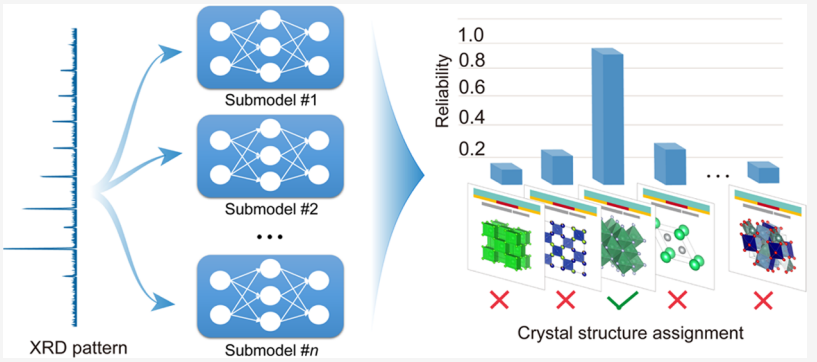
JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
pubs.acs.org/JACS
Article

Crystal Structure Assignment for Unknown Compounds from X-ray Diffraction Patterns with Deep Learning

Litao Chen,[†] Bingxu Wang,[†] Wentao Zhang,[†] Shisheng Zheng, Zhefeng Chen, Mingzheng Zhang, Cheng Dong, Feng Pan,* and Shunning Li*

Cite This: J. Am. Chem. Soc. 2024, 146, 8098–8109



ABSTRACT: Determining the structures of previously unseen compounds from experimental characterizations is a crucial part of materials science. It requires a step of searching for the structure type that conforms to the lattice of the unknown compound, which enables the pattern matching process for characterization data, such as X-ray diffraction (XRD) patterns. However, this procedure typically places a high demand on domain expertise, thus creating an obstacle for computer-driven automation. Here, we address this challenge by leveraging a deep-learning model composed of a union of convolutional residual neural networks. The accuracy of the model is demonstrated on a dataset of over 60,000 different compounds for 100 structure types, and additional categories can be integrated without the need to retrain the existing networks. We also unravel the operation of the deep-learning black box and highlight the way in which the resemblance between the unknown compound and a structure type is quantified based on both local and global characteristics in XRD patterns. This computational tool opens new avenues for automating structure analysis on materials unearthed in high-throughput experimentation.

INTRODUCTION

Autonomous synthesis and characterizations in high-throughput experimentation have rapidly emerged as a means to accelerate the experimental cycle for material innovation in various fields.[1] However, along with the advent of this new material research paradigm, there arises a daunting challenge regarding the analysis of a large portfolio of data generated by characterization techniques, among which X-ray diffraction (XRD) is arguably the most important one. The conventional method for XRD interpretation is based on pattern matching between the sample and the reference patterns in XRD databases. This procedure requires the joint optimization of multiple parameters[2−6] and can be quite an arduous task that may take a well-trained expert a considerably large amount of time.[7,8] This is especially true when some of the constituent compounds are not yet stored in the existing XRD databases, either because of newly discovered crystal structures or because of a compositional deviation from solid-solution compounds. Hence, there is a dire need to develop methods to expedite and automate the crystal structure identification for such unknown compounds, which may constitute a basic step toward the practical application of self-driving laboratories[9−11] for materials science.

Deep-learning algorithms have recently gained increasing popularity for assisting XRD interpretation, as they can effectively extract the latent information that is often hard to manually capture from the diffraction patterns.[12−26] Early pioneering studies have primarily focused on the autonomous inference of structural attributes from the XRD patterns, including lattice parameters,[27] space group,[28−32] and crystallographic dimensionality.[33] These attributes can be fed to the traditional rule-based approaches to speed up the process of

Received: November 1, 2023
Revised: February 24, 2024
Accepted: February 27, 2024
Published: March 13, 2024

https://doi.org/10.1021/jacs.3c11852
J. Am. Chem. Soc. 2024, 146, 8098−8109

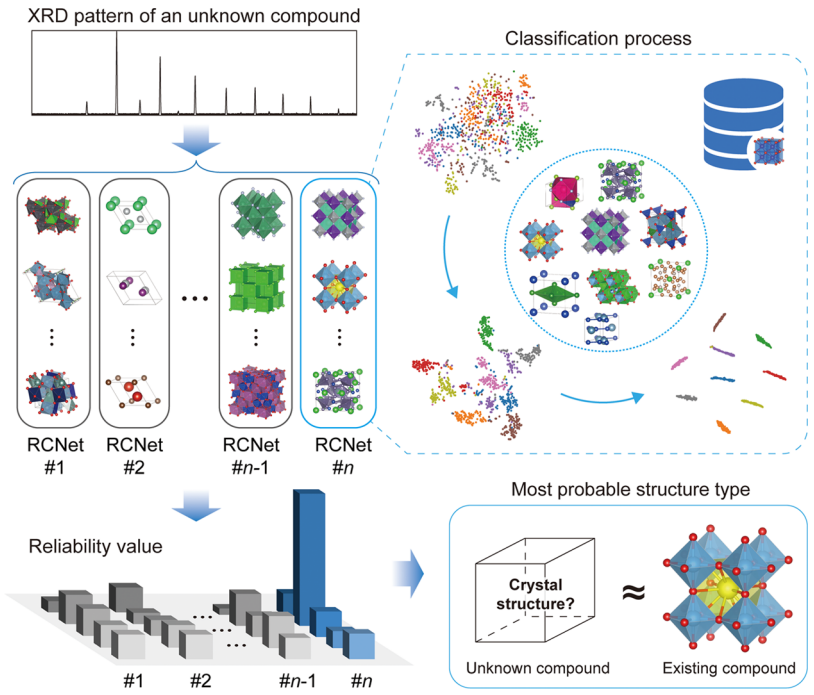


Figure 1. Overview of the CrySTINet model for XRD interpretation. CrySTINet is based on a union of ResNet Confidence Networks (RCNets), each of which performs an isolate classification task based on a distinct group of structure types and then outputs a reliability value to measure how similar the XRD patterns are between the unknown compound and a structure type of existing compounds in the corresponding dataset. The highest reliability value among all RCNets indicates the most probable structure type for this unknown compound.

XRD analysis. However, the ultimate goal of autonomous XRD interpretation is to identify the constituent phases without human intervention and not merely to infer the structural attributes. Only recently has this goal been partially accomplished, with the use of deep convolutional neural networks (CNNs) trained on simulated XRD datasets.[34−37] It was demonstrated via class activation mapping analysis[38] that the CNN models can learn the distribution characteristics of main diffraction peaks upon an adequate amount of XRD data, which allows the accurate discrimination of compounds from experimental data. Despite the inspiring results in these studies, there remains a significant unmet need for deciphering the structures of new compounds unseen in both experimental and simulated XRD datasets, especially in a wide compositional range. If such unseen compounds (not included in the labels of the training data) were to emerge in experiments, false interpretation would naturally be expected from most of the previous CNN models. Although a model has been recently developed for the discrimination of a specific type of structure (e.g., perovskites) without being restricted to the reported compounds,[39] it can only focus on the classification of a single category and is difficult to be directly extended to other crystal structures. Up to now, a universal model for XRD interpretation of various kinds of new compounds has not yet been reported.

In conventional XRD analysis tools, there is one way to help identify the structures of these new compounds: searching for the isoconfigurational structures.[40−43] For example, high pressure could induce the phase transition of certain compounds from a four-coordinate wurtzite to a six-coordinate rock-salt structure.[44,45] The high-pressure phase that may not yet be stored in the databases can be intuitively assigned to a rock-salt prototype for Rietveld refinement, and by pattern matching, one can determine whether the high-pressure phase has the isoconfigurational structure of the rock-salt prototype. If not, another prototype will be suggested, followed by a new pattern matching process, and this operation is repeated until the correct isoconfigurational structure is found. While this approach has proved indispensable in conventional XRD analysis, it has not yet been incorporated into any of the previous deep-learning algorithms. Thus, fulfilling the ultimate goal of autonomous XRD interpretation involves the task of an automatic search for the isoconfigurational structures, which can be translated into a problem of identifying the most probable structure type for the unknown compound from an XRD pattern.

The present work introduces a model named crystal structure-type identification network (CrySTINet) on the basis of convolutional residual neural networks (ResNet)[46] to automate the identification of the most probable structure types for inorganic compounds from XRD patterns. The applicability and potential of this model are demonstrated on a dataset of 63,963 compounds extracted from the Inorganic Crystal Structure Database (ICSD),[47] with all of these compounds belonging to the top 100 most popular structure types in ICSD. This model can reach a promising accuracy of 80.0% without requiring any prior information on the composition of the materials, and it is extensible for a larger number of structure types according to the need of the task. We also tracked the internal behavior of the network and provided a detailed account of how we score the structural resemblance between the unknown compound and a structure type from the XRD patterns. This work lays the foundation for the autonomous XRD interpretation module in the high-throughput experimentation for material discovery.

https://doi.org/10.1021/jacs.3c11852
J. Am. Chem. Soc. 2024, 146, 8098−8109

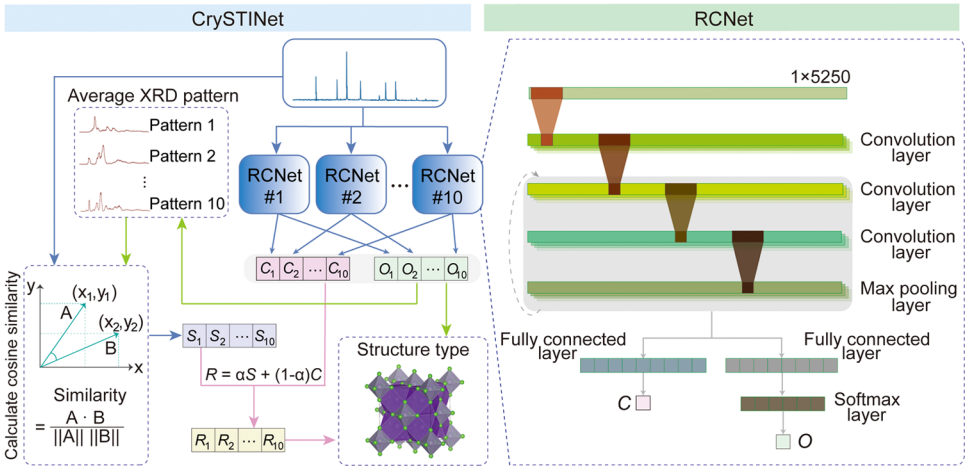


Figure 2. The architecture of CrySTINet and RCNet. The CrySTINet takes the one-dimensional vector of the XRD pattern as an input, which is fed into all of the RCNets in the framework. Each RCNet employs a residual architecture to classify the pattern as one of the structure types (O) included in the training dataset and simultaneously outputs a confidence value (C). After that, a cosine similarity value (S) is calculated according to the input pattern and the average XRD pattern for the structure type O predicted by the corresponding RCNet. A series of reliability values (R) are derived based on S and C, which combine with the classification results of the RCNets to arrive at the final decision on the recommended structure type.

RESULTS

Architecture of CrySTINet. Following the work of Allmann and Hinek,[48] compounds in ICSD can be categorized into different structure types according to the space group, Wyckoff sequence, Pearson symbol, c/a ratio, β ranges, and some composition-relevant information. This kind of categorization has already been adopted by the latest ICSD release, with the structure-type notation directly stored in the crystallographic information file for each compound. We extracted this information from all ICSD entries and used it as the basis for the classification output. Any two compounds that belong to the same structure type can be regarded as sharing an isoconfigurational structure. According to the number of compounds belonging to each category, the top 100 most popular structure types were identified, corresponding to a total of 63,963 inorganic compounds. The one-dimensional XRD patterns of all of these compounds were taken as the dataset for CrySTINet. Unlike the previous studies[27−37] that used a restricted compositional pool in the training and evaluation of their deep-learning models for XRD interpretation, our model can cover a wide compositional range that includes nearly all elements in the periodic table.

Beyond the 100 popular structure types, there are a lot more species that each correspond to relatively few kinds of compounds. It means that these structure types can emerge only at distinctive compositions, but despite this, they should not be entirely left out of consideration in the autonomous XRD interpretation process. Whenever there is the possibility for some uncommon structure types to emerge at a combination of elements specified in the experiments, these structure types should be added to the deep-learning model. Therefore, it is necessary that this model can be conveniently adjusted and extended to additional categories of structure types. To achieve this goal, the framework of CrySTINet has been crafted as a union of submodels named ResNet Confidence Networks (RCNets), each trained from a different subset of structure types (Figure 1). These subsets, here denoted as Datasets #1 to #n, are derived from the grouping of all of the structure types required for classification. When an input pattern is fed into the RCNets, each one of them will arrive at a distinct classification decision according to the corresponding dataset. Along with this classification result, a reliability value, which indicates the probability of the match between the decided structure type and the unknown compound, will be outputted by each RCNet. The structure type with the highest reliability value among all RCNets will be selected as the one finally recommended for the unknown compound. We note that the initial set of RCNets are trained on the 100 popular structure types, while any additional species could be incorporated in a supplementary network without retraining the initial set of RCNets. Utilizing this architectural feature, we can readily extend the model to any of the less popular crystal structures without sacrificing its high recognition ability on the common structure types.

The architecture of RCNet in this work (Figure 2) is adapted from those designed for image-related tasks in computer vision.[46] We have deployed a method proposed by Devries et al.[49] to output the confidence value (C) via a confidence estimation branch added after the penultimate layer of the network. This value reflects to some extent how much the XRD pattern of the unknown compound resembles the characteristic pattern of a structure type, as predicted by an RCNet. However, we find that this confidence value is not reliable enough to identify the most probable structure type among all RCNets, which stems from a weak ability to discriminate similar structures that belong to datasets of different RCNets. To overcome this limitation, we explored the combined use of confidence value and cosine similarity value (S), the latter of which is calculated between the input pattern and the general feature of the simulated patterns for the output structure type (denoted as O) by each RCNet. In an aim to guarantee a sufficiently large difference between any two structure types in different datasets, we have performed an unsupervised clustering for the 100 popular structure types before the training of RCNets (see the Methods section and Supporting Note 1 for details), dividing them into 10 subsets, with each subset serving as the dataset to train one RCNet. After the output structure type O of each RCNet for the unknown compound is acquired, the simulated patterns of all

https://doi.org/10.1021/jacs.3c11852
J. Am. Chem. Soc. 2024, 146, 8098−8109

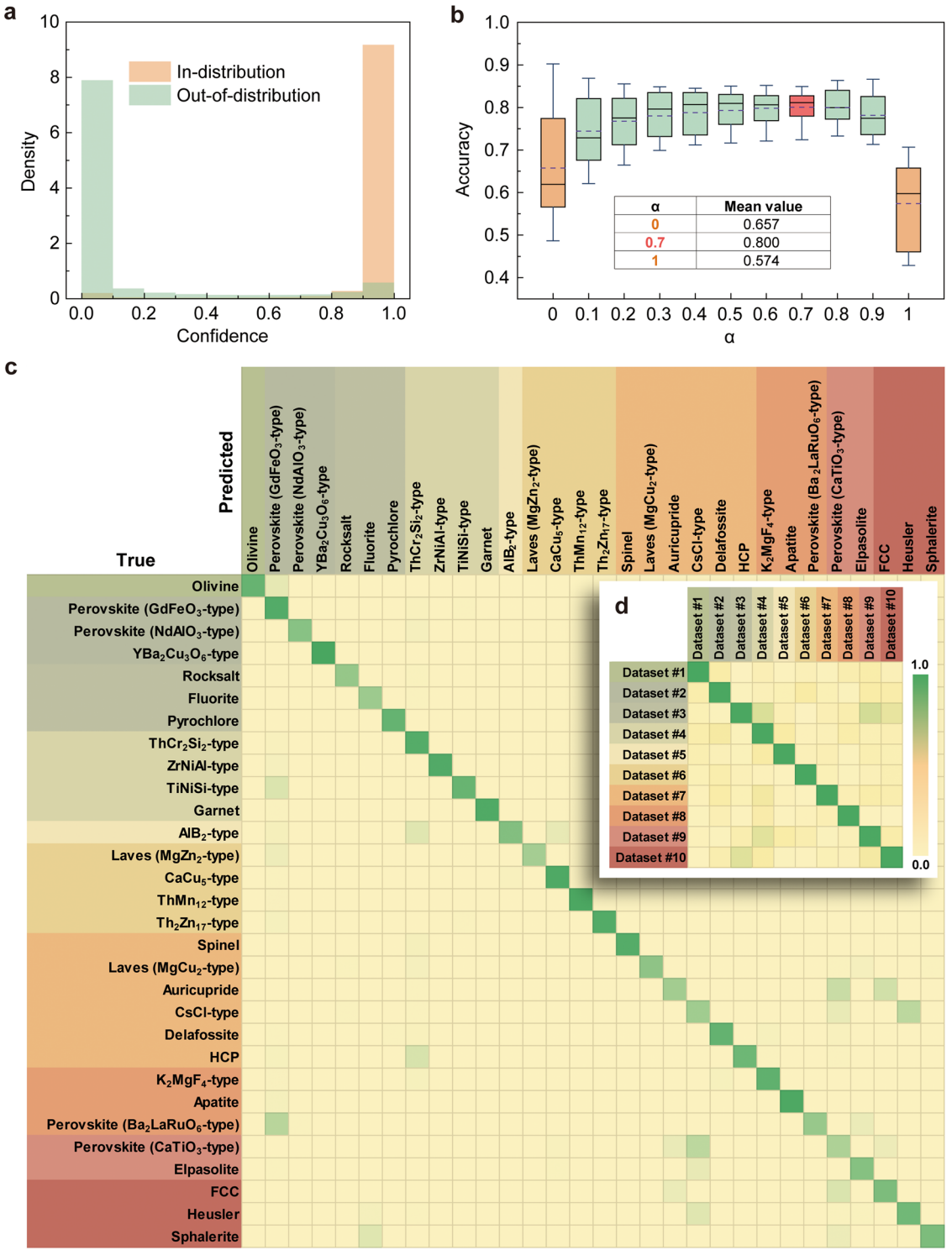


Figure 3. Performance of CrySTINet. (a) In- and out-of-distribution patterns of the confidence value for RCNet #1. (b) Test accuracy of CrySTINet as a function of α in the formula for calculating the reliability value, $R = \alpha S + (1 - \alpha)C$. (c) Confusion matrix for the top 30 most popular structure types. Entries along the descending diagonal represent the accuracies for the corresponding structure types, while other entries indicate the percentages of misclassification. (d) Performance breakdown for subsets of structure types. The entries represent the percentages out of the test XRD patterns that are classified into a certain dataset (row) given a ground-truth label (column).

compounds in this structure type will be used to obtain an average XRD pattern. This pattern is taken as the general feature to calculate the similarity value S for the corresponding output O. We note that for both the training process of RCNets and the calculation of S, all of the datasets of simulated patterns have been augmented via the hypothetical perturbations on the physical information on crystal structures, which allows us to take into account some experimental complexities, such as strain, domain size, and preferred crystal orientation (Supporting Note 2). Thereafter, a linear combination of S and C with a formula of $\alpha S + (1 - \alpha)C$ is employed to determine the reliability value (R), a variable that

https://doi.org/10.1021/jacs.3c11852
J. Am. Chem. Soc. 2024, 146, 8098−8109

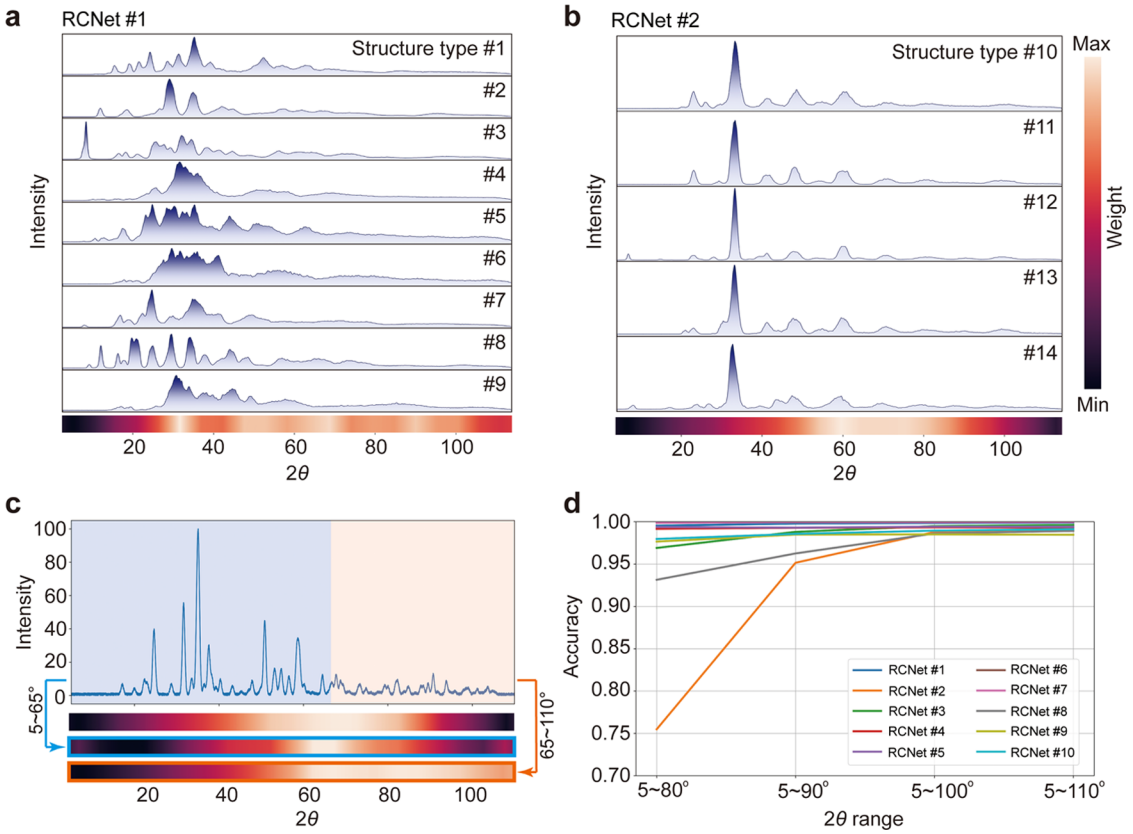


Figure 4. Model analysis via Grad-CAM. (a) The average XRD pattern over compounds belonging to one of the structure types (#1−#9) in Dataset #1 and the average Grad-CAM heatmap for all of the compounds of Dataset #1 in RCNet #1. (b) The average XRD pattern over compounds belonging to one of the structure types (#10−#14) in Dataset #2 and the average Grad-CAM heatmap for all of the compounds of Dataset #2 in RCNet #2. (c) The XRD pattern of $Fe_{1.10}Mg_{0.75}Mn_{0.15}SiO_4$ in data set #1 and its Grad-CAM heatmaps in RCNet #2 when using different 2θ ranges for structure-type classification: 5−110° (top), 5−65° (middle), and 65−110° (bottom). The use of 5−110 and 65−110° ranges will lead to falsely high confidence values with this compound classified into a structure type in Dataset #2. In contrast, the 5−65° setting yields a much reduced confidence value in RCNet #2, and the compound is correctly classified into Dataset #1, revealing that an attentional bias toward the high-angle XRD peaks may cause misclassifications between datasets in the model. **d**, The accuracy of each RCNet as a function of the angle region used for model training, demonstrating the indispensable role of high-angle XRD peaks in maximizing the accuracy for isolate RCNets.

has been adopted in our model to effectively predict the ranking of probable structure types for the unknown compound.

Model Performance. The simulated XRD pattern data of all compounds for the 100 popular structure types are randomly divided into the training set, validation set, and testing set with a ratio of 7:1:2. First, we evaluate the performance of structure-type classification in an isolate RCNet. Here, we take RCNet #1 as an example, where the testing data come from the same repository (Dataset #1) as the training data. In this setting, the test accuracy reaches a promising level of 99.89%, with a confidence value of generally 0.9−1.0 (we denoted this value as the in-distribution confidence, as shown in Figure 3a). Results of other RCNets are provided in Supporting Note 3. Then, we evaluate the classification across different RCNets. This time, we use other subsets of structure types (Datasets #2−10) as the testing data for RCNet #1 (note that its training is still on Dataset #1) to measure the confidence value, which is called out-of-distribution confidence (Figure 3a). This value is mainly distributed between 0 and 0.1, sufficiently low to discriminate the structures in other subsets from those in Dataset #1. However, there is still a non-negligible possibility of the C value reaching above 0.9 in this setting. It means that the classification into a false subset of structure types will potentially take place, which leads to an overall accuracy of only 65.7% for the combinatorial classification of all RCNets. This precludes us from using the confidence value alone for the identification of the structure type. Therefore, we conceived a reliability value as mentioned above that combines the confidence value (which captures the characteristic patterns of different structure types with sole reliance on a single dataset) and the similarity value (which reinforces the differentiation among all datasets based on the general feature of XRD patterns). By variation of α, we can seek to optimize the model performance (Figure 3b). With α = 0.7, a test accuracy of 80.0% is reached for the classification task of CrySTINet, which makes it a practical surrogate model to replace the conventional labor-intensive trial-and-error process in searching for the isoconfigurational structure of an unknown compound. We note that after making this recommendation for the isoconfigurational structure, additional processes using rule-based pattern matching techniques could be further incorporated to achieve XRD analysis as fine-grained as possible.

A performance breakdown of the model for individual structure types is presented in the confusion matrix in Figure 3c, where the top 30 most popular structure types are taken as examples. The confusion matrix indicates the percentage of trials of a given compound that can be classified into one of the

https://doi.org/10.1021/jacs.3c11852
J. Am. Chem. Soc. 2024, 146, 8098−8109

given structure types. The highest percentages fall into the cells along the descending diagonal, meaning that most of the classifications match the expectation. Misclassification generally occurs when two structure types belonging to different datasets are alike, for example, $GdFeO_3$-type perovskite in Dataset #2 and Ba_2LaRuO_6-type perovskite in Dataset #8. Nevertheless, our prior unsupervised clustering of the structure types before model training can prevent the case where numerous structures in one dataset are similar to those in the other. Thus, we could see that the percentages of misclassification between Datasets #2 and #8 are relatively low (Figure 3d). This helps to improve the overall accuracy of the CrySTINet model, especially for handling compounds with diverse possible structure types.

We further examined the accuracy of CrySTINet for experimentally measured XRD patterns. A total of 80 different patterns of inorganic compounds were extracted from the RRUFF database,[50] and each of them has been manually and unambiguously identified as belonging to one of the top 100 most popular structure types. With these compounds employed for the examination, CrySTINet yielded an overall accuracy of 81.3% (Supporting Note 4). Most of the correct classifications are supported by a reliability value exceeding 0.9, and the test involves diverse kinds of structures without any constraint in the compositional space. Here, we have to mention that some of the test data in the RRUFF database will correspond to known compounds in ICSD and exhibit high similarity to specific samples of the simulated XRD data. This could lead to an overestimated accuracy in the experimental test set. In spite of that, our results can validate the practical applicability of simulated XRD dataset in establishing a robust deep-learning model for the interpretation of experimentally measured XRD patterns.

Prediction Interpretability. The interpretability of predictions made by a neural network is essential for the assessment of whether the high accuracy comes from the proper extraction of discriminative features or from the exploitation of artifacts in the data.[51] To better interpret the classification decisions of CrySTINet, we can rely on gradient-weighted class activation mapping (Grad-CAM),[52] which is one of the best techniques for generating saliency maps, i.e., locating the attentional regions in the input feature space that are relevant for positive predictions.[38,52−55] The Grad-CAM allows for visualization of the attention maps on a trained CNN-based architecture and can be applied to XRD patterns to highlight the most important angle regions where the model mainly focuses on when developing a prediction.

RCNets #1 and #2 are taken as examples to illustrate the prediction interpretability, as shown in Figure 4a,b. For the dataset in RCNet #1, Grad-CAM identifies several angle regions that are representative of significant difference among the XRD patterns of all structure types (#1−#9) in Dataset #1. Notably, the neural network shows particular attention across the strong peaks at low 2θ angles (30−35°). In contrast, RCNet #2 focuses primarily on the 55−85° region, with little attention paid to the low-angle region. This can be rationalized by the close similarity of the low-angle peaks (nearly identical position and height) for all five structure types (#10−#14) in Dataset #2. The ability of RCNets to automatically locate the correct angle regions that underlie variation in XRD patterns of different structure types is indicative that our model has extracted informative features for the separation of categories in the latent space. The importance of capturing the informative features in XRD patterns has also been demonstrated in recent studies for automating the discrimination of specific structure types like perovskites.[39]

The Grad-CAM analysis could also bring transparency to the model, providing insights into why it fails to reach a desired level of accuracy without the incorporation of the similarity value S. We exemplify this issue on $Fe_{1.10}Mg_{0.75}Mn_{0.15}SiO_4$ (its simulated XRD pattern is shown in Figure 4c), which belongs to the olivine structure type in Dataset #1. In an ideal case, a high confidence value C is expected for RCNet #1, while low values are expected for the others. However, RCNet #2 actually returns a confidence value of 0.97, exceeding that of RCNet #1. $Fe_{1.10}Mg_{0.75}Mn_{0.15}SiO_4$ is therefore misclassified as the $GdFeO_3$-type perovskite structure type in Dataset #2 if we rely solely on the confidence value. We note that the Grad-CAM heatmap for $Fe_{1.10}Mg_{0.75}Mn_{0.15}SiO_4$ in RCNet #2 indicates a high attention weight in the 2θ range of 50−80° (top heatmap in Figure 4c), which overlaps significantly with that of the structure types in Dataset #2 (Figure 4b). This suggests that the false classification is related to the high-angle peaks of the pattern of $Fe_{1.10}Mg_{0.75}Mn_{0.15}SiO_4$. Here, we attempted to verify the contribution of different angle regions by masking portions of the input XRD pattern. When the range of 65−110° is retained, the compound is still classified as the $GdFeO_3$-type perovskite structure type, and the Grad-CAM heatmap shows a rather scattered distribution (bottom heatmap in Figure 4c). In contrast, when the range of 5−65° is retained, a correct classification into olivine structure type can be reached, with Grad-CAM highlighting the peaks close to 65° (middle heatmap in Figure 4c). The fact that truncating the high-angle XRD peaks could be of benefit in reducing the out-of-distribution confidence value for RCNets implies that an overfocused attention of the network on high-angle peaks is probably one of the main reasons behind the misclassifications determined by the confidence value.

Actually, in conventional XRD analysis, the manual pattern matching process is typically conducted by drawing the rough analogy between the patterns of the unknown compound and an existing compound with a focus on the strong peaks that often emerge at low 2θ angles. The concentration of attention on the low-angle region is also witnessed on submodels such as RCNet #1 (Figure 4a), but sometimes the low-angle peaks in a dataset are so similar in position and height that the model attention has to be steered to high-angle peaks, such as RCNet #2 (Figure 4b). We note that this is fully compatible with our expert knowledge. Nevertheless, while in theory, the high-angle peaks still carry sufficient information to differentiate all of the structure types, the missing information from low-angle regions indeed hampers the capability to identify out-of-distribution samples (i.e., structure types not included in the training set for this submodel) in the combinatorial classification of all RCNets. The confidence value output by the network is derived under such missing information, which could be a major problem for improving the model accuracy. To mitigate this problem, we may try tuning the 2θ range of the input data, as inspired by the example of $Fe_{1.10}Mg_{0.75}Mn_{0.15}SiO_4$. However, as shown in Figure 4d, the accuracy of some isolated RCNets will be seriously compromised when the upper bound of the 2θ range is set below 90°. Therefore, adjusting the 2θ range is clearly not a viable solution. The above results also imply that using a single large ResNet model instead of a union of submodels will potentially lead to higher accuracy, but it would

https://doi.org/10.1021/jacs.3c11852
J. Am. Chem. Soc. 2024, 146, 8098−8109

lose the ability to be conveniently adjusted and extended to the additional datasets.

Alternatively, we try to employ a variable that carries information about the global similarity between an unknown compound and a structure type. This variable is the similarity value S as mentioned above, which complements the confidence value C and reinforces the differentiation among different datasets of XRD patterns. We note that the confidence value occasionally fails to reflect the peak information in some angle regions, as demonstrated by the Grad-CAM analysis above, while the similarity value never does, because it is calculated from the whole XRD pattern of the unknown compound (Figure 5). On the other hand, the similarity value does not contain any bias with respect to the characteristic XRD peaks that are deemed critical by the neural networks, whereas the confidence value does. Moreover, the similarity value could only serve as a remedy for the confidence value because the latter performs better in an overall sense (Figure 3b). Consequently, the combination of both variables into a single score, namely, the proposed reliability value R, could enable a more reliable quantification of the resemblance between the unknown compound and a structure type and therefore provides the potential for precise XRD interpretation of unknown compounds using a union of submodels. To further elaborate on this, we have performed ablation tests to decouple the contribution of every RCNet to the model accuracies with $\alpha = 0$ and 0.7 (Supporting Note 5). It turns out that when RCNet #2 is eliminated from the model of $\alpha = 0$, an increase in the accuracy of ∼17% is achieved on Dataset #1. However, this value drops to ∼4% with $\alpha = 0.7$, meaning that misclassifications between RCNets #1 and #2 are substantially reduced. Hence, the capture of information on both characteristic and general features of XRD patterns, which not only circumvents the overwhelming bias in favor of specific angle regions but also well accommodates the domain expertise in XRD analysis, can enable efficient optimization of the model for structure-type identification.

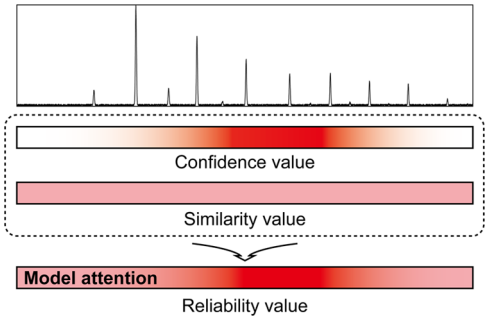


Figure 5. Schematic representation of the combinatorial use of confidence value and similarity value. The color bars reflect the degree of attention.

Compounds Not Belonging to the Popular Structure Types. While CrySTINet adopts an architecture that can conveniently incorporate additional categories, it is essential to know when we need to do so in practical application. This can be translated into a problem of distinguishing between materials in and out of the top 100 most popular structure types. Here, we use a simple screening step according to the highest R-value (R_{max}) outputted by all RCNets for an input XRD pattern. When R_{max} is higher than a threshold (R^0), then

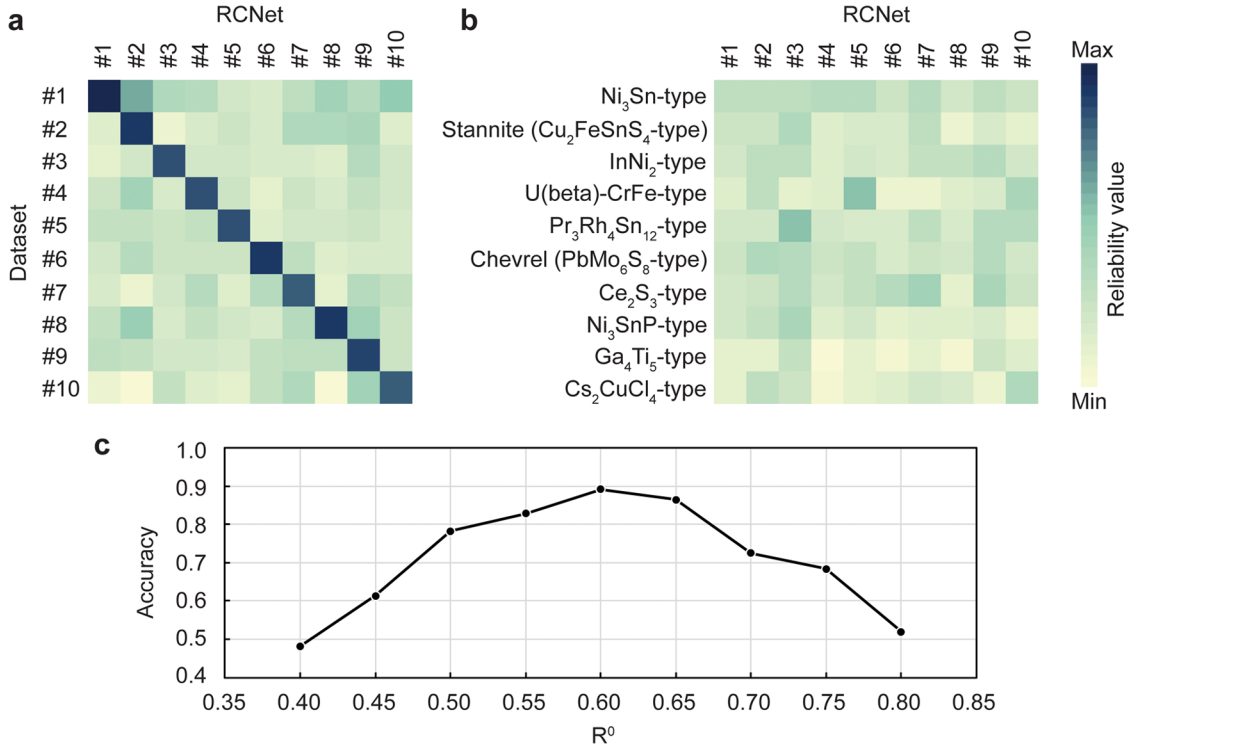


Figure 6. Identification of compounds not belonging to the top 100 most popular structure types. (a) Average reliability values for compounds belonging to the top 100 most popular structure types. (b) Average reliability values for compounds belonging to the top 101−110 most popular structure types. (c) Classification accuracy of CrySTINet for distinguishing compounds not belonging to the top 100 most popular structure types. R^0 corresponds to the threshold in the screening step.

https://doi.org/10.1021/jacs.3c11852
J. Am. Chem. Soc. 2024, 146, 8098−8109

the input can be classified as belonging to the top 100 most popular structure types (in-distribution); otherwise, it is classified as belonging to other structure types (out-of-distribution). Here, we use the No. 101−110 most popular structure types for the test. As shown in Figure 6a,b, the samples belonging to the top 100 most popular structure types provide an average R_{max} much higher than that of the No. 101−110 most popular structure types. The accuracy of binary classification between in-distribution and out-of-distribution samples as a function of R^0 is provided in Figure 6c. A high accuracy of 89.1% is reached at R^0 = 0.6, indicating that this simple screening step is quite effective in distinguishing materials not belonging to the top 100 most popular structure types.

Hence, if an unknown material outputs an R_{max} higher than 0.6, it can be classified as belonging to the top 100 most popular structure types. Then, CrySTINet predicts the exact structure type (among the 100 categories) for this unknown material with an accuracy of 80.0%. This unknown material could be a solid-solution compound that has not been documented yet or a new phase that has a crystal structure closely related to an existing compound among the 63,963 ICSD entries considered in our work. On the other hand, when the R_{max} is lower than 0.6, we need to identify the structure type via a manual pattern matching process. Then, we can decide whether we should train and incorporate an additional RCNet into the framework so that CrySTINet would be able to analyze this kind of material afterward.

DISCUSSION

So far, most of the previous deep-learning models for crystal structure assignment from XRD patterns were confined to a particular compositional space, and once the composition is changed, the entire model has to be restrained, thereby risking a reduction in prediction accuracy. Furthermore, they are often restricted to the classification of existing compounds and can hardly be generalized to compounds outside their databases. Although some recent studies succeeded in symmetry classification in a large chemical space, it is hardly feasible to utilize these prior models for phase identification because the phases/structures of all of the inorganic compounds are much more numerous than their space groups.[56−59] To solve these issues, it is required that the model be extensible according to the need of the task. The ResNet-based structure-type prediction protocol introduced in this work is developed based on the above idea, using a carefully designed parameter, reliability value, to tackle the challenge. The CrySTINet model is much more convenient for retraining than the conventional large-scale CNN models because only the training of a single submodel, i.e., an RCNet, would be required. This convenience does not sacrifice the accuracy too much, as demonstrated in the example of the top 100 most popular structure types in ICSD, covering a compositional space much larger than those of the previous studies concerning autonomous phase identification. These merits make CrySTINet a practical tool for material structure determination through the pipeline of material discovery in self-driving laboratories.

Additionally, with the present model serving as the structure-type recommendation platform, a combination with the Rietveld refinement method, as well as the deep-learning algorithms in previous studies on precise phase identification,[15,36] is therefore advocated to offer more robust XRD interpretation, especially in the analysis of multiphase mixtures. Drawing inspiration from the recent development of multimodal learning,[60,61] another intriguing future direction for autonomous XRD interpretation could be to integrate domain knowledge from multiple information sources including textual,[62−64] image,[65,66] and other types of data into the deep-learning model.

Another thing worth mentioning is that there are a lot of duplicates in ICSD, i.e., similar entries corresponding to a slight difference in stoichiometry, temperature, and neutron/X-ray sources. Duplicates in both training and validation sets will lead to the problem of information leakage, which may overestimate accuracy and result in the poor generalizability of the model. However, removal of the duplicates always requires some parametric thresholds, and the setting of these thresholds is highly subjective and ambiguous, which makes it a challenging task to apply deduplication of ICSD to the field of XRD interpretation since the XRD patterns could be very sensitive to the perturbation of crystal structures. Because our data augmentation process can to some extent reduce the scale of similarity between duplicates, we find that the problem of information leakage in this work is not very serious and that its influence on the analysis of unknown materials is quite limited (Supporting Note 6). Nevertheless, we still concur with the need for future research on a standardized duplicate detection method (probably via machine learning) to solve this thorny problem.

The remaining challenges in deep-learning-based analysis of XRD data pertain to the collection of experimental datasets for fine-tuning of the model. We note that simulated XRD patterns are employed for training in this work and in previous reports, and it is questionable whether these patterns can span the diversity present in real experimental measurements. Although the results here show that a model trained on simulated data can generalize well to experimental data, its performance is still not guaranteed. Moreover, there is also a trade-off between the diversity and uncertainty in experimental data. We believe that an active learning protocol established on an automated experimental data acquisition system would be a prime target for future research on the autonomous deciphering of unknown materials.

As a final remark, we note that while applying CrySTINet for material discovery in self-driving laboratories could give a better chance of identification of new compounds, this model is not capable of confirming whether the unknown material has a new crystallographic structure. CrySTINet can identify a compound that does not belong to specific common structure types as long as the compounds of these structure types have been included in training. However, as it is hardly applicable to incorporate all existing structure types into a single model, we cannot rule out the possibility that the material investigated actually corresponds to an existing ICSD entry with a less common structure type. Nevertheless, CrySTINet could still facilitate material discovery by differentiating compounds that are analogous to existing ones but with different compositions. It can accommodate the cases of compositional disorder (i.e., with the fractional occupancy of crystallographic positions), a facet that has been rarely addressed in previous studies for the goal of autonomous crystal structure assignment, especially in a wide compositional range. Actually, the trouble of a compositional disorder, along with the generally poor prediction accuracy for multiphase samples, has recently raised a debate in the community as to whether artificial intelligence could truly enable material discovery by replacing the conventional

https://doi.org/10.1021/jacs.3c11852
J. Am. Chem. Soc. 2024, 146, 8098−8109

human-operated analysis of characterization data.[67,68] There is no doubt that the material analysis step has become a decisive ingredient for the success of automated laboratories in the area of inorganic synthesis. To further advance this important field, it would be prudent to fund efforts to extend our proposed framework of multiple submodels to the design of deep-learning methods for the unsupervised Rietveld refinement, which is now a major bottleneck to the reliable screening of novel materials in automated laboratories.

CONCLUSIONS

In this work, we introduced a deep-learning model, CrySTINet, that is capable of identifying the structure types for unknown compounds from XRD patterns. This computational tool could be incorporated as a component for XRD interpretation of materials discovered in high-throughput experimentation. Being based on a union of neural networks, CrySTINet is extensible with respect to the collections of structure types employed according to the need, and this extension does not require the retraining of existing networks, making it possible to envision a translational application of the model in various material domains with minimal need for modification. Our results give a working example using the top 100 most popular structure types in ICSD to verify the accuracy achieved by the model. In addition, with the aid of Grad-CAM, we looked into the deep-learning black box and disclosed the essential role of a balance between characteristic and general or, in other words, local and global—features in quantifying the similarity between two patterns, which could reduce the misclassifications resulting from overfocused attention on specific angle regions by the neural networks. We believe that this tactic can spur future studies on deep-learning models for the automated analysis of XRD patterns as well as spectroscopic measurements including infrared spectra, Raman spectra, and nuclear magnetic resonance.

METHODS

Dataset of Simulated XRD Patterns. The information on the structure type was extracted from the crystallographic information files (CIFs) included in ICSD.[47,48] This process was streamlined by a Python program, and the data were stored in an SQLite database. Among the 182,674 inorganic compounds in ICSD, 37,274 entries were excluded here due to the lack of structure-type information in their CIFs. The remaining 145,400 compounds were assigned to 9384 structure types, among which the top 100 most popular structure types were employed to set up the CrySTINet model in our work. This corresponds to a total of 63,963 compounds, accounting for 44% of all compounds. The structure types of the remaining compounds in ICSD are very diverse, with each type mostly having less than 100 entries of inorganic compounds. The representative crystal structures for the top 100 most popular structure types are displayed in Supporting Table 1.

The XRD patterns (in a 2θ range of 5−110°) of all of the 63,963 compounds were simulated according to their structural information in CIFs. To add some experimental complexities to the simulated patterns, data augmentation was performed similarly to a previous study[36] by adding small changes to existing data or creating new synthetic data from existing ones. More specifically, we altered the physical information on crystal structures by imposing strain along each crystal axis, which results in shifts in peak position. We also considered variations of peak width and peak intensity to mimic the effects of grain size and preferred orientation, respectively. Moreover, random noise signals were introduced to the background, which could lead to a slightly ragged curve like that obtained in experiments. After the above procedure, we randomly chose around 10 patterns for each compound. For the less popular structure types (i.e., they are composed of fewer compounds), the number of selected patterns for each compound would be relatively higher so as to reach a better balance in dataset size among different structure types. Finally, the dataset for training and evaluation of our model was augmented to 617,041 simulated XRD patterns.

As the CrySTINet model is composed of 10 RCNets in this study, the top 100 most popular structure types were partitioned into 10 groups, and the compounds in each group were used as the dataset for training one specific RCNet. To increase the difference between any two of the 10 datasets, we relied on the unsupervised clustering according to the average XRD patterns of the 100 structure types. The average XRD pattern for a structure type was calculated by averaging over all of the simulated patterns after data augmentation for the compounds corresponding to this structure type. The vector of the average XRD pattern was downscaled to two dimensions via Uniform Manifold Approximation and Projection (UMAP),[69,70] which is a nonlinear dimensionality reduction algorithm based on stream shape learning. The affinity propagation clustering algorithm[71] was employed to divide the dataset into 10 heaps, as depicted in Supporting Figure 1. From the clustering results as well as a careful inspection of all of the 100 structure types, we found 9 pairs of species that exhibit particular similarity in the crystal structure for each pair, as illustrated in Supporting Figure 2. To attain a higher accuracy in the structure-type identification, the two species in each pair were merged in the classification task during both training and evaluation of the model.

Experimental XRD Patterns. We extracted 80 XRD patterns from the RRUFF database,[50] all of which belong to the top 100 most popular structure types. These experimental patterns exhibit irregular background noises that may obscure or even mask the true diffraction signal, affecting the accuracy of qualitative and quantitative XRD interpretation. To remove the background noises, we used the adaptive iteratively reweighted penalized least-squares method,[72] which can solve the background noise estimation without any user intervention. The patterns were further smoothed via the Savitzky−Golay filtering[73] to minimize the effects of noises and interference while preserving the characteristic peaks. The RRUFF IDs of all of the patterns employed in this work are tabulated in Supporting Table 4.

Model Architecture and Training Details. The CrySTINet model in this work is composed of a union of RCNets, which are based on the ResNet architecture.[46] Each RCNet consists of an initial convolutional layer, 16 residual blocks, and a final fully connected classification layer. Each residual block contains 2 convolutional layers and a pooling layer. The detailed configuration of this neural network is given in Supporting Table 3. The network was optimized using the Adam optimization algorithm[74] as the optimizer, with cross-entropy loss and confidence loss as the loss functions. 1024 XRD patterns were used each time as the minibatch to train a total of 400 epochs with a learning rate of 0.001. The ratio of training, validation, and testing data size was set to 7:1:2 for all of the simulated XRD datasets in this work. The XRD patterns were fed into the RCNets as one-dimensional vectors with a length of 5250, and after passing through the residual blocks, they were fed forward into a prediction module containing a prediction branch and a confidence branch. The prediction branch outputs the predicted category, while the confidence branch outputs the confidence value. The accuracy corresponds to the cases when a new material, which we know belongs to the top 100 most popular structure types, is correctly classified in the right one among these categories. All RCNets were implemented and trained on a laboratory GPU computing cluster, which consists of 62 nodes, each containing an 8-core Intel(R) Xeon(R) CPU E5−2623 v4 and 4 Nvidia GeForce GTX 1080 graphics cards. In terms of software, all codes in this paper are in Python, and the deep-learning model was built based on the Pytorch framework version 1.7.1. Based on the above hardware and software platforms, the average training time for completing 400 epochs is around 100 min.

Class Activation Mapping. Gradient-weighted class activation mapping (Grad-CAM) was used in this work for a visualization of the

https://doi.org/10.1021/jacs.3c11852
J. Am. Chem. Soc. 2024, 146, 8098−8109

weights of the neural network, highlighting the regions of the input patterns to reflect the interest of the RCNets. With Grad-CAM, we can intuitively comprehend the working logic of the neural networks. First, we calculated the gradients of the last convolutional layer flowing into the network to assign the values of importance to each neuron for specific attention decisions. Then, the importance weights of the filters of the last convolutional layer were generated by the global average pooling of the gradients, and each filter was multiplied by its importance weight to generate the Grad-CAM heatmap. The Grad-CAM heatmap for a structure type was obtained by averaging the heatmaps for all of the compounds involved in this class. The resolution of the Grad-CAM heatmap is consistent with the dimension of the input vector of an XRD pattern.

■ ASSOCIATED CONTENT

Data Availability Statement

All codes needed to replicate these results are available in the Supporting Information and from https://github.com/PKUsam2023/CrySTINet.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.3c11852.

Unsupervised clustering for structure types; data augmentation of simulated XRD patterns; distribution of confidence values; optimization of model performance; model performance on experimental dataset; ablation tests on the submodels; and influence of information leakage (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Feng Pan − *School of Advanced Materials, Peking University, Shenzhen Graduate School, Shenzhen 518055, People's Republic of China;* orcid.org/0000-0002-8216-1339; Email: panfeng@pkusz.edu.cn

Shunning Li − *School of Advanced Materials, Peking University, Shenzhen Graduate School, Shenzhen 518055, People's Republic of China;* orcid.org/0000-0002-5381-6025; Email: lisn@pku.edu.cn

Authors

Litao Chen − *School of Advanced Materials, Peking University, Shenzhen Graduate School, Shenzhen 518055, People's Republic of China*

Bingxu Wang − *School of Advanced Materials, Peking University, Shenzhen Graduate School, Shenzhen 518055, People's Republic of China*

Wentao Zhang − *School of Advanced Materials, Peking University, Shenzhen Graduate School, Shenzhen 518055, People's Republic of China*

Shisheng Zheng − *School of Advanced Materials, Peking University, Shenzhen Graduate School, Shenzhen 518055, People's Republic of China*

Zhefeng Chen − *School of Advanced Materials, Peking University, Shenzhen Graduate School, Shenzhen 518055, People's Republic of China*

Mingzheng Zhang − *School of Advanced Materials, Peking University, Shenzhen Graduate School, Shenzhen 518055, People's Republic of China*

Cheng Dong − *School of Advanced Materials, Peking University, Shenzhen Graduate School, Shenzhen 518055, People's Republic of China;* orcid.org/0000-0003-2093-7573

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.3c11852

Author Contributions

†L.C., B.W., and W.Z. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge financial support from the Basic and Applied Basic Research Foundation of Guangdong Province (2023A1515011391, 2020A1515110843), the Soft Science Research Project of Guangdong Province (No. 2017B030301013), the National Natural Science Foundation of China (22109003), and the Major Science and Technology Infrastructure Project of Material Genome Big-science Facilities Platform supported by Municipal Development and Reform Commission of Shenzhen.

■ REFERENCES

(1) Alberi, K.; Nardelli, M. B.; Zakutayev, A.; Mitas, L.; Curtarolo, S.; Jain, A.; Fornari, M.; Marzari, N.; Takeuchi, I.; Green, M. L.; Kanatzidis, M.; Toney, M. F.; Butenko, S.; Meredig, B.; Lany, S.; Kattner, U.; Davydov, A.; Toberer, E. S.; Stevanovic, V.; Walsh, A.; Park, N.-G.; Aspuru-Guzik, A.; Tabor, D. P.; Nelson, J.; Murphy, J.; Setlur, A.; Gregoire, J.; Li, H.; Xiao, R.; Ludwig, A.; Martin, L. W.; Rappe, A. M.; Wei, S.-H.; Perkins, J. The 2019 materials by design roadmap. *J. Phys. D: Appl. Phys.* **2019**, *52* (1), No. 013001.

(2) Altomare, A.; Camalli, M.; Cuocci, C.; Giacovazzo, C.; Moliterni, A.; Rizzi, R. EXPO2009: Structure solution by powder data in direct and reciprocal space. *J. Appl. Crystallogr.* **2009**, *42*, 1197−1202.

(3) Andreev, Y. G.; Lightfoot, P.; Bruce, P. G. Structure of the polymer electrolyte poly(ethylene oxide)3: LiN(SO2CF3)2 determined by powder diffraction using a powerful Monte Carlo approach. *Chem. Commun.* **1996**, No. 18, 2169−2170.

(4) Albesa-Jové, D.; Kariuki, B. M.; Kitchin, S. J.; Grice, L.; Cheung, E. Y.; Harris, K. D. M. Challenges in Direct-Space Structure Determination from Powder Diffraction Data: A Molecular Material with Four Independent Molecules in the Asymmetric Unit. *ChemPhysChem* **2004**, *5* (3), 414−418.

(5) Stanev, V.; Vesselinov, V. V.; Kusne, A. G.; Antoszewski, G.; Takeuchi, I.; Alexandrov, B. S. Unsupervised phase mapping of X-ray diffraction data by nonnegative matrix factorization integrated with custom clustering. *npj Comput. Mater.* **2018**, *4*, No. 43, DOI: 10.1038/s41524-018-0099-2.

(6) Long, C. J.; Bunker, D.; Li, X.; Karen, V. L.; Takeuchi, I. Rapid identification of structural phases in combinatorial thin-film libraries using x-ray diffraction and non-negative matrix factorization. *Rev. Sci. Instrum.* **2009**, *80* (10), No. 103902, DOI: 10.1063/1.3216809.

(7) McCusker, L. B.; Von Dreele, R. B.; Cox, D. E.; Louer, D.; Scardi, P. Rietveld refinement guidelines. *J. Appl. Crystallogr.* **1999**, *32* (1), 36−50.

(8) Rietveld, H. M. The Rietveld Method: A Retrospection. *Z. Kristallogr. - Cryst. Mater.* **2010**, *225* (12), 545−547.

(9) Hase, F.; Roch, L. M.; Aspuru-Guzik, A. Next-Generation Experimentation with Self-Driving Laboratories. *Trends Chem.* **2019**, *1* (3), 282−291.

(10) MacLeod, B. P.; Parlane, F. G. L.; Morrissey, T. D.; Hase, F.; Roch, L. M.; Dettelbach, K. E.; Moreira, R.; Yunker, L. P. E.; Rooney, M. B.; Deeth, J. R.; Lai, V.; Ng, G. J.; Situ, H.; Zhang, R. H.; Elliott, M. S.; Haley, T. H.; Dvorak, D. J.; Aspuru-Guzik, A.; Hein, J. E.; Berlinguette, C. P. Self-driving laboratory for accelerated discovery of thin-film materials. *Sci. Adv.* **2020**, *6* (20), No. eaaz8867, DOI: 10.1126/sciadv.aaz8867.

(11) Abolhasani, M.; Kumacheva, E. The rise of self-driving labs in chemical and materials sciences. *Nat. Synth.* **2023**, *2* (6), 483−492.

(12) Guo, Y.; Liu, Y.; Oerlemans, A.; Lao, S.; Wu, S.; Lew, M. S. Deep learning for visual understanding: A review. *Neurocomputing* **2016**, *187*, 27−48.
(13) Janiesch, C.; Zschech, P.; Heinrich, K. Machine learning and deep learning. *Electron. Mark.* **2021**, *31* (3), 685−695.
(14) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521* (7553), 436−444.
(15) Lee, J. W.; Park, W. B.; Kim, M.; Singh, S. P.; Pyo, M.; Sohn, K. S. A data-driven XRD analysis protocol for phase identification and phase-fraction prediction of multiphase inorganic compounds. *Inorg. Chem. Front.* **2021**, *8* (10), 2492−2504, DOI: 10.1039/D0QI01513J.
(16) Suzuki, Y.; Hino, H.; Hawai, T.; Saito, K.; Kotsugi, M.; Ono, K. Symmetry prediction and knowledge discovery from X-ray diffraction patterns using an interpretable machine learning approach. *Sci. Rep.* **2020**, *10* (1), No. 21790, DOI: 10.1038/s41598-020-77474-4.
(17) Jha, D.; Kusne, A. G.; Al-Bahrani, R.; Nguyen, N.; Liao, W. K.; Choudhary, A.; Agrawal, A. *Peak Area Detection Network for Directly Learning Phase Regions from Raw X-ray Diffraction Patterns*, International Joint Conference on Neural Networks (IJCNN); IEEE: Budapest, HUNGARY, 2019.
(18) Agatonovic-Kustrin, S.; Wu, V.; Rades, T.; Saville, D.; Tucker, I. G. Ranitidine hydrochloride X-ray assay using a neural network. *J. Pharm. Biomed. Anal.* **2000**, *22* (6), 985−992.
(19) Schuetzke, J.; Benedix, A.; Mikut, R.; Reischl, M. Enhancing deep-learning training for phase identification in powder X-ray diffractograms. *IUCrJ.* **2021**, *8* (3), 408−420.
(20) de Castro, P. B.; Terashima, K.; Echevarria, M. G. E.; Takeya, H.; Takano, Y. XERUS: An Open-Source Tool for Quick XRD Phase Identification and Refinement Automation. *Adv. Theory Simul.* **2022**, *5* (5), No. 2100588, DOI: 10.1002/adts.202100588.
(21) Chakraborty, A.; Sharma, R. A deep crystal structure identification system for X-ray diffraction patterns. *Visual Comput.* **2022**, *38* (4), 1275−1282.
(22) Szymanski, N. J.; Bartel, C. J.; Zeng, Y.; Diallo, M.; Kim, H.; Ceder, G. Adaptively driven X-ray diffraction guided by machine learning for autonomous phase identification. *npj Comput. Mater.* **2023**, *9* (1), No. 31, DOI: 10.1038/s41524-023-00984-y.
(23) Venderley, J.; Mallayya, K.; Matty, M.; Krogstad, M.; Ruff, J.; Pleiss, G.; Kishore, V.; Mandrus, D.; Phelan, D.; Poudel, L.; Wilson, A. G.; Weinberger, K.; Upreti, P.; Norman, M.; Rosenkranz, S.; Osborn, R.; Kim, E.-A. Harnessing interpretable and unsupervised machine learning to address big data from modern X-ray diffraction. *Proc. Natl. Acad. Sci. U. S. A.* **2022**, *119* (24), No. e2109665119, DOI: 10.1073/pnas.2109665119.
(24) Surdu, V.-A.; Győrgy, R. X-ray Diffraction Data Analysis by Machine Learning Methods—A Review. *Appl. Sci.* **2023**, *13* (17), No. 9992, DOI: 10.3390/app13179992.
(25) Zhao, X.; Luo, Y.; Liu, J.; Liu, W.; Rosso, K. M.; Guo, X.; Geng, T.; Li, A.; Zhang, X. Machine Learning Automated Analysis of Enormous Synchrotron X-ray Diffraction Datasets. *J. Phys. Chem. C* **2023**, *127* (30), 14830−14838.
(26) Le, N. Q.; Pekala, M.; New, A.; Gienger, E. B.; Chung, C.; Montalbano, T. J.; Pogue, E. A.; Domenico, J.; Stiles, C. D. Deep Learning Models to Identify Common Phases across Material Systems from X-ray Diffraction. *J. Phys. Chem. C* **2023**, *127* (44), 21758−21767.
(27) Chitturi, S. R.; Ratner, D.; Walroth, R. C.; Thampy, V.; Reed, E. J.; Dunne, M.; Tassone, C. J.; Stone, K. H. Automated prediction of lattice parameters from X-ray powder diffraction patterns. *J. Appl. Crystallogr.* **2021**, *54* (6), 1799−1810.
(28) Park, W. B.; Chung, J.; Jung, J.; Sohn, K.; Singh, S. P.; Pyo, M.; Shin, N.; Sohn, K.-S. Classification of crystal structure using a convolutional neural network. *IUCrJ* **2017**, *4* (4), 486−494.
(29) Suzuki, Y.; Hino, H.; Takeichi, Y.; Hawai, T.; Kotsugi, M.; Ono, K. Machine Learning-based Crystal Structure Prediction for X-Ray Microdiffraction. *Microsc. Microanal.* **2018**, *24* (S2), 144−145.
(30) 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* (24), No. 245120.
(31) Zaloga, A. N.; Stanovov, V. V.; Bezrukova, O. E.; Dubinin, P. S.; Yakimov, I. S. Crystal symmetry classification from powder X-ray diffraction patterns using a convolutional neural network. *Mater. Today Commun.* **2020**, *25*, No. 101662.
(32) Ziletti, A.; Kumar, D.; Scheffler, M.; Ghiringhelli, L. M. Insightful classification of crystal structures using deep learning. *Nat. Commun.* **2018**, *9* (1), No. 2775, DOI: 10.1038/s41467-018-05169-6.
(33) Oviedo, F.; Ren, Z.; Sun, S.; Settens, C.; Liu, Z.; Hartono, N. T. P.; Ramasamy, S.; DeCost, B. L.; Tian, S. I. P.; Romano, G.; Gilad Kusne, A.; Buonassisi, T. Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. *npj Comput. Mater.* **2019**, *5* (1), No. 60, DOI: 10.1038/s41524-019-0196-x.
(34) Wang, H.; Xie, Y.; Li, D.; Deng, H.; Zhao, Y.; Xin, M.; Lin, J. Rapid Identification of X-ray Diffraction Patterns Based on Very Limited Data by Interpretable Convolutional Neural Networks. *J. Chem. Inf. Model.* **2020**, *60* (4), 2004−2011.
(35) 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* (1), No. 86, DOI: 10.1038/s41467-019-13749-3.
(36) Szymanski, N. J.; Bartel, C. J.; Zeng, Y.; Tu, Q.; Ceder, G. Probabilistic Deep Learning Approach to Automate the Interpretation of Multi-phase Diffraction Spectra. *Chem. Mater.* **2021**, *33* (11), 4204−4215.
(37) Maffettone, P. M.; Banko, L.; Cui, P.; Lysogorskiy, Y.; Little, M. A.; Olds, D.; Ludwig, A.; Cooper, A. I. Crystallography companion agent for high-throughput materials discovery. *Nat. Comput. Sci.* **2021**, *1* (4), 290−297.
(38) Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. *Learning Deep Features for Discriminative Localization*, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 27−30 June, 2016; pp 2921−2929.
(39) Massuyeau, F.; Broux, T.; Coulet, F.; Demessence, A.; Mesbah, A.; Gautier, R. Perovskite or Not Perovskite? A Deep-Learning Approach to Automatically Identify New Hybrid Perovskites from X-ray Diffraction Patterns. *Adv. Mater.* **2022**, *34* (41), No. 2203879, DOI: 10.1002/adma.202203879.
(40) Parthe, E.; Gelato, L. M. The standardization of inorganic crystal-structure data. *Acta Crystallogr., Sect. A* **1984**, *40* (MAY), 169−183.
(41) Parthe, E.; Gelato, L. M. The best unit-cell for monoclinic structures consistent with b-axis unique and cell choice-1 of International Tables for Crystallography (1983). *Acta Crystallogr., Sect. A* **1985**, *41* (MAR), 142−151.
(42) Bergerhoff, G.; Berndt, M.; Brandenburg, K.; Degen, T. Concerning inorganic crystal structure types. *Acta Crystallogr., Sect. B: Struct. Sci.* **1999**, *55*, 147−156.
(43) Burzlaff, H.; Malinovsky, Y. A procedure for the classification of non-organic crystal structures. 1. Theoretical background. *Acta Crystallogr., Sect. A* **1997**, *53*, 217−224.
(44) Chia-Chun, C.; Herhold, A. B.; Johnson, C. S.; Alivisatos, A. P. Size dependence of structural metastability in semiconductor nanocrystals. *Science* **1997**, *276* (5311), 398−401, DOI: 10.1126/science.276.5311.398.
(45) Xiao, T.; Nagaoka, Y.; Wang, X.; Jiang, T.; LaMontagne, D.; Zhang, Q.; Cao, C.; Diao, X.; Qiu, J.; Lu, Y.; Wang, Z.; Cao, Y. Nanocrystals with metastable high-pressure phases under ambient conditions. *Science* **2022**, *377*, No. 870, DOI: 10.1126/science.abq7684.
(46) He, K.; Zhang, X.; Ren, S.; Sun, J. *Deep Residual Learning for Image Recognition*, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 27−30 June, 2016; pp 770−778.
(47) Bergerhoff, G.; Hundt, R.; Sievers, R.; Brown, I. D. The inorganic crystal structure data base. *J. Chem. Inf. Comput. Sci.* **1983**, *23* (2), 66−69.
(48) Allmann, R.; Hinek, R. The introduction of structure types into the Inorganic Crystal Structure Database ICSD. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **2007**, *63*, 412−417.

(49) Devries, T.; Taylor, G. W. Learning Confidence for Out-of-Distribution Detection in Neural Networks (accessed October 1, 2023), 2018 DOI: 10.48550/arXiv.1802.04865.
(50) Lafuente, B.; Downs, R. T.; Yang, H.; Stone, N. 1. The power of databases: The RRUFF project. In *Highlights in Mineralogical Crystallography*; Armbruster, T.; Danisi, R. M., Eds.; De Gruyter (O): Berlin, München, Boston, 2016.
(51) Montavon, G.; Samek, W.; Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digital Signal Process.* **2018**, *73*, 1−15.
(52) Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. *Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization*, 2017 IEEE International Conference on Computer Vision (ICCV), 22-29 October, 2017; pp 618−626.
(53) Zeiler, M. D.; Fergus, R. *Visualizing and Understanding Convolutional Networks*, Computer Vision − ECCV 2014; Fleet, D.; Pajdla, T.; Schiele, B.; Tuytelaars, T., Eds.; Springer International Publishing: Cham, 2014; pp 818−833.
(54) Pope, P. E.; Kolouri, S.; Rostami, M.; Martin, C. E.; Hoffmann, H. *Explainability Methods for Graph Convolutional Neural Networks*, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 15−20 June, 2019; pp 10764−10773.
(55) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60* (6), 84−90.
(56) Lee, B. D.; Lee, J.-W.; Ahn, J.; Kim, S.; Park, W. B.; Sohn, K.-S. A Deep Learning Approach to Powder X-Ray Diffraction Pattern Analysis: Addressing Generalizability and Perturbation Issues Simultaneously. *Adv. Intell. Syst.* **2023**, *5* (9), No. 2300140, DOI: 10.1002/aisy.202300140.
(57) Dong, R.; Zhao, Y.; Song, Y.; Fu, N.; Omee, S. S.; Dey, S.; Li, Q.; Wei, L.; Hu, J. DeepXRD, a Deep Learning Model for Predicting XRD spectrum from Material Composition. *ACS Appl. Mater. Interfaces* **2022**, *14* (35), 40102−40115.
(58) Schuetzke, J.; Szymanski, N. J.; Reischl, M. Validating neural networks for spectroscopic classification on a universal synthetic dataset. *npj Comput. Mater.* **2023**, *9* (1), No. 100, DOI: 10.1038/s41524-023-01055-y.
(59) Lee, B. D.; Lee, J. W.; Park, W. B.; Park, J.; Cho, M. Y.; Singh, S. P.; Pyo, M.; Sohn, K. S. Powder X-Ray Diffraction Pattern Is All You Need for Machine-Learning-Based Symmetry Identification and Property Prediction. *Adv. Intell. Syst.* **2022**, *4* (7), No. 2200042, DOI: 10.1002/aisy.202200042.
(60) Ektefaie, Y.; Dasoulas, G.; Noori, A.; Farhat, M.; Zitnik, M. Multimodal learning with graphs. *Nat. Mach. Intell.* **2023**, *5* (4), 340−350.
(61) Srivastava, N.; Salakhutdinov, R. Multimodal learning with deep Boltzmann machines. *J. Mach. Learn. Res.* **2014**, *15* (1), 2949−2980, DOI: 10.5555/2627435.2697059.
(62) Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; Lowe, R. Training language models to follow instructions with human feedback (accessed October 1, 2023), 2022 DOI: 10.48550/arXiv.2203.02155.
(63) Kononova, O.; He, T. J.; Huo, H. Y.; Trewartha, A.; Olivetti, E. A.; Ceder, G. Opportunities and challenges of text mining in materials research. *Iscience* **2021**, *24* (3), No. 102155.
(64) Nie, Z. W.; Zheng, S. S.; Liu, Y. J.; Chen, Z. F.; Li, S. N.; Lei, K.; Pan, F. Automating Materials Exploration with a Semantic Knowledge Graph for Li-Ion Battery Cathodes. *Adv. Funct. Mater.* **2022**, *32* (26), No. 2201437, DOI: 10.1002/adfm.202201437.
(65) DeCost, B. L.; Holm, E. A. A computer vision approach for automated analysis and classification of microstructural image data. *Comput. Mater. Sci.* **2015**, *110*, 126−133.
(66) Kaufmann, K.; Zhu, C. Y.; Rosengarten, A. S.; Maryanovsky, D.; Harrington, T. J.; Marin, E.; Vecchio, K. S. Crystal symmetry determination in electron diffraction using machine learning. *Science* **2020**, *367* (6477), 564.
(67) Szymanski, N. J.; Rendy, B.; Fei, Y.; Kumar, R. E.; He, T.; Milsted, D.; McDermott, M. J.; Gallant, M.; Cubuk, E. D.; Merchant, A.; Kim, H.; Jain, A.; Bartel, C. J.; Persson, K.; Zeng, Y.; Ceder, G. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **2023**, *624*, 86−91.
(68) Leeman, J.; Liu, Y.; Stiles, J.; Lee, S. B.; Bhatt, P.; Schoop, L. M.; Palgrave, R. G. Challenges in high-throughput inorganic material prediction and autonomous synthesis (accessed February 9, 2024). 2024 DOI: 10.26434/chemrxiv-2024-5p9j4.
(69) Ghojogh, B.; Crowley, M.; Karray, F.; Ghodsi, A. Uniform Manifold Approximation and Projection (UMAP). In *Elements of Dimensionality Reduction and Manifold Learning*; Ghojogh, B.; Crowley, M.; Karray, F.; Ghodsi, A., Eds.; Springer International Publishing: Cham, 2023; pp 479−497.
(70) McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction (accessed October 1, 2023), 2018 DOI: 10.48550/arXiv.1802.03426.
(71) Frey, B. J.; Dueck, D. Clustering by passing Messages Between Data Points. *Science* **2007**, *315*, 972−976.
(72) Zhang, Z. M.; Chen, S.; Liang, Y. Z. Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst* **2010**, *135* (5), 1138−1146.
(73) Savitzky, A.; Golay, M. J. E. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **1964**, *36* (8), 1627−1639, DOI: 10.1021/ac60214a047.
(74) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization (accessed October 1, 2023), 2014 DOI: 10.48550/arXiv.1412.6980.