

# DTI模型测试

(1) 其中部分数据没有证据支撑，比如某个分子C记录提示其靶点为T，但没有证据证明化合物C对靶点T的活性，这里drugbank的这个数据构造机制不明；建议额外增加一些有活性实验值的数据来作评测集，比如ChEMBL上下载，这个我们也会另外找一批有IC50值的数据来评测，再确认一下模型性能；

- (1) 用我们准备的ChEMBL数据继续测试，那个数据有pIC50；
- (2) 暂时先全部用抑制剂的靶点（如CDK2/4/6，BACE1，PDE9A，HDAC2，PDE4D）测试，不用其它的（激活剂等）

如果能找出对应的IC50进行排名，跟测试结果比较更有说服力


(2) drugbank这11167个里，有不常规的像[Li+], [Zn++], 或者18F类似的放射性分子，这类情况，模型的预测确实会有问题（和模型学习的数据类型差别很大，不恰当的比喻，就像学习的是钥匙能不能开锁，这里来了一根牙签），对于这部分分子量比较小或其他明显特征的化合物，应该排除或者在模型训练数据中手工加上；

- (3) 如果要排名的话测试前去除小分子量的化合物；

总体上，这次的drugbank数据集，可能对药物重定向、图谱的预测是OK的，但是对于靶点亲和力作用/药物活性可能不太适合，我们再增加更吻合的活性数据来评测确认（模型的适用场景不一样，如果证实图谱的评测数据准确，可以药物重定向任务用图谱，大规模的虚拟筛选用DTI模型）；

药物之所以有活性就是因为有亲和力，虽然有亲和力不一定有活性，IC50就是因为亲和力引起的，drugbank可以用来测试DTI模型

还有像这个化合物，在drugbank里面显示是CDK2靶点抑制剂，但是实际IC50非常大，drugbank的药物-靶点对应这个数据，可能有一些问题，或者机制我们不明确

Target/Host (Institution)	Ligand	Target/Host Links	Ligand Links	Trg + Lig Links	Ki nM	$\Delta G^\circ$ kcal/mole	IC50 nM	Kd nM	EC50 nM
Cyclin-dependent kinase 2  Homo sapiens (Human))   Astex Therapeutics Ltd  Curated by ChEMBL	BDBM24626    (6-chloropyrazin-2-amine   6-chloropyrazin-2-amine)	PDB MMDB  NCI pathway Reactome pathway KEGG  UniProtKB/SwissProt UniProtKB/TrEMBL  DrugBank antibodypedia GoogleScholar	Purchase  ChEMBL DrugBank MMDB PC cid PDB Article PubMed UniChem  Patents		n/a	n/a	3.50E+5	n/a	n/a
					Assay Description Inhibition of CDK2				J
					More data for this Ligand-Target Pair				B

(1) 如果是只考虑是不是药物的话，可以考虑去除这些活性比较低的化合物；当然也可以测试化合物的IC50预测准不准，这样不管化合物IC50多大多小都没关系，主要看预测的准确度就可以

(2) 找已经上市的药物，这些化合物IC50肯定会比较小

针对以上问题，改进测试方案：

=====

原测试靶点：

5-hydroxytryptamine receptor 2A
Carbonic anhydrase 2
Cyclin-dependent kinase 2
Estrogen receptor
Histamine H1 receptor
Muscarinic acetylcholine
Prothrombin
Trypsin-1

新增靶点：PDE9A， PDE4D， HDAC2， CDK4， CDK6， ACHE

=====

**测试内容1：**设想需要从drugbank中筛选药物，查看针对该靶点的药物能不能**排名**比较靠前，且观察下预测的PIC50值是不是比较大（例如是否大于5或者6）这里假设除药物外的其它化合物对该靶点是没有作用的，虽然这个假设不一定100%正确，但在一定程度上是可以的

**筛选数据：**在drugbank中11167基础上删除分子量过小或者过大的化合物

**测试数据：**以下数据中找每个靶点对应的化合物

drugbank\_target\_drug\_info\_202110.xlsx

**备注：**可以在已经测试得到的数据基础上，让向冰算下化合物分子量，去除掉分子量小于200和大于1000的化合物，然后再排序

### 测试内容2:

(1) ZINC中下载的某靶点的化合物全部预测一遍，跟模型PIC50结果比较，计算R2，RMSE；

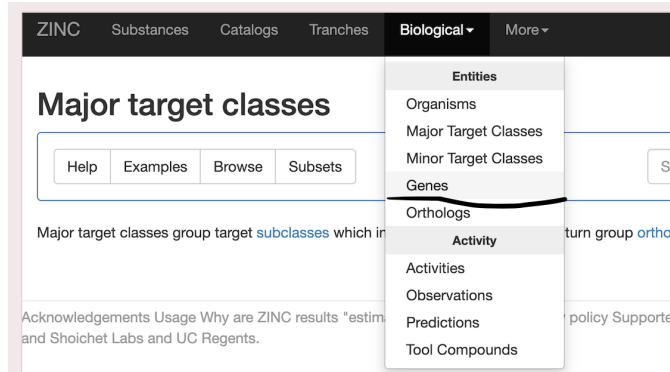
(2) 把模型结果按PIC50为6进行切分，大于6为1，小于5为0，ZINC数据集也按照相同的方法划分为1和0，再根据混淆矩阵计算AUC，accuracy, sensitivity, specificity等指标；

(3) 更具体一些的研究下化合物：可以把针对某靶点活性比较高的化合物（例如affinity>8）抽取出来，看下模型预测出来的结果是多少（如果模型效果好的情况下，预测值也应该比较高）；另一方面可以把针对靶点活性比较差的化合物（例如affinity<8），也抽取出来，看下模型预测出来的结果是多少（如果模型效果好的情况下，预测值也应该比较低）

**测试数据：**来源于zinc中每个靶点的数据

**备注：**测试内容2中数据的下载方法：

1. 登陆网站：<https://zinc.docking.org/>
2. biological → genes:



4. input gene name and search:

**Genes**

Help Examples Browse Subsets

CDK2 Search

Genes group [orthologs](#) within specific [organism](#) classes using the Uniprot Gene symbol.

5. select gene:

Name	Description	Organism	Sub Class (Major Class)	Orthologs	Observations	Substances	Purchasable	Predicted
CDK2	Cyclin-dependent kinase 2	Eukaryotes	kinase (enzyme)	1	1552	1279	169	257908
CDK2AP1	Cyclin-dependent kinase 2-associated protein 1	Eukaryotes	unclassified (Unclassified)	1	18	18	0	39876
CDK2AP2	Cyclin-dependent kinase 2-associated protein 2	Eukaryotes	unclassified (Unclassified)	1	0	0	0	0

6. select observations:

**CDK2 (Cyclin-dependent kinase 2)**

In: [eukaryotic](#) [kinase](#) [liganded](#) [purchasable](#)

3D models for docking

DB2 format [db2.gz](#)

mol2 format [mol2.gz](#)

SDF format [sdf.gz](#)

Relations

# Orthologs	# Observations	# Substances	# Purchasable	# Predicted
1	1552	1279	169	257908

7. download:

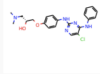
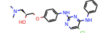
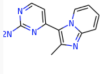
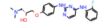
1

1552

/ genes / CDK2 / observations

Filters

Lookup

Target	Description	pKi (L.E.)	Substance	Structure	Citation
CDK2_HUMAN	Cyclin-dependent kinase 2	6.30 (0.30)	ZINC000013538076		5737; PMID12941312 Bloorg. Med. Chem. Lett. 2003
CDK2_HUMAN	Cyclin-dependent kinase 2	6.30 (0.30)	ZINC000013538079		5737; PMID12941312 Bloorg. Med. Chem. Lett. 2003
CDK2_HUMAN	Cyclin-dependent kinase 2	5.40 (0.44)	ZINC000012354765		5748; PMID12941325 Bloorg. Med. Chem. Lett. 2003
CDK2_HUMAN	Cyclin-dependent kinase 2	5.15 (0.25)	ZINC000013537935		5737; PMID12941312 Bloorg. Med. Chem. Lett. 2003

其它待解决问题：训练数据集中含有与测试数据中相同的化合物