CADD408

数据清洗

2023年4月13日

文档属性

|  |  |
| --- | --- |
| 属性 | 内容 |
| 文档主题 | 数据查询简介 |
| 文档副标题 |  |
| 文档版本 | 0.4 |
| 文档日期 | 2023.04.13 |
| 文档状态 | 草案 |
| 作者 | 李果 |

文档变更

|  |  |  |  |
| --- | --- | --- | --- |
| 版本 | 修订日期 | 修订人 | 描述 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

1.数据查询简介0.5.docx 收集数据

2.使用方方格子合并处理数据

赵云阳买过正版，盗版比较好用，不过现在已经不支持扩展包，可以安装好了去实验室电脑里找，安装了扩展包的，复制粘贴即可本体够用。



红色部分为扩展包

3.使用wash\_toYang进行数据清洗

具体使用自己调试，代码为殷红岩师姐所写

包括标准化，去离子什么的

1. 关于代码环境配置问题

可见00Python安装

已经配置好的环境可用CADD教学论文配置的环境

<https://projects.volkamerlab.org/teachopencadd/all_talktorials.html>

1. 关于编码问题

大多数电脑是utf-8，但也有其它编码，比如我的是国标，因此代码会有encoding="gbk"，这与你的系统有关

6.抑制剂数据库的清洗

统一数据库中使用的单位为 nM，将 SMILES 结构式标准化，去除含水含溶剂分子中的水或溶剂，并删除重复的抑制剂，返回原始文献调整该操作中报错的分子，删除了分子量 900 以上的分子，原因是这些大分子配体为由长链连接的孪生分子，其大概率发生折叠进入口袋与受体发生完全不一样的结合，清洗含有盐离子或溶剂的化合物。由于数据库中存在区间数据（interval data），即在文献中没有给出具体的值，只知道变量的取值范围而不知道确切的取值。核对全部数据，如一化合物同时存在区间数据与精确数据，则删除区间数据，对于其它相同结构式不同活性的化合物，如其重复次数超过 3 次的化合物则取活性值的中位数，对于重复次数小于 3 次的化合物取活性值的最小值。IC50值的大小从 0.00026 nM 到68000 nM，横跨 8 个数量级，如图 2-2，同时计算化合物的多项理化性质，诸如分子量、脂水分配系数、重原子数、氢键供体数、氢键受体数等，记录边界点的化合物，保证其在之后的训练集中。

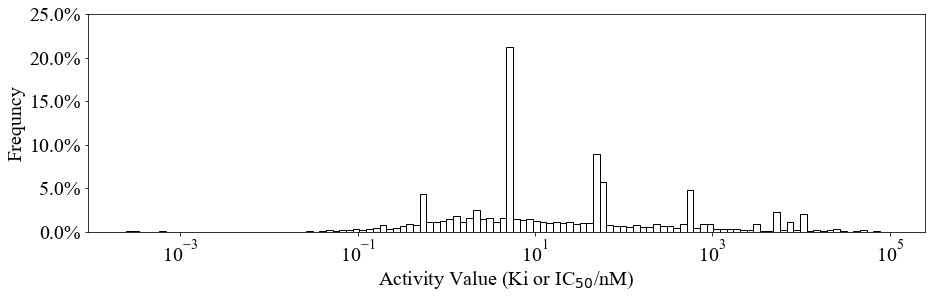


图 2-2 7446个 BTK 抑制剂 IC50值分布直方图（计算化学空间相似性.ipynb）

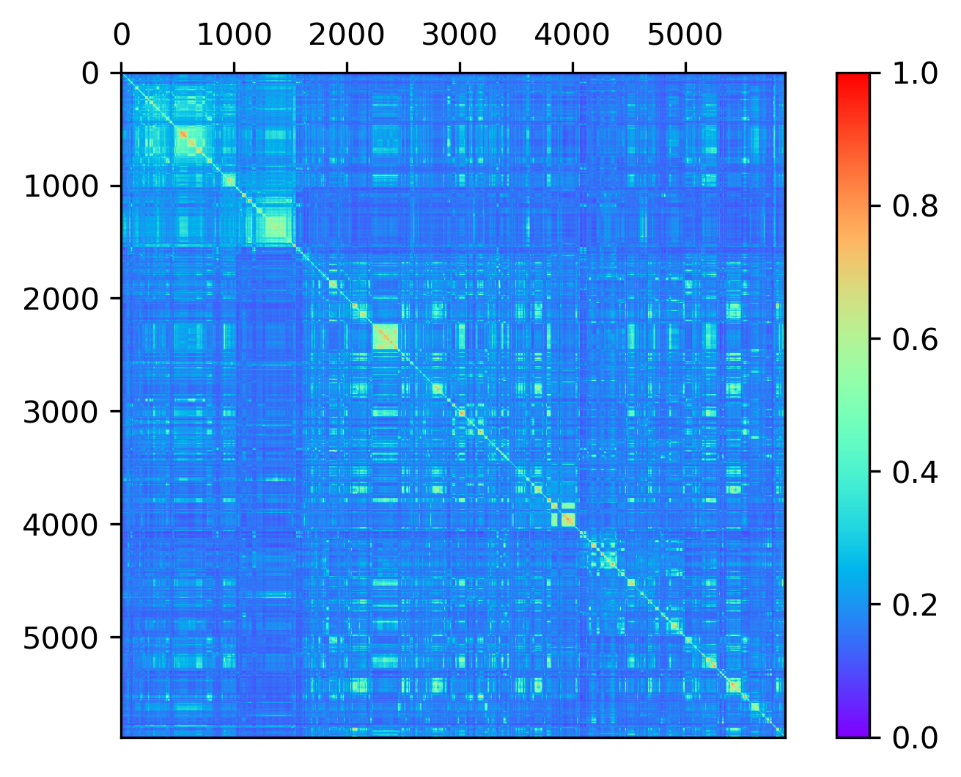


图 2-3 数据库中的化合物基于1024位 ECFPs 的 Tanimoto相似性的化学多样性热图（计算化学空间相似性.ipynb）

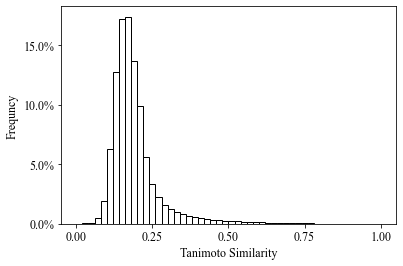
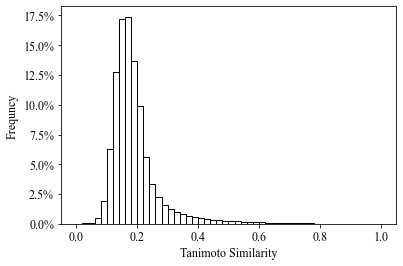


图 2-4 基于1024位 ECFP 的 Tanimoto 系数频率分布直方图

（a）原始数据集（b）不含子结构数据集（计算化学空间相似性.ipynb）

这三个图都用计算化学空间相似性.ipynb跑

我做的比较复杂，

不可忽视的步骤其实是：

首先跑wash\_toYang

其次跑01去除手性、顺反02删除截断数据，这俩也可以不跑，看你自己的想法。

删除了分子量 900 以上的分子，原因是这些大分子配体为由长链连接的孪生分子，其大概率发生折叠进入口袋与受体发生完全不一样的结合（这个操作可以自己定）

最后是算一个边界化合物

从图 2-5 可以看出，训练集化合物所在的化学空间较好的覆盖了测试集的化合物。脂水分配系数与分子量最大的要在训练集里，你可在这一步就把这些分子挑出来。

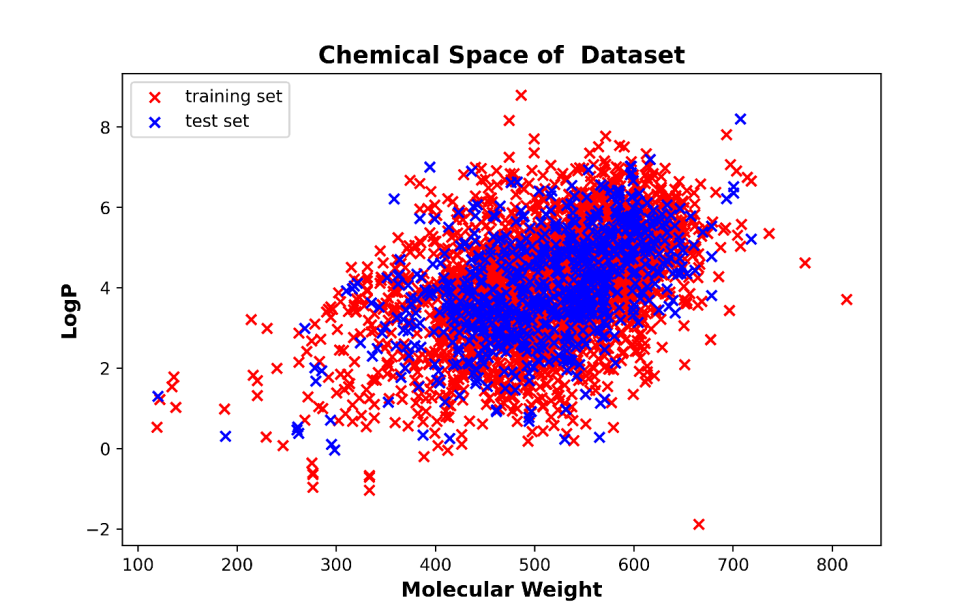


图 2-5 训练集和测试集中化合物的分子量（Weight）和正辛醇/水分配系数（XlogP）散点图（训练集和测试集中化合物的分子量（Weight）和正辛醇水分配系数（XlogP）散点图.py）

7.子结构划分

这个可不跑，这个是我自己课题的东西

1. 分箱 阈值讨论 跑03分箱.py即可