基于机器学习的材料发现项目可行性方案

gogogo

2025 年 9 月 7 日

1 项目框架

1.1 问题切入

- 选择一个我们研究的**核心性质**，例如：带隙、形成能、超导转变温度（T_c）。
- 定义**模型评判标准**，例如：稳定半导体，要求带隙 > 2 eV 且形成能接近凸包。

1.2 数据收集

从公开的第一性原理数据库获取材料结构与性质数据：

- Materials Project (MP)，支持 API 调用和 pymatgen 工具。
- OQMD (Open Quantum Materials Database)。这玩意是最权威的之一但是我打不开链接
- AFLOW (Automatic Flow for Materials Discovery)。
- JARVIS-DFT (NIST)。
- NOMAD (Novel Materials Discovery Laboratory)。

1.3 数据预处理

- 清理重复数据和缺失值。
- 材料特征化方法：
  - 基于成分：Magpie 特征、mat2vec 嵌入（使用 `matminer` 工具包）。
  - 基于晶体结构：图神经网络（GNN）构建原子-键的周期图。

1.4 基线模型

- 使用简单监督学习模型进行基线实验：
  - 随机森林（Random Forest）。
  - XGBoost。
- 在验证集上评估性能（MAE、R^2），作为后续模型的比较基准。

1.5 图神经网络模型

- **CGCNN**：晶体图卷积神经网络。
- **MEGNet**：引入状态变量和迁移学习。
- **ALIGNN**：利用线图结构引入键角信息。
- **M3GNet**：引入三体相互作用，可作为通用势能面。

1.6 主动学习与筛选

- 使用训练好的代理模型预测大规模候选材料的性质。
- 引入**不确定性估计**（如集成模型、dropout）来识别有前景但模型不确定的样本。
- 结合**贝叶斯优化**或主动学习迭代，与 DFT 验证形成闭环。

1.7 最终交付

- 一个完整的模型（基线 + GNN）。
- 一份候选材料清单，满足预设目标。
- 可视化结果：
  - 特征重要性排序图。
  - 预测值 vs 真实值散点图。
  - 候选材料排名表格。

2 扩展功能

在基本框架之上，可以进一步加入：

- 结合文本优化，利用文献进行相关延伸或补充。
- 迁移学习（在大数据库上预训练，在小数据集上微调）。这一点我们就可以根据交大的强势方向进行特化，比如针对电池材料，二维超导材料等特定领域进行微调。
- 生成模型（如 VAE、GAN）实现材料逆向设计。