**化学反应产率预测自动化建模系统**

环境：

开发环境：

编程语言：Python

硬件： —— 性能需求

显存

内存

数据:

原始数据：

底物，产物，条件（Reagent、溶剂、时间、温度、压强），产率

SMILES~~，名字，CAS号~~

读入原始数据，读入用户输入的数据文件，读取其中的哪些列

文件类型：csv, xls, xlsx, txt

文件大小：

小于100MB?

数据集划分：

按比例划分 – 指定随机数种子

按已分好的文件

文件编码：统一转为UTF-8

校验：空值、多于字符、数据类型校验、SMILES有效性

描述符计算：

SMILES：拼接反应SMILES （A,B,C）🡪 A.B>>C ：深度学习

物理化学性质：根据SMILES计算物理化学性质，挑选可用rdkit用的性质，机器学习 化合物本不需要表示，仅用物理化学性质表示化合物

分子指纹：机器学习

~~图结构表示:GNN~~

~~DFT描述符:系统先不做计算~~

描述符存储：

Inchikey作为查询的key

先到文件查找描述符，如果没有，再去计算

过采样算法，平衡数据分布

降维算法

如果在训练中用到了降维模型，要把降维模型跟预测模型一起保存到文件

如果用户指定使用降维算法，在超参数搜索空间添加降维后的维度作为超参数

模型

分类模型&回归模型：对下面的模型归类，哪些可以做分类，哪些可以做回归

不同模型的初始化、超参数的制定

不同模型接受的输入数据格式

模型保存、管理：

保存格式：pickle

命名规范，如何把训练数据-评价指标与模型对应

机器学习：输入矩阵

线性回归(回归)

逻辑回归（分类）

SVM（回归、分类）

决策树（回归、分类）

随机森林（回归、分类）

XGBoost（回归、分类）

… …

深度学习：输入反应SMILES，用BERT做embedding

BERT（回归）

T5（回归、分类）

综合模型：确认输入是矩阵还是表格

AutoGluon

… …

超参数

不做调参，用户指定超参数

调参工具：wandb，NNI

调参依据，评价指标的选择

需要指定超参数的空间：

问题：不同模型、算法有不同的超参数和搜索空间

预测：

新一批数据预测，评价指标计算，散点图绘制？

评价：

分类评价指标

回归评价指标

用户：

1. 训练模型：
   1. 指定数据集，指定反应中每种角色对应的列
   2. 数据集划分方法：
      1. 指定划分比例
      2. 指定已划分好的文件
   3. 指定描述符
   4. 训练分类模型还是回归模型
   5. 指定是否使用过采样算法
   6. 是否使用降维算法
   7. 选择模型
   8. 指定调参的配置（尝试参数组合数量、并行数量……）
   9. 输出超参数组合-评价指标-模型路径？
2. 测试模型（推理）
   1. 指定数据集，指定反应中每种角色对应的列
   2. 数据集划分方法：
      1. 指定划分比例
      2. 指定已划分好的文件
   3. 指定模型
   4. 输出测试结果（原始数据+预测值）

原始数据读入、处理 1人

描述符计算、存储、读取+过采样+降维 2人

模型初始化、模型保存、读取、测试（推理），评价指标计算 3人

超参数搜索空间定义、实验定义&启动、最优模型的选择、保存 1人