Competition 7: Application of AI Algorithms in Indexing Unknown Phases in New Materials

1. Competition Background

The development of new materials is the core driving force behind scientific progress and industrial transformation, and precise analysis of crystal structures is of vital importance. X-ray powder diffraction (XRD) technology is one of the key methods for revealing the crystal structures of materials, with more than 90% of crystal structure determinations for new functional materials relying on this technique. Indexing, as the first step in powder XRD data analysis, converts diffraction peak sequences into Miller indices and back-calculates unit cell parameters, serving as the prerequisite for analyzing complex phase compositions. However, the applicability of existing indexing algorithms is limited for low-symmetry crystals and diffraction patterns containing impurity peaks, which restricts the discovery of new materials. Therefore, employing machine learning and artificial intelligence algorithms to address these challenges holds significant importance for advancing new material research and promoting industrial applications.

2. Competition Application Scenario

Indexing, the critical first step in powder XRD data analysis, converts sequences of diffraction peaks into Miller indices and deduces unit cell parameters—forming the foundation for resolving complex phase compositions. However, existing indexing algorithms often struggle with low-symmetry crystals and diffraction patterns containing impurity peaks, limiting their effectiveness in novel material discovery. Practical measurements further introduce complications such as low signal-to-noise ratios, poorly resolved peak shapes, and peak position shifts due to instrumental resolution limitations, zero-point drift, sample misplacement, and inhomogeneous sample preparation. Moreover, unknown phases frequently occur in multiphase mixtures where diffraction peaks overlap or appear broad, and the inherent nonlinear relationship between unit cell parameters and diffraction angles makes it difficult for conventional methods to reach a globally optimal solution.These challenges underscore the significance of applying machine learning and artificial intelligence algorithms to overcome such limitations, thereby accelerating the development and industrial application of new materials.In R&D laboratories of material enterprises, researchers routinely rely on powder XRD to analyze the structure of newly synthesized materials. However, impurities or low crystal symmetry often hinder accurate interpretation using traditional indexing approaches. AI-enhanced algorithms can process complex XRD data rapidly and precisely, delivering reliable unit cell parameters and Miller indices. These insights facilitate a deeper understanding of crystal structures, enabling performance optimization and the design of

superior materials for applications across electronics, energy storage, aerospace, and other fields.Similarly, in academic and research institutions, the accurate resolution of crystal structures remains a central challenge in the exploration of new materials. Precise indexing plays an essential role in advancing the study of fundamental physicochemical properties, thereby driving progress in materials science.

3. Competition Information

3.1 Organizing Unit:

Prepared by experts appointed by the Competition Organizing Committee.

3.2 Competition Advisor:

Associate Professor Zhenjie Feng serves as the technical advisor.

3.3 Supporting Units:

This competition is supported by the Global Global Campus Artificial Intelligence Algorithm Elite Competition Organizing Committee, Suzhou Laboratory, and other institutions, providing application scenarios, datasets, and computing platforms.

4. Competition Task

Participants are required to use the Cu-target X-ray diffraction dataset and its annotations to design and implement AI algorithms that predict unit cell parameters (a, b, c, α, β, γ), assign Miller indices (hkl) to each diffraction peak, and identify impurity peaks.

In crystallography, the interplanar spacing (d-spacing) corresponding to a diffraction peak is calculated based on the unit cell parameters (a, b, c, α, β, γ) and Miller indices (h, k, l). Although the formula varies across different crystal systems, a general triclinic formula can be used, which is applicable to all crystal systems.

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

where:

a, b, c are the unit cell edge lengths;

α, β, γare the unit cell angles (in radians);

h, k, l are the Miller indices;

V is the unit cell volume, calculated as

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

The tasks for participants are as follows:

(1) Unit cell parameter prediction: predict the unit cell parameters (a, b, c, α, β, γ).

(2) Miller index assignment: assign Miller indices (hkl) to each diffraction peak and identify impurity peaks.

5. Dataset and Data Description

5.1 Data Source

The data is sourced from the Crystallography Open Database (https://www.crystallography.net), covering seven crystal systems with representative diversity.

5.2 Data Scale

The dataset contains simulated powder diffraction patterns and corresponding annotated diffraction peaks, with a total of 7,000 entries. Among these, 4,900 (70%) are used as the training set for model training; 1,050 (15%) as the test set for model tuning and algorithm evaluation; and 1,050 (15%) as the validation set for final performance assessment.

5.3 Data Format

The dataset includes XRD simulated spectrum files (.xy) and annotation files (.json).

Simulated spectrum files are Cu-target XRD patterns (wavelengths 1.5406 and 1.5444) obtained by simulation, with added noise, impurity peaks (not included for monoclinic and triclinic systems; 1-2 added for other crystal systems), background (exponential decay), zero-point drift (0.001-0.1°), and sample displacement (diffractometer radius 200 mm, displacement 0.001-0.1 mm). The intensities have been normalized, with the maximum intensity set to 100.

The annotation files include: space group numbers and unit cell parameter information; diffraction peak information (peak positions, intensities, and corresponding Miller indices HKL); additionally introduced diffraction peaks; and applied zero-point drift and sample displacement (see files and documentation for details).Sample data can be downloaded at:

https://workdrive.zoho.com.cn/folder/q4a331755316b567a4a2197e3223d887b2a78

6. Algorithm Requirements

6.1 Model Type

Participants are encouraged to adopt deep learning algorithms for unit cell parameter prediction and Miller index assignment, such as Transformers and their variants.

6.2 Innovativeness

Participants are encouraged to propose innovative algorithmic architectures or improve existing methods to enhance the accuracy of unit cell parameter prediction and Miller index assignment.

6.3 Scalability

The algorithm should be highly scalable, capable of running on computing devices with different configurations, and should maintain stable performance when handling large-scale datasets. For example, the algorithm should be able to operate efficiently both on standard

workstations and cloud servers, and its performance should not degrade significantly as the dataset size increases.

7. Performance Metrics

7.1 Primary Metrics

(1) Root Mean Square Error (RMSE) of unit cell parameters: used to evaluate the accuracy of unit cell parameter prediction.

(2) Indexing Accuracy: defined as the number of correctly indexed diffraction peaks (where the angular deviation is less than 0.2°, regarded as a successful match) divided by the total number of diffraction peaks for the given phase. This metric evaluates the accuracy of Miller index assignment.

7.2 Secondary Metrics

(1) Detection Time: the time taken by the algorithm to predict unit cell parameters and assign Miller indices for a single XRD pattern. It is calculated by summing the detection and classification times for all XRD patterns in the test set, and dividing by the total number of patterns.

(2) Model Size: the file size of the trained model, serving as an important indicator of model complexity and storage requirements. A smaller model size suggests lower complexity, reduced storage costs, and easier deployment across different devices and environments.

8. Functional Requirements

8.1 Accuracy

The algorithm must achieve high accuracy in predicting unit cell parameters, ensuring precise results even for low-symmetry XRD patterns or those containing impurity peaks. For Miller index assignment, the predicted results should closely match the ground truth.

8.2 Reliability

When processing XRD data from different crystal systems, the algorithm should run stably and produce reliable results. Even in the presence of impurity peaks, zero-point drift, or sample displacement, the algorithm should maintain accurate and consistent predictions of unit cell parameters and Miller index assignments, without large fluctuations.

9. Development Environment

9.1 Software Environment

The recommended programming environment is Python, preferably version 3.6 or above, due to its rich ecosystem of scientific computing libraries and deep learning framework support.

TensorFlow 2.x or PyTorch 1.x are recommended as the deep learning frameworks, as both are widely used in the field, providing efficient computational performance and abundant

APIs that facilitate model construction, training, and deployment.

For peak-finding algorithms, the SciPy library is recommended; SciPy is a core library for scientific computing in Python, covering algorithms and functions across multiple domains such as signal processing and image processing.

9.2 Hardware Environment

Participants may use either local workstations or cloud-based computing platforms for development and training. Local workstations should be equipped with NVIDIA GPUs (e.g., GTX 10 series or above, or RTX series) to accelerate deep learning computations. Cloud platforms such as Alibaba Cloud Tianchi, XunCloud TI Platform, or Baidu AI Studio may also be used, as these platforms provide a variety of resource configurations that participants can flexibly select according to their needs.

10. Evaluation Criteria

10.1 Input Data Format Requirements

Algorithms must correctly read the XRD spectra in .xy format and the corresponding annotations in .json format provided by the organizers. Participants should use appropriate parameters and peak-finding algorithms to extract diffraction peaks from the spectra. For the Json annotation files, the algorithms must accurately parse unit cell parameters, diffraction peaks, and related information.

10.2 Output Data Format Requirements

The output data must match the input file formats. Some parameters may be optional, but the following must be included. Output files should be named after the corresponding XRD spectrum and saved in .json format. The number of output files must match the number of input files. Missing outputs will result in no score.

(1) Unit cell parameter output: Output in JSON format, for example: "crystal_info": {"a": 2.7941, "b": 2.7941, "c": 2.7941, "alpha": 90.0, "beta": 90.0, "gamma": 90.0}.

(2) Peak assignment output: (Note: evaluation will not use participant-calculated angles; instead, the submitted unit cell parameters and hkl indices will be used to calculate angles). Output in JSON format as [two_theta, intensity, d_spacing, hkl], for example:"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) Zero-point drift (0.001-0.1°) and sample displacement (diffractometer radius 200

mm, displacement 0.001-0.1 mm): optional to submit.

10.3 Scoring Formula

Scores are determined by comparing algorithm outputs with the ground-truth annotations, based on performance evaluation metrics (e.g., RMSE and accuracy).

Final Score = 0.9*Accuracy-0.1*RMSE

11. Problem-Solving Approach

11.1 Feature Extraction

Leverage the powerful feature extraction capabilities of convolutional neural networks by designing different combinations of convolutional and pooling layers to extract peak positions and related features. Use Transformer architectures to capture relative positional relationships among diffraction peaks.

11.2 Model Training

Build models with appropriate deep learning frameworks (e.g., TensorFlow or PyTorch), and set suitable training parameters such as learning rate, number of iterations, and batch size. During training, employ cross-validation with the validation set to evaluate and tune the model, preventing overfitting.

11.3 Model Ensemble and Optimization

Experiment with ensemble approaches that combine models of different structures or training stages, using methods such as voting or weighted averaging to integrate predictions and improve final accuracy. Optimize models according to performance evaluation metrics by adjusting architectures or increasing training data volume.

11.4 Task Decomposition

If handling the task with a single model proves challenging, participants may decompose the problem into appropriate sub-tasks and integrate their results to accomplish the overall task.

12. Reference Resources

12.1 Fundamentals of Powder Diffraction and Crystallography

Principles of X-ray Diffraction (3rd edition), authored by B.D. Cullity et al. This classic textbook in the field of X-ray diffraction systematically explains fundamental principles, experimental methods, and practical applications, including Bragg's law for powder diffraction and the mechanisms of diffraction peak formation, providing a solid theoretical foundation for understanding powder XRD data processing and indexing.

12.2 Fundamentals of Artificial Intelligence and Machine Learning

Deep Learning (authored by Goodfellow et al.): This book, written by renowned experts in the field of deep learning, is considered an authoritative textbook. It systematically

introduces the fundamental concepts of deep learning, its mathematical foundations, neural network models (such as CNNs, RNNs, and Transformers), and optimization methods. It helps participants gain a deeper understanding of the essence of deep learning algorithms, thereby enabling them to better apply these methods to unit cell parameter prediction and Miller index assignment tasks.

12.3 Relevant Papers

Participants are encouraged to search in academic databases such as IEEE Xplore and the ACM Digital Library for the latest research papers on unit cell parameter prediction based on powder diffraction patterns, such as “Powder Diffraction Indexing as a Pattern Recognition Problem: A New Approach for Unit Cell Determination Based on an Artificial Neural Network” and “Convolutional Neural Networks to Assist the Assessment of Lattice Parameters from X-ray Powder Diffraction”. These papers will help participants understand the state-of-the-art techniques and research methods in the field.

13. Submission Requirements

13.1 Algorithm Code

Submit complete algorithm code, including all components such as data preprocessing, model training, and inference. The code must be written in Python, follow the PEP8 coding standard, and include clear comments and documentation to ensure reviewers can understand and execute it.

13.2 Technical Report

Submit a detailed technical report that includes the algorithm design approach, model architecture diagrams, experimental setup (e.g., training parameters, data augmentation methods), performance analysis (detailed analysis of both primary and secondary metrics), as well as the innovation points and limitations of the algorithm. The technical report must be in PDF format, with a minimum of 3,000 words.

13.3 Result Files

The output data format must match the input file format. Some parameters may be optional, but the following parameters must be included. Each output file should be named after the corresponding XRD spectrum, in “.json” format. The number of output results must match the number of input data files. Missing submissions for any input file will result in no score being awarded. See Section 10, Evaluation Criteria, for details.

13.4 Model Files

Submit the trained model files, along with instructions for loading and using the model, including the required runtime environment and dependencies. It is recommended to convert

the model files into ONNX format. The model must be runnable in the designated test environment and capable of generating prediction results.

14. Updates and Q&A

14.1 Updates and Q&A

This section specifies whether the task may be updated and how participants can seek assistance when encountering problems. Each task will have a dedicated QQ group for participant Q&A (available after official registration).

14.2 Task Intelligent Assistant

Each task will also establish an online intelligent assistant to facilitate participant inquiries.

15. Competition Process and Award Settings

15.1 Registration Stage

Participants complete registration on the official competition website, submit individual or team information, and obtain the preliminary dataset download link.

15.2 Preliminary Stage

Participants design algorithmic models using the training dataset provided by the organizers and validate/debug their methods using the preliminary test set. During the preliminary stage, the number of daily submissions is unlimited; however, the leaderboard is refreshed every hour.

15.3 Semifinal (Provincial) Stage

After the preliminary stage concludes, the semifinal stage begins, with the semifinal dataset download link released. Only teams that submitted valid results in the preliminary stage may advance. During the semifinal stage, participants use the provided data to tune their models and submit inference results on the semifinal test data. The semifinal lasts for 3 days, with each team allowed a maximum of 2 submissions per day. The semifinal leaderboard is refreshed every hour.

15.4 Semifinal (Provincial) Results Announcement

Semifinal results are announced on the official competition website. The number of teams advancing to the semifinal serves as the baseline for prize allocation. Awards for first, second, and third prizes at the provincial level (with certificates issued) are selected according to the award ratio specified for the provincial competition. Submissions whose performance falls below the baseline reference score provided by the organizers will be deemed invalid and will not be awarded. Winners of the first and second prizes advance to the national finals.

15.5 Final (National) Stage

(1) Online Evaluation: For teams advancing to the finals, based on semifinal leaderboard results, the number of teams serves as the baseline for prize allocation. According to the award ratio specified for the national competition, a shortlist of first-prize candidates and winners of second and third prizes (with certificates issued) will be determined.

(2) Final Submission: National first-prize candidate teams must submit their technical documents, algorithm code and model files, demonstration videos, and supplementary materials within the specified time. No modifications or additions will be accepted after the submission deadline.

(3) Final Review: A professional review panel will reproduce and evaluate the submitted works of first-prize candidate teams. If any issues arise during the review, participants may be required to provide explanations.

(4) On-site Defense: First-prize candidate teams submit revised technical documents, algorithm code and model files, demonstration videos, and supplementary materials, and participate in the national on-site final defense. The final list of first-prize winners and their rankings will be determined based on both algorithm performance scores and defense performance. Failure to attend the on-site defense will be considered a waiver of the award. First-prize winners will receive honorary certificates.

16. Additional Notes

16.1 Fairness

Any form of cheating is strictly prohibited, including but not limited to data leakage, overlap between model pre-training data and test data, or plagiarism of others' code. Once discovered, the participant's qualification will be immediately revoked, and relevant responsibilities will be pursued.

16.2 Intellectual Property

All submissions must be original and must not have won awards in other competitions or been publicly published. The organizers reserve the right to display and promote submitted works for competition-related activities; however, the intellectual property rights remain with the participants.

17. Contact Information

Competition Q&A Group (QQ): 879542469

Email: whitestar@shu.edu.cn

Official Registration Website: www.aicomp.cn