Robustness test of the *spacegroupMining* model for determining space groups from atomic pair distribution function data

Ling Lan,[a] Chia-Hao Liu,[a]* Qiang Du[a,b]* and Simon J. L. Billinge [a,c]*

[a]*Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027, United States*, [b]*Data Science Institute, Columbia University, New York, NY 10027, United States*, and [c]*Condensed Matter Physics and Materials Science Department, Brookhaven National Laboratory, Uptown, NY 11973, United States*. *E-mail: sb2896@columbia.edu*

Robustness Test, Machine Learning, Data Mining, Space Group, Pair Distribution Function

Abstract

Machine learning models based on convolutional neural networks have been used for predicting space groups of crystal structures from their atomic pair distribution function (PDF). However, the PDFs used to train the model are calculated with using a fixed set of parameters that reflect specific experimental conditions, and the accuracy of the model when given PDFs generated with different choices of these parameters is unknown. In this paper, we report that the results of the top-1 accuracy and top-6 accuracy are robust when applied to PDFs of different choices of experimental parameters $r_{\max}$, $Q_{\max}$, Q_{damp} and atomic displacement parameters.

1. Introduction

Recently it was shown (Liu *et al.*, 2019) that a convolutional neural network (CNN) machine learning model could predict the space group of a material from its atomic

pair distribution function (PDF) (Billinge, S. J. L., 2019; Egami & Billinge, 2012) with good accuracy. This model is called SPACEGROUPMINING and was recently deployed as a web application on the pdfitc.org website (Yang *et al.*, 2021).

The atomic pair distribution function (PDF) method is a total scattering technique for determining local order in nanostructured materials. Theoretically, the PDF gives the scaled probability of finding two atoms in a material a distance r apart and is related to the density of atom pairs in the material (Billinge, S. J. L., 2019; Egami & Billinge, 2012).

The model of (Liu *et al.*, 2019) was trained, as shown in the red section in Figure 1, using calculated PDFs, denoted by $G(r, \Omega)$ here, where Ω indicates the set of parameters that define experimental details of the measurement. These model experimental parameters that affect the quality of the data, such as the maximum range of Q, $Q_{\max}$, where Q is the modulus of the scattering vector (this affects peak widths in the PDF), the r-range of the input PDF, $r_{\max}$, and so on (Proffen & Billinge, 1999). These are listed in full in Table 1. Although a specific set of values were used to train the model, in general, different parameter values might be relevant in a scientist's measured PDFs. We denote these as $G(r, \Omega')$, where the prime on the Ω indicates that some experimental values in the set are different from the ones used in the training. A natural and important question is whether the trained model with $G(r, \Omega)$ could provide reasonable predictions on materials associated with $G(r, \Omega')$. If the accuracy of the model predictions on $G(r, \Omega')$ is close to its performance on the PDFs, $G(r, \Omega)$, that the model learned from, we believe that the model is robust. To be more explicit, this robustness test investigates how input data quality translates to performance, while the input data distribution does not shift, i.e., the materials we use to train and test are not varying. In this paper we assess how well the model performs when it is tested on PDFs that were calculated using experimental parameters different from

those for the training set (blue section in Figure 1). We conclude that overall it performs quite well with respect to $r_{\max}$, $Q_{\max}$, Q_{damp} and atomic displacement factor (ADP), or U_{iso}, of the measurement, hence providing evidence to the robustness of the CNN machine learning model developed in (Liu *et al.*, 2019).

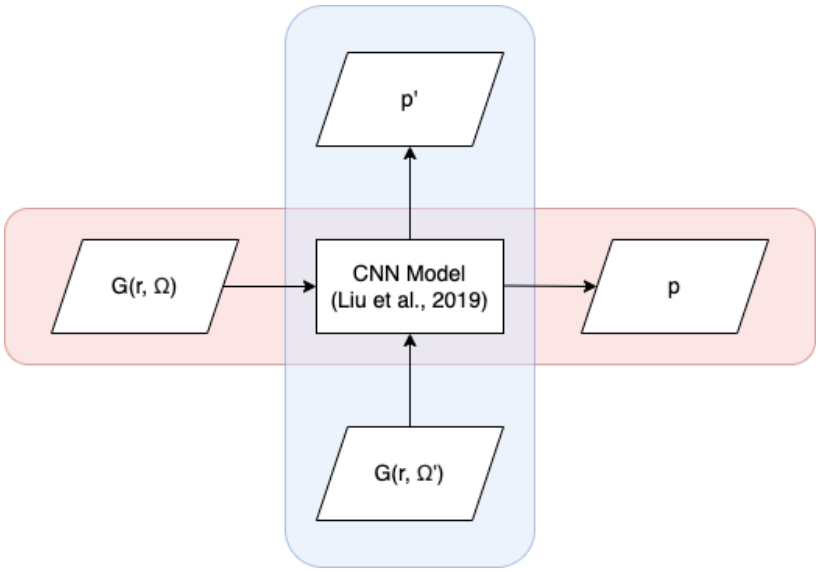

Fig. 1. $G(r, \Omega)$ are the calculated PDFs w.r.t. the original experimental parameters used to train the CNN model, while $G(r, \Omega')$ are PDFs with varying parameters. p and p' are the corresponding model outputs from which we could measure the accuracy of the model predictions. The red (blue) section depicts the training (testing) process respectively.

2. Method

2.1. Data and Model

Our main objective is to test the robustness of the originally trained model. However, we are not able to identify the exact datasets that constituted the training set in the original training. In order to avoid testing robustness by inadvertently using a dataset that might be part of the original model's training data, our first step is to rebuild the model again.

The input PDF data are calculated from $98,830$ structures in the 45 most heavily

represented space groups in the ICSD (Belsky *et al.*, 2002) structural database. The PDFs are calculated from crystallographic information framework (CIF) (Hall *et al.*, 1991) files obtained from ICSD using the diffpy-cmi (Juhás *et al.*, 2015) package with parameters Ω defined in Table 1. The parameters are the same as the ones used in (Liu *et al.*, 2019), except that the grid size is $\pi/40$ in our experiment (the paper used $\pi/23$), so that we could calculate the PDF with higher $Q_{\max}$'s. 80% of the data is considered as training samples, and the rest is treated as test samples. The choice of $r_{\min}$, $r_{\max}$, and r_{grid} in Table 1 discretizes the input PDFs to 1D signal sequences of dimension $209 \times 1 \times 1$. We further normalize the PDF input, $G(r, \Omega)$, to ensure that it lies between 0 and 1 for each entry.

Table 1. *Experimental parameters used to calculate the PDFs used for training, validation and testing of the model. Here Q_{min} and Q_{max} define the range of Q that was included in the Fourier transform to obtain the PDF (Q is the modulus of the scattering vector), r_{min} and r_{max} are the minimum and maximum range of r of the PDF and r_{grid} was the size of the bins in the numerical determination of G. U_{iso} is the atomic displacement factor of the atoms. Q_{damp} and Q_{broad} are parameters that determine the r-dependence of the damping of the PDF signal, mostly due to the Q-space resolution of the measurement. All these parameters are defined in detail in (Egami & Billinge, 2012; Proffen & Billinge, 1999).*

$r_{\min}$ (Å)	1.5
$r_{\max}$ (Å)	30.0
r_{grid} (Å)	$\pi/40$
$Q_{\min}$ (Å^{-1})	0.5
$Q_{\max}$ (Å^{-1})	23.0
U_{iso} (Å^2)	0.008
Q_{damp} (Å^{-1})	0.04
Q_{broad} (Å^{-1})	0.01

To rebuild the model, we use the architecture based on the convolutional neural network (CNN) used in Liu's paper (Liu *et al.*, 2019). The output, p, of the model is a 45×1 vector, which represents the probability of the input PDF being in each of the 45 space groups considered in our study. We use weighted categorical cross entropy loss,

$$\text{Loss} = -\sum_{i=1}^{45} w_i \cdot p_{true_i} \cdot \log p_i, \tag{1}$$

to mitigate the effects of unbalanced data, where the weight w_i is defined as the number of structures in the training set over the number of structures of each space group in the training dataset. Adaptive moment estimation (Adam) with a mini-batch size of 64 is used to train the model. Furthermore, we modify the learning rate as an exponential decay, $l = 5 \times 10^{-4} e^{-0.025 \times \text{epoch}}$. The model is trained using Keras on a single Nvidia Tesla P100 GPU.

An accuracy of 67.7% from top-1 prediction and 90.2% from top-6 predictions is achieved. The performance of our reconstructed model is similar to the one shown in the original paper, which was 70.0% top-1 accuracy and 91.9% top-6 accuracy. The model rebuilt here is used, without any further retraining, in subsequent robustness tests on datasets involving PDFs having different parameter values, as illustrated schematically in Fig. 1.

2.2. Robustness Test

In order to test the robustness, we consider the four experimental parameters that are used to calculate the PDFs from the structural CIFs, which are Q_{max}, r_{max}, Q_{damp}, and U_{iso}. The other parameters in Table 1 are not expected to affect the accuracy greatly and were not explicitly tested. Variations in Q_{min} produce no effect until low-Q Bragg peaks are lost and then result in long-wavelength damped sinusoidal oscillations in the PDF that appear like an oscillating background to the signal. The data are interpolated onto a different r_{grid} during the process and so the user r_{grid} will not affect the outcome, and most users are expected to have data calculated to an arbitrarily small r_{min}. Finally, Q_{broad} has a very small effect on the width of peaks in rather high r-regions that are unlikely to be uploaded. The effects of Q_{damp} are to damp the signal at high-r and are expected to have a similar effect to reducing r_{max}; however, this is an important parameter in general as it can also approximate

the finite size of crystallites in a measurement and we expect users to want to apply the approach to the nanoparticulate data so we have run robustness tests on Q_{damp}.

To carry out the tests we randomly choose structures from the testing set (10% of the testing samples are chosen), and compute their PDFs while varying each of these parameters separately between limits that are chosen to bracket values that are experimentally reasonable. These calculated PDFs are then given to the trained model, without model retraining despite the changes in parameter values of the input PDFs, to predict the space group, and the model accuracy is computed as a function of the experimental parameter value.

First we consider the robustness against a variation in r_{max}. The model was trained with an r_{max} of 30 Å and we want to test its performance when given PDFs computed (or measured) over a narrower r-range, from 10 Å to 30 Å every 2 Å. Variations in r_{max} will change the length of the PDF vector, which is not allowed in our model. Since we are only considering r-ranges that are shorter than 30 Å, to keep the dimension of all input PDFs consistent, the data are padded with zero's up to the value of $r_{\mathrm{max}} = 30$ Å before being interpolated on to the $209 \times 1 \times 1$ grid using quadratic interpolation.

To test the Q_{max} sensitivity, computed PDFs in the range of $12 \leq Q_{\mathrm{max}} \leq 30$ Å^{-1} in steps of 3 Å^{-1} were tested against the trained model. For the Q_{damp}, we tested on computed PDFs in the range of $0 \leq Q_{\mathrm{damp}} \leq 0.15$ Å^{-1} in steps of 0.03 Å^{-1}. Finally for the ADP, U_{iso}, from $0.005 \leq U_{\mathrm{iso}} \leq 0.01$ Å^2 in steps of 0.001 Å^2, where the model was trained on values $Q_{\mathrm{max}} = 23$ Å^{-1}, $Q_{\mathrm{damp}} = 0.04$ Å^{-1} and $U_{\mathrm{iso}} = 0.008$ Å^2, respectively.

3. Results

3.1. Robustness with respect to r_{max}

Figure 2 shows the top-6 accuracy against a variation in $r_{\max}$ from 10 Å to 30 Å. When $r_{\max}$ is larger than 20 Å, top-6 accuracy is always above 87.1%, which is close to the optimal value of 90.2%. It is recommended to give the model a PDF with a $r_{max} \geq 30$ Å. However, the robustness test shows that if the signal is from data over a narrower range, such as a nanoparticle whose signal dies on a shorter length-scale, the model can still be categorized into space group with reasonably good accuracy, though the performance drops off more quickly below an $r_{\max}$ of 20 Å or so.

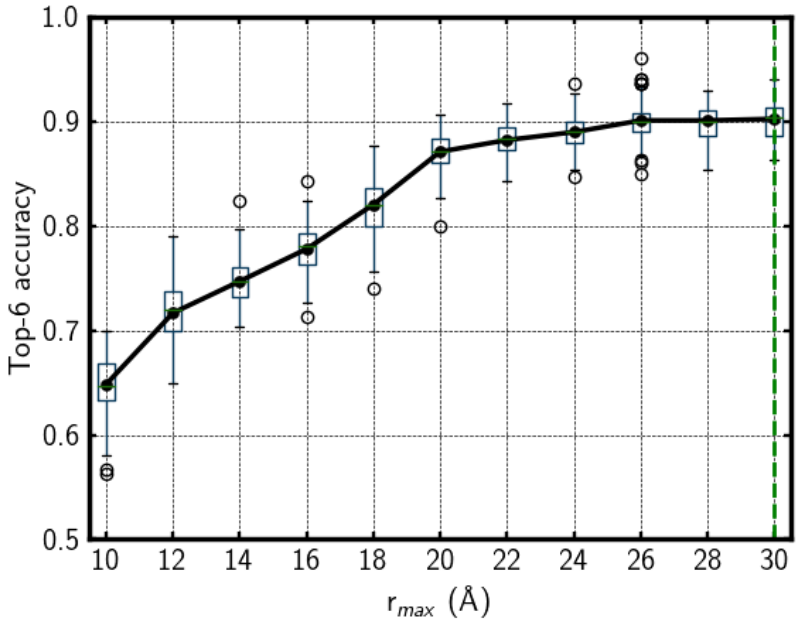

Fig. 2. The black dots represent the top-6 accuracy as $r_{\max}$ is varied. The value of $r_{\max}$ = 30 Å used to train the model is shown as a vertical green dashed line. The box plot at each $r_{\max}$ value shows the uncertainties on the top-6 accuracy (see text for details). The box extends from the first quartile (Q1) to the third quartile (Q3) of the distribution, with a green horizontal line at the median. The whiskers from the box extend by 1.5 times the inter-quartile range (IQR). If the maximum/minimum of the dataset is smaller/larger than median$\pm$1.5$\times$IQR, then the whiskers extend to the maximum/minimum. The hollow circles are those past the end of the whiskers.

3.2. Robustness with respect to Q_{max}

Next we consider the robustness of the model when PDFs are generated using different $Q_{\max}$ values. As shown in Figure 3, when $Q_{\max}$ is larger than 18 Å^{-1}, top-6 accuracy is above 81.1%. The bump around 23 Å^{-1} makes sense, as the model favors the $Q_{\max}$ value that it is trained on. But the performance with $Q_{\max}$ values deviated from 23 Å^{-1} is still fairly good over the entire range of values considered, the accuracy never falls below 77.7%, and so the model is quite robust against variations in $Q_{\max}$.

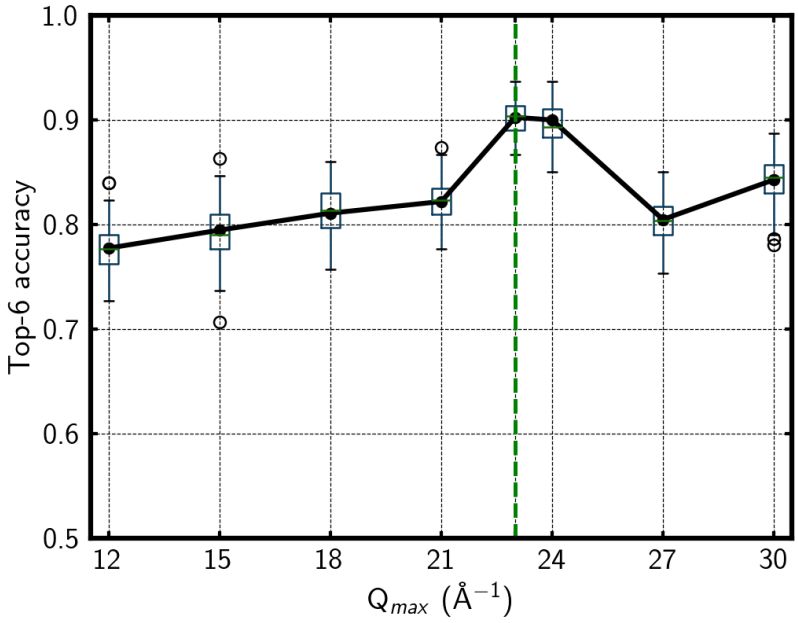

Fig. 3. The black dots represent the top-6 accuracy as $Q_{\max}$ is varied. The value of $Q_{\max} = 23$ Å^{-1} used to train the model is shown as a vertical green dashed line. The box plot at each $Q_{\max}$ value is plotted in the same way as the ones in Figure 2.

3.3. Robustness with respect to Q_{damp}

Figure 4 shows the top-6 accuracy against a variation in Q_{damp} in the range of $0 \leq Q_{\text{damp}} \leq 0.15$ Å^{-1} in steps of 0.03 Å^{-1}.

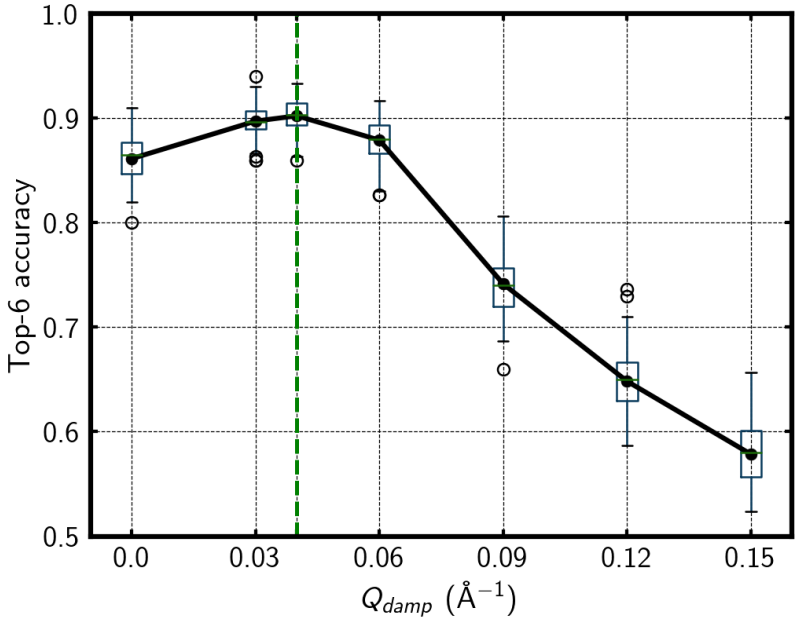

Fig. 4. The black dots represent the top-6 accuracy as Q_{damp} is varied. The value of $Q_{\mathrm{damp}} = 0.04$ Å^{-1} used to train the model is shown as a vertical green dashed line. The box plot at each Q_{damp} value is plotted in the same way as the ones in Figure 2.

When Q_{damp} is smaller than 0.06 Å^{-1}, the top-6 accuracy is always above 86.1%. However, the performance drops off fairly quickly above $Q_{\mathrm{damp}} = 0.06$ Å^{-1} or so. When $Q_{\mathrm{damp}} = 0.15$ Å^{-1}, the PDF signal practically vanishes in the region above $r = 20$ Åand so we might expect the accuracy to be similar to that of $r_{\mathrm{max}} = 20$ Å. We find that the accuracy of $Q_{\mathrm{damp}} = 0.15$ Å^{-1} falls to 57.8%, which is significantly lower than the value of 87.1% of $r_{\mathrm{max}} = 20$ Å. This is presumably because Q_{damp} damps the signal progressively over the entire signal and therefore the model is more sensitive to Q_{damp} variations. However, we note that the accuracy with Q_{damp} values deviated from 0.06 Å^{-1} never falls below 57.8%, which can still give acceptable results in many cases.

3.4. Robustness with respect to Atomic Displacement Parameter, U_{iso}

Finally (Fig. 5), we consider robustness against variations in $U_{\rm iso}$. The results are even less sensitive to the choice of ADP. When $U_{\rm iso}$ of the PDFs were in the range 0.005 Å^2 to 0.01 Å^2, the top-6 accuracy is always above 87.3%.

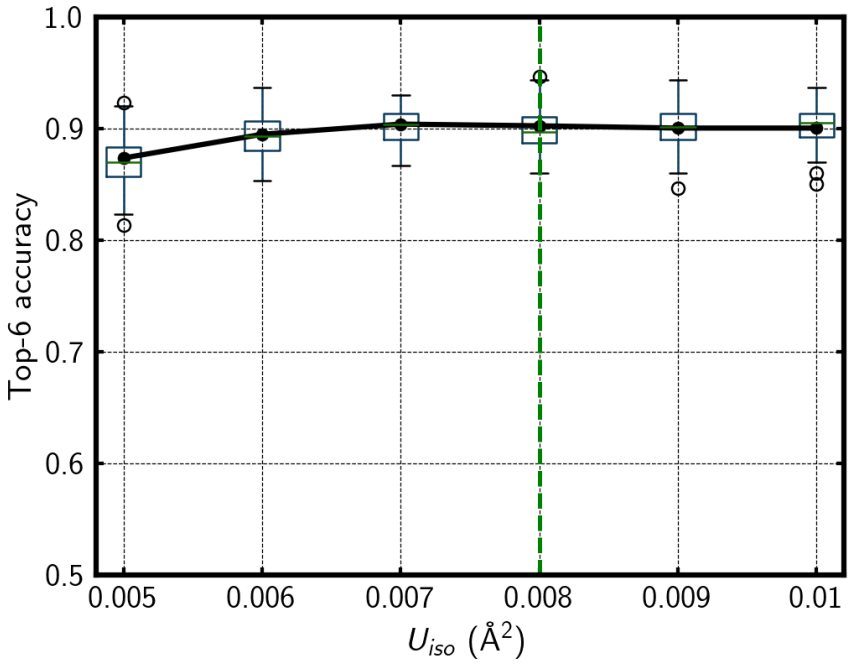

Fig. 5. The black dots represent the top-6 accuracy as $U_{\rm iso}$ is varied. The value of $U_{\rm iso}$ = 0.008 Å^2 used to train the model is shown as a vertical green dashed line. The box plot at each $U_{\rm iso}$ value is plotted in the same way as the ones in Figure 2.

The numbers from all the robustness tests are reproduced in the supplementary information associated with this paper.

4. Discussion

The main goal of this work is to investigate the model performance when given PDFs computed using different experimental parameters than the model was trained with. We found that the model predictions are quite robust for reasonable ranges of parameters, giving hope that the tool can be used by experimentalists when they have PDFs

measured under different conditions. Here we speculate on some more detailed aspects of the findings.

Two conditions must be satisfied for the model to perform well. First, the input data must contain sufficient information to do the differentiation by space group. Second, the values of the learned CNN model parameters must be able to correctly classify based on this information, even when it has been distorted by the use of different experimental parameters. We briefly discuss the effect of each experimental parameter on the information content of the PDF. These parameters are discussed in detail in (Egami & Billinge, 2012). Lowering $Q_{\rm max}$ can result in a loss of information from the missing high-Q region, and results in a broadening of the peaks in the PDF. Increases in $U_{\rm iso}$ also broaden the PDF peaks but are coming from increased static or thermal disorder in the sample itself. Because of peak overlap in the PDF, especially in the high-r region, broadened peaks always result in a loss of structural information. Clearly, lower the range of data used ($r_{\rm min}$ to $r_{\rm max}$) decreases the information content of the supplied data. These are parameters under the control of the experimenter though, as they are parameters that they can set during the data analysis to produce the PDF. A lower Q-space resolution of the measurement (resulting in a higher $Q_{\rm damp}$) can also lower the information content of the data due to Bragg peak overlap, especially in the high-Q region of the data before it is Fourier transformed to obtain the PDF. In the PDF this appears as an approximately Gaussian fall off in the structural signal with increasing-r. Any intrinsic nanocrystallinity in the sample, such as finite nanoparticles or loss of structural coherence in the form of smaller crystallites or domains, has a similar effect as $Q_{\rm damp}$ on the PDF. The fall-off in the signal will have a different functional form in this case (for example a power-law in the case of spherical domains/particles) but roughly speaking appears in a rather similar way as the Gaussian dropoff modelled by $Q_{\rm damp}$ and so we did not explicitly separate

these factors in this robustness test. We tested Q_{damp} over a range of values that simulated structural coherence down to a ~ 2 nm diameter. Finally, Q_{min} and Q_{broad} are expected to have only a very small effect on the accuracy. Q_{broad} is only relevant for data with very asymmetric Bragg peaks, for example, coming from time-of-flight neutron data, and even then, only at very high values of r that tend to be higher than the values we have been giving to the CNN. Q_{min} is often determined by the shadow of the beamstop in an experiment and will only affect the data if any low-angle signal is lost due to this. In that case, it results in very long wavelength undulations in the background of the resulting PDF that will not affect the model's ability to classify by spacegroup.

The observed robustness indicates that measured PDFs generally contain sufficient information to make this space-group determination, even when the data content is degraded somewhat by reduced real-space resolution (lower Q_{max} higher U_{iso}) and a more limited r-range of the data (lower r_{max} and higher Q_{damp}). The accuracy falls off more rapidly when there is a loss of information in the PDF (lower Q_{max}, r_{max}, higher Q_{damp}, U_{iso}); however, we note that the accuracy also falls off when we give the model a dataset with higher resolution or r-range, and therefore increased information content. The fall-off in accuracy in these cases must be due to the less than optimal learned CNN parameter values. This could be addressed by retraining the model with a wider range of experimental parameters, but it seems that it may not be required, except perhaps for the case of small nanoparticles (represented by large Q_{damp} values above 0.08 in this study). r_{min} and r_{max} is largely under the control of the experimentalist (it is a setting in most PDF data analysis programs), but more importantly, the range of r-that the signal persists over depends on the crystallite/domain size of the sample and whether it is nanocrystalline. This suggests that training a new model suitable for small nanoparticles (i.e., data signal ranges up to 1 or 2 nm) may

be warranted. We will look into this in the future and consider deploying it at the PDFitc website.

Another way that information in data is degraded is the presence of noise. Noise may be random or correlated. We have not systematically tested the robustness of the model to the presence of added noise in the data because it is difficult to reliably mimic the actual errors that are present in real data. A more meaningful measure of this is to establish how well the model works on actual datasets from known materials. This was reported in the original paper (Liu *et al.*, 2019). There it was found that of 15 experimental PDFs the model gave a correct prediction in the top-6 from 12 cases. This is not a large sample, but is an 80% accuracy. Given that the datasets were already obtained with experimental parameters that are not necessarily those that the model was trained with, this is comparable, if somewhat degraded, performance to the test data without noise that we report here. The SPACEGROUPMINING model is apparently also quite robust against the effects of measurement noise.

5. Conclusions

The use of deep learning to do complex classifications from data is a potentially useful approach that is becoming more widespread in materials science, crystallography and diffraction. Inherent in the process is that the model was trained on a particular set of data and its applicability to do the classification on data that is, in some way, different, for example, measured with different resolutions or over different ranges, might limit its ability to make accurate predictions. In general, the model may be retrained on a wider set of data that incorporates cases of different ranges, resolutions and so on. However, here, for the case of the SPACEGROUPMINING model that is deployed on pdfitc.org, we simply explored its robustness in making accurate predictions on different range and resolution data without retraining the model. The main result is

that the model is quite robust and performs well without having to be retrained in most cases. Modest reductions in prediction accuracy were observed, but it still performed well given a rather wide, but reasonable, range of resolution and range parameters, suggesting that it is not of great urgency to retrain it. We note that retraining it with a more diverse set of training data, whilst increasing accuracy for parameter values away from the original training values, it may decrease the prediction accuracy for PDFs with the original set of parameter values, where those values were chosen as being somewhat representative of values in many rapid acquisition x-ray PDF studies. Through this work, it has been shown that, without additional retraining, the spacegroupMining@pdfitc model still performs with reasonable accuracy for a relatively wide range of experimental parameters, and can thus be used as a robust computational tool.

Funding information

This work in the Billinge group was supported by the U.S. National Science Foundation through grant DMREF-1922234 and CCF-1704833.

References

Belsky, A., Hellenbrandt, M., Karen, V. L. & Luksch, P. (2002). *Acta Crystallographica Section B Structural Science*, **58**(3), 364–369.

Billinge, S. J. L. (2019). In *Nanometre-Scale Structure from Powder Diffraction: Total Scattering and Atomic Pair Distribution Function Analysis*, edited by C. Gilmore *et al.*, vol. H. Buffalo, NY, USA: International Union of Crystallography.

Egami, T. & Billinge, S. J. L. (2012). *Underneath the Bragg Peaks: Structural Analysis of Complex Materials*. No. 16 in Pergamon Materials Series. Amsterdam: Elsevier, 2nd ed.

Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *Acta Crystallographica Section A: Foundations of Crystallography*, **47**(6), 655–685.

Juhás, P., Farrow, C., Yang, X., Knox, K. & Billinge, S. (2015). *Acta Crystallographica Section A: Foundations and Advances*, **71**(6), 562–568.

Liu, C.-H., Tao, Y., Hsu, D., Du, Q. & Billinge, S. J. L. (2019). *Acta Crystallographica Section A: Foundations and Advances*, **75**(4), 633–643.

Proffen, T. & Billinge, S. J. L. (1999). *J. Appl. Crystallogr.* **32**, 572–575.

Yang, L., Culbertson, E. A., Thomas, N. K., Vuong, H. T., Kjær, E. T. S., Jensen, K. M. Ø., Tucker, M. G. & Billinge, S. J. L. (2021). *Acta Crystallographica Section A: Foundations and Advances*, **77**(1), 2–6.

Supplemental Materials

Table 2. *Top-6 accuracy and top-1 accuracy when r_{max} is chosen from 10 Å to 30 Å.*

$r_{\max}$ (Å)	10	12	14	16	18	20	22	24	26	28	**30**
Top-6 accuracy	0.648	0.717	0.747	0.778	0.820	0.871	0.882	0.890	0.901	0.901	**0.902**
Top-1 accuracy	0.285	0.367	0.433	0.449	0.511	0.552	0.600	0.617	0.652	0.671	**0.677**

Table 3. *Top-6 accuracy and top-1 accuracy when Q_{max} is chosen from 12 $Å^{-1}$ to 30 $Å^{-1}$.*

$Q_{\max}$ ($Å^{-1}$)	12	15	18	21	**23**	24	27	30
Top-6 accuracy	0.777	0.795	0.811	0.822	**0.902**	0.900	0.805	0.84
Top-1 accuracy	0.516	0.591	0.597	0.604	**0.677**	0.663	0.598	0.610

Table 4. *Top-6 accuracy and top-1 accuracy when Q_{damp} is chosen from 0 $Å^{-1}$ to* 0.15 $Å^{-1}$.

Q_{damp} ($Å^{-1}$)	0	0.03	**0.04**	0.06	0.09	0.12	0.15
Top-6 accuracy	0.861	0.897	**0.902**	0.879	0.741	0.648	0.578
Top-1 accuracy	0.602	0.659	**0.677**	0.579	0.390	0.294	0.234

Table 5. *Top-6 accuracy and top-1 accuracy when ADP is chosen from 0.005 $Å^2$ to 0.01 $Å^2$.*

adp ($Å^2$)	0.005	0.006	0.007	**0.008**	0.009	0.01
Top-6 accuracy	0.873	0.895	0.904	**0.902**	0.900	0.900
Top-1 accuracy	0.618	0.649	0.664	**0.677**	0.666	0.641