CHEMISTRY OF MATERIALS

pubs.acs.org/cm

Article

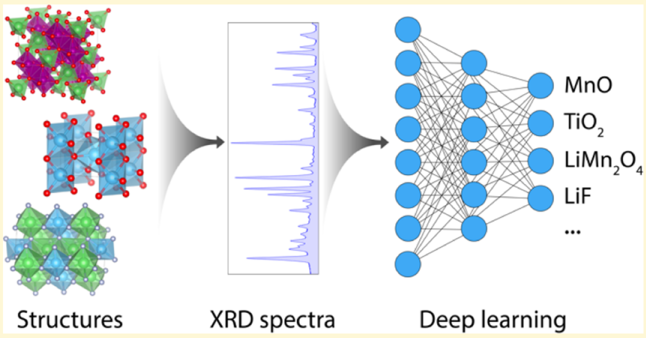
Probabilistic Deep Learning Approach to Automate the Interpretation of Multi-phase Diffraction Spectra

Nathan J. Szymanski, Christopher J. Bartel, Yan Zeng, Qingsong Tu, and Gerbrand Ceder*

Cite This: *Chem. Mater.* 2021, 33, 4204−4215

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Autonomous synthesis and characterization of inorganic materials require the automatic and accurate analysis of X-ray diffraction spectra. For this task, we designed a probabilistic deep learning algorithm to identify complex multi-phase mixtures. At the core of this algorithm lies an ensemble convolutional neural network trained on simulated diffraction spectra, which are systematically augmented with physics-informed perturbations to account for artifacts that can arise during experimental sample preparation and synthesis. Larger perturbations associated with off-stoichiometry are also captured by supplementing the training set with hypothetical solid solutions. Spectra containing mixtures of materials are analyzed with a newly developed branching algorithm that utilizes the probabilistic nature of the neural network to explore suspected mixtures and identify the set of phases that maximize confidence in the prediction. Our model is benchmarked on simulated and experimentally measured diffraction spectra, showing exceptional performance with accuracies exceeding those given by previously reported methods based on profile matching and deep learning. We envision that the algorithm presented here may be integrated in experimental workflows to facilitate the high-throughput and autonomous discovery of inorganic materials.

■ INTRODUCTION

The development of high-throughput and automated experimentation has ignited rapid growth in the amount of data available for materials science and chemistry.[1,2] Unlocking the physical implications of resulting datasets, however, requires detailed analyses that are traditionally conducted by human experts. In the synthesis of inorganic materials, this often entails the manual interpretation of X-ray diffraction (XRD) spectra to identify the phases present in each sample. Past attempts to automate this procedure using peak indexing[3,4] and full profile matching[5,6] algorithms have been limited by modest accuracy, in large part because measured spectra usually deviate from their ideal reference patterns (e.g., due to defects or impurities). Consequently, the analysis of XRD spectra widely remains a manual task, impeding rapid materials discovery and design. To alleviate this bottleneck, deep learning based on convolutional neural networks (CNNs) has recently emerged as a potential tool for automating the interpretation of diffraction spectra with improved speed and accuracy.[7,8]

Previous work has demonstrated that CNNs can be used to perform symmetry classification[9−12] and phase identification[13,14] from XRD spectra of single-phase samples. Given the lack of well-curated diffraction data obtained experimentally, training is most commonly performed on labeled sets of simulated spectra derived from known crystalline materials, for example, in the Inorganic Crystal Structure Database (ICSD).[15] However, because many factors can cause differences between observed and simulated diffraction peaks, this approach can be problematic for extension to experimentally measured XRD spectra. Vecsei et al. demonstrated that a neural network trained on simulated spectra produced an accuracy of only 54% for the classification of experimentally measured diffraction spectra extracted from the RRUFF database.[10] To overcome this limitation, simulated spectra can be augmented with perturbations designed to emulate possible artifacts.[12] For example, Oviedo et al. trained a CNN using simulated spectra augmented with random changes in their peak positions and intensities, which were chosen to account for texture and epitaxial strain in the thin films being studied. The resulting model correctly classified the space group for 84% of diffraction spectra measured from 115 metal halide samples.[7] Based on past work, we propose that generalization of deep learning to handle complex XRD spectra requires a more complete data augmentation procedure

Received: March 27, 2021
Revised: May 10, 2021
Published: May 26, 2021

https://doi.org/10.1021/acs.chemmater.1c01071

that properly accounts for all the artifacts and complexities that frequently arise during sample preparation and synthesis.

To extend the application of CNNs to mixtures of materials, Lee et al. constructed a training set of multi-phase spectra that were simulated using linear combinations of single-phase diffraction spectra from 38 phases in the quaternary Sr−Li−Al−O space.[8] Their model performed well in the identification of high-purity samples, with 98% of all phases correctly labeled based on 100 three-phase spectra. However, the combinatorial nature of their technique requires an exceptionally high number of training samples (nearly two million spectra from 38 phases), which restricts the inclusion of experimental artifacts via data augmentation. Moreover, because the number of training samples increases exponentially with the number of reference phases, the breadth of the composition space that can be efficiently considered is limited. Proposing an alternative approach, Maffettone et al. designed an ensemble model trained on simulated single-phase spectra to yield a probability distribution of suspected phases for a given spectrum.[13] From this distribution, the authors infer that high probabilities suggest that the corresponding phases are present in the mixture. While this method avoids combinatorial explosion and thus allows many experimental artifacts to be included during training, it sometimes leads to confusion as obtaining comparable probabilities for two phases does not necessarily imply that both are present. Rather, it may simply mean that the algorithm has difficulty distinguishing between the two phases. An improved treatment of multi-phase spectra therefore necessitates an approach that (i) allows artifacts to be incorporated across many phases and (ii) distinguishes between probabilities associated with mixtures of phases as opposed to similarities between single-phase reference spectra.

In this work, we introduce a novel deep learning technique to automate the identification of inorganic materials from XRD spectra of single- and multi-phase samples. In our approach, training spectra are generated with physics-informed data augmentation, whereby experimental artifacts (strain, texture, and domain size) are used to perturb diffraction peaks. The training set is built not only from experimentally reported stoichiometric phases but also from hypothetical solid solutions that account for potential off-stoichiometries. An ensemble CNN is trained to yield a distribution of probabilities associated with suspected phases, which is shown to be a surrogate for prediction confidence. We extend this probabilistic model to the analysis of multi-phase mixtures by developing an intelligent branching algorithm that iterates between phase identification and profile subtraction to maximize the probability over all phases in the predicted mixture. As a representative example to assess the efficacy of our approach, we trained and tested a model on diffraction spectra derived from materials in the broad Li−Mn−Ti−O−F composition space given their structural diversity and technological relevance (e.g., for Mn-based battery cathodes).[16] By also systematically testing on a dataset of experimentally measured XRD spectra designed to sample complexities that often arise during synthesis, we show that our model achieves considerably higher accuracy than state-of-the-art profile matching techniques and previously developed deep learning-based methods. The improved performance demonstrated here should be generalizable to any alternative chemical space (beyond Li−Mn−Ti−O−F) through application of the same data augmentation and training procedures to any given set of phases from the space of interest.

■ METHODS

Stoichiometric Reference Phases. The identification of inorganic materials from their XRD spectra relies on the availability of suitable reference phases that can be compared to samples of interest. In this work, we focus on the Li−Mn−Ti−O−F chemical space (and subspaces) and retrieved all the 1216 corresponding entries from the ICSD.[15] For the identification of stoichiometric materials, we excluded 386 entries with partial occupancies from this set. To remove duplicate structures from the remaining 830 entries, all unique structural frameworks were identified using the pymatgen structure matcher.[17] For each set of duplicates, the entry measured most recently under conditions closest to ambient ones (20 °C and 1 atm) was retained. Based on these selection criteria, 140 unique stoichiometric materials listed in Table S1 were tabulated and used as reference phases. The code used to apply these selection criteria and create a set of unique reference phases from ICSD entries in any given composition space is available at https://github.com/njszym/XRD-AutoAnalyzer.

Non-stoichiometric Reference Phases. Although many solid solutions are available in the ICSD, they generally cover a narrow composition range while leaving others sparse. We therefore designed an algorithm to extend the space of non-stoichiometric reference phases using empirical rules to construct hypothetical solid solutions between the available stoichiometric materials. To determine which phases may be soluble with one another, all combinations of the 140 stoichiometric references phases in the Li−Mn−Ti−O−F space were enumerated, and two criteria were considered for each pair. First, solubility requires that the two phases adopt similar structural frameworks, which was verified using the pymatgen structure matcher.[17] Second, based on the Hume-Rothery rules,[18] the size mismatch between any ions being substituted with one another should be ≤15%. To estimate the ionic radii of all species comprising each phase, oxidation states were assigned using the composition-based oxidation state prediction tool in pymatgen.[17] In cases where mixed oxidation states are present (e.g., $Mn^{3+/4+}$), we chose to focus on the state(s) that minimizes the difference between the radii of the ions being substituted and therefore increases the likelihood for solubility. As will be shown by our test results, including more reference phases does not lead to a substantial decrease in accuracy; hence, it is preferable to overestimate solubility such that more structures are created as potential references.

Based on the 140 stoichiometric reference phases in the Li−Mn−Ti−O−F space, 43 pairs of phases were found to satisfy both solubility criteria described above. The phases in each pair were treated as end members, from which interpolation was used to generate a uniform grid of three intermediate solid solution compositions. For example, between spinel $LiMn_2O_4$ and $LiTi_2O_4$, intermediate compositions take the form $LiMn_{2-x}Ti_xO_4$ with $x \in \{0.5, 1.0, 1.5\}$. The lattice parameters of hypothetical solid solutions were linearly interpolated between those of the corresponding end members in accordance with Vegard's law.[19] Atomic positions and site occupancies were similarly obtained by interpolating between equivalent sites in the end members. This procedure gave a total of 129 hypothetical solid solution states from the 43 pairs of soluble phases. Excluding 14 duplicates resulted in 115 distinct solid solutions, as listed in Table S2. The code for generating hypothetical solid solutions for an arbitrary group of reference phases is available at https://github.com/njszym/XRD-AutoAnalyzer.

Data Augmentation. From the reference phases in the Li−Mn−Ti−O−F space, we built an augmented dataset of simulated XRD spectra with the goal of accurately representing experimentally measured diffraction data. Physics-informed data augmentation was applied to produce spectra that sample possible changes in peak positions, intensities, and widths. Shifts in peak positions (2θ) were derived by creating modified unit cells with up to ±4% strain in each lattice parameter. This was done by applying strain tensors to the lattice parameter matrix $(\vec{a}, \vec{b}, \vec{c})$ that preserve the space group of each structure. Internal cell coordinates were left unchanged, so that only peak positions were affected. Peak widths were broadened by

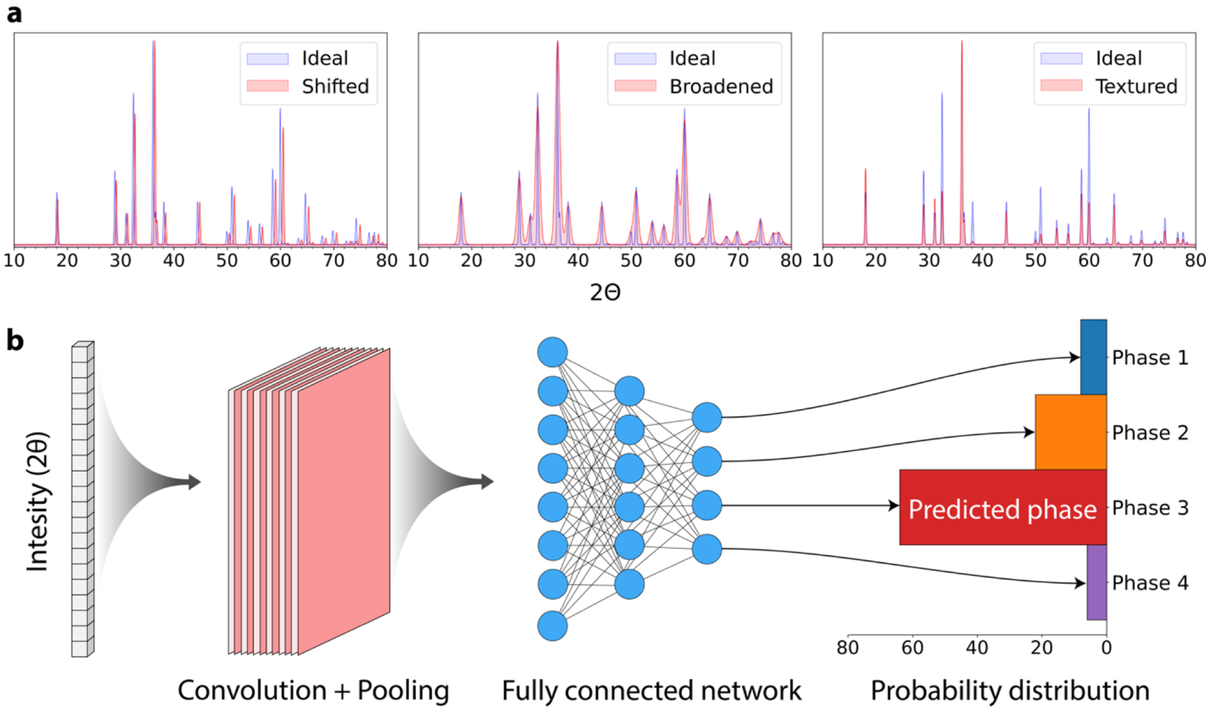


Figure 1. (a) Illustration of the data augmentation procedure designed to sample possible experimental artifacts including peak shift associated with cell strain, peak broadening related to small domain size, and peak intensity variation caused by texture. (b) Schematic of the deep learning pipeline used to map XRD spectra onto a probability distribution of suspected phases.

simulating domain sizes ranging from 1 nm (broad) to 100 nm (narrow) through the Scherrer equation.[20] Peak intensities were varied to mimic the preferred orientation along the preferred crystallographic planes (hereafter referred to as texture). This was done by performing scalar products between the peak indices and randomly selected Miller indices (*hkl*), followed by a normalization that scaled peak intensities by as much as ±50% of their initial values.

The bounds used for each artifact are chosen such that perturbations to simulated spectra are large enough to capture possible experimental complexities but not so large that they produce spectra that are unlikely to ever arise in experiment. Although it is difficult to rigorously define the range of artifacts that may occur, we used our prior experience and physics-based intuition to determine the extent of strain, texture, and domain size described in the previous paragraph. We note that larger variations may arise when substantial off-stoichiometry is present; however, this situation was treated separately by the addition of non-stoichiometric solid solutions as reference phases. In Figure 1a, we illustrate the effect of each of the three experimental artifacts on the XRD spectrum of spinel Mn_3O_4 as an example. Each artifact was applied separately to the simulated spectrum by taking 50 random samples from a normal distribution (e.g., between −4 and +4%), resulting in 150 augmented spectra per reference phase (50 samples for each of the three artifacts). Applying this procedure to all 255 references phases, including both experimentally reported stoichiometric materials and hypothetical solid solutions, resulted in 38,250 simulated diffraction spectra. Further details regarding data augmentation and spectrum simulation are provided in the Supplementary Note 1. The code to perform data augmentation for an arbitrary group of reference phases is available at https://github.com/njszym/XRD-AutoAnalyzer. Although all spectra used here were derived using Cu Kα radiation, any wavelength can be specified by the user. Therefore, our model can be applied to spectra measured using a variety of in-lab diffractometers or synchrotron light sources.

Convolutional Neural Network. The workflow used to classify a given XRD spectrum is displayed in Figure 1b. Similar to previous work,[8] diffraction spectra are treated as one-dimensional vectors that contain 4501 values for intensity as a function of 2θ. The range of 2θ is set from 10 to 80°, which is commonly used for scans with Cu Kα radiation (λ = 1.5406 Å). The intensities (represented as 4501 valued vectors) serve as input to a CNN that consists of six convolutional layers, six pooling layers, and three fully connected layers. Training was carried out with five-fold cross-validation using 80% of the simulated diffraction spectra, with the remaining 20% reserved for testing (i.e., excluded from training and validation). Details regarding the architecture of the CNN and the hyperparameters used during training are given in the Supplementary Note 2. The code used for training is also available at https://github.com/njszym/XRD-AutoAnalyzer. To classify spectra outside of the training set, an ensemble approach was used whereby 1000 individual predictions are made with 60% of connections between the fully connected layers randomly excluded (i.e., using dropout) during each iteration. The probability that a given phase represents the spectrum is then defined as the fraction of the 1000 iterations where it is predicted by the CNN. The resulting distribution may be treated as a ranking of suspected phases in the sample, with the corresponding probabilities providing measures of confidence.

Intelligent Branching Algorithm. Given that the CNN was trained only on single-phase XRD spectra, additional methods were developed to automate the identification of materials in multi-phase mixtures. In our workflow, we use an iterative procedure where phase identification is followed by profile fitting and subtraction. Once a phase is identified by the CNN, its diffraction peaks are simulated and fit to the spectrum in question using dynamic time warping (DTW), a well-known technique for correlating features in time series.[21] In contrast to Rietveld refinement, which is typically conducted manually using expert intuition regarding the structure and composition of each phase, DTW is readily automated, as it requires no physical input other than a user-specified window in which features can be correlated. For this work, we use a window of $\Delta 2\theta$ = 1.5° since larger peak shifts are typically not expected. After DTW has been applied to fit the simulated spectrum along 2θ, its diffraction peaks are scaled to minimize the average difference between the simulated and measured intensities. Using an average difference rather than focusing only on the largest peaks, we aim to avoid scaling errors caused by overlapping peaks between different phases. Following this scaling

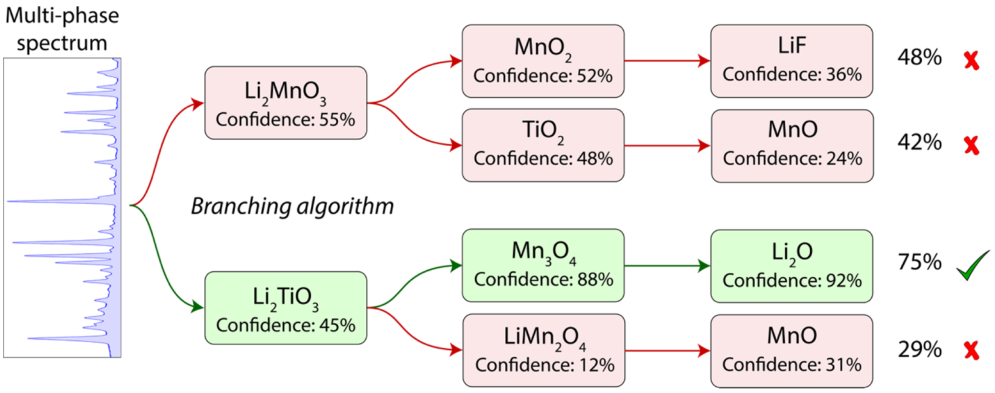


Figure 2. Schematic illustrating possible pathways enumerated by the branching algorithm for multi-phase identification. This method iteratively performs single-phase predictions followed by profile stripping, at each step tabulating the probability associated with each phase. This process is repeated until all intensities fall below 5% of the original maximum value. From all branches developed, the one with the highest average probability (highlighted green above) across all levels is chosen as the most likely set of phases present in the mixture.

process, the profile of the identified phase is subtracted to produce a modified spectrum that is representative of the mixture minus the phase that has already been identified. In other words, all known peaks are iteratively removed from the spectrum. This process is repeated until all significant peaks are attributed to a reference phase, that is, the cycle is halted once all intensities fall below 5% of the initially measured maximum intensity. Further details regarding the techniques used to perform profile fitting and subtraction are described in the Supplementary Note 3, and the corresponding code is available at https://github.com/njszym/XRD-AutoAnalyzer.

Following the iterative procedure outlined above, one could identify a multi-phase mixture using the collection of most probable phases given by the model at each step. However, because the spectrum is affected by all prior phases that have been identified, such a method over-prioritizes the first iteration of phase identification. In cases where the first phase predicted by the CNN is incorrect, the spectrum resulting from profile fitting and subtraction will contain diffraction peaks that do not accurately represent the remaining phases in the sample. All subsequent analyses will therefore be less likely to identify these phases. To improve upon this approach, we developed an intelligent branching algorithm that gives equal importance to each iteration of phase identification. In Figure 2, we illustrate how the algorithm evaluates several possible sets of phases to classify a diffraction spectrum derived from a mixture of Li_2TiO_3, Mn_3O_4, and Li_2O. At each step, the CNN generates a list of suspected phases along with their associated probabilities. As opposed to considering only the most probable phase at each iteration, the branching algorithm investigates all phases with non-trivial probabilities (≥10%). By following the spectrum associated with the subtraction of each suspected phase, a "tree" is constructed to describe all combinations of phases predicted by the model. Once each route has been fully exhausted, the branch with the highest average probability is chosen as the final set of predicted phases (e.g., the green phases highlighted in Figure 2). In this way, the algorithm maximizes the likelihood that predictions are representative of *all* phases contained in the actual mixture, as opposed to over-prioritizing the first iteration of phase identification. We found that this is an essential feature to predict multi-phase spectra correctly.

EXPERIMENTAL MEASUREMENTS

To further validate our model, we built an experimental dataset from a series of measurements designed to sample complexities that often arise during synthesis. A total of 10 materials, listed in the Supplementary Note 4 with details regarding the experimental procedures, were chosen to span a range of structures and compositions in the Li−Mn−Ti−O−F space. For a benchmark on pristine single-phase spectra with no intended artifacts, we conducted precise diffraction measurements on each of the 10 materials using carefully prepared, high-purity samples. The following modifications were then separately introduced such that each batch of samples contained one anticipated artifact: (i) samples were overlaid with Kapton tape during characterization to produce a diffuse background signal with a magnitude as large as 200% of the highest diffraction peak intensity; (ii) rapid scan rates (30°/minute) were used to generate noisy baseline signals with magnitudes reaching 5% of the maximum diffraction peak intensity; (iii) peak shifts as large as 0.4° were imposed by preparing thick pellets such that specimens were leveled slightly above the sample holder; and (iv) broad peaks with full widths at half maxima as large as 1.5° were obtained by ball milling. Several additional materials were also made to sample changes in the composition and site occupancy. Six samples of spinel $LiMnTiO_4$ were synthesized at temperatures of 900, 950, and 1000 °C followed by quenching or slow cooling based on previously reported procedures.[22] These samples were intended to contain differences in relative diffraction peak intensities owing to varied distributions of cation site occupancies. Non-stoichiometry was studied using four disordered rocksalt phases, each with a different composition made via solid-state synthesis. For the classification of multi-phase XRD spectra, 10 two- and three-phase mixtures (listed in the Supplementary Note 4) were prepared from combinations of materials in the Li−Mn−Ti−O−F space that were chosen to include spectra with a substantial amount of peak overlap. The mixtures contained equal weight fractions of all constituent phases. To isolate the effects of multiple phases, these measurements were conducted on samples for which no experimental artifacts were purposefully incorporated.

RESULTS

Identification of Stoichiometric Phases. As a first test case, we evaluated the performance of our model on simulated single-phase XRD spectra derived from the 140 stoichiometric reference phases in the Li−Mn−Ti−O−F space. Accordingly, the CNN was trained on 80% of the 21,000 generated spectra (140 materials × 150 augmentations) that were augmented to include physics-informed perturbations to their diffraction peak positions, widths, and intensities. The remaining 4200 spectra were reserved for testing. To assess the ability of the CNN to handle artifacts not considered during training, the test set was also supplemented with spectra having diffuse and noisy background signals. A diffuse background was simulated by adding an XRD spectrum measured from amorphous silica to the diffraction peaks of the stoichiometric materials. A total of 10 spectra were created for each phase (1400 spectra in total), with the maximum intensity produced by silica ranging from 100 to 300% of the maximum peak intensity of the reference phase. Another set of 1400 spectra were simulated by

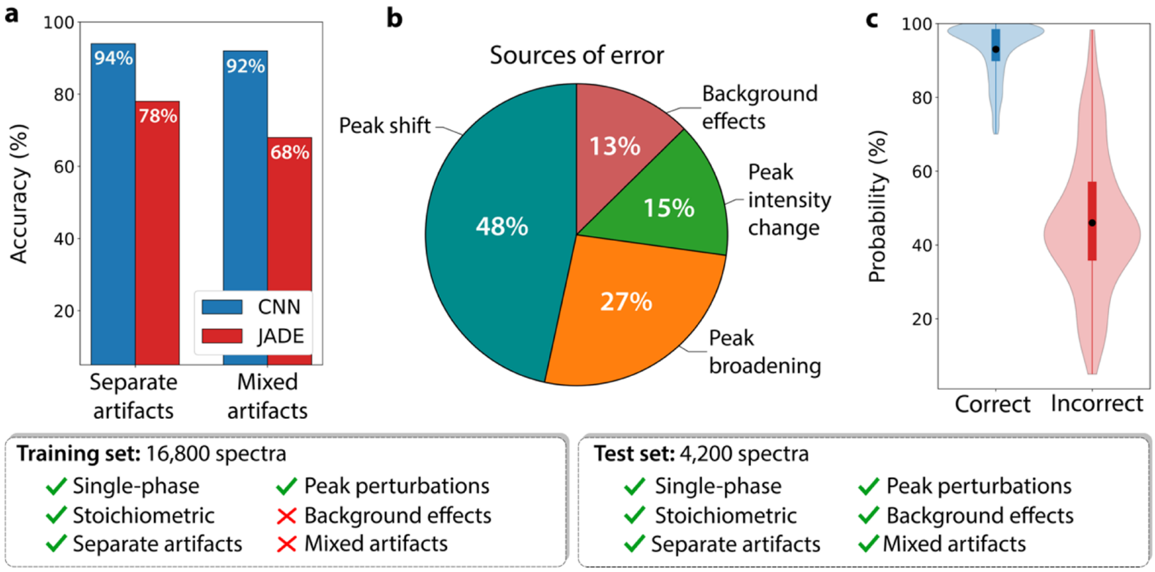


Figure 3. (a) Accuracies given by the CNN and JADE when applied to simulated spectra containing (i) individual artifacts applied separately and (ii) mixed artifacts applied altogether. (b) Sources of error in the CNN are illustrated by calculating the fraction of misclassifications that occur for spectra containing each separate artifact. (c) Distributions of probabilities given by the CNN when correct and incorrect classifications are made during testing on spectra containing mixed artifacts. Violin plots illustrate the density of probabilities, whereas embedded boxes extend from the lower to upper quartiles. Black dots are used to denote the average probability in each case.

adding Gaussian noise with magnitudes ranging from 1 to 5% of the maximum diffraction peak intensity. Before being passed to the CNN, these 2800 spectra were pre-processed using the baseline correction and noise filtering algorithms described in the Supplementary Note 5. This procedure is designed to replicate artifacts formed when imperfect corrections are made during pre-processing, which occasionally leads to the disappearance of minor peaks or leaves behind residual intensities related to amorphous impurities. Previous work has dealt with diffuse and noisy background signals by training on spectra with added baseline functions (e.g., polynomials).[9,13] However, because these functions are randomly selected rather than derived from possible impurities or defects, they are unlikely to accurately represent experimental measurements.[14] With this in mind, our current approach relies only on physics-informed data augmentation to improve the match between simulated and experimentally measured spectra.

The performance of our model is compared to that of a known standard, the JADE software package from MDI.[23] JADE is a widely used program that can automate phase identification with conventional profile matching techniques.[5] During testing, JADE was employed without any manual intervention to ensure a consistent comparison with the CNN, as we are assessing the capability of our approach to perform phase identification as part of an autonomous platform. We emphasize that our model is not designed to replace manual techniques such as Rietveld refinement but rather to provide more rapid and reliable predictions regarding phase identities. For this task, we applied both the trained CNN and JADE to the test set of simulated diffraction spectra that sample possible experimental artifacts *separately*, as discussed in the Methods. In Figure 3a, we compare the resulting accuracy of each method quantified as the fraction of phases correctly identified. Across the simulated test spectra, the CNN achieves a high accuracy of 94%. In contrast, JADE correctly identifies only 78% of phases when applied to the same set of spectra. To further verify the effectiveness of the CNN, an additional 1400 spectra were simulated with mixed artifacts such that each spectrum contains all aforementioned perturbations to its diffraction peaks (shifting, broadening, and texture) and a diffuse and noisy background signal. This incorporates an additional level of complexity not included in the training set, where each spectrum contained just one type of perturbation. When applied to the new test set with mixed artifacts, the accuracy of the CNN decreases only by 2% (from 94 to 92%), whereas the accuracy of JADE decreases by 10% (from 78 to 68%).

The tests show promising results for the CNN, although its performance is not without error. We look into the underlying causes of the occasional misclassifications that occur by dividing the simulated test spectra into four major categories: those augmented via the individual application of peak shifts, peak broadening, peak intensity change, and background effects (including diffuse and noisy baselines). The training set remains unchanged from the previous paragraph. In Figure 3b, we show the fraction of misclassifications that arise from each perturbation category. Of the 7000 total test spectra, 418 are misclassified by the CNN. The largest portion (48%) of misclassifications occurs for spectra containing peak shifts, which we attribute to the overlapping of diffraction peaks between similar phases. This most commonly occurs between isomorphic phases, and as a result, the CNN gives a higher accuracy for the identification of the structure (96%) as opposed to the composition (92%). We investigated the effects of increasing the bounds on strains that were used during training (beyond ±4%); however, a decrease in accuracy was observed as larger strains were incorporated. For example, training on spectra derived from structures with strain as large as ±6% led to a lower accuracy of 86% when applied to the test set containing spectra with as much as ±4% strain. More details regarding the effects of strain are illustrated in Figure S1. Relative to peak shifts caused by strain, spectra with broad peaks lead to fewer misclassifications, comprising 27% of

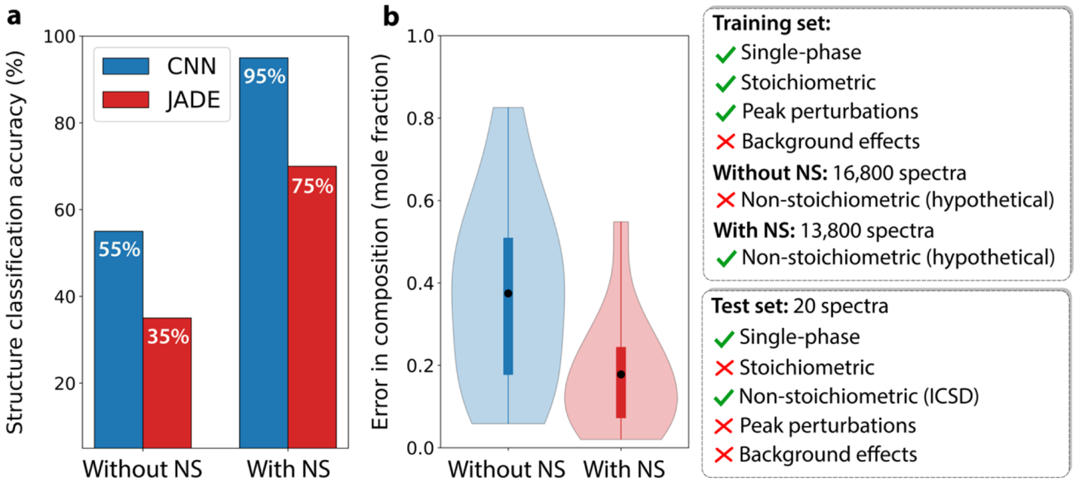


Figure 4. (a) For a set of diffraction spectra derived from 20 experimentally reported solid solutions, the fractions of structures correctly identified by the CNN and JADE are shown in two cases: (i) when the training set includes only stoichiometric reference phases (without NS, where NS denotes non-stoichiometry) and (ii) when the training set is augmented with hypothetical solid solutions (with NS). (b) For the same set of spectra, differences between true compositions and those predicted by the CNN are quantified by their mole fraction difference. Violin plots illustrate the full distribution of errors, whereas embedded boxes range from lower to upper quartiles. Black dots are used to denote the average probability given in each case.

errors. For this effect, misclassification occurs more frequently in low-symmetry structures, as they contain many diffraction peaks that tend to overlap with one another upon broadening. Of the 113 spectra that are incorrectly classified by the CNN due to peak broadening, 82 are from phases with monoclinic or triclinic symmetry. The remaining artifacts, including texture and background effects, show a relatively weak influence on the accuracy of the CNN. Because both of these artifacts cause changes in relative peak intensities, the distribution of misclassifications suggests that peak intensities have a more subtle role in the identification of stoichiometric single phases.

To assess the reliability of predictions made by our model, we examined the probability distributions given by the ensemble CNN. In Figure 3c, we compare the probabilities of correct and incorrect classifications made when the CNN is applied to the simulated spectra containing mixed artifacts. All correct classifications are accompanied by a probability greater than 70%, with an average of 93%, whereas incorrect classifications show a wide range of probabilities with a much lower average of 46%. This dichotomy suggests that probabilities are akin to confidence in the prediction and may be used as a reliable metric to gauge the likelihood that a classification is correct. If, for example, predictions are constrained to those with a probability above 70% (which comprise 84% of all spectra in the test set), then, the accuracy increases from 92 to 96%. On the other hand, when the probability is lower than 70%, we propose that the model should raise a "red flag," signifying that manual intervention is needed to clarify the identity of the underlying phase. Interestingly, even when an incorrect classification is made regarding the most probable phase, the correct phase is present within the top three suspected phases for 99% of all test spectra. Therefore, although manual intervention may occasionally be required to handle complex spectra, the problem is greatly simplified by allowing the user to choose from a small set of probable phases.

Incorporating Non-stoichiometry. To determine whether the accuracy of our model extends to non-stoichiometric materials, we built a test set of XRD spectra simulated from 20 experimentally reported solid solutions in the Li−Mn−Ti−O−F chemical space. These materials, listed in Table S3, were manually selected from the ICSD to ensure that their compositions are different (greater than 0.05 mol fraction) from those of the stoichiometric phases already considered in the previous section. To isolate the effects of non-stoichiometry, diffraction spectra were simulated without including any experimental artifacts. We first restricted the training set to include only diffraction spectra derived from stoichiometric materials to illustrate the necessity of including additional reference phases with non-stoichiometry (i.e., from hypothetical solid solutions). Similarly, JADE was applied to the new test set containing solid solutions while restricting its reference database to contain only stoichiometric phases. In doing so, neither method can be used to predict the exact compositions of the solid solutions. Instead, their prediction accuracy can be resolved into two components: (i) Is the predicted structure isomorphic to the true structure? (ii) How similar are the predicted and true compositions? Isomorphism was verified using the pymatgen structure matcher.[17] Differences in compositions were quantified using the mole fraction distance between the barycentric coordinates of each phase in the Li−Mn−Ti−O−F chemical space (i.e., with each constituent element representing a vertex). For example, the compositional difference between $LiMnO_2$ and $LiMn_{0.5}Ti_{0.5}O_2$ is quantified as 0.125 mol fraction since 0.5 out of four elements are interchanged in the formula unit.

In Figure 4a, we show the fraction of non-stoichiometric materials with structures correctly identified by the CNN and JADE when only stoichiometric reference spectra are used for training or profile matching. This case is labeled as "without NS" where NS denotes non-stoichiometry. The CNN correctly classifies the structures of 11/20 spectra, whereas JADE gives only 7/20 correct structural classifications. For the same set of spectra, we illustrate the differences between true compositions and those predicted by the CNN in Figure 4b. Errors in the predicted compositions range from 0.05 to 0.82 mol fraction, with an average value of 0.38. Therefore, when only stoichiometric reference phases are used, neither the deep

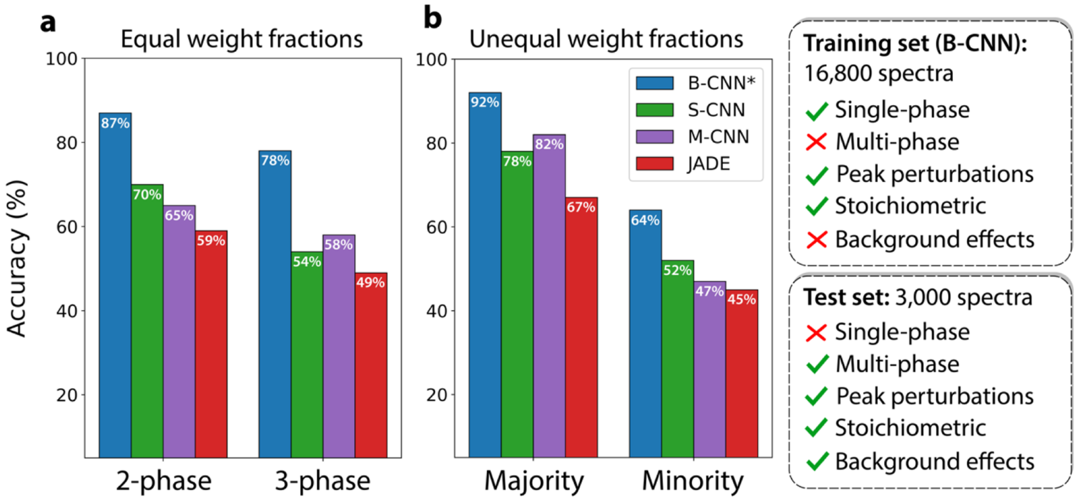


Figure 5. (a) Fractions of phases correctly identified by the B-CNN (*introduced in this work) when applied to simulated diffraction spectra of two- and three-phase mixtures with equally distributed weight fractions. For comparison, accuracies obtained using two methods based on previous work (S-CNN[13] and M-CNN[8]) are shown, in addition to results from JADE. (b) These same techniques are applied to diffraction spectra of two-phase mixtures with unequally distributed weight fractions of 10−30 and 70−90%. Accuracies are divided into the identification of majority and minority phases.

learning algorithm nor conventional profile matching techniques can be utilized to reliably predict the structure or composition of non-stoichiometric materials from their diffraction spectra. This conclusion supports our initial expectations given that substantial off-stoichiometry is known to cause large changes in the positions and intensities of diffraction peaks. Although data augmentation is useful (and necessary) to account for relatively weak deviations from ideality, it is not capable of extrapolating to larger changes well beyond those included in the training set.

A proper treatment of non-stoichiometry necessitates additional reference phases with compositions that more closely match experimentally observed solid solutions. To this end, we introduced XRD spectra simulated from hypothetical solid solutions spanning the Li−Mn−Ti−O−F space into the training set. In addition to the 21,000 spectra obtained from the 140 stoichiometric materials, 17,250 new spectra were derived from 115 hypothetical solid solutions (115 materials × 150 augmentations). Perturbations were applied via the data augmentation procedure described in the Methods, and 80% of the resulting diffraction spectra were used to re-train the CNN. For comparison, the same set of hypothetical solid solutions was also added to the reference database used by JADE. Both updated models were then applied to the test set containing 20 diffraction spectra simulated from the experimentally reported non-stoichiometric materials. The fraction of structures correctly identified by each method is displayed in Figure 4a, labeled as "with NS". In contrast to earlier results, the CNN and JADE achieve much higher accuracies of 95 and 70%, respectively. These improvements in performance are realized without sacrificing much accuracy in the classification of stoichiometric materials—our updated model correctly identifies 89% of phases across the test set containing simulated diffraction spectra with mixed artifacts, a decrease of only 3% compared to the CNN trained only on stoichiometric phases (Figure 3a). In Figure 4b, we present the updated distribution of errors in compositions given by the CNN trained with non-stoichiometric phases. Differences between the predicted and true compositions now range from 0.02 to 0.54 mol fraction, with an average value of 0.18. Hence, these results highlight the advantages of including non-stoichiometric reference phases, which nearly doubles the number of correctly identified structures and reduces compositional errors by ∼50% when classifying experimentally reported solid solutions.

Multi-phase Classification. Extending the CNN to characterize mixtures of materials, we constructed three new test sets, each containing 1000 simulated multi-phase diffraction spectra. These tests were designed to mimic samples with multiple phases by creating linear combinations of single-phase diffraction peaks derived from 140 stoichiometric reference phases in the Li−Mn−Ti−O−F chemical space. The first two sets consider mixtures generated from randomly selected two- and three-phase combinations with equal weight fractions of the reference phases. In the last set, we probe the effects of impurity phases by simulating two-phase spectra where the weight fractions of the majority and minority phases are randomly set to constitute 70−90 and 10−30% of the mixture, respectively. In all three test cases, data augmentation is applied using mixed artifacts (peak shifting, broadening, texture, and a diffuse and noisy background signal), so that the resulting spectra provide a realistic representation of experimental measurements.

In addition to our newly developed branching algorithm (denoted as B-CNN hereafter), multi-phase identification was performed using three other techniques for comparison: (i) based on the work of Maffettone et al.,[13] a "single-shot" approach (S-CNN) was employed such that the two or three materials with the highest probabilities are chosen for each two- or three-phase mixture, respectively; (ii) by training the CNN explicitly on simulated multi-phase spectra (M-CNN) as described in the work of Lee et al.,[8] entire mixtures of phases are directly predicted as opposed to separately identifying individual phases; and (iii) using JADE to obtain a list of suspected phases for each mixture based on profile matching, the two or three highest-ranked materials are chosen for two- and three-phase spectra, respectively. Given that method (ii) requires many possible linear combinations of single-phase spectra to produce a sufficient number of multi-phase spectra for training, only ideal diffraction spectra were used without

Table 1. Fractions of Materials Correctly Identified with the CNN and JADE When Applied to Experimentally Measured XRD Spectra Designed to Sample Possible Artifacts Arising during Sample Preparation and Synthesis[a]

experimental procedure	anticipated artifact	CNN	JADE
Single-Phase			
pristine samples	none	10/10	9/10
Kapton tape overlaid	diffuse baseline	9/10	8/10
rapid XRD scan	noisy baseline	10/10	7/10
thick samples	shifts in 2θ	5/6	2/6
ball milled	broadening	5/5	4/5
partially disordered	intensity variation	5/6	4/6
solid solutions	non-stoichiometry	4/4	3/4
Multi-Phase			
two-phase mixtures	none	10/10	7/10
three-phase mixtures	none	13/15	9/15
	overall accuracy	93.4%	71.4%

[a] For diffraction spectra of non-stoichiometric materials, a classification is considered correct if the predicted structure is isomorphic to the true structure.

applying any data augmentation. Further details regarding this technique are supplied in the Supplementary Note 6.

In Figure 5a, we show the fraction of phases correctly identified by each of the four methods when tested on two- and three-phase mixtures with equally distributed weight fractions. Among all of the techniques considered here, our newly developed B-CNN algorithm achieves by far the highest accuracy, correctly identifying 87 and 78% of all materials from two- and three-phase spectra, respectively. This outperforms previously reported methods based on deep learning, with the S-CNN[13] and M-CNN[8] giving accuracies of 70% (54%) and 65% (58%) in the classification of two-phase (three-phase) mixtures, respectively. Despite their similarity in performance, these two approaches highlight separate limitations. Recall that the M-CNN does not utilize data augmentation to expand the diversity of its training set and therefore often fails when applied to diffraction spectra containing large perturbations arising from experimental artifacts. In contrast, the S-CNN accounts for possible artifacts through physics-informed augmentation (as in our approach) and consequently is more robust against changes in the diffraction spectra. However, since the S-CNN identifies all phases in a "single shot" without subtracting known diffraction peaks, it leads to misclassifications when similar reference phases produce comparable probabilities for a given spectrum. The B-CNN improves upon both shortcomings using an iterative process of single-phase identification and profile subtraction to achieve higher accuracy. Furthermore, by maximizing the probability over all phases in the predicted mixture, the B-CNN ensures that the first iteration of phase identification is not over-prioritized. If only the most probable phase is evaluated at each step without maximizing probability over the entire mixture, lower accuracies of 78 and 69% are given across two- and three-phase mixtures, respectively.

In Figure 5b, we compare the accuracy of each approach for the classification of majority/minority two-phase mixtures. The B-CNN again outperforms all other evaluated approaches. However, the reliability of our model varies substantially in the identification of majority versus minority phases. The B-CNN correctly classifies 92% of all majority phases, matching its performance across single-phase spectra and therefore suggesting the presence of impurity phases has little to no effect on majority-phase identification. Identifying minority phases, on the other hand, presents a greater challenge, as reflected by a lower accuracy of 64% given by the B-CNN. We note that most misclassifications occur due to imperfect applications of profile subtraction that occasionally leave behind residual intensities or subtract some diffraction peaks associated with the minority phase of interest. Despite this limitation in the *identification* of minority phases, the model generally performs reliably in their *detection*. Recall that the number of phases in a mixture is determined by halting the B-CNN when all diffraction intensities fall below 5% of the initially measured maximum intensity. With this cutoff, the B-CNN correctly reports the presence of a second phase in 93% of the two-phase mixtures with unequally distributed weight fractions. For comparison, when the B-CNN is applied to simulated single-phase spectra with mixed artifacts (Figure 3a) using the same cutoff intensity of 5%, the number of phases is overestimated in only 9% of the samples. The key component enabling a reliable prediction for the number of phases is the approach of profile subtraction. Here, known diffraction peaks are fit to the spectrum through DTW, so that their subtraction yields a new spectrum that accurately represents the mixture minus the phase(s) that has already been identified. This capability is particularly useful in the optimization of synthesis procedures, where it is of interest to know whether the formation of a targeted product is accompanied by some impurity phases.

Application to Experimental Spectra. As a final demonstration of the generalizability of our approach, the B-CNN was applied to experimentally measured spectra in the Li−Mn−Ti−O−F chemical space. In Table 1, we list the fraction of phases correctly identified using the CNN versus JADE, with results categorized by the artifacts and number of phases included for each class of spectra (previously described in the Experimental Measurements). For the classification of pristine diffraction spectra, the CNN correctly identifies all 10 phases considered. Interestingly, JADE incorrectly classifies one material (Li_2TiO_3) from this category. Upon further inspection, the error is attributed to large deviations in the relative peak intensities between the measured and ideal spectra of Li_2TiO_3 (shown in Figure S2), possibly caused by stacking faults in the sample.[24] In the analysis of spectra with diffuse and noisy background signals, the CNN correctly identifies all but one material (anatase TiO_2), likely due to the fact that it exhibits significant diffraction peaks at low values of 2θ where the amorphous background is strong. JADE is found to be more sensitive to background effects, as it yields five misclassifications across these 20 spectra. These misclassifications occur because JADE fails to index peaks that blend in with the background signal and have low intensities or broad widths after a baseline correction is applied. The CNN is more robust against these perturbations since it is trained on spectra having diffraction peaks with varied intensities and widths.

For spectra containing peak shifts, the CNN correctly identifies five out of six phases. In contrast, JADE struggles to handle changes in peak positions, identifying only two phases from this category. This highlights a key weakness of profile matching techniques, which fail when there is a weak overlap between measured and simulated diffraction peaks owing to a shift in 2θ. Fortunately, because the CNN can handle these changes through data augmentation, its performance remains reliable in the classification of spectra with peak shifts. When diffraction peaks are broadened, the CNN and JADE correctly identify five and four phases, respectively, from the five measured spectra. The single misclassification from JADE occurs for Li_2MnO_3 owing to strong overlapping of its neighboring diffraction peaks, an effect which is accounted for by the CNN during training. For the six spectra with changes in their peak intensities, the CNN correctly classifies five phases, while JADE identifies four. The misclassification made by the CNN occurs because the varied peak intensities closely resemble those of a hypothetical solid solution ($Li_{0.5}Mn_{1.5}TiO_4$) that is isomorphic to the true phase ($LiMnTiO_4$). Across non-stoichiometric materials, the CNN correctly predicts all four materials to adopt the rocksalt structure, whereas JADE finds only three phases to be rocksalt. For both methods, the predictions are facilitated by the introduction of hypothetical solid solutions; without including these additional reference phases, neither the CNN nor JADE predicts any of the four samples to be rocksalt-structured.

For the classification of multi-phase mixtures, JADE provides limited accuracy. Only 7/10 and 9/15 phases are correctly identified from two- and three-phase spectra, respectively. Such limitations in accuracy can be attributed to the inability of profile matching techniques to distinguish between diffraction peaks produced by several phases, which often overlap with one another. The B-CNN adeptly overcomes these limitations and correctly identifies 10/10 and 13/15 phases in the two- and three-phase mixtures, respectively. Hence, the benefits provided by deep learning are highlighted by the noticeable disparity between the performances of the CNN and JADE, especially when applied to multi-phase spectra. This advantage is vital to assist in targeted synthesis, considering that attempts to produce novel inorganic materials are frequently impeded by the appearance of multiple impurity phases. Our deep learning approach can therefore be used to identify not only desired products but also impurity phases, which provides insights into why a given synthesis procedure failed and informs future attempts.

The results from testing the CNN on experimentally measured spectra (Table 1) closely match the performance on simulated spectra (Figures 3−5). For example, in spectra where we include a single type of artifact, the CNN correctly identifies 94% of phases from both simulated and experimentally measured single-phase spectra. This lends credence to the simulation-based test cases that are rich in data (e.g., a total of 4200 single-phase test spectra were derived from stoichiometric materials) and suggests that the simulated spectra used for training and testing provide a realistic representation of experimental measurements.

■ DISCUSSION

In summary, we developed an improved deep learning technique that can reliably automate the identification of inorganic materials from XRD spectra. A key advantage of our approach is the physics-informed data augmentation procedure that accounts for several experimental artifacts commonly observed after sample preparation and synthesis. Conventional profile matching techniques often fail when material variations cause large differences between observed and simulated diffraction peaks, requiring manual intervention to analyze any irregularities and identify the samples of interest. In contrast, our CNN learns these differences during training and therefore can autonomously perform phase identification from complex spectra. These benefits are highlighted by the test results presented in this work, which show that the performance of profile matching quickly deteriorates as larger perturbations are applied to the diffraction spectra, whereas the CNN remains reliable in the presence of such perturbations. Furthermore, even though our model is trained only on spectra that account for three types of artifacts (strain, texture, and domain size), it is demonstrated to successfully generalize to spectra outside of the training set. For example, our algorithm achieves a high accuracy for the identification of spectra with diffuse and noisy baseline signals and for samples containing unexpected artifacts (e.g., possible stacking faults in Li_2TiO_3).

Of the artifacts considered in our work, changes in peak positions are shown to be the most challenging to deal with, comprising nearly half of all misclassifications made by the CNN when applied to the simulated diffraction spectra of single-phase stoichiometric materials. Because peak positions are derived from the spacings between crystallographic planes and therefore the lattice parameters of the material, it is difficult to distinguish between isomorphic phases when their structures have a significant degree of strain. We find that our model provides an optimal treatment of changes in peak positions by including samples with as much as ±4% strain in the training set, which is unlikely to be exceeded in experiment unless the materials contain substantial off-stoichiometry. Indeed, tests involving an increased magnitude of strain in the training set led to decreased accuracy during testing owing to degeneracies between the diffraction spectra of similar phases. In general, the bounds used for data augmentation should reflect the experimental system at hand; for example, larger perturbations may be beneficial in cases where certain artifacts are expected to dominate (e.g., epitaxial strain in thin films). When using the approach supplied in our repository (https://github.com/njszym/XRD-AutoAnalyzer), these bounds can be manually specified for any given set of reference phases. To avoid degeneracy of spectra in the training set, the number of reference phases should be constrained to include only those that are expected to arise in experiment—for synthesis, these can be chosen to reflect the composition space spanned by the precursors used and the possibility of reactions with oxygen, water, or CO_2 in air.

The importance of peak positions is further highlighted by our tests involving non-stoichiometric materials. Varying the composition of a material typically leads to changes in its lattice parameters, which in turn shifts the positions of its diffraction peaks. As a result, when the CNN is trained only with stoichiometric reference phases, it frequently fails to identify the structures of non-stoichiometric materials. Because the model is trained to identify individual phases rather than their symmetry, it does not necessarily learn the subtle relationships between peak positions imposed by the space group of each structure. Instead, it considers the positions of all peaks and makes a comparison with known phases in the training set. Therefore, when non-stoichiometry causes large shifts in the positions of diffraction peaks, the CNN will struggle if it has no reference phase available with comparable peak positions. With this in mind, we improved the treatment of non-stoichiometric materials by building a library of hypothetical solid solutions following Vegard's law. After adding their diffraction spectra to the training set, the CNN correctly identifies the structures for 95% of the non-stoichiometric materials considered during testing. We note that this approach is successful because the lattice parameters of most solid solutions follow Vegard's law with only minor deviations.[25] When deviations do occur, data augmentation ensures that the match between hypothetical and experimentally observed phases need not be exact for the model to maintain a high level of accuracy for the identification of the material's structure.

Despite the improved prediction of the structure enabled by introducing hypothetical solid solutions to the training set, predicting the compositions of non-stoichiometric materials remains challenging. This limitation can be understood by considering the effects of non-stoichiometry on diffraction peak intensities, which are influenced by the structure's internal cell coordinates and site occupancies. Given the similarity of structural frameworks between materials forming solid solutions, changes in cell coordinates are usually small and therefore do not contribute significantly to differences in peak intensities. Changes in site occupancies, however, strongly influence peak intensities owing to the distinct scattering factors of substituted species. As opposed to changes in lattice parameters that can be described by Vegard's law, an automatic prediction of site occupancy is more difficult to achieve because site occupancies can redistribute in solid solutions. For example, partial inversion (i.e., swapping Wyckoff positions) between lithium and transition metal ions has been observed in spinel $LiMn_{2-x}Ti_xO_4$.[26] Such differences give rise to errors in predicted compositions and not structures because site occupancies control peak intensities while leaving peak positions relatively unaffected. Hence, we re-iterate that our approach is not designed to give precise refinements of composition but rather to provide a reliable prediction of the structure and an estimate of the composition.

Beyond the scope of this work, future efforts may be conducted to design a more accurate prediction of site occupancies so that refinement can be carried out autonomously. A recent report by Mattei et al. has shown some progress toward this end, providing an approach to enumerate many possible distributions of site occupancies with the goal of identifying the best match with experimental measurements.[27] As their approach requires the structural framework of the suspected phase to be known prior to refinement, our model may prove useful in coordination with their algorithm. The results from our CNN may also provide a useful starting point for manual Rietveld refinement as they contain necessary information regarding the composition and structure of each phase identified in a spectrum. An estimation of the lattice parameters can be given for these phases based on their corresponding entries in the ICSD. Furthermore, because DTW measures the shift in 2θ between experimental and simulated diffraction peaks, it is possible that our model can provide a more precise estimation of the lattice parameters by relating peak shifts with strain parameters through Bragg's law. Demonstrating this capability is outside the scope of the current report but may be considered in future work.

When samples contain more than one material, new challenges arise as diffraction peaks often overlap and can be difficult to distinguish. To handle multi-phase spectra, we designed a branching algorithm that iterates between phase identification and profile subtraction to identify the combination of phases that maximizes the average probability given by the CNN. This approach yields exceptionally high accuracy across simulated and experimentally measured multi-phase XRD spectra, exceeding the performance of profile matching techniques and recently published methods based on deep learning. The advantages of our branching algorithm can be summarized by two main points. First, by training only on single-phase spectra, we avoid the combinatorial explosion of training samples that would arise if multi-phase spectra were instead used. Because the number of pristine reference spectra is kept low, many experimental artifacts can be included through physics-informed data augmentation, which ensures that the model is robust against perturbations in diffraction spectra caused by defects or impurities. Second, our algorithm avoids confusion between phases with similar reference spectra by identifying phases in a one-by-one manner and iteratively subtracting their diffraction peaks from the spectrum until all non-negligible intensities have been accounted for. The removal of known peaks prevents the algorithm from overestimating the number of phases in a sample, which would otherwise occur if the probability distribution given by the CNN was assumed to represent a mixture of phases (e.g., assuming that all phases with a probability ≥50% exist in a given sample).

■ CONCLUSIONS

We have demonstrated that a deep learning algorithm based on a CNN can be trained to identify inorganic materials from complex diffraction spectra. Physics-informed data augmentation was shown to accurately account for possible experimental artifacts in measured diffraction spectra, therefore improving the generalizability of the CNN. Simulated spectra derived from hypothetical solid solutions were also added to the training set, which improves the performance of the model when dealing with off-stoichiometric samples. For samples containing multiple phases, an iterative process of phase identification and profile subtraction was designed to maximize the probability given by the CNN over all phases in the predicted mixture, which performs well when applied to multi-phase spectra. The proposed accuracy of our deep learning approach was validated with respect to simulated and experimentally measured diffraction spectra.

Although our current tests focus on materials in the Li−Mn−Ti−O−F space, the algorithm developed here (provided below) can be applied to any arbitrary composition space given a set of reference phases, which can be extracted from existing crystallographic databases. Based on the 255 reference phases considered in this work, the entire process of spectrum simulation, data augmentation, and model training was completed in 20 h on a single compute node with 16 CPUs. Because the number of training samples required by our method scales linearly with the number of reference phases, new models can be created on much broader composition spaces without requiring excessive amounts of time or computational resources. The compositions included during training should be chosen to reflect anticipated elements in the samples being characterized, and therefore, it is generally not necessary to include *all* compositions in a single model. Once a model is trained for a given chemical space, it can be applied rapidly and automatically to each experimental XRD spectrum to predict what phases are in the sample. Additionally, new reference phases can be introduced to the model at any time without requiring the regeneration of training spectra for existing phases. Given the efficiency of our approach and the promising results illustrated throughout this work, we suggest that the algorithm developed here may be used to effectively accelerate materials discovery by incorporating automatic phase identification to support high-throughput and autonomous experimental workflows.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.chemmater.1c01071.

> List of the compositions and structures used during training and testing, how the magnitudes of peak shifts included during training affect the accuracy of the resulting model during testing, visualization of the differences between measured and experimental spectra for Li_2TiO_3, and data augmentation, spectrum simulation, and training procedures (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Gerbrand Ceder − *Department of Materials Science & Engineering, UC Berkeley, Berkeley, California 94720, United States; Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States;* orcid.org/0000-0001-9275-3605; Email: gceder@berkeley.edu

Authors

Nathan J. Szymanski − *Department of Materials Science & Engineering, UC Berkeley, Berkeley, California 94720, United States; Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States;* orcid.org/0000-0003-2255-9676

Christopher J. Bartel − *Department of Materials Science & Engineering, UC Berkeley, Berkeley, California 94720, United States; Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States;* orcid.org/0000-0002-5198-5036

Yan Zeng − *Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States*

Qingsong Tu − *Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States;* orcid.org/0000-0002-2345-799X

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.chemmater.1c01071

Notes

The authors declare no competing financial interest.

A public repository containing the methods discussed in this work can be found at https://github.com/njszym/XRD-AutoAnalyzer. This includes the code used to perform data augmentation, generation of hypothetical solid solutions, training of the CNN, and application of the CNN to classify XRD spectra using the probabilistic branching algorithm. A pre-trained model is available for the Li−Mn−Ti−O−F chemical space.

All XRD spectra used for testing can be found on Figshare. Reported accuracies can be reproduced by applying our pre-trained model to these spectra.

■ ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, under Contract No. DE-AC02-05-CH11231 within the Joint Center for Energy Storage Research (JCESR) program. We also acknowledge support from the National Science Foundation Graduate Research Fellowship under grant no. 1752814.

■ REFERENCES

(1) Stein, H. S.; Gregoire, J. M. Progress and prospects for accelerating materials science with automated and autonomous workflows. *Chem. Sci.* **2019**, *10*, 9640−9649.

(2) Ludwig, A. Discovery of new materials using combinatorial synthesis and high-throughput characterization of thin-film materials libraries combined with computational methods. *npj Comput. Mater.* **2019**, *5*, 70.

(3) Altomare, A.; et al. Advances in powder diffraction pattern indexing: N-TREOR09. *J. Appl. Crystallogr.* **2009**, *42*, 768−775.

(4) Le Meins, J.-M.; Cranswick, L. M. D.; Le Bail, A. Results and conclusions of the internet based "Search/match round robin 2002". *Powder Diffr.* **2003**, *18*, 106−113.

(5) Gilmore, C. J.; Barr, G.; Paisley, J. High-throughput powder diffraction. A new approach to qualitative and quantitative powder diffraction pattern analysis using full pattern profiles. *J. Appl. Crystallogr.* **2004**, *37*, 231−242.

(6) Iwasaki, Y.; Kusne, A. G.; Takeuchi, I. Comparison of dissimilarity measures for cluster analysis of X-ray diffraction data from combinatorial libraries. *npj Comput. Mater.* **2017**, *3*, 4.

(7) Oviedo, F.; et al. Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. *npj Comput. Mater.* **2019**, *5*, 60.

(8) Lee, J.-W.; Park, W. B.; Lee, J. H.; Singh, S. P.; Sohn, K.-S. A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic XRD powder patterns. *Nat. Commun.* **2020**, *11*, 704.

(9) Park, W. B.; et al. Classification of crystal structure using a convolutional neural network. *IUCrJ* **2017**, *4*, 486−494.

(10) Vecsei, P. M.; Choo, K.; Chang, J.; Neupert, T. Neural network based classification of crystal symmetries from x-ray diffraction patterns. *Phys. Rev. B* **2019**, *99*, 245120.

(11) Suzuki, Y.; et al. Symmetry prediction and knowledge discovery from X-ray diffraction patterns using an interpretable machine learning approach. *Sci. Rep.* **2020**, *10*, 21790.

(12) Aguiar, J. A.; Gong, M. L.; Tasdizen, T. Crystallographic prediction from diffraction and chemistry data for higher throughput classification using machine learning. *Comput. Mater. Sci.* **2020**, *173*, 109409.

(13) Maffettone, P. M.; et al. Crystallography companion agent for high-throughput materials discovery. *Nat. Comput. Sci.* **2021**, *1*, 290−297.

(14) Wang, H.; et al. Rapid Identification of X-ray Diffraction Patterns Based on Very Limited Data by Interpretable Convolutional Neural Networks. *J. Chem. Inf. Model.* **2020**, *60*, 2004−2011.

(15) Belsky, A.; Hellenbrandt, M.; Karen, V. L.; Luksch, P. New developments in the Inorganic Crystal Structure Database (ICSD):

accessibility in support of materials research and design. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 364−369.

(16) Clément, R. J.; Lun, Z.; Ceder, G. Cation-disordered rocksalt transition metal oxides and oxyfluorides for high energy lithium-ion cathodes. *Energy Environ. Sci.* **2020**, *13*, 345.

(17) Ong, S. P.; et al. Python Materials Genomics (pymatgen): A Robust, Open-Source Python Library for Materials Analysis. *Comput. Mater. Sci.* **2013**, *68*, 314−319.

(18) Hume-Rothery, W.; Powel, H. M. On the theory of superlattice structures in alloys. *Z. Kristallogr.-Cryst. Mater.* **1935**, *91*, 23−47.

(19) Vegard, L. Die Konstitution der Mischkristalle und die Raumfüllung der Atome. *Z. Phys.* **1921**, *5*, 17−26.

(20) Patterson, A. L. The Scherrer formula for X-ray particle size determination. *Phys. Rev.* **1939**, *56*, 978−982.

(21) Berndt, D. J.; Clifford, J. Using Dynamic Time Warping to Find Patterns in Time Series. *KDD Workshop*, 1994; pp 359−370.

(22) Murphy, D. T.; Schmid, S.; Hester, J. R.; Blanchard, P. E. R.; Miiller, W. Coordination Site Disorder in Spinel-Type LiMnTiO4. *Inorg. Chem.* **2015**, *54*, 4636−4643.

(23) *JADE Pro*; Materials Data MDI: Livermore, CA, USA, 2019.

(24) Watanabe, A.; et al. Structural analysis of imperfect Li2TiO3 crystals. *J. Alloys Compd.* **2020**, *819*, 153037.

(25) Gschneidner, K. A.; Vineyard, G. H. Departures from Vegard's Law. *J. Appl. Phys.* **1962**, *33*, 3444.

(26) Krins, N.; et al. LiMn2−xTixO4 spinel-type compounds (x ≤ 1): Structural, electrical and magnetic properties. *Solid State Ionics* **2006**, *177*, 1033−1040.

(27) Mattei, G. S.; et al. Enumeration as a Tool for Structure Solution: A Materials Genomic Approach to Solving the Cation-Ordered Structure of Na3V2(PO4)2F3. *Chem. Mater.* **2020**, *32*, 8981−8992.