SAIL-25 夏令营 Starter Task

——by chenyc、xudh

任务一：编程实验

背景

药物血脑屏障通透性预测（Blood-Brain Barrier Penetration，BBBP）在药物开发的早期阶段具有重要作用，能快速判断哪些分子有潜力进入大脑。该预测可以尽早淘汰掉那些注定无法进入大脑的分子，避免在它们身上浪费后续昂贵的研究经费和时间。因此建立可靠的预测模型对于优化药物设计、提高药物研发成功率至关重要。随着计算生物学和人工智能技术的发展，越来越多基于机器学习的预测工具被开发出来，这些工具结合分子结构特征与已有的实验数据，能够较为准确地模拟药物分子在体内的穿透行为。通过这些方法，研究人员可以在化合物合成之前就对其血脑屏障通透性进行评估，从而更有针对性地开展后续实验。这类预测也有助于理解影响血脑屏障穿透性的关键分子特性，推动新型递送系统的设计与应用，为治疗神经退行性疾病、脑肿瘤等难治性脑部疾病提供新的可能性。

问题定义

对于这个任务，可以简单将其抽象为对分子的二分类任务。根据分子的性质，将其分成两类，即是否可以穿过血脑屏障。

数据集

数据集储存在文件"BBBP.csv"中，其结构如下：

"num"	分子 id
"name"	分子名
"p_np"	标签，分为 0/1
"smiles"	分子的 smiles 表达式

要求

1.分子通常可以被表示为“分子指纹”和“分子图”，请使用其中一种表征，利用提供的分子数据，搭建并训练模型，对其进行二分类，计算分类指标（如 AUROC、F1 等）。

2.在 AI4S 的场景中，正例和负例常常不是相等的，通常其中一些类是少数类，本任务即是如此。请分析该任务 0/1 标签是否存在不平衡的情况，并思考如何在模型训练中应对。

3.根据上述要求，撰写实验报告，记录问题的解决方法和思考感悟。

Bonus（该题作为附加题）

4.请分别用分子指纹和分子图两种分子表征训练模型，并进行比较。

Hints

1.smiles 是一种表示分子的方法，python 库 rdkit 有读取 smiles 的函数。

2.分子通常可以被表示为“分子指纹”和“分子图”，前者通常输入 MLP 中，后者则输入 GNN 中。分子指纹和分子图通常都可以由 smiles 得到，且 rdkit 中有计算分子指纹的函数。相对而言，使用分子指纹进行分类更简单直接易于实现，使用分子图则需要构图并搭配 GNN 使用，会略微复杂。

任务二：文献阅读

背景

文献阅读是科研的重要基础。为考察大家的文献阅读与分析能力，这里提供了五篇近年的生信领域的顶会/顶刊论文，供大家进行阅读。

论文列表

1.Retrieval-Augmented Language Model for Knowledge-Aware Protein Encoding. ICML 2025

2.UniMoMo: Unified Generative Modeling of 3D Molecules for De Novo Binder Design. ICML 2025

3.Boltzmann-Aligned Inverse Folding Model as a Predictor of Mutational Effects on Protein-Protein Interactions. ICLR 2025

4.KGARevion: An AI Agent for Knowledge-Intensive Biomedical QA. ICLR 2025

5.Traversing chemical space with active deep learning for low-data drug discovery. NCS 2024

要求

1.请任选其中一篇进行深入阅读，对论文的背景、方法和实验设计等进行深入分析。

2.制作文献分享 PPT，我们将根据提交内容安排后续线上的文献分享，欢迎大家积极参与。

提交方式

请将任务一的实验报告+代码和任务二的 PPT 压缩打包，一并发送至邮箱：xudh6@mail2.sysu.edu.cn。