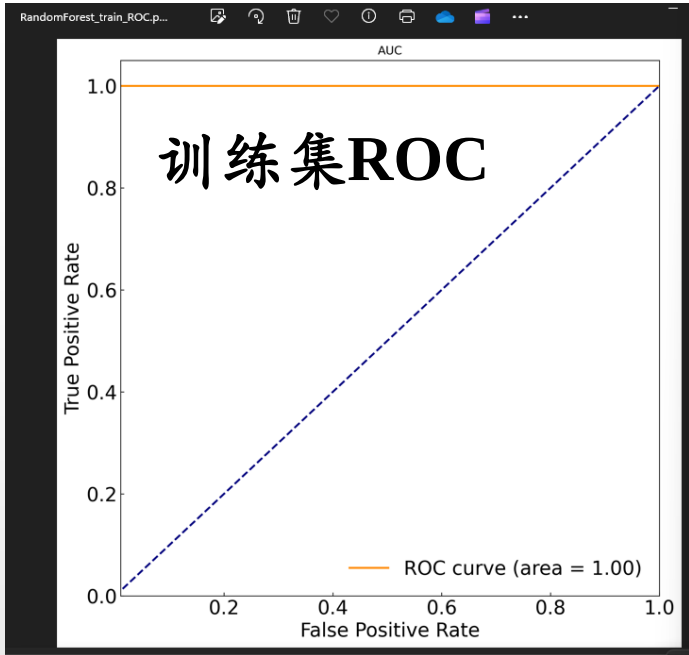


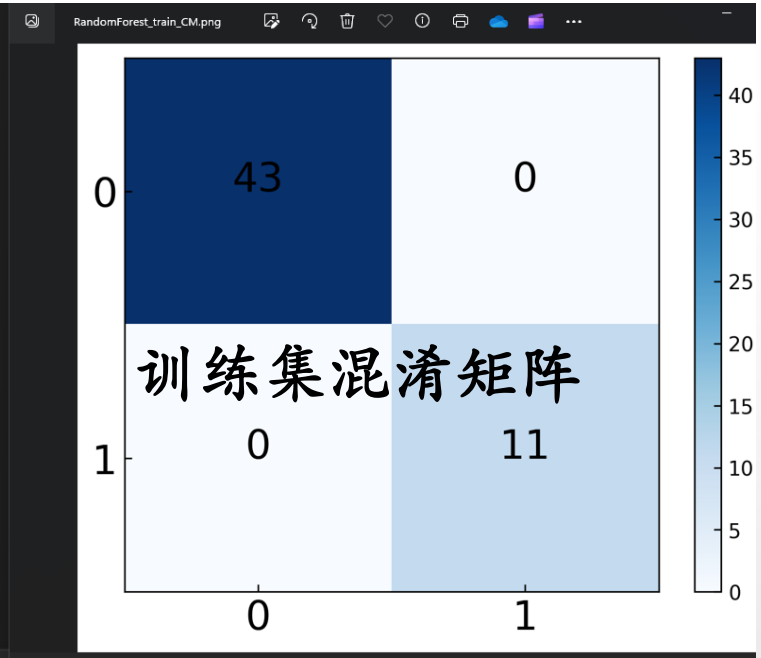
反复调参，直至测试集准确

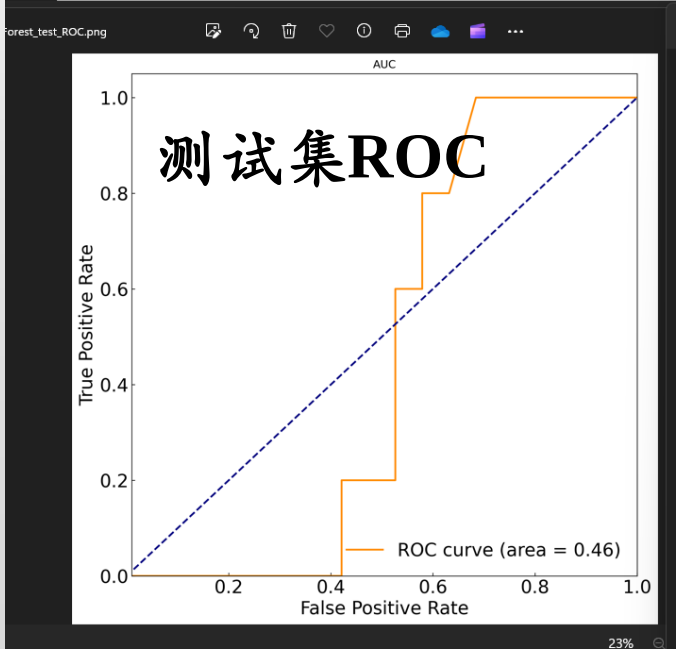
准确率提升方法:

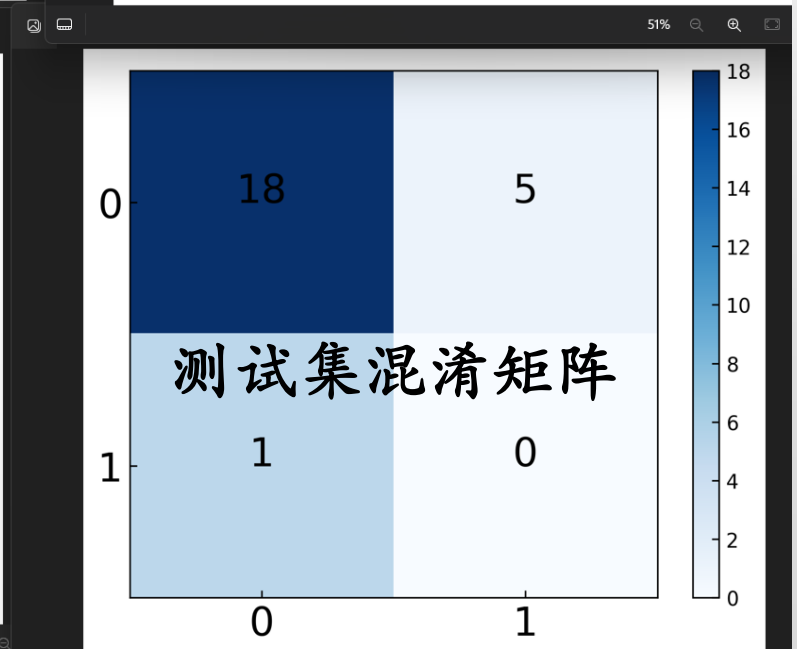
1. Customized 描述符!!!
2. 机器学习（算法与超参数）

5.4 机器学习分类任务（e.g.,稳定与否）









反复调参，直至测试集准确

Again，准确率提升方法：

1. Customized 描述符！（重点）

2. 机器学习（算法与超参数）

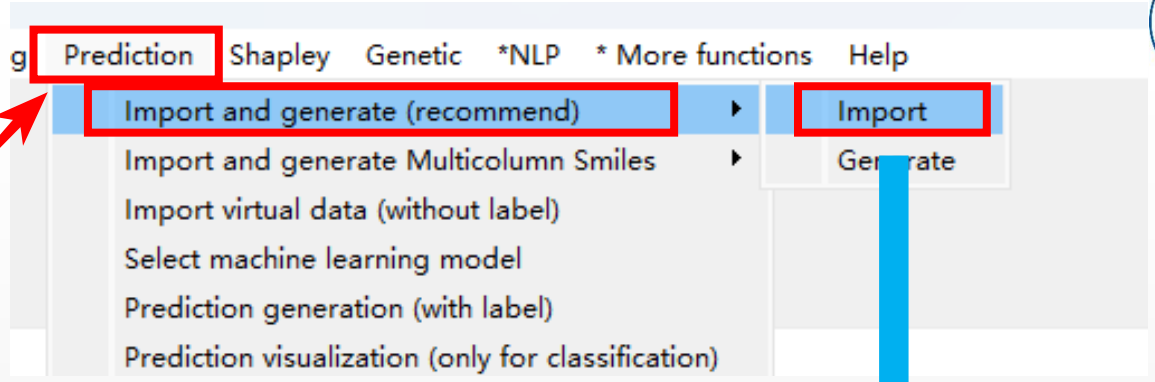
6 虚拟空间预测

预测虚拟空间

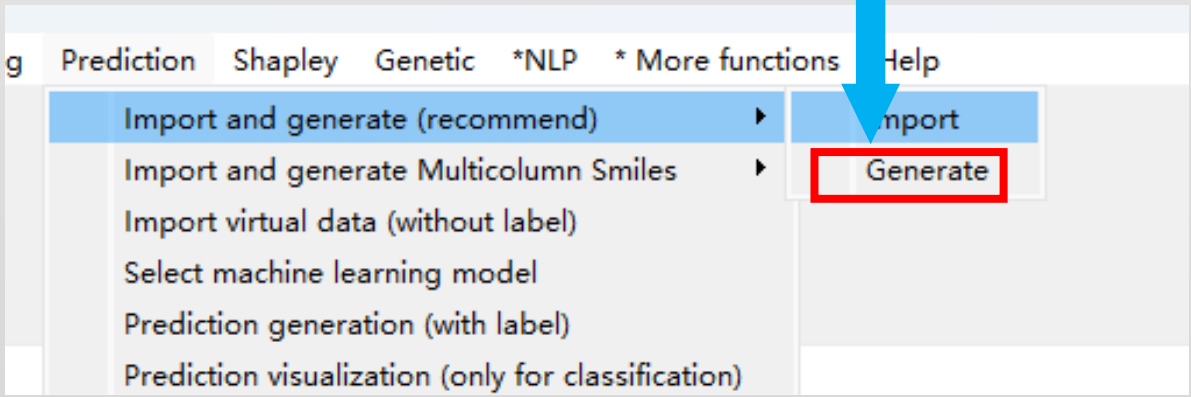
如果需要辅助特征化
从import and generate做起
按照下述步骤：
Prediction → Import and generate (保持与前一致） → Import → Generate → Import virtual data（without label) →
Select machine learning model → Prediction generation (with label)

如果不需要辅助特征化、已经有特征化好的虚拟空间，直接跳过本步骤：
直接选择**Prediction → Import virtual data（without label) → Select machine learning model → Prediction generation (with label)**

虚拟空间csv保持和原始csv格式一致
但最右一列为空（需要预测）
最右一列需有列名

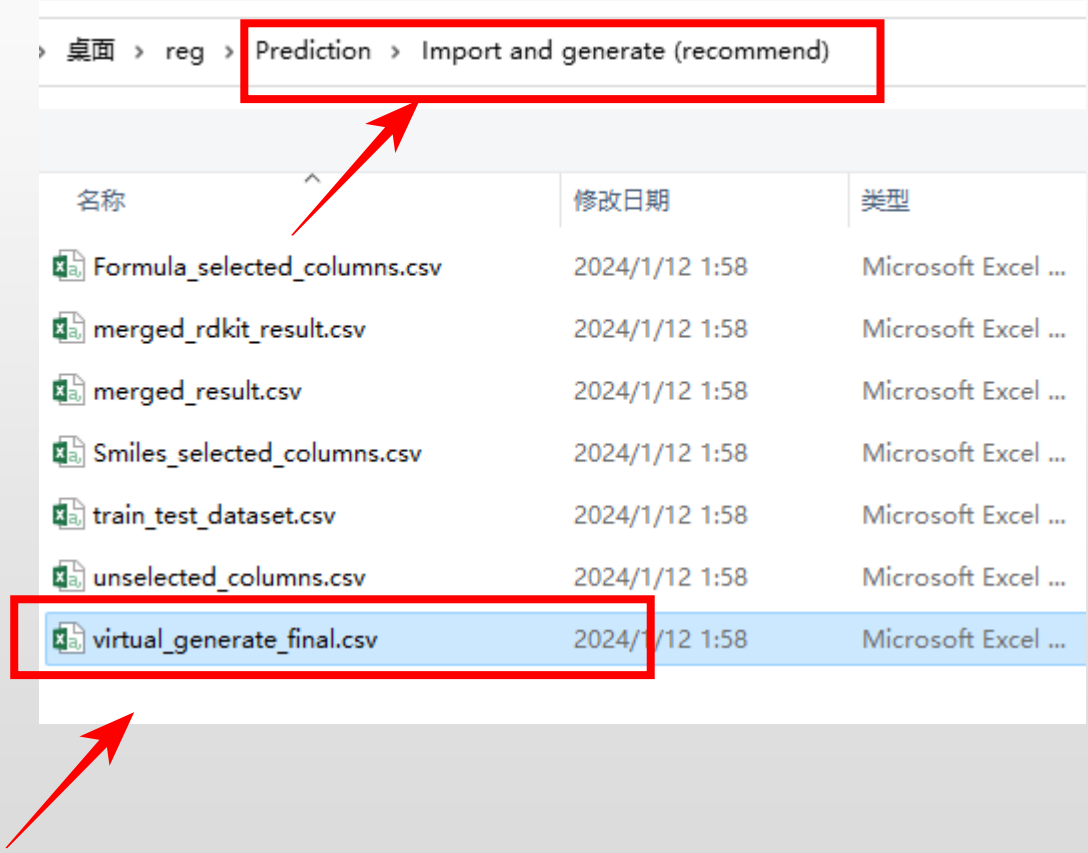


	1	2	3	4
1	layer1_Formula	layer2_Formula	SMILES_Smiles	target
2	CH3NH3SnF2Cl	CH3NH3SnI3	C1=CC(=CC(=C1)Br)/C=C/C(=O)O	
3	CsSnClBr2	CH3NH3SnI3	[NH4+].[Cl-]	
4	CH3NH3PbI3	(NH2)2CHSnI3	C(=N)N.[I-]	
5	CsSnF2Cl	(NH2)2CHSnI3	[Li+].[Cl-]	
6	(NH2)2CHSnI3	CH3NH3SnI3	[Na+].[Cl-]	
7	CsSnBr2I	CsPbI3	[Mg+2].[Cl-].[Cl-]	
8	CsSnBr2I	CsPbFCl2	[Cl-].[K+]	



预测虚拟空间

Import virtual data（without label)
Virtual_generate_final.csv



预测虚拟空间

Prediction → Import virtual data（without label) → Select machine learning model →

Prediction generation (with label)

预测的数据（稳定性、效率等）
最终数据

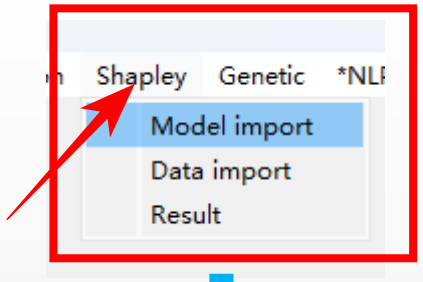
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	_formula_1	_formula_1	_formula_1	_formula_1	_formula_1	_formula_2	_formula_2	_formula_2	_formula_2	organic_da	organic_da	organic_da	organic_da	Output
2	8.388889	19.5697	396.1492	2.655833	2.656556	22.16667	53.7101	2.394167	0	62	-107	3	-66	0.938381
3	11.28	29.14595	301.974	2.366	0.86512	22.16667	53.7101	2.394167	0	69	8	-31	16	0.435465
4	25.72222	63.54232	477.6058	2.425	2.972236	21.83333	52.98795	2.464167	0	5	1	-11	0	0.519687
5	19.6	48.63574	217.054	2.774	1.02928	21.83333	52.98795	2.464167	0	17	5	21	20	0.441674
6	21.83333	52.98795	473.7317	2.464167	2.873903	22.16667	53.7101	2.394167	0	-11	16	49	8	0.652173
7	8.48	22.20927	345.024	2.266	0.63616	9.12	25.21449	2.22	0	65	2	-27	22	0.637941
8	8.48	22.20927	345.024	2.266	0.63616	26	67.2406	2.684	0	41	27	16	-79	0.889752
9														

待实验验证！
待模拟验证！

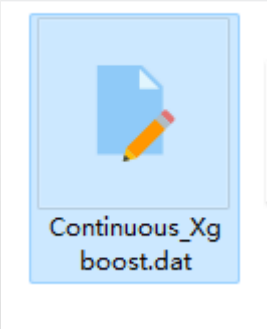
7 Shap特征分析

Shap特征分析

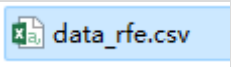
点击result:自动生成Shapley图

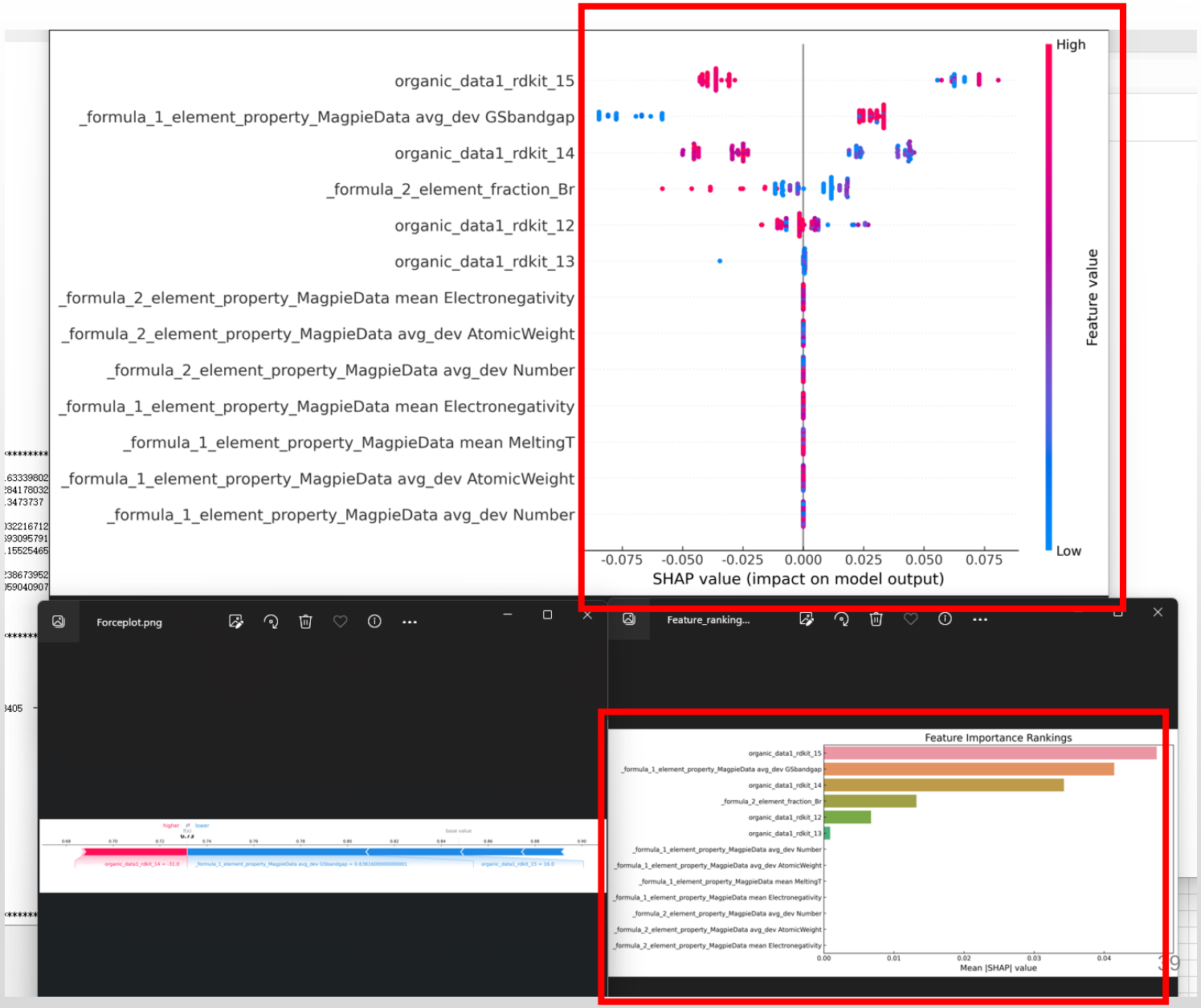


选择保存的机器学习模型.dat



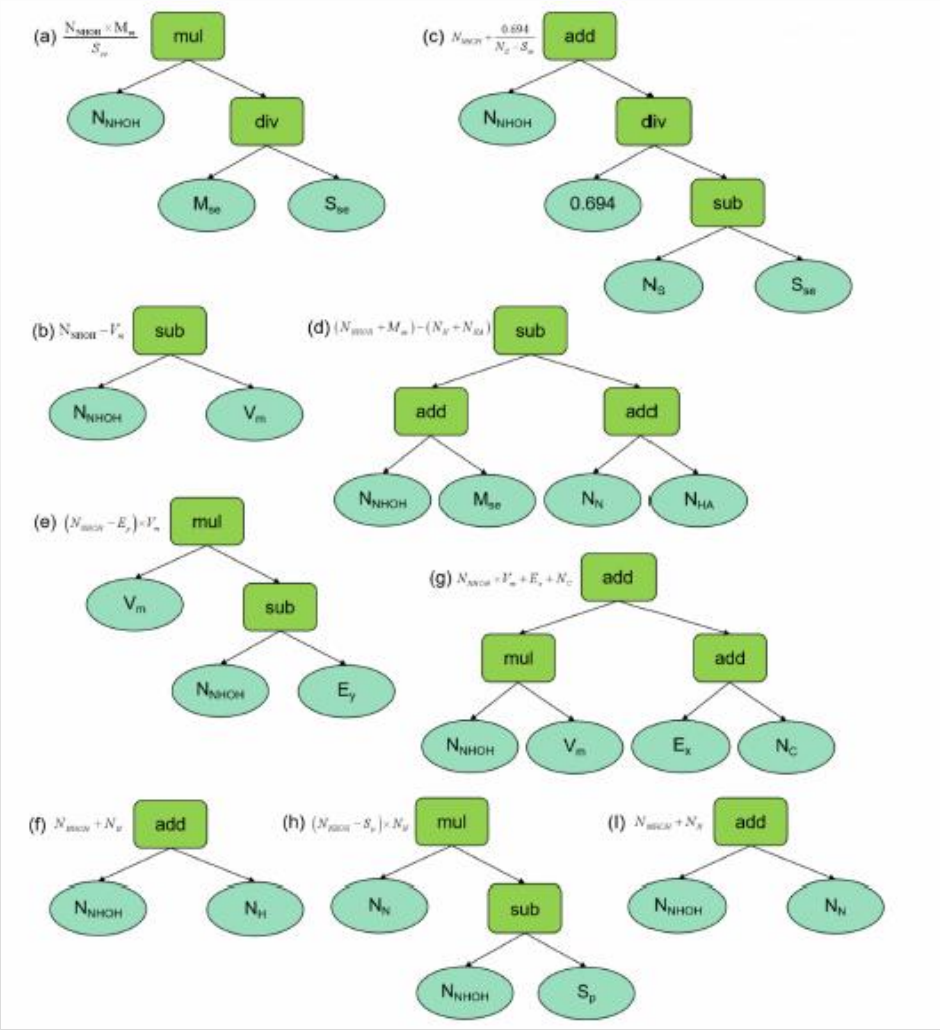
Data import:特征选择后的 Data_rfe.csv





8 遗传算法（符号回归/分类）

遗传算法:Mutation



Hoist Mutation

Point Mutation

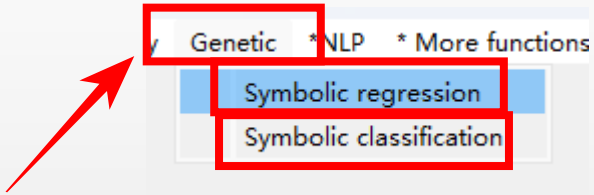
Crossover

Subtree mutation

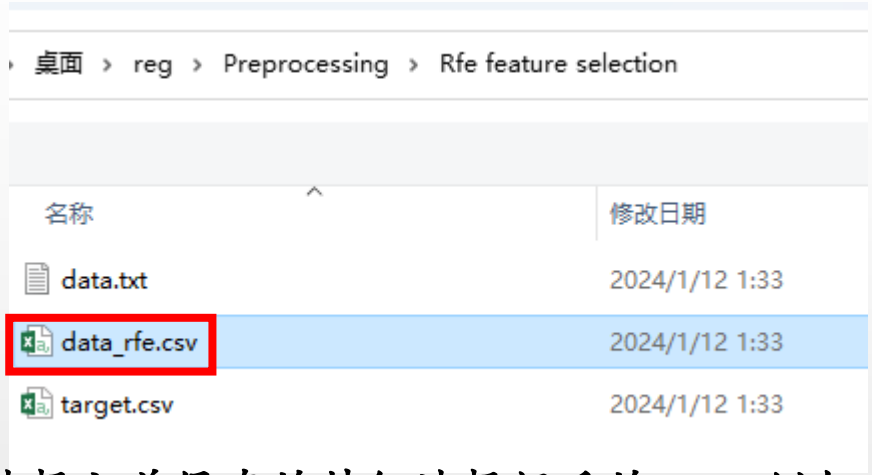
symbol	mathematical formula	correlation r
F1	$N_c(E_2^2 - M^3)$	0.50
F2	$N_c(E_2 - M)$	0.49
F3	$\frac{E_2^3}{M^2 + M_{ae}^2}$	0.48
F4	$(N_c^3 - \sqrt{E_2})(E_2 - M)$	0.51
F5	$E_2^2 - R$	0.46
F6	$E_2^2 - N_c M$	0.51
F7	$N_c^2 - N_c(M - E_9^2)$	0.51
F8	$\frac{E_2^2}{E_6^2\left(M^2 + E_4^2 \frac{N_c^2}{N_{ar}^2}\right)}$	0.50
F9	$E_4^2\left(E_2^2 + \frac{M^3}{M^2 + N_{ar}}\right)$	0.51
F10	$2(E_2^3 - M^3 - M_{ae}^3)\ \ln\ (E_2)$	0.50
F11	$N_c(N_c^2 - M - R)$	0.52
F12	$\frac{B_j E_2^2}{E_{10}(M^2 - E_4)}$	0.52

遗传算法建立模型：导入csv数据集

以特征选择后的Data_rfe.csv为例



符号回归与符号分类按钮



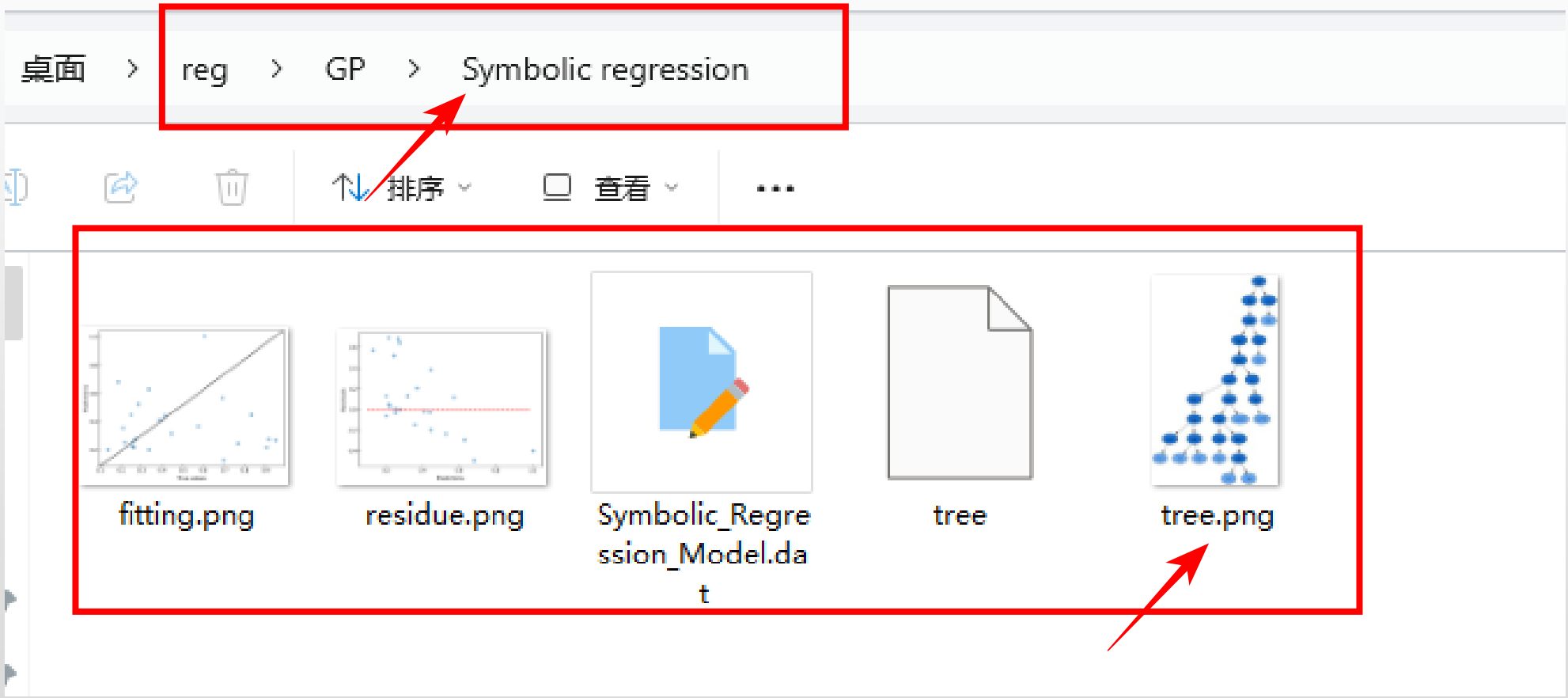
选择之前保存的特征选择好后的csv，例如data_rfe.csv

	Population Average		Best Individual			
Gen	Length	Fitness	Length	Fitness	OOB Fitness	Time Left
0	9.22	1.78342	5	0.245292	N/A	20.65s
1	5.89	0.570097	14	0.220634	N/A	20.99s
2	6.23	0.480999	14	0.213933	N/A	20.49s
3	7.54	0.45332	9	0.209441	N/A	21.24s
4	9.61	0.450839	15	0.203901	N/A	19.17s
5	10.74	0.361537	18	0.198648	N/A	18.18s
6	10.06	0.268182	21	0.189408	N/A	16.53s

遗传算法建模进行中！（目前默认超参数不可改，后期待续）

遗传算法建立模型：生成数据图

若未自动打开，查看图路径：自定义保存路径→ GP → Symbolic regression （分类图存在Symbolic classification）

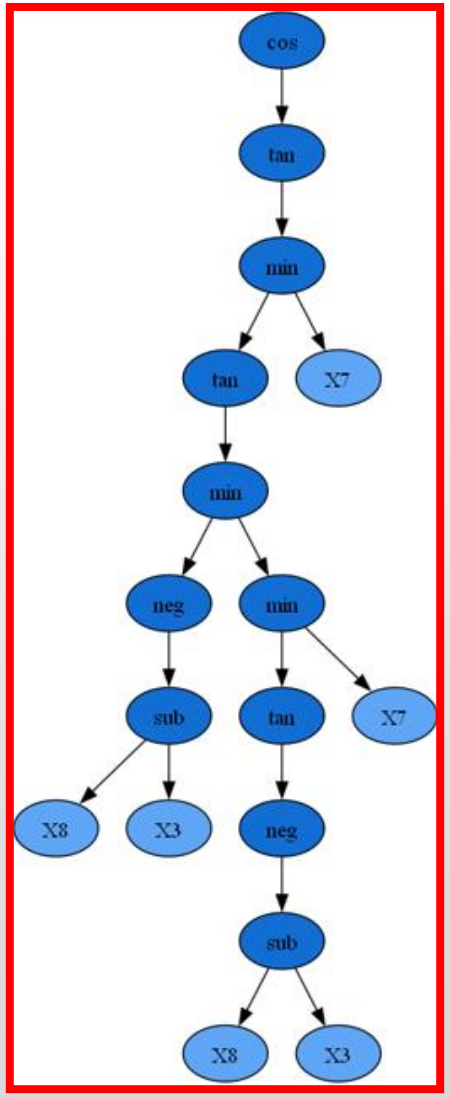


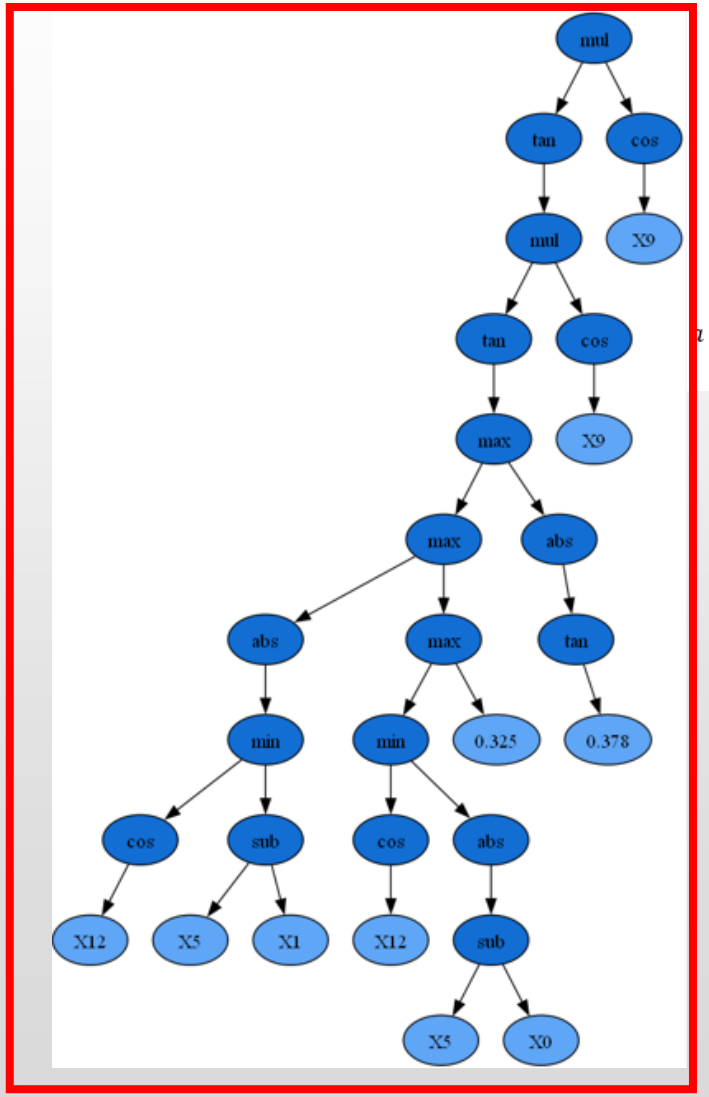
遗传算法建立模型：符号树图 Tree

可解释性！

科学！

白匣子！！





$$stability = f_1 + f_2 + f_3 \times f_4$$

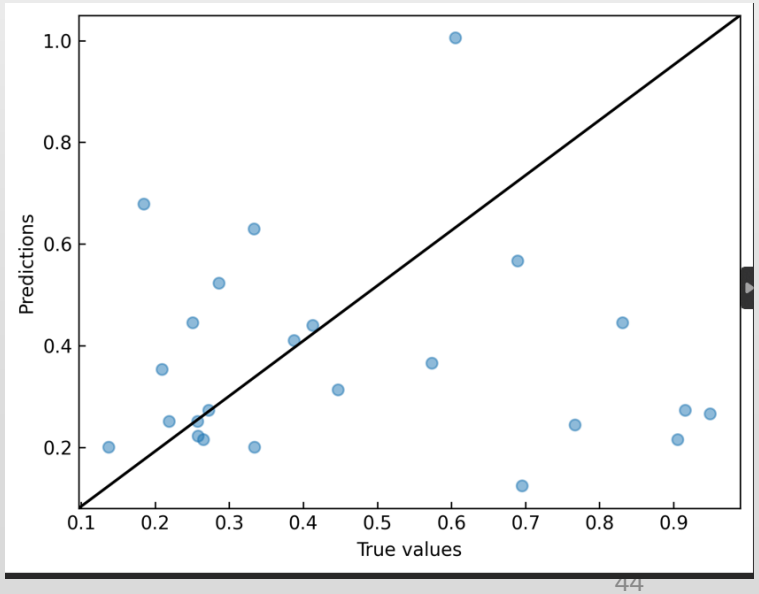
$$f_1 = R_A^2 + a + E_S + R_D$$

$$f_2 = \frac{D_{Ma}^2(E_B + A_S + b)}{h + b} + \frac{f}{h} + \frac{R_C}{D_{Fp}} + D_{Ma}(2d + g + 2C_{Me} + 2b)$$

$$f_3 = R_A^2 + a + \frac{fD_{Ma}}{b} + \frac{D_{Ma}(d + b)}{D_{Ma} + d + \frac{f}{d}}$$

$$f_4 = D_{Ma}(d + g + 2C_{Me}) + D_{Ma}^2(C_{Me} + b + A_s + 1) + \frac{f}{d} + d + C_S$$

$$a = \frac{R_A R_D}{R_A + E_S}, b = \frac{R_D}{D_{MI}}, d = C_{Me} + R_C, f = D_{Fp+}D_s, g = \frac{D_{Fp}}{E_B + b}, h = \frac{D_{Fp+}D_s}{E_b + b}$$



未完待续（其它软件功能

NLP/Featurizer/Algorithms/DFT/SLME/CSP

/Linux/CGCNN等）

Acknowledgement and Conclusions

NSFC
National Natural Science Foundation of China

江苏省科学技术厅
Jiangsu Provincial Department of Science and Technology

- 材料设计中小数据问题仍然严重（无论模拟和实验：成本高，欠拟合，过拟合）
- 各类数据来源和算法各有利弊
- **Natural language processing (自然语言处理) facilitates materials prediction**
- 遗传算法**Genetic programming**帮助科学解释（公式/**decouple**解耦)
- 基于文本描述符的多模材料设计方法：准确率、成本与可解释性的较好均衡!
- **NJmat**软件开发（描述符与算法)

Thanks !