# 项目题目：

**基于监督流形学习的精准分子三维构象生成方法研究**

# 背景介绍

精确预测小分子的三维构象对药物设计、材料科学及计算化学领域至关重要。传统方法依赖于力场优化或间接距离预测，这种方式存在误差累积及泛化能力不足的问题。无监督流形学习方法如UMAP等虽能捕获高维数据内在结构，但其缺乏监督信息导致无法精准预测具有物理意义的分子构象。

# 研究目标

本项目旨在开发一种监督流形学习框架，将分子图拓扑结构精准映射到三维构象空间，以实现精确、物理一致且泛化性能良好的构象预测。

# 实验设计与方法细节

## （一）数据准备与预处理

- 从现有分子数据库（如QM9、GEOM）获取小分子数据，包括分子的2D拓扑结构（SMILES或图结构）及实验验证或DFT计算的3D构象。
- 使用RDKit工具包对数据进行标准化，包括氢原子补全、原子类型统一及初步几何优化，确保数据质量与一致性。
- 数据集划分为训练集、验证集及测试集，比例约为8:1:1，以验证模型的泛化性能。

## （二）高维拓扑图构建

- 对每个分子构建图结构，节点表示原子（原子类型、杂化轨道等），边表示化学键（键类型、键级等）。
- 基于图距离（如shortest-path距离）生成高维邻接关系矩阵，使用指数衰减函数 $w_{ij}^{H} = \exp(-d_{ij}/\sigma_H)$ 计算节点间拓扑相似性，作为高维流形结构。

## （三）低维几何图构建

- 网络预测的三维坐标用于构建低维图结构，通过欧氏距离 $d_{ij}^{L} = \|x_i - x_j\|$ 计算几何邻接矩阵，并计算相似度 $w_{ij}^{L} = \exp(-d_{ij}^{L}/\sigma_L)$。
- 该图与高维拓扑图共同用于监督学习中的结构一致性损失。

## （四）监督流形嵌入模型构建

- 采用图神经网络（如GIN或GAT）提取分子图结构信息，经过数层GNN后输出每个原子的嵌入向量，并通过全连接层映射为三维坐标向量。
- 设计总损失函数由两个部分组成：
- **流形保持损失（结构一致性）**：

$$L_{manifold} = \sum_{i,j} (w_{ij}^{H} - w_{ij}^{L})^2$$

- **坐标重建损失（MSE）**：

$$L_{MSE} = \sum_i \|\hat{x}_i - x_i^{\text{true}}\|^2$$

总损失为：

$$L = \lambda_1 L_{manifold} + \lambda_2 L_{MSE}$$

其中 $\lambda_1, \lambda_2$ 为权重系数，控制结构保持与精确坐标监督之间的权衡。

### （五）模型训练与优化策略

- 使用PyTorch Geometric实现图神经网络结构，训练中采用Adam优化器，初始学习率设置为0.001，并使用余弦退火学习率调度策略。
- 引入批归一化（BatchNorm）与Dropout抑制过拟合。
- 模型通过早停机制监测验证集损失自动终止。
- 对超参数（如GNN层数、隐藏层宽度、$\lambda_1, \lambda_2$ ）使用贝叶斯优化或网格搜索调整。

### （六）模型评估与验证

- 使用RMSD（Root Mean Square Deviation）和MAE（Mean Absolute Error）作为构象预测精度评价指标。
- 采用T-SNE或UMAP对预测构象的嵌入向量可视化，验证其是否具备流形结构一致性。
- 与传统距离矩阵回归、力场优化（如UFF/MMFF）以及最先进构象生成模型（如GeoDiff、ConfGF）进行系统对比评估。

## 创新点

- 首次明确使用监督式流形学习方法优化分子图结构与三维几何构象的显式一致性。
- 设计结构-几何双重损失函数，兼顾几何精度与拓扑合理性。
- 提供物理直观且高度可解释的构象生成过程。

## 项目的重要性

本项目提出的监督流形学习方法能够有效提升分子三维构象预测的精确度、泛化性能及物理合理性，将对药物虚拟筛选、材料计算设计及计算生物学研究提供强大的方法学支持。