AI4Science 研究进展分享

基于深度学习的 XRD 晶体结构解析

黄维然 副教授

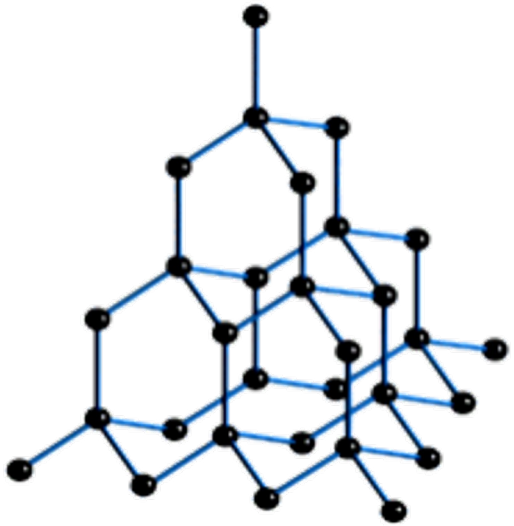
上海交通大学计算机学院

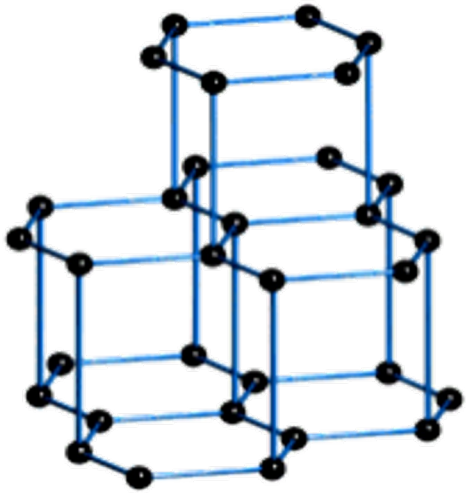
2025 年 2 月 23 日

晶体结构解析

金刚石

石墨





晶体结构解析具有十分重要的科学意义

生物与医学：理解生物大分子的功能和作用机制、针对疾病蛋白质结构设计药物

材料科学：预测材料的物理、化学和力学性质，从而设计出具有特定性能的新材料

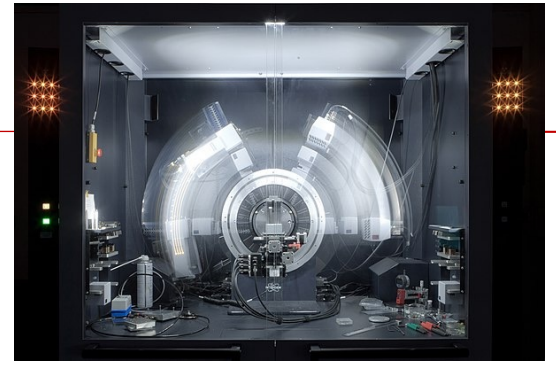
物理与电子学：精确控制晶体结构，可以制造出高性能的晶体管、集成电路和光电器件

化学与分子科学：晶体结构可以帮助科学家理解化学反应中分子之间的相互作用

地质科学、能源技术、环境科学 等等……

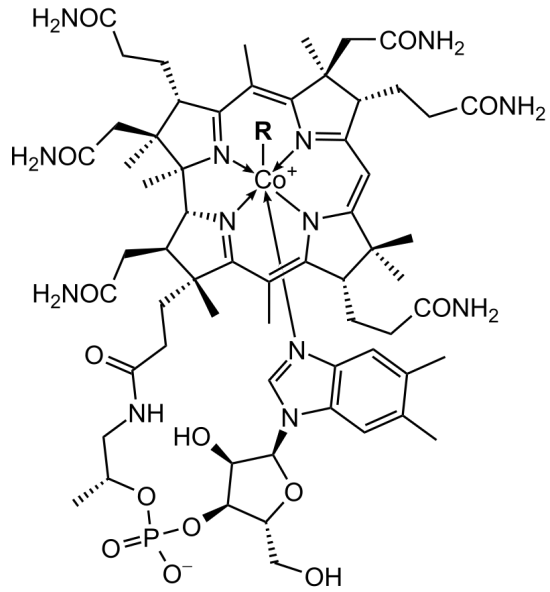
X 射线衍射 (X-Ray Diffraction)

- X 射线衍射（XRD）是一种常用的晶体结构解析方法
- 与 XRD 晶体结构解析相关的诺贝尔奖大约有 18 个



X 射线晶体仪

1962 年诺贝尔生理学与医学奖，利用 XRD 发现 DNA 的双螺旋结构



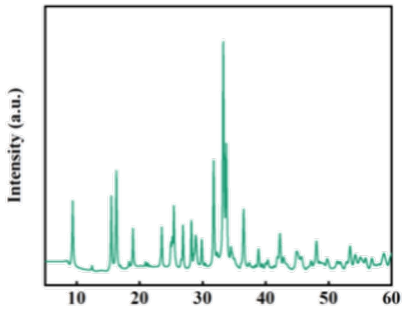
R = 5'-deoxyadenosyl, CH_3, OH, CN

1964 年诺贝尔化学奖，利用 XRD 测定青霉素、维生素 B_{12} 晶体结构

XRD 晶体结构解析步骤

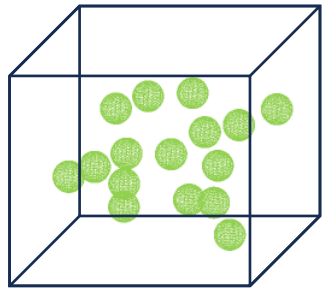
待解析晶体

X 射线



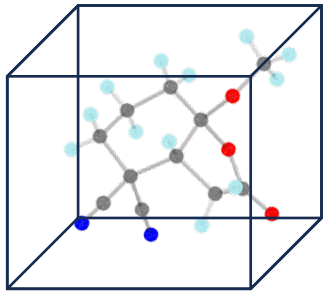
衍射强度分布

相位解析



电子密度图

计算机辅助
的人工求解

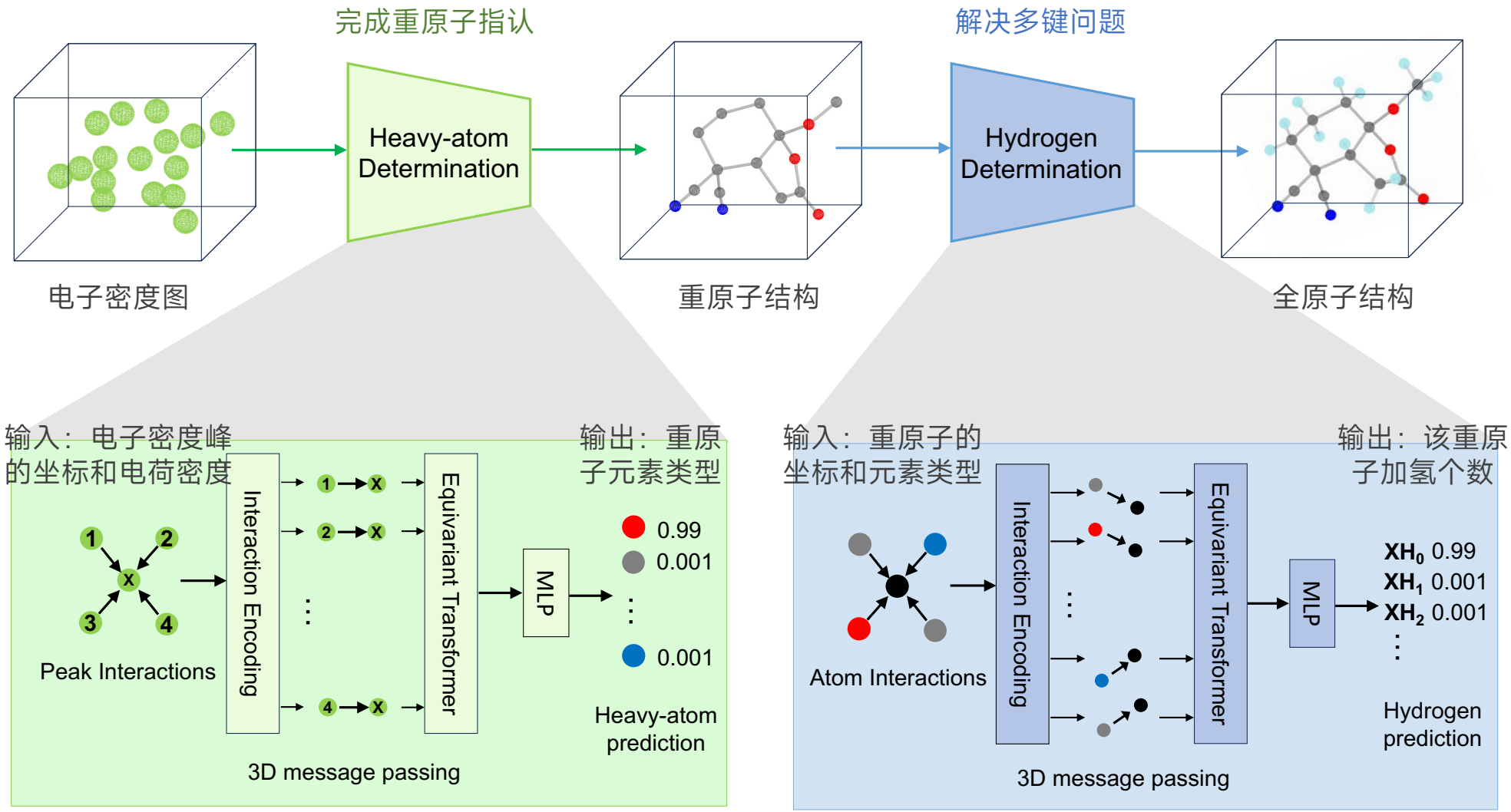


晶体结构

计算理论衍射强度分布，比对修正，反复迭代

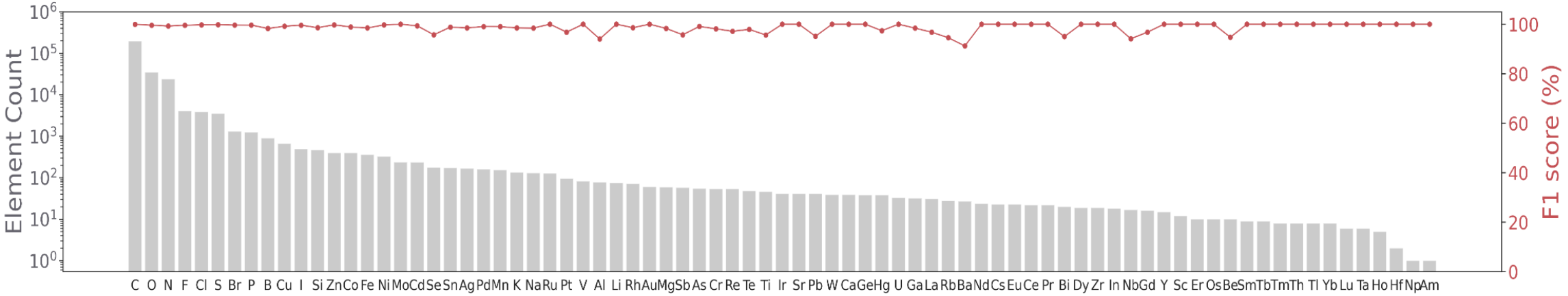
- **高效性**：人工求解非常依赖经验（比如原子指认、加氢），求解耗时较长（约几个小时）
- **准确性**：非 AI 求解算法获得“完全正确”结构的准确率很低（~48%），约有 6% 重原子指认错误
- **鲁棒性**：衍射数据含有噪声时，计算机辅助算法容易过拟合，导致准确率低

基于深度学习的求解算法——CrystalX

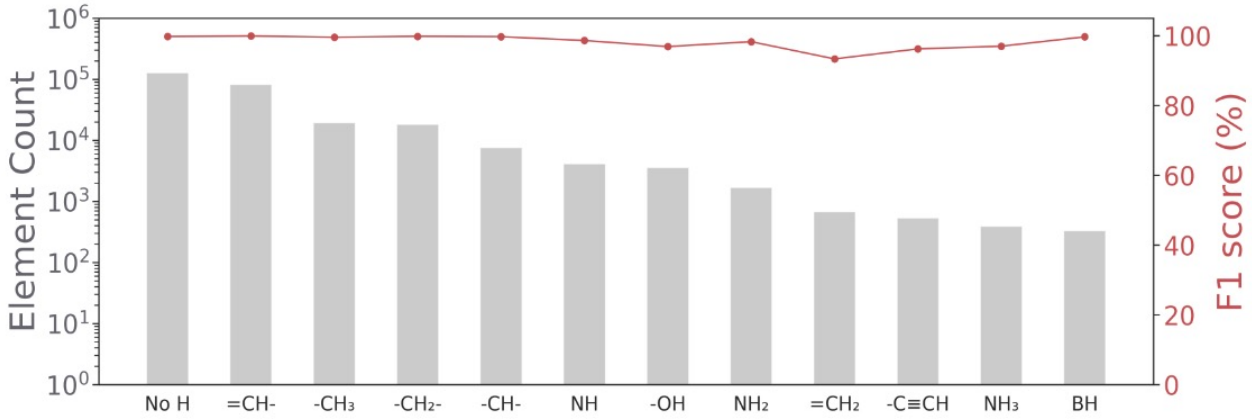


二阶段算法：使用两个几何深度学习的 GNN，根据电子密度图来确定重原子元素类型和加氢个数

CrystalX 算法性能（准确性）



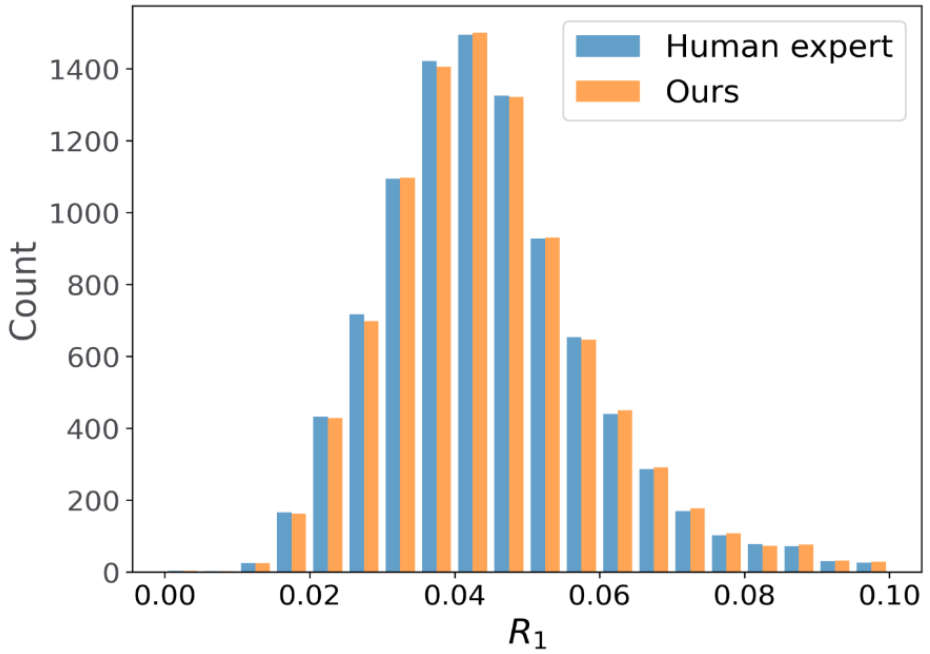
重原子指认的准确率：越往右的元素训练时见过的次数越少，因此准确率会比左边的元素稍低

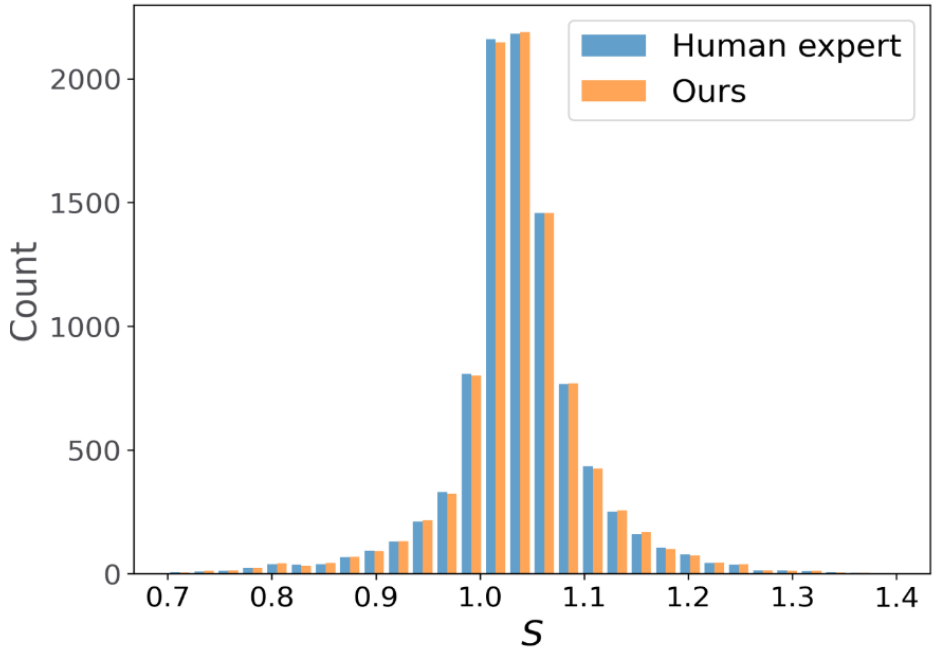


预测加氢个数的准确率：

左图展示了出现频次最高的几种加完氢后的基团，以及对应的准确率

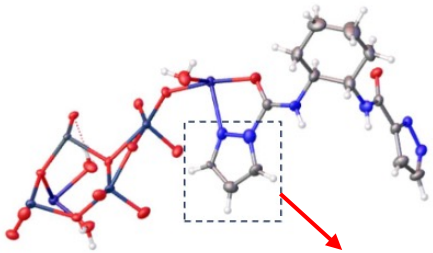
CrystalX 算法性能（准确性）





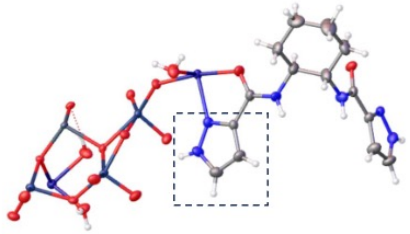
综合性能：在晶体学指标上（R_1 因子，拟合优度）达到 人类专家相当的水平

CrystalX 算法性能（准确性）

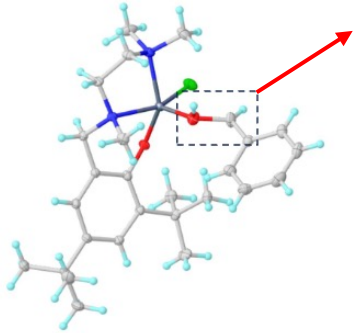


R_1=3.21%

原子指认错误！

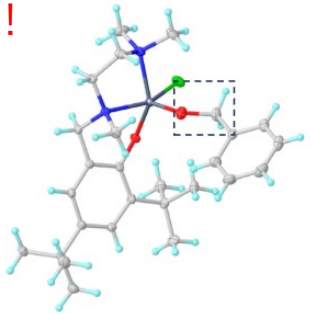


R_1=3.02% (↓0.19%)

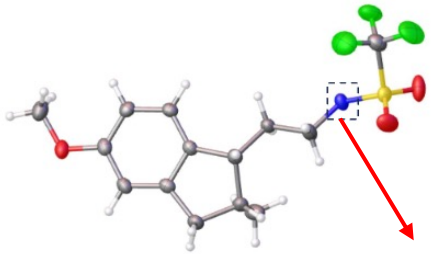


加氢个数错误！

R_1=3.27%

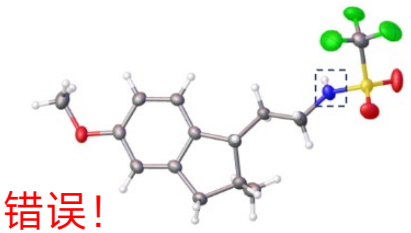


R_1=3.01% (↓0.26%)



加氢个数错误！

R_1=2.82%

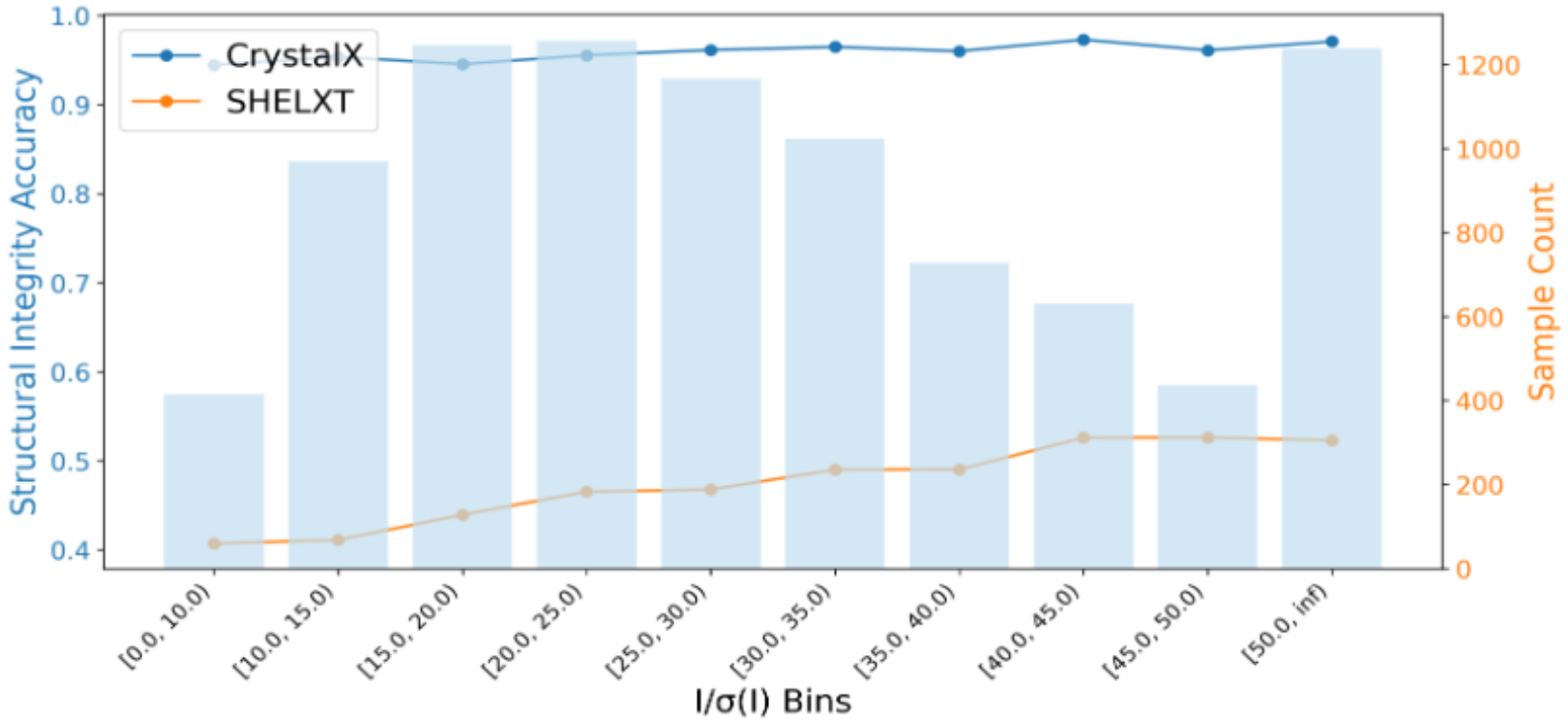


R_1=2.62% (↓0.20%)

文献纠错：我们还在 1500+ 篇 SCI 1 区期刊中找出 9 处错误的晶体解析，包括化学 TOP 1 期刊 JACS 上的 2 处错误。

CrystalX 算法性能（鲁棒性）

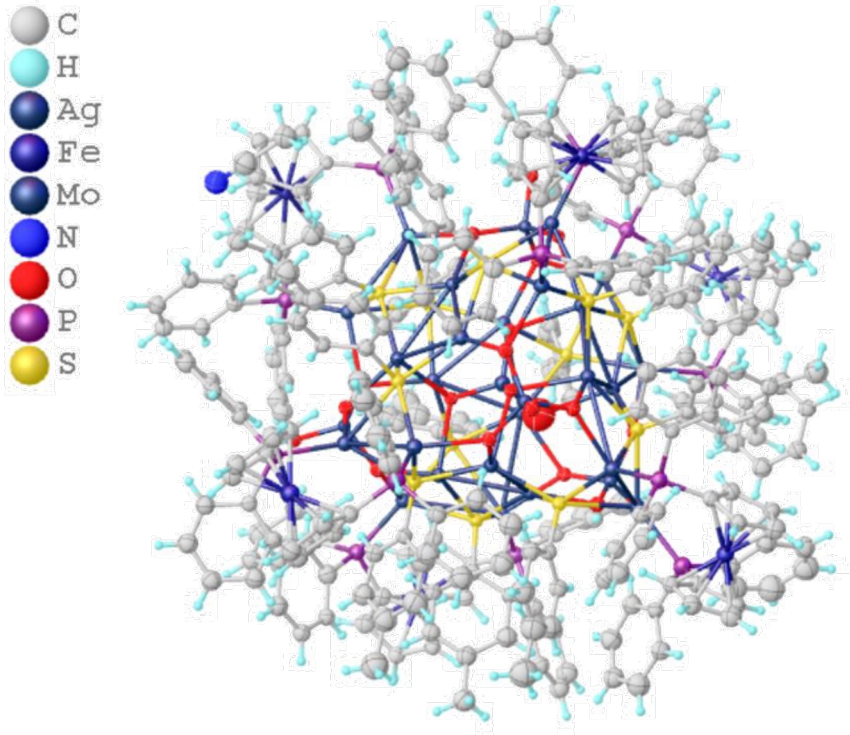
模型性能和衍射数据的信噪比关系



CrystalX 算法的平均准确率 超过 95%，大幅超过 SHELXT 传统方法

而且，在衍射数据 噪声较大 的情况下，仍能保持 良好的准确性

CrystalX 算法性能（高效性）



只需 几十秒 就能 完全正确 地解析出包含 370 个重原子 的复杂庞大分子结构

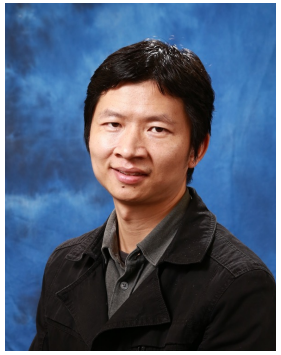
晶体解析时间：小时 → 秒级（专家级准确率）

未来的改进方向

- 构建从衍射数据到晶体结构的“端到端”模型
  - 避免相位解析过程中引入额外噪声
- 扩展到 XRD 多晶结构解析
  - 晶体朝向杂乱无章，因此衍射分布图中的峰会非常多（复杂噪声）

感谢本项目所有的合作者：

郑凯鹏
上海交通大学 / 创智学院



欧阳万里
上海 AI Lab

钟翰森
上海 AI Lab / 创智学院

李玉强
上海 AI Lab / 创智学院

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

感谢您的观看！

Thank you!

SJTU
饮 水 思 源 · 爱 国 荣 校