

ARTICLE OPEN



Identification of crystal symmetry from noisy diffraction patterns by a shape analysis and deep learning

Leslie Ching Ow Tiong^{1,2}, Jeongrae Kim^{1,2}, Sang Soo Han¹✉ and Donghun Kim¹✉

The robust and automated determination of crystal symmetry is of utmost importance in material characterization and analysis. Recent studies have shown that deep learning (DL) methods can effectively reveal the correlations between X-ray or electron-beam diffraction patterns and crystal symmetry. Despite their promise, most of these studies have been limited to identifying relatively few classes into which a target material may be grouped. On the other hand, the DL-based identification of crystal symmetry suffers from a drastic drop in accuracy for problems involving classification into tens or hundreds of symmetry classes (e.g., up to 230 space groups), severely limiting its practical usage. Here, we demonstrate that a combined approach of shaping diffraction patterns and implementing them in a multistream DenseNet (MSDN) substantially improves the accuracy of classification. Even with an imbalanced dataset of 108,658 individual crystals sampled from 72 space groups, our model achieves $80.12 \pm 0.09\%$ space group classification accuracy, outperforming conventional benchmark models by 17–27 percentage points (%p). The enhancement can be largely attributed to the pattern shaping strategy, through which the subtle changes in patterns between symmetrically close crystal systems (e.g., monoclinic vs. orthorhombic or trigonal vs. hexagonal) are well differentiated. We additionally find that the MSDN architecture is advantageous for capturing patterns in a richer but less redundant manner relative to conventional convolutional neural networks. The proposed protocols in regard to both input descriptor processing and DL architecture enable accurate space group classification and thus improve the practical usage of the DL approach in crystal symmetry identification.

npj Computational Materials (2020)6:196; <https://doi.org/10.1038/s41524-020-00466-5>

INTRODUCTION

High-throughput material synthesis and characterization have been popular topics of research during the past few decades and have accelerated the discovery of novel materials^{1–5}. Although various characterization methods exist, identifying the crystal symmetry, that is, the way the atoms are arranged in space, is inarguably the first and most important process in material characterization. This is because the crystallographic structure of a material plays an important role in determining the material properties (structure–property relationship)^{6,7}. For a concrete example, consider the magnetism of iron: body centred cubic Fe is ferromagnetic, while face centred cubic Fe shows paramagnetic behaviors⁸. The most effective way to classify crystal symmetries is to find the group representing all transformations under which a system is invariant, namely, its space group (SG). In three dimensions, there are 230 distinct types of SGs when chiral copies are considered^{9–11}; these SGs are formed from the combinations of the 32 point groups with the 14 Bravais lattices¹². Manually determining the SG to which a target material belongs is a tedious and highly inefficient task due to the brute-force nature of the search algorithms, which are based on matching diffraction patterns (DPs) to those in a database, such as the Crystallography Open Database or the Inorganic Crystal Structure Database^{6,13–17}. Thus, there is a strong and timely need for robust and automated assessment tools for crystal symmetry determination.

Techniques based on X-ray and electron-beam diffraction are the most related to the identification of crystal symmetries. The latest generation of tools for diffraction experiments allows the simultaneous collection of large volumes of data^{18,19}, the handling of which calls for big data techniques and machine-learning-based approaches. Several recent works have introduced regression

models or deep learning (DL) models for material characterization. Ryan et al.²⁰ used deep neural networks to effectively distinguish chemical elements based on the topology of their crystallographic environment. Liu et al.²¹ refined atomic pair distribution functions in a convolutional neural network (CNN) to classify SGs. For similar purposes, Park et al.²², Vecsei et al.²³, Wang et al.²⁴, Oviedo et al.²⁵, and Aguiar et al.²⁶ used powder X-ray diffraction (XRD) 1D curves, for which information such as peak positions, intensities, and full-widths at half-maximum are mainly treated as the key input descriptors. In addition, Ziletti et al.²⁷ (in a parent work of this study), Aguiar et al.²⁸, Kaufmann et al.²⁹, and Ziatdinov et al.³⁰ developed DL models by extracting features from electron-beam based 2D DPs. These studies clearly show that DL methods can effectively reveal correlations between diffraction data and crystal symmetry. Despite their promise, however, most of these studies have been limited to identifying relatively few classes or crystal systems into which a material can be grouped. DL-based methods of crystal structure determination work perfectly for problems with a small number of symmetry classes (fewer than 10); however, they suffer from a drastic drop in accuracy for more difficult problems involving classification into tens or hundreds of symmetry classes (e.g., up to 230 SGs), severely limiting their practical usage. A DL model that is capable of identifying hundreds of classes with a sufficiently high accuracy will be needed to realize a robust, automated, and ultimately self-driving microscopy system or laboratory^{31–33}.

In this work, considering the limitations imposed by the spotty and noisy distributions of raw DPs, we propose a solution, namely, shaped DPs in a multistream DenseNet (MSDN). Our method greatly enhances the accuracy of SG classification. Even for an imbalanced dataset of 108,658 crystals sampled from 72 SGs, the

¹Computational Science Research Center, Korea Institute of Science and Technology (KIST), Seoul 02792, Republic of Korea. ²These authors contributed equally: Leslie Ching Ow Tiong, Jeongrae Kim. ✉email: sangsoo@kist.re.kr; donghun@kist.re.kr

model achieves $80.12 \pm 0.09\%$, exceeding the performance of benchmark methods by 17–27 percentage points (%p). We find that the shaping strategy enhances the uniqueness of the raw DPs; hence, even small observable differences between raw images of symmetrically close crystal systems (e.g., monoclinic vs. orthorhombic or trigonal vs. hexagonal) become pronounced. In addition, the introduction of the MSDN allows the patterns to be captured in a richer but less redundant manner than is possible in a standard CNN. Owing to their substantial performance enhancements, our proposed methodological protocols show promise for improving the practical usage of DL approaches in crystal symmetry determination.

RESULTS

Shaped DPs in a multistream DenseNet

Raw DPs are spotty and noisy and, thus, difficult to learn from. To enhance the capabilities of DL, we propose two ideas: one is to shape the DPs, and the other is to implement them in a multistream DL network (Fig. 1). The former strategy is to refine the raw DPs by selectively connecting nodes, which transforms them into shaped DPs. One can expect three possible benefits from shaped DPs: (1) the learning objective becomes more solid; (2) by controlling the shaping criteria, it is possible to maximize the uniqueness of each DP; and (3) the added lines may amplify critical information such as lattice parameters (length, angles, etc.). We hypothesize that these benefits will result in improved DL of crystal symmetries.

Shaped DPs are produced as follows. First, raw DPs are collected from three orthogonal zone axes (the x -, y -, and z -axis) in the Condor software with an incident beam wavelength λ of 3.5×10^{-12} m³⁴. The raw DPs are represented as image-based input (or a pixel-based matrix), as illustrated in Fig. 1b, and each pixel is given different values between 0 and 255 in grayscale order from black to white, which is a common method in computer vision theory. The raw DPs (\mathbf{R}^*) are typically composed of several nodes (the position of diffraction peaks) and thus can be represented as $\mathbf{R}^* = \{N_{*,1}, N_{*,2}, \dots, N_{*,n}\}$, where N_* denotes each node, which is simply a collection of multiple pixels; n is the number of nodes; and $*$ denotes each axis. Here, the pixel value threshold for the node identification is 50, which means that only pixels with values ≥ 50 constitute each node. We find that the overall DL performance depends on this threshold for the peak identification, and the threshold of 50 was found to be the optimal value after testing integers between 30 and 100 (Supplementary Fig. 1).

Once nodes are identified, the distances between node pairs are then calculated, that is, $dist_{*,i-j} = d(N_{*,i}, N_{*,j})$, where $N_{*,i}$ and $N_{*,j}$ are two arbitrary nodes and $d(\cdot)$ is the Euclidean distance function. We draw interpolated lines only for node pairs with a distance smaller than a certain threshold, that is, $1.7 \times \min(dist_{*,i-j})$, where $\min(\cdot)$ returns a minimum value after searching over all possible $i-j$ pairs. The overall performance as a function of this threshold (prefactor ranging from 1.3 to 1.9) was also evaluated, and a prefactor of 1.7 was found to be the optimal value (Supplementary Fig. 2). This finding is likely attributed to the fact that the shapes become too complex with a larger threshold value, whereas the shapes are not clearly formed with a smaller threshold value. The colors R, G, and B are used for lines in images of the x -, y -, and z -axis. The red (R) lines are shown in our exemplary schematic (Fig. 1b). Thus, the shaped DP, or \mathbf{S}^* , is calculated as $\mathbf{R}^* + \sum \text{lineplot}(N_{*,i}, N_{*,j})$, where the sum \sum is taken over the selected node pairs and $\text{lineplot}(\cdot)$ is the interpolation function. As shown in the scheme of the DP shaping process (Fig. 1c), the $\text{lineplot}(\cdot)$ function is dependent on the node sizes; as a result, the line thickness will differ for different node pairs. Additional information related to the DP shaping protocols is provided in the “Methods” and in Supplementary Fig. 3. As seen in the examples from several SGs

presented in Fig. 1c and Supplementary Fig. 4, the shaped DPs are more solid and much less noisy than the raw versions. The resulting shapes comprise composition information that describes the particular regions of interest that are useful for representing DPs in more unique manners. In a broader approach, although not attempted in this study, the line drawing process may be performed differently by using a response function rather than a sharp cutoff. Herein, the response function means a functional form of the conversion between node distances and the line strength, which should be a good approach in diversifying the line characteristics (thickness or opacity). Developing the entire shaping protocols within the neural networks can be a better encoding than using the manually crafted shape representations (such as sharp cutoffs), and thus is left for possible future study.

For the further processing of multiple inputs (DPs collected from the three-zone axes), we propose a multistream network, namely, an MSDN, as shown in Fig. 1d. In the MSDN, three substream DenseNets are applied in parallel to each shaped DP; these DenseNets share all of their parameters (weights \mathbf{W} and biases \mathbf{b}). The idea of sharing parameters imposes prior knowledge that the inputs to each substream (\mathbf{S}_{R} , \mathbf{S}_{G} , and \mathbf{S}_{B}) are processed concurrently by the network, which substantially reduces the number of parameters in the MSDN; there is some speedup of the optimization process. In addition, this method warrants different layers in each stream be functionally equivalent after training, which is known to be beneficial to preventing extrapolation biases; that is, the network can adapt better to out-of-domain examples than networks without shared parameters^{35,36}. In addition, the MSDN utilizes the design concept of DenseNet³⁷, in which all layers are densely connected (Fig. 1d); in contrast, in a standard CNN, the features in each convolutional (conv) layer are used as input to the next layer without communication. DenseNet uses a different connectivity pattern by introducing direct connections from any layer to all subsequent layers, which improves the information flow between layers³⁷. Each layer has access to all the preceding feature maps in its block and thus to the network’s collective knowledge. We refer to this network architecture as Dense because all the layers are connected to one another (dense connectivity). The superior performance of DenseNets over standard CNNs has been previously reported in the field of image learning and classification^{37–39}. Likewise, in the present study on the processing of DP images, the proposed MSDN is expected to create rich patterns while maintaining the low complexity of information, thus enabling better classification performance.

The MSDN concurrently accepts and processes shaped DPs, that is, \mathbf{S}_{R} , \mathbf{S}_{G} , and \mathbf{S}_{B} , to extract a better feature representation from each substream for SG classification. Specifically, each layer in each DenseNet receives the inputs from all preceding layers and passes its features to all subsequent layers, meaning that the final output layer has direct supervision over every single layer. As a result, the network offers stronger feature propagation for the extraction of collective knowledge in the inference process. Regarding the network configuration, the MSDN used in this study consists of four dense-block (DB) layers and three transition layers in each substream network, as shown in Fig. 1d and Supplementary Table 1.

Dataset

A large-scale collection of DPs for 108,658 materials sampled from 72 SGs was acquired. These 72 SGs (out of a total of 230) were selected based on the criterion that each group should be represented by at least 295 materials in the Materials Project (MP) library⁴⁰, as shown in Fig. 2a. There are too few materials (mostly < 100) available for the remaining SGs in the MP library, which were therefore excluded for DL training and testing. The selected SGs include 2 triclinic, 12 monoclinic, 22 orthorhombic, 13

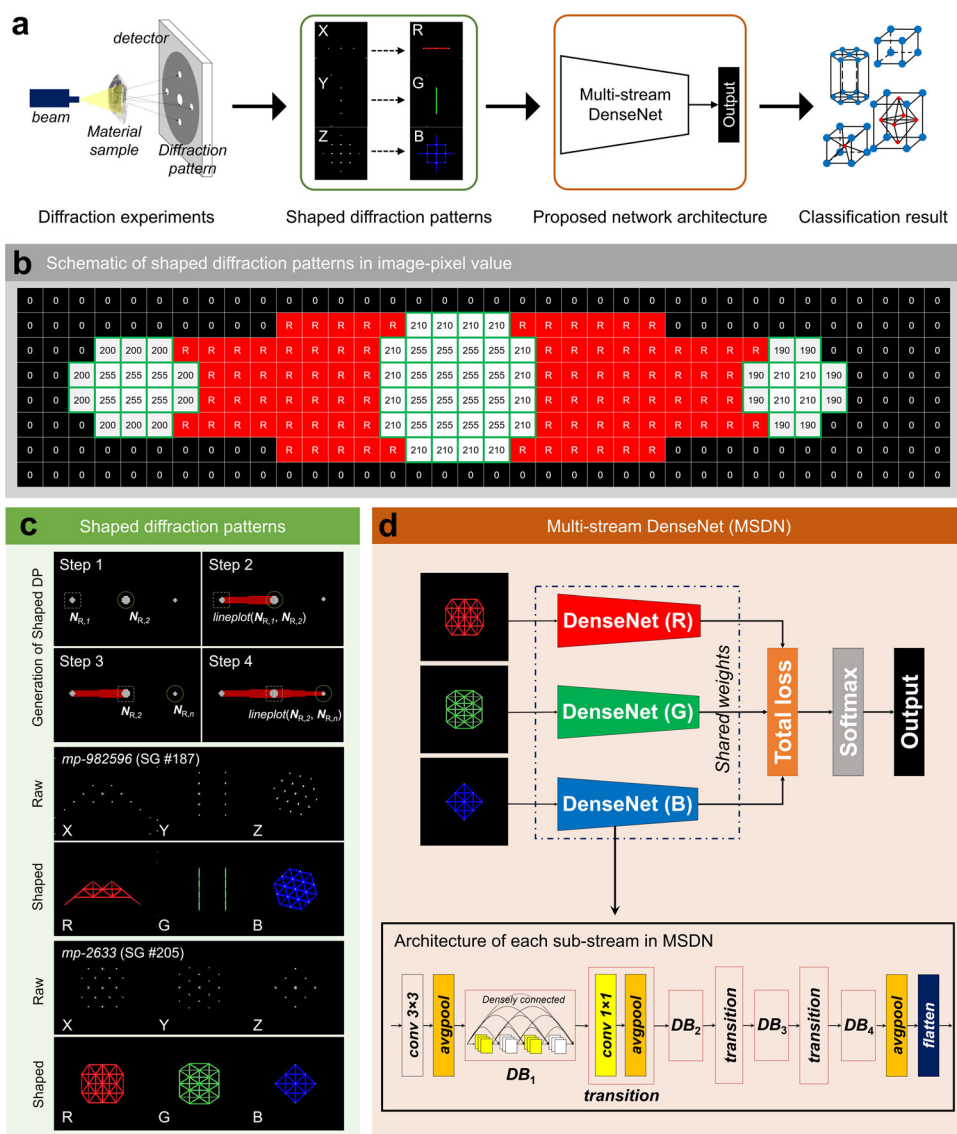


Fig. 1 Shaped diffraction patterns in an MSDN. **a** A scheme that describes the automated determination of crystal symmetry based on diffraction experiments. **b** A scheme describing the visualization of a shaped DP in image-pixel values. **c** A scheme describing the generation process for shaped DPs as well as two exemplary results from space groups #187 and #205. Note that in the generation scheme, the line thickness depends on the node size, which makes the shapes more unique. **d** The network architecture of the MSDN.

tetragonal, 6 trigonal, 8 hexagonal, and 9 cubic crystal systems. Because we downloaded the full list of materials for each SG, the dataset is highly imbalanced, ranging from 295 materials for SG #223 to 8700 materials for SG #14. For the following DL experiments on SG classification, we constructed datasets consisting of 8, 20, 49, and 72 SGs, as shown in Fig. 2b. The number of materials in each SG is tabulated in Supplementary Table 2.

Classification experiments with varying numbers of SGs

We conducted DL experiments to study the classification of SGs (Fig. 3). To evaluate the impact of our strategy (shaped DPs in an MSDN), we performed comparisons with other benchmark models, that is, spot DPs in AlexNet⁴¹, DenseNet³⁷, ResNet⁴², and VGGNet⁴³. Spot DPs, which were originally proposed in the work of Ziletti et al.²⁷, are the superimposed version of the raw DPs from R/G/B color channels. See the scheme in Supplementary Fig. 5 for an exemplary illustration of spot DPs. The key parameter in our experiments was the number of SGs into which materials

could be classified; we considered 8, 20, 49, and 72 (Fig. 2b). In each case, the dataset was divided into the data for learning (training/validation) and the data for testing an 80/20 ratio, with no overlap.

Before explaining the ML test results, we provide evidence that our models are not overfitted. Figure 3a shows the evolutions of the accuracy and total loss during the model's learning process (increasing epochs). The performance of accuracy and total loss show the convergent curves in both the training and cross-validation phases, which indicates that the model is not overfitted. In addition, we performed principal component analysis (PCA) on several SG classes for data visualization (Supplementary Fig. 6). PCA visualization shows that our training-test splitting is random and nonbiased, and that the training and test datasets are not nearest neighbors in the latent space. It is observed that a number of correct prediction datapoints are located far away from training clusters in the latent space. This indicates that the high test accuracy is not a result of the simple memorization of common structures in our training datasets and consequently proves that the model is not overfitted.

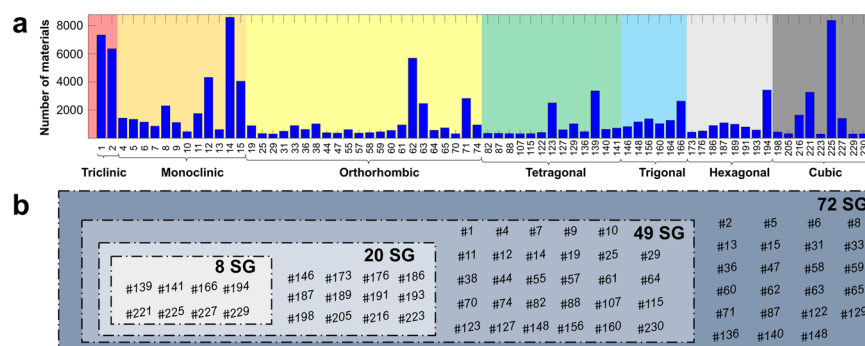


Fig. 2 Population distribution of the diffraction pattern dataset. **a** The number of materials in each space group, along with the crystal system information. The background colors represent seven types of crystal systems: triclinic in red, monoclinic in orange, orthorhombic in yellow, tetragonal in green, trigonal in blue, hexagonal in light gray, and cubic in dark gray. **b** The usage of our dataset for the experiments.

In Fig. 3b and Supplementary Table 3, to begin with the smallest-scale dataset (with eight SGs), both our approach and the other benchmark models work excellently: ours shows $99.04 \pm 0.06\%$ accuracy, while the others also achieve accuracies of above 94.5%. Notably, we have well reproduced the results of the state-of-the-art work of Ziletti et al.²⁷ (over $98.63 \pm 0.09\%$ for 8 SGs), which indicates that our experiments are reliable.

Proceeding to more difficult problems, that is, larger-scale datasets (20, 49, and 72 SGs), we observe that our strategy of shaped DPs in an MSDN performs substantially better than the benchmark models. Note that we performed transfer learning by utilizing the pretrained weights from the 8 SG classification task to retrain the 20 SG, 49 SG, and 72 SG datasets. In Fig. 3b and Supplementary Table 3, our method achieves excellent top-1 classification accuracies of $99.04 \pm 0.06\%$, $95.01 \pm 0.10\%$, $83.16 \pm 0.10\%$, and $80.12 \pm 0.09\%$ on the 8 SG, 20 SG, 49 SG, and 72 SG datasets, respectively. On the other hand, the other models based on spot DPs considerably underperform: even the leading model among the benchmarks (spot DPs in Ziletti et al.'s network) exhibits an accuracy of below 63% for the 72 SG dataset. This result proves the relatively high tolerance of our model to an increasing number of SGs for classification, which is a critical requirement for its practical usage. We additionally measured the performance achieved with shaped DPs in a multistream VGGNet (MSVGG) in order to distinguish the contributions from the “shaped DP” and “MSDN” aspects of the proposed strategy. For the case of the 72 SG dataset, the total enhancement of 17%p can be divided into a 10%p contribution from the shaped DPs and the remaining 7%p of the contribution from the MSDN, confirming that both strategies play critical roles.

Unlike in Fig. 3b, in which only the top-1 classification performance is considered, the top- k ($k = 1 - 5$) ranking accuracy is presented in Figs. 3c–f (for the 8, 20, 49, and 72 SG datasets, respectively). We observe that for all cases, our strategy of shaped DPs in an MSDN performs the best regardless of the k value, followed by shaped DPs in an MSVGG. This once again confirms the superiority of shaped DPs over the conventional spot DPs as the descriptors used for crystal symmetry determination. For the smaller datasets (8 and 20 SGs), the classification is almost perfect (accuracy >99%) even at the top-2 ranking. For the larger datasets (49 and 72 SGs), the accuracy remains above 95% at the top-4 ranking (49 SG dataset) or the top-5 ranking (72 SG dataset).

As an important note, our classification task is different from the standard image classification task due to the hierarchical nature of crystal symmetries. Crystallography is inherently hierarchical and continuous for any real material system^{44,45}. Under enough noise, for example, a slight tetragonal distortion to a cubic structure will be indistinguishable from a standard cubic structure. This situation then requires the probabilistic classifications with the uncertainty quantifications. To overcome the hierarchical nature of crystal

symmetry, we conducted probabilistic classifications for all the tasks in this study using the Monte Carlo (MC) dropout method^{46,47}. This approach allows us to estimate the mean probability and standard deviations over the test datasets. Figure 4 shows several examples of probabilistic classifications. The probabilities are computed via 500 passes of each image with MC dropout active. For many of the misclassified cases (see the red dotted boxes in Fig. 4), the second most likely solution turned out to be the actual SG. We also observe that the confidence in the incorrect predictions is much lower (<50%) than that of the correct prediction cases (generally >80%), and the misclassification seems reasonable based on these probability values.

Data augmentation and transfer learning for highly underrepresented datasets

Our main results are limited in that the tasks did not attempt classifications into all 230 SGs due to the statistically insufficient number of materials (<250) for many classes in the current MP database. To overcome the problem of small data size, we performed data augmentation on some highly underrepresented classes. The new 44 SG datasets are selected based on the criterion that each group is represented by only 100–250 materials in the MP library (Supplementary Table 4). The data augmentation process for these 44 SGs is performed as follows. We replace the constituent elements in available materials with other random elements in a combinatorial manner so that new hypothetical (physically implausible) materials are generated. Figure 5a shows the population distributions of the original vs. generated datasets for the new 44 SGs.

After the data augmentation step was completed, we performed two types of transfer learning using the pretrained weights on the 72 SG dataset: (1) retrain with the original data (without data augmentation), vs. (2) retrain with the original and the generated data (with data augmentation). For both cases, the dataset was divided into the data for learning (training/validation) and the data for testing at an 80/20 ratio, with no overlap. Note that the testing dataset includes only the set of patterns computed from the MP database and are thus physically plausible. Figure 5b, c show the data usage for transfer learning and the resulting DL performances. We achieve 87.5% and 68.5% top-1 accuracies for the model trained with and without data augmentation, respectively, on the new 44 SG datasets. We also conducted testing on a larger-scale dataset that contains all 116 SGs (72 + 44), rather than the new 44 SG datasets themselves. In this test, we achieve 79.1% and 74.9% top-1 accuracies with and without data augmentation, respectively. The data augmentation method is indeed helpful in expanding our study to include a highly underrepresented dataset (in this case, the classes that are represented by much fewer than 250 materials) and substantially improved the classification accuracies by up to 19%p. The success

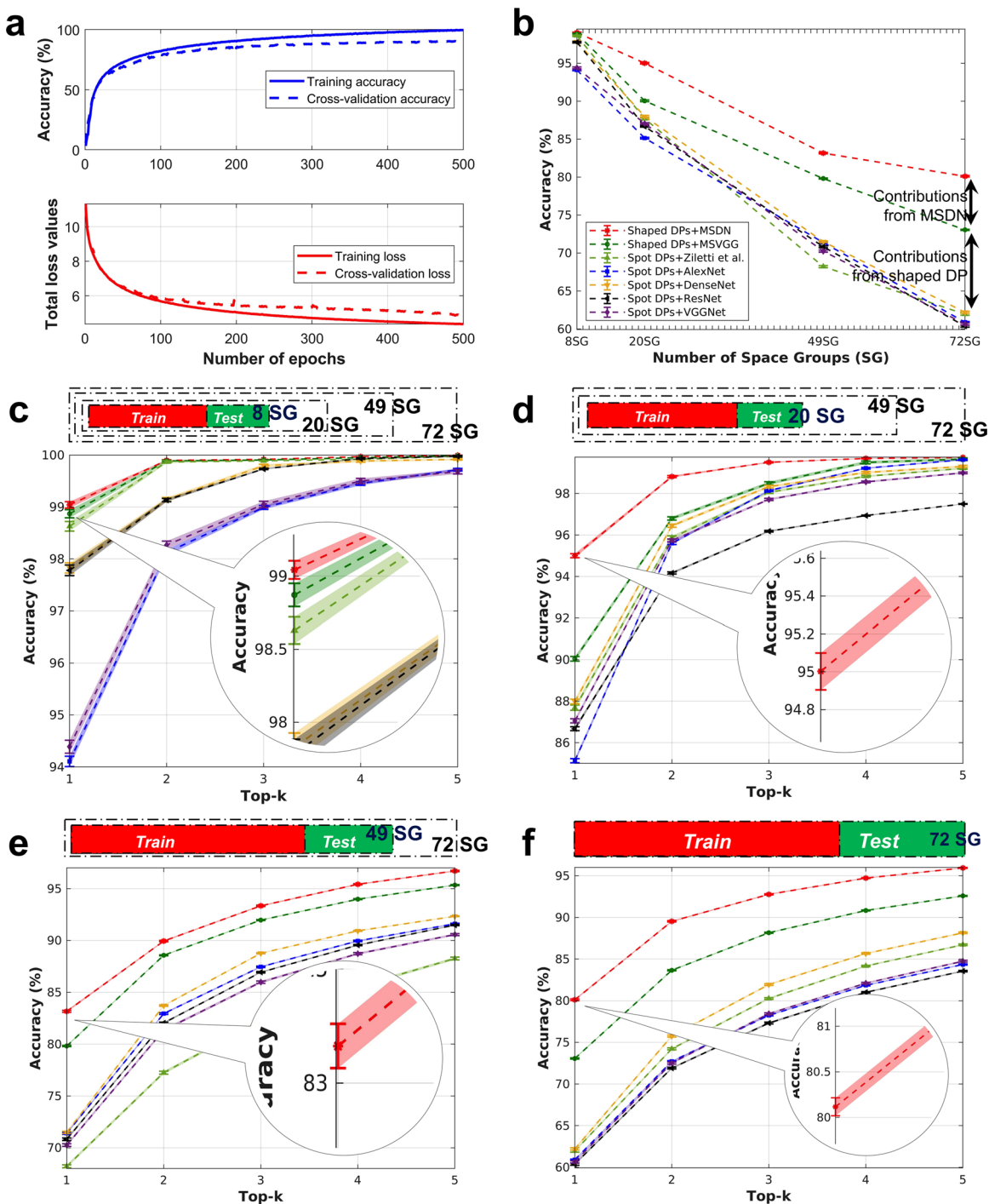


Fig. 3 Space group classification performance. **a** The learning process of MSDN using 72 SG dataset. The classification accuracy and total loss is shown both in training and cross-validation phases. **b** Top-1 accuracy as a function of the number of space groups for classification with standard deviation. **c–f** Top- k accuracies for the datasets consisting of 8 SGs (**c**), 20 SGs (**d**), 49 SGs (**e**), and 72 SGs (**f**). The top- k accuracy refers to the percentage of cases in which the correct class label appears among the top- k probabilities. The circled insets magnify the results of the “shaped DPs + MSDN” model for a clearer vision of error bars.

of transfer learning indicates that the learned features from the pretrained model are generic and powerful, rather than overfitted.

Classification results for individual SGs

We investigated the classification results for individual SGs. Only the 49 SG and 72 SG cases were analyzed (Fig. 6a, b). An interesting observation for both benchmarks and our model is that the accuracy is generally higher for SGs in high-symmetry

crystal systems. The classification process tends to work much better for cubic/hexagonal/trigonal systems than for monoclinic/orthorhombic ones. Triclinic systems are an exception, largely due to the insufficient number of materials belonging to these systems. In Fig. 6c, d, while the benchmarks show the highest accuracy for cubic systems, the accuracy of our model is the highest for trigonal and hexagonal systems rather than cubic systems. In particular, for the 49 SG dataset, it is observed that for

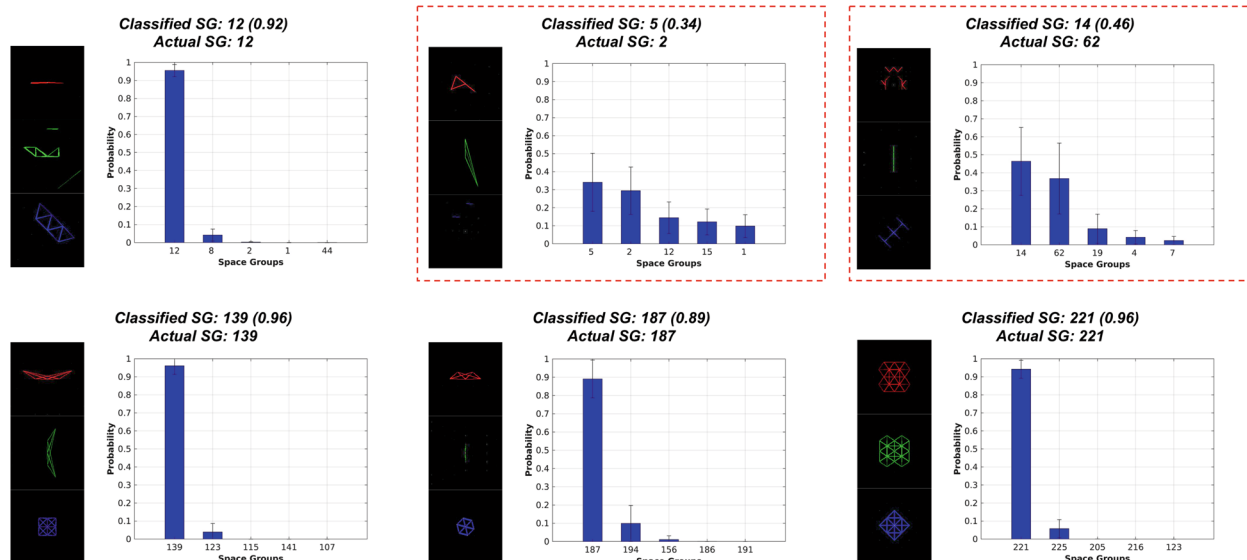


Fig. 4 Probabilistic classification examples. Six examples are randomly selected from the 72 SG test dataset. Each example is shown with the bar graph where the top-5 classification probabilities are shown with the standard deviations. The probabilities are computed via 500 passes of each image with MC dropout active. The probability of the most likely solution is shown in the parentheses in each example. Misclassified images are highlighted by the red dotted box.

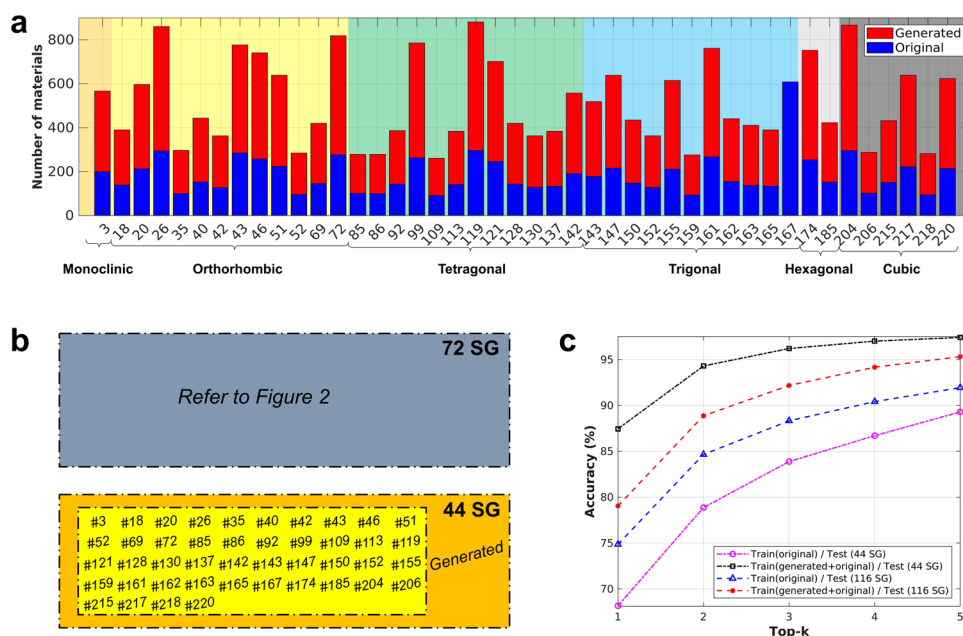


Fig. 5 Transfer learning performances for the new 44 SG dataset. **a** The population distribution of the diffraction pattern dataset for the new 44 SGs. **b** The usage of our dataset for the transfer learning experiments. **c** Transfer learning results with and without the data augmentations.

all SGs corresponding to trigonal and hexagonal systems (#146–#194), the classification accuracy is excellent, being over 90%.

The accuracy improvements in our model over the benchmarks appear to be universal for most SGs. To identify the source of these improvements, we now decompose the contributions for each crystal system (Fig. 6c, d). The model named spot DPs + Ziletti et al. is selected as the representative benchmark here due to its relatively high performance. Triclinic systems are excluded from the analysis due to the statistically insufficient number of materials. The enhancements in accuracy are ranked as follows: trigonal (24.1%p) > monoclinic (19.7%p) > hexagonal (18.1%p) \approx tetragonal (18.11%p) > orthorhombic (13.7%p) > cubic (4.8%p),

where the values in parentheses are the average values for the 49 and 72 SG datasets. The contribution for cubic systems is much smaller than those for the other crystal systems.

Next, we focus on further characterizing the incorrect classifications obtained from the benchmark (spot DPs + Ziletti et al.) and our model (shaped DPs + MSDN). In Fig. 6e, f, for instance, the [monoclinic, orthorhombic] coordinate in the matrices represents the materials belonging to an SG corresponding to a monoclinic system that were incorrectly classified as belonging to an orthorhombic system. In the comparisons between the benchmark and our model, the most prominent changes are observed in two areas, that is, the monoclinic/orthorhombic and trigonal/

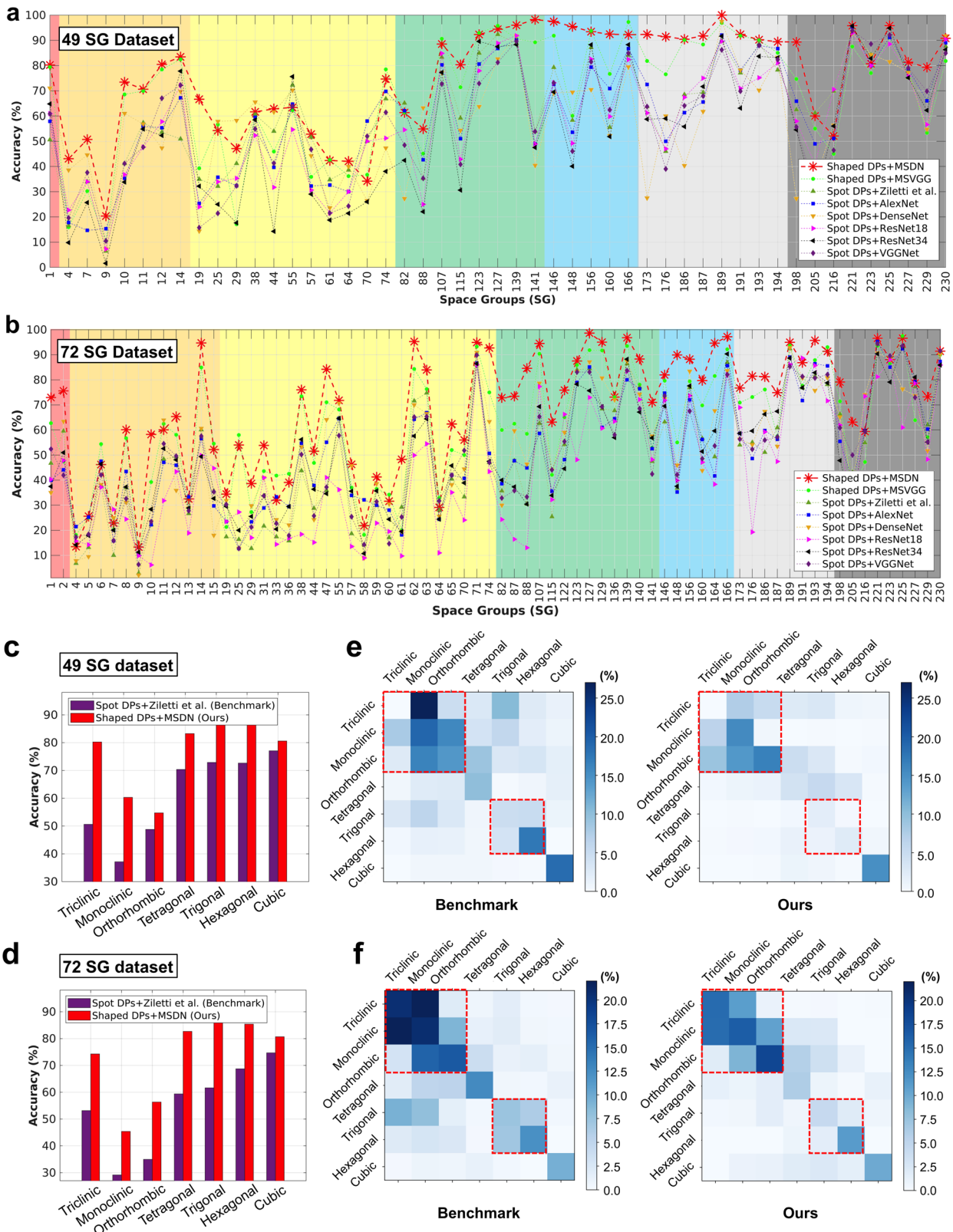


Fig. 6 Decomposition analysis to identify the origins of performance improvement. **a, b** Classification results for individual space groups from the 49 SG (**a**) and 72 SG (**b**) datasets. The background colors represent the seven types of crystal systems, as in Fig. 2a. **c, d** Average classification accuracy by crystal system type for the 49 SG (**c**) and 72 SG (**d**) datasets. **e, f** Matrices showing the distribution rates (%) of incorrect predictions for the 49 SG (**e**) and 72 SG (**f**) datasets. If the rate is, for example, 20% for the [monoclinic, orthorhombic] coordinate in a matrix, this means that 20% of the materials belonging to monoclinic systems in our dataset are incorrectly classified as belonging to SGs corresponding to orthorhombic systems. Red dotted boxes highlight the regions that are considerably different between the benchmark and our model.

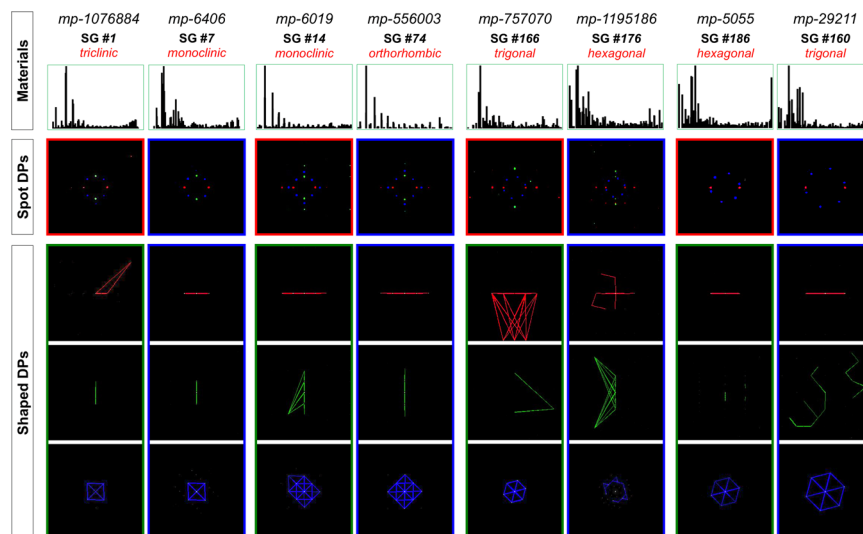


Fig. 7 Case studies in which spot DPs fail and shaped DPs succeed in yielding correct SG classifications. The top row provides the material information of the test samples, which are available in the MP library, including the MP i.d., SG #, and powder X-ray diffraction data. The chemical formula of each material is as follows: mp-1076884 ($\text{Sr}_2\text{Ca}_2\text{Fe}_7\text{CoO}_{20}$), mp-6406 ($\text{Na}_2\text{MgSiO}_4$), mp-6019 (Sr_2YNbO_6), mp-556003 (CaTiO_3), mp-757070 (BaCa_4), mp-1195186 ($\text{RbLa}_2\text{C}_6\text{N}_6\text{O}_{16}$), mp-5055 (Na_6MnS_4), and mp-29211 ($\text{V}_4\text{Cu}_3\text{S}_8$). The next four rows show the spot DPs and shaped DPs of each material. The green and red boxes indicate success and failure cases, respectively, for SG classification, and the blue boxes refer to the reference data in the training set. Best viewed in an electronic version.

hexagonal pairs. This indicates that the benchmark model often finds it difficult to correctly classify SGs corresponding to monoclinic vs. orthorhombic systems or to trigonal vs. hexagonal systems, whereas our model performs much better in resolving this confusion. We speculate that such confusion may occur mainly between symmetrically close crystal systems. For instance, monoclinic and orthorhombic systems are very close in terms of lattice symmetry, differing only in the lattice angle requirements (90° angle requirements). Therefore, similar spot distributions in spot DPs can possibly arise even from materials from different crystal systems, which may undermine the performance of spot-DP-based benchmark models.

To further justify our observation that our model (shaped DPs + MSDN) can largely resolve the confusion between symmetrically close systems, we scrutinize the DPs of several test samples. Figure 7 shows exemplary cases in which spot DPs fail and shaped DPs succeed in yielding correct SG classifications. For the first two example pairs of mp-1076884 (SG #1, triclinic) vs. mp-6406 (SG #7, monoclinic) and mp-6019 (SG #14, monoclinic) vs. mp-556003 (SG #74, orthorhombic), the raw and spot DPs are both too similar (almost identical) to be easily differentiated. This is consistent with the powder XRD data available in the MP library in which the peak locations and intensities are alike. However, the shaped DPs look substantially different, enabling the correct SG classification of these samples. In appearance comparisons of the shaped DPs, we find that the shaped DPs appear more symmetric for the higher-symmetry crystal system, as seen in the R-channel image for the first example pair (triclinic vs. monoclinic) and the G- and B-channel images for the second example pair (monoclinic vs. orthorhombic). The result indicates that the shape analysis can distinguish even small differences (barely observable by human eyes) in node position, size, and brightness, which are likely to be induced by the different level of lattice symmetries of crystal systems.

For the latter two example pairs of mp-757070 (SG #166, trigonal) vs. mp-1195186 (SG #176, hexagonal) and mp-5055 (SG #186, hexagonal) vs. mp-29211 (SG #160, trigonal), although the raw and spot DPs do look slightly different, the benchmark models unfortunately do not predict the correct SGs for these samples. In the shaped DPs, however, these subtle differences are maximized.

Notably, the distance information of adjacent node pairs, which is often related to the lattice parameters, is greatly amplified in the shaped DPs, as observed in the R and B channels of the fourth example pair. From these case studies, we find that the shaping strategy enhances the uniqueness of the raw DPs more than the superimposition strategy used to produce the spot DPs does; hence, even small observable differences in pattern between symmetrically close crystal systems (e.g., monoclinic vs. orthorhombic or trigonal vs. hexagonal) become pronounced.

In addition to the shaping strategy, the MSDN architecture also contributes to performance improvements; here, we would like to discuss the benefits of this network. Figure 8 visualizes both the conv layers from the MSVGG and the DB layers from the MSDN for selected diffraction images. Several additional examples are presented in Supplementary Figs. 7 and 8. The visualization results show that the patterns captured in the MSDN are clearer, richer, and less redundant than those in the MSVGG. Indeed, several feature patterns in the MSVGG are redundant, such as those for samples A, C, and D (highlighted in the red dotted boxes), while such redundant feature patterns are not found in the MSDN. This is likely because the MSDN reuses the features from previous layers to prevent redundancy within the network (Supplementary Fig. 9).

We also compared the computational and memory efficiency of the MSVGG and MSDN. The MSDN is superior to the MSVGG in terms of both space complexity (total number of parameters) and time complexity (FLOPs: floating-point operations per second). The numbers of parameters and FLOPs are 128.85 M and 515.37 M, respectively, for the MSVGG, while they are much smaller at 1.54 M (84 times smaller) and 5.75 M (90 times smaller), respectively, for the MSDN. In fact, the number of parameters of the MSVGG is enormous because every single layer has its own weights and biases (\mathbf{W} and \mathbf{b}) to be learned. In the MSDN, this complexity is avoided by optimizing the parameters and simplifying the connectivity between layers because it is unnecessary to learn redundant feature maps. Such a large difference is possible because the MSDN can receive direct supervision for the propagation of the error signal from the preceding layers to the final layer. These comparisons indicate

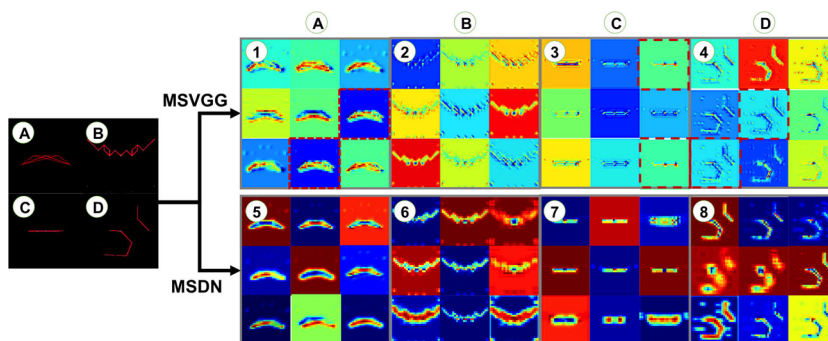


Fig. 8 Benefits of the MSDN over the MSVGG in processing DP images. For selected exemplary diffraction images A, B, C, and D, the block layers of the MSVGG (1, 2, 3, and 4) and MSDN (5, 6, 7, and 8) are visualized. The third conv block of the MSVGG and the DB_2 layer of the MSDN are shown for comparison. The red dotted box indicates redundant (almost identical) feature maps. Best viewed in an electronic version.

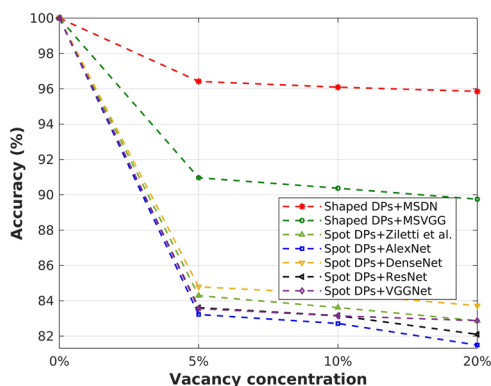


Fig. 9 Robustness of our model to defects. Classification accuracy as a function of vacancy concentrations. The experiment was performed using only test samples that were classified correctly in non-defect cases, which agrees with 100% accuracy for the 0% defect concentration case for all models.

that DP image processing is extremely fast and efficient in our MSDN model.

DISCUSSION

The main results above are limited in that any noise effects (such as defects) that could occur in real-world experiments were not considered. Defects exist everywhere in the form of grain boundaries, dislocations, vacancies, and local inclusions, and may have a large impact on macroscopic properties of materials. We performed additional DL experiments for defective structures to understand whether or not our model is robust to the perturbations such as intrinsic defects (e.g., vacancies). Figure 9 shows the classification accuracy as a function of vacancy concentration. This experiment was performed using only 8800 test samples that were classified correctly in the experiment using 72 SG datasets in Fig. 3f (non-defect case). Vacancies were produced by randomly removing 5%, 10%, and 20% of the atoms in the total system. We observe that our model (shaped DPs + MSDN) maintains a ~96% accuracy compared to the non-defect case even with vacancy concentrations up to 20%, while others exhibit lower accuracies of ~80–84%, which indicates the relatively stronger robustness of our model to defect generations.

This study is entirely focused on simulated data, and thus does not consider other possible defects that likely occur in real-world experimental detector images, including shot-noise, background scattering, parasitic scattering, and other relevant missing data

(intermodule gaps, beamstop, other masked items). Our shaping scheme is simple to apply for simulated data (or high-quality experimental data with strong peaks); however, it might be challenging to apply in the case of real data where noises constitute undesired peaks. We have plans to apply the proposed protocols to experimental data where various methods of image denoising and filtering could be attempted. If the presented strategy also works for real experimental data, the impact of this work will be greatly enhanced.

Finally, we would like to discuss the origins of the limited classification performances. The presented performance is high for <20 SGs; however, it drops to 80.1% for 72 SGs. Throughout the study, we find that it was extremely difficult to reach an accuracy >90% if we have 72 classes or more. We speculate that the main challenge is related to the DPs themselves. Via the analysis of a bulk volume of DPs, many examples were found where similar patterns (almost identical patterns) appear from different SGs, and even from different crystal systems. Four exemplary pairs are shown in Fig. 7. These pairs are mostly found between crystal systems of the nearest symmetry, such as triclinic vs. monoclinic, monoclinic vs. orthorhombic, trigonal vs. hexagonal, and tetragonal vs. cubic. The reason is because, for instance, tetragonal vs. cubic systems differ only in the lattice parameter requirement $a = b \neq c$ vs. $a = b = c$, respectively. Even though such small differences in lattice parameters change the appearance of DPs only slightly, they should be classified into different crystal systems. We believe that the observed similarities between patterns may limit the accuracies in the classification problems involving tens or hundreds of groups. The other challenge is related to the number of materials in each SG (limited experimental data size). Via the transfer learning studies for the highly underrepresented datasets shown in Fig. 5, we concluded that a sufficient number of materials warrants improved classification accuracies. For the new 44 SGs that are represented by a small number of materials (in the range 100–250), we achieve an accuracy of only 68.5% without data augmentation, but a much improved accuracy of 87.5% with data augmentation. Thus, we believe that there is still room for improving the quality of training by using the larger input dataset.

In summary, we propose methodological protocols for enhanced DL-based determination of crystal symmetry, namely, shaped DPs in an MSDN. Our methods greatly improve the SG classification accuracy. Even for an imbalanced dataset of 108,658 crystals sampled from 72 SGs, our approach achieves an accuracy of $80.12 \pm 0.09\%$, outperforming benchmark models based on conventional spot DPs by 17–27%p. Both the shaped DP strategy (~10%p) and the MSDN architecture (~7%p) make considerable contributions to performance improvement. The shaping strategy enhances the uniqueness of the raw DPs; hence, even small observable differences between the raw images of symmetrically

close crystal systems (e.g., monoclinic vs. orthorhombic or trigonal vs. hexagonal) become pronounced in the shaped versions. We additionally find that the MSDN architecture captures the patterns in a richer but less redundant manner than is possible in a standard CNN. This work provides protocols in regard to both input descriptor processing and the DL architecture and, as a result, enables the robust and automated classification of SGs, which we hope will facilitate the practical usage of the DL approach in crystal symmetry determination.

METHODS

Generating and shaping DPs

First, using the MP library⁴⁰, the coordinates of a standard conventional cell are prepared⁴⁸. Next, these are converted into the Protein Data Bank format to satisfy the input-feeding requirement of Condor settings. In the Condor software, a wavelength of $\lambda = 3.5 \times 10^{-12}$ m is used for the incident beam. Three different zone axes (x -, y -, and z -axis) are considered. To produce the shaped DPs, we initialize the first node in \mathbf{R} , which is assigned to $N_{R,i}$, $N_{G,i}$, and $N_{B,i}$ of the dotted i^{th} circle (Supplementary Fig. 3). Then, we detect the neighboring j^{th} node ($N_{R,j}$, $N_{G,j}$, and $N_{B,j}$) and calculate the distance between the i^{th} and j^{th} nodes. For each node pair with a distance smaller than a specified threshold ($1.7 \times \min(\text{dist}_{ij})$), the algorithm will plot a line between the nodes; otherwise, the algorithm will skip this step. For the line colors, red (R), green (G), and blue (B) are used for each x -, y -, and z -axis DP, respectively. After plotting is performed, the shaped DP outcome is created as shown in step K, Supplementary Fig. 3.

DL experiment

For the DL experiments related to Fig. 3b–f, the dataset was divided into 80% of the data for learning (training and validation) and 20% for testing, with no overlap. We then divided the images in the learning set by SG for cross-validation purposes. The cross-validation procedure was designed as follows: (1) randomly shuffle the learning set; (2) split it into ten groups; (3) take one group as the validation set and the remaining groups as the training set; (4) repeat step 3 every 100 epochs and summarize the model evaluation scores. For the testing scheme, the test set images were used to evaluate the performance of our network.

For the proposed model (shaped DPs + MSDN), we used Adam optimizer⁴⁹ with a learning rate of 1.0×10^{-5} and a weight decay and momentum of 1.0×10^{-7} and 0.9, respectively. The MSDN consists of four DB layers and three transition layers in each substream (Fig. 1c). The structure of a dense block is illustrated in Supplementary Fig. 9. Let DB be a dense block with l layers H_l , composed of conv, rectified linear unit and dropout⁵⁰ layers:

$$\text{DB} = H_l([x_0, x_1, \dots, x_{l-1}]), \quad (1)$$

where $x_0 \sim x_{l-1}$ represent feature outputs and $[...]$ is defined as a concatenation operator. Then, a transition layer is implemented in every block that performs 1×1 conv and average pooling operations, where 1×1 conv means that the filter size of the conv layer is 1×1 . Supplementary Table 1 shows the configuration of the proposed network in detail. During training, we defined a total loss (ℓ_{total}) function consisting of a sum of the softmax cross-entropies ℓ of logit vectors and their respective encoded labels, as follows:

$$\ell_{\text{total}} = \ell(F_R) + \ell(F_G) + \ell(F_B), \quad (2)$$

$$\ell(F_*) = - \sum_t^T \sum_c^C L_{tc} \log[\delta_{SG}(F_*)_{tc}], \quad (3)$$

$$\delta_{SG}(F_*)_{tc} = \frac{\exp(F_*)_{tc}}{\sum_c^C \exp(F_*)_{tc}}, \quad (4)$$

where $*$ denotes the zone axis information (one of the color R, G, and B), F is a flatten layer, L denotes the class labels, T is the number of training samples, C is the number of classes, and $\delta_{SG}(\cdot)$ is the output layer, implemented with the softmax function. The ℓ_{total} function provides joint supervision for the training process of the MSDN; it can robustly aggregate the descriptors from the different substreams.

For the alternative model (shaped DPs + MSVGG), we used Adam optimizer with a learning rate of 1.0×10^{-5} and a weight decay and momentum of 1.0×10^{-7} and 0.9, respectively. This network consists of

24 shared conv layers, 15 maxpool layers, and 3 fc layers; more details of the layer configuration are provided in Supplementary Table 5. We again implemented the ℓ_{total} function in Eq. (2) to robustly aggregate the descriptors from the different substreams. For all other benchmark networks, we also used Adam optimizer with a learning rate of 1.0×10^{-4} and a weight decay and momentum of 1.0×10^{-6} and 0.9, respectively.

DATA AVAILABILITY

The data samples of the shaped DP descriptors are shared on the following Zenodo link: <https://doi.org/10.5281/zenodo.4030041>.

CODE AVAILABILITY

The codes for generating shaped DPs and the pretrained model of MSDN are available in the GitHub repository (<https://github.com/tiongleslie/crystal-structure-classification>). All codes are written in Python 3.7 and the architecture of MSDN is implemented using TensorFlow r1.13.

Received: 25 May 2020; Accepted: 18 November 2020;

Published online: 17 December 2020

REFERENCES

- Correa-Baena, J. P. et al. Accelerating materials development via automation, machine learning, and high-performance computing. *Joule* **2**, 1410–1420 (2018).
- 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.* **5**, 70 (2019).
- Lan, Y. et al. Materials genomics methods for high-throughput construction of COFs and targeted synthesis. *Nat. Commun.* **9**, 5274 (2018).
- Mirsaneh, M. et al. High throughput synthesis and characterization of the $\text{Pb}_{(n)}\text{Nb}_{(2)}\text{O}_{(5+n)}$ ($0.5 < n < 4.1$) system on a single chip. *Acta Mater.* **59**, 2201–2209 (2011).
- Kelty, M. L. et al. High-throughput synthesis and characterization of nanocrystalline porphyrinic zirconium metal–organic frameworks. *Chem. Commun.* **52**, 7854–7857 (2016).
- Markvardsen, A. J. et al. ExtSym: a program to aid space-group determination from powder diffraction data. *Appl. Crystallogr.* **41**, 1177–1181 (2008).
- Roy, B., Reddy, M. C. & Hazra, P. Developing the structure–property relationship to design solid state multi-stimuli responsive materials and their potential applications in different fields. *Chem. Sci.* **9**, 3592–3606 (2018).
- Medvedeva, N. I., Van Aken, D. & Medvedeva, J. E. Magnetism in bcc and fcc Fe with carbon and manganese. *J. Phys. Condens. Matter* **22**, 316002 (2010).
- Krivovichev, S. V. Structure description, interpretation and classification in mineralogical crystallography. *Crystallogr. Rev.* **23**, 2–71 (2017).
- Smyth, M. S. & Martin, J. H. J. X Ray crystallography. *J. Clin. Pathol. Mol. Pathol.* **53**, 8–14 (2000).
- Bruno, A. E. et al. Classification of crystallization outcomes using deep convolutional neural networks. *PLoS ONE* **13**, 1–16 (2018).
- Hahn, T. *International Tables for Crystallography* (International Union of Crystallography, 2006).
- Stokes, H. T. & Hatch, D. M. FINDSYM: Program for identifying the space group symmetry of a crystal. *J. Appl. Crystallogr.* **38**, 237–238 (2005).
- Hick, D. et al. AFLOW-SYM: platform for the complete, automatic and self-consistent symmetry analysis of crystals. *Acta Crystallogr. A* **74**, 184–203 (2018).
- Togo, A. & Tanaka, I. Spglib: a software library for crystal symmetry search. Preprint at <https://arxiv.org/abs/1808.01590> (2018).
- Neumann, M. A. X-cell: a novel indexing algorithm for routine tasks and difficult cases. *Appl. Crystallogr.* **36**, 356–365 (2003).
- Coelho, A. A. An indexing algorithm independent of peak position extraction for X-ray powder diffraction patterns research papers. *Appl. Crystallogr.* **50**, 1323–1330 (2017).
- Lo, B. T. W., Ye, L. & Tsang, S. C. E. The contribution of synchrotron X-ray powder diffraction to modern zeolite applications: a mini-review and prospects. *Chem* **4**, 1778–1808 (2018).
- Jesse, S. et al. Big data analytics for scanning transmission electron microscopy ptychography. *Sci. Rep.* **6**, 1–8 (2016).
- Ryan, K., Lengyel, J. & Shatruk, M. Crystal structure prediction via deep learning. *J. Am. Chem. Soc.* **140**, 10158–10168 (2018).

21. Liu, C., Tao, Y., Hsu, D. & Billinge, S. J. L. Using a machine learning approach to determine the space group of a structure from the atomic pair distribution function research papers. *Acta Crystallogr. A* **75**, 633–643 (2019).
22. Park, W. B. et al. Classification of crystal structure using a convolutional neural network. *IUCr Mater. Comput.* **4**, 486–494 (2017).
23. Vecsei, P. M., Choo, K., Chang, J. & Neupert, T. Neural network-based classification of crystal symmetries from X-ray diffraction patterns. *Phys. Rev. B* **99**, 245120 (2018).
24. 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.* <https://doi.org/10.1021/acs.jcim.0c00020> (2020).
25. Oviedo, F. et al. Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. *npj Comput. Mater.* **60**, 1–9 (2019).
26. 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.* **173**, 109409 (2020).
27. Ziletti, A., Kumar, D., Scheffler, M. & Ghiringhelli, L. M. Insightful classification of crystal structures using deep learning. *Nat. Commun.* **9**, 2775 (2018).
28. Aguiar, J. A., Gong, M. L., Unocic, R. R., Tasdizen, T. & Miller, B. D. Decoding crystallography from high-resolution electron imaging and diffraction datasets with deep learning. *Sci. Adv.* **5**, 1–10 (2019).
29. Kaufmann, K. et al. Crystal symmetry determination in electron diffraction using machine learning. *Science* **31**, 564–568 (2020).
30. Ziatdinov, M. et al. Deep learning of atomically resolved scanning transmission electron microscopy images: chemical identification and tracking local transformations. *ACS Nano* **11**, 12742–12752 (2017).
31. Dyck, O., Jesse, S. & Kalinin, S. V. A self-driving microscope and the atomic forge. *Mater. Matters* **44**, 669–670 (2019).
32. Coley, C. W. et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **365**, eaax1566 (2019).
33. MacLeod, B. P. et al. Self-driving laboratory for accelerated discovery of thin-film materials. *Sci. Adv.* **6**, 20, <https://doi.org/10.1126/sciadv.aaz8867> (2019).
34. Hantke, M. F., Ekeberg, T. & Maia, F. R. N. C. Condor: a simulation tool for flash X-ray imaging. *Appl. Crystallogr.* **49**, 1356–1362 (2016).
35. Tiong, L. C. O., Lee, Y. & Teoh, A. B. J. Periocular recognition in the wild: Implementation of RGB-OCLBCP dual-stream CNN. *Appl. Sci.* **9**, 1–17 (2019).
36. Savarese, P. & Maire, M. Learning implicitly recurrent CNNs through parameter sharing. In *International Conference on Learning Representations (ICLR)*, 1–15 (2019).
37. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 4700–4708 (IEEE, 2017).
38. Zhang, H., Guo, Y., Wang, X., Yuan, J. & Ding, Q. Multiple Feature Reweight DenseNet for image classification. *IEEE Access* **7**, 9872–9880 (2019).
39. Tao, Y., Xu, M., Lu, Z. & Zhong, Y. DenseNet-based depth-width double reinforced deep learning neural network for high-resolution remote sensing image per-pixel classification. *Remote Sens.* **10**, 779 (2018).
40. Persson, K. Materials Project. <https://materialsproject.org> (2020).
41. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Twenty-sixth Conference on Neural Information Processing Systems*, 1097–1105 (2012).
42. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778 (IEEE, 2015).
43. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)* (eds Bengio, Y. & LeCun, Y.), 1–14 (2015).
44. Moeck, P. Towards generalized noise-level dependent crystallographic symmetry classifications of more or less periodic crystal patterns. *Symmetry* **10**, 133 (2018).
45. Moeck, P. On classification approaches for crystallographic symmetries of noisy 2D periodic patterns. *IEEE Trans. Nanotechnol.* **18**, 1166–1173 (2019).
46. Gal, Y. & Ghahramani, Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning, PMLR* (eds Balcan, M. F. & Weinberger, K. Q.), Vol. 48, 1050–1059 (JMLR.org, University of Cambridge, 2016).
47. Vasudevan, R. K. et al. Mapping mesoscopic phase evolution during E-beam induced transformations via deep learning of atomically resolved images. *npj Comput. Mater.* **4**, 1–9 (2018).
48. Setyawan, W. & Curtarolo, S. High-throughput electronic band structure calculations: challenges and tools. *Comput. Mater. Sci.* **49**, 299–312 (2010).
49. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *International Conference on Learning Representations (ICLR)* (eds Bengio, Y. & LeCun, Y.), 1–13 (2015).
50. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn.* **15**, 1929–1958 (2014).

ACKNOWLEDGEMENTS

This work was supported by the Samsung Research Funding and Incubation Center of Samsung Electronics under Project Number SRFC-MA1801-03.

AUTHOR CONTRIBUTIONS

D.K. and S.S.H. conceived and designed the research. L.C.O.T. and J.K. performed the research, including data collection and deep learning tasks. All authors contributed to manuscript writing.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at <https://doi.org/10.1038/s41524-020-00466-5>.

Correspondence and requests for materials should be addressed to S.S.H. or D.K.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020