Perovskite or Not Perovskite? A Deep-Learning Approach to Automatically Identify New Hybrid Perovskites from X-ray Diffraction Patterns

Florian Massuyeau, Thibault Broux, Florent Coulet, Aude Demessence, Adel Mesbah, Romain Gautier

Perovskite or Not Perovskite? A Deep Learning Approach to Automatically Identify New Hybrid Perovskites from X-ray Diffraction Patterns

Florian Massuyeau,[1,], Thibault Broux,[1] Florent Coulet,[1] Aude Demessence,[2] Adel Mesbah,[2] Romain Gautier[1,*]*

[1] Nantes Université, CNRS, Institut des Matériaux de Nantes Jean Rouxel, IMN, F-44000 Nantes, France

[2] Université Claude Bernard Lyon 1, UMR CNRS 5256, Institute of Researches on Catalysis and Environment of Lyon (IRCELYON), Villeurbanne, France

E-mail: Florian.Massuyeau@cnrs-imn.fr and Romain.Gautier@cnrs-imn.fr

Keywords: deep learning, hybrid perovskite, X-ray diffraction

Determining the crystal structure is a critical step in the discovery of new functional materials. This process is time consuming and requires extensive human expertise in crystallography. Here, we developed a machine learning based approach which enables to determine automatically if an unknown material is of perovskite type from powder X-ray diffraction patterns. After training a deep learning model on a dataset of known compounds, the structure types of new unknown compounds could be predicted using their experimental powder X-ray diffraction patterns. This strategy was used to distinguish perovskite-type materials in a series of new hybrid lead halides. After validation, our approach was shown to accurately identify perovskites (accuracy of 92% with convolutional neural network). From the identification of the key features of the patterns used to discriminate perovskite vs. non-perovskite, crystallographers could learn how to quickly identify low dimensional perovskites from X-ray diffraction patterns.

1. Introduction

The structure determination is crucial in the discovery of new materials. The crystal structure guides the materials' properties such as absorption, conductivity or magnetism. When the structure-property relationships are well established, chemists can predict the property from the crystal structure. This determination from X-ray diffraction typically consists in four steps: (i) indexing, (ii) determination of the space group, (iii) solving the structure and (iv) refining the structural model. This four-step process, which is time consuming and requires specific skills in crystallography, is often considered as the main

bottleneck to accelerate the discovery of new materials. Recently, machine learning approaches have emerged to determine the symmetry (step (ii)) from their diffraction patterns.[1] However, the application of such methods are limited as the knowledge of the symmetry without any other information on the crystal structure is rarely useful for the prediction of properties. In other recent works, machine learning models have been shown to rapidly identify known compounds (or their symmetries) from their X-ray diffraction patterns.[2–8] However, no method has ever been reported to rapidly and efficiently identify structural characteristics of an unknown material from its powder X-ray diffraction pattern. Such method could pave the way to the development of high-throughput approaches to discover new materials.

To predict the properties, the information on the structure type of a new unknown material is very often sufficient. For example, hybrid lead halides of perovskite type have recently shown a great potential in optoelectronic applications owing to structure related properties such as defects tolerance, diffusion of exciton, ...[9–15] This discovery leads many research groups to reinvestigate the hybrid metal halide system in order to discover new low-dimensional hybrid perovskite compounds with enhanced properties (better chemical stability in solar cells, generation of white emission for solid-state lighting, …) (Figure 1).[13,14,16–21]

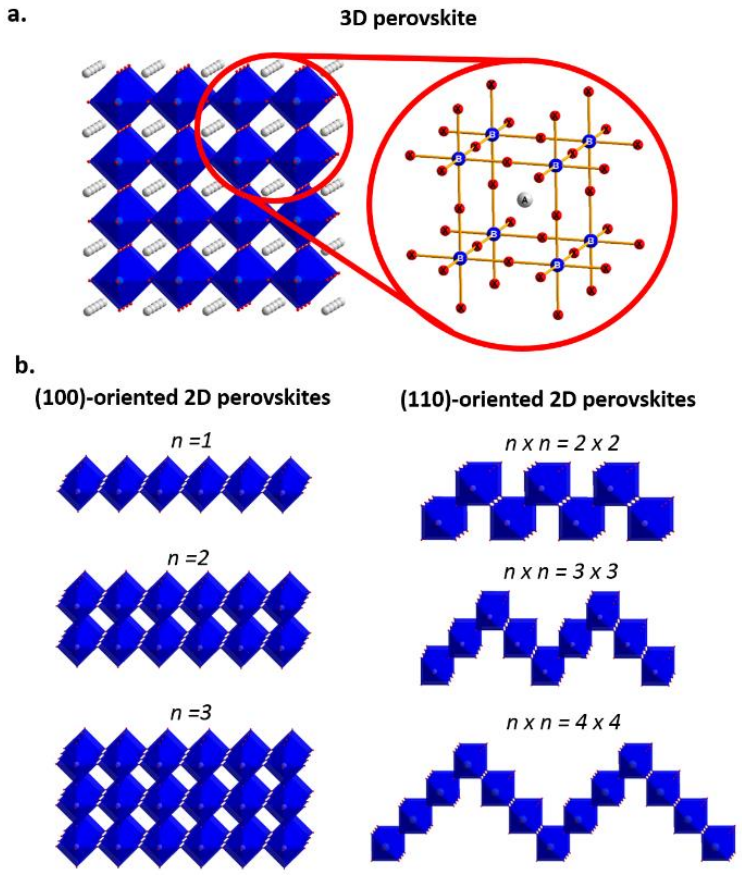

Figure 1. Representation of the perovskite structure type. (a) View of the 3D ABX_3 perovskite structure and (b) Examples of perovskite type layers.

Synthesizing hybrid metal halides is a straightforward process which can be carried out in a high-throughput manner.[22,23] However, no feature in the diffraction patterns can be simply identified and

used to discriminate perovskite from non-perovskite materials (the majority of hybrid lead halides) and identify interesting candidates for a specific property because the crystal structures of hybrid perovskites (especially the low dimensional ones) are very diversified. Such compounds can crystallize in 3D, 2D or 1D structures and many different categories of perovskite frameworks exist (see few examples for perovskites layers in Figure 1b).[24] For this reason, the discrimination between perovskite and non-perovskite type compounds currently requires the complete structure determination from X-ray diffraction. In this article, we demonstrated that Machine Learning (ML) models can be used to automatically determine whether an unknown compound is of a specific structure type and illustrated this approach with the hybrid perovskite system.

2. Results and discussion

2.1. Identification of perovskite structures by ML

To train ML algorithms at distinguishing between perovskite and non-perovskite structures, a preliminary step consisted in automatically sort perovskite and non-perovskite structures of hybrid lead halides from a dataset of crystallographic files extracted from the Cambridge Structural Database (see Experimental section and Figure S1). Each of these structures were identified as perovskite type or non-perovskite type using the definition by Mercier.[24] Thus, a perovskite-type compound can be 1D, 2D or 3D network built from only corner sharing PbX_6 octahedra. Powder X-ray diffraction (XRD) patterns were simulated for each of these structures.

Concerning the convolutional neural network (CNN) approach (Figure 2 and Experimental section), the accuracy of our predictive approach was tested on 200 simulated powder patterns and reached a mean value of 0.92 (see Experimental section for details). An increase of the accuracy is observed when reducing the patterns data (i.e. increasing the step size - Figure S2). The accuracy for different XRD 2θ ranges is compared in Figure 3a. This figure represents the result of CNN trainings with reduced 2θ range XRD patterns as inputs (with lower and upper 2θ range as indicating by the x- and y-axis, respectively). For small XRD ranges (results along the diagonal of Figure 3a), the accuracy remains close to the random weighted accuracy of 0.53. Three zones in the maps allow to better identify which XRD 2θ ranges are of importance for the accurate discrimination of *perovskite* vs. *non-perovskite*. The *zone 1* represented in Figure 3a is a region of high accuracy (> 0.85). The accuracy dropped significantly for XRD 2θ ranges starting after ca. 16° or ending before ca. 16°. Thus, an important feature at 16° is required to obtain an accurate prediction. The *zone 2* represented in Figure 3a is a region of moderate accuracy (≈ 0.75). This result shows that the important feature at 16° is not sufficient to discriminate accurately between *perovskite* and *non-perovskite*. The XRD 2θ range should either start below 8° and includes the feature at 16° or includes the feature at 16° and ends above 22° (*zone 1*). Thus, three features are important for the CNN at ca. 8° (A), 16° (B) and 22° (C). The *zones 2, 3* and *4* represented in Figure 3a are regions of moderate to low accuracies (< 0.75). These regions include only A (*zone 3*), B (*zone 2*) or C (*zone 4*). This result shows that the features A, B and C alone cannot discriminate accurately

between *perovskite* and *non-perovskite*. The combination of features (*zone 1* = A+B or B+C or A+B+C) leads to much higher accuracies. Thus, the analysis of the CNN results brings the following information: (a) each of the three features is not sufficient by itself to accurately predicts the perovskite, (b) the feature B must be considered to optimize the accuracy but must be at least accompanied with either the feature A or C. In addition, the results of the confusion matrix (Figure 3b-e) show that the small ranges of data lead the CNN to predict systematically a *non-perovskite* (the number of true and false negatives are high while the number of true and false positives are low). The data ranges and step size can also be optimized in order to obtain a fast and accurate prediction. Thus, a high accuracy (≈ 0.85) can be obtained for a data range as small as from 8° to 17° and with a step size of 0.2-0.35°. Such XRD patterns could be collected within one minute with a lab diffractometer enabling a high-throughput characterization of the structure type for new materials.

To compare this deep learning approach with other ML models, a random forest (RF) model was developed and showed an accuracy of 0.89 at discriminating perovskite vs. non-perovskites. The evolution of accuracy versus the step size was analyzed (Figure S3). Similarly to the CNN model, an increase of the accuracy can firstly be observed from 0.02° (2θ) to 1.7°, reaching an optimal value of about 0.9 in the range from 1.7° to 5.4°, followed by a constantly decreasing accuracy reaching a value as low as the random weighted accuracy of 0.53. These evolutions of accuracies with step size are normal as reducing the pattern size (i.e., increasing the step size) can negatively impact the accuracy if the models cannot recognize some specific signatures but increasing the pattern size can also negatively impact the accuracy prediction if the model is not enough sophisticated. Thus, increasing the step size from 0.02° to ca. 5° allows the reduction of parameters without losing information and increases the performance of the RF model. However, reducing the parameters with a step size above 5° leads to an important loss of information and the decrease of the accuracy. Besides this, the RF model calculates, for each XRD pattern, the probability of the structure to be of perovskite type. The distribution of these probabilities were analyzed in Figure S4. From this analysis, one can note that the model is more efficient at predicting correctly non-perovskite structures than perovskite structures. This result was expected, as non–perovskite structures are more numerous in the dataset. In addition, we can note that predicting correctly perovskite structures is optimal for a step size in the range 2°-10° (in agreement with the optimal step size of 2.18° for the highest overall accuracy).

The feature importance profile (step size hyperparameter of 2.18°) for the random forest were evaluated and showed three ranges which have importance for the discrimination of *perovskite* vs. *non-perovskite* for this ML model: (i) from 9 Å to 11 Å (i.e. approximately from 8° to 10°(2θ)) equivalent to the feature A of the CNN, (ii) from 5 Å to 7 Å (i.e. approximately from 12.5° to 17.5° (2θ)) equivalent to the feature B of the CNN, and (iii) from 3 Å to 5 Å (i.e. approximately from 17.5° to 29.5°) equivalent to the feature C of the CNN (Figure 3f). This result is validated by the analysis of the feature importance according to the step size and the data range (Figure S5 and S6 respectively). Using the full data from 3° to 50°, the three features A, B and C are shown to be important for the accuracy of the discrimination. When

data from 3° to 12° are excluded (i.e. feature A is excluded), the feature importance features only includes B and C. When data from 3° to 17° are excluded (i.e. features A and B are excluded), the feature importance only includes C.

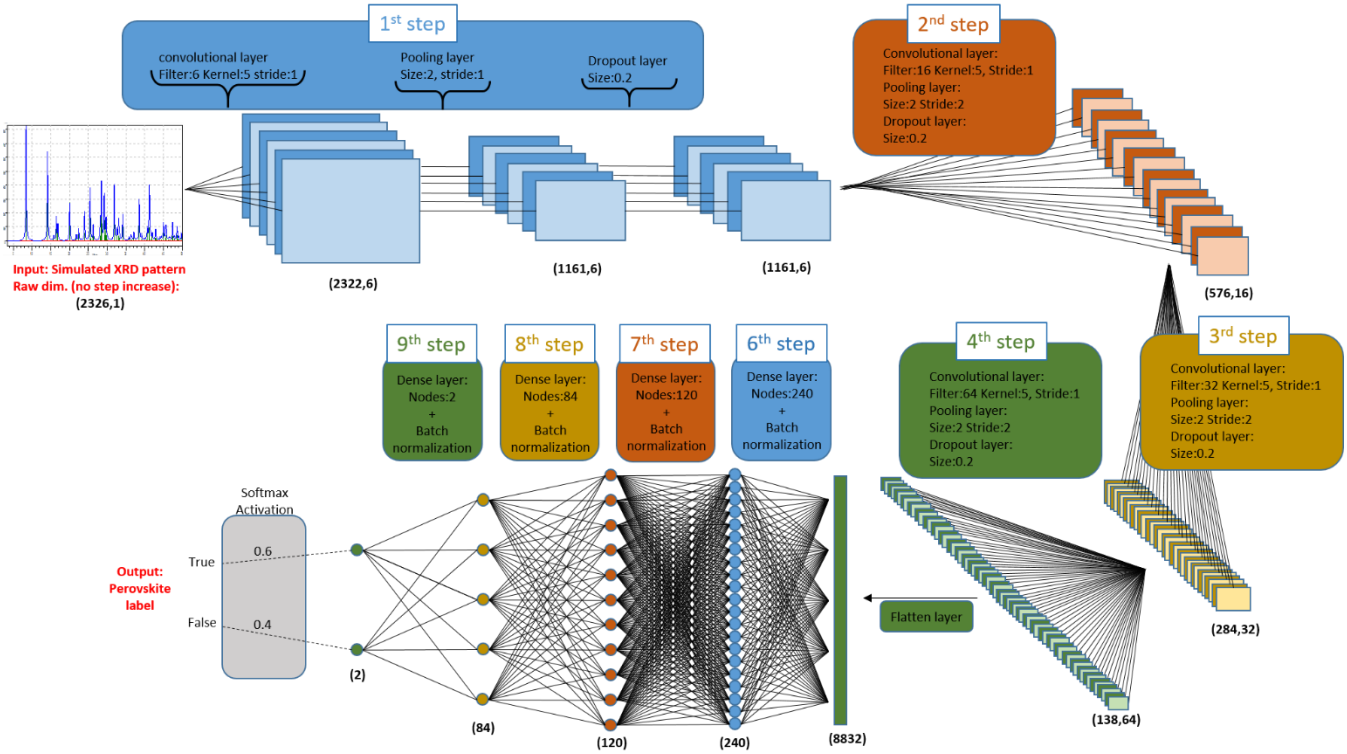

Figure 2. View of the architecture of the convolutional neural network used to classify perovskite vs. non-perovskite from the X-ray diffraction patterns.

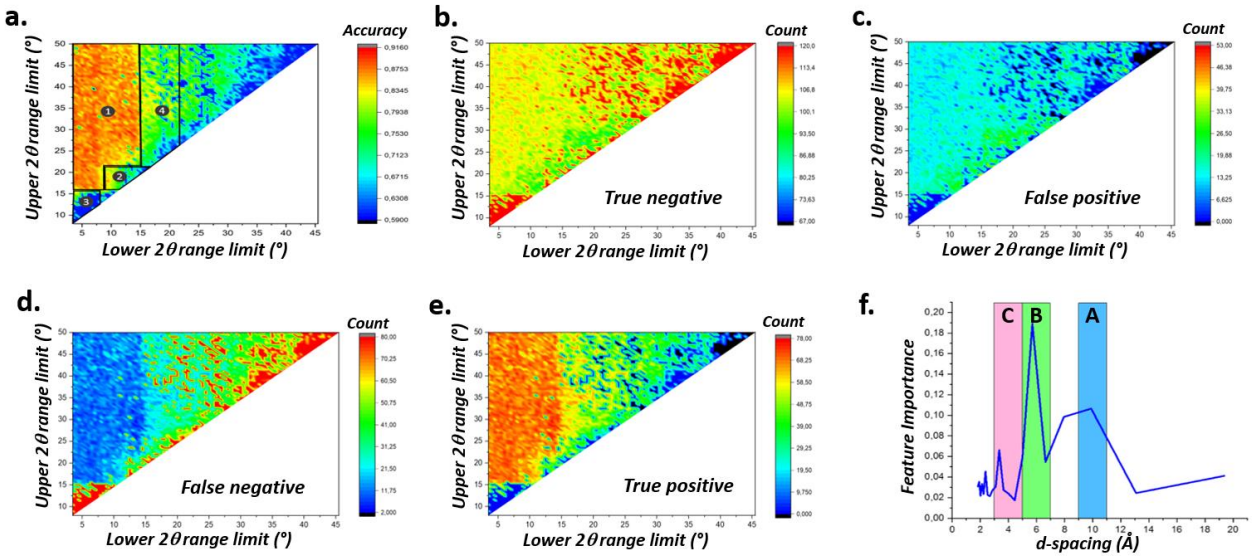

Figure 3. ML approach on the classification of perovskite vs. non-perovskite., a. Accuracy of the CNN prediction according to the range of XRD patterns (2θ) in the dataset, b. CNN confusion matrix *True Negative* according to the range of XRD patterns (2θ) in the dataset, c. CNN confusion matrix *False Positive* according to the range of XRD patterns (2θ) in the dataset, d. CNN confusion matrix *False Negative* according to the range of XRD patterns (2θ) in the dataset, e. CNN confusion matrix *True Positive* according to the range of XRD patterns (2θ) in the dataset, f. Feature importance of the XRD patterns (*d-spacing* (Å)) for the classification from random forest (Step: 2.18°(2θ)).

2.2. Signification of the important XRD features identified by ML

The accuracy of both approaches is relatively high (about 90%). In addition, both strategies enable to interpret the automatic sorting of perovskite vs. non-perovskite. To unravel what the three important features A, B and C identified by the ML models correspond to, the powder X-ray diffraction patterns of all compounds have been summed up according to the halogen and the structure type (i.e. perovskite or non-perovskite) (Figure 4a-b). We can observe for the perovskite patterns the appearance of a large peak, shifting from about 16° to 14° from Cl to I halogen respectively. These peaks can be assigned to the feature B and correspond respectively to interplanar spacings of 6.30 Å, 5.88 Å and 5.52 Å. In addition, the non-perovskites summed patterns show an unresolved broadband from about 5° to 15° (i.e. interplanar spacing from 8 Å to 17.6 Å) assigned to the feature A and a broad peak from about 17.5° to 29.5° (i.e. interplanar spacing from 3 Å to 5 Å) assigned to feature C. In the XRD patterns, the most intense diffraction peaks are associated to the crystallographic planes within which the electron density variations are maximal relatively to other planes. Thus, the plane should be dense in atoms and these atoms should have a high scattering factors. As the scattering cross sections are larger for heavier atoms, the most intense peaks on the diffraction patterns are the ones related to the distance between crystallographic planes of Pb atoms for the hybrid lead halides (Figure 4c). For this reason, we decided to analyze the statistics of Pb-Pb distances in the crystal structures of the datasets *Perovskites* and *Non-perovskites*. The Pb-Pb distances were calculated using the CSD Python API. While the mean minimum Pb-Pb distances for *perovskites* are between 5.7 Å to 6.3 Å according to the used halogens, the mean minimum Pb-Pb distances are around 4.2 Å for *non-perovskites* (Figure 5a-c.). For the *non-perovskites* group, some compounds with large Pb-Pb distances are also reported. These distributions can be explained by a simple analysis of the structural characteristics of each group. While *perovskites* are built from corner sharing PbX_6 octahedra, *non-perovskites* are typically built from either face-sharing (Short Pb-Pb distances), edge-sharing (Short Pb-Pb distances) or isolated PbX_6 octahedra (Long Pb-Pb distances).

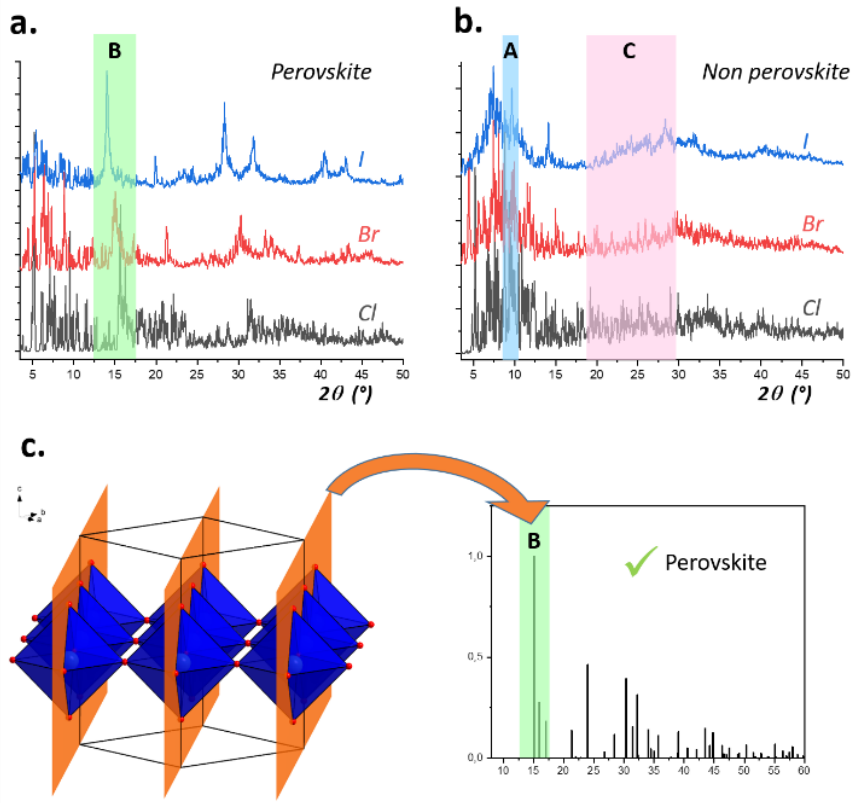

Figure 4. Relationships between the structural characteristics and the feature importance of the XRD patterns. a. Sum of all X-ray diffraction patterns of perovskite type compounds according to the halogen, b. Sum of all X-ray diffraction patterns of non-perovskite type compounds according to the halogen, c. Simulation of the XRD pattern of a layered (100) perovskite structural model. The most intense peak corresponding to dense crystallographic planes of PbBr atoms is an important feature for the ML to discriminate between perovskite and non-perovskite.

To relate these statistical results to the *perovskite* / *non-perovskite* classifications by RF and CNN, a decision tree (entropy criteria, max deep = 3) relating Pb-Pb distances to the structure type *perovskite* vs. *non-perovskite* was built (Figure 5d). For this purpose, different Pb-Pb distance groups (]3 Å , 5 Å];]5 Å , 7 Å];]7 Å , 9 Å];]9 Å , 11 Å];]11 Å , 13 Å];]13 Å , 15 Å];]15 Å , 17 Å];]17 Å , 20 Å]) were considered. For each structure, the frequency of Pb-Pb distances in each group is quantified and normalized. Then, the decision tree classifies *perovskite* vs. *non-perovskite* according to the frequency of Pb-Pb distances of each group. From this result, one can observe that only two groups (]3 Å , 5 Å] and]5 Å , 7 Å]) have a significant importance for the classification (Figure 5d). More precisely, if a significant frequency of Pb-Pb distances is reported in the group]3 Å , 5 Å], the structure is classified as *non-perovskite*. If this frequency remains low in]3 Å , 5 Å] and the frequency is relatively high in]5 Å , 7 Å], the structure is classified as *perovskite*. However, if the frequency remains low in]3 Å, 5 Å] and the frequency is relatively low in]5 Å , 7 Å], the structure is classified as *non-perovskite*. This classification agrees with the statistical analysis of minimum Pb-Pb distances (Figure 5a-c) and enables to rationalize the CNN and the RF classification from XRD patterns in the regions from 3 Å to 5 Å corresponding to the feature C and from 5 Å to 7 Å corresponding to B. However, the feature A (region from 9 Å to 11 Å) cannot be explained on the basis of this analysis of Pb-Pb distances (Figure 3f). This region corresponds to large interplanar spacings and cannot be related to minimum distances between atoms of the crystal structure. Instead, such region would be directly related to general information on

the crystal lattice such as the unit-cell and the symmetry. Thus, a statistical analysis on the unit-cell volume and symmetry (crystal system and space-group) was performed (Figure 5e-f and Figure S7). Such analysis shows that the *non-perovskite* structures are more likely to exhibit large unit-cells and lower symmetries than *perovskite* structures. Such differences would result in the presence of more XRD peaks at low angles for *non-perovskite* than for *perovskite*. This explanation is also confirmed through the observation of the summed XRD patterns of *non-perovskites* in which a broadband peak is observed at low 2θ angles. Thus, the CNN and RF algorithms would use the information at low angle (feature A) corresponding to statistical differences in symmetry and unit-cell volume to classify *non-perovskites* vs. *perovskites*. It is also interesting to add that the three features identified by the ML models are not correlated to each others.

In order to confirm our assignments of each of the three features (A to non-perovskite, B to perovskite and C to non-perovskite), a dataset of 100 XRD patterns was built and analyzed by a scientist. Using this assignment, the scientist correctly discriminated the perovskites vs. non-perovskites with an accuracy of 72 %. Even if this performance is lower than the one of ML (the mathematical models statistically weight the key features leading to more accurate predictions than humans can do), these results further confirm the right assignments of A, B and C. In addition, it is important to note that, prior to this identification of the three features using the ML models, the discrimination of perovskites vs. non perovskites from powder XRD patterns was not be possible without a time consuming structure determination.

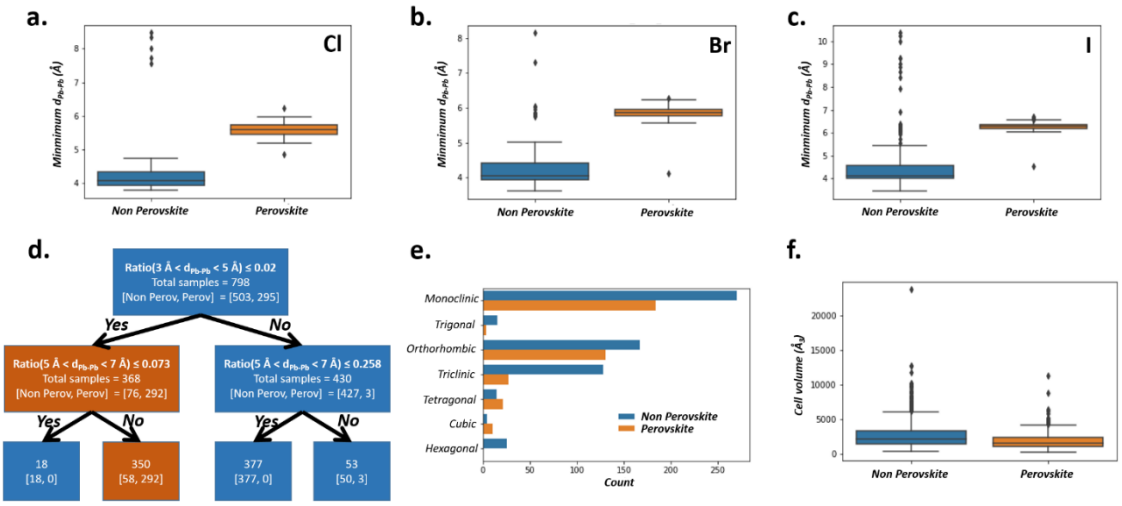

Figure 5. Structural analysis of perovskite vs. non-perovskite. a. Minimum Pb-Pb distances in the crystal structures of hybrid lead chlorides, b. Minimum Pb-Pb distances in the crystal structures of hybrid lead bromides, c. Minimum Pb-Pb distances in the crystal structures of hybrid lead iodides, d. Decision tree for the classification of *perovskite* vs. *non-perovskite* using ranges of Pb-Pb distances from (i) 3 Å to 5 Å and (ii) from 5 Å to 7 Å, e. Classification of perovskite vs. non-perovskite according to the crystal system, f. statistical analysis of the unit-cell volumes.

2.3. Performances on experimental XRD data

In addition to the tests on 200 simulated XRD patterns, the ML approaches were further validated on experimental XRD data. For this purpose, the XRD patterns of 23 freshly synthesized samples were

collected (Table S1). Nine were previously published compounds (i.e. reported in the CSD) while 14 were new compounds with unknown crystal structures. The crystal structures of these unknown compounds were determined through traditional structure determination from the single-crystal or synchrotron powder X-ray diffraction patterns (Figure S8 and S9). For the structure type identification, a Savitzky-Golay smoothing was used with the same parameters for all experimental raw patterns to remove the noise and background signal using Origin 2020 software (Figure S10 and S11). Of the 23 synthesized samples, the classification through RF and CNN reached an accuracy of 0.78 and 0.73 (mean value, see Figure 6), respectively. Thus, these accuracies remain relatively high even if different effects (e.g. preferential orientation, different signal / noise ratio, …) could have impacted the classification based on experimental datasets.

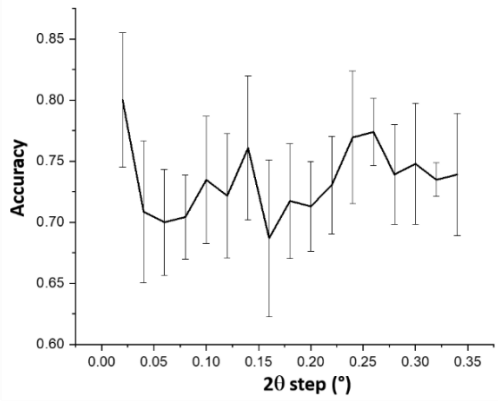

Figure 6. Evolution of the accuracy according to the step size of the XRD patterns for the CNN model on the experimental XRD patterns.

3. Conclusion

In summary, the identification of the perovskite structure types from the powder X-ray diffraction patterns of hybrid lead halides was carried out through RF and CNN models. The analysis of feature importance and the structural characteristics of perovskite structures vs. others allows to identify the XRD peaks which enable to discriminate perovskites vs. non-perovskites. The distances between Pb ions as well as volume of unit-cell are identified as a key structural factors related to the diffraction peaks of importance for the ML. Thus, we believe such ML approaches of identification of perovskites could be generalized to other chemical systems and structure types. Not only these techniques can be used to automatically identify the structure types from rapidly collected XRD patterns but the identification of the important features also enable to augment the scientists' ability in discriminating between different structure types of new materials.

4. Experimental Section

Machine learning: The dataset built in 2020 consists in 998 crystal structures from the Cambridge Structural Database. Pymatgen and Robocrystallographer python libraries were used to sort perovskites and non-perovskites.[25,26] Pymatgen enabled to create the structures

from the CIF to automatically identify the atomic neighbors of Pb atoms while Robocrys identified how the PbX_6 octahedra were linked to each other's (i.e. isolated, corner-sharing, edge-sharing or face-sharing). From these characteristics, perovskite and non-perovskite could be automatically sorted with the following rule: *A perovskite-type compound must be built from only corner sharing PbX_6 octahedra*. Compounds with edge-sharing, face-sharing, or isolated PbX_6 octahedra are not perovskites. The correctness of this sorting was validated manually. The resulting datasets are composed of 375 perovskite-type compounds (50 chlorides, 105 bromides and 220 iodides), and 623 non-perovskite-type compounds (58 chlorides, 139 bromides and 426 iodides) (Figure S1). The powder X-ray diffraction pattern for each compound of our database was simulated from 3.5° to 50° 2θ using the freely available software package xrayutilities considering a Cu Kα radiation (the most commonly used radiation for powder X-ray diffraction).[27] Two supervised classification machine learning algorithms were investigated and compared: (i) the convolutional neural network (CNN) using Tensorflow (Python library),[28] and Keras API,[29] and (ii) the random forest (RF) using Scikit-Learn (Python library).[30] For both approaches, the training and test datasets represented 80% (798 structures) and 20% (200 structures) of the data, respectively. Concerning the deep learning approach, the CNN is composed of 23 layers (Figure 2) and simulated patterns act as 1-dimensional input for this convolutional strategy. The kullback-Leibler divergence was used as loss function with an Adam optimization. A training with a batch size of 20 samples and maximum epoch of 30 was carried out using an early-stopping function to avoid overtraining. To obtain a reliable accuracy value, the training of the CNN was repeated ten times for each step size. Concerning the RF approach, a 5-fold cross-validation coupled with a randomized search (1000 iterations) approach was used to tune the hyperparameters and optimize the model. An optimal mean accuracy for the test dataset of 0.89±0.03, which is slightly lower than the CNN performance, is obtained over the 5-fold cross-validation for the best iteration corresponding to the five following hyperparameters (four hyperparameters from the random forest and one custom-built hyperparameter): (i) *Number of trees: 100*, (ii) *Maximum number of levels in tree: 10, (iii) Minimum number of samples on a leaf: 2, (iv) Minimum number of samples to split a node: 10, (v) Step size for the XRD patterns (customized one) :2.18°*. The step-size custom-built hyperparameter is constructed by reducing the XRD pattern data while preserving its overall shape. The analysis of accuracy vs. step-size (Figure S3) was carried out by maintaining all other hyperparameters constant (number of trees: 100, maximum number of levels in tree: 10, minimum number of samples on a leaf: 2, minimum number of samples to split a node: 10 constants).

Data and code availability: The data are provided in the SI in the folder 'perov-files\data'. Cif files from CSD can be found in the folder 'perov-files\data\raw', experimental XRD patterns in the folder 'perov-files\data\external' and calculated features and ML results in the folder 'perov-files\data\processed'. The files provided in the folder 'perov-files\data\interim' were used to obtain the halide label and to eliminate duplicate structures. Codes are provided in the folder 'perov-files'. 'build_features.py' and 'Features-calc.ipynb' calculate the different structural features. The accuracies for the CNN for simulated and experimental patterns were obtained with'cnn-10training.ipynb'. Results for the different trainings and different step sizes are in the folder 'perov-files\data\processed\data-CNN'. Confusion matrices for CNN are obtained with 'CNN-model.py'. As this calculation is very time-consuming, this code has been parallel processed using bash command (CNN_bash.sh and command line in CNN-bash-command.txt). Results of this code are on 'CNN-results.txt'. Pb-Pb distances and decision trees are obtained with 'PbPbdist.ipynb' and 'PbPb-histo-tree.ipynb' respectively. The optimal hyperparameters, accuracies and feature importances for different step sizes, as well as optimal accuracy for experimental patterns for RF model are obtained with 'RF-red.ipynb'. The calculated feature importances are provided in the folder 'perovfiles\data\processed\data-RF'. Accuracies and feature importances obtained for RF by modifying the 2θ range of the XRD pattern are calculated with 'RF-databegin.ipynb', and results are provided in the folder 'perov-files\data\processed\data-RF-begin'.

Synthesis of hybrid lead halides: To evaluate the ML performances on experimental XRD powder patterns, 23 samples were synthesized. The compounds were synthesized by mixing HX (HCl, 37%, HBr 48% or HI 57%) with an amine and a Pb source (Pb metal, $PbCl_2$, $PbBr_2$ or PbI_2): *Compound 1* – 120 mg Pb and 145µl 4-(Aminomethyl)piperidine in 10mL HI at room temperature. *Compound 2* – 120 mg Pb and 147 µL N,N,N' –Trimethylethylenediamine in 10mL HI at room temperature with addition of ethanol for precipitation. *Compound 3* – 120 mg Pb and 123 µL 4-Aminopiperidine in 10mL HBr under reflux. *Compound 4* – 120 mg Pb and 123 µL 4-Aminopiperidine in 10mL HI at room temperature. *Compound 5* – 120 mg Pb and 145 µL 4-(Aminomethyl)piperidine in 10mL HCl at room temperature. *Compound 6* – 120 mg Pb and 123 µL 4-Aminopiperidine in 10mL HCl at room temperature with addition of ethanol for precipitation. *Compound 7* – 120 mg Pb and 432 mg 1,4-Diazobicyclo-[2,2,2]-octane in 10mL HBr at room temperature. *Compound 8* – 120 mg Pb and 282 µL (2-Methylbutyl)amine in 10mL HBr at room temperature. *Compound 7* – 120 mg Pb and 432 mg 1,4-Diazobicyclo-[2,2,2]-octane in 10mL HBr at room temperature. *Compound 9* – 120 mg Pb and 282 µL (2-

Methylbutyl)amine in 10mL HI at room temperature. *Compound 10* – 120 mg Pb and 147 μL N,N,N' -Trimethylethylenediamine in 10mL HBr at room temperature with addition of ethanol for precipitation. *Compound 11* – 120 mg Pb and 229 μL Piperidine in 10mL HI at room temperature. *Compound 12* – 120 mg Pb and 229 μL pyridine in 10mL HBr at room temperature with addition of ethanol for precipitation. *Compound 13* – 120 mg Pb and 91 μL pymiridine in 10mL HI at room temperature with addition of ethanol for precipitation. *Compound 14* – 200 mg $PbCl_2$ and 125 mg 1,2 aminoethylpiperazine in 10mL HCl at room temperature with addition of ethanol for precipitation. *Compound 15* – 200 mg $PbBr_2$ and 125 mg 1,2 aminoethylpiperazine in 5mL HBr at room temperature with addition of ethanol for precipitation. *Compound 16* – 1 g Pb and 1 g trans-2,5-dimethylpiperazine in 20mL HBr under reflux. *Compound 17* – 1 g Pb and 1 g trans-2,5-dimethylpiperazine in 20mL HCl under reflux. *Compound 18* – 1 g $PbCl_2$ and 1 ml 2,2'-(ethylenedioxy)bis(ethylamine) in 20mL HCl. *Compound 19* – 139 mg $PbCl_2$ and 180 mg 4,4'-Azopyridine in 10 mL HCl under reflux and with addition of ethanol for precipitation. *Compound 20* – 184 mg $PbBr_2$ and 100 mg 4,4'-Azopyridine in 10 mL HBr under reflux and with addition of ethanol for precipitation. *Compound 21* – 230 mg PbI_2 and 180 mg 4,4'-Azopyridine in 10 mL HI under reflux and with addition of ethanol for precipitation. *Compound 22* – 500 mg Pb and 500 μL methylamine in 3 mL HCl in an autoclave (180°C for 24 H / cooling ramp at 10°C /hour). *Compound 23* – 500 mg Pb and 500 μL methylamine in 1 mL HBr in an autoclave (180°C for 24 H / cooling ramp at 10°C /hour).

Structure determination: The crystal structures of the new hybrid metal halide compounds have been determined through single-crystal X-ray diffraction, laboratory/synchrotron powder X-ray diffraction. For structures determined by single-crystal X-ray diffraction, data were collected on a Bruker-Nonius Kappa CCD diffractometer with monochromated Mo Kα radiation. Absorption corrections were carried out with SADABS.[31] Direct methods were used to determine the crystal structures with SHELXT.[32] The refinement of the crystal structures with anisotropic displacement parameters was carried out using SHELXL-2013.[33] The CIFs were compiled with Olex2.12.[34] Additional symmetry elements were checked with the program PLATON.[35] For structures determined by powder X-ray diffraction, synchrotron data were collected at the CRISTAL beamline of the synchrotron SOLEIL using the 2 circle diffractometer equipped with the Mythen detector (λ = 0.72844 Å). Each PXRD pattern was collected in a transmission mode (Debye-Sherrer) in a capillary (diameter of 0.7 mm). The 2θ angular range is 1° - 70° and the total counting time is 2 minutes. LaB_6 was collected in similar

conditions to refine the wavelength value and extract the instrumental function. Laboratory data were collected on a D8 Bruker diffractometer with the CuK-L3 radiation (Ge (111) monochromator) and a Lynxeye 1D detector in the 5–90° 2θ range. The unit cell parameters of each compound were indexed using the X-Cell algorithm of the Materials Studio program.[36] The inorganic part of the structure was obtained by direct methods using EXPO 2014,[37] and when it is possible the organic part was solved under the form of a rigid body using the direct space methods (simulated annealing) implemented in the Reflex Module of materials studio (BIOVIA). The final Rietveld refinement of the solved structures was performed using the Fullprof_Suite package.[38] The unit cell parameters and the structural features of each compound are reported in Table S1 and Figure S9. CCDC 2168613-2168613 and 2168619 contain the supplementary crystallographic data for this paper. These data can be obtained free of charge from The Cambridge Crystallographic Data Centre via www.ccdc.cam.ac.uk/data_request/cif.

Supporting Information

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

Acknowledgements

This work was supported by the National Agency for Research (ANR-16-CE08-0003-01 and ANR-21-ERCC-0009-01) and Region Pays de la Loire (Etoiles montantes en Pays de la Loire 2017, project “Découverte de pérovskites hybrides assistée par ordinateur”). Calculations were conducted at Centre de Calcul Intensif des Pays de la Loire (CCIPL), Université de Nantes.

Received: ((will be filled in by the editorial staff))

Revised: ((will be filled in by the editorial staff))

Published online: ((will be filled in by the editorial staff))

References

[1] W. B. Park, J. Chung, J. Jung, K. Sohn, S. P. Singh, M. Pyo, N. Shin, K.-S. Sohn, *IUCrJ* **2017**, *4*, 486.

[2] A. Ziletti, D. Kumar, M. Scheffler, L. M. Ghiringhelli, *Nat Commun* **2018**, *9*, 2775.

[3] P. M. Vecsei, K. Choo, J. Chang, T. Neupert, *Phys. Rev. B* **2019**, *99*, 245120.

[4] F. Oviedo, Z. Ren, S. Sun, C. Settens, Z. Liu, N. T. P. Hartono, S. Ramasamy, B. L. DeCost, S. I. P. Tian, G. Romano, A. Gilad Kusne, T. Buonassisi, *npj Comput Mater* **2019**, *5*, 1.

[5] J. A. Aguiar, M. L. Gong, R. R. Unocic, T. Tasdizen, B. D. Miller, *Science Advances* **2019**, *5*, eaaw1949.
[6] H. Wang, Y. Xie, D. Li, H. Deng, Y. Zhao, M. Xin, J. Lin, *J. Chem. Inf. Model.* **2020**, *60*, 2004.
[7] J.-W. Lee, W. B. Park, J. H. Lee, S. P. Singh, K.-S. Sohn, *Nat Commun* **2020**, *11*, 86.
[8] P. M. Maffettone, L. Banko, P. Cui, Y. Lysogorskiy, M. A. Little, D. Olds, A. Ludwig, A. I. Cooper, *Nat Comput Sci* **2021**, *1*, 290.
[9] S. D. Stranks, G. E. Eperon, G. Grancini, C. Menelaou, M. J. P. Alcocer, T. Leijtens, L. M. Herz, A. Petrozza, H. J. Snaith, *Science* **2013**, *342*, 341.
[10] M. A. Green, A. Ho-Baillie, H. J. Snaith, *Nature Photonics* **2014**, *8*, 506.
[11] S. D. Stranks, H. J. Snaith, *Nat Nano* **2015**, *10*, 391.
[12] D. Shi, V. Adinolfi, R. Comin, M. Yuan, E. Alarousu, A. Buin, Y. Chen, S. Hoogland, A. Rothenberger, K. Katsiev, Y. Losovyj, X. Zhang, P. A. Dowben, O. F. Mohammed, E. H. Sargent, O. M. Bakr, *Science* **2015**, *347*, 519.
[13] H. Tsai, W. Nie, J.-C. Blancon, C. C. Stoumpos, R. Asadpour, B. Harutyunyan, A. J. Neukirch, R. Verduzco, J. J. Crochet, S. Tretiak, L. Pedesseau, J. Even, M. A. Alam, G. Gupta, J. Lou, P. M. Ajayan, M. J. Bedzyk, M. G. Kanatzidis, A. D. Mohite, *Nature* **2016**, *536*, 312.
[14] J.-C. Blancon, H. Tsai, W. Nie, C. C. Stoumpos, L. Pedesseau, C. Katan, M. Kepenekian, C. M. M. Soe, K. Appavoo, M. Y. Sfeir, S. Tretiak, P. M. Ajayan, M. G. Kanatzidis, J. Even, J. J. Crochet, A. D. Mohite, *Science* **2017**, *355*, 1288.
[15] H. Yuan, L. Qi, M. Paris, F. Chen, Q. Shen, E. Faulques, F. Massuyeau, R. Gautier, *Advanced Science* **2021**, *8*, 2101407.
[16] D. B. Mitzi, S. Wang, C. A. Feild, C. A. Chess, A. M. Guloy, *Science* **1995**, *267*, 1473.
[17] E. R. Dohner, E. T. Hoke, H. I. Karunadasa, *J. Am. Chem. Soc.* **2014**, *136*, 1718.
[18] B. Saparov, D. B. Mitzi, *Chem. Rev.* **2016**, *116*, 4558.
[19] L. Mao, Y. Wu, C. C. Stoumpos, M. R. Wasielewski, M. G. Kanatzidis, *J. Am. Chem. Soc.* **2017**, *139*, 5210.
[20] R. Gautier, F. Massuyeau, G. Galnon, M. Paris, *Advanced Materials* **2019**, *31*, 1807383.
[21] L. Mao, P. Guo, M. Kepenekian, I. Hadar, C. Katan, J. Even, R. D. Schaller, C. C. Stoumpos, M. G. Kanatzidis, *J. Am. Chem. Soc.* **2018**, *140*, 13078.
[22] S. Brochard-Garnier, M. Paris, R. Génois, Q. Han, Y. Liu, F. Massuyeau, R. Gautier, *Advanced Functional Materials* **2019**, *29*, 1806728.
[23] R. W. Epps, M. S. Bowen, A. A. Volk, K. Abdel-Latif, S. Han, K. G. Reyes, A. Amassian, M. Abolhasani, *Advanced Materials* **2020**, *32*, 2001626.
[24] N. Mercier, *Angewandte Chemie* **2019**, *131*, 18078.
[25] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, G. Ceder, *Computational Materials Science* **2013**, *68*, 314.
[26] A. M. Ganose, A. Jain, *MRS Communications* **2019**, *9*, 874.
[27] D. Kriegner, E. Wintersberger, *Xrayutilities, Http:// Xrayutilities.Sourceforge.Net.*, **2013**.
[28] *Tensorflow, Https://Arxiv.Org/Abs/1603.04467*, **n.d.**
[29] *Chollet, F., & Others. (2015). Keras. GitHub. Retrieved from Https://Github.Com/Fchollet/Keras*, **n.d.**
[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *Journal of Machine Learning Research* **2011**.
[31] G. M. Sheldrick, *SADABS*, University of Göttingen: Germany, **2002**.
[32] G. M. Sheldrick, *Acta Cryst A, Acta Cryst Sect A, Acta Crystallogr A, Acta Crystallogr Sect A, Acta Crystallogr A Cryst Phys Diffr Theor Gen Crystallogr, Acta Crystallogr Sect A Cryst Phys Diffr Theor Gen Crystallogr* **2015**, *71*, 3.

[33] G. M. Sheldrick, *Acta Cryst C, Acta Cryst Sect C, Acta Crystallogr C, Acta Crystallogr Sect C, Acta Crystallogr C Cryst Struct Commun, Acta Crystallogr Sect C Cryst Struct Commun* **2015**, *71*, 3.
[34] O. V. Dolomanov, L. J. Bourhis, R. J. Gildea, J. A. K. Howard, H. Puschmann, *Journal of Applied Crystallography* **2009**, *42*, 339.
[35] Spek, A. L., *PLATON ; Utrecht University: Utrecht, The Netherlands, 2001*, **n.d.**
[36] M. A. Neumann, *J Appl Cryst* **2003**, *36*, 356.
[37] A. Altomare, C. Cuocci, C. Giacovazzo, A. Moliterni, R. Rizzi, N. Corriero, A. Falcicchio, *J Appl Cryst* **2013**, *46*, 1231.
[38] T. Roisnel, J. Rodríquez-Carvajal, *Materials Science Forum* **2001**, *378–381*, 118.

A deep learning approach is developed to automatically and accurately assign the structure type from the X-ray diffraction patterns of new hybrid lead halides. An accuracy of 92% is obtained and the new insights provided by the model enable to augment the scientists' ability in discriminating manually between different structure types of new materials.

Florian Massuyeau,[1,], Thibault Broux,[1] Florent Coulet,[1] Aude Demessence,[2] Adel Mesbah,[2] Romain Gautier[1,*]*

Perovskite or Not Perovskite? A Deep Learning Approach to Automatically Identify New Hybrid Perovskites from X-ray Diffraction Patterns