赛题七：AI 算法在新材料为未知相指标化中的应用

一、赛题背景

新材料的发展是推动科技进步与产业转型的核心动力，对新材料晶体结构的精确解析至关重要。X 射线粉末衍射（XRD）技术是揭示材料晶体结构的关键技术之一，90% 以上的新型功能材料晶体结构解析依赖该技术。指标化作为粉末 XRD 数据分析的第一步，将衍射峰序列转化为晶面指数并反推晶胞参数，是解析复杂相组成的前提。然而，现有指标化算法对低对称性晶体、含有杂质峰的衍射图等适用性不足，制约了新材料的发现。因此，利用机器学习与人工智能算法解决这些问题，对推动新材料研发及产业化应用具有重要意义。

二、赛题应用场景

指标化作为粉末 XRD 数据分析的第一步，将衍射峰序列转化为晶面指数并反推晶胞参数，是解析复杂相组成的前提。然而，现有指标化算法对低对称性晶体、含有杂质峰的衍射图等适用性不足，制约了新材料的发现。在实际测量中，数据质量受仪器分辨率、零点漂移、样品放置偏差、样品制备不均等因素影响，存在信噪比低、峰形不清、峰位置移动等问题。同时，未知相材料多为多相混合，衍射峰重叠或模糊，且晶胞参数与衍射角度间是非线性关系，传统解析方法难以获得全局最优解。因此，利用机器学习与人工智能算法解决这些问题，对推动新材料研发及产业化应用具有重要意在材料研发企业的实验室中，科研人员对新合成的材料进行结构分析时，常借助粉末 XRD 技术获取衍射数据。但由于材料可能存在杂质、晶体对称性低等情况，传统指标化方法难以准确处理这些数据。通过人工智能算法，可对复杂的粉末 XRD 数据进行快速、精准的指标化处理，为科研人员提供准确的晶胞参数和衍射峰面指数，帮助他们深入了解材料晶体结构，进而优化材料性能，开发出更具优势的新材料，应用于电子、能源、航空航天等多个领域。在高校和科研机构的材料研究项目中，对于新型材料的探索也面临着同样的晶体结构解析难题。准确的指标化结果有助于科研人员深入研究材料的物理化学性质，推动基础材料科学的发展。

三、出题信息

（一）出题单位

本赛题由大赛组委会组织专家出题。

（二）赛题顾问

本赛题由冯振杰副教授担任技术总顾问。

（三）支持单位

本赛题由全球校园人工智能算法精英大赛组委会、苏州实验室等单位提供应用场

景、数据集、设备平台等支持。

四、赛题任务

参赛者需使用 Cu 靶 X 射线衍射数据集及其标注信息，设计并实现人工智能算法，预测晶胞参数（a,b,c,α,β,γ），并标定每一个衍射峰的晶面指数（hkl）并识别杂峰。

在晶体学中，衍射峰位置对应的晶面间距（d-spacing）是根据晶胞参数 (a, b, c, α, β, γ) 及晶面指数 (h, k, l) 计算得出的。不同晶系下公式有所不同，但可以使用通用三斜晶系公式，适用于任意晶系。

$$\frac{1}{d^2}=\frac{1}{V^2}[h^2b^2c^2\sin^2\alpha+k^2a^2c^2\sin^2\beta+l^2a^2b^2\sin^2\gamma+2hkabc^2(\cos\alpha\cos\beta-\cos\gamma)+2hlab^2c(\cos\gamma\cos\alpha-\cos\beta)+2kla^2bc(\cos\beta\cos\gamma-\cos\alpha)]$$

其中：

a, b, c 为晶胞边长；

α, β, γ 为晶胞内角（单位：弧度）；

h, k, l 为晶面指数；

V 为晶胞体积，其计算公式为：

$$V=abc\sqrt{1-\cos^2\alpha-\cos^2\beta-\cos^2\gamma+2\cos\alpha\cos\beta\cos\gamma}$$

参赛者的任务如下：

1. 晶胞参数预测：预测晶胞参数（a,b,c,α,β,γ）。

2. 晶面指数分配：为每个衍射峰分配晶面指数（hkl）,并识别杂峰位置。

五、数据集及数据说明

（一）数据来源

数据来源于晶体学开源数据库 Crystallography Open Database (https://www.crystallography.net)，涵盖了七类晶系，具有一定的代表性。

（二）数据规模

数据包含计算相应的粉末衍射图及衍射峰标注信息，共 7000 条数据，其中训练集 4900（70%），用于参赛者训练算法模型；测试集 1050（15%），用于模型调优和算法评估；验证集 1050（15%），用于最终结果测评。

（三）数据格式

数据包含 XRD 模拟图谱文件（.xy）和信息标注文件（.json）。

模拟图谱文件为模拟计算的 Cu 靶（波长 1.5406，1.5444）XRD 图谱,其中添加了噪声、杂峰（单斜三斜晶系不添加，其余晶系 1-2 个）、背景（指数衰减背景）、零点漂移（0.001-0.1°），样品偏移（衍射仪测角仪半径 200mm,偏移量 0.001-0.1mm），其强度已经进行了归一化处理（最大强度为 100）。

信息标注文件包含：空间群号及晶胞参数信息、衍射峰信息（衍射峰位置、强度以及其对应的晶面指数 HKL）、额外添加的衍射峰信息、添加的零点漂移和样品偏移等（详细见文件及说明），示例数据可以在 https://workdrive.zoho.com.cn/folder/q4a331755316b567a4a2197e3223d887b2a78 下载。

六、算法设计要求

（一）模型类型

鼓励参赛者采用深度学习算法进行晶胞参数预测和晶面指数分配，如 Transformer 及其变体。

（二）创新性

鼓励提出创新的算法架构或改进现有算法，以提高晶胞参数预测和晶面指数分配的准确率。

（三）可扩展性

算法应具备良好的可扩展性，能够在不同配置的计算设备上运行，且在处理大规模数据时性能稳定。例如，算法应能够在普通工作站和云端服务器上高效运行，并且在增加数据量时模型性能不会出现明显下降。

七、性能指标要求

（一）主要指标

1. 晶胞参数均方根误差（RMSE）：用于评估晶胞参数预测的精度。

2. 晶面指标化准确率（Accuracy）：即指标化正确的晶面数量（角度差值小于 0.2°，即认为匹配成功）除以该物相总衍射峰数量，用于评估晶面指数的分配准确率。

（二）次要指标

1. 检测时间：算法对单个 XRD 图谱进行晶胞参数预测和晶面指数分配的时间。计算方式为将测试集中所有 XRD 图谱的检测与分类时间相加，再除以 XRD 图谱总数。

2. 模型大小：训练得到的模型文件大小，是衡量模型复杂 度和存储需求的重要指标。较小的模型大小表明模型复杂度较低，存储成本和部署难度也相对较低，更便于在不同设备和环境中应用。

八、功能要求

（一）准确性

算法预测晶胞参数时须具备高准确性，确保对低对称或有杂峰的 XRD 图谱也能精准预测。对于晶面指数的分配，预测结果应与真实情况高度吻合。

（二）可靠性

面对属于不同晶系的 XRD 数据，算法应能稳定运行，输出可靠结果。即使存在杂峰、零点漂移、样品放置偏差等干扰因素，算法也不应出现大幅波动，而是保持对晶胞参数预测与晶面指数分配的准确性和稳定性。

九、开发环境

（一）软件环境

编程环境推荐使用 Python，建议使用 Python3.6 及以上版本，因其具有丰富的科学计算库和深度学习框架支持。

深度学习框架推荐使用 TensorFlow2.x 或 PyTorch1.x，这两个框架在深度学习领域广泛应用，具有高效的计算性能和丰富的 API，便于模型的搭建、训练和部署。

寻峰算法推荐使用 Scipy 库，Scipy 是 Python 中用于科学计算的核心库，涵盖信号处理、图像处理等多领域的算法和函数。

（二）硬件环境

参赛者可使用本地工作站或云端计算平台进行开发和训练。 本地工作站需配备 NVIDIAGPU（如 GTX10 系列及以上，或 RTX 系列）以加速深度学习计算；云端平台可选择阿里云天池、讯云 TI 平台、百度 AIStudio 等，这些平台提供了多种配置的计算资源，方便参赛者根据需求灵活选择。

十、成绩评价

（一）输入数据格式要求

参赛者的算法应能正确读取主办方提供的‘xy’ 格式的 XRD 图谱以及 JSON 格式的标注文件。此外参赛者应该使用合适的参数和寻峰算法提取图谱中的衍射峰；对于 JSON 标注文件，要准确提取晶胞参数、衍射峰等信息。

（二）输出数据格式要求

输出数据格式应与输入文件格式相同，部分参数可选填，但是一定要包含以下的参数，文件名应以对应 XRD 图谱的名字命名，格式是“.json”。输出结果数量与输入数据数量相同，用户提交结果问题时不应该漏交（否则没有得分）：

1. 晶胞参数输出：以 JSON 格式输出，例如： "crystal_info": {"a": 2.7941,"b": 2.7941,"c": 2.7941,"alpha": 90.0,"beta": 90.0,"gamma": 90.0}。

2. 晶胞参数分配输出，（注意在评判时不会采用选手计算的角度，而是采取选手提交的晶胞参数与 hkl 来计算角度）：同样以 JSON 格式输出[two_theta, intensity, d_spacing, hkl]，例如： "peaks": [

[45.8943, 100.0000, 1.9757, [1, 1, 0]],

[84.9541, 33.9401, 1.1407, [2, 1, 1]],

[121.3375, 26.9661, 0.8836, [3, 1, 0]],

[66.9229, 15.9598, 1.3971, [2, 0, 0]],

[102.4779, 12.9417, 0.9879, [2, 2, 0]]

]

3.零点漂移（0.001-0.1°），样品偏移（衍射仪测角仪半径 200mm,偏移量 0.001-0.1mm）可以选择提交。

（三）成绩计算公式

成绩将根据算法输出结果与真实标注数据对比，依据性能评估指标（如 RMSE、准确率 Accuracy 等）进行打分。

比赛得分=0.9*Accuracy-0.1*RMSE

十一、解题思路

（一）特征提取

利用卷积神经网络强大的特征提取能力，通过设计不同的卷积层、池化层组合，提取衍射峰的位置等特征。利用 Transformer 架构提取衍射峰之间位置的相对信息。

（二）模型训练

选择合适的深度学习框架（如 TensorFlow、PyTorch）搭建模型，设置合理的训练参数，如学习率、迭代次数、批量大小等。在训练过程中，采用交叉验证方法，利用验证集数据对模型进行评估和调优，防止模型过拟合。

（三）模型融合与优化

可尝试将多个不同结构或训练阶段的模型进行融合，如采用投票法或加权平均法，综合多个模型的预测结果，提高最终预测的准确性。同时，根据性能评估指标，对模型进行针对性优化，如调整模型结构、增加训练数据量等。

（四）任务分解

如果使用单一模型来处理问题比较困难，参数者可以将问题分解为合适的子任务，通过整合子任务的结果完成任务。

十二、参考资源

（一）粉末衍射与晶体学基础

《X 射线衍射学原理》（第 3 版）：作者是 [英] B.D. Cullity 等。这本书是

X 射线衍射领域的经典教材，系统阐述了 X 射线衍射的基本原理、实验方法和应用实例，包括粉末衍射的布拉格定律、衍射峰的形成原理等基础知识，为理解粉末 XRD 数据处理与指标化提供坚实的理论基础。

（二）人工智能与机器学习基础

《深度学习》（Goodfellow 等著）：本书由深度学习领域的知名专家撰写，是深度学习领域的权威教材。它系统地介绍了深度学习的基本概念、数学基础、神经网络模型（如 CNN、RNN、Transformer 等）及其优化方法，帮助参赛者深入理解深度学习算法的本质，从而更好地将其应用于晶胞参数预测和晶面指数分配任务。

（三）相关论文

在 IEEEXplore、ACMDigitalLibrary 等学术数据库中搜索关于基于粉末衍射图谱晶胞参数预测的最新研究论文，如 “Powder Diffraction Indexing as a Pattern Recognition Problem: A New Approach for Unit Cell Determination Based on an Artificial Neural Network” ,“Convolutional Neural Networks to Assist the Assessment of Lattice Parameters from X-ray Powder Diffraction”等，了解当前该领域的前沿技术和研究方法。

十三、提交要求

（一）算法代码

提交完整的算法代码，包括数据预处理、模型训练、预测推理等各个环节的代码。代码需使用 Python 语言编写，并遵循 PEP8 编程规范，具备清晰的注释和文档说明，以便评审人员理解和运行。

（二）技术报告

提交详细的技术报告，内容包括算法设计思路、模型架构图、实验设置（如训练参数、数据增强方法等）、性能分析（对主要指标和次要指标的详细分析）以及算法的创新点和不足之处。 技术报告格式采用 PDF，字数不少于 3000 字。

（三）结果文件

输出数据格式应与输入文件格式相同，部分参数可选填，但是一定要包含以下的参数，文件名应以对应 XRD 图谱的名字命名，格式是“.json”。输出结果数量与输入数据数量相同，用户提交结果问题时不应该漏交（否则没有得分）。详细见（十、成绩评价）。

（四）模型文件

提交训练好的模型文件，并提供模型的加载和使用说明，包括所需的运行环境、依赖库等信息，建议将模型文件转为 ONNX 模型格式。模型文件应能够在指定的测

试环境中正常运行并输出预测结果。

十四、更新与答疑

（一）更新与答疑

说明赛题是否可能会进行更新，以及参赛者在遇到问题时的咨询方式。每个赛题单独设立选手答疑 QQ 群（选手正式报名后可看到）。

（二）赛题智能助手

每个赛题将建立一个在线智能助手，方便选手进行赛题咨询。

十五、比赛流程及奖项设置

（一）报名阶段

参赛者在比赛官方网站上完成报名注册，提交个人或团队信息，获取初赛数据下载链接。

（二）初赛阶段

参赛者利用赛事方提供的训练数据集进行算法模型设计，利用提供的初赛测试集进行相应方法的验证与调试。初赛阶段参赛者每天提交结果的次数不限，但是初赛排行榜每隔 1 小时刷新一次。

（三）复赛(省赛)阶段

初赛结束后进入复赛阶段，开放复赛数据下载链接。仅有初赛阶段提交有效结果的参赛团队可以进入复赛。复赛期间，参赛者利用赛事方提供的复赛阶段数据进行算法模型调试，提交对复赛测试数据的推理结果。复赛阶段持续 3 天，每个参赛队伍每天仅能提交 2 次。复赛排行榜每隔 1 小时刷新一次。

（四）复赛(省赛)成绩公布

在比赛官方网站上公布复赛成绩。以进入复赛参赛团队数量作为计奖基数，按照不超过大赛省赛设奖比例，评选出复赛一、二、三等奖(颁发省赛获奖证书)。评选复赛奖过程中，参赛者提交的算法性能低于赛事方提供的基线参考分数的判定为无效成绩，不予授奖。复赛一、二等奖晋级参加国赛总决赛。

（五）决赛(国赛)阶段

1. 决赛线上评选。晋级决赛的参赛团队，依据复赛排行榜结果，以进入决赛参赛团队数量作为计奖基数，按照不超过大赛国赛设奖比例，评选出国赛一等奖候选名单及国赛二、三等奖获奖名单(颁发国赛二、三等奖证书)。

2. 决赛作品提交。国赛一等奖候选团队在规定时间内提交技术文档、算法代码和模型文件、演示视频、补充材料等。提交截止后，不再接受任何形式的修改和补充。

3. 决赛审核阶段。由专业评审团队对国赛一等奖候选参赛团队的参赛作品进行结果复现与审核。评审过程中如有疑问，可要求参赛者进行解释说明。

4. 决赛线下答辩。国赛一等奖候选团队在规定时间内提交完善后的技术文档、算法代码和模型文件、演示视频、补充材料，参加国赛线下总决赛复核答辩，最终依据算法性能得分和线下答辩得分确定国赛一等奖获奖名单及其排名(未参加线下复核答辩视同放弃奖项)。国赛一等奖颁发荣誉证书。

十六、其他说明

（一）公平性

严禁任何形式的作弊行为，包括但不限于数据泄露、模型预训练数据与测试数据重叠、抄袭他人代码等。一经发现，立即取消参赛资格，并追究相关责任。

（二）知识产权

参赛者提交的作品必须为原创，未在其他比赛中获奖或公开发表。比赛主办方有权对参赛作品进行展示、宣传等相关活动，但知识产权仍归参赛者所有。

十七、联系方式

赛项交流 QQ 群：879542469

邮　　箱：whitestar@shu.edu.cn

报名官网：www.aicomp.cn