## 基于Python的chembl库靶点活性小分子数据处理及筛选 应用案例简介

1. 这是一个用Python处理实验数据并用于虚拟筛选的Python应用案例。

2. 该应用案例所研究的对象是在chembl数据库网站上的靶点活性小分子数据库。实验目的是为了区分分子的active和inactive。

Cheml数据库是欧洲生物信息研究所(EuroeaBioiformaticItitute,EBI)开发的一个在线的免费的活性小分子数据库。它通过从大量文献中收集各种靶点及化合物的生物活性数据，为科学家们提供了一个非常便利的信息查询平台。目前，该数据库共收集了1.2万个靶点，180.0万个化合物，共有110万条生物活性信息。

该案例所处理的数据的下载方式详看本文件夹中给出的ppt文档：ChEMBL上下载数据.pptx

3. 该应用案例涉及到为约5000多个分子计算分子描述符。没有做并行，所以算起来会比较慢。

4. 该应用案例的文件1\_trainModel.ipynb 和2\_screening.ipynb，这两个 后缀名为.ipynb, 这是使用jupyter notebook编写的python脚本文件，需要大家安装Anaconda并使用该套件中的工具：jupyter notebook打开。

为了大家使用方便，已将这两个文件转换成Pycharm中能打开的.py后缀名文件：1\_trainModel.py 和 2\_screening.py

5. 该应用案例的1\_trainModel.ipynb文件需要第三方库：pandas

6. 该应用案例的2\_screening.ipynb文件需要第三方库rdkit； mordred； molvs ；cPickle

7. 所有第三方库的安装版本在文档 requirments.txt中指定了。使用如下方法安装txt文件中配置好版本的库包：  
pip install -r requirements\_my\_project.txt

如在教育网，可使用清华镜像资源安装：

pip3 install -i https://pypi.tuna.tsinghua.edu.cn/simple -r requirements\_my\_project.txt

1. 该案例仅做教学使用，请勿在网上传播，谢谢！
2. 两个原始csv文件的都包含表头。

本文档撰写人：王芬

邮箱：wangfen@hust.edu.cn