政府许可

提交的稿件由阿贡国家实验室（“Argonne”）的运营方 UChicago Argonne, LLC 创作。阿贡国家实验室是美国能源部科学办公室下属的实验室，根据合同编号 DE-AC02-06CH11357 运营。美国政府为自身及代表其行事的其他方保留一项已付清、非独占、不可撤销的全球许可，可通过政府或代表政府复制该文章、编制衍生作品、向公众分发副本，并公开表演和展示该文章。能源部将根据美国能源部公共访问计划，向公众开放这些联邦资助研究的结果。http://energy.gov/downloads/doe-public-access-plan。

arXiv:2603.23367v1 [cond-mat.mtrl-sci] 24 Mar 2026

AlphaDiffract：粉末X射线衍射数据的自动化晶体学分析

Nina Andrejevic[1*†]、Ming Du[1*†]、Hemant Sharma[1]、James P. Horwath[1]、Aileen Luo[1]、Xiangyu Yin[1]、Michael Prince[1]、Brian H. Toby[1],和 Mathew J. Cherukara[1*]

[1]先进光子源，阿贡国家实验室，美国伊利诺伊州莱蒙特 *60439* [†]贡献相同的共同第一作者。

[*]电子邮箱：nandrejevic@anl.gov，mingdu@anl.gov，mcherukara@anl.gov

摘要

利用粉末X射线衍射（PXRD）数据进行材料鉴定和结构理解，是材料科学领域长期面临的挑战，也是发现和表征新型材料的基础。实现完整结构解析的前提是准确确定晶格，包括晶格参数和晶体学对称性。传统方法具有迭代性，通常需要专家参与；尽管现有深度学习方法已展现出潜力，但如何从实验数据中一次性全面确定晶格，仍是一个关键目标。在此，我们提出 AlphaDiffract，这是一种深度学习框架，可直接从粉末X射线衍射图谱中预测晶系、空间群和晶格参数，达到当前最先进的性能。AlphaDiffract采用ConvNeXt架构的一维改编版。ConvNeXt是一种现代卷积神经网络，融合了变换器的关键设计原则；该框架还为每种晶体学性质配备了专用预测头。模型使用迄今规模最大的基于物理原理的数据集进行训练，该数据集包含超过3100万条模拟衍射图谱，由对ICSD和Materials Project数据库中312,267个精选结构进行增强生成。尤为重要的是，该模型对实验数据展现出强大的泛化能力：在RRUFF数据集上，晶系预测准确率达到81.7%，空间群预测准确率达到66.2%，并且还能预测全部六个晶格参数。通过提供一个能够根据粉末X射线衍射数据快速、准确确定晶格的统一模型，AlphaDiffract为运用深度学习推动高通量材料发现迈出了重要一步。

1 引言

粉末X射线衍射（粉末X射线衍射）是材料科学中的一项基础表征技术，对于发现和理解新型材料至关重要。确定晶格，即将一维衍射图谱解码为三维晶胞几何结构，通常是根据粉末X射线衍射数据进行完整结构解析的前提 [1]。虽然单晶测量中这属于常规操作，但粉末衍射面临着重大挑战。传统指标化算法可利用高质量数据完成这一任务 [2–4], ，实验技术的进步也进一步提高了可靠性；例如，同步辐射光源可消除样品偏移不确定性，而高分辨率测量即使面对间距很近的衍射反射，也能准确确定峰位 [5]。确定晶格参数后，概率方法可根据系统消光和衍射条件，协助识别相容的空间群 [6]。然而，在许多现实情形中，这些方法仍难以奏效，尤其是在材料含有杂质，或实验条件导致峰展宽和重叠时。

近年来，深度学习方法已展现出根据粉末X射线衍射数据进行自动化晶体学分析的潜力。卷积神经网络（卷积神经网络）是应用最广泛的架构 [7–16] ，因为它们通过多尺度感受野同时捕捉局部峰特征和全局图谱结构。多层感知机 [8, 11, 17] 和变换器 [11] 等其他架构也已得到探索。这些方法通常在基于无机晶体结构数据库 [18] 和 Materials Project [19], 等晶体学数据库生成的大型合成数据集上进行训练，并经常采用数据增强策略来提高稳健性 [12, 15]。主要由实验图谱构成的 RRUFF数据库 [20] 已成为评估其对真实世界数据泛化能力的关键基准 [8, 12, 15, 16]。近年来，包括大语言模型 [21, 22] 和扩散模型 [16], 在内的生成式方法也已被应用

于根据粉末X射线衍射图谱获得完整的结构预测，不过前者高度依赖于对化学式的了解，而后者则需要分别预测组成和晶格参数。
尽管取得了这些进展，从粉末X射线衍射数据确定晶格仍面临重大挑战。补充表1详细比较了近期基于粉末X射线衍射的预测方法，重点介绍其架构、训练数据来源和预测输出。现有大多数研究仅关注晶系和空间群分类[7–9, 12, 13, 15]。预测晶格参数的方法，要么针对每种布拉维晶格或晶系分别训练模型 [10, 17], ，因而需要预先了解对称性；要么依赖化学组成作为输入，并且需要对实验数据进行后续精修 [14]。此外，主要使用理想化合成数据训练的模型，往往难以应对真实的实验条件，包括噪声、峰展宽和仪器效应。如何构建统一框架，在同时预测晶系、空间群和晶格参数的同时，仍能在实验数据上保持稳健的性能，仍是一项尚待解决的挑战。

为应对这些挑战，我们提出了AlphaDiffract，这是一个用于根据粉末X射线衍射图谱确定晶格的统一深度学习框架。我们的方法在RRUFF数据集上实现了最先进的性能，晶系准确率达到81.7%，空间群准确率达到66.2%，同时预测全部六个晶格参数。在借鉴先前研究成果的基础上，我们的框架融入了三项关键创新：

1. 尽管卷积神经网络仍是分析粉末X射线衍射图谱等序列科学数据的一种常用且有效的选择，但要设计一种既擅长识别单个衍射峰等局部特征，又能捕捉编码晶体学对称性的复杂长程依赖关系的架构，仍颇具挑战。为此，我们采用了 ConvNeXt 架构的一维改编版本 [23]。ConvNeXt 是一种现代卷积神经网络架构，改编自 ResNet [24], ，并融入了变换器的关键设计特征，从而提升了参数效率和对长程相互作用的建模能力。这些改进使 ConvNeXt 非常适合粉末X射线衍射数据分析：该架构在保留卷积神经网络用于模式识别的空间归纳偏置的同时，也继承了变换器处理序列数据的关键优势。

2. 我们的模型使用模拟粉末X射线衍射图谱进行训练，这些图谱由无机晶体结构数据库和 Materials Project数据库中的结构组合生成；同时，我们保留 RRUFF数据库作为独立基准，用于评估泛化能力。按照方法部分5.1中详述的整理流程，合并后的无机晶体结构数据库与 Materials Project数据集包含 312,267 个晶体结构。对于每种结构，我们模拟了 100 个增强衍射图谱，并施加随机扰动以模拟微应变和晶粒尺寸展宽等物理效应（见第2.1节）。这种数据增强策略最终生成了一个包含超过 3,100 万条衍射图谱的训练集。据我们所知，这是迄今为止为该任务编制的最大、最全面的数据集。

3. 我们方法的一项关键设计特征是采用单一的端到端模型，仅根据粉末X射线衍射图谱，同时预测晶系、空间群和晶格参数。这种统一的模型架构简化了晶体学分析流程，无需为每项任务分别使用专门的模型，并能够在单个推理步骤中完成晶格的全面确定。

在以下各节中，我们将详细介绍数据生成流程、模型架构与训练，以及在模拟和实验基准数据集上的全面评估。

2 结果

2.1 数据准备与基于物理的模拟

为生成与实验结果相匹配的粉末衍射图谱，我们通过 GSASIIscriptable API [25, 26]使用 GSAS-II 晶体学软件包。GSAS-II 专为里特维尔德分析而设计，因此能够对真实仪器数据进行定量准确的模拟。我们对来自 NIST ICSD 数据库[27] 和 Materials Project [19] 的结构进行模拟，以创建训练数据集。所有模拟均采用能量为 20 keV 的单色 X 射线源，在 2θ 至 20° 的 5° 范围内，以 8192 个等间距点进行。为降低模型对不同衍射仪器间变化的实验参数的敏感度，我们针对每种结构生成了 100 组模拟，并随机设置仪器参数和样品参数。每次模拟均在表 5 所列范围内随机抽取微应变和晶粒尺寸（两者都会导致洛伦兹展宽）以及高斯仪器展宽参数，从而生成具有多样峰形、可代表不同实验条件的图谱 [28]。训练期间，还按照方法第 5.2 节所述，对每个图谱动态添加额外的泊松噪声和高斯噪声，随后将强度标准化至 0 到 100，再输入模型。

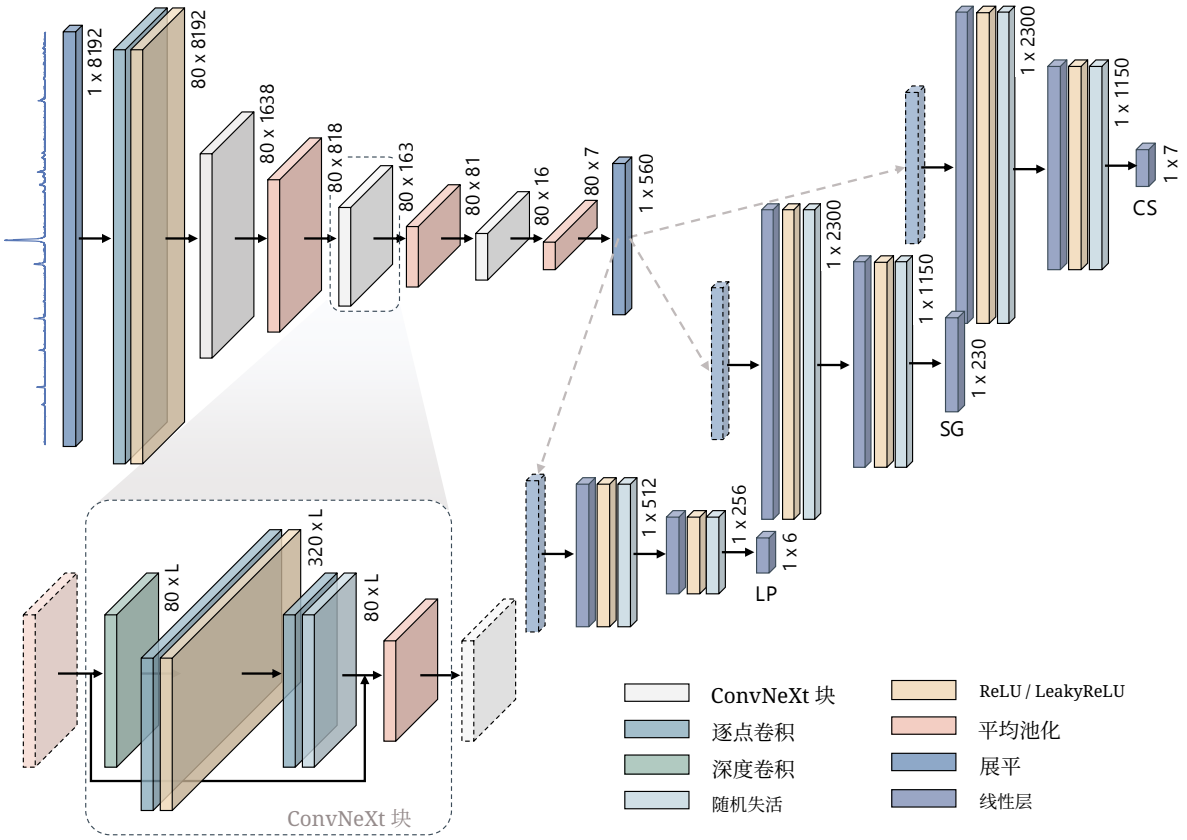


图 1：**AlphaDiffract 模型架构**。AlphaDiffract 模型由一维 ConvNeXt 主干网络组成，该网络通过一系列 ConvNeXt 块处理输入的粉末X射线衍射图谱，并逐步进行下采样。每个 ConvNeXt 块的组成如左下角插图所示。提取的特征被馈送至三个独立的预测头：晶系 (CS) 分类器、空间群 (SG) 分类器和晶格参数 (LP) 回归器。每个预测头均采用多层感知机架构，其层维度如图所示。

2.2 AlphaDiffract 架构

AlphaDiffract 模型通过结合三个专用预测头的 1D ConvNeXt 骨干网络，从粉末X射线衍射图谱中提取晶体学信息。如图 1 所示，骨干网络充当特征提取器，为三个不同的多层感知机（MLP）预测头提供输入，以进行针对性预测。该端到端模型能够同时预测晶系（CS）、空间群（SG）和全部六个晶格参数（LP）。模型训练和评估所用的晶格参数均取自尼格利约化晶胞。尼格利约化可提供唯一的标准晶胞表示，从而避免因晶胞设定选择不同而产生歧义。

特征提取器是对 ConvNeXt 架构 [23], 的一维改编。ConvNeXt 是一种卷积神经网络（CNN），融合了视觉变换器的若干原理，例如深度可分离卷积、倒置瓶颈设计和较大的卷积核尺寸。输入的粉末X射线衍射图谱是一个 1×8192 向量，经过一系列 ConvNeXt 块处理。每个块如图 1 插图所示，先使用深度卷积捕获空间模式（即峰形和局部排列），再通过逐点卷积学习通道间交互。该结构能够以较高的参数效率对图谱中的关系进行建模。每个块之后，平均池化会逐步对特征图进行下采样，从而扩大网络的感受野，使其能够捕获相距较远的衍射峰之间的长程依赖关系。骨干网络的最终输出是一个 560 维特征向量，它编码了原始衍射图谱中的关键特征，以支持后续的分类和回归任务。

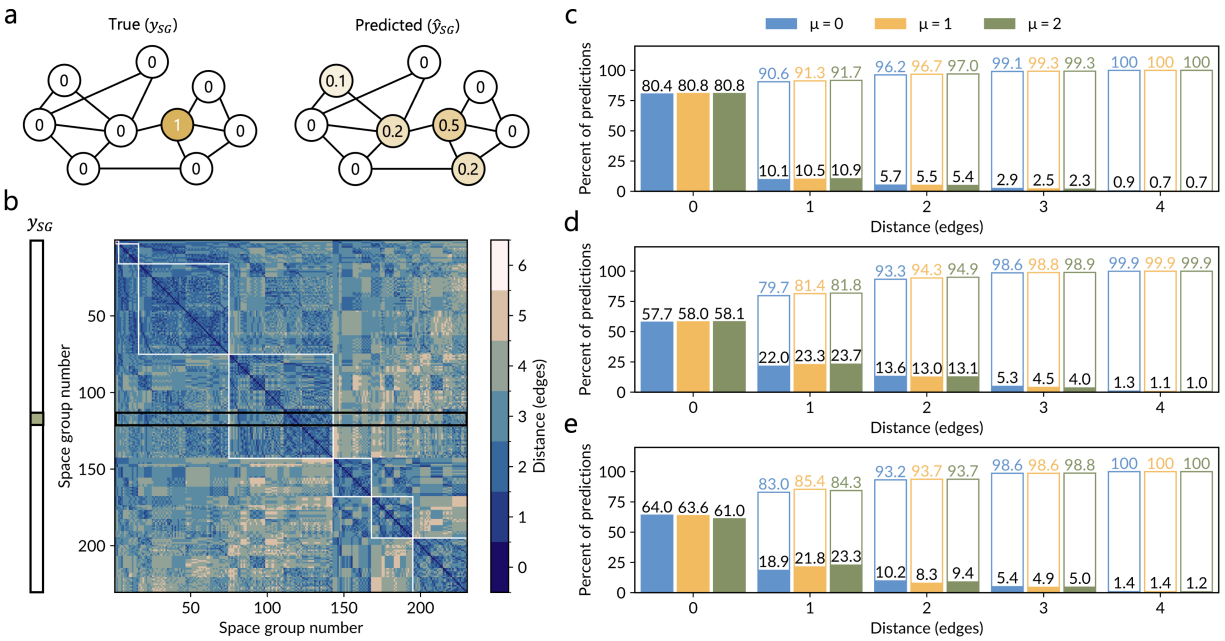


图2：**使用图地球搬运工距离评估空间群预测。a.** 代表性子群图上真实（y_{SG}）和预测（$\hat{y}_{\mathrm{SG}}$）空间群概率分布的示意图，其中节点表示空间群，边表示极大子群关系。真实标签将概率1赋给单个节点（黄色），而预测分布通常会将概率分散到多个节点。**b.** 根据空间群之间的极大子群关系计算得到的距离矩阵，颜色深浅表示连接每对空间群所需的最小图边数。垂直向量（y_{SG}）表示独热编码真实标签，它选取相应行，用于计算预测分布的GEMD损失。**c-e.** 三个数据集的预测误差分布随图距离（边数）的变化情况：**c.**ICSD验证集、**d.** Materials Project验证集，以及**e.** RRUFF测试集。实心柱表示所有预测（包括距离为0的正确预测）中，与真实空间群的图距离分别为各距离的百分比，展示三种不同的GEMD损失权重（$\mu = 0$、1、2）。距离为0表示预测正确（预测空间群= 真实空间群）。带标注数值的空心柱表示截至该距离（含该距离）的相应累积百分比。随着 μ 值增大，预测结果越来越集中在距真实空间群较近的图距离上。

该特征向量被传入三个专用的多层感知机预测头，每个预测头针对一项特定任务进行设计：用于晶系分类的 7 节点输出、用于空间群分类的 230 节点输出，以及用于晶格参数回归的 6 节点输出。这种多任务方法使模型能够在骨干网络中学习共享表示，同时针对每种不同的晶体学性质微调预测结果。有关架构参数的详细信息，请参见方法第5.3节。

2.3 物理感知损失函数

AlphaDiffract 的一项关键创新是在标准交叉熵之外，为空间群分类引入了一种新型损失函数。交叉熵对所有误分类一视同仁，而空间群通过对称性层级在结构上彼此相关：关系密切的子群或超群更为相似，因此某些误分类的严重程度低于其他误分类。为纳入这些关系，我们采用图地球搬运距离（GEMD）损失，将传统的地球搬运距离应用于极大子群图。该图中的节点表示空间群，边连接具有直接极大子群关系的群，反映它们在结构上的相似性。这种方法受到 Vecsei等人 的研究启发；该研究使用极大子群距离作为指标评估空间群预测性能 [29]，而我们进一步将这一概念发展为一种损失函数，在训练过程中约束这些结构关系。GEMD损失使用预先计算的 230 × 230 距离矩阵 D，其中每个元素D_{ij} 编码了空间群 i 与 j 之间在此图上的跳数（图2**a** 和**b**）。这些跳数由毕尔巴鄂晶体学服务器 [31]上提供的 MAXSUB程序 [30] 确定。如图所示

图2**a**中，真实标签将概率1赋给单个空间群，而预测分布通常会将概率分散到多个节点上。GEMD损失计算将概率质量从预测分布迁移到真实分布所需的代价，该代价由矩阵 D中的距离加权。此方法根据预测与真实空间群之间的晶体学差异来惩罚误分类：在子群图中，预测结果距离真实空间群越远，受到的惩罚就越大。损失函数包含一个超参数 μ，用于控制GEMD项相对于交叉熵的权重，使我们能够在训练过程中平衡这两个目标。有关损失函数的更多细节，请参阅方法第5.4节。

为了确定 GEMD 损失项的最佳权重，我们仅使用晶系和空间群预测头（不包括晶格参数），分别采用 μ (0, 1, 2) 三种不同取值训练模型，其中 $\mu = 0$ 对应于仅使用标准交叉熵。我们在主要的 RRUFF测试集以及无机晶体结构数据库和 Materials Project验证集上评估这些模型（图 2**c-e**）。在这些图中，图距离为零表示预测完全正确；非零距离则表示分类错误，即预测的空间群与真实空间群在晶体学上相隔子群图中所示数量的边。我们发现，加入 GEMD损失会使误差分布向距真实空间群更近的图距离移动；也就是说，当模型预测错误时，它倾向于预测在子群层级中与真实空间群晶体学上更接近的空间群。这一规律在 RRUFF测试集以及无机晶体结构数据库和 Materials Project验证集中均一致。累积百分比表明，使用 $\mu = 1$时，RRUFF 上超过 85% 的预测结果正确，或与真实空间群仅相差一个对称生成元（相隔 1 条边）。根据这些结果，我们选择 $\mu = 1$，用于训练包含全部三个预测头的完整模型。

2.4 AlphaDiffract 在分类任务上的表现

确定 $\mu = 1$ 为最优 GEMD损失权重后，我们训练了一个包含全部三个预测头（晶系、空间群和晶格参数）的10模型集成。图3**a**显示，在 RRUFF 数据集上，晶系和空间群的准确率均随集成规模增大而提升，但超过8—10个模型后收益递减，这证明我们选择10模型集成是合理的。表1和表2汇总了纯分类模型（分类）和分类与回归联合模型（分类 + 回归）在三个数据集上的分类性能，同时报告了10个模型各自的结果（取平均值）以及通过汇总其预测得到的集成结果。我们还列出了朴素多数类分类器的基线性能以供参考。在我们的 RRUFF测试集上，集成模型的晶系分类准确率和空间群分类准确率分别达到81.74 ± 0.78%和66.21 ±0.71%；不确定性反映了集成效应与增强变异性的综合影响（见方法第5.6节）。这些结果显著高于相应的基线准确率（26.98%和3.27%），表明模型学到了有意义的晶体学模式，而非利用类别不平衡。与此前在 RRUFF 数据上评估的研究相比，我们的结果也颇具优势。在晶系分类方面，我们的集成模型优于 Lee等人的 Large FCN（74.24%） [12] 和 Salgado 等人的 NPCNN（74 %） [15], , 而在

	模型变体	ICSD	Materials Project	RRUFF
基线	–	20.13	20.63	26.98
帕克等人 [7]	–	**94.99**	–	–
维切等人 [8]	密集	73	–	70
Vecsei等 [8]	卷积	85	–	56
Lee等 [11]	FCN	92.12	**82.17**	–
Lee等 [12]	Large FCN	92.10	–	74.24
Salgado等 [15]	Large，NPCNN	–	–	74
Choudhary [21]	DGPT公式	18.16 (19.80)†	30.10 (32.10)†	28.75 (26.57)‡
Ours	分类 ($\mu = 1$)	90.60 ± 0.38†	76.75 ± 0.60†	76.70‡
Ours	分类 +回归 (平均值)	90.52 ± 0.38†	76.47 ± 0.60†	78.62‡
Ours	分类 +回归 集成	92.22 ± 0.40†	79.00 ± 0.64†	**81.74** ± **0.78**‡

表1：**AlphaDiffract 和参考模型的晶系分类准确率**。部分参考研究训练了多个采用不同架构和/或数据集的模型；此处仅展示在训练集中不包含 RRUFF 数据时表现最佳的模型变体。“基线”表示始终预测训练数据中占比最高类别的朴素分类器。对于我们的方法，“平均值”表示 10 个训练模型各自独立性能的平均值，而“集成”则指这些模型的集成平均预测。误差条表示集成不确定性和增强不确定性的汇总 95%置信区间。

†在我们的验证集数据上进行评估。对于参考文献 21 中 DGPT公式 模型的推理，我们从每个数据集的验证集中选取了 1000 个代表性样本进行评估。括号内的分数表示在未添加泊松噪声或高斯噪声的合成 PXRD 图谱上的结果。

‡在我们的测试集数据上进行评估。

	模型变体	ICSD	Materials Project	RRUFF
基线	–	7.01	6.36	3.27
帕克等人 [7]	–	81.14	–	–
Vecsei 等人 [8]	密集	57	–	54
Vecsei 等 [8]	卷积	76	–	42
Liang等 [17]	–	78.77*	–	–
Lee等 [11]	FCN	79.67	**69.01**	–
Lee等人 [12]	Large FCN	**84.85**	–	58.82
Salgado等人 [15]	Large, NPCNN	–	–	66
Choudhary [21]	DGPT公式	3.99（5.43）†	10.05 (13.38)†	5.46 (5.48)‡
Ours	分类 ($\mu = 1$)	80.79 ± 0.54†	58.04 ± 0.73†	63.62‡
Ours	分类 +回归（平均值）	79.96 ± 0.55†	57.41 ± 0.73†	64.55‡
Ours	分类 +回归 集成	81.75 ± 0.59	59.54 ± 0.79†	**66.21** $\pm$ **0.71**‡

表2：**AlphaDiffract 和参考模型的空间群分类准确率**。部分参考研究训练了多个采用不同架构和/或数据集的模型；我们仅展示未将 RRUFF 数据纳入训练集时表现最佳的模型变体。“基线”表示一种朴素分类器，它始终预测训练数据中最常见的类别。对于我们的方法，“平均值”表示 10 个训练模型的独立性能平均值，而“集成”则指这些模型的集成平均预测。误差条表示集成不确定性和增强不确定性汇总后的 95% 置信区间。

*Weighted 对 mo 求平均值 dels 规格ialized f或 eac h 布拉瓦is lattice.

†在我们的验证集数据上评估。对于参考文献 21 中 DGPT-formula 模型的推理，我们从每个数据集的验证集中选取了 1000 个代表性样本进行评估。括号中的分数表示在未添加泊松噪声或高斯噪声的合成 PXRD 图谱上的结果。

‡在我们的测试集数据上评估。

在空间群分类方面，它超过了 Lee等人的 Large FCN（58.82%），表现与 Salgado等人的 NPCNN（66%）相当。

为了与主要在合成数据上评估的模型进行比较，我们还报告了在无机晶体结构数据库和 Materials Project 验证集上的结果；我们的集成模型在这两个验证集上的晶系准确率分别为 $92.22 \pm 0.40\%$ 和 $79.00 \pm 0.64\%$，空间群准确率分别为 $81.75 \pm 0.59\%$ 和 $59.54 \pm 0.79\%$。此前的一些研究在 ICSD 合成数据上取得了略高的准确率，包括 Park等人 （94.99% 晶系） [7] 以及 Lee等人 （84.85% 空间群） [12],；在 Materials Project 合成数据上，它们的表现也超过了我们的模型，例如 Lee等人 （82.17% 晶系，69.01% 空间群） [11]例如，。不过，在解读这些比较时，应考虑训练所用的数据生成流程。我们的训练数据经过了大量增强，涵盖广泛的样品和仪器展宽参数范围，并包含根据 RRUFF 图谱得出的真实泊松噪声和高斯噪声水平（表 5 和补充图4）。虽然一些先前研究描述了其合成数据生成中的参数变化，但文献中并不总是充分说明具体范围和噪声模型，这可能影响训练鲁棒性及对实验数据的泛化能力。无论是在我们的工作还是先前研究中，合成数据集与实验数据集之间的性能差距都凸显了准确模拟真实世界粉末X射线衍射测量完整复杂性的难度。值得注意的是，在分类之外加入晶格参数回归（分类 + 回归）后，分类性能与纯分类模型（分类）相比保持不变或略有提升，表明额外的回归任务不会损害分类准确率。将 10 个独立训练的模型集成为一个集成模型（分类 + 回归集成）后，所有数据集上的性能进一步提升；在 RRUFF 上，晶系分类和空间群分类的最佳结果分别达到 81.74% 和 66.21%。图3**b** 展示了集成模型在全部三个数据集上的空间群预测误差随图距离变化的分布，表明 GEMD损失能成功地将误差集中在距离真实空间群较近的位置；在 RRUFF 数据中，超过 87% 的预测结果与真实空间群相距不超过 1 条边。

2.5 AlphaDiffract 在回归 任务

对于晶格参数预测，我们评估了集成模型在相同三个数据集上的回归性能（表 3，图 3**c**-**e**）。作为参考，我们还纳入了朴素模型的基线性能：该模型对所有输入均预测训练数据中的平均晶格参数。在 RRUFF 数据集上，集成模型的晶格长度平均绝对误差（MAE）为 2.11 ± 0.02 Å，晶格角度平均绝对误差为 2.72 ± 0.03^{o}；对应的平均绝对百分比误差（MAPE）分别为 $23.50 \pm 0.25\%$ 和 $2.91 \pm 0.03\%$，相比基线预测有明显改善。决定系数（R^2）在晶格长度方面为 0.39 ± 0.01，在晶格角度方面为 0.25 ± 0.01。补充表 2 还报告了按各晶格参数分别评估的指标。我们注意到，尽管这些结果显著优于朴素基线和现有的统一模型方法，但

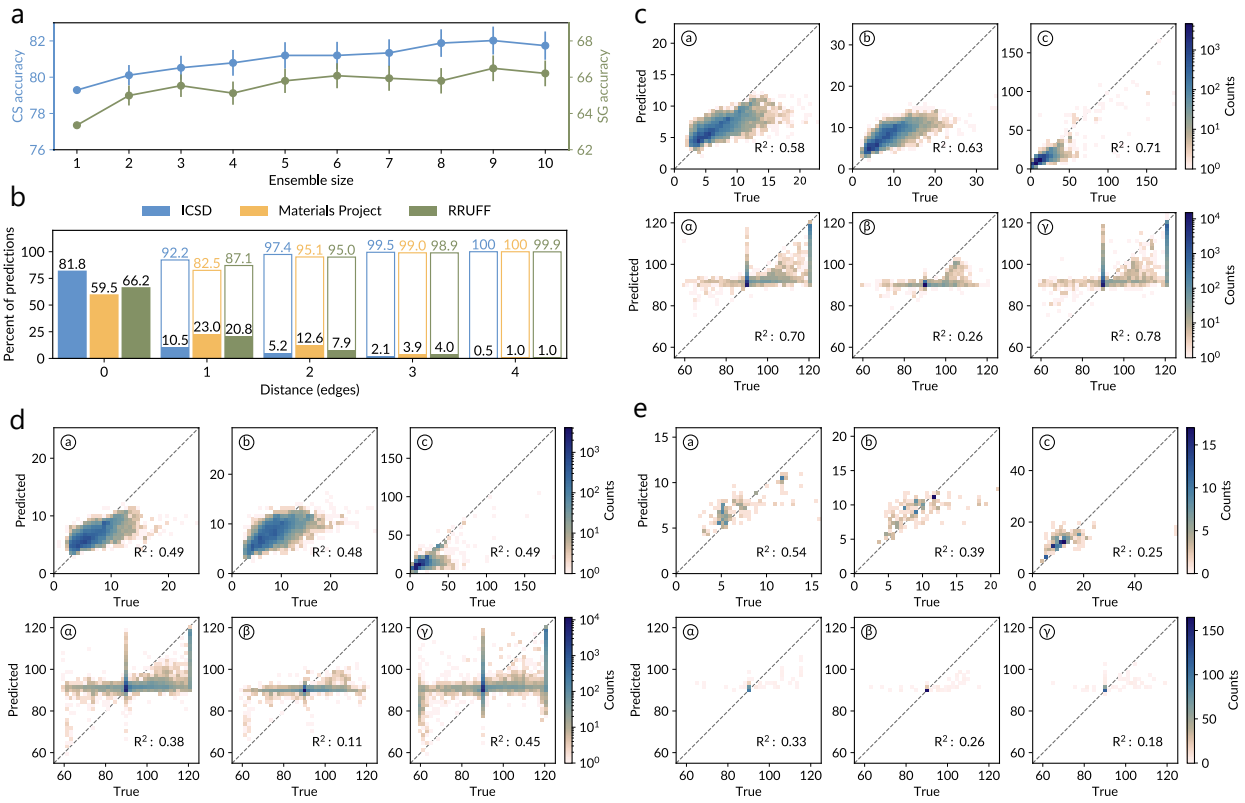


图3：**AlphaDiffract 集成模型的性能**。**a.** RRUFF数据集上晶系（CS）和空间群（SG）预测准确率随集成规模的变化。误差条表示集成模型预测的不确定性。**b.** 在三个数据集上评估的10模型集成中，预测误差相对于真实空间群的图距离（边数）分布。未填充的柱及其标注值表示预测结果落在距真实空间群该图距离范围内的累计百分比（即截至该距离（含该距离）的所有填充柱之和）。**c-e.** 三个数据集上10模型集成的预测晶格参数与真实晶格参数对比的奇偶图：**c.** 无机晶体结构数据库，**d.** Materials Project，以及**e.** RRUFF。每个面板展示三个晶格长度（a、b、c；上排）和三个晶格角（α、β、γ；下排）的预测结果。虚线表示完全一致。热图颜色表示点密度。R^2 值表示拟合优度。

其绝对精度，尤其是晶格长度的 ~20% MAPE，尚不足以直接用作 Pawley 或 Le Bail 拟合等全谱精修方法的初始值；这类方法通常要求晶胞参数精度达到百分之几以内。因此，AlphaDiffract 的预测更适合被视为快速、粗略的晶格估计，可用于指导后续精修，而不能取代精修。

在合成数据上，集成模型也取得了出色的性能：在无机晶体结构数据库上，长度和角度的平均绝对误差分别为 1.59 Å 和 1.68^o（R^2= 0.64 和 0.58）；在 Materials Project 上则分别为 1.92 Å 和 3.54^o（R^2 = 0.48 和 0.32）。Chitturi 等人在无机晶体结构数据库数据上报告了更低的平均绝对百分比误差（9.2%）；不过，他们的方法针对七大晶系分别训练独立模型。由于 AlphaDiffract 为所有晶系采用单一的统一模型，因此对于晶系可能事先未知的晶格测定任务，它或许更能反映实际应用场景。先验地。图 **3c-e** 展示了三个数据集上集成模型预测的晶格参数与真实值的对比图。图中显示，预测值总体上与真实值具有良好的相关性，不过与合成的无机晶体结构数据库衍射图谱相比，RRUFF 数据的离散程度更大。晶格角的 R^2 值明显低于晶格长度，这可以解释为：除单斜晶系和三斜晶系外，各晶系的角度均受对称性约束而取固定值（例如 90°）。模型经常能正确预测这些固定值，从而形成图 **3c-e**中可见的密集簇，但这同时也会导致较低的 R^2 值，因为该指标衡量的是解释方差。合成数据与实验数据在晶格方面的性能差距

数据集	模型	晶格长度			晶格角度		
		平均绝对误差 (Å)	平均绝对百分比误差 (%)	R^2	平均绝对误差 (°)	平均绝对百分比误差 (%)	R^2
ICSD	基线	3.18	44.89	0.00	4.56	4.48	0.00
	Chitturi 等 [10]	–	**9.2***	–	–	–	–
	Liang等 [17]	–	–	0.45**	–	–	0.19**
	Choudhary [21]	2.85 (2.79)†	34.32 (33.66)†	-0.03 (-0.06)†	5.76（6.23）†	5.82 (6.33)†	-2.07 (-2.80)†
	Ours	**1.59 ± 0.00**†	19.77 ± 0.04†	**0.64 ± 0.00**†	**1.68 ± 0.01**†	**1.69 ± 0.01**†	**0.58 ± 0.00**†
Materials Project	基线	2.89	34.76	-0.01	5.22	5.55	0.00
	Chitturi 等 [10]	–	–	–	–	–	–
	Liang等 [17]	–	–	–	–	–	–
	Choudhary [21]	2.69 (2.69)†	28.14 (27.04)†	-0.01 (-0.01)†	5.97 (6.29)†	6.25 (6.61)†	-1.54 (-1.99)†
	Ours	**1.92 ± 0.00**†	**22.59 ± 0.04**†	**0.48 ± 0.00**†	**3.34 ± 0.01**†	**3.61 ± 0.01**†	**0.32 ± 0.00**†
RRUFF	基线	2.86	31.03	-0.08	4.19	4.49	-0.07
	Chitturi 等人 [10]	–	–	–	–	–	–
	Liang等人 [17]	–	–	–	–	–	–
	Choudhary [21]	2.53 (2.40)‡	25.94 (25.32)‡	-0.11 (-0.12)‡	5.32（5.39）‡	5.73 (5.72)‡	-2.34 (-2.31)‡
	Ours	**2.11 ± 0.02**‡	**23.50 ± 0.25**‡	**0.39 ± 0.01**‡	**2.72 ± 0.03**‡	**2.91 ± 0.03**‡	**0.25 ± 0.01**‡

表 3：**AlphaDiffract 与参考模型的晶格参数预测误差**。误差以平均绝对误差（MAE）、平均绝对百分比误差（MAPE）和决定系数（R^2）衡量。为便于直接比较，我们仅分析采用回归指标而非匹配率等分类指标量化预测准确性的研究。由于针对 RRUFF 进行测试的研究较少，我们也列出在其他数据集上测试的研究，仅供参考。“基线”指一种朴素模型，它对所有输入均预测训练数据中的平均晶格参数。误差条表示集成模型和数据增强预测结果的汇总标准差。

*针对每种晶系分别专门训练的模型的加权平均值。请注意，这些结果对应于合并后的无机晶体结构数据库/CSD 数据集。**针对每种布拉维晶格分别专门训练的模型的加权平均值。

†在我们的验证集数据上进行评估。对于参考文献 21 中 DGPT-formula 模型的推理，从每个数据集的验证集中选取了 1000 个代表性样本进行评估。括号中的分数表示在未添加泊松噪声或高斯噪声的合成 PXRD 图谱上取得的结果。

‡在我们的测试集数据上进行评估。

参数回归呈现出与分类相似的趋势，这反映了基于存在复杂实验伪影的真实测量结果进行精确的定量预测更加困难。

2.6 推理时间统计

为评估 AlphaDiffract 的计算效率，我们测量了各个模型在三个评估数据集上的推理时间（表4）。单个模型处理每个粉末X射线衍射图谱的平均推理时间为 1.15–1.39 毫秒，在所有数据集上的中位时间为 1.04–1.08 毫秒。这些结果相当于在单个图形处理器上每秒处理 700–870 个样本，因此 AlphaDiffract 非常适用于实时分析和超高通量筛选应用。在规模各异（734 至 10,000 个样本）的数据集上，推理时间保持一致，表明其性能不受数据集规模影响，十分稳定。采用由 10 个模型组成的集成方法时，由于各模型依次生成预测，每个图谱的推理时间近似线性增加至 11.5–13.9 毫秒（每秒 70–90 个样本）。即使推理时间增加了十倍，与每个图谱通常需要数秒至数分钟的 Rietveld 精修相比，集成模型的推理时间仍微不足道。这种快速推理使单模型和集成方法都能作为自动化 Rietveld 精修流程中的高效预处理步骤，在不产生显著额外计算开销的情况下，提供晶系、空间群和晶格参数的初始估计，从而用于晶体学工作流程。

数据集	样本数量	每个样本的平均时间（毫秒）	每个样本的中位数时间（毫秒）	吞吐量（样本数 / 秒）
ICSD	10,000	1.33	1.08	751
Materials Project	10,000	1.39	1.08	719
RRUFF	734	1.15	1.04	871

表4：**推理时间统计**。各评估数据集上的单模型推理时间统计。所有测量均在单个 NVIDIA H100 GPU 上以批量大小 1 执行。

3 讨论

尽管 AlphaDiffract 性能出色，仍存在若干局限，有待未来改进。一个主要挑战是训练数据中固有的类别不平衡。ICSD 和 Materials Project 等晶体学数据库存在严重偏斜：少数常见空间群（例如， $P2_1/c$, $Pnma$）所占比例远高于其他类别，而数十种空间群则十分罕见。补充图1显示了这种不平衡，它也是不同数据集之间性能差异的原因之一。如补充图 5 所示，不同晶系的性能差异显著：立方晶系的准确率最高（在 ICSD、Materials Project 和 RRUFF 数据集上，晶系预测准确率分别为 99.7%、99.3% 和 82.4%，空间群预测准确率分别为 88.1%、76.8% 和 78.2%），而低对称性晶系的表现则更不稳定。三斜晶系和单斜晶系是最具挑战性的类别之一，在所有数据集上的准确率都明显较低。Materials Project 中这类低对称性晶系所占比例更高，其中许多是通过计算预测得到的结构或训练数据中较少见的亚稳态结构；这种组成差异直接导致该数据集的整体分类准确率较低。不过，训练数据偏向更常见、对称性更高的空间群，也与实际材料中遇到这些空间群的可能性相符。

此外，当前模型并未明确考虑择优取向（织构）这一常见的实验伪影，而它会显著改变相对衍射峰强度。由于模型依赖整个图谱（包括强度），因此可能会受到具有强织构的实验数据误导。补充图 6 比较了来自相同晶体结构的实验与合成 RRUFF 图谱的空间群分类性能，结果显示，合成图谱的准确率比实验图谱高 15–17%。这一性能差距可能部分归因于实验测量中存在、而我们的训练数据扰动中未包含的织构效应。例如，通过 GradCAM 进行的注意力分析 [32]（补充图7和图8）显示，在处理实验数据与模拟数据时，模型有时会关注图谱中截然不同的区域，尤其是对于呈现织构效应的结构；这表明模型会根据强度变化调整其分类策略。尽管我们的数据增强提供了一定的鲁棒性，但未来的迭代可考虑将织构模拟纳入训练数据，或设计对强度比不那么敏感的架构。

关于晶格参数预测，我们希望强调，所报告的误差（RRUFF中长度的∼20%平均绝对百分比误差、角度的 ∼3%平均绝对百分比误差）代表着朝向统一晶格测定迈出的重要一步，但与实际Rietveld精修初始化所需的精度相比仍有不足。相较于Chitturi等人提出的晶系专用方法（平均绝对百分比误差为9.2%），这一性能差距反映了完全统一模型所固有的难度：该模型必须先隐式推断晶系，再预测晶格参数。提高这一准确性是未来研究的重要方向，例如采用序列式或分层条件化架构。

未来工作将集中于以下几个方向。首先，我们计划探索替代架构，以改进多任务学习。例如，可以对分支式多任务网络进行序列训练：首先预测晶系和空间群，然后利用该预测结果指导更准确、针对特定晶系的晶格参数预测。最后，我们预计，通过在训练数据中将衍射峰引入随机位置，可以降低模型对额外相存在的敏感性，从而提供一种能够对混合物中的衍射图谱进行索引的独特工具。其次，尽管我们的模型能够预测晶系、空间群和晶格参数，但它目前尚不能确定原子位置——而这是基于粉末衍射数据完成端到端晶体结构测定所需的最后一步。不过，我们的工作解决了这一更广泛挑战中的关键子问题。近期方法已经探索了这一逆向设计问题的不同范式。例如，Riesel等人的Crystalyze模型采用扩散模型，在生成原子位置之前，需要准确的晶格参数、组成和原子数来初始化晶胞 [16]。AlphaDiffract的晶格参数预测可以改进这一初始化步骤，而晶系和空间群预测则可用于施加对称性约束，并引导扩散模型生成与底层晶体学对称性一致的原子位置。替代性的基于大语言模型的方法，例如DiffractGPT [21] 和deCIFer [22], ，遵循不同的原理，但接受可选的条件信息（例如，deCIFer的空间群信息），而我们的预测结果有望进一步增强这些信息。将我们的预测结果与这些生成式方法相结合，是未来朝着完全自动化结构解析发展的一个有前景方向。尽管如此，如本文所示，仅凭实验PXRD数据准确预测晶格参数、晶系和空间群，已经解决了晶体学分析中的一个根本性瓶颈。这一能力无需预先了解组成，也无需针对不同晶系分别建立模型，即可实现快速结构表征，推动高通量材料发现目标的实现。

4 结论

我们介绍了 AlphaDiffract，这是一种用于从粉末X射线衍射图谱确定晶格参数的统一深度学习框架。该框架采用一维ConvNeXt架构，并使用包含超过3100万张模拟图谱的基于物理的训练集

该训练集纳入了真实的实验效应；AlphaDiffract可从单张衍射图谱同时预测晶系、空间群以及全部六个晶格参数。

在 RRUFF 基准测试中，我们的集成模型在晶系分类和空间群分类上的准确率分别达到 81.7% 和 66.2%，同时在基于实验PXRD数据的晶格参数回归方面树立了新标准。这一性能与以往专门针对分类的最先进方法相当或更优，表明统一预测对称性和晶格参数是可行的，且不会牺牲准确率。我们的物理感知型 GEMD损失还使预测误差集中在对称性层级中接近真实空间群的位置，即使未能实现精确匹配，也能提供具有晶体学意义的预测。

从粉末衍射数据测定晶体结构仍然颇具挑战，现有的从头算方法受数据质量要求和适用范围有限的制约 [33–37]。AlphaDiffract 通过仅依据衍射图谱直接预测晶格参数和候选空间群，跳过显式的峰搜索和指标化步骤，从而解决了这一工作流程中的关键问题。当预测的空间群与真实值不同时，它们通常是仅相差一个对称生成元的直接子群或超群；这类信息既有助于传统结构测定工具，也有助于晶体学数据库检索 [38]。随着深度学习方法不断成熟并融入现有晶体学工作流程，它们正日益为粉末衍射结构解析中长期存在的挑战提供实用解决方案。

5 方法

5.1 数据整理

用于训练的材料结构以 CIF 文件的形式下载 [39]自两个主要晶体学数据库：从 ICSD 下载了 198,778 个结构 [27]，从 Materials Project 下载了 153,169 个结构 [19]。未尝试排除重复结构。作为第一步精修，我们确保所有 CIF 文件都能成功解析为 pymatgen[40] 结构，以便进行后续质量检查并编写尼格利约化结构。解析后，ICSD 中有 186,765 个结构通过验证，Materials Project 中有 153,169 个。
精修的第二阶段使用 pymatgen 的空间群分析器 [40]进行了多项质量检查。我们确认，在位点容差范围为 0.01 Å 至 0.1 Å、角度容差范围为 0.1^o 至 5^o时，常规晶胞的空间群保持一致，其中上限为 Materials Project [40]使用的默认值。此外，在比较预期的晶胞长度和角度等价关系时，我们还验证了尼格利约化晶胞在相对容差为 10^{-3} 的条件下与晶系对称性保持一致。我们还采用了实际限制条件，要求结构的晶胞体积小于 100,000 $Å^3$，且每个晶胞包含的原子不超过 500 个。应用这些筛选标准后，最终数据集包含 ICSD 中的 165,634 个有效结构，以及 Materials Project 中的 146,633 个有效结构。

用于测试的材料结构和粉末X射线衍射图谱下载自 RRUFF数据库 [20], , 该数据库包含 1,362 个粉末X射线衍射图谱（均已进行背景校正，使基线平坦，但未作其他修改）以及 2,901 个 DIF 文件，后者包含通过精修实验粉末X射线衍射图谱获得的结构信息。本研究使用的 RRUFF 图谱是 RRUFF数据库直接提供的背景校正版，数据库维护人员已扣除背景，使基线平坦。除下文所述的波长转换外，作者未进行其他背景扣除或处理。初步筛选剔除了缺少必要信息（晶胞参数、空间群或波长数据）的 DIF 文件，最终得到 2,572 个有效文件。其中，1,880 个 DIF 文件可成功解析为 pymatgen 晶格，但不一定能解析为完整结构，因为部分文件缺少原子坐标或包含非标准原子符号。

进一步精修验证表明，在 10^{-3} 的相对容差范围内（比较预期的晶胞长度和角度等价关系），且晶胞体积小于 100,000 $Å^3$的结构中，有1,867个结构的晶格与晶系对称性一致。其中，837个结构具有与最初下载的1,362个图谱相对应的粉末X射线衍射图谱。对DIF文件和粉末X射线衍射文件进行交叉验证后确认，有745个结构的晶系和晶格参数相匹配。由于许多实验粉末X射线衍射图谱使用Cu K_α 辐射（8.04 keV）测量，而我们的模型假设能量为20 keV，因此我们依据布拉格定律，将图谱转换至所需的 2θ 范围。转换后，有734个结构的图谱在新的 2θ 范围内不含缺失数据。在这734个结构中，有240个具有完整的原子位置数据，从而能够进行GSAS-II [25] 粉末X射线衍射模拟，并获得与DIF文件中报告的峰一致的结果。由于数据不完整，该验证过程在部分情况下需要手动确定原点和非标准设置。具体而言，许多RRUFF条目使用了非标准的空间群符号，因此需要将其转换为GSAS-II所需的完整赫尔曼–莫金符号。对于具有歧义对称中心位置的空间群（原点1和原点2设置），我们通过系统检查特殊位置、原子间距以及对结构的目视检查，确定了正确的原点。

我们利用全部734个结构进行晶系和空间群分类，但仅使用其中240个结构进行晶格参数回归，因为可靠的参数提取需要从完整结构中得到的尼格利约化晶胞，而不能仅依赖晶格参数。需要指出的是，尽管RRUFF数据库被广泛认为是实验粉末X射线衍射数据库，但其中部分条目是根据单晶数据计算得到的粉末衍射图谱。在用于分类的734个图谱中，我们的测试数据集包含59个此类条目（8.0%）；在用于回归的240个图谱中包含55个（22.9%）。鉴于这些条目被纳入RRUFF，我们将其视为与RRUFF一致的真实世界数据。
补充图**1a**、**b**和**d** 分别展示了用于晶系和空间群分类的最终无机晶体结构数据库、Materials Project 和 RRUFF 数据集中的晶系和空间群分布。此外，补充图2和3分别展示了用于晶格参数回归的最终无机晶体结构数据库、Materials Project 和 RRUFF 数据集中各个尼格利约化晶胞的晶格长度和角度分布。

5.2 噪声模拟

为逼真地模拟合成 PXRD 图谱中的噪声，我们对 RRUFF数据库中粉末X射线衍射数据的噪声特性进行了表征。我们识别了每个图谱中的背景区域，并用其估算噪声参数（补充图**4a**）。从这些背景区域中，我们提取了两项关键噪声特性： $\lambda_{\max}$，用于量化泊松噪声；以及 σ_{rel}，用于量化相对强度尺度上的高斯噪声。具体而言， σ_{rel} 设为背景区域的标准差，而 $\lambda_{\max}$ 设为背景区域平均值的倒数。参数 $\lambda_{\max}$ 用作泊松分布的均值，噪声强度从该分布中采样，并叠加到无噪声的模拟图谱上，从而体现X射线探测固有的计数统计。参数 σ_{rel}表示施加于归一化强度（即将每个图谱重新缩放至 [0,1]之后）的高斯噪声标准差，因此衡量的是相对于图谱最大强度的噪声，而非绝对尺度上的噪声。RRUFF数据集中 $\lambda_{\max}$ 和 σ_{rel} 值的分布分别见补充图**4b-c** 和**f-g**。为验证噪声模型，我们将不同噪声水平下的 RRUFF 图谱与叠加了相应泊松噪声和高斯噪声的无噪声模拟图谱进行了比较（补充图**4d-e** 和**h-i**）。实验图谱同时含有泊松噪声和高斯噪声成分（其中高斯噪声似乎占主导），但从视觉比较来看，整体噪声特性和趋势符合预期，这表明我们的两步噪声模型足以生成逼真的合成训练数据。

含噪声的粉末X射线衍射强度通过以下连续步骤获得。首先，加入泊松噪声并重新缩放强度：

$$I_{\text{Pois}}(2\theta) = \frac{\max(I_{\text{clean}})}{\lambda_{\max}} \cdot \text{Poisson}\left(\lambda_{\max} \cdot \frac{I_{\text{clean}}(2\theta)}{\max(I_{\text{clean}})}\right) \tag{1}$$

然后，对归一化强度加入高斯噪声：

$$I_{\text{Gauss}}(2\theta) = \frac{I_{\text{Pois}}(2\theta) - \min(I_{\text{Pois}})}{\max(I_{\text{Pois}}) - \min(I_{\text{Pois}})} + \mathcal{N}(0, \sigma_{\text{rel}}) \tag{2}$$

最后，将该图案重新归一化为 [0,1]，并缩放回原始强度范围：

参数	最小值	最大值
样品与仪器展宽（静态）		
微应变	500	10,000
晶粒尺寸 (μm)	0.1	1
U (cdeg^2)	0	3
V (cdeg^2)	-1	0
W (cdeg^2)	0	4
噪声增强（动态）		
$\lambda_{\max}$	1	100
σ_{rel}	10^{-3}	10^{-1}

表 5：**粉末X射线衍射图谱增强的参数**。对于每种结构，我们使用 GSAS-II 在给定范围内对样品参数和仪器参数进行均匀采样，生成 100 个增强的粉末X射线衍射图谱（静态增强）。训练期间，还会在数据加载器中对噪声参数 ($\lambda_{\max}$, σ_{rel}) 进行均匀采样（动态增强）。单位“cdeg”指百分之一度（度/100）。

$$I_{\text{noisy}}(2\theta) = \left[\frac{I_{\text{Gauss}}(2\theta) - \min(I_{\text{Gauss}})}{\max(I_{\text{Gauss}}) - \min(I_{\text{Gauss}})}\right] \cdot (\max(I_{\text{Pois}}) - \min(I_{\text{Pois}})) + \min(I_{\text{Pois}}) \tag{3}$$

其中， $I_{\text{clean}}(2\theta)$ 是干净的模拟强度， $I_{\text{Pois}}(2\theta)$ 是应用泊松噪声后的强度， $I_{\text{Gauss}}(2\theta)$ 是添加高斯噪声后的归一化强度，$\text{Poisson}(\lambda)$ 表示从均值为 λ 的泊松分布中采样，而 $\mathcal{N}(0, \sigma_{\text{rel}})$ 表示从均值为零、标准差为 σ_{rel} 的高斯分布中采样。

训练过程中，通过数据加载器动态应用噪声增强：对每个训练样本，随机均匀采样 $\lambda_{\max}$，取值范围为 1 到 100；并随机均匀采样 σ_{rel}，取值范围为 10^{-3} 到 10^{-1}。尽管与补充图4所示的经验分布相比，这些范围偏向噪声更大的样本，但我们发现，这种偏差在实际中提升了模型在真实实验数据上的表现，这可能是因为它增强了模型对噪声的鲁棒性。

5.3 模型架构

AlphaDiffract 模型由一个特征提取器组成，该提取器接收粉末X射线衍射图谱并将其映射到特征空间。提取出的特征会传递给各个预测头，每个预测头专门用于预测一种目标属性。

5.3.1 特征提取器

特征提取器由 3 个经过一维适配的 ConvNeXt [23] 模块组成。每个 ConvNeXt 模块之后都会通过一个平均池化层将图像下采样 2 倍。图 1 展示了 ConvNeXt 块的架构，其中 [23] 中的二维卷积层被替换为对应的一维卷积层，但保留了 ConvNeXt 的关键模块设计特征，包括倒置瓶颈以及位于上游的逐通道卷积层。与 ConvNeXt 类似，该模块采用残差连接，将输入与最后一个逐点卷积层的输出相加。干路（即非残差）分支采用随机深度 [41]，其比率为 0.3。每个模块的输入/输出通道、卷积核大小和池化步幅等其他参数列于表 6 中。输入维度为 8192 时，特征提取器输出的特征维度为 560。

5.3.2 预测头

每个量（晶系、空间系、等等）都由各自的预测头进行预测。预测头是一个多层感知机，可将特征提取器的特征映射为具有特定量维度的向量。预测头的输入、输出和隐藏维度如表 7 所示。对于分类量，输出向量会经过 Softmax层，以得到每个类别的预测概率。对于晶格参数，输出是 6 维向量，分别对应晶胞的长度和角度 a、 b、 c、 α、 β 和 γ；输出会经过 Sigmoid函数处理，然后缩放至设定的下限和上限之间：

$$y_i = \sigma(p_i)(UB_i - LB_i) + LB_i \tag{4}$$

其中， i 索引 6 个晶格参数的各个元素， p_i 是预测头输出中对应参数 i的元素。UB_i 和LB_i 表示上限和下限；对于长度，上限和下限分别为 500 Å 和 0；对于角度，则分别为 0° 和 180°。

5.4 损失函数

我们的总体损失函数结合了三个针对不同预测任务设计的组成部分。对于晶系（CS）和空间群（SG）分类，这两项任务均采用交叉熵损失。空间群损失还可选择性地加入由超参数 μ 加权的图地球搬运工距离（GEMD）损失。对于同时进行晶格参数回归的组合模型，我们还会加入用于预测晶格参数（LP）的均方误差（MSE）损失。批次损失通过对 N 个样本取平均值计算得出：

模块	输入通道	输出通道	卷积核大小 (k)	池化步幅 (s)
1	1	80	100	5
2	80	80	50	5
3	80	80	25	5

表 6： **特征提取器的模块级架构**。

$$\mathcal{L} = \mathcal{L}_{CS} + \mathcal{L}_{SG} + \mathcal{L}_{LP} = \frac{1}{N}\sum_{i=1}^{N} \ell^{(i)} \tag{5}$$

其中，单个样本的损失定义如下：

$$\ell^{(i)} = -\sum_{j=1}^{7} y_{CS,j}^{(i)} \log(\hat{y}_{CS,j}^{(i)}) - \sum_{j=1}^{230} y_{SG,j}^{(i)} \log(\hat{y}_{SG,j}^{(i)}) + \mu \sum_{k=1}^{230}\sum_{j=1}^{230} y_{SG,j}^{(i)} D_{jk} \hat{y}_{SG,k}^{(i)} + \frac{1}{6}\sum_{j=1}^{6} (y_{LP,j}^{(i)} - \hat{y}_{LP,j}^{(i)})^2 \tag{6}$$

其中， $y^{(i)}$ 表示真实标签或数值，而$\hat{}y^{(i)}$ 表示该 i^{th} 数据样本的模型预测。第三项表示 GEMD损失，它利用空间群的层次结构，根据距离矩阵 D所表示的极大子群图中的距离，对错误分类进行惩罚；其中D_{jk} 编码了空间群 j 和 k 之间经过极大子群图的跳数。

5.5 模型训练

模型使用 AdamW优化器 [42] 进行训练，学习率为 2×10^{-4}，权重衰减因子为 0.01。我们采用了无周期学习率调度方案 [43]，在六个训练轮次（半个周期）内将学习率从标称值的 10% 线性提高至完整值，随后再降回原值；之后每个周期的振幅均减半。每个模型均使用单个 NVIDIA H100 GPU 进行训练，批量大小为 64。将 ICSD 和 Materials Project数据库中各自总训练数据的 10% 用作验证集。

5.6 不确定性估计

我们量化了神经网络预测中的两种独立不确定性来源。首先，集成不确定性源于独立训练的模型实例之间的差异。其次，由于对无机晶体结构数据库和 Materials Project 数据集中的每个唯一结构进行了 100 倍数据增强，涵盖样品和仪器展宽参数及噪声的变化，我们将增强不确定性定义为同一结构生成的粉末X射线衍射图谱之间预测结果的差异。我们按如下方式将这些不确定性转化为正文所报告的分类准确率和回归指标的误差条。

5.6.1 晶系和空间群分类

对于每项预测，我们首先计算各集成成员的类别概率标准差， $\sigma_{model}^{(i)}$。然后，使用 Delta 方法将这一逐样本不确定性传播至总体准确率指标，并计算标准误：

$$\sigma_{model} = \frac{1}{N}\sqrt{\sum_{i=1}^{N} \left(\sigma_{model}^{(i)}\right)^2}, \tag{7}$$

其中， N 为样本总数。95%置信区间为准确率 $\pm z_{0.975} \cdot \sigma_{model}$，其中 $z_{0.975} = 1.96$。接下来，通过首先计算正确分类的数据增强结果所占比例，来计算不同数据增强所产生的不确定性：

$$a^{(j)} = \frac{1}{100}\sum_{k=1}^{100} (y_k^{(j)} = \hat{y}_k^{(j)}). \tag{8}$$

那么，这些结构的平均准确率的标准误为：

$$\sigma_{aug} = \frac{s_a}{\sqrt{n}}, \tag{9}$$

预测量	输入维度	隐藏维度	输出维度
晶系	560	2300 → 1150	7
空间群	560	2300 → 1150	230
晶格参数	560	512 → 256	6

表 7：**多层感知机预测头的维度**。

其中， s_a 是 $\{a^{(j)}\}$ 的样本标准差， $n = N/100$是唯一结构的数量。95%置信区间采用t分布，表示为准确率 $\pm t_{0.975,\ n-1} \cdot \sigma_{aug}$。

5.6.2 晶格参数回归

我们使用标准误差传播，将集成不确定性和增强不确定性传播至回归指标。设N 为测试样本总数， n 为唯一结构的数量，每个结构 j 均进行 100 倍增强。对于属于结构 j的每个增强样本 i ，模型不确定性 $\sigma^{(i)}_{model}$ 表示集成成员预测值的标准差。增强不确定性 $\sigma^{(j)}_{aug}$ 反映了结构 j的 100 种增强模式下集成平均预测值的变化。为了估计用于评估晶格参数预测的各项回归指标的误差，我们首先通过平方和开根号合并这两个不确定性来源，

$$\sigma^{(j)}_{total} = \sqrt{\frac{1}{100}\sum_{k=1}^{100}\left(\sigma^{(j)}_{model,k}\right)^2 + \left(\sigma^{(j)}_{aug}\right)^2}. \tag{10}$$

然后使用 Delta 方法将总不确定性传播至各项指标。对于平均绝对误差（平均绝对误差），其不确定性为

$$\sigma_{\mathrm{MAE}} = \frac{1}{n}\sqrt{\sum_{j=1}^{n}\left(\sigma^{(j)}_{total}\right)^2}. \tag{11}$$

对于平均绝对百分比误差（平均绝对百分比误差），其不确定性按真实值的倒数缩放:

$$\sigma_{\mathrm{MAPE}} = \frac{100}{n}\sqrt{\sum_{j=1}^{n}\left(\frac{\sigma^{(j)}_{total}}{|y^{(j)}|}\right)^2}. \tag{12}$$

对于决定系数（R^2），我们使用导数 $\partial R^2/\partial \hat{y}^{(j)} = -2(\hat{y}^{(j)} - y^{(j)})/SS_{tot}$，其中 $SS_{tot} = \sum_{i=1}^{n}(y^{(j)} - \bar{y})^2$，得到

$$\sigma_{R^2} = \sqrt{\sum_{j=1}^{n}\left(\frac{2(\hat{y}^{(j)} - y^{(j)})}{SS_{tot}}\sigma^{(j)}_{total}\right)^2} \tag{13}$$

这里， $y^{(j)}$ 和$\hat{y}^{(j)}$ 分别表示结构的真实值和预测值 j （对其数据增强取平均）。

数据可用性

晶体结构可从 Materials Project [19] 和 RRUFF [20] 数据库（开放访问）以及 ICSD [27]（需要许可证）获取。根据公开数据集得出的模型权重和预测结果将在发表后提供。完整的模型权重和预测结果可在合理要求下向作者索取。

代码可用性

本研究使用的代码可在 https://github.com/AdvancedPhotonSource/OpenAlphaDiffract 获取。

致谢

作者感谢 Laurent Chapon博士就如何将机器学习方法应用于粉末X射线衍射数据的晶体学分析进行了深入讨论。本研究使用了先进光子源的资源。先进光子源是美国能源部（DOE）科学办公室用户设施，由阿贡国家实验室代表能源部科学办公室运营，依据合同编号 DE-AC02-06CH11357；本研究工作得到美国能源部科学办公室基础能源科学部的支持，合同编号 DE-AC02-06CH11357。

作者贡献

M.J.C. 构思了本研究，监督研究工作，并参与了数据分析和稿件撰写。N.A. 开发了对称性约束训练方法，牵头开展数据处理，并参与模型训练和

评估、分析和稿件撰写。M.D. 牵头开发并实现模型架构，并参与数据处理、模型训练与评估、分析和稿件撰写。H.S. 开发了 PXRD 模拟流程，并生成了用于训练的衍射图谱。J.P.H. 进行了晶体结构数据整理，并参与 PXRD 图谱生成。A.L. 参与了数据处理、分析和稿件撰写。X.Y. 针对既有方法开展了基准测试，并参与分析。M.P. 开发了面向用户的模型部署工具，将代码库容器化，并进行了验证测试。B.H.T. 提供了晶体学专业知识，进行了数据整理，参与 PXRD 图谱生成，并协助撰写稿件。所有作者均审阅并批准了最终稿件。N.A. 和 M.D. 对本研究贡献相同。

竞争性利益

作者声明不存在竞争性利益。

补充信息

有关通信和材料索取事宜，请联系 Nina Andrejevic (nandrejevic@anl.gov)、Ming Du (mingdu@anl.gov) 或 Mathew J. Cherukara(mcherukara@anl.gov)。

参考文献

[1] H. M. Rietveld.《核结构和磁结构的轮廓精修方法》。载于：应用晶体学杂志2 (1969年)，第 65–71 页。网址：http://www.ccp14.ac.uk/ccp/web-mirrors/hugorietveld/ xtal/paper2/paper2.html。[2] A. Coelho.《通过迭代使用奇异值分解对粉末衍射图谱进行指标化》。载于：应用晶体学杂志36.1 (2003年)，第 86–95 页。ISSN：0021-8898。DOI：doi: 10.1107/ S0021889802019878。网址：http://dx.doi.org/10.1107/S0021889802019878。[3] A. Boultif 和 D. Louer.《用逐次二分法对低对称性晶格的粉末衍射图谱进行指标化》。载于：应用晶体学杂志24 (1991年)，第 987–93 页。ISSN：0021-8898。网址：%3CGo%20to%20ISI%3E://INSPEC:4063785。[4] P. E. Werner、L. Eriksson 和 M. Westdahl.《TREOR：一种适用于所有对称性的半穷举式试错粉末指标化程序》。载于：应用晶体学杂志18 (1985年)，第 367–70 页。ISSN：0021-8898。网址：%3CGo%20to%20ISI%3E://INSPEC:2583001。[5] D. E. Cox、B. H. Toby 和 M. M. Eddy.《利用同步辐射采集粉末衍射数据》。载于：澳大利亚物理学杂志41.2 (1988年)，第 117 页。[6] AJ Markvardsen 等人。《一种基于概率的方法，用于根据粉末衍射数据判定空间群》。载于：晶体学基础57.1 (2001年)，第 47–54 页。[7] W B Park 等人。《使用卷积神经网络对晶体结构进行分类》。载于：*IUCrJ* 4.4 (2017年)，第 486–494 页。[8] Pascal Marc Vecsei 等人。《基于神经网络对X射线衍射图谱中的晶体对称性进行分类》。en。载于：*Phys. Rev. B.* 99.24 (2019年6月)。[9] Alexander N. Zaloga 等人。《使用卷积神经网络对粉末X射线衍射图谱中的晶体对称性进行分类》。载于：*MaterialsToday Communications*25 (2020年)，第 101662 页。DOI：10.1016/j. mtcomm.2020.101662。[10] Sathya R Chitturi 等人。《根据X射线粉末衍射图谱自动预测晶格参数》。载于：*J.Appl. Crystallogr.* 54.6 (2021年)，第 1799–1810 页。[11] Byung Do Lee 等人。《粉末X射线衍射图谱就是机器学习对称性识别与性质预测所需的一切》。载于：*Advanced Intelligent Systems* 4.7 (2022年)。[12] Byung Do Lee 等人。《粉末X射线衍射图谱分析的深度学习方法：同时解决泛化能力和扰动问题》。载于：*Advanced Intelligent Systems* 5.9 (2023年)。[13] Nicola Corriero 等人。《CrystalMELA：一个用于晶系判定的新型晶体学机器学习平台》。载于：*J. Appl. Crystallogr.*56.2 (2023年)，第 409–419 页。[14] Juan Iván Gómez-Peralta、Xim Bokhimi 和 Patricia Quintana.《利用卷积神经网络辅助评估X射线粉末衍射中的晶格参数》。载于：《物理化学杂志*A*》127.36 (2023年)，第 7655–7664 页。

[15] Jerardo E Salgado 等人。《使用深度学习模型对大规模X射线衍射数据进行自动分类》。en。载于：Npj Comput. Mater. 9.1 (2023年12月)，第 214 页。[16] Eric A Riesel 等人。《利用生成式机器学习从粉末衍射图谱测定晶体结构》。en。载于：J. Am. Chem. Soc. (2024年9月)。[17] Haotong Liang 等人。《CRYSPNet：通过神经网络预测晶体结构》。载于：Physical Review Materials 4.12 (2020年)，第 123802 页。[18] Dejan Zagorac 等人。《无机晶体结构数据库的最新进展：理论晶体结构数据及相关功能》。载于：Applied Crystallography 52.5 (2019年)，第 918–925 页。[19] Anubhav Jain 等人。《评论：材料项目：以材料基因组方法加速材料创新》。载于：APL materials 1.1 (2013年)。[20] Barbara Lafuente 等人。《数据库的力量：RRUFF项目》。载于：Highlights in mineralogical crystallog- raphy 1 (2015年)，第 25 页。[21] Kamal Choudhary.《DiffractGPT：利用生成式预训练Transformer根据X射线衍射图谱测定原子结构》。en。载于：J. Phys. Chem. Lett. (2025年2月)，第 2110–2119 页。[22] Frederik Lizak Johansen 等人。《deCIFer：使用自回归语言模型从粉末衍射数据预测晶体结构》。载于：arXiv预印本 arXiv:2502.02189 (2025年)。[23] Zhuang Liu 等人。《面向2020年代的卷积网络》。载于：arXiv (2022年)。[24] Kaiming He 等人。《用于图像识别的深度残差学习》。载于：arXiv (2015年)。[25] B. H. Toby 和 R. B. Dreele.《GSAS-II：现代开源通用晶体学软件包的诞生》。载于：应用晶体学杂志 46.2 (2013年)，第 544–549 页。[26] Jackson H. O'Donnell 等人。《GSAS-II 的脚本接口》。载于：应用晶体学杂志 51.4 (2018年)，第 1244–1250 页。DOI：10.1107/S1600576718008075。[27] 美国国家标准与技术研究院。NIST无机晶体结构数据库，NIST标准参考数据库第3号。美国国家标准与技术研究院，马里兰州盖瑟斯堡，20899。DOI：10.18434/M32147。网址：https://doi.org/10.18434/M32147。[28] James A. Kaduk 和 Joel Reid.《Rietveld仪器轮廓系数的典型值》。载于：Powder Diffraction 26.01 (2011年)，第 88–93 页。ISSN：1945-7413。DOI：doi:10.1154/1.3548128。网址：http://dx.doi.org/10.1154/1.3548128。[29] Pascal Marc Vecsei 等人。《基于神经网络对X射线衍射图谱中的晶体对称性进行分类》。载于：Physical Review B 99.24 (2019年)，第 245120 页。[30] Gemma de la Flor、Hans Wondratschek 和 Mois I Aroyo.《毕尔巴鄂晶体学服务器上亚周期群极大子群的完整在线数据库》。载于：Applied Crystallography 58.2 (2025年)。[31] Mois Ilia Aroyo 等人。《毕尔巴鄂晶体学服务器：I. 数据库与晶体学计算程序》。载于：Zeitschrift für Kristallographie-Crystalline Materials 221.1 (2006年)，第 15–27 页。[32] Ramprasaath R Selvaraju 等人。《Grad-cam：通过基于梯度的定位对深度网络进行可视化解释》。载于：IEEE国际计算机视觉会议论文集。2017年，第 618–626 页。[33] A. Altomare 等人。《EXPO：用于完整粉末图谱分解和晶体结构解析的程序》。载于：应用晶体学杂志 32 (1999年)，第 339–340 页。ISSN：0021-8898。DOI：10 .1107 / s0021889898007729。网址：%3CGo%20to%20ISI%3E://WOS:000079981900026。[34] Vincent Favre-Nicolin 和 Radovan Cerny.《FOX，"用于晶体学的自由对象"：一种通过粉末衍射从头测定结构的模块化方法》。载于：应用晶体学杂志 35.6 (2002年)，第 734–743 页。ISSN：0021-8898。DOI：doi:10.1107/S0021889802015236。网址：http://dx.doi.org/10.1107/S0021889802015236。[35] William I. F. David 等人。《DASH：一种根据粉末衍射数据测定晶体结构的程序》。载于：应用晶体学杂志 39.6 (2006年)，第 910–915 页。ISSN：0021-8898。DOI：doi:10.1107/ S0021889806042117。网址：http://dx.doi.org/10.1107/S0021889806042117。[36] Stefan Habermehl、Carina Schlesinger 和 Martin U. Schmidt.《利用基于互相关函数的图谱比较，通过全局优化方法从头测定未索引粉末数据的结构》。载于：《晶体学报B辑》78.2 (2022年)，第 195–213 页。ISSN：2052-5206。DOI：doi : 10.1107/S2052520622001500。网址：https://doi.org/10.1107/S2052520622001500。[37] Peter W Stephens 和 Ashfia Huq.《PSSP：用于直接空间模拟退火的开源粉末结构解析程序》。载于：Transactions of the American Crystallographic Assocation 37 (2002年)，第 125–142 页。[38] James A. Kaduk.《将无机晶体结构数据库用作问题求解工具》。载于：《晶体学报B辑》58.3 Part 1 (2002年6月)，第 370–379 页。DOI：10.1107/S0108768102003476。网址：https://doi.org/10.1107/S0108768102003476。

[39] I. David Brown 和 Brian McMahon。《CIF：晶体学的计算机语言》。载于：*Acta Crystallograph-ica Section B* 58.3 第1部分（2002年6月），第317–324页。[40] Shyue Ping Ong 等人。《Python Materials Genomics (pymatgen)：用于材料分析的稳健开源 Python 库》。载于：*Computational MaterialsScience* 68（2013年），第314–319页。[41] GustavLarsson、MichaelMaire 和 Gregory Shakhnarovich。《FractalNet：无需残差的超深神经网络》。载于：*arXiv [cs.CV]*（2016年5月）。[42] Ilya Loshchilov 和 Frank Hutter。《解耦权重衰减正则化》。载于：*arXiv [cs.LG]*（2017年11月）。[43] Leslie N Smith。《用于训练神经网络的循环学习率》。载于：*2017 IEEE*冬季计算机视觉应用会议（*WACV*）。IEEE。2017年，第464–472页。

《AlphaDiffract：粉末X射线衍射数据的自动化晶体学分析》补充材料

Nina Andrejevic[1*†]、Ming Du[1*†]、Hemant Sharma[1],James P. Horwath[1],Aileen Luo[1],Xiangyu Yin[1],Michael Prince[1],Brian H. Toby[1],和 Mathew J. Cherukara[1*]

[1]美国阿贡国家实验室先进光子源，莱蒙特，伊利诺伊州 *60439* †共同第一作者，贡献相同。

*电子邮箱：nandrejevic@anl.gov，mingdu@anl.gov，mcherukara@anl.gov

1 深度学习方法在粉末X射线衍射分析中的文献综述

1.1 架构设计

一维卷积神经网络（卷积神经网络）一直是粉末X射线衍射分析中最主要的架构选择 [S1– S10]。卷积神经网络非常适用于粉末X射线衍射数据，因为它们既能通过归纳偏置捕捉局部特征（单个衍射峰），又能通过分层下采样捕捉长程关系（衍射峰模式）——这两者对结构识别都至关重要。典型实现会将卷积神经网络用作特征提取器，把粉末X射线衍射图谱映射为潜在表示，随后接入一个或多个多层感知机（MLP）预测头，将特征映射为分类或回归输出。这些特征提取器通常采用多尺度架构，并在模块之间使用池化或步幅卷积，以增大感受野。然而，Salgado等 [S9] 发现，在保留步幅卷积的同时移除池化层，能提升某些数据集上的表现；他们认为这是由于信息压缩减少所致，但代价是感受野缩小，且下游预测头的参数量增加。在粉末X射线衍射分析中，卷积神经网络的一个关键考量是平移等变性：输入发生空间平移时，输出也会相应平移。结合池化操作 [S11], 后，这可能导致平移不变性——这对图像分类有利，但对粉末X射线衍射而言却有问题，因为图谱在 2θ 空间中的平移代表不同的结构。若在卷积神经网络特征提取器之后接入多层感知机预测头，则可缓解这一问题，因为多层感知机本质上具有平移敏感性。一些研究 [S2, S5, S12] 探索了纯多层感知机架构，以利用这种敏感性，但结果不一：Vecsei 等 [S2]发现，与卷积神经网络相比，多层感知机在实验数据上的表现更好，但在合成数据上的表现更差。实际应用中，多层感知机面临计算方面的挑战，因为其参数量大，且矩阵运算开销高。变换器在自然语言处理 [S13]和计算机视觉 [S14], 领域取得了巨大成功，但在粉末X射线衍射中的应用仍然有限。Lee等 [S5] 发现，即使经过粉末X射线衍射专用预训练，基于Transformer的模型表现仍不如卷积神经网络；他们认为这是因为变换器对数据需求量大：缺乏类似卷积神经网络的归纳偏置时，它们需要数量多得多的训练样本。然而，近期关于 DiffractGPT 的研究 [S15] 表明，大型预训练语言模型（例如，Mistral 7B [S16]）可以有效微调，用于基于粉末X射线衍射的晶格参数和原子位置预测，这为利用在海量语言数据上预训练的模型指明了一个很有前景的方向。

1.2 训练数据来源与规模

ICSD [S17] 和 Materials Project [S18] 是训练所用的主要晶体学数据库。剔除有问题的结构后，这两个数据库都提供了大规模训练集，并已得到广泛采用[S1–S6, S9, S10, S12]。ICSD 和 Materials Project 提供模拟数据，而 RRUFF 数据库 [S19] 主要提供实验粉末X射线衍射图谱，可作为评估模型在真实条件下性能的基准[S2, S6, S9, S10]。据报道，规模最大的训练集包含来自 ICSD 和 Materials Project 的 263,000 个结构 [S5]。为模拟晶格应变和峰展宽等实验缺陷，通常会针对每种结构生成多个经过数据增强的图谱，并施加不同的扰动。例如，Lee 等人 [S6] 为每种结构生成了20个数据增强样本，从而得到370万个训练样本。

1.3 预测目标与模型范围

大多数研究侧重于晶体对称性分类，预测晶系和空间群。一些研究将范围扩展到晶格参数 [S4、S8、S10、S12、S15] 等连续量，以及带隙和形成能等材料性质 [S5]。然而，许多方法需要针对不同任务或结构类型分别构建专用模型。例如，Chitturi 等人 [S4] 训练了七个独立模型——每个晶系对应一个模型——用于晶格参数

预测，而 Lee等人 [S5] 则使用不同的模型进行分类和性质回归。虽然专用模型简化了训练，但能够同时预测多个量的统一模型可简化推理流程。

2 无机晶体结构数据库与 Materials Project 数据库之间的结构相似性

为评估结构相似性和数据库重叠情况，我们采用 CrystalNN位点指纹方法 [S20]（该方法通过结构序参数表征结构，并在 Matminer 中实现 [S21]），对无机晶体结构数据库和 Materials Project数据库中的每个结构进行表征。我们使用 scipy [S22], 实现的层次聚类，对组成和空间群完全相同的结构的指纹进行聚类，采用完全连接法和欧氏距离度量。按照 Materials Project文档 [S18]提出的惯例，如果结构在距离阈值 0.9 内属于同一簇，则将其归类为“等价”。通过这项分析，我们将合并后的数据集划分为仅出现在无机晶体结构数据库中的结构、仅出现在 Materials Project 中的结构，以及同时出现在两个数据库中的结构；随后又将各数据集细分为唯一结构及其结构等价物（补充图**1c**）。这种聚类方法表明，最终整理后的数据集包含 65,748 个无机晶体结构数据库唯一结构、93,418 个 Materials Project 唯一结构，以及两个数据库共有的 39,239 个唯一结构；此外，无机晶体结构数据库中的 18,683 个结构、Materials Project 中的 12,616 个结构，以及两个数据库共有的 82,545 个结构，均与各自数据集中的其他结构等价。

Reference	Architecture	Training database	Test database	Number of training inorganic structures	Number of training organic structures	Total number of training structures	Total number of training diffraction patterns	Predicted quantities	Remarks
Park *et al.* [S1]	1D CNN	ICSD	ICSD	120,000	0	120,000	120,000	CS, SG, EG	
Vecsei *et al.* [S2]	1D CNN, MLP	ICSD	ICSD, RRUFF	114,404	0	114,404	114,404	CS, SG	
Zaloga *et al.* [S3]	1D CNN	ICSD	ICSD	153,603	0	153,603	–	CS, SG	
Liang *et al.* [S12]	MLP	ICSD	ICSD	110,813	0	100,000	100,000	BL, SG, LP	After predicting BL, trained separate models to predict SG and LP for each BL. Train/test are split by year of publication.
Chitturi *et al.* [S4]	1D CNN	ICSD, CSD	ICSD, CSD	∼137,000	∼825,000	960,000	960,000	LP	Trained a separate model for each CS. Inorganic/organic numbers are calculated from Fig. 1 of cited paper.
Lee *et al.* [S5]	1D CNN, transformer (structures), MLP (properties)	ICSD, MP	ICSD, MP	263,000	0	263,000	263,000	CS, EG, SG, E_g, E_f, E_h	
Lee *et al.* [S6]	1D CNN	ICSD	RRUFF	187,131	0	187,131	3,742,620	CS, SG, EG	Listed numbers are based on 20 texture-involving perturbations generated for each structure. A separate dataset of 10 texture-free patterns per structure was also generated.
Corriero *et al.* [S7]	1D CNN	POW_COD	Private database	21,783	261,223	283,006	283,006	CS	Listed numbers are total data size; train/test ratio unavailable.
Gómez-Peralta *et al.* [S8]	1D CNN	COD	COD	0	83,000	83,000	332,000	LP	
Salgado *et al.* [S9]	1D CNN	ICSD	RRUFF	171,006	0	171,006	1,200,000	CS, SG	
Riesel *et al.* [S10]	1D CNN	MP-20	MP-20, AMCSD, RRUFF	36,185	0	36,185	144,740	CP, LP, NA	Train on MP-20 subset of MP dataset containing only structures with 1-20 atoms in the primitive unit cell.
Choudhary [S15]	Mistral 7B	JARVIS-DFT	JARVIS-DFT	72,990	0	72,990	72,990	LP, AP	
Ours	**1D ConvNeXt**	**ICSD, MP**	**RRUFF**	**312,267**	**0**	**312,267**	**31,226,700**	**CS, SG, LP**	

Supplementary Table 1: **Survey of deep learning methods for PXRD analysis.** A survey on the data sources, sizes of training data, and predicted quantities for models that predict crystal structures and material properties from PXRD diffraction patterns. We only include works where the models were trained on a comprehensive collection of structures rather than ones only trained for certain materials systems. "Number of structures" refers to the distinct structures in the training data, while "number of diffraction patterns" means the actual number of samples in the dataset that include all augmented variants. Since multiple augmented variants can be generated from each structure, numbers in this columns are greater than or equal to the numbers of structures. "–" means the figures are not explicitly mentioned in the papers and are hard to deduce with certainty. Abbreviations and symbols: CS – crystal system, SG – space group, EG – extinction group, LP – lattice parameters, BL – Bravais lattice, AP – atomic positions, E_g – bandgap, E_f – formation energy, E_h – energy above the convex hull, CP – chemical composition, NA – number of atoms.

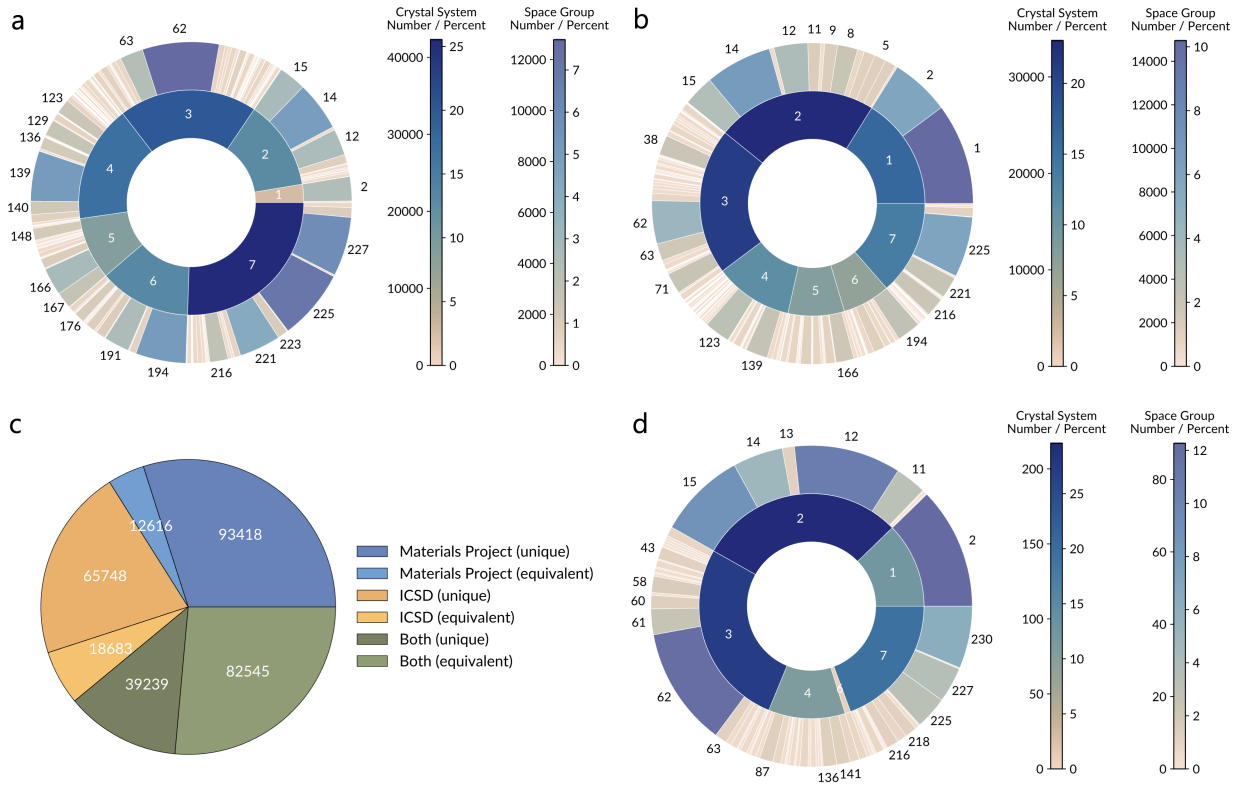
	模型变体	测试数据	指标	指标值					
				a	b	c	α	β	γ
Choudhary [S15]	DGPT公式	JARVIS-DFT	MAE	**0.17**	**0.18**	**0.27**	–	–	–
Choudhary [S15]	DGPT公式	无机晶体结构数据库†	MAE	1.72 (1.66)	2.40 (2.28)	4.44 (4.44)	5.00 (5.16)	2.76 (3.34)	9.51 (10.20)
Choudhary [S15]	DGPT公式	Materials Project†	MAE	1.55 (1.48)	2.17 (2.13)	4.36 (4.47)	4.96 (5.45)	2.85 (3.53)	10.11 (9.89)
Choudhary [S15]	DGPT公式	RRUFF‡	MAE	1.72 (1.66)	2.40 (2.28)	4.44 (4.44)	5.00 (5.16)	2.76 (3.34)	9.51 (10.20)
Ours	分类 +回归 集成	无机晶体结构数据库†	MAE	1.14 ± 0.00	1.33 ± 0.00	2.29 ± 0.01	**2.02 ± 0.01**	**0.61 ± 0.00**	**2.41 ± 0.01**
Ours	分类 +回归 集成	Materials Project†	MAE	1.22 ± 0.00	1.60 ± 0.00	2.93 ± 0.01	3.66 ± 0.01	1.63 ± 0.00	4.72 ± 0.02
Ours	分类 +回归 集成	RRUFF‡	MAE	1.37 ± 0.02	1.76 ± 0.02	3.22 ± 0.05	2.95 ± 0.05	2.31 ± 0.03	2.90 ± 0.07
Chitturi等 [S4]	全范围	无机晶体结构数据库, CSD	MAPE	**9.20**			–	–	–
Choudhary [S15]	DGPT公式	无机晶体结构数据库†	MAPE	27.11 (26.97)	34.63 (33.32)	41.23 (40.69)	4.88 (5.06)	3.05（3.66）	9.54（10.27）
Choudhary [S15]	DGPT公式	Materials Project†	MAPE	23.17（22.88）	29.59（28.17）	31.67（30.06）	5.10 (5.62)	3.15 (3.91)	10.51 (10.31)
Choudhary [S15]	DGPT公式	RRUFF‡	MAPE	22.72 (23.18)	27.92 (25.40)	27.20 (27.38)	4.38 (4.45)	3.81 (4.04)	9.00 (8.68)
Ours	分类 +回归 集成	无机晶体结构数据库†	平均绝对百分比误差	19.52 ± 0.05	18.54 ± 0.06	21.23 ± 0.08	**2.03 ± 0.01**	**0.66 ± 0.00**	**2.39 ± 0.01**
Ours	分类 +回归 集成	Materials Project†	MAPE	21.58 ± 0.05	21.92 ± 0.06	24.27 ± 0.08	3.88 ± 0.01	1.83 ± 0.00	5.12 ± 0.02
Ours	分类 +回归 集成	RRUFF‡	平均绝对百分比误差	22.21 ± 0.31	19.87 ± 0.28	28.42 ± 0.64	3.14 ± 0.06	2.51 ± 0.03	3.06 ± 0.08
Liang等 [S12]	–	ICSD	R^2	0.56*	0.34*	0.46*	0.43*	0.14*	0.01*
Choudhary [S15]	DGPT公式	无机晶体结构数据库†	R^2	-0.04 (0.03)	0.05 (0.12)	-0.08 (-0.31)	-0.55 (-0.66)	-4.50 (-6.41)	-1.15 (-1.32)
Choudhary [S15]	DGPT公式	Materials Project†	R^2	0.11 (0.12)	-0.02 (0.03)	-0.13 (-0.18)	-0.82 (-1.00)	-2.77 (-3.93)	-1.02 (-1.03)
Choudhary [S15]	DGPT公式	RRUFF‡	R^2	0.04 (-0.02)	-0.35 (-0.19)	-0.04 (-0.14)	-0.89 (-0.77)	-1.28 (-1.11)	-4.85 (-5.06)
Ours	分类 +回归 集成	无机晶体结构数据库†	R^2	**0.58 ± 0.00**	**0.63 ± 0.00**	**0.71 ± 0.01**	**0.70 ± 0.00**	**0.26 ± 0.00**	**0.78 ± 0.00**
Ours	分类 +回归 集成	Materials Project†	R^2	0.49 ± 0.00	0.48 ± 0.00	0.49 ± 0.00	0.38 ± 0.00	0.11 ± 0.00	0.45 ± 0.00
Ours	分类 +回归 集成	RRUFF‡	R^2	0.54 ± 0.01	0.39 ± 0.01	0.25 ± 0.02	0.33 ± 0.02	0.26 ± 0.02	0.18 ± 0.03

补充表 2：**AlphaDiffract 和参考模型的晶格参数预测误差**。误差以平均绝对误差（MAE）、平均绝对百分比误差（MAPE）和决定系数（R^2）量化。为便于直接比较，我们仅分析使用回归指标而非匹配率等分类指标量化预测准确性的研究。由于在 RRUFF 上进行测试的研究很少，我们也列出在其他数据集上测试的研究，仅供参考。误差条表示集成和数据增强预测结果汇总后的标准差。

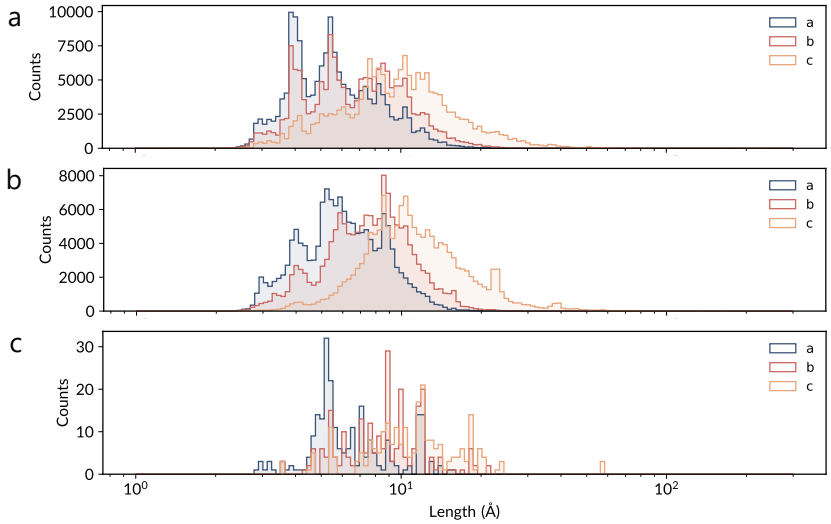
*针对每种布拉维晶格的专用模型的加权平均值。

†在我们的验证集数据上进行评估。使用 DGPT-formula 模型进行推理时，从每个数据集的验证集中选取 1000 个具有代表性的样本进行评估。括号中的分数表示在未添加泊松噪声或高斯噪声的合成 PXRD 图谱上的结果。

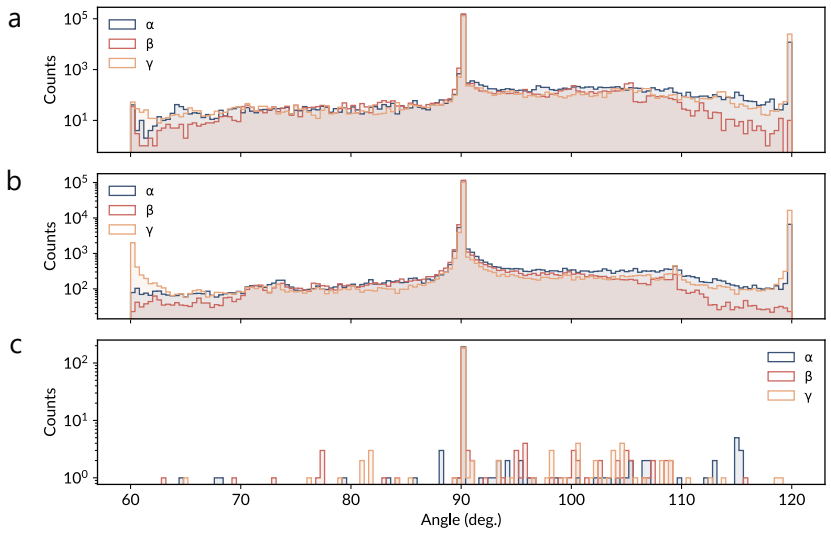
‡在我们的测试集数据上进行了评估。



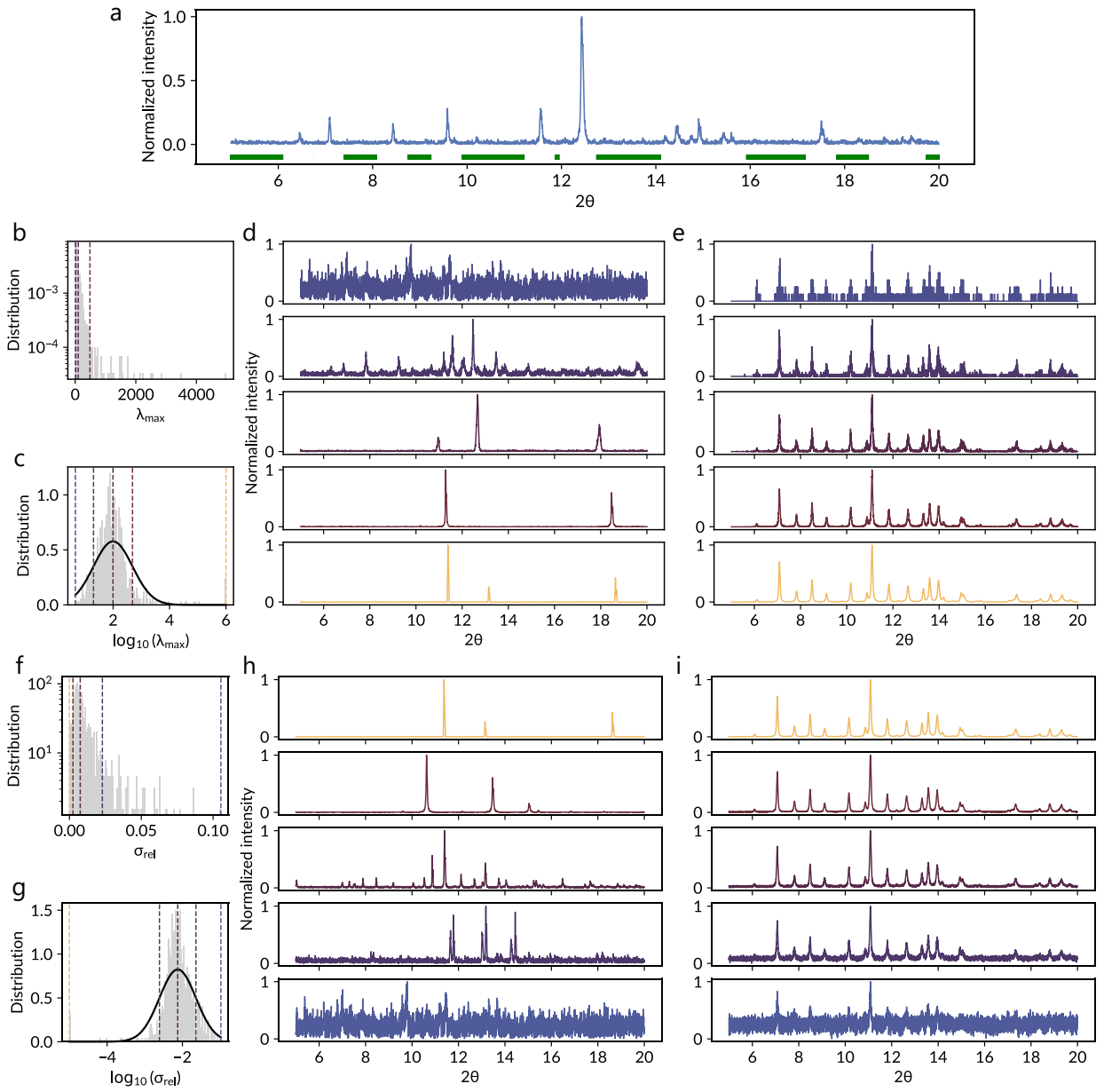
补充图1：**晶体学数据库中晶系和空间群的分布及结构唯一性分析**。**a.** ICSD数据库展示了晶系（内环）及其对应空间群（外环）的分布，颜色深浅表示结构的数量和百分比。**b.**Materials Project数据库以相同的层级表示方式展示晶系和空间群。**c.** 饼图量化了无机晶体结构数据库和Materials Project数据库中的结构唯一性和冗余，其中“唯一”结构在晶体学上不同，“等效”结构则根据第2节详述的结构相似性度量，被认定为与某个唯一结构相似。**d.** RRUFF数据库按照与图**a** 和**b**相同的可视化方案展示晶系和空间群的分布。七大晶系为：1-三斜晶系、2-单斜晶系、3-正交晶系、4-四方晶系、5-三方晶系、6-六方晶系和7-立方晶系。色标同时表示各数据库中晶系和空间群的绝对计数及相对百分比。



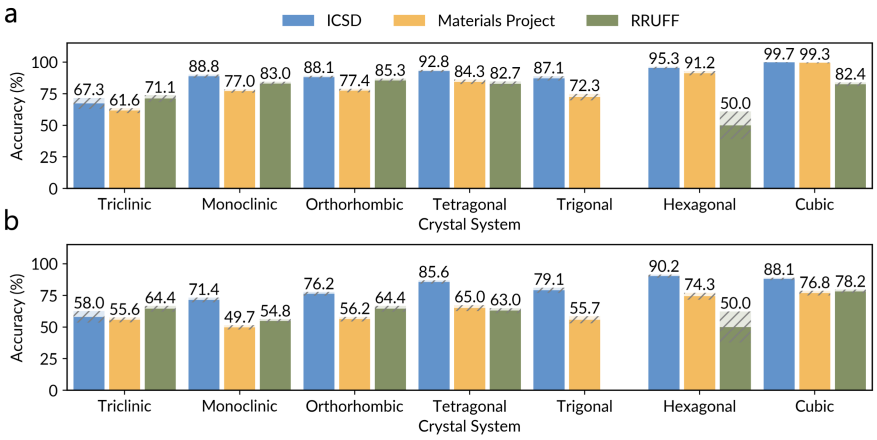
补充图 2：**晶体学数据库中的晶格长度分布。a.** 最终无机晶体结构数据库数据集中结构的尼格利约化晶胞晶格长度（a、b、c）直方图。**b.** 最终 Materials Project 数据集中结构的相应晶格长度分布。**c.** 用于回归的 RRUFF数据集中结构的晶格长度分布。



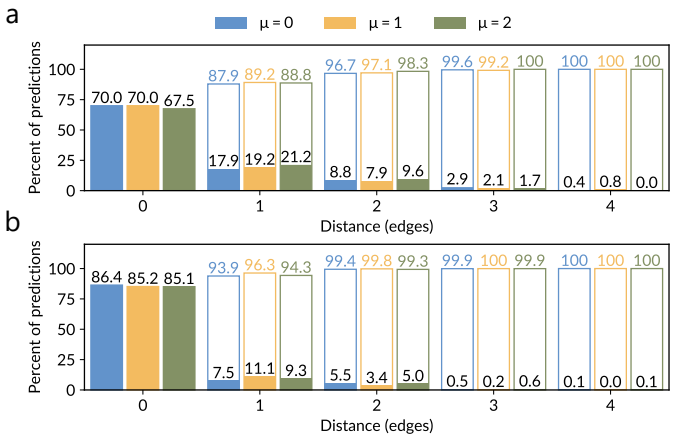
补充图 3：**晶体学数据库中的晶格角分布。a.** 最终无机晶体结构数据库数据集中结构的尼格利约化晶胞晶格角（α、 β、 γ）直方图。**b.** 最终 Materials Project 数据集中结构的相应晶格角分布。**c.** 用于回归的 RRUFF数据集中结构的晶格角分布。



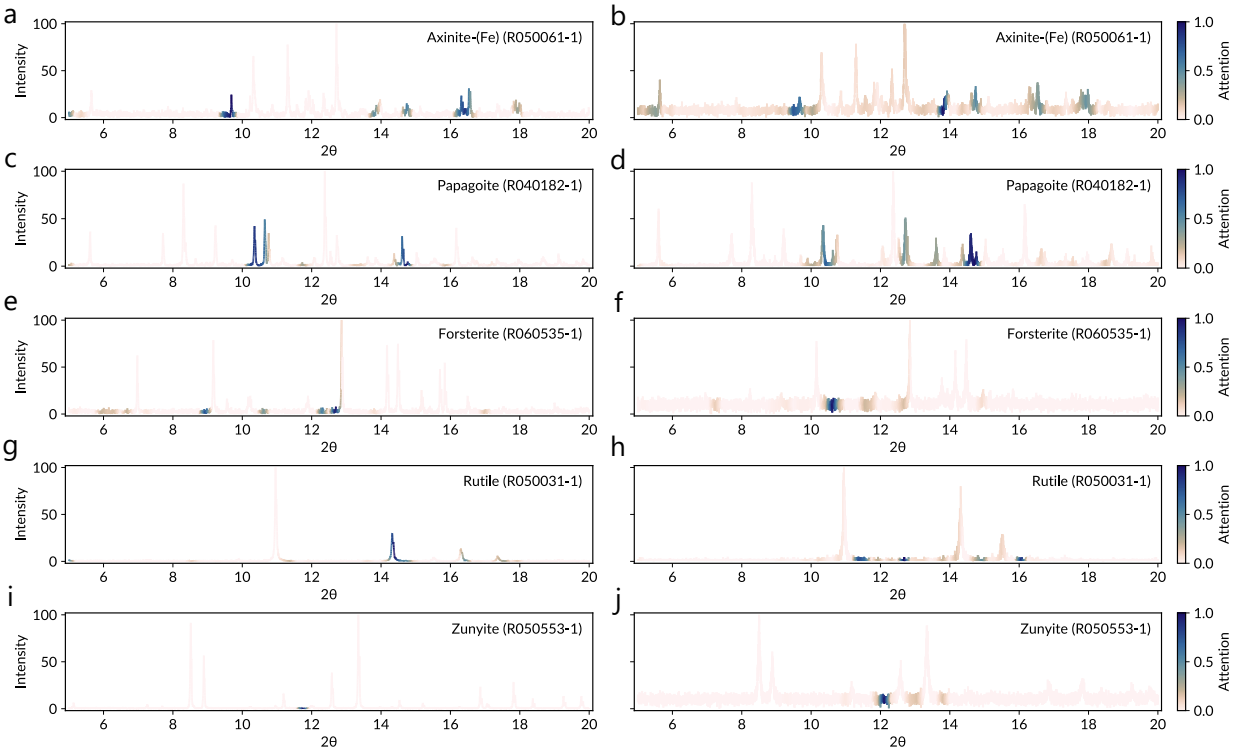
补充图4：**X射线衍射图谱中噪声的表征与模拟**。**a.** 用于噪声估计的代表性粉末X射线衍射图谱，其中背景区域以绿色下划线标示。**b-c.** RRUFF数据集中 λ_{max}值的分布，分别以线性（**b**）和对数（**c**）尺度表示；其中 λ_{max} 作为泊松分布的均值，用于抽样生成模拟所需的噪声值。虚线标示了选取用于可视化的代表性 λ_{max} 值。**d.** RRUFF数据库中选取的实验粉末X射线衍射图谱，展示了代表性的泊松噪声水平。**e.** 通过从均值对应于代表性 λ_{max} 值的分布中抽样并施加泊松噪声生成的模拟粉末X射线衍射图谱。**f-g.** RRUFF数据集中 σ_{rel} 值的分布，分别以线性（**f**）和对数（**g**）尺度表示；其中 σ_{rel} 表示用于抽样生成模拟所需噪声值的高斯分布的标准差。虚线标示了选取用于可视化的代表性 σ_{rel} 值。**h.** 选取的实验RRUFF图谱，展示了代表性的高斯噪声水平。**i.** 施加高斯噪声的模拟粉末X射线衍射图谱，噪声水平对应于代表性 σ_{rel} 值。所有粉末X射线衍射图谱均以归一化强度为纵轴、 2θ （度）为横轴绘制。



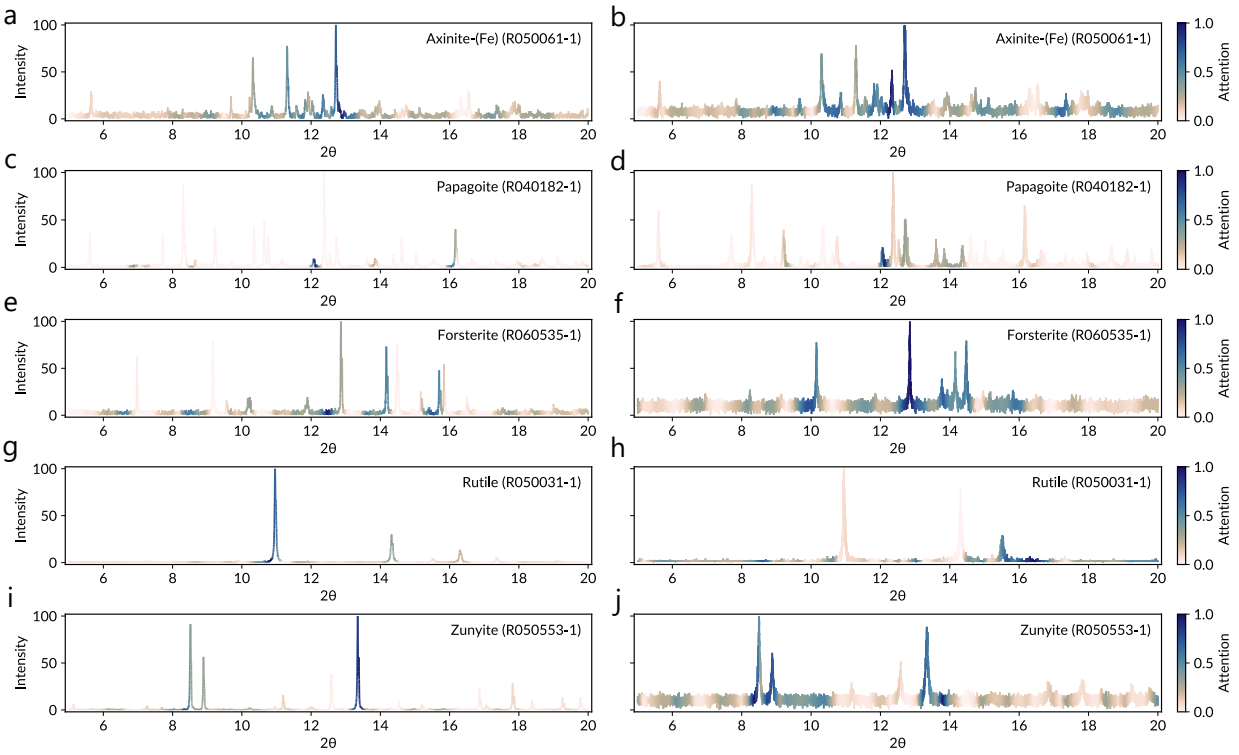
补充图 5：**各晶系的晶系和空间群分类准确率**。在无机晶体结构数据库（蓝色）、Materials Project（橙色）和 RRUFF（绿色）数据集上评估的晶系分类准确率（%）：**a.** 晶系分类和**b.** 空间群分类。每组展示了模型在七大晶系上的表现。带斜线的柱表示集成模型与增强不确定性的组合结果。立方晶系在所有数据集中均取得最高准确率，而对称性较低的晶系（三斜晶系、单斜晶系）表现出更大的波动，尤其是在 Materials Project 和 RRUFF 数据集中。



补充图 6：**基于实验和合成 RRUFF 数据的空间群预测评估**。对于具有完整结构的 RRUFF 数据子集（240 个样本），以**a.** 实验和**b.** 合成粉末X射线衍射图谱作为输入时，预测误差分布随真实空间群的图距离（边数）变化的情况。实心柱表示采用三种不同 GEMD 损失项权重（$\mu = 0$、1、2）时各距离对应的预测百分比；空心柱上的标注值表示累积百分比。



补充图7：**用于晶系分类的特征重要性 GradCAM 注意力图**。左列显示来自 RRUFF数据库的实验粉末 PXRD 图谱，右列显示相同矿物结构的合成衍射图谱。各行代表不同的晶系：**a-b.** 三斜晶系，**c-d.** 单斜晶系，**e-f.**正交晶系，**g-h.** 四方晶系，**i-j.** 立方晶系。颜色刻度表示从最后一个 ConvNeXt 块中提取的归一化注意力权重。



补充图 8：**用于空间群分类的特征重要性 GradCAM 注意力图**。左列显示来自 RRUFF数据库的实验粉末 PXRD 图谱，右列显示相同矿物结构的合成衍射图谱。各行代表不同的晶系：**a-b.** 三斜晶系，**c-d.** 单斜晶系，**e-f.**正交晶系，**g-h.** 四方晶系，**i-j.** 立方晶系。颜色刻度表示从最后一个 ConvNeXt 块中提取的归一化注意力权重。

参考文献

[S1] W B Park 等人。“使用卷积神经网络进行晶体结构分类”。载于：IUCrJ 4.4（2017年），第486–494页。[S2] Pascal Marc Vecsei 等人。“基于神经网络的X射线衍射图谱晶体对称性分类”。英文。载于：Phys. Rev. B. 99.24（2019年6月）。[S3] Alexander N. Zaloga 等人。“使用卷积神经网络对粉末X射线衍射图谱进行晶体对称性分类”。载于：Materials Today Communications 25（2020年），第101662页。DOI：10.1016/j. mtcomm.2020.101662。[S4] Sathya R Chitturi 等人。“根据X射线粉末衍射图谱自动预测晶格参数”。载于：J. Appl. Crystallogr. 54.6（2021年），第1799–1810页。[S5] Byung Do Lee 等人。“机器学习对称性识别与性质预测，只需粉末X射线衍射图谱”。载于：Advanced Intelligent Systems 4.7（2022年）。[S6] Byung Do Lee 等人。“粉末X射线衍射图谱分析的深度学习方法：同时应对泛化能力与扰动问题”。载于：Advanced Intelligent Systems 5.9（2023年）。[S7] Nicola Corriero 等人。“CrystalMELA：用于晶系判定的新型晶体学机器学习平台”。载于：J. Appl. Crystallogr. 56.2（2023年），第409–419页。[S8] Juan Iván Gómez-Peralta、Xim Bokhimi 和 Patricia Quintana。“利用卷积神经网络辅助评估X射线粉末衍射的晶格参数”。载于：《物理化学杂志A》127.36（2023年），第7655–7664页。[S9] Jerardo E Salgado 等人。“利用深度学习模型自动分类大规模X射线衍射数据”。英文。载于：Npj Comput. Mater. 9.1（2023年12月），第214页。[S10] Eric A Riesel 等人。“利用生成式机器学习从粉末衍射图谱测定晶体结构”。英文。载于：J. Am. Chem. Soc.（2024年9月）。[S11] Min Lin、Qiang Chen 和 Shuicheng Yan。“Network In Network”。载于：arXiv *[cs.NE]*（2013年12月）。[S12] Haotong Liang 等人。“CRYSPNet：通过神经网络预测晶体结构”。载于：Physical Review Materials 4.12（2020年），第123802页。[S13] Ashish Vaswani 等人。“Attention Is All You Need”。载于：arXiv（2017年）。[S14] Alexey Dosovitskiy 等人。“一幅图像胜过16×16个词：用于大规模图像识别的变换器”。载于：arXiv（2020年）。[S15] Kamal Choudhary。“DiffractGPT：使用生成式预训练Transformer从X射线衍射图谱测定原子结构”。英文。载于：J. Phys. Chem. Lett.（2025年2月），第2110–2119页。[S16] Albert Q Jiang 等人。“Mistral 7B”。载于：arXiv *[cs.CL]*（2023年10月）。[S17] Dejan Zagorac 等人。“无机晶体结构数据库的最新进展：理论晶体结构数据及相关特征”。载于：Applied Crystallography 52.5（2019年），第918–925页。[S18] Anubhav Jain 等人。“评述：材料项目：加速材料创新的材料基因组方法”。载于：APL materials 1.1（2013年）。[S19] Barbara Lafuente 等人。“数据库的力量：RRUFF项目”。载于：Highlights in mineralogical crystallography 1（2015年），第25页。[S20] Nils ER Zimmermann 和 Anubhav Jain。“用于量化配位环境与晶体结构相似性的局域结构序参量和位点指纹”。载于：RSC advances 10.10（2020年），第6063–6081页。[S21] Logan Ward 等人。“Matminer：用于材料数据挖掘的开源工具包”。载于：Computational Materials Science 152（2018年），第60–69页。[S22] Pauli Virtanen 等人。“SciPy 1.0：用于科学计算的Python基础算法”。载于：Nature methods 17.3（2020年），第261–272页。