CADD408

数据查询及数据预处理

2023年4月20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 | 0.6 |
| 文档日期 | 2023.4.13 |
| 文档状态 |  |
| 作者 | 李果 |

文档变更

|  |  |  |  |
| --- | --- | --- | --- |
| 版本 | 修订日期 | 修订人 | 描述 |
| 0.2 | 2021.7.27 | 李果 |  |
| 0.4 | 2021.12.8 | 李果 |  |
| 0.6 | 2023.4.13 | 李果 |  |
|  |  |  |  |

[1. 定量构效关系（Quantitative Structure - Activity Relationship）](#_Toc14564)

[2. 数据类型](#_Toc26513)

[2.1. 1.大分子信息（靶点信息）](#_Toc17227)

[2.1.1. 蛋白质晶体结构（如果存在晶体结构）](#_Toc23624)

[2.1.2. 蛋白质的序列](#_Toc8837)

[2.2. 2.小分子信息（抑制剂）](#_Toc6947)

[2.2.1. 小分子与靶点的结合数据](#_Toc4726)

[2.2.2. 小分子的结构](#_Toc32620)

[3. 大分子数据--晶体结构（PDB网站）](#_Toc7276)

[4. 大分子数据—序列信息（NCBI网站）](#_Toc17333)

[5. uniorot数据库的使用](#_Toc5418)

[6. 蛋白质和小分子化合物描述符](#_Toc17485)

[7. 小分子结合数据（chembl）](#_Toc8208)

[8. 小分子结合数据（pubchem）](#_Toc5681)

[9. 小分子结合数据（BindingDB）](#_Toc2240)

[10. 小分子结合数据（Reaxys）](#_Toc7857)

[11. 其它数据库](#_Toc3760)

[12. 谷歌专利没有的专利下载](#_Toc31018)

[13. 数据预处理的脚本处理步骤](#_Toc23177)

[14. 数据预处理](#_Toc9078)

# 定量构效关系（Quantitative Structure - Activity Relationship）

定量构效关系(QSAR)是一种借助分子的理化性质参数或结构参数，以数学和统计学手段定量研究有机小分子与生物大分子相互作用、有机小分子在生物体内吸收、分布、代谢、排泄等生理相关性质的方法。

# 数据类型

## 1.大分子信息（靶点信息）

### 蛋白质晶体结构（如果存在晶体结构）

### 蛋白质的序列

## 2.小分子信息（抑制剂）

### 小分子与靶点的结合数据

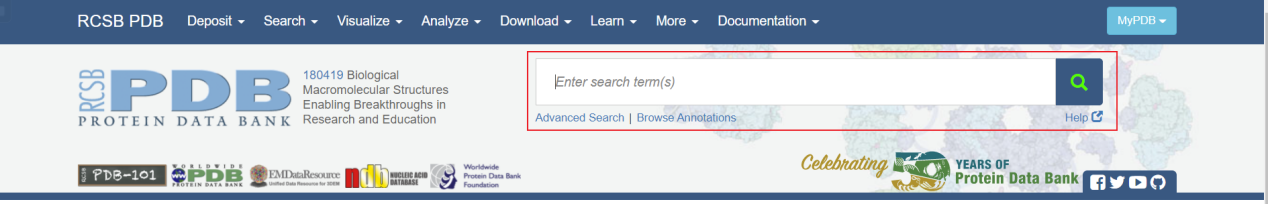
### 小分子的结构

# 大分子数据--晶体结构（PDB网站）

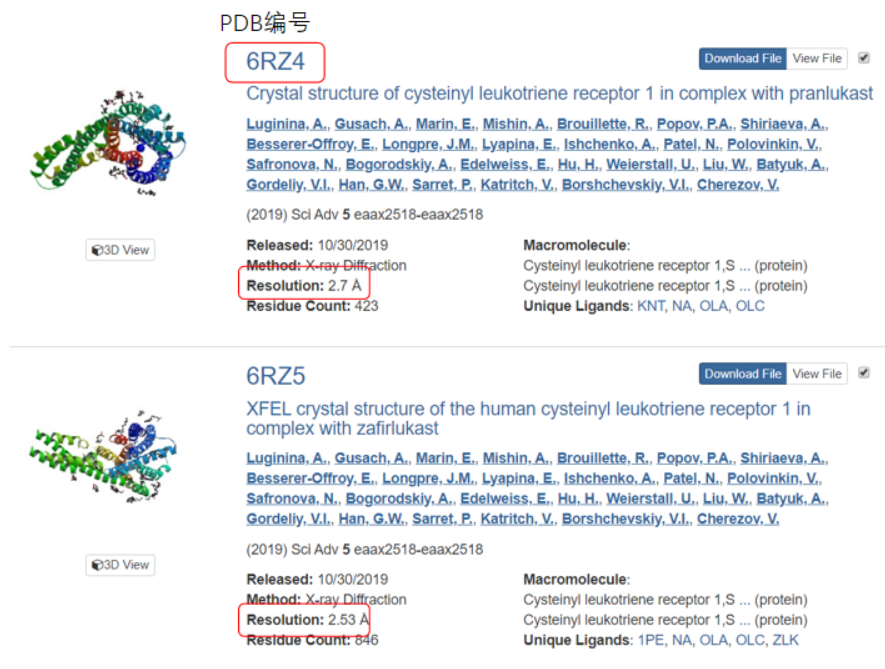
<http://www.rcsb.org/>



输入靶点

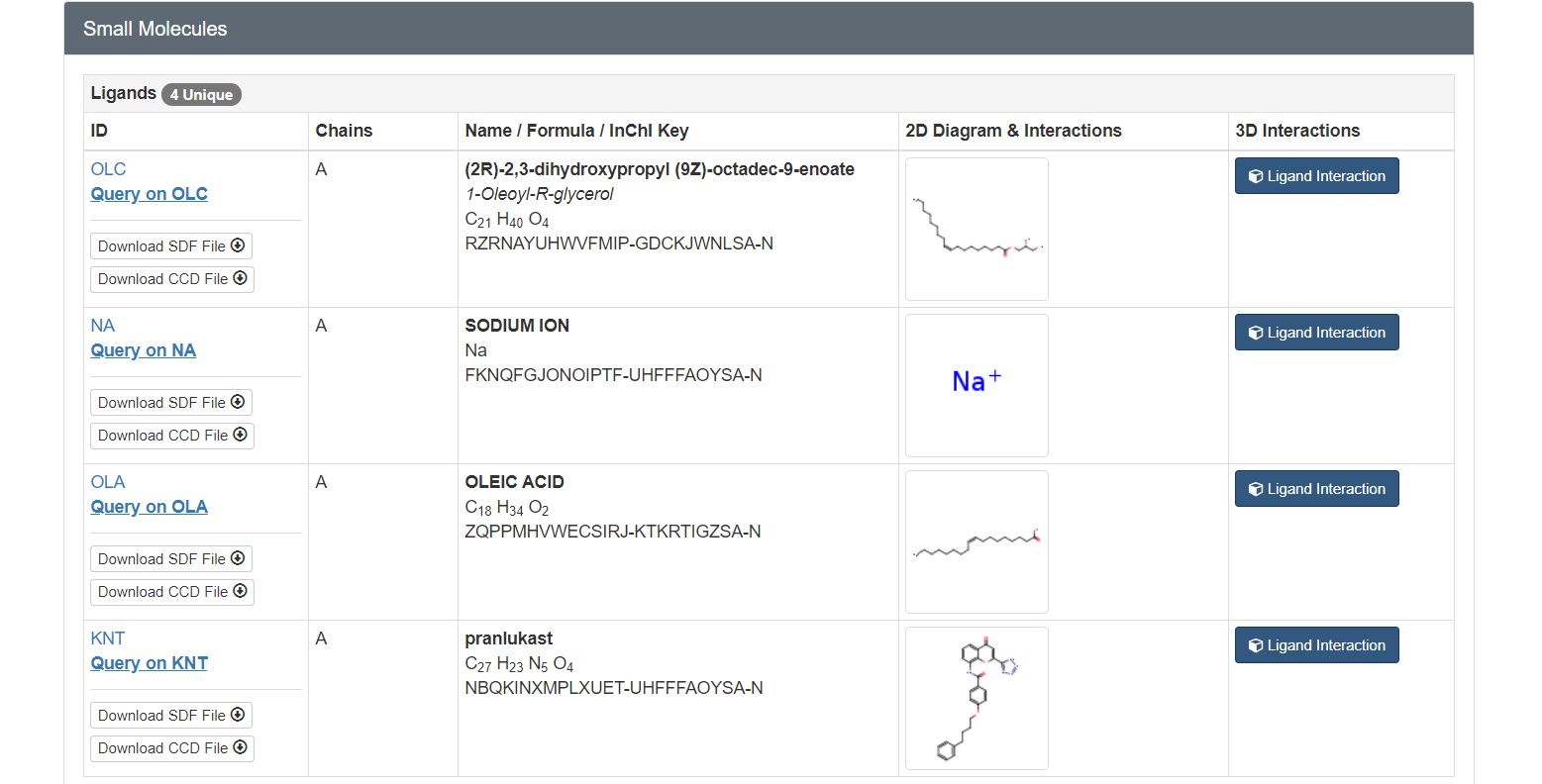


**细节：**分辨率越小越好，尽量选择3.0Å以下的,小于2Å更好，这里的Å念ai，表示10的-10次方，这里用作表示误差



要看蛋白名称Molecule是否为该靶点，Organism是否人源

要看PDB是否显示IC50，才能确定是否为活性口袋

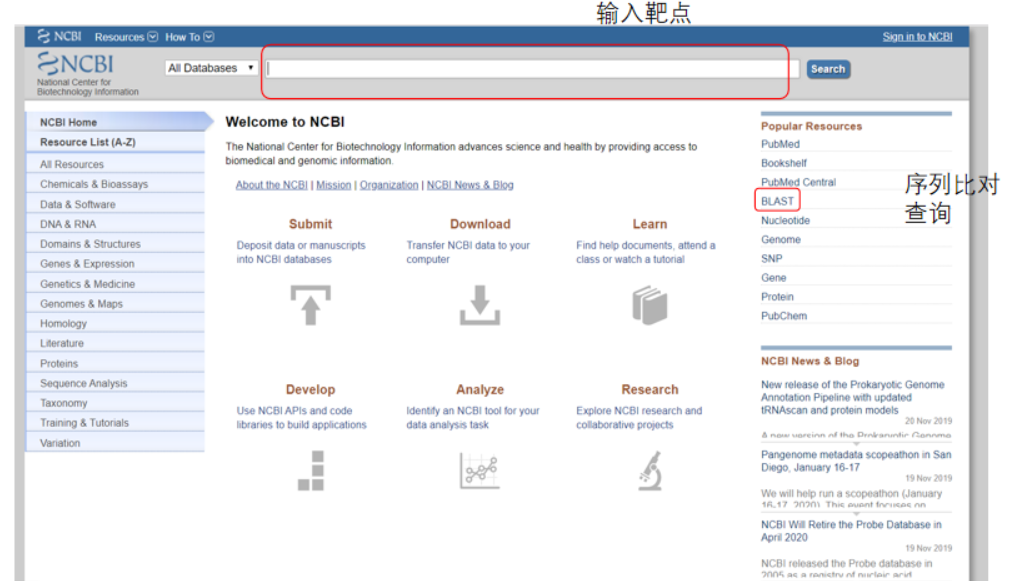


一般选择最大的那个,下载打开后可以查看对应的氨基酸序列

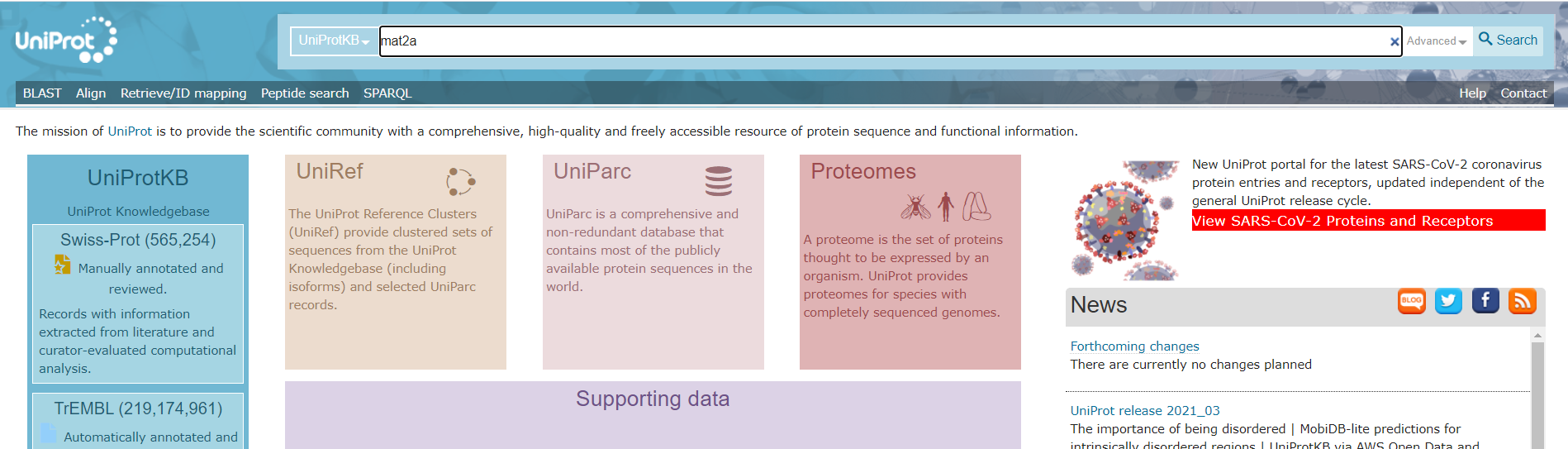
如果有Entity ID: 2则为复合物，会有其它多肽

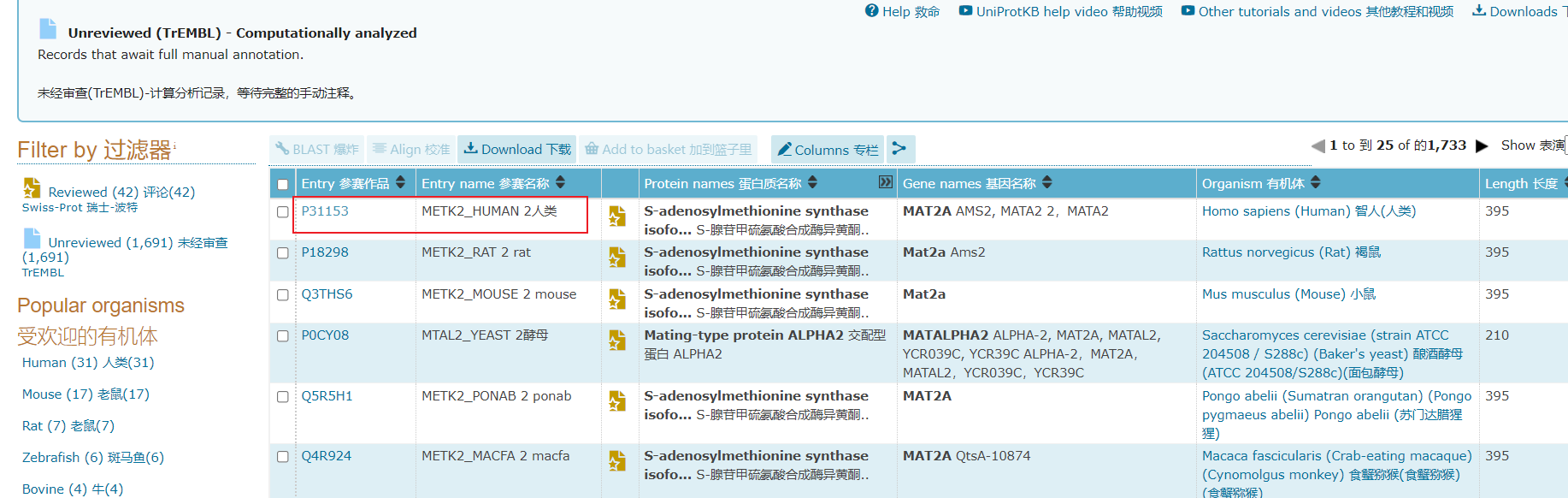
去水，关注蛋白与配体

# 大分子数据—序列信息（NCBI网站）



# uniorot数据库的使用





这些都是需要查看的

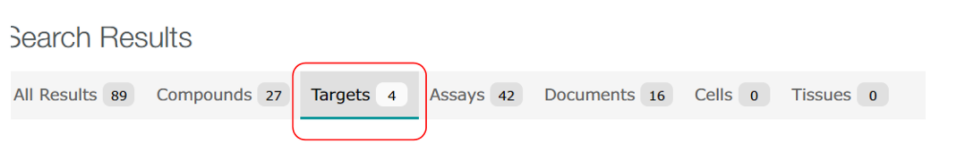
氨基酸序列sp表示人工校正后

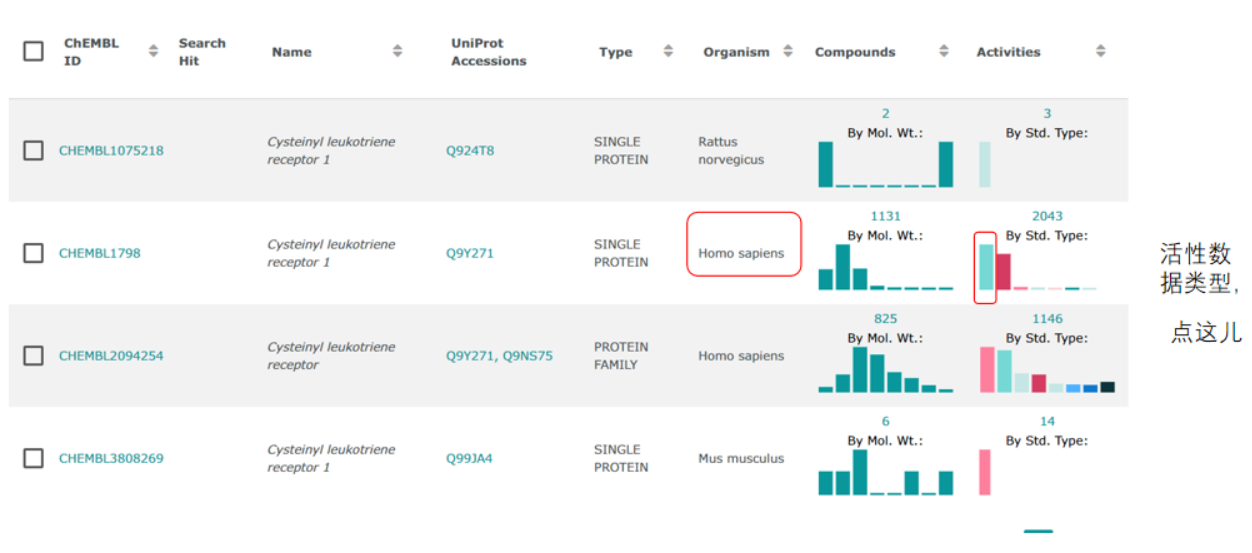
# 蛋白质和小分子化合物描述符

http://home.scbdd.com/index.php?s=/Home/Article/detail/id/10.html

# 小分子结合数据（chembl）

<https://www.ebi.ac.uk/chembl/>

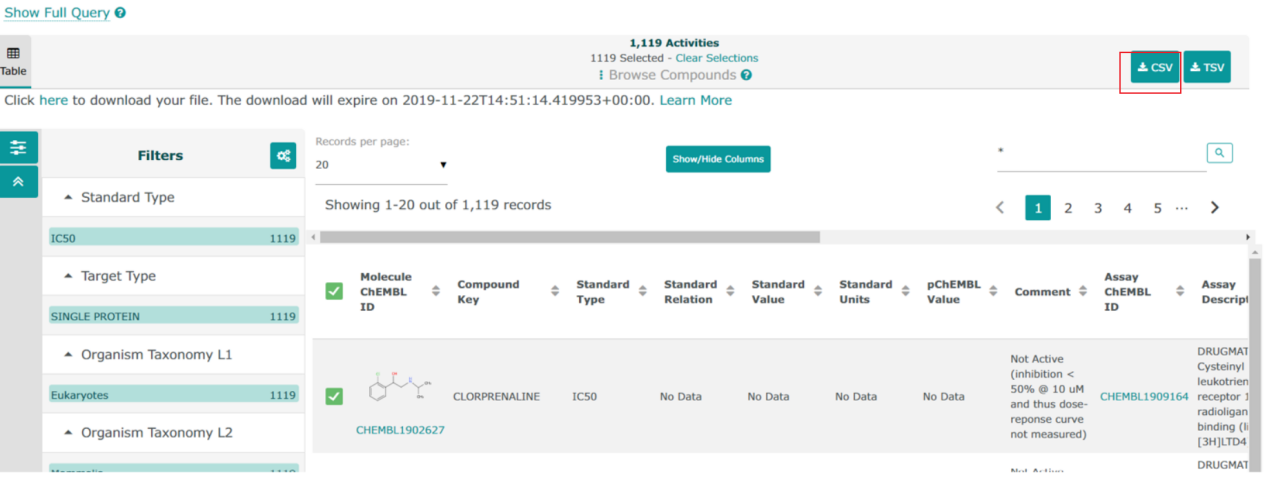




一般选择IC50，因为是二分类问题，也可以选择不同的标准混搭

JCI\_2020\_ADMET2.0

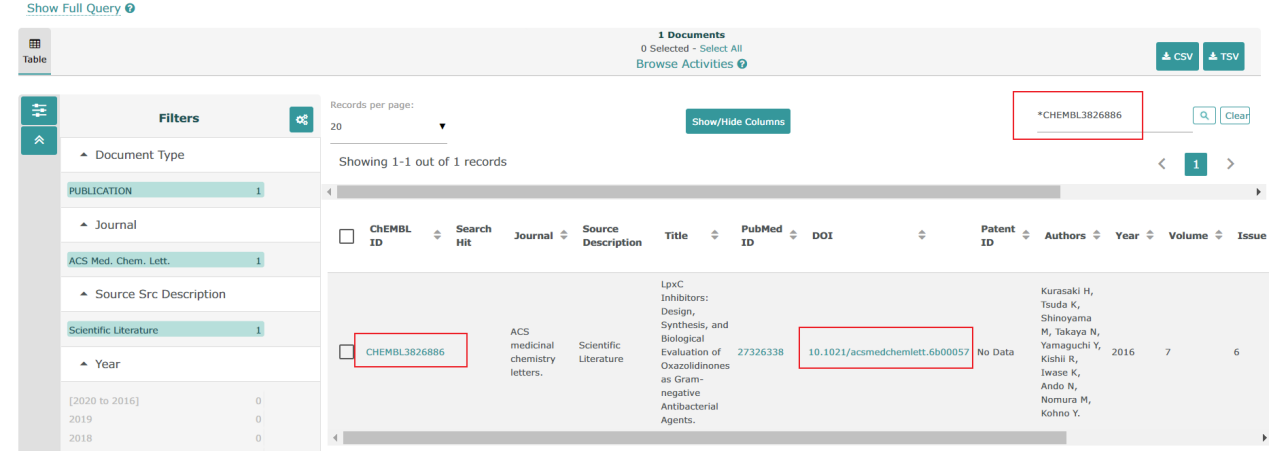
先下载所有数据，后期我们再对数据进行核对,这里可以选择show columns把文献杂志名与文献年限下载



下载的是csv分列后对表格进行整理

保留Molecule ChEMBL ID，Assay ChEMBL ID，Smiles，Standard Relation，Standard Value，Standard Units，Assay Description

1. 通过这个Assay ChEMBL ID（Document ChEMBL ID），返回原文献，找到原化合物进行比对
2. 查看Assay Description，对照Assay ChEMBL ID,这个步骤可以下载全部文献，用vlookup替换为doi，然后用python下载。（见3）

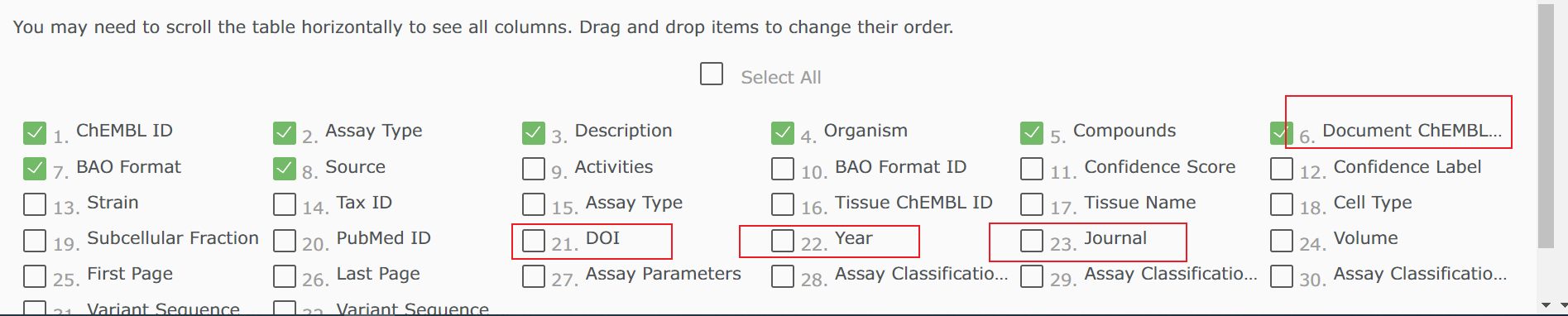


1. 进入靶点









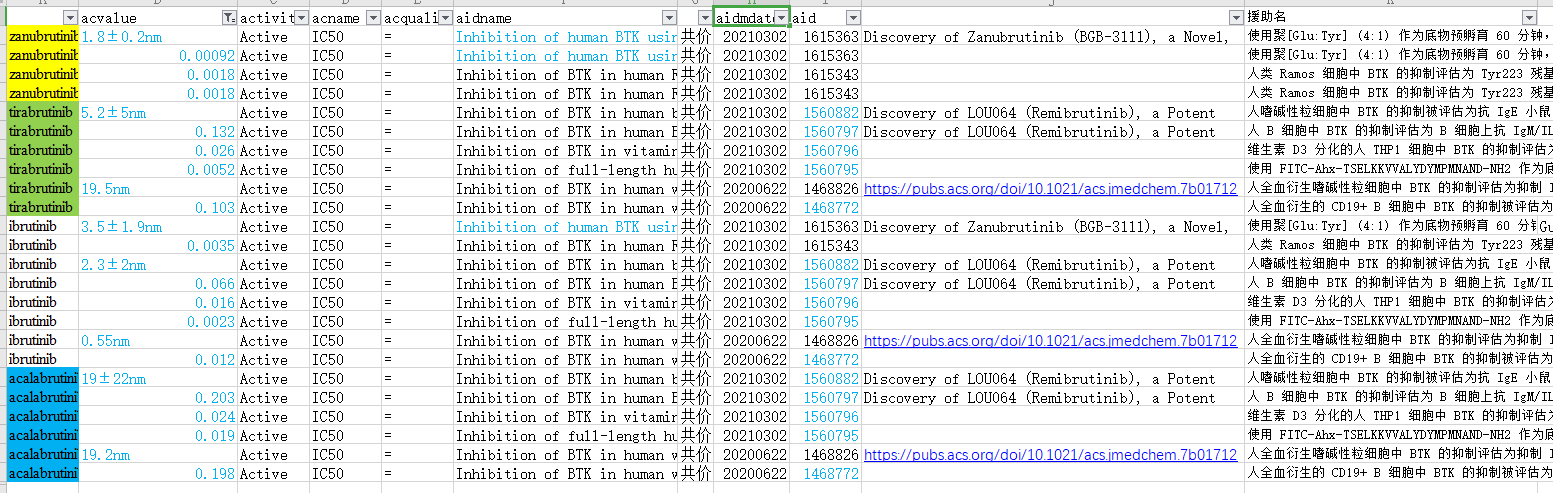
即可下载对应doi

Smiles 转2d

1. *# -\*- coding: utf-8 -\*-*
2. from rdkit.Chem import PandasTools as pt
3. from rdkit import Chem
4. import pandas as pd
5. input\_path = r'D:\Users\79864\Desktop\work\抗生素调研\数据收集\1.csv'
6. save\_path = r'D:\Users\79864\Desktop\work\抗生素调研\数据收集\Pseudomonas aeruginosa - 副本\_2D.csv'
7. test = pd.read\_csv(input\_path)
8. test['ROMol'] = test['Smiles'].apply(lambda x: Chem.MolFromSmiles(x))
9. pt.SaveXlsxFromFrame(test, save\_path, molCol='ROMol', size=(300, 400))

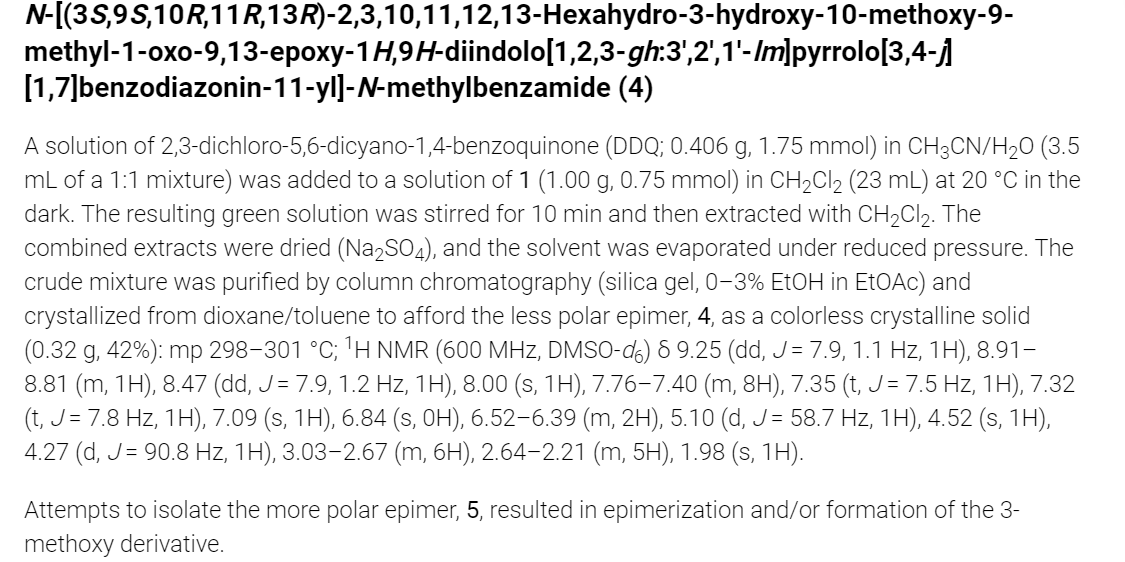
# 小分子结合数据（pubchem）

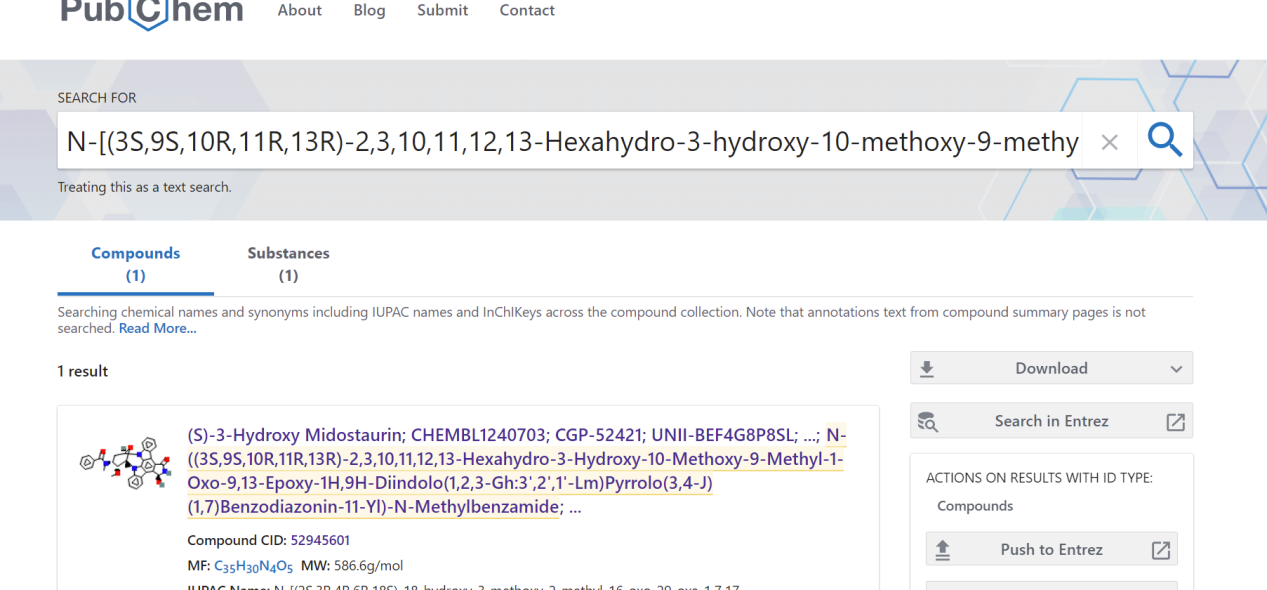
<https://pubchem.ncbi.nlm.nih.gov/>



保留aidmdate

pubchem还可以查询化合物名，比如文献出现这样的,如何对照smiles

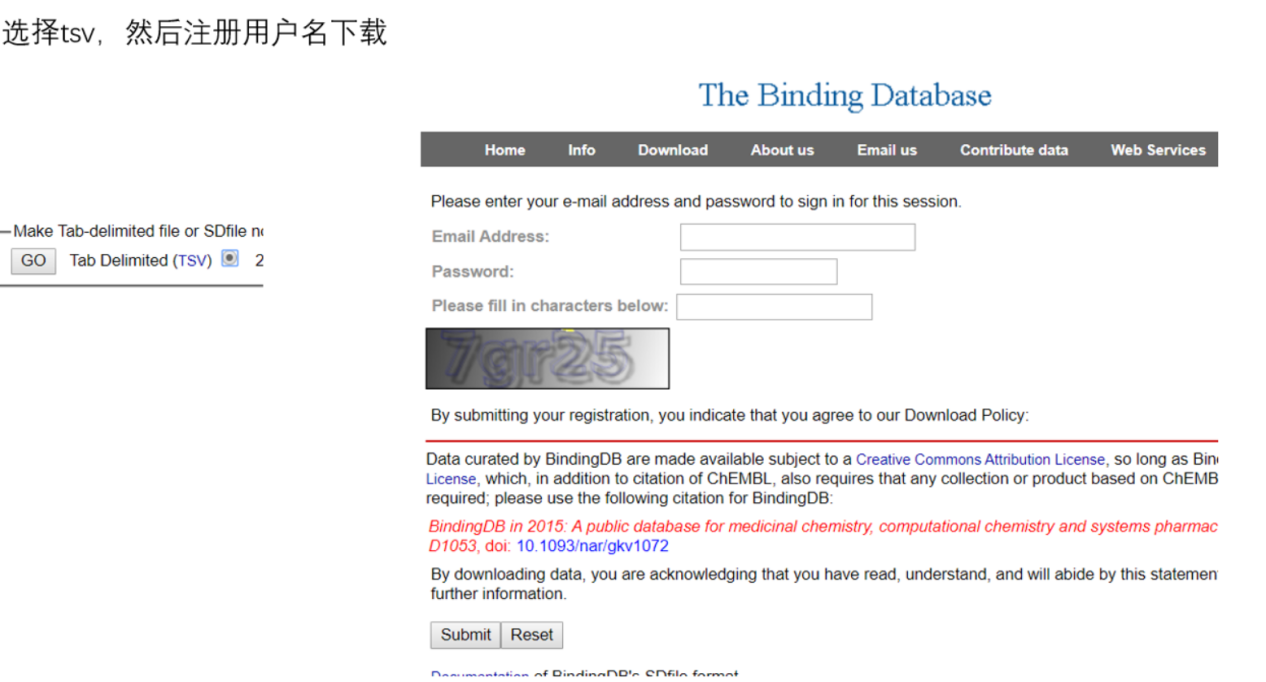




# 小分子结合数据（BindingDB）

<http://www.bindingdb.org/bind/index.jsp>



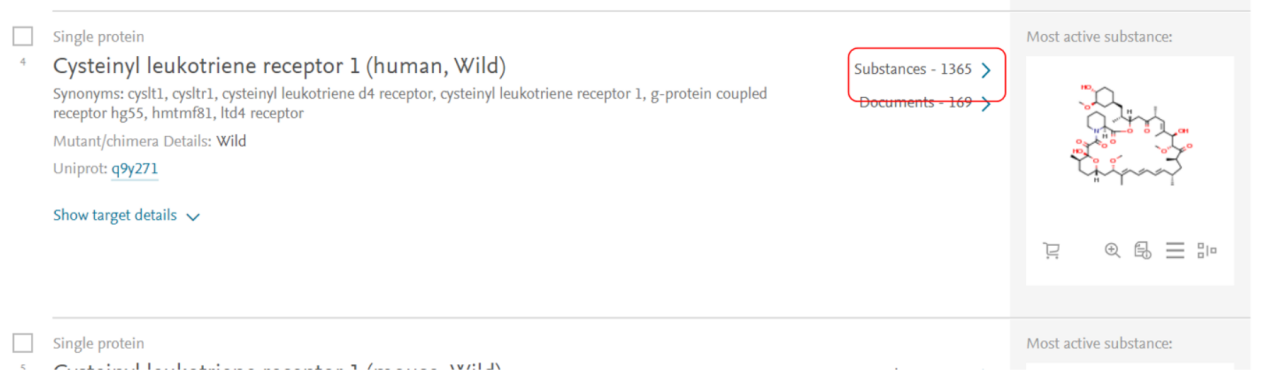


处理方式和chembl一样，返回原始文献核对数据。但是bindingdb没有测活的方法，可以后期手动添加

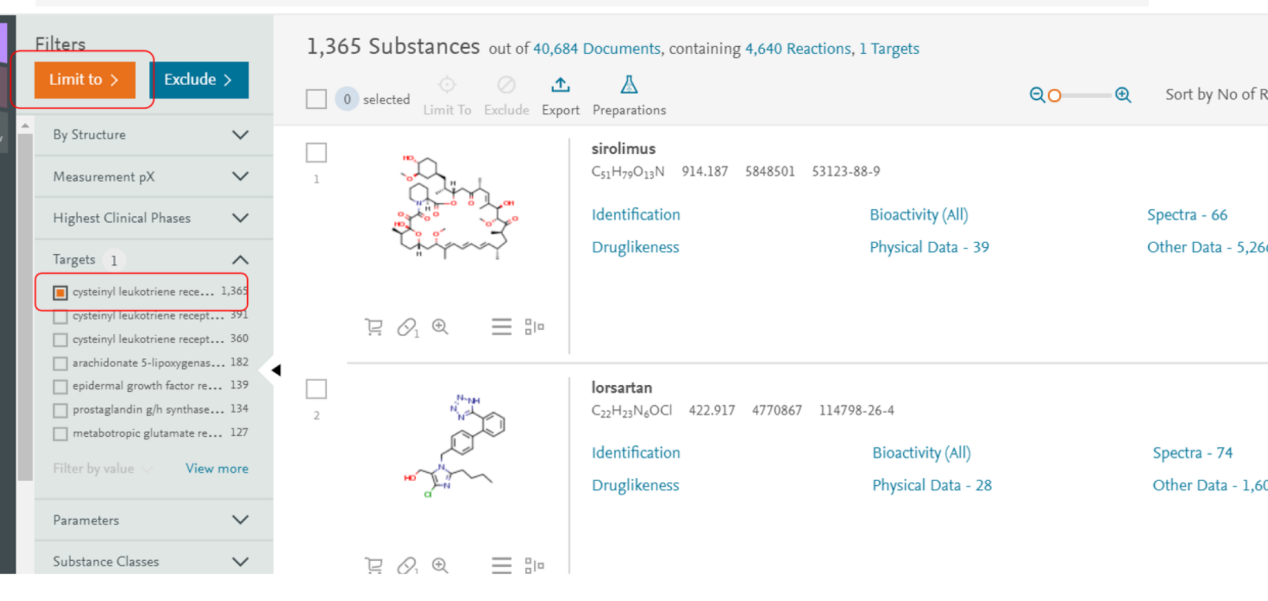
# 小分子结合数据（Reaxys）

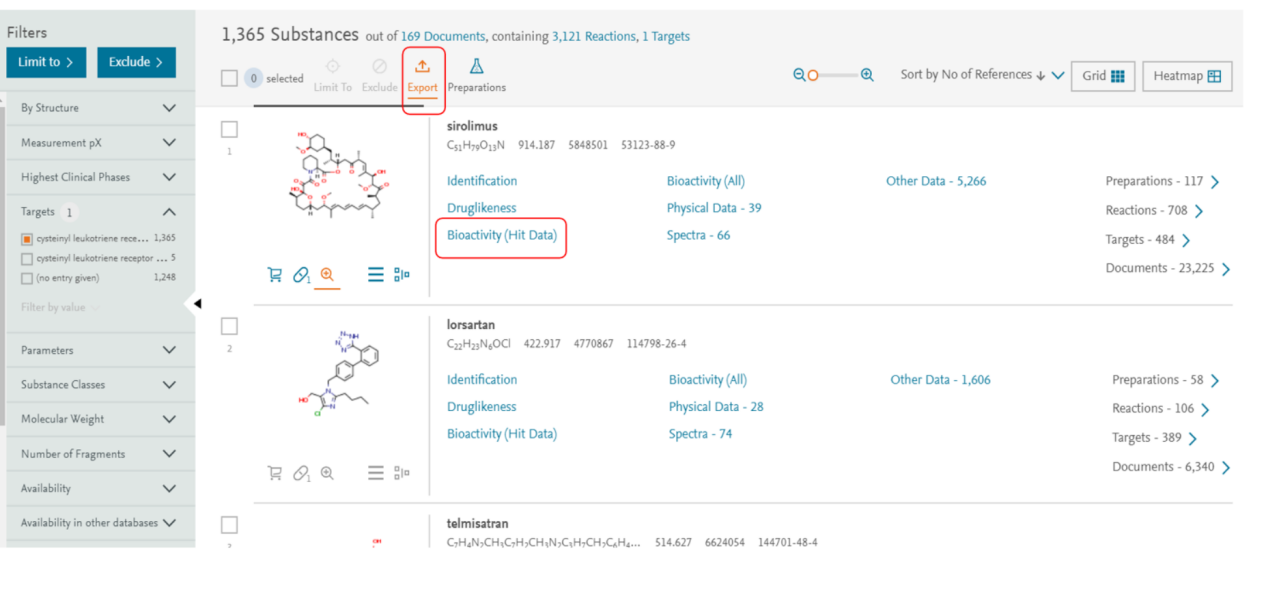
例子1

第二个选项



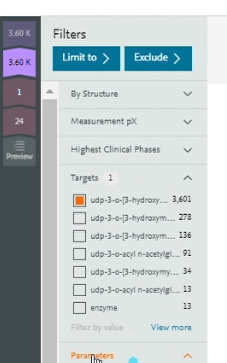
这里一定要再选一次靶点，直到document小于substance

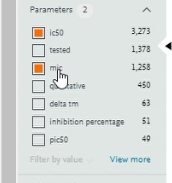


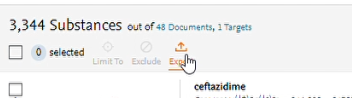


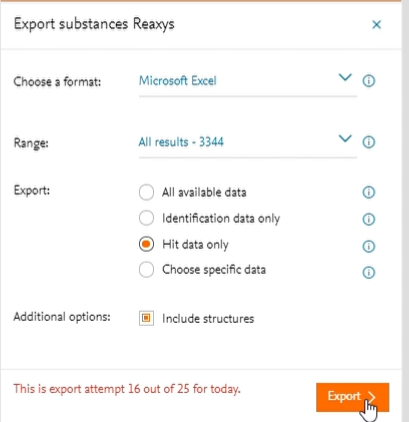
例子2 要什么就点什么不要多点，你全选反而下载不了全选的东西











Format选择Excel，将以上3个数据库合并一共只保留3列（分类模型），smiles,ic50,以及化合物的在各个数据库的编号。

定量模型保留4列，多添加一列测活方法。

去重思路：

1.去掉相同一模一样的SMILES

2.同一个化合物经过不一样的软件生成的smiles有可能不相同，所以先用rdkit读取所有化合物，然后输出统一格式的smiles，删除一样的smiles，如果不同化合物库的不一样，返回原文献进行比对。

# 其它数据库

<https://www.genecards.org/cgi-bin/carddisp.pl?gene=BTK&keywords=BTK>

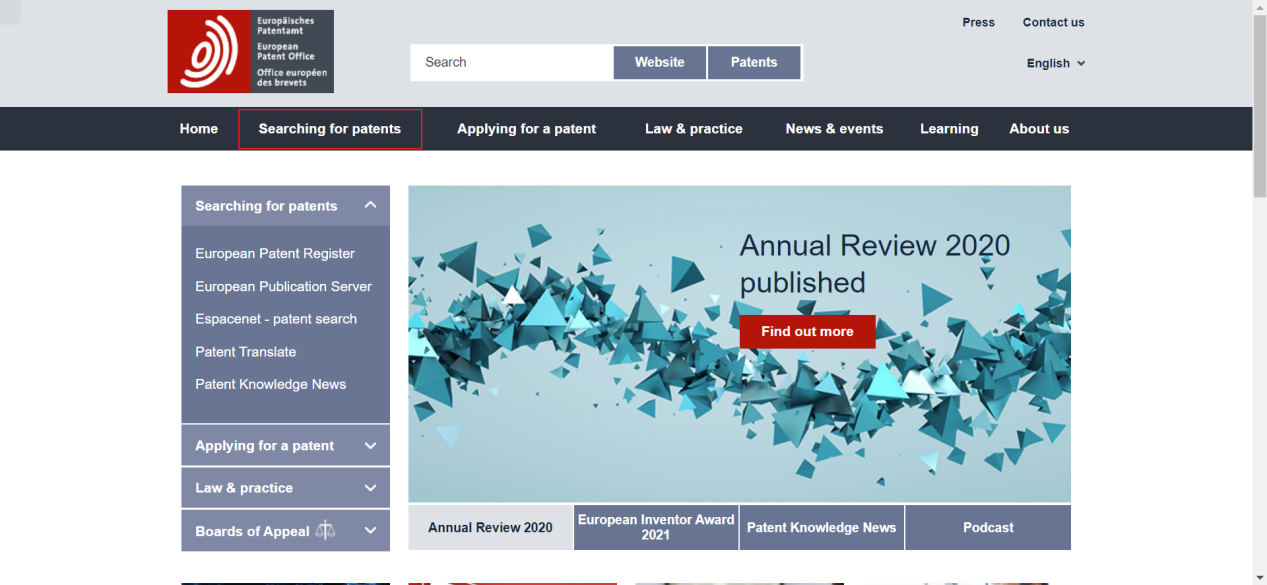
<https://go.drugbank.com/>

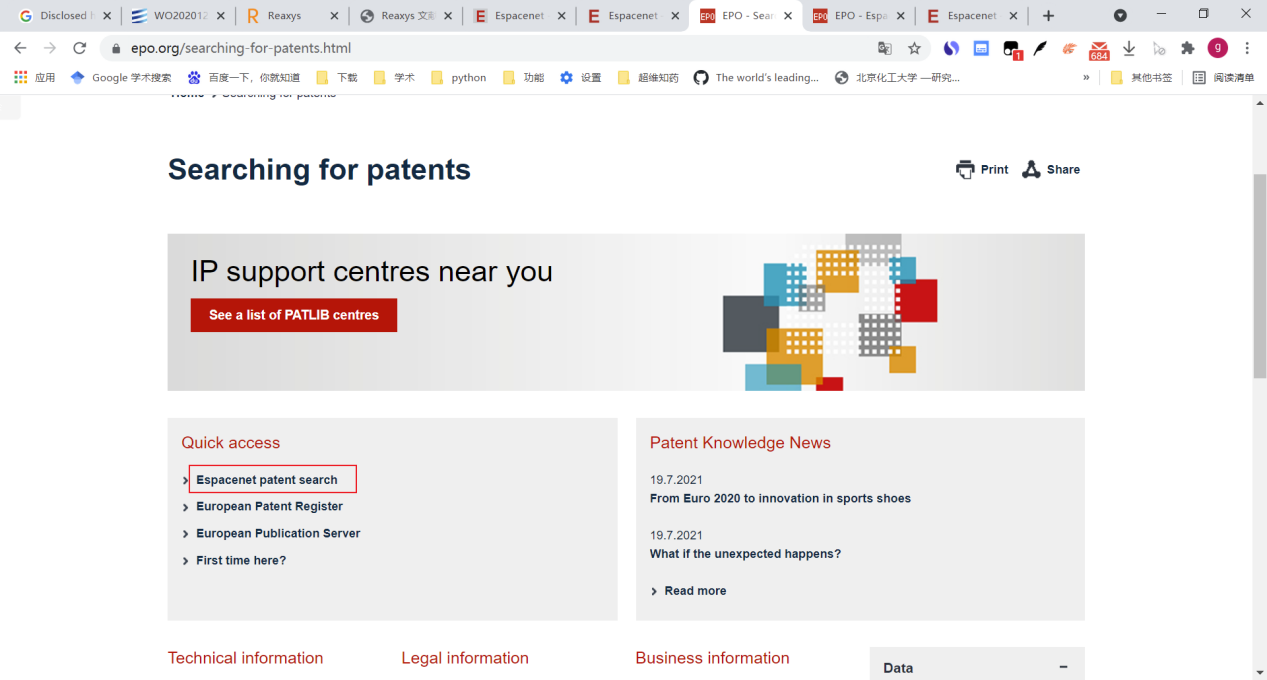
https://www.x-mol.com/

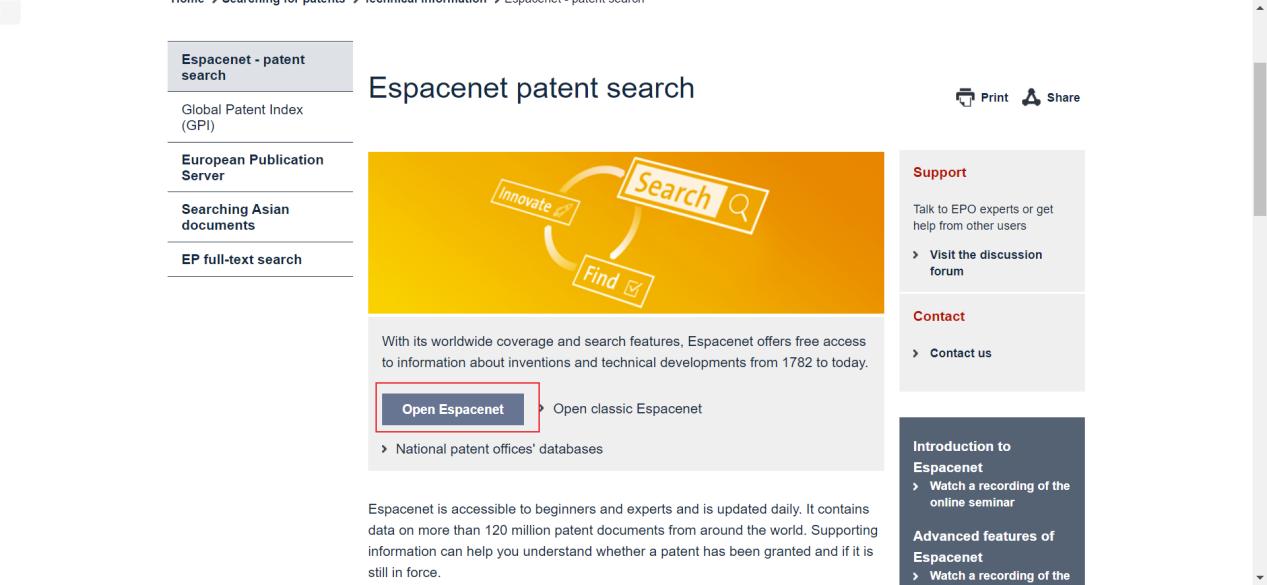
# 谷歌专利没有的专利下载

https://patents.google.com/

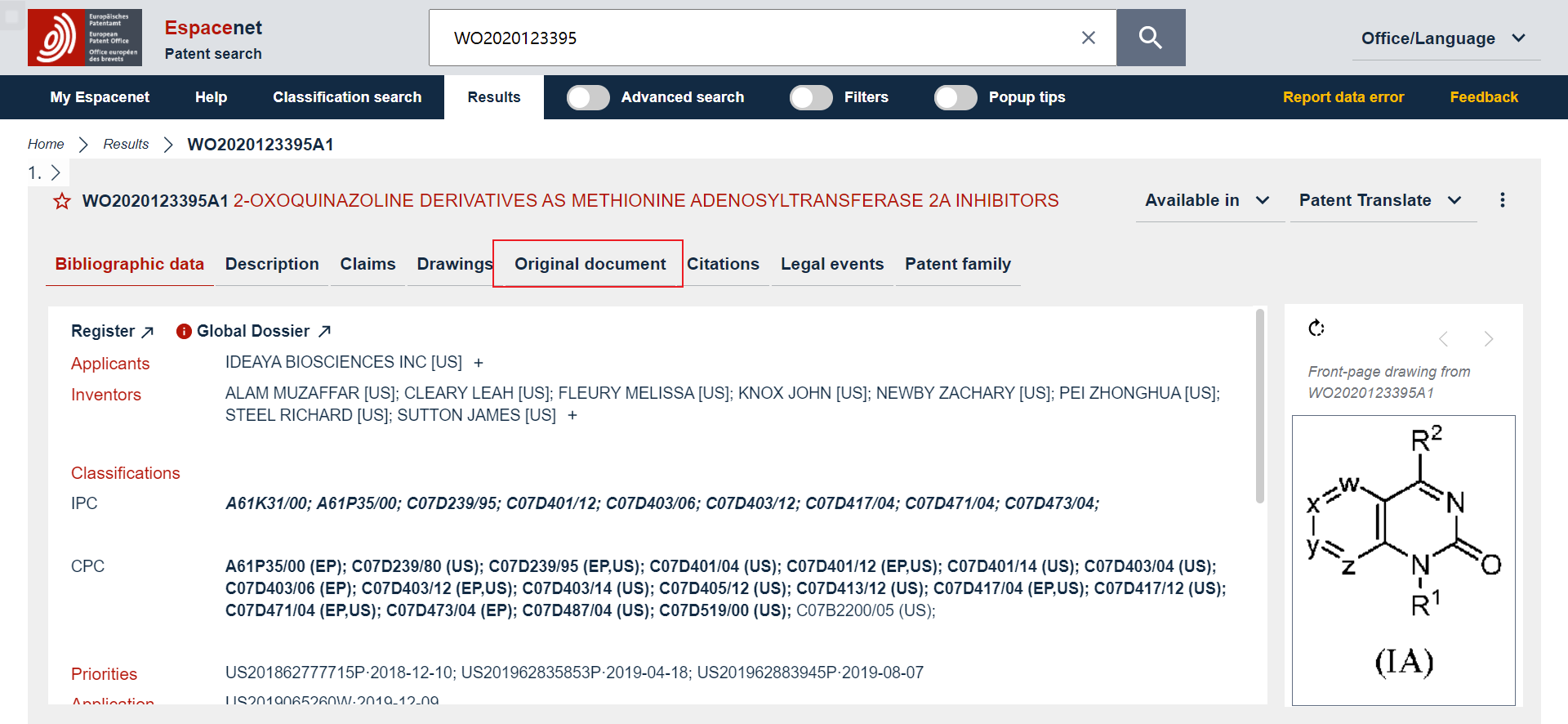
<https://www.epo.org/index.html>

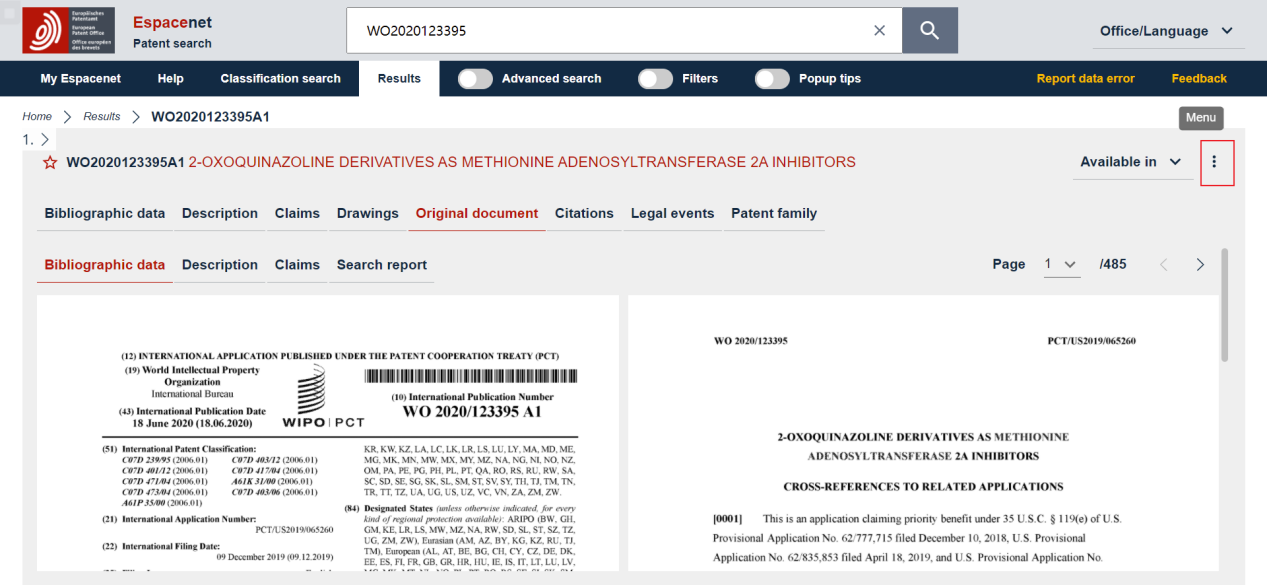






<https://worldwide.espacenet.com/>





Ps 建议下载汉语的同族专利，方便阅读，以同族专利最新的专利为准

<http://www.drugfuture.com/cnpat/cn_patent.asp>

问题：有些专利号检索不到怎么办？

解决：加0

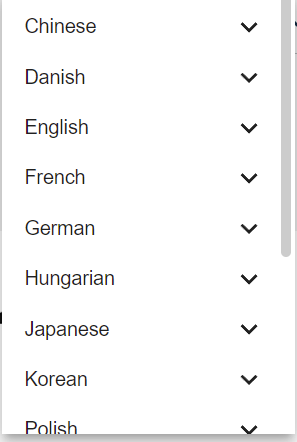
比如

US开头的专利需要在专利/后加0至7到8位，如US2012/82702变为US2012/082702

US2009/82330变为US2009/082330

才能检索到

# 数据预处理的脚本处理步骤

1. 欧洲专利局查找专利号（见上12.谷歌专利没有的专利下载）
2. 看是否有中文同族专利，如果有，迅速浏览查找测活方法、是否人源，如果没有则浏览英文专利
3. 比如这样，英语专利看两wo与us最新的两，一般是一样的，但时间差得久可能不一样，具体意义如下：

S1— 表示专利的状态为已授权。

A代表的是还没授权的出版文本，B代表的是已经授权的版本，具体为：

A1—附有检索报告的欧洲专利申请说明书；

A2—未附检索报告的欧洲专利申请说明书；

A3—单独出版的检索报告。

B1——欧洲专利说明书；

B2——经修改后再次公告出版的欧洲专利说明书；

B3——指经过实质性审查授予专利权的，后经限制性修改程序修改后再次公告出版的欧洲专利说明书。

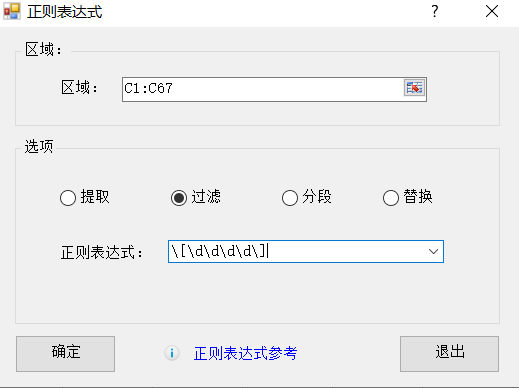
1. 根据专利号去谷歌专利复制为txt，使用正则表达式提取，如常见的：

Example 34

Synthesis of (R)-2-(3-(4-amino-3-(2-fluoro-4-phenoxyphenyl)-1H-pyrazolo[3,4-d]pyrimidin-1-yl)piperidine-1-carbonyl)-4-(4-ethyl-3-oxopiperazin-1-yl)-4-methylpent-2-enenitrile

re.findall('(Example [1-9]\d\*\n.\*?\n.\*?)\n',data)

1. 再打开excel进行具体的正则表达式优化，因为取了三行可能会混入其它的东西，人工统一使用方方格子正则表达式，常见的有





1. 将得到的id，iupac名称两行运行NameToSmiles脚本，得到id与smiles
2. 复制谷歌专利中ic50的表格，简单处理后使用vlookup对应smiles与id还有ic50，也就是使用6步骤得到的作为字典，查找出下载数据中smiles对应的id，再用这个id，以步骤7得到的ic50表格作为字典，得到这个id对应的ic50，判断该ic50是否与数据库中相同，如果相同，则逻辑闭环，可判定为有效，如果有一步不同，根据id查找。
3. Tips

转出的smiles可以高亮重复项，提前看看有没有错，也可以全部转图片看看有木有错，因为有些命名有问题，所以出错概率还是很大的，一半能实现闭环吧。

# 数据预处理

1.smiles列标准化

先将下载的smile列标准化，之后方便处理，调用rdkit即可。可用rdkit的MolFromSMILES检查是否合规。

2.cheml数据的smiles转2d图片生成

需要留出哪些行？

1.总表

专利号、文献doi（导出生成，数据库都支持下载检索编号对应的doi）；化合物数量；备注；测活方法（可预先生成，比如chembl中有一列就是测活方法）；实际专利号（使用最新的的同族专利替换）、文献年限；专利公司、期刊名（vlookup生成）、是否人源。

Ps:冻结首行,专利公司一致，一般测活也一致，可以锁定关键词检索，比如已知公司A的专利1用了电泳，则在公司A的专利2中搜索电泳看是不是电泳。

2.分表

文献中的标号；smiles；备注；活性误差；是否为ic50（Medchem: Measurement Parameter）；单位（Unit）（nM）；大于小于等于（Qualitative value）；活性值（Quantitative value）；文献来源（cheml、reaxys）；DOI

问题

1. 有些图片中带波浪线，是高能键么？

不是， 波浪线表示这个甲基的立体化学结构不明，既可以朝上（实楔线）也可以朝下（虚楔线）。这是立体化学结构的一种表示方法。看成一种键就可以了。

1. 遇到放射性元素怎么办？

删除，去除包含不常见元素的分子（例如包含金属）

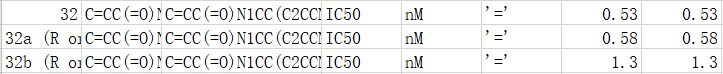
1. 遇到化合物有手性或者没标明？

看是否计算3d描述符，如果不计算3d描述符，则不影响。

附：下面是俩化合物，活性不一致，但是它俩描述符一致，怎么办？

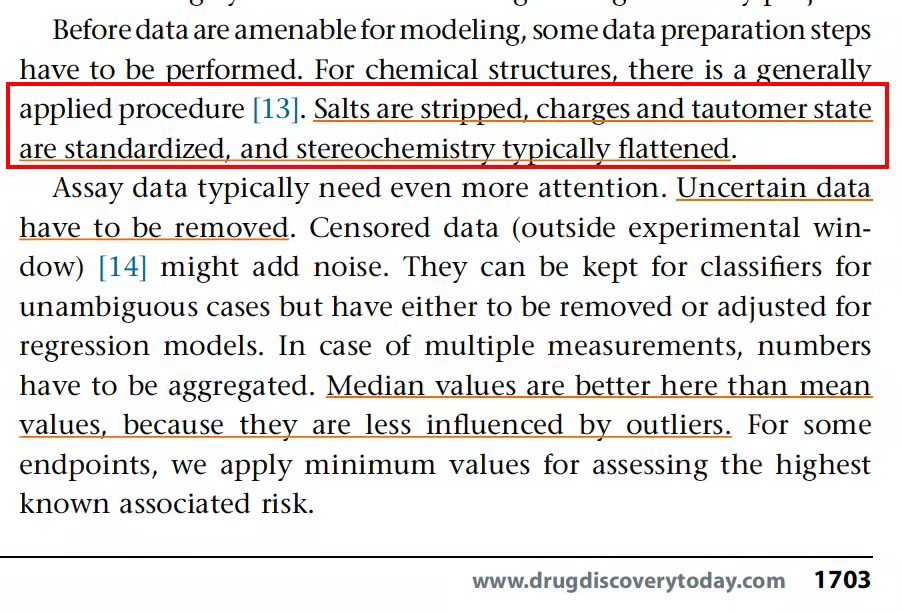
|  |  |  |
| --- | --- | --- |
| #9 | B | C=CC(=O)N[C@H]1CC[C@@H](c2cc(Nc3ccc(N4CCOCC4)cn3)c3ncnn3c2)C1 |
| #9 | A | C=CC(=O)N[C@@H]1CC[C@H](c2cc(Nc3ccc(N4CCOCC4)cn3)c3ncnn3c2)C1 |

两个都保留（暂定），根据拜耳公司发的综述（见下），他是把所有的手性中心都去掉，均处理为外消旋体，但只用2d描述符与分子指纹，去不去手性中心其实影响不大。（一般在药物研发中，为了降低研发成本，选择去除手性中心。但我们作为先导药物设计，没有过多这些顾虑，外消旋体有时可能会更稳定活性更强或者毒性更小）

比如：

32a与32b为手性，32为消旋（混合物），这里消旋活性反而更强。

情况2：筛选描述符之后，描述符相同，活性不同的状况很常见，但这个时候需要我的训练不要过拟合，根据筛选的描述符将多数活性预测对了就行。只要数据正确就保留。



在数据可用于建模之前，必须进行一些数据准备步骤。对于化学结构，有一个普遍适用的范式[13]。去除盐类，电荷和同分异构体状态被标准化（标准化smiles结构式），立体化学通常被平坦化（去除手性中心）。化验数据通常需要更加注意。必须删除不确定的数据。截断的数据（比如10-100nM、大于20nM）（实验窗口之外）[14] 可能会增加噪音。对于不确定的情况，它们可以被保留在分类模型中（分类模型可以保留这种截断数据），但对于回归模型，必须被删除或调整。在多次测量的情况下，必须对数字进行汇总，在这里，中位数比平均值更好，因为它们受离群值的影响更小。而对于某些端点，我们采用最低值来评估最高的已知相关风险。（这里我的理解是就像是，我们一开始取最高活性最低ic50一样，因为对于假阳性与假阴性来说，假阴性更不能容忍，所以取最低ic50）来源：Bayer’s in silico ADMET platform:a journey of machine learning over the past two decades

Ps 误差是不可避免的，机器学习需要算法，算力，数据，至少在数据上，我们能做的其实很少，比如表征端，很多测活，他给的就是外消旋体，面对这种数据，即使我用上GNN用图论进行表征，效果也并不会变好，因为数据来源本身表征的就是混合物，跟你能不能表示出手性顺反无关，因此我的建模也一定不可能区分手性，这是由于数据本身决定的，其次数据端，测定活性时，不同人不同实验室是有可能得到完全不同的影响的，因此我们使用高低活性模糊化数据，这样得到的数据才更有意义，但同时不可避免的有信息损失，综上建模实际上是摸索出一个大致的规律。

1. 顺反问题，

例1，N#CC(=CC1CC1)C(=O)N1CCCC(n2nc(-c3ccc(Oc4ccccc4)cc3)c3c(N)ncnc32)C1

N#C/C(=C/C1CC1)C(=O)N1CCCC(n2nc(-c3ccc(Oc4ccccc4)cc3)c3c(N)ncnc32)C1

数据库里分了顺反，文献里没有，名称也没有，图跟数据库的顺反不一致

如果去斜杠，要去两个/ \，二者2d描述符一致。

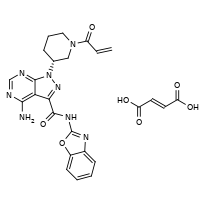
例2

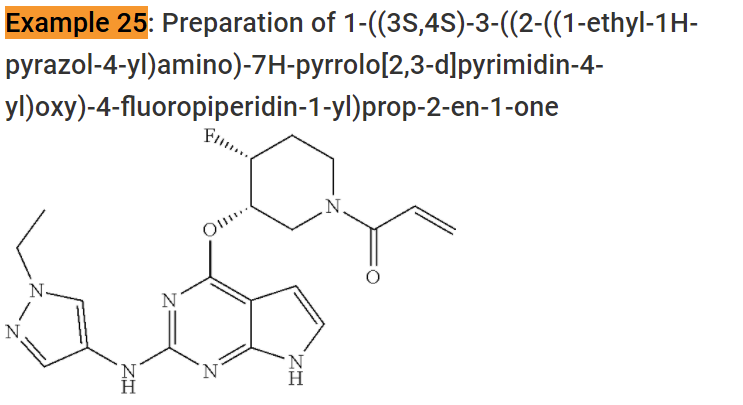
NC1=C2C(=NC=N1)N(CCN1C(=O)SC(=CC3=CC=CC=C3)C1=O)N=C2C1=CC=C(OC2=CC=CC=C2)C=C1

NC1=NC=NC2=C1C(=NN2CCN1C(=O)S\C(=C/C2=CC=CC=C2)C1=O)C1=CC=C(OC2=CC=CC=C2)C=C1-

算ECFP4，同手性

1. 盐类，如

化合物A的富马酸盐

1. US2020/308177中有测试 化合物31显示 No activity可以加入数据库吗？
2. 图片文字不一致，文献自己出现手性时自我前后矛盾

看看是不是前后有一样的图片，比如例36与例37图片一致，而名称不一致，则以名称为准。

1. 配位键画不出来？

这种化合物键型定义不明确，最好不用，不然后期算描述符也可能出错

1. 哪种ic50不能要？

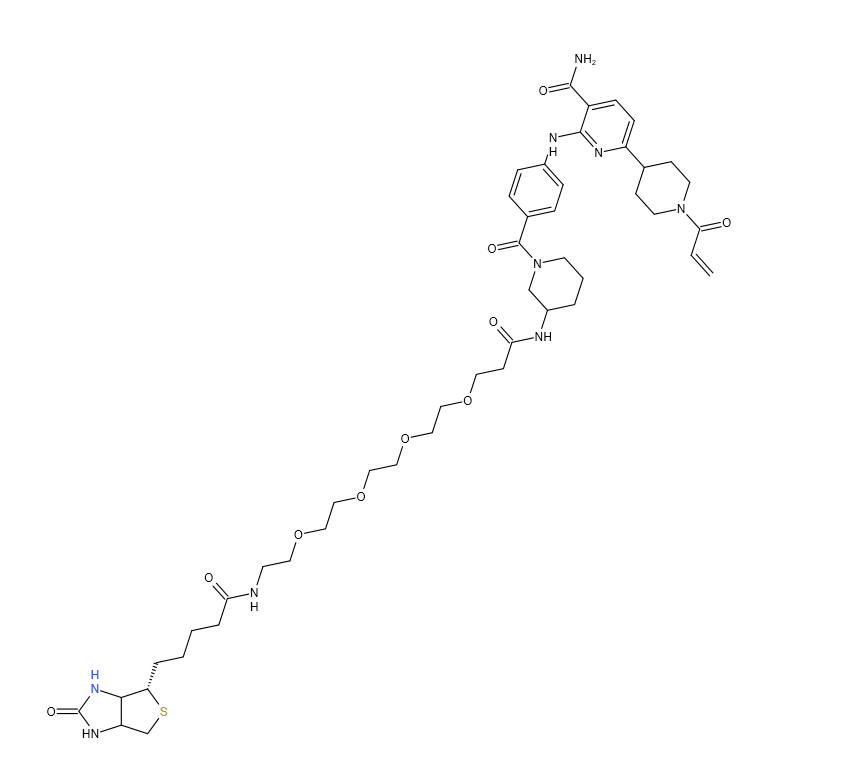
细胞的，比如提到人全血巴拉巴拉，百分比的，单位是%，测活方法不限，但需要保证是体外测活，因为体内测活会存在许多干扰。

1. 什么叫要去查对应方法，比如

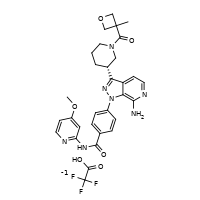
物化学评价

[0287] 式A化合物的BTK抑制活性在美国宾夕法尼亚州马尔文市大峡谷干道的反应生物公司(React1n B1logy Corporat1n,, One Great Valley Parkway, Malvern, PA, USA)测定。使用人BTK酶，底物为20 μ M的肽底物[KVEKIGEGTYGVVYK]。测定用的ATP浓度为10 μ Μ，星形孢菌素用作标准品，IC50为3.94nΜ。

这种，就得去查，或者有些给了个测试编号，你得具体去查是用的什么测活方法。

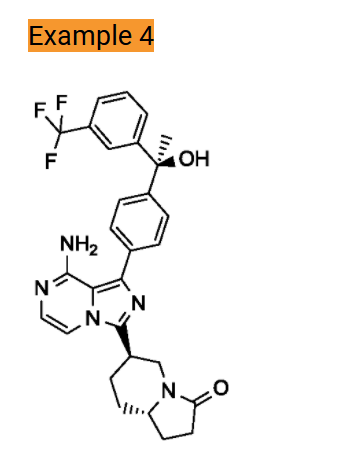
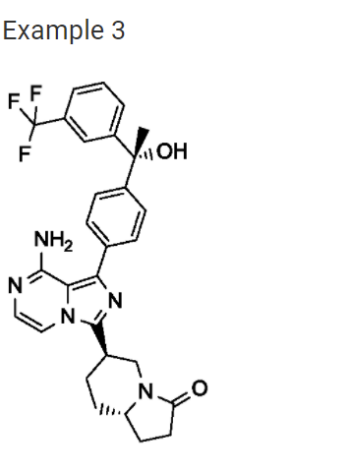
1. 

去除重原子数小于10大于60的分子，这种分子量过大或者过小都有可能是不同机理导致。

1. 文献中是，数据库中是，是否算错误？

不算，之后统一可以做盐的清洗，你也可以化结构时直接把盐去掉。

1. 活性悬崖是什么？



Example 3 883.5nM

Example 4 0.62nM

1. 为什么不把后续操作整合成一个软件或者程序？

这个问题可以替换成现在能用全机器人合成化合物了，一周全自动出结果，为什么不用机器人代替研究生？我们都知道所谓机器学习，需要算法，算力，数据，一旦合并会极大地提高算法与算力的成本，首先是算法的事，既然要合并，我保证程序的稳定性就是至关重要的，我不能三步一报错，因此我除了要考虑各个脚本的架构以外，我还需要考虑到各个报错，而且一个好的软件一定是少用if的，因为用if就会出现，if不到的情况，更倾向于使用for循环，那好了，复杂度直接全部由2n变n2了，好的架构二分法优化成nlogn，就是这些都需要很厉害的几个架构师进行优化。那算力部分呢，可以估算一下O（大O复杂度），我的数据比如有6000，那么我会有600个描述符，这就是600\*6000的底了，每一个算法20几个参数总有吧，每个参数不一定都是布尔类型吧，你再多加几个算法试试（一般计算机每秒算力在10的8次方，超级计算机在10的18次方每秒）。所以为什么机器学习或者深度学习领域十分看师承，因为调参这个东西更多看经验，现如今还没有直接能达到目的的算力，因此我们的调参都是有优化空间，比如95.1的准确率我花了大力气，变成了95.2，不值当，还不如去开发新算法，我们能保证的是我这些参数已经是超过99%的其它参数结果了，而不是保证我的准确率一定能达到99%。再者，我马上就能给你说出一个准确率高达90%以上的测活算法，那就是任何化合物，我都说没有活性，因为有活性的化合物在总共所有化合物总的比例一定是极低的，那么这个算法有意义吗，没有任何意义，我们的训练集永远是高低活性抑制剂，而绝非全体化合物，不然你的训练结果一定是这种没有任何意义的模型。此外，就像上面3问的情况2一样，也有可能出现根据皮尔森相关系数法筛选后的描述符，比如2048位ecfp4筛选剩余600，可能就会出现，相同描述符不同活性值的情况，因为不一致的描述符在数据集整体中方差较小，那么这种保留的噪声，也就决定了一定不可能达到100%的正确率了。

算法（深度学习、迁移学习等）

算力（量子计算机，摩尔定律下每年计算机运算速度的增长）

数据 数据的表征（图论、CV、NLP） 数据的采集（机器合成，纯度提高，测活精确，高产低成本地合成新分子获得测活数据）

1. 既然大多数化合物是无活性的，那么我加入这些大多数化合物是否可行？

不行，有文献这样做过，结论是反直觉的，加入了不同骨架的化合物，模型效果好了，但实际并没有用，因为加入了构造的与原有数据集相同骨架的化合物，整体也没有变化。个人猜测是因为在我的训练集本身就是有倾向的，实验做测试时，就是在相同骨架上来回改，因此我的模型也是如此，加入了构造的其它骨架只能是噪声。

Tips

1. 快捷键
2. 开启wps重复项高亮
3. Vlookup函数可一个一个删除
4. 安装Enhanced Image Viewer谷歌浏览器插件，可以将打开图片时的黑色变成白色，方便ocr。
5. 设置 输入步骤，可以记录步骤，快速生成文档。打开选择打印输出为pdf即可

阅读本文或者阅读本文后请查看下面链接

Doi：https://doi.org/10.1186/s13321-019-0351-x

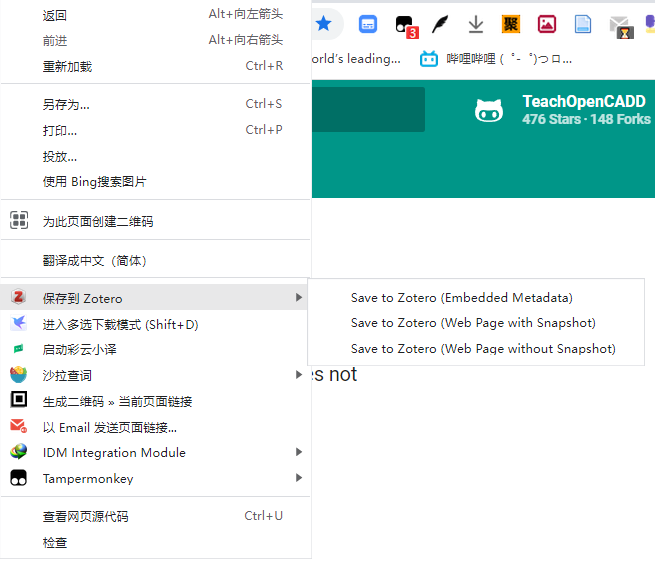
https://projects.volkamerlab.org/teachopencadd/all\_talktorials.html

<https://projects.volkamerlab.org/teachopencadd/talktorials/T001_query_chembl.html>

<https://github.com/volkamerlab/teachopencadd>

1. 关于文献管理器

不推荐endnote，因为没有版权，盗版会有奇怪的卡顿与bug，强烈推荐使用Zotero，能和谷歌浏览器或者微软浏览器插件连用。详细使用可搜索B站，补全该文档

1. 可以下载收藏网页  
   可以存网页快照，方便。
2. 保存文献的同时，能下载文献原始版本，支持JAVA，可以运行代码，批量下载文献或者更改文献信息。
3. 支持团队合作可创建组，团队分享文献及网站，推荐实验室可以自己建组，我没空弄了，收藏网站或者论文
4. 开源正版