厦門大學
XIAMEN UNIVERSITY

《数据挖掘与分析》课程大作业

主讲：张仲楠 教授

2025年05月

作业任务

Xiamen University

1. 药物脂溶性预测（回归任务）
2. 药物副作用预测（分类任务）
3. 化合物库代表性分子筛选（聚类任务）
4. 药物-靶点相互作用预测（链接预测任务）

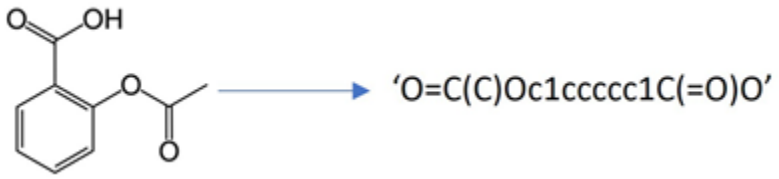
以上四个任务任选三个完成

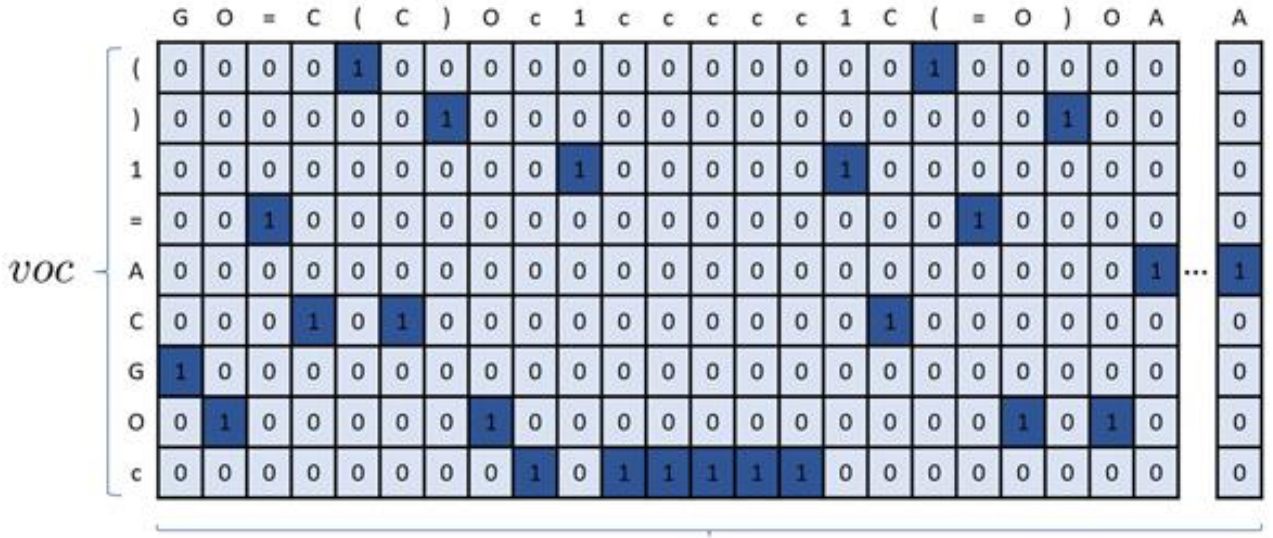
分子表示

Xiamen University

1. SMILES表示：字符串描述分子结构

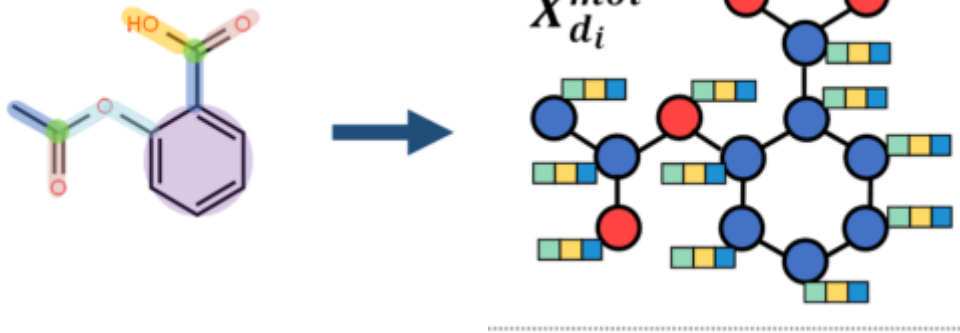
编码成one-hot向量输入模型





2. 分子图表示：描述分子的拓扑结构

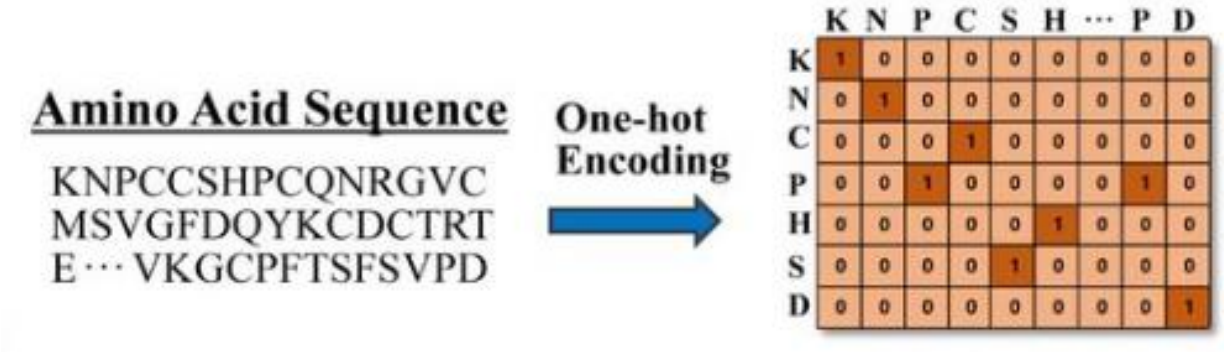
使用图神经网络（GNN）提取图的特征



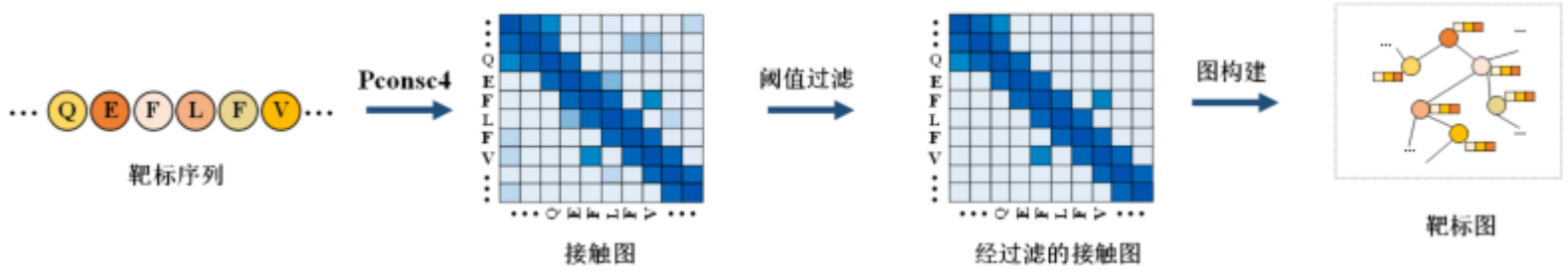
蛋白质表示

1. 蛋白质序列表示：由氨基酸组成，描述蛋白质结构

   编码成one-hot向量输入模型



2. 蛋白二维接触图



1、药物脂溶性预测（回归任务）

Xiamen University

- 药物脂溶性预测：脂溶性（Lipophilicity）是衡量一个分子在脂相和水相之间分布倾向的重要理化性质，是药物分子极其关键的理化性质之一，脂溶性直接影响药物ADME（吸收、分布、代谢和排泄）性质。该任务旨在构建预测模型，输入药物的结构信息（如 SMILES 或分子图），预测其脂溶性值，从而实现对脂溶性的高效建模与精准估计。

1、药物脂溶性预测（回归任务）

Xiamen University

■ 数据集介绍：

Lipophilicity数据集来自ChEMBL数据库，提供了化合物在pH7.4条件下辛醇/水分配系数的实验结果。

smiles	exp
CC(C)NCC(O)COc1	-1.703482254
Cc1ccc(NC(=N)N)	-2.892589231
CC(C)C(=O)NCCNC	-1.909924438
C=CCNS(=O)(=O)c	0.7490508866057364
C=CCOc1ccccc10C	-1.620905381
Cc1ccc(O)c(O)c1	-0.654755962
CCCCN(CCCC)CCC(	-0.19232547

■ 数据集结构

- 1_Lipophilicity

-train：训练集，含有3360个分子辛醇/水分配系数结果

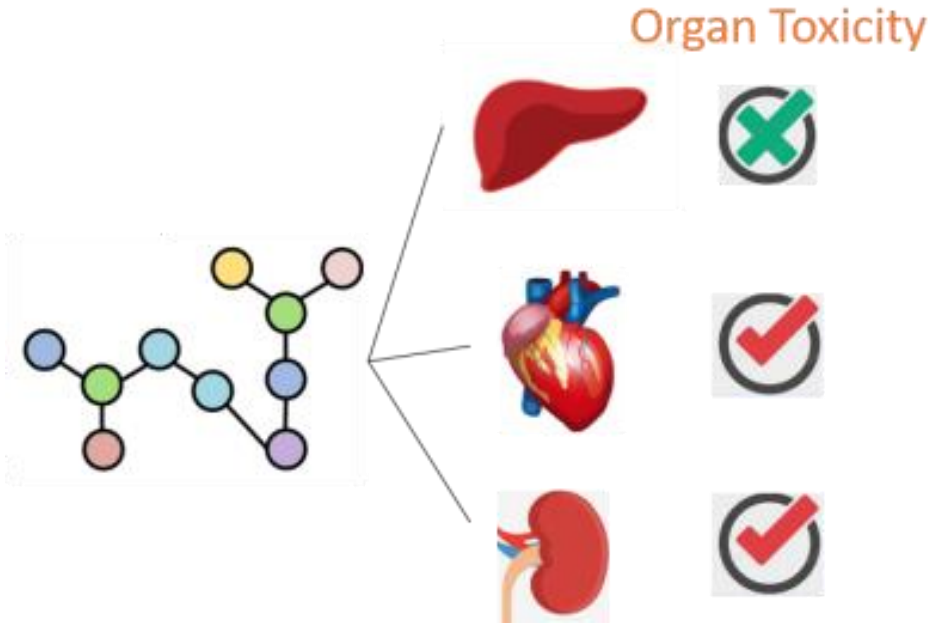
-test：测试集，含有840个分子辛醇/水分配系数结果

■ 实验要求

RMSE指标值<0.7

2、药物副作用预测（分类任务）

■药物副作用预测（Drug Side Effect Prediction）：是指利用药物的分子结构信息，预测其在临床使用过程中可能引发的非预期生理反应（副作用）。该任务通常采用多标签分类模型，通过学习药物结构与已知副作用之间的关联，判断其是否会诱发特定类型的不良反应。药物副作用预测有助于在药物研发早期识别潜在安全风险，提高新药的临床成功率，降低后期撤药风险与试验成本。



2、药物毒性预测（分类任务）

Xiamen University

■ 数据集介绍

副作用资源(SIDER)数据集：是一个上市药物和药物不良反应数据库，SIDER 数据集按照将药物副作用分为27个系统器官类别，包括肝胆系统、代谢系统、眼部疾病、胃肠系统等。每条样本对应一个药物分子（以 SMILES 表示），并通过 0/1 标签标识其是否在某一器官系统类别中具有已知副作用，标签值为 1 表示该药物已被报告在对应的器官系统中存在不良反应；标签值为 0 表示当前无已知不良反应关联。

■ 数据集目录结构：

- SIDER

-train：训练集，包含1141个分子

-test: 测试集，包含286个分子

	smiles	Hepatobil	Metabolis	Product i	Eye disor	Investiga	Musculosk	Gastroint	Social ci	I
2	NCCNCCNCC	1	1	0	0	1	1	1	0	
3	C1[Tl]	0	0	0	1	1	0	1	0	
4	C[N+](C)(	0	1	0	1	1	1	1	0	
5	NCCCC(=O)(	0	0	0	0	0	0	0	0	
6	NCC(=O)CC(	1	0	0	1	1	0	1	0	
7	CC(=O)O.O-	0	1	0	0	1	0	0	0	
8	CC(=O)[O-	0	1	0	0	0	0	0	0	

2、药物毒性预测（分类任务）

Xiamen University

■ **实验要求：**

- 实验评估指标包括：

  AUROC（Area Under the ROC Curve）

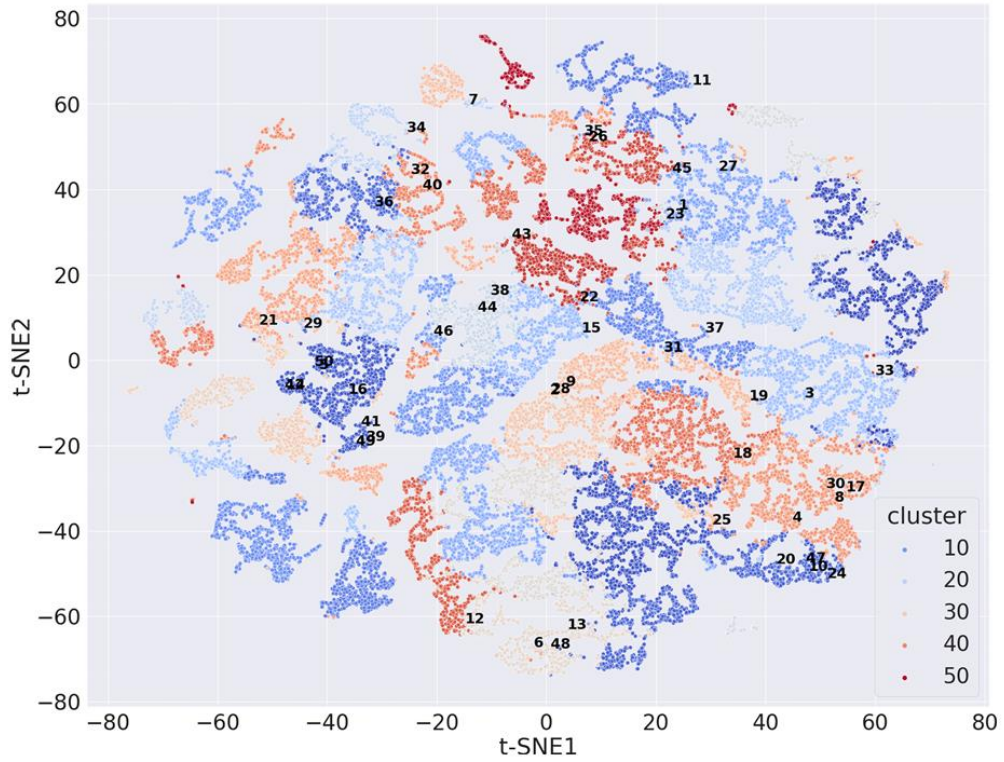
  AUPR（Area Under the Precision-Recall Curve）

- 除了数值结果，还需提供相应的曲线图（如 ROC 曲线与 PR 曲线）
- 实验结果中，AUROC 值应达到 0.8 以上

3、化合物库代表性分子筛选（聚类任务）

Xiamen University

■在从大规模分子数据中选取结构多样且具有代表性的分子子集，覆盖整个化学空间，提高虚拟筛选与实验验证效率，降低冗余与成本。



3、化合物库代表性分子筛选（聚类任务）

Xiamen University

■数据集：

- johnson数据集中包含47217 种化合物的SMILES信息。

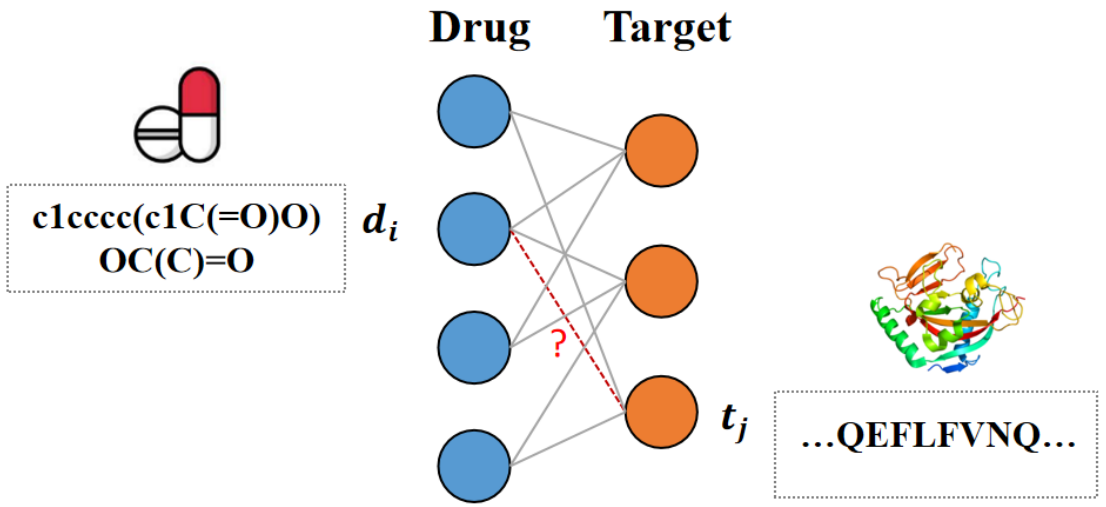
	SMILES
2	CN(C)C(=O)CC1CC2(CCN(CC2)C(=O)N2CCCC2)Oc2ccccc12
3	Cc1c([nH]c2CC(CC(=O)c12)c1ccco1)C(=O)OCC1CCCO1
4	CNC(=O)CN1CCC11CCN(C1)C(=O)c1ccn(C)n1
5	Cn1cc(cn1)N1CCC2(CCN(C2)C(=O)c2ccncc2)C1=O
6	CC(C)CN1CC2CN(CC2C1)S(=O)(=O)c1ccccc1
7	Fc1cccc(c1)-c1ccc2c(nnn2c1)C1CCCN1Cc1ccncc1

■实验要求：

- 包含分子表示的 t-SNE 可视化图；
- 指标要求： Silhouette>0.3 、 Calinski-Harabasz 指数>10000， 以及 Davies-Bouldin<1。

4、药物-靶点相互作用预测（链接预测任务）

■ 药物-靶点相互作用预测（Drug-target interactions）：基于药物分子结构和蛋白质序列信息，预测两者是否存在结合或生物活性关系。该任务可用于新药筛选以及作用机制研究等，是药物发现中的重要环节。



4、药物-靶点相互作用预测（链接预测）

Xiamen University

■ 数据集目录结构：

– HUMAN

-train：训练集，4197条药物（SMILES）-靶点（Protein）相互作用数据，其中interaction=1表示药物与靶点之间存在已知的、正向的相互作用; interaction=0,表示当前没有证据表明该药物与该靶点存在相互作用。

-test: 测试集, 1200条药物-靶点相互作用数据。

	A	B	C
1	SMILES	Protein	interaction
2	CS(=O)(=O)c1cc	MDSLVVLVL	0
3	CCOC(=O)[C@H](	MLEKFCNST	0
4	Cn1c(=O)c(Oc2c	MDSLVVLVL	0
5	CC(C)C[C@@H](N	MALGRLSSR	0
6	COc1ccc(S(=O)(	MALIPDLAM	0
7	CN(C)CCCN1c2cc	MPTVDDILE	0
8	COc1cc(N2CCC(N	MATGGRRGA	1
9	COc1cc(O)c(/C=	MAPLCPSPW	0
10	C[C@H](O)C(=O)	MGMACLTMT	0
11	CC(C)(O)c1cncc	MALRAKAEV	1
12	CC(C)CN(Cc1nc2	MAASRRSQH	1

■ 实验要求：

- 需根据给定的训练集使用基于图的方法进行训练，并在测试集上完成药靶相互作用预测，即看作是链接预测任务。
- 指标要求：AUC>0.86，AUPR> 0.8。

提交要求

Xiamen University

(1) 需根据每个作业任务完成对应的实验与分析，并将过程与结果整理形成一份完整的书面报告(word文档)，书面报告至少应当包括如下内容:封面，目录，概述，总体设计，算法的详细介绍，实验软硬件环境，实验详细步骤，实验结果与分析，参考文献。

(2) python源代码及所有支撑文件，代码readme。提交的源代码中要有详细明确的注释，代码要能够直接正常运行。

(3) 以上所有文档打包上传到lnt.xmu.edu.cn，压缩包命名方式完整学号_姓名.rar(zip)。准备充分了再上传，尽量避免重复上传。

(4) 截至时间:2025.7.5 23:59:59。以系统时间为准。