阿布都赛米·阿布都外力 学号：2020182631 考试号：180150129

**实验名称：2D-QSAR**

**实验目的：**

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。

**实验原理：**

使用Discovery Studio 软件进行2D-QSAR模型的构建、外部数据集检测、未知活性化合物预测。

本实验所用软件环境：

DS Version：19.1.0.18287

PP Version：19.1.0.1963

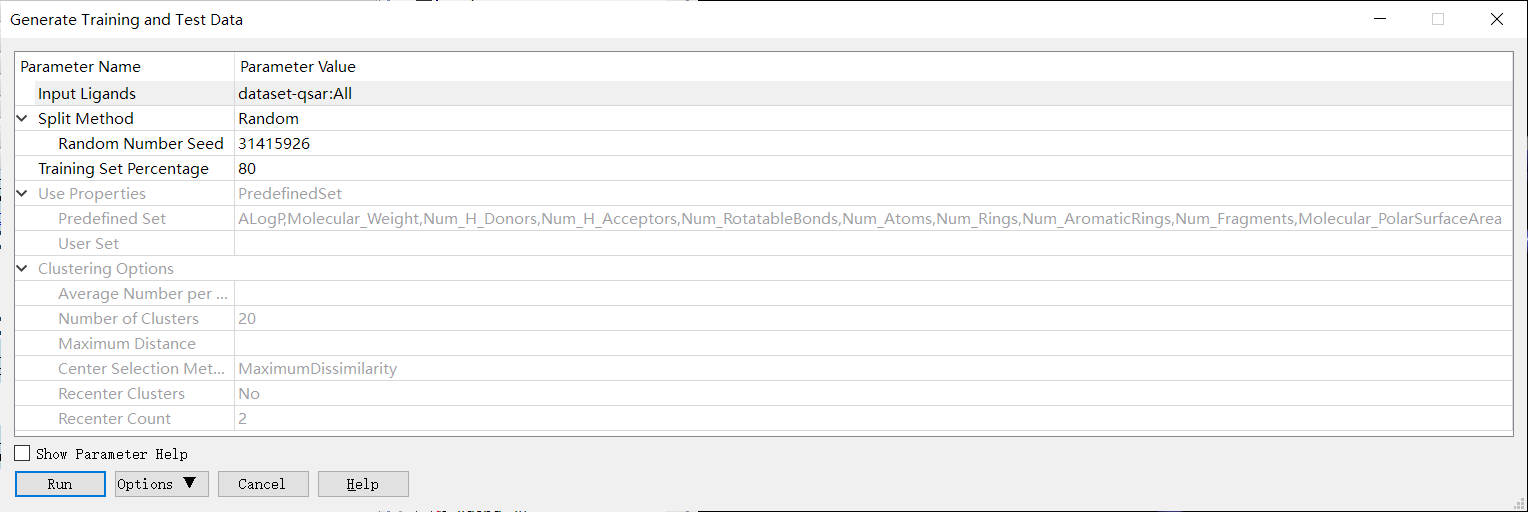
DS Client Version：19.1.0.18287

OS Distribution：Windows

OS Version：10.0.19044

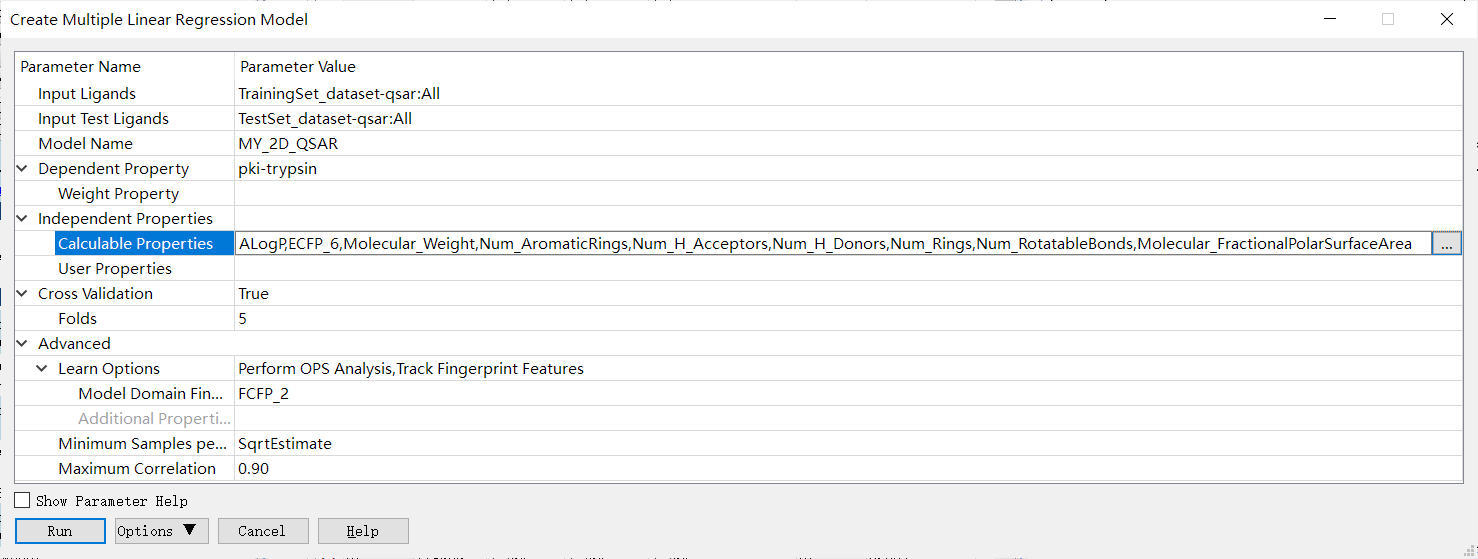
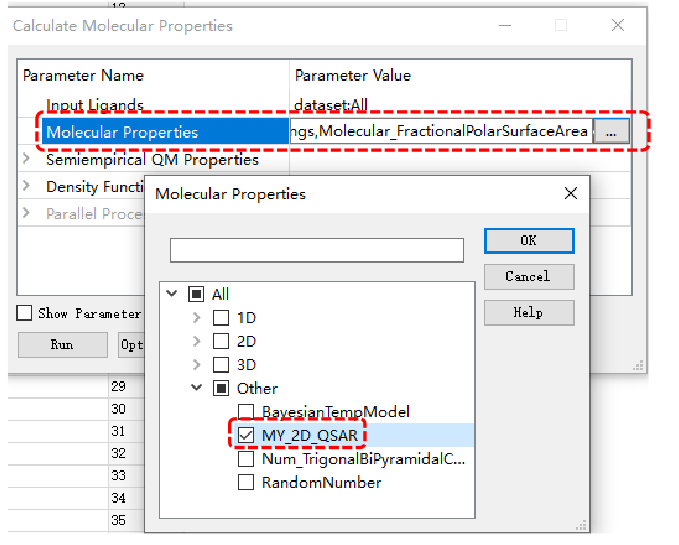
**实验步骤：**

1. 已知活性数据收集：本实验使用指导老师提供的dataset-qsar.sdf数据集。

 2. 数据集准备（训练集与测试集拆分等）：点击Discovery Studio软件上的Small Molecules🡪Create QSAR Model🡪Generate Training and Test Data进行训练集与测试集拆分。设置参数如下：

3. 分子描述属性计算（传统分子描述符、分子指纹等）：本实验使用指导老师提供的dataset-qsar.sdf数据集中的pki-trypsin描述符。

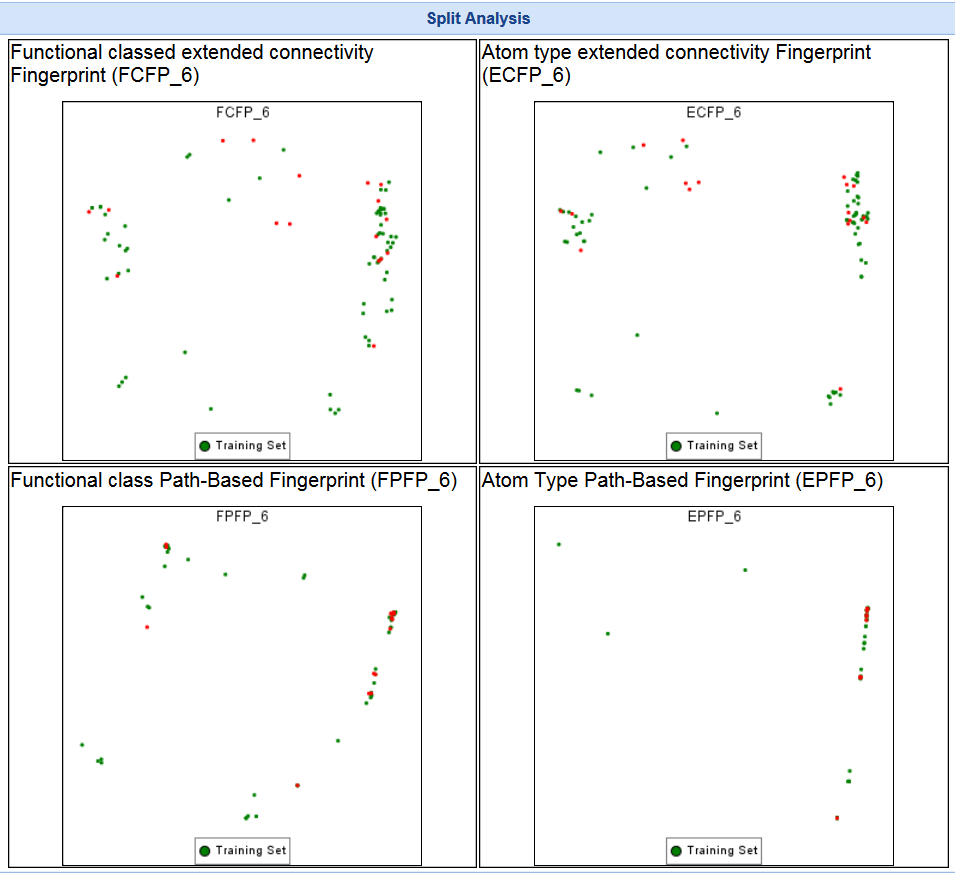
4. 模型的参数设置与构建：点击Discovery Studio软件上的Small Molecules🡪 Create QSAR Model 🡪 Create Multiple Linear Regression Model进行多元线性回归模型的构建。设置参数如下：

 5. 未知活性化合物预测：点击Discovery Studio软件上的Small Molecules🡪Calculate Molecular Properties🡪 Calculate Molecular Properties🡪进行未知活性化合物预测。设置参数如下：（本实验没做）

**实验结果：**

1. 数据集准备的结果：

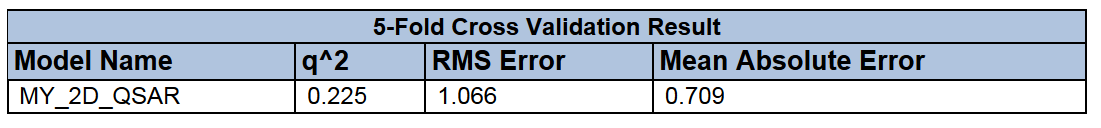
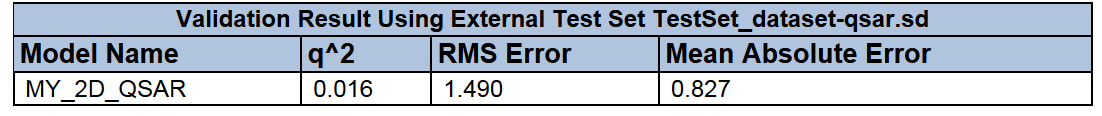
Status：Success Elapsed Time: 00:00:08

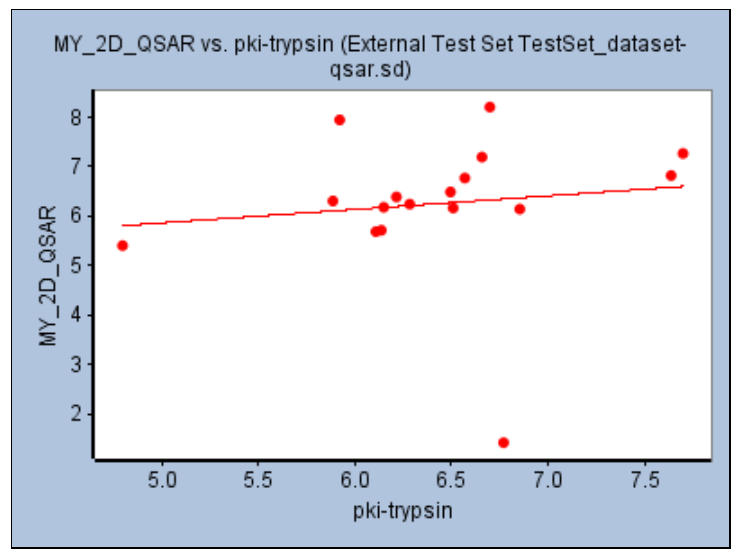
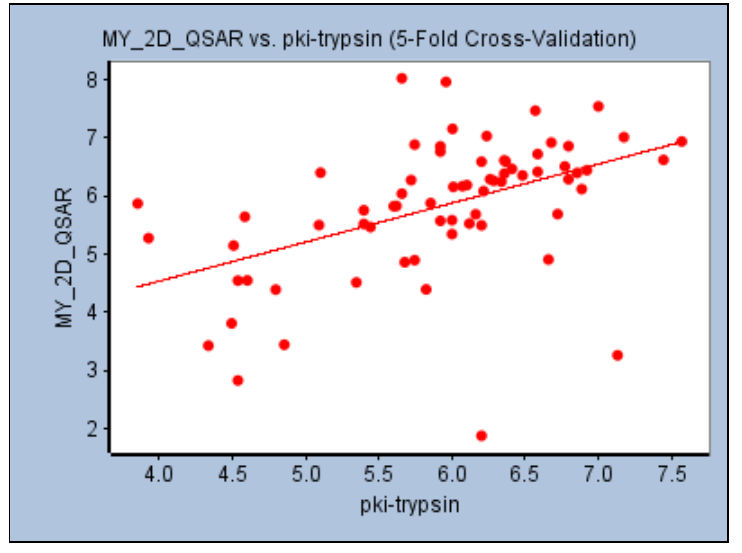
 Summary：Data split: 69 in training set, 17 in test set.

2. 模型的参数设置与构建的结果：

Status：Success Elapsed Time: 00:00:07

Summary: Multiple Linear Regression Model MY\_2D\_QSAR has been created. It contains the following equation:

[ MY\_2D\_QSAR ] = 0.3379 + 0.2242 \* [ ALogP ] + 0.4999 \* [ Count&lt;ECFP\_6:-1897341097&gt; ] + 0.3375 \* [ Count&lt;ECFP\_6:-1910270391&gt; ] + 0.3373 \* [ Count&lt;ECFP\_6:-1100000244&gt; ] + 0.3366 \* [ Count&lt;ECFP\_6:-1074141656&gt; ] + 0.3366 \* [ Count&lt;ECFP\_6:1559650422&gt; ] + 0.336 \* [ Count&lt;ECFP\_6:642810091&gt; ] + 0.337 \* [ Count&lt;ECFP\_6:-182236392&gt; ] + 0.3382 \* [ Count&lt;ECFP\_6:-606302475&gt; ] + 0.3366 \* [ Count&lt;ECFP\_6:1572579716&gt; ] + 1.616 \* [ Count&lt;ECFP\_6:-992506539&gt; ] + 0.8966 \* [ Count&lt;ECFP\_6:-167460056&gt; ] - 0.3814 \* [ Count&lt;ECFP\_6:734603939&gt; ] + 2.365 \* [ Count&lt;ECFP\_6:-797085356&gt; ] + 0.01107 \* [ Count&lt;ECFP\_6:944467641&gt; ] + 0.3375 \* [ Count&lt;ECFP\_6:2099970318&gt; ] + 0.338 \* [ Count&lt;ECFP\_6:770157610&gt; ] + 0.3361 \* [ Count&lt;ECFP\_6:-2024255407&gt; ] + 0.3356 \* [ Count&lt;ECFP\_6:1996767644&gt; ] + 0.3343 \* [ Count&lt;ECFP\_6:-786013480&gt; ] + 0.3362 \* [ Count&lt;ECFP\_6:1997021792&gt; ] + 0.3362 \* [ Count&lt;ECFP\_6:-175146122&gt; ] + 0.3367 \* [ Count&lt;ECFP\_6:1298504034&gt; ] + 0.3355 \* [ Count&lt;ECFP\_6:858184972&gt; ] + 0.3348 \* [ Count&lt;ECFP\_6:-932108170&gt; ] - 1.037 \* [ Count&lt;ECFP\_6:292958156&gt; ] - 0.2933 \* [ Count&lt;ECFP\_6:-756348342&gt; ] - 0.5317 \* [ Count&lt;ECFP\_6:-1102925512&gt; ] - 0.4482 \* [ Count&lt;ECFP\_6:-757679000&gt; ] - 0.3015 \* [ Count&lt;ECFP\_6:-1332781180&gt; ] + 0.007167 \* [ Count&lt;ECFP\_6:865482986&gt; ] - 1.887 \* [ Count&lt;ECFP\_6:-178525456&gt; ] + 2.371 \* [ Count&lt;ECFP\_6:-666950485&gt; ] + 1.096 \* [ Count&lt;ECFP\_6:-103562730&gt; ] + 0.3359 \* [ Count&lt;ECFP\_6:-1302110264&gt; ] + 0.3366 \* [ Count&lt;ECFP\_6:1095683433&gt; ] + 0.3368 \* [ Count&lt;ECFP\_6:-768126022&gt; ] + 0.3339 \* [ Count&lt;ECFP\_6:2007300961&gt; ] + 0.3338 \* [ Count&lt;ECFP\_6:397284699&gt; ] + 0.3349 \* [ Count&lt;ECFP\_6:1451403962&gt; ] + 0.3348 \* [ Count&lt;ECFP\_6:2071685859&gt; ] + 0.6994 \* [ Count&lt;ECFP\_6:-658363709&gt; ] - 1.242 \* [ Count&lt;ECFP\_6:-1950934120&gt; ] + 0.3364 \* [ Count&lt;ECFP\_6:1155958977&gt; ] + 0.3363 \* [ Count&lt;ECFP\_6:-952707428&gt; ] + 0.3349 \* [ Count&lt;ECFP\_6:-1278685991&gt; ] + 0.3359 \* [ Count&lt;ECFP\_6:1079175434&gt; ] + 0.3368 \* [ Count&lt;ECFP\_6:-2135040425&gt; ] + 1.166 \* [ Count&lt;ECFP\_6:-1059365320&gt; ] + 0.8697 \* [ Count&lt;ECFP\_6:-857146788&gt; ] - 0.8452 \* [ Count&lt;ECFP\_6:-1683911134&gt; ] + 0.6952 \* [ Count&lt;ECFP\_6:661073749&gt; ] - 0.8734 \* [ Count&lt;ECFP\_6:864518973&gt; ] + 2.252 \* [ Count&lt;ECFP\_6:-1686976258&gt; ] + 0.3479 \* [ Count&lt;ECFP\_6:1564392544&gt; ] + 0.1261 \* [ Count&lt;ECFP\_6:1571214559&gt; ] - 8.271e-002 \* [ Count&lt;ECFP\_6:191790798&gt; ] + 1.464 \* [ Count&lt;ECFP\_6:-177264675&gt; ] - 2.462 \* [ Count&lt;ECFP\_6:859796174&gt; ] - 0.837 \* [ Count&lt;ECFP\_6:-1952026932&gt; ] - 2.36 \* [ Count&lt;ECFP\_6:-1506130950&gt; ] + 1.878 \* [ Count&lt;ECFP\_6:1670941296&gt; ] - 1.776 \* [ Count&lt;ECFP\_6:-1884411803&gt; ] - 1.475 \* [ Count&lt;ECFP\_6:-1487746661&gt; ] + 1.205 \* [ Count&lt;ECFP\_6:-572965350&gt; ] - 0.5657 \* [ Count&lt;ECFP\_6:-1867561664&gt; ] - 0.7463 \* [ Count&lt;ECFP\_6:864909220&gt; ] + 1.442 \* [ Count&lt;ECFP\_6:1021725999&gt; ] + 0.3452 \* [ Count&lt;ECFP\_6:-740847217&gt; ] + 1.671 \* [ Count&lt;ECFP\_6:1595399376&gt; ] - 0.4076 \* [ Count&lt;ECFP\_6:515773057&gt; ] - 1.199 \* [ Count&lt;ECFP\_6:1338334141&gt; ] - 0.4369 \* [ Count&lt;ECFP\_6:-408473190&gt; ] - 1.817e-002 \* [ Molecular\_Weight ] + 0.004699 \* [ Num\_AromaticRings ] + 0.4961 \* [ Num\_H\_Acceptors ] + 0.6824 \* [ Num\_H\_Donors ] - 0.4032 \* [ Num\_Rings ] - 7.2e-002 \* [ Num\_RotatableBonds ] - 8.947 \* [ Molecular\_FractionalPolarSurfaceArea ]



Y=0.943x+0.340 R-Square=0.943

Y=0.273x+4.494 R-Square=0.016

**讨论：**

模型精度的影响因素、化合物结构分析等