

Accelerating Materials Discovery: Automated Identification of Prospects from X-Ray Diffraction Data in Fast Screening Experiments

Jan Schuetzke, Simon Schweidler, Friedrich R. Muenke, Andre Orth, Anurag D. Khandelwal, Ben Breitung, Jasmin Aghassi-Hagmann, and Markus Reischl*

New materials are frequently synthesized and optimized with the explicit intention to improve their properties to meet the ever-increasing societal requirements for high-performance and energy-efficient electronics, new battery concepts, better recyclability, and low-energy manufacturing processes. This often involves exploring vast combinations of stoichiometries and compositions, a process made more efficient by high-throughput robotic platforms. Nonetheless, subsequent analytical methods are essential to screen the numerous samples and identify promising material candidates. X-ray diffraction is a commonly used analysis method available in most laboratories which gives insight into the crystalline structure and reveals the presence of phases in a powder sample. Herein, a method for automating the analysis of XRD patterns, which uses a neural network model to classify samples into nondiffracting, single-phase, and multi-phase structures, is presented. To train neural networks for identifying materials with compositions not matching known crystallographic structures, a synthetic data generation approach is developed. The application of the neural networks on high-entropy oxides experimental data is demonstrated, where materials frequently deviate from anticipated structures. Our approach, not limited to these materials, seamlessly integrates into high-throughput data analysis pipelines, either filtering acquired patterns or serving as a standalone method for automated material exploration workflows.

1. Introduction

X-ray diffraction (XRD) has long been regarded as an indispensable tool for the characterization of material samples, which is capable of analyzing a wide array of substances, ranging from metals, ceramics, polymers, to thin films and nano-structured materials.^[1] One of the key factors behind the prevalent use of XRD is its ability to provide a comprehensive analysis of various distinct properties. For instance, the XRD technique allows for determining the material's phase composition, crystal structure, lattice parameters, texture, and strain, among other characteristics.^[2] Moreover, the diffraction analysis is a non-destructive technique, safeguarding the integrity of the material for further studies. Given these advantages, XRD instruments are ubiquitously present and essential for materials research workflows.


In the field of materials discovery, the primary goal is to develop materials with enhanced or unique properties that can outperform existing materials. Due to the inherent limitations of existing materials in

aspects such as performance, cost, and sustainability, the development of new substances is imperative for propelling technological advancements and elevating living standards. A prevalent approach to discovering these novel materials includes the intentional addition of foreign atoms or ions to existing components. This can lead to enhanced properties, such as thermal stability or electrical conductivity, or it can serve to replace scarce or environmentally harmful substances. One of the most effective methods for identifying such novel materials is the combinatorial approach, in which a multitude of different substances are systematically combined in varying proportions and configurations for rapid screening of vast material composition spaces.^[3] Nevertheless, a large fraction of these configurations unfortunately results in materials that exhibit inconsistent and inhomogeneous properties that are not desirable.^[4] Here, XRD is an essential tool for the identification of amorphous, phase-pure, and multi-phase samples, as well as further characterization of the crystalline properties for the produced materials.

The analysis of the data generated from the XRD technique, however, poses a considerable challenge. In the traditional

J. Schuetzke, F. R. Muenke, A. Orth, M. Reischl
Institute for Automation and Applied Informatics
Karlsruhe Institute of Technology
Hermann-von-Helmholtz Platz 1, 76344 Eggenstein-Leopoldshafen,
Germany
E-mail: markus.reischl@kit.edu

S. Schweidler, A. D. Khandelwal, B. Breitung, J. Aghassi-Hagmann
Institute of Nanotechnology
Karlsruhe Institute of Technology
Hermann-von-Helmholtz Platz 1, 76344 Eggenstein-Leopoldshafen,
Germany

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/aisy.202300501>.

© 2023 The Authors. Advanced Intelligent Systems published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/aisy.202300501

analysis of powder XRD patterns, similarity metrics such as the figure-of-merit (FOM) are typically used to compare measured signals with reference phases, as obtained from databases such as the ICSD or the COD.^[5,6] However, the presence of experimental artifacts, such as measurement noise and background signals, complicates the analysis process and necessitates manual preprocessing steps.^[2] Additionally, the incorporation of multiple elements into a single-crystal structure in newly developed multicomponent materials can lead to significant lattice distortions and reflection shifts, posing a challenge due to crucial deviations from the reference phases. Given the exponential surge in data volume generated by newly developed high-throughput systems,^[7,8] manual analysis of powder XRD data using the traditional FOM method becomes highly time-consuming and practically unfeasible. Consequently, the automation of XRD analysis becomes a necessity, enabling researchers to efficiently process and interpret large datasets, accelerating the pace of material discovery.

As an alternative to the manual data analysis, artificial neural networks have demonstrated promising results in the accurate and fast interpretation of unprocessed powder XRD data. Neural networks are trainable mathematical models that use interconnected neurons, layers, weights, and activation functions to map input data to output predictions. During training, the network adjusts the weights to minimize the difference between its predictions and the desired output, gradually learning to recognize complex patterns and make accurate predictions for new data. Within the domain of XRD analysis, Park et al. first developed a neural network to determine the crystal system, extinction group, and space group for scans of phase-pure samples.^[9] Since then, multiple publications have demonstrated the successful application of network models for the classification of single-phase^[10] or multi-phase samples.^[11–13] Beyond phase identification tasks, neural networks have shown promising performance in other applications, such as the determination of scale parameters or lattice constants from the XRD scans.^[14]

Expanding upon the foundational research, the application of neural networks to XRD data has extended to include their use for novel material discovery in experimental settings. For instance, Velasco et al. used a neural network to determine the crystal structure of unique compositions in complex multicomponent systems.^[7] Furthermore, Massuyeau et al. introduced a neural network capable of differentiating between perovskite and non-perovskite materials through their XRD patterns.^[15] Similarly, various studies have leveraged neural networks in XRD analysis for the identification of diverse novel materials, such as A15-type phases and quasicrystals.^[16–18] Additionally, Szymanski et al. deployed a neural network to identify target and intermediate phases in material synthesis experiments, enabling their optimization algorithm to determine the most suitable precursors and experimental parameters for the effective synthesis of the target phase.^[19]

To effectively train neural networks for phase identification in XRD datasets, previous research has predominantly utilized simulated training data.^[9–19] This approach involves generating synthetic diffraction patterns from crystallographic database entries, incorporating variations and experimental artifacts characteristic of actual experimental patterns, to ensure that models trained on simulated data effectively transfer their performance to actual experimental scans. The primary aim of this methodology is

to address the difficulty of obtaining a sufficiently large dataset containing high-quality XRD scans with their specific phase identification results, crucial for training the neural network. As an alternative to simulating the training data, Velasco et al. acquired phase-pure signals of the essential structures in their study and systematically altered these signals to enlarge the data basis.^[7]

Nonetheless, the existing studies on applying neural networks for the analysis of powder XRD patterns present some limitations. First, the exemplary data needed to train the network models is typically not at hand for novel materials. While databases provide reference materials from past studies, they are not typically equipped with information on newly synthesized materials. Alternatively, the method of altering measured patterns from phase-pure samples, as introduced by Velasco et al.,^[7] requires the synthesis of pristine samples, which is not a trivial task for complex materials. Second, an appropriate network structure is required to handle the peculiarities of the diffraction patterns. For instance, we evaluated commonly used neural network structures for the analysis of XRD in a recent study and identified deficiencies in detecting minor peaks in the diffraction patterns,^[20] which extend to the recognition of multi-phase samples in the material discovery data. Additionally, amorphous phases have not been considered in prior works, so modifications to the architecture of established networks are necessary to handle such components.

Therefore, we present a universal approach for the rapid identification of prospects from XRD data using a neural network structure. The model categorizes the samples into nondiffracting (including amorphous) and crystalline samples and accurately distinguishes between referenced and highly distorted structures that exhibit nonideal properties, such as the formation of multi-phase compounds. Training data is generated by simulating diffraction patterns based on a theoretical description of the desired structure in the form of a crystallographic information file (cif), eliminating the need for an initial production of pristine reference samples. The simulation of XRD patterns and training of the neural network takes less than 5 min on consumer-level hardware, so models are readily available for use cases at hand. In this work, we demonstrate the application of our approach on distinct material structures: multimetallic spinels and doped copper oxides.

2. Results

To train a neural network for the automated analysis of the acquired powder XRD patterns, we present a universal data generation pipeline that simulates realistic signals. Accordingly, **Figure 1** provides an overview of our presented method. First, synthetic patterns are generated based on variations of a description of a structure in the form of cif. Our model generates realistic variations of the base structure without the requirement of modeling the exact lattice and occupancy changes, providing a general approach to represent altered structures. Generally, for each variation, the position of the peaks, the ratio of peak heights, and the shape of the peaks are varied to account for naturally occurring variations. Prior research demonstrated that such variations are crucial to generate adequate training data for the application of neural networks to measured XRD patterns.^[12]

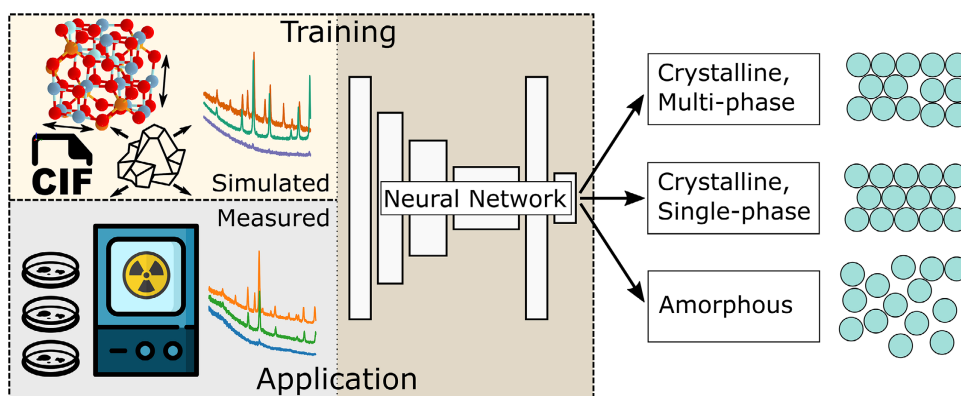


Figure 1. Concept of the presented method for training and application of a neural network to automatically categorize XRD patterns in materials discovery experiments. The neural network is trained with simulated signals that depict the range of variation in the experimental data. Subsequently, the model classifies measured XRD patterns into the amorphous, single-phase, and multi-phase categories. The spinel structure was obtained from the materials project.^[27]

In the context of doping experiments, for instance, the structure variations are depicted by lattice contractions and expansions that are reflected in the synthetic diffraction patterns without specifying the exact type and concentrations of the doping material. Furthermore, altered scattering factors of the incorporated species are reflected by varying the intensity ratios of the peaks in the pattern, and the width of the peaks is randomly chosen to mirror the varying crystallite sizes and defects. To depict multi-phase samples, the simulated patterns of the varied structures are complemented with arbitrary, additional diffraction peaks placed randomly. Finally, samples that lack a periodic atomic arrangement, including amorphous materials, are represented by patterns that only contain a diffuse background intensity without characteristic reflections.

Utilizing the simulated data, a specialized neural network is trained for automated classification of the XRD patterns. While this model is developed to categorize the three distinct classes encountered in fast screening experiments, we have elected to divide the classification task into two separate predictions. The initial model output discerns between nondiffracting and crystalline samples, while the second output differentiates between single-phase and multi-phase patterns. Both outputs use a sigmoid activation function (scaled between 0 and 1), allowing the predicted values to be interpreted as probability estimates for their respective classification tasks. In this context, the initial output predicts the sample's crystallinity, while the secondary probability estimate quantifies the likelihood of a multi-phase compound's presence. Should the initial output's predicted value fall below 0.5, the sample is designated as non-diffracting (amorphous), irrespective of the secondary output.

In the following sections, the adaptability of our approach is presented by applying trained neural network models to experimental data. Therefore, the doping of copper oxides and composition variations to form a spinel-type structure are tested in fast screening experiments, which enable the compilation of large and diverse datasets for the evaluation of our method. The networks have been trained using our generalized data generation pipeline with reference structures obtained from the ICSD,^[5] and

no modifications are required to apply the presented approach for the different datasets.

2.1. Spinel Structures

First, the described method is applied to identify spinel-type MgAl_2O_4 structures that incorporate a multitude of different elements. The respective materials class is called "high-entropy oxides", related to a high configurational entropy that is formed when many different elements are incorporated into a single-phase structure. Between the different elements, interactions arise, called cocktail effects, which can give these materials unique properties that can differ completely compared to the parent materials. In this study, the parent structure Fe_3O_4 (Fe(II) Fe(III) $_2\text{O}_4$) was used and the divalent and trivalent Fe replaced by other elements, forming, for example, $(\text{CuMg})(\text{FeMnCr})_2\text{O}_4$. The samples are produced and characterized on a high-throughput platform, which allows for parallel synthesis and analysis of 99 specimens (11×9 grid) using a robotic synthesis platform and a high-throughput sample holder for a Ga-Jet X-ray source.^[7] Accordingly, a cif (ICSD code 13 859), representing the spinel structure of MgAl_2O_4 , is used to generate the synthetic training data.

A brief, manual screening of the acquired data reveals the different outcomes of the unique precursor combinations, which include amorphous or multi-phase compounds instead of the intended, phase-pure spinel structure. **Figure 2a** illustrates exemplary XRD patterns of the three classes, which have been shifted on the intensity axis for clarity of presentation. In the case of the amorphous/nondiffracting class (gray), the XRD patterns mainly exhibit a background signal, with minor diffraction peaks observed in a few cases. Here, the 9×11 grid contained positions that were unoccupied, producing diffraction signals devoid of reflections, which were grouped with the amorphous class during analysis to simplify the process. In contrast, the single-phase XRD pattern (blue) shows diffraction peaks that stand out from the noise and background, with the major peak being located at $29^\circ 2\theta$. Likewise, the multi-phase samples (red) exhibit

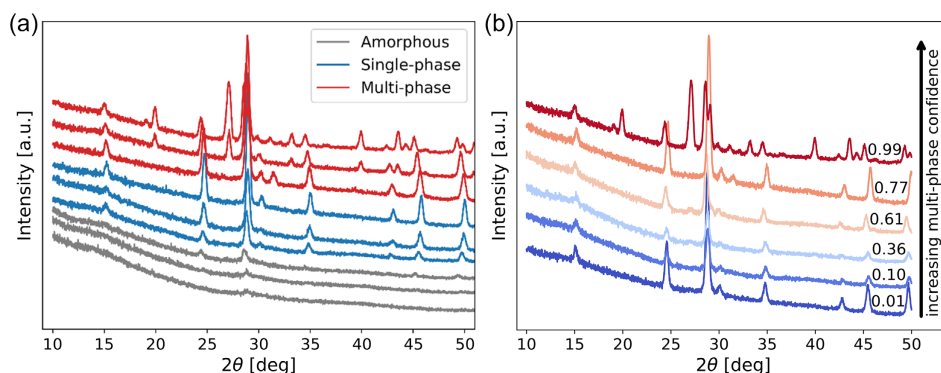


Figure 2. Experimental XRD patterns from the MgAl_2O_4 spinel-type structure experimental series. a) The diffraction patterns show examples of amorphous/nondiffracting (gray), single-phase (blue), and multi-phase (red) structures. b) Predicted confidences (blue: low; red: high) by our model grouping the exemplary XRD patterns into single- and multi-phase. The patterns with muted colors are close to the multi-phase detection threshold (0.5). The predicted value corresponds with the prominence of those extra peaks.

the same peaks as the single-phase structure in addition to other unassociated diffraction peaks.

While manual screening of the patterns is only feasible for limited sizes, the developed neural network analysis approach was applied to categorize the patterns within seconds. Figure 2b shows XRD patterns for identified crystalline samples and the corresponding multi-phase probability estimates (blue: low, red: high), as predicted by the model (second output). A visual assessment of the measured patterns alongside the predicted confidence scores demonstrates that the network has learned to detect multi-phase samples based on the presence and prominence of additional peaks. Signals that align precisely with the diffraction pattern of the identified structure are categorized as single phase (represented in blue), whereas irregular intensity baselines yield heightened multi-phase confidence predictions (muted blue to red colors). Regarding the light-blue pattern in Figure 2b, the marginally elevated background between 27° and 28° is ambiguous and could be due to noise or minor impurity peaks. Consistently, the corresponding pattern is classified as single-phase, since the multi-phase confidence lies below the detection threshold. In comparison, the light-red pattern exhibits even higher intensities in this range and is assigned to the multi-phase class by the model. Following this trend, patterns with even more distinct impurity peaks yield higher multi-phase probability estimates. Similarly, the model accurately identified the amorphous samples, as well as patterns that resulted from the empty grid positions.

2.2. Doped Copper Oxides

Using the identical robotic platform for sample preparation and subsequent characterization via XRD, the doping of copper oxide (CuO) is examined in a second experimental series. While the first experiment examined the unique precursor combinations to form the spinel structure, we additionally show how our approach can be used for automatic determination of the dopant concentration thresholds, thus avoiding the formation of multi-phase compounds and preserving the material's desired properties. Doped copper oxides have been produced and analyzed in various studies,^[21–24] but typically, only a few compositions are

tested. Depending on the concentration, the dopant material is either fully incorporated or forms an impurity phase, but diverging results have been reported with respect to the critical dopant concentration for synthesizing phase-pure samples. For example, Al-Amri and colleagues reported phase-pure oxides with Ni doping concentrations ranging from 1 to 7%,^[23] while Meneses et al. detected impurity phases for the same Ni-doped CuO nanoparticles, even for low concentrations of the dopant.^[21] While both studies analyzed the structures of the identical nanoparticles, the differing groups used distinct experimental routes and configurations to produce the doped structures (i.e., differing temperatures for the synthesis).

Doped copper oxides in the form of $\text{Cu}_{1-x}(\text{Zn,Ni,Mn})_x\text{O}$ were produced with systematically incremented doping concentrations ranging between 0% and 25%. Moreover, the materials were generated and examined within three discrete experimental conditions, with samples being calcined at temperatures of either 500, 600, or 700 °C.

Figure 3a shows several XRD patterns from the fast-screening experiments with varying dopants and compositions. While the ideal CuO sample exhibits two major diffraction peaks at about 28° and 30° 2θ (for the Ga-jet radiation source) and further, minor reflections at 25° and 39° , the doped copper oxides show additional reflections due to impurity phases, as highlighted by the black triangles on top of the respective patterns. High concentrations of the dopants (here 25%) cause the sample to form multi-phase compounds with ZnO , NiO , or MnO_2 being present, in addition to the intended CuO phase.^[21,22,24] The detection of those additional phases, however, remains a challenging task due to overlapping diffraction peaks of the doped copper oxide structure and the impurity phase. For example, the additional ZnO phase peaks are almost overlapping with one of the major peaks from the CuO structure (around 28°). In this series, all materials exhibit clear reflections in the XRD signal, so there are no amorphous samples produced for the doped copper oxides.

We applied our novel method for the doped copper oxides by generating varied structures from the CuO ICSD entry (code 16 025) and training a neural network for the classification of the respective XRD patterns. Using our trained network, we were able to analyze the 225 synthesized samples (75 unique

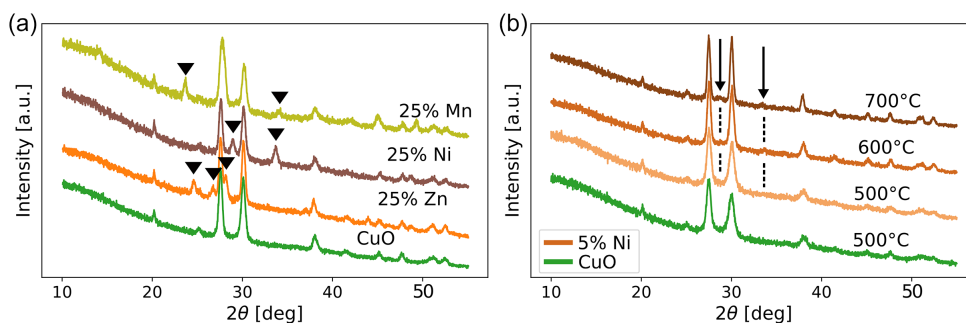


Figure 3. Experimental XRD patterns of pure and doped copper oxides (CuO). a) The dopants Zn, Ni, and Mn have been tested with concentrations up to 25% and cause additional peaks in the diffraction signal due to forming multi-phase compounds, as highlighted by the black triangles. b) Depending on the synthesis temperature, the XRD signals show either single-phase or multi-phase patterns for identical compositions. Here, the XRD patterns for the CuO samples with 5% Ni doping show signs of additional diffraction peaks at about 29° and 33° 2θ (as indicated by the black arrows and dashed lines), only if the material was synthesized with temperatures higher than 500 °C.

compositions and 3 distinct temperatures) within milliseconds. In addition to the outstanding speed of the automated analysis, the model proved to be sensitive to those additional reflections, even if the extra peak positions aligned almost perfectly with the diffraction patterns of the doped copper oxides. The additional peaks display as shoulders of the diffraction peaks, and the neural network was able to identify those nonsymmetric peaks for the accurate detection of multi-phase compounds. Notably, our presented method does not require detailed information regarding occurring impurities or the possibility of overlapping peaks and is even applicable to further dopants without the need for retraining, highlighting the versatility and adaptability of our approach.

The fast-screening experiment and analysis of the XRD patterns revealed interesting properties of the copper oxide with respect to incorporating dopants, especially while considering the temperature during the synthesis process. Figure 3b shows the diffraction patterns for the sample with 5% Ni dopant (in shades of brown), together with the signal obtained for the pure CuO specimen (green). While the XRD pattern of the sample synthesized at 500 °C concurs with the signal of the phase-pure material, the patterns of the samples synthesized with higher temperatures exhibit minor, additional diffraction peaks at 29° and 33° 2θ. Despite the relatively indistinguishable peaks submerged within the noise, the trained network demonstrated proficiency in accurately classifying these patterns. This simultaneously confirms both results from Al-Amri et al. and Meneses et al. which observed the presence and absence of additional phases for Ni-doped copper oxides that have been produced at similar temperatures.^[21,23]

Accordingly, **Figure 4** shows the classification of our network for the XRD patterns with respect to synthesis temperature and dopant. The colors correspond to the output of the network, which ranges between 0 (single-phase, blue) and 1 (multi-phase, red), with the impurity classification threshold at 0.5 (white). For Cu_{1-x}(Zn,Ni, or Mn)_xO and 500 °C, about 7% dopant can be incorporated into the copper oxide, while still forming a phase-pure material (blue region). The output of the network correlates with the significance of the additional peaks, so there's an initial dopant concentration region with only minor additional peaks (white, multi-phase threshold), before the impurity phases are distinctly detectable (red). The lighter shades of blue correspond to single-phase predictions with elevated multi-phase probabilities, that we identified as patterns with higher noise levels or minor irregularities of the baseline intensities. For higher temperatures, the detected multi-phase threshold decreased for all three dopants, so, presumably, lower concentrations can be incorporated without forming multi-phase compounds.

To verify the predictions of the neural network, some XRD scans have been analyzed manually. Using the Rietveld refinement method, the weight percentages of the primary and impurity phases were determined to identify those samples that contain multi-phase structures. Instead of performing the refinement for all scans, the model's prediction allowed for the selection of a subset of the patterns. Therefore, only the samples calcined at 500 and 700 °C have been evaluated, as it was determined that multi-phase thresholds in the 600 and 700 °C test series exhibited substantial similarities. Moreover, according to the prediction of the model, only the dopant concentration

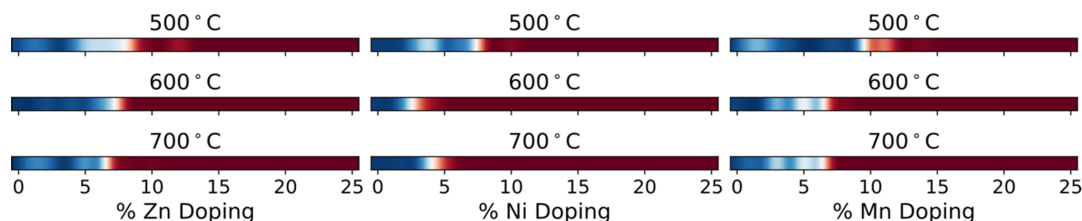


Figure 4. The predictions of our neural network separating the XRD patterns of dopants for CuO into single-phase samples (blue) and multi-phase compounds (red) are here visualized for all temperatures. The intensity of the color corresponds to the confidence of the neural network. For 500 °C, about 7% dopant can be incorporated and the thresholds shift for higher temperatures.

range of 1–10% held crucial significance for the formation of multi-phase structures (1–8% for the 700 °C samples). This model-driven insight notably reduced the number of samples necessitating manual analysis, streamlining our focus to the most pertinent data subsets. Manual analysis showed agreement with the results predicted by the model. Detailed information for the Rietveld refinement can be found in the Supporting Information.

An unexpected observation of our tests is that the threshold for dopant concentrations that yield multi-phase compounds declines with increased synthesis temperatures. Concurrently, samples synthesized at these higher temperatures display narrower peak shapes attributable to larger crystallite sizes. While narrow peaks stand out from the noise, broader diffraction peaks can merge indistinguishably with noise and background. This indicates that at lower synthesis temperatures, it is not that higher dopant concentrations were incorporated, but rather the resulting impurity phases became undetectable due to the broad peak shapes. However, neither the neural network model nor the manual Rietveld refinement identified suspected impurity phases at lower dopant concentrations in the 500 °C samples. Such a limitation underscores the requirement for extended acquisition times, which enhance signal-to-noise ratios and could consequently facilitate the detection of impurity phases.

3. Conclusion

To facilitate material discovery experiments, we present a method for the automatic analysis of XRD patterns in fast screening experiments. The XRD technique provides information about the crystalline structure of the analyzed sample and allows the distinction between single-phase and multi-phase structures. Single-phase materials are of particular interest because they possess uniform properties and behavior, which can be critical for certain applications. The neural network we developed automatically separates the produced samples into three categories: non-diffracting/amorphous, single-phase, and multi-phase.

We demonstrate the fitness of our approach on two distinct experimental series: spinels ($Fd-3m$) and doped copper oxides ($C2/c$). Using our unified data generation approach and a cif-file of the desired structure, models were trained for automated analysis of the XRD scans. The accuracy of the predicted classifications was validated manually through Rietveld refinement and visual examination of XRD patterns. While a quantitative Rietveld refinement analysis necessitates the identification of precise phases to ascertain weight percentages and detect impurities, our method operates at a more general level, bypassing the need to explicitly define impurity phases. Consequently, the speed of materials discovery experiments can be significantly enhanced using our universal approach. This method swiftly filters out unsuitable materials, ensuring that only prospective materials advance to subsequent stages of analysis or are considered for future experimental series.

Moreover, our methodology lessens the burden of manual analysis in expedited screening experiments. Given that the model's output aligns with the significance of additional reflections, experts can cherry-pick samples with high multi-phase probability estimates, which exhibit distinct diffraction peaks, thereby

facilitating the phase identification process. Alternatively, manual examination of dopant concentration thresholds can be strategically limited to samples near the predicted multi-phase detection boundary, rather than analyzing the entire spectrum. Therefore, our method not only paves the way for full automation of the analysis process but can also effectively complement human expertise and promotes a synergistic relationship between AI and human experts for more nuanced and efficient investigations.

4. Experimental Section

Generation of Training Data: To train neural networks for identifying prospective material samples, it is essential to have reference data. As the materials synthesized in our experiments were novel, experimental data for these materials did not exist and, therefore, must be simulated. Crystalline materials are characterized by their structure, as described by the lattice, and the atoms that constitute the crystal. Databases such as the ICSD or the COD store this information and provide it in the form of text files or database entries,^[5,6] which are parsed from the database in commercial software for the analysis of the experimental data. One example for such text files is the crystallographic information file (cif) format, which contains information about the crystals, including lattice parameters, space group, and coordinates for each atom in the unit cell. In the materials discovery experiments described here, the reference material and its structure are known, which serves as a starting point for generating training data.

A variety of software packages and libraries are available for handling crystallographic information, including parsers for cifs. We chose to build on the well-established Python library *pymatgen* for generating synthetic data.^[25] While it is possible to accurately describe the resulting structures of the synthesis (regardless of stability) given exact information about experimental parameters like doping material and size of substituting atoms, we decided to take a more general approach. When foreign atoms or ions substitute positions within the structure, or when they are incorporated, the lattice is influenced by factors like the atom size of elements present in the precursors. These factors can either compress or extend the lattice, thereby impacting its overall structure. Thus, we introduced random variation (up to 1%) to the lattice parameters, while maintaining restrictions defined by the crystal system of the reference structure.

In addition to parsing crystallographic information, the *pymatgen* package also provides tools to simulate X-ray powder diffraction patterns using the *XRDCalculator* object. This tool is designed to calculate the positions and intensities of diffraction peaks, using the specified structure and wavelength as input. While the variation in lattice dimensions accounts for the shift in peak positions, it is equally essential to represent the changes in relative intensities that arise due to differences in form factors introduced by foreign species. Given the uncertainty in the variance range of form factors, we chose to model peak intensity variations with a separate effect, preferred orientation, which occurs when certain particle orientations are overrepresented, thus altering the XRD pattern's relative intensities. Accordingly, preferred orientation is introduced to the training set to account for these variations in relative intensities.

Finally, the width and shape of the diffraction peaks in the recorded signal depend on the sizes of the crystals in the powder sample. The relation between crystallite size and full width at half maximum of the peaks is described by the Scherrer equation,^[26] so we generated synthetic powder diffraction patterns with varied peak shapes related to grain sizes between 10 and 100 nm. In consideration of the diverse instrumental broadening effects arising from the use of different equipment, our data simulation pipeline used a pseudo-Voigt diffraction peak profile to encapsulate the distinct optical characteristics inherent to each instrument, thereby accounting for the diverse appearances of peaks observed in the acquired signals. Additionally, Gaussian and Poisson noise were added to the simulated patterns to accurately represent the variation of the measured

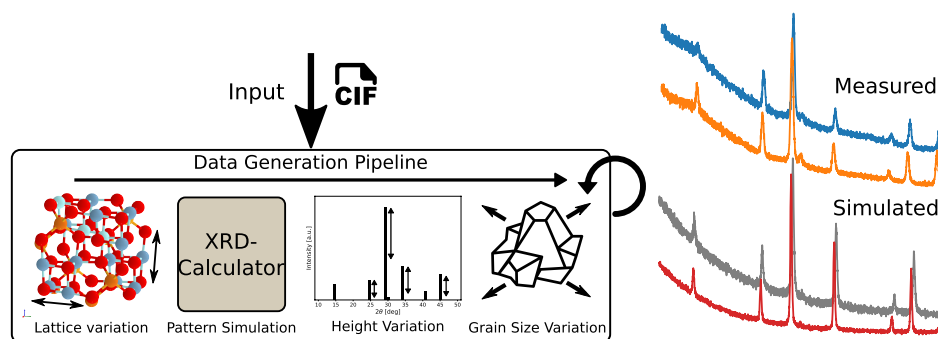


Figure 5. Pattern simulation approach. Based on a provided cif (here: MgAl_2O_4 , ICSD Code 13 859), artificial patterns are generated that depict the typical variation that occurs in experiments that test the doping of base materials. To account for variations and experimental artifacts, the lattice, the simulated peak heights, and the crystallite sizes are varied, and a baseline intensity and noise are added.

signals. Moreover, the baseline of the measured XRD patterns was simulated using Chebyshev polynomials, as is common practice to replicate XRD data.^[12]

Accordingly, **Figure 5** provides an overview of our simulation approach. The presented data generation pipeline takes a cif-file as input (either from a database or a description of an arbitrary structure) and generates multiple variants by varying the lattice parameters, texture, and crystallite sizes. Additionally, artificial noise and a baseline intensity were added to account for experimental artifacts. By comparison of the simulated and measured patterns, it is shown that the simulation approach depicts the realistic variation that occurs in such fast-screening experiments.

In addition to generating single-phase XRD pattern training data, the automatic discrimination system must be capable of handling multi-phase and amorphous XRD patterns. Generating amorphous XRD patterns is straightforward; instead of adding a baseline intensity to the simulated pattern, the background function alone can act as an example of an amorphous structure. Multi-phase structures, on the other hand, are based on single-phase patterns that are complemented with additional, random peaks. We added a few diffraction peaks at random positions to generate multi-phase examples while ensuring that those positions do not overlap completely with the peak positions of the single-phase pattern. As a result, our simulated dataset represents the three classes to identify: amorphous phases, single-phase patterns, and multi-phase patterns.

Network Architecture: Several neural network architectures were proposed for the analysis of XRD data, with many utilizing a convolutional neural network (CNN) structure.^[9–14] These CNNs use convolutional layers to apply a kernel that slides linearly across the input, identifying position-independent features, such as diffraction peaks, that stand out from background noise. By doing so, the CNNs can suppress baseline intensity and noise while matching varying shapes of diffraction peaks. Pooling operations often follow the convolutional layers to reduce input dimensionality, thereby minimizing peak position variations.

However, we conducted a recent study that revealed the lack of sensitivity with respect to identifying minor peaks in patterns for established network structures.^[20] The detection of multi-phase peaks is of great importance for this work, necessitating modifications to the network architecture to improve performance in minor peak identification. Although single-phase structure peaks regularly occur in the training data, multi-phase peaks are inserted at random positions, making them outliers from the expected results. To identify minor outliers, a common strategy involves scaling data points according to mean and standard deviation, resulting in an amplification of irregular peak intensities. Nevertheless, scaling cannot be applied to the raw input, which includes noise, background, and peak position shifts.

Thus, we present a modified network which is illustrated in **Figure 6**. This network takes Min-Max scaled XRD patterns as input, so the first layer

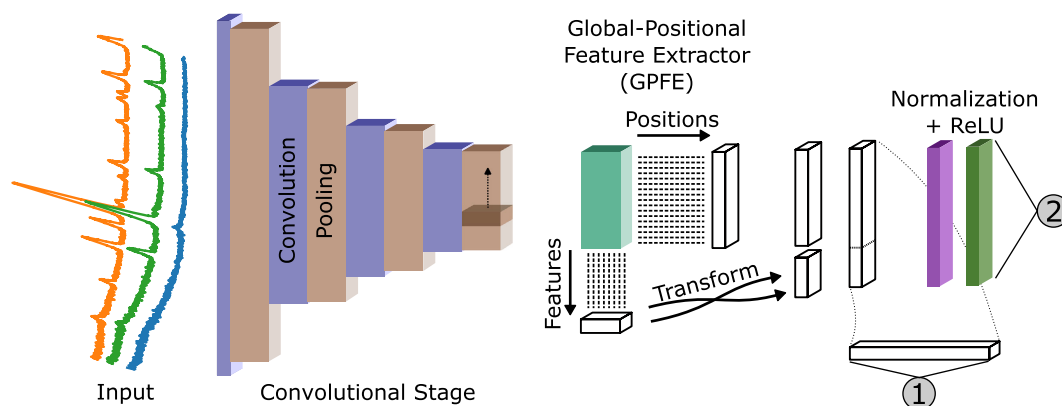


Figure 6. Network architecture for the model presented in this work. The normalized XRD patterns are fed into the network and multiple pairs of convolutional layers and max-pooling operations condense the input to fewer data points. Subsequently, the relevant features are compressed in the global-positional feature extractor (GPFE), which simultaneously identifies peaks with respect to the positions in the signal and globally relevant features, such as exceptional peak shapes. By concatenating the two separate types of features, the network conditions the information for the following classification. The first output (1) classifies nondiffracting versus crystallite structures based on the activations of the extracted features. To amplify the anomalies of multi-phase samples, a normalization and subsequent ReLU layer are used for rescaling of the intensities before the second output (2), that distinguishes between single-phase and multi-phase signals.

of the network matches the dimensionality of the signals. For instance, XRD signals collected from our robotic platform were measured from 10° to 60° 2θ with a step width of 0.015, resulting in 3334 data points. The input then passes through the convolutional stage, which contains multiple convolutional layers and pooling operations to identify peaks and reduce the dimensionality of the input. The exact configuration of weights in the respective layers depends on the properties of the data. Here, we used a kernel size of 17 in 4 convolutional layers and 32 filters to identify the relevant features, but different parameters could be necessary for other instrument configurations (e.g., larger kernels for patterns with smaller step sizes).

Following the convolutional stage of the network is our custom global-positional feature extractor (GPFE) that combines positional features and unique textures that appear globally in the patterns. The position of the peaks is crucial for identifying crystalline structures; therefore, it's essential to maintain the integrity of positional information to differentiate between various patterns. Additionally, the detection of exceptional features, such as the almost-overlapping peaks of the $Zn_xCu_{1-x}O$ structures, necessitates additional paths in the network that are not related to the positions. Hence, our GPFE extracts both types of features simultaneously and combines the diverse information for the following layers. Utilizing GlobalMaxPooling layers, we identified the largest activation both across the channel dimension (thereby preserving positional information) and within each channel (thus pinpointing unique features). The resultant information was then condensed, serving as a compressed input for the subsequent layers. A more detailed explanation of the GPFE's functionality can be found in the Supporting Information.

The network splits the three-class categorization task into two separate outputs. First, the model distinguished between nondiffracting (amorphous, empty sample holders) and crystalline structures depending on the extracted activations of the positional and global features. For signals without relevant reflections, the feature maps should be mostly zero, as noise and the baseline intensity were filtered from the inputs. Therefore, the model identifies patterns that match the defined reference structure based on the positions of the extracted diffraction peaks.

Multi-phase samples, on the other hand, exhibit XRD signals with additional reflections, which are an anomaly from the typical phase-pure pattern. Thus, the normalization layer that scales the respective features according to the mean and standard deviation is placed right before the second output that classifies single-phase and multi-phase samples. By amplifying the exceptional activations, the network facilitates the detection of anomalies, hence, crucially improving the accuracies of analyzing XRD patterns in fast-screening experiments. Additionally, a rectified linear unit (ReLU) is used to clip intensities below the learned means that are not relevant for the identification of additional reflections, which stabilizes the training process.

While our neural network architecture presents a significant deviation from established structures, the adaptations were necessary for robust classification of those subtle multi-phase peaks. In Table S3, Supporting Information, we compared our developed network to a similar network presented by Szymanski et al. for the identification of XRD patterns.^[13,19] Our model performed better on both presented datasets. While the performance of their model nearly matches ours on the spinels dataset, the network by Szymanski and colleagues fails to successfully extract the almost overlapping peaks that appear in the doped copper oxides dataset, resulting in considerably worse performance metrics. Even for the spinel dataset, the reference model mostly detected the patterns with clear multi-phase peaks, while failing to correctly classify those signals with only minor impurity reflections. This highlights the application of our GPFE and the subsequent normalization, which allows for the detection of unique textures and subtle peaks.

Details: Spinel-type oxide synthesis: Water-based nitrate salt precursor solutions (0.2 mol L^{-1} in distilled water) of $\text{Cu}(\text{NO}_3)_2 \cdot 2.5\text{H}_2\text{O}$ (Sigma-Aldrich, 98%), $\text{Cr}(\text{NO}_3)_3 \cdot 9\text{H}_2\text{O}$ (Sigma-Aldrich, 99.99%), $\text{Fe}(\text{NO}_3)_3 \cdot 9\text{H}_2\text{O}$ (Sigma-Aldrich, 98%), $\text{Mg}(\text{NO}_3)_2 \cdot 6\text{H}_2\text{O}$ (Sigma-Aldrich, 99.99%), $\text{Mn}(\text{NO}_3)_2 \cdot 4\text{H}_2\text{O}$ (Sigma-Aldrich, 98%), and $\text{Zn}(\text{NO}_3)_2 \cdot 6\text{H}_2\text{O}$ (Sigma-Aldrich, 98%) were used for the synthesis of the respective spinel oxide compounds.

Cu doping study: Water-based nitrate salt precursor solutions (0.2 mol L^{-1} in distilled water) of $\text{Cu}(\text{NO}_3)_2 \cdot 2.5\text{H}_2\text{O}$ (Sigma-Aldrich, 98%), $\text{Zn}(\text{NO}_3)_2 \cdot 6\text{H}_2\text{O}$ (Sigma-Aldrich, 98%), $\text{Ni}(\text{NO}_3)_2 \cdot 6\text{H}_2\text{O}$ (Sigma-Aldrich, 99.99%), and $\text{Mn}(\text{NO}_3)_2 \cdot 6\text{H}_2\text{O}$ (Sigma-Aldrich, 99.99%) were used for the respective Cu doping study.

For both studies, the nitrate salt solutions were mixed in different combinations in a standard 360 μL 96-well plate using an automated pipetting robot (Opentrons OT-2). To initiate coprecipitation, the respective precursor solutions were mixed with ammonia (Sigma Aldrich, 28–30%) at a ratio of 1:2 on a carrier substrate (two-sided polished (100) Si wafers) suitable for calcination and further XRD analysis. The spinel-type oxides were calcined at 700°C for 5 h in air and the Cu doping study materials were calcined at 500, 600, and 700°C in air for 3 h. For both studies, a constant heating rate of 300°C h^{-1} and naturally cooling down to room temperature inside the oven were used.

Automated X-ray Diffraction (XRD): Automated XRD measurements were performed at a STOE Stadi P diffractometer, equipped with a Ga-jet X-ray source (Ga-K β radiation, 1.2079Å) and a custom-built XY stage for automated sample measurement. XRD patterns were obtained in transmission mode. Patterns were collected between 10° and 60° 2θ with a step size of 0.015° . The powder samples on the (100) Si wafer were fixed with Kapton film, and the Si wafer was held by an in-house designed holder.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

deep learning, fast screening, high throughputs, X-ray diffraction

Received: August 22, 2023

Revised: December 8, 2023

Published online:

- [1] J. Epp, in *Materials Characterization using Nondestructive Evaluation (NDE) Methods*, Woodhead Publishing, Sawston, UK **2016**, pp. 81–124.
- [2] V. Pecharsky, P. Zavaliy, *Fundamentals of Powder Diffraction and Structural Characterization of Materials*, Springer, New York, NY **2005**.
- [3] X.-D. Xiang, X. Sun, G. Briceno, Y. Lou, K.-A. Wang, H. Chang, W. G. Wallace-Freedman, S.-W. Chen, P. G. Schultz, *Science* **1995**, 268, 1738.
- [4] L. Zhang, Y. Wang, J. Lv, Y. Ma, *Nat. Rev. Mater.* **2017**, 2, 17005.
- [5] A. Belsky, M. Hellenbrandt, V. L. Karen, P. Luksch, *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, 58, 364.
- [6] S. Gražulis, D. Chateigner, R. T. Downs, A. F. T. Yokochi, M. Quirós, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck, A. Le Bail, *J. Appl. Crystallogr.* **2009**, 42, 726.

- [7] L. Velasco, J. S. Castillo, M. V. Kante, J. J. Olaya, P. Friederich, H. Hahn, *Adv. Mater.* **2021**, *33*, 2102301.
- [8] J. Chen, S. R. Cross, L. J. Miara, J.-J. Cho, Y. Wang, W. Sun, (Preprint) arXiv: 2304.00743, Submitted: April **2023**.
- [9] W. B. Park, J. Chung, J. Jung, K. Sohn, S. P. Singh, M. Pyo, N. Shin, K. Sohn, *IUCrj* **2017**, *4*, 486.
- [10] H. Wang, Y. Xie, D. Li, H. Deng, Y. Zhao, M. Xin, J. Lin, *J. Chem. Inf. Model.* **2020**, *60*, 2004.
- [11] J. Lee, W. B. Park, J. H. Lee, S. P. Singh, K. Sohn, *Nat. Commun.* **2020**, *11*, 86.
- [12] J. Schuetzke, A. Benedix, R. Mikut, M. Reischl, *IUCrj* **2021**, *8*, 408.
- [13] N. J. Szymanski, C. J. Bartel, Y. Zeng, Q. Tu, G. Ceder, *Chem. Mater.* **2021**, *33*, 4204.
- [14] H. Dong, K. T. Butler, D. Matras, S. W. Price, Y. Odarchenko, R. Khatry, A. Thompson, V. Middelkoop, S. D. Jacques, A. M. Beale, A. Vamvakeros, *npj Comput. Mater.* **2021**, *7*, 74.
- [15] F. Massuyeau, T. Broux, F. Coulet, A. Demessence, A. Mesbah, R. Gautier, *Adv. Mater.* **2022**, *34*, 2203879.
- [16] C. Liu, K. Kitahara, A. Ishikawa, T. Hiroto, A. Singh, E. Fujita, Y. Katsura, Y. Inada, R. Tamura, K. Kimura, R. Yoshida, *Phys. Rev. Mater.* **2023**, *7*, 093805.
- [17] N. Q. Le, M. Pekala, A. New, E. B. Cienger, C. Chung, T. J. Montalbano, E. A. Pogue, J. Domenico, C. D. Stiles, *J. Phys. Chem. C* **2023**, *127*, 21758.
- [18] H. Uryu, T. Yamada, K. Kitahara, A. Singh, Y. Iwasaki, K. Kimura, K. Hiroki, N. Miyao, A. Ishikawa, R. Tamura, S. Ohhashi, C. Liu, R. Yoshida, *Adv. Sci.* **2023**, 2304546, <https://doi.org/10.1002/adv.202304546>.
- [19] N. J. Szymanski, P. Nevatia, C. J. Bartel, Y. Zeng, G. Ceder, *Nat. Commun.* **2023**, *14*, 6956.
- [20] J. Schuetzke, N. J. Szymanski, M. Reischl, *npj Comput. Mater.* **2023**, *9*, 100.
- [21] S. Al-Amri, M. Shahnawaze Ansari, S. Rafique, M. Aldahri, S. Rahimuddin, A. Azam, A. Memic, *Curr. Nanosci.* **2015**, *11*, 191.
- [22] I. Khan, S. Khan, H. Ahmed, R. Nongjai, *AIP Conf. Proc.* **2013**, *1536*, 241.
- [23] Y. Lv, L. Li, P. Yin, T. Lei, *Dalton Trans.* **2020**, *49*, 4699.
- [24] C. T. Meneses, J. G. Duque, L. G. Vivas, M. Knobel, *J. Non-Cryst. Solids* **2008**, *354*, 4830.
- [25] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, G. Ceder, *Comput. Mater. Sci.* **2013**, *68*, 314.
- [26] P. Scherrer, in *Kolloidchemie Ein Lehrbuch*, Springer Berlin, Heidelberg, Germany **1920**, pp. 387–388.
- [27] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, *APL Mater.* **2013**, *1*, 011002.