Machine Learning Recognition of hybrid lead halide perovskites and perovskite-related structures out of X-ray diffraction patterns

Marchenko E.I.[a,b], Korolev V.V.[c], Kobeleva E. A.[b], Belich N.A. [a], Udalova N.N. [a], Eremin N.N. [b,e], Goodilin E.A.[a,d], Tarasov A.B.[a,d]*

[a] *Laboratory of New Materials for Solar Energetics, Department of Materials Science, Lomonosov Moscow State University; 1 Lenin Hills, 119991, Moscow, Russia;*
[b] *Department of Geology, Lomonosov Moscow State University; 1 Lenin Hills, 119991, Moscow, Russia*
[c] *MSU Institute for Artificial Intelligence, Lomonosov Moscow State University; 119192, Moscow, Russia*
[d] *Department of Chemistry, Lomonosov Moscow State University; 1 Lenin Hills, 119991, Moscow, Russia*
[e] *Institute of Geology of Ore Deposits, Petrography, Mineralogy, and Geochemistry, Russian Academy of Science, Moscow, Russia*
e-mail: alexey.bor.tarasov@yandex.ru

Supporting Information

Evaluation of machine learning models

The evaluation of machine learning algorithms plays a crucial role in assessing their performance. However, it can be challenging, especially in scenarios where limited or no access to real-world data exists. In such cases, additional human effort is required to assess the model performance. In classification tasks, the evaluation is typically carried out by splitting the dataset into a training set and a test set. The machine learning algorithm is trained on the training set, while the test set is used to calculate performance indicators that assess the performance of the model. One common challenge faced by machine learning algorithms is the availability of limited training and test data. This can impact the algorithm's generalization capabilities and lead to overfitting or underfitting. It's crucial to have enough data for training and testing to have a good machine-learning model. Evaluating the performance of a machine learning model involves considering multiple factors. While there is no perfect indicator applicable to every scenario, several important factors are considered. These factors include:

• Accuracy: Measures the proportion of correctly classified instances, providing an overall assessment of the algorithm's performance.

• Precision: Evaluates the algorithm's ability to correctly identify positive instances within a given class, indicating its effectiveness in minimizing false positives.

• Recall: Assesses the algorithm's ability to identify all positive instances within a given class, indicating its effectiveness in minimizing false negatives.

• F1 score: Combines precision and recall into a single metric, providing a balanced measure of a classifier's performance

It is important to select appropriate evaluation metrics based on the specific problem domain and objectives of the machine learning algorithm. The choice of evaluation metrics should align with the desired outcomes and provide meaningful insights into the algorithm's performance. The performance of the machine learning model, could be compromised due to the issues such as class imbalance, for example, if the data set is dominated by the class-2D over the classes 0D and 1D. In this case, it is more advantageous to have a model that is able to predict the positive instances for each of those classes with high accuracy rather than using a metric that assesses overall predictions. Accuracy measures the overall correctness of the predictions, which is not a suitable performance metric when class distribution is imbalanced. In recall the score is calculated on the true positives and false negatives values. False negatives describe the ability of the model to identify the positive instances correctly. Recall is not crucial for this type of classification but may be important for certain scenarios such as medical diagnosis, where false negative class is more important. In this classification model, the focus is on achieving accurate predictions for each class that best match the definitions of precision where incorrectly predicted positive instance gives a significant reduction in the model performance. Therefore, precision over recall and accuracy were selected as performance metrics.

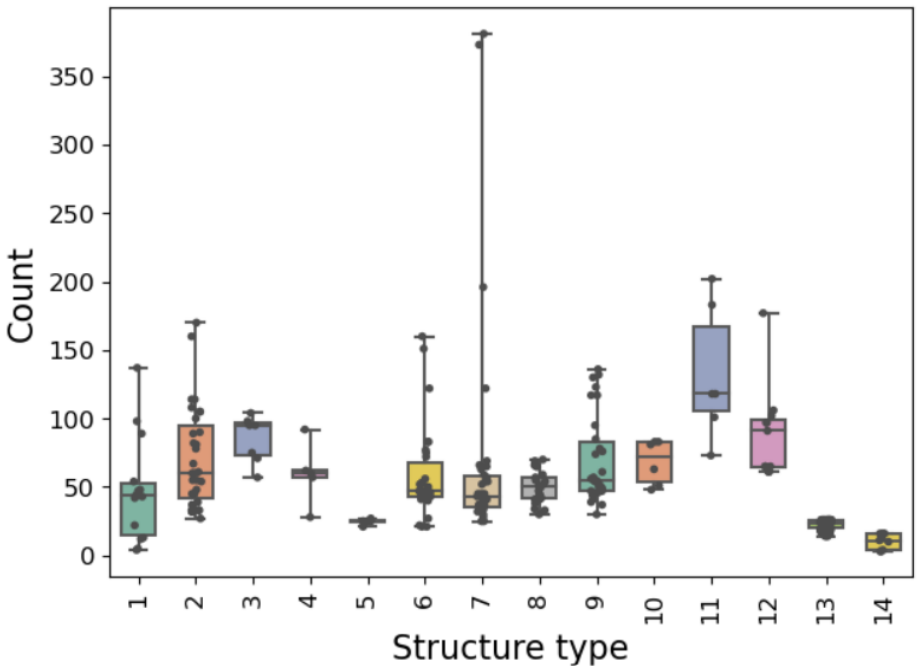


Figure S1. Boxplot of the number of reflections on the simulated XRD pattern in the angle range 3-30 ° 2Θ. The x axis shows the different structure types according to Table 1.

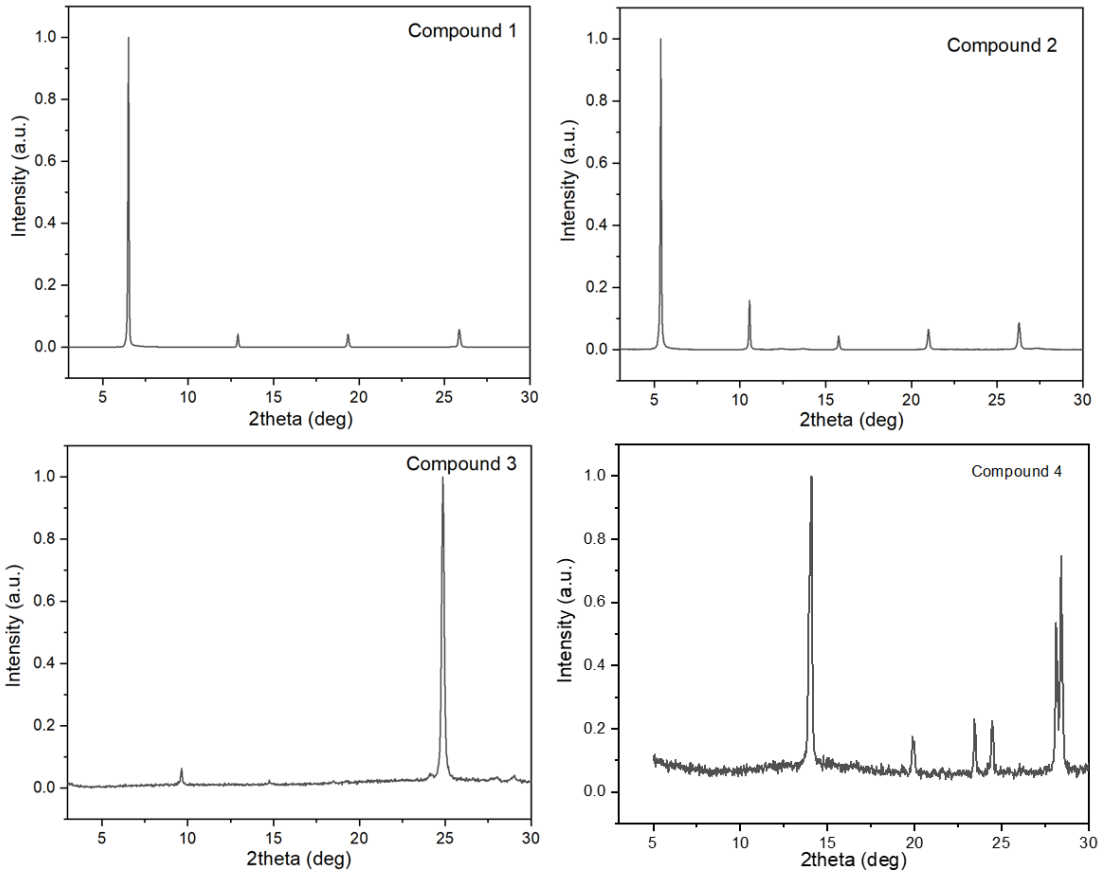


Figure S2. Experimental XRD patterns of compounds 1-4 (Table 2) from thin films at Cu Kα radiation using for testing of ML algorithms.

Synthesis of hybrid lead halide thin films

Compound 1: 1M solution of BAI and PbI_2 in 2:1 molar ratio was spin-coated at 6000 rpm for 20s without antisolvent. Film was annealed at 100 ºC for 10 min.

Compound 2: 1M solution of (F-PMA)I and PbI_2 in 2:1 molar ratio was spin-coated at 6000 rpm for 20s without antisolvent. Film was annealed at 100 ºC for 10 min.

Compound 3: 1M solution of GUAI, CsI and PbI_2 in 1:1:1 molar ratio was spin-coated at 6000 rpm for 20s without antisolvent. Film was annealed at 100 ºC for 10 min.

Compound 4: 1.5M solution of MAI and PbI2 in equimolar ratio in DMF/DMSO (4:1 v/v) was spin-coated onto previously cleaned glass substrates at 6000 rpm with quick addition of 100 µl chlorobenzene antisolvent at 10th second of spinning. Film then was annealed at 100ºC for 30 min.

The crystal structures determined by X-ray diffraction

The unit cell parameters and space groups were determined for synthesized compounds. The data obtained are consistent with the literature data on the structure refining of these compounds.

The crystal structures of these compounds belong to the 4 most common structure types: 2D inorganic substructure of (100) type with n=1 and (110) corrugated 2x2 motifs, 1D inorganic substructure with chains of octahedra connected along vertices and 3D crystal structures with perovskite structure type (Figure S3).

Compound 1. Composition – $[CH_3(CH_2)_3NH_3]_2PbI_4$ (BA_2PbI_4). The crystal structure was refined using the ordered model previously published by [1] in the *Pbca* space group. The refinement of the collected powder pattern led to lattice constants with *a* = 8.8632 Å, *b* = 8.6814 Å, *c* = 27.5692 Å and R_p = 2.56, R_{wp} = 8.28, GOF = 0.05.

Compound 2. Composition - $[FC_6H_5CH_2NH_3]_2PbI_4$ ($F\text{-}PMA)_2PbI_4$). The crystal structure was refined using the ordered model previously published by [2] in the $P2_1/c$ space group. The refinement of the collected powder pattern led to lattice constants with *a* = 8.6973 Å, *b* = 9.2461 Å, *c* = 27.5309 Å, β=97.603 ° and R_p = 3.11, R_{wp} = 11.15, GOF = 0.01.

Compound 3. Composition - $Cs[C(NH_2)_3]PbI_4$ CsGUAPbI$_4$). The crystal structure was refined using the ordered model previously published by [3] in the *Pnnm* space group. The refinement of the collected powder pattern led to lattice constants with a = 12.7425 Å, b = 18.6066 Å, c = 12.1767 Å and R_p = 3.29, R_{wp} = 12.48, GOF = 0.15.
Compound 4. $MAPbI_3$ (MA^+ – methylammonium). The crystal structure was refined previously using the disordered model previously published by [4] in the *P4/mcm* space group with cell parameters a = 8.85728 Å and c = 12.65104 Å. The XRD pattern of compound 4 is in good agreement with the model proposed in the [4] (Figure S3).

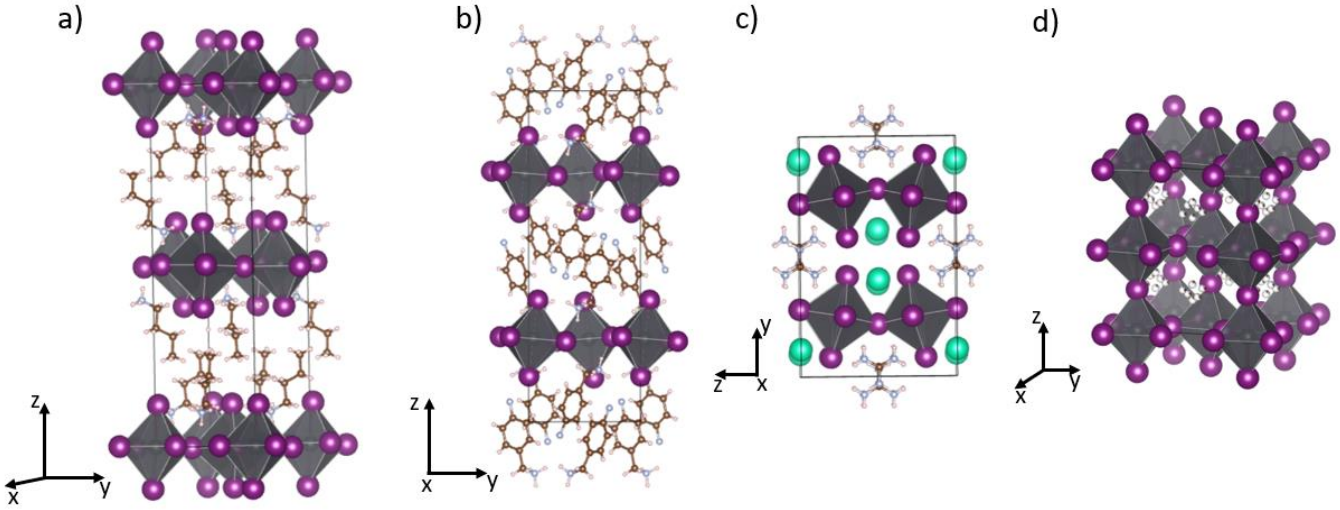


Figure S3. The crystal structures of synthesized compounds: a) BA_2PbI_4 – 2D (100), b) $(F\text{-}PMA)_2PbI_4$ – 2D (100), c) CsGUAPbI$_4$ 2D - (100), d) $MAPbI_3$ – 3D.

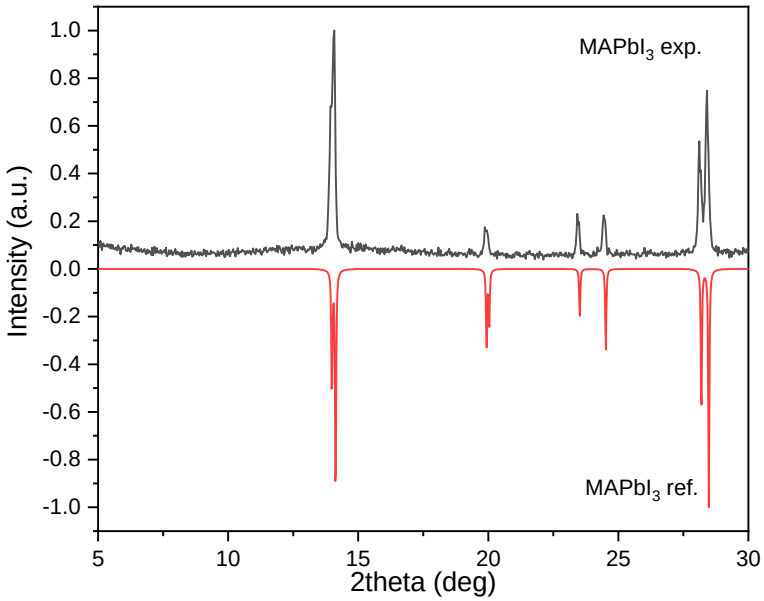


Figure S4. Comparison of $MAPbI_3$ experimental XRD pattern from this work (black) and the XRD pattern of model proposed in [4] (red).

Table S1. Comparison of the performance of different ML classification algorithms.

Dimensionality (3 classes)

	accuracy	balanced accuracy	precision	recall	F1 score	MCC	accuracy (exp. data)
DecisionTreeClassifier	0.86±0.05	0.84±0.08	0.87±0.06	0.86±0.05	0.86±0.06	0.78±0.09	11/11
RandomForestClassifier	0.888±0.03	0.84±0.04	0.89±0.04	0.888±0.030	0.882±0.031	0.81±0.06	11/11
ExtraTreesClassifier	0.884±0.027	0.82±0.05	0.890±0.029	0.884±0.027	0.871±0.035	0.81±0.05	8/11
XGBClassifier	0.90±0.06	0.88±0.08	0.90±0.06	0.90±0.06	0.89±0.06	0.83±0.10	10/11
CatBoostClassifier	0.89±0.06	0.87±0.07	0.90±0.06	0.89±0.06	0.89±0.06	0.82±0.10	11/11

Dimensionality (4 classes)

	accuracy	balanced accuracy	precision	recall	F1 score	MCC
DecisionTreeClassifier	0.76±0.07	0.67±0.13	0.81±0.06	0.76±0.07	0.78±0.07	0.65±0.10
RandomForestClassifier	0.85±0.06	0.66±0.09	0.84±0.07	0.85±0.06	0.83±0.08	0.76±0.10
ExtraTreesClassifier	0.846±0.033	0.65±0.06	0.84±0.05	0.846±0.033	0.83±0.05	0.76±0.05
XGBClassifier	0.85±0.05	0.69±0.06	0.85±0.05	0.85±0.05	0.84±0.05	0.76±0.07
CatBoostClassifier	0.86±0.06	0.70±0.09	0.87±0.06	0.86±0.06	0.86±0.05	0.78±0.09

Type of structure (14 classes)

	accuracy	balanced accuracy	precision	recall	F1 score	MCC
DecisionTreeClassifier	0.71±0.05	0.66±0.06	0.723±0.031	0.71±0.05	0.696±0.034	0.68±0.05
RandomForestClassifier	0.75±0.06	0.66±0.09	0.74±0.05	0.75±0.06	0.73±0.05	0.72±0.07
ExtraTreesClassifier	0.76±0.07	0.69±0.09	0.74±0.06	0.76±0.07	0.74±0.06	0.73±0.08
XGBClassifier	0.74±0.06	0.68±0.10	0.74±0.05	0.74±0.06	0.72±0.06	0.71±0.07
CatBoostClassifier	0.73±0.05	0.67±0.07	0.746±0.014	0.73±0.05	0.72±0.04	0.70±0.06

Type of structure (4 classes)

	accuracy	balanced accuracy	precision	recall	F1 score	MCC	accuracy (exp. data)
DecisionTreeClassifier	0.82±0.08	0.83±0.11	0.86±0.10	0.82±0.08	0.81±0.10	0.75±0.12	9/11
RandomForestClassifier	0.85±0.12	0.85±0.15	0.87±0.13	0.85±0.12	0.84±0.14	0.79±0.16	8/11
ExtraTreesClassifier	0.89±0.05	0.83±0.11	0.89±0.06	0.89±0.05	0.88±0.05	0.84±0.08	9/11
XGBClassifier	0.82±0.09	0.78±0.11	0.81±0.10	0.82±0.09	0.79±0.11	0.75±0.13	9/11
CatBoostClassifier	0.85±0.08	0.85±0.11	0.87±0.08	0.85±0.08	0.83±0.10	0.79±0.11	7/11

Connection of octahedra (6 classes)

	accuracy	balanced accuracy	precision	recall	F1 score	MCC
DecisionTreeClassifier	0.827±0.028	0.67±0.12	0.82±0.05	0.827±0.028	0.82±0.04	0.72±0.05
RandomForestClassifier	0.82±0.04	0.57±0.12	0.78±0.06	0.82±0.04	0.79±0.05	0.70±0.07
ExtraTreesClassifier	0.84±0.04	0.65±0.09	0.81±0.05	0.84±0.04	0.81±0.04	0.73±0.07
XGBClassifier	0.842±0.029	0.64±0.11	0.83±0.04	0.842±0.029	0.828±0.031	0.74±0.05
CatBoostClassifier	0.85±0.06	0.72±0.16	0.86±0.07	0.85±0.06	0.84±0.07	0.76±0.10

Table S2. Comparison of the performance of different ML classification algorithms on the enlarged dataset.

Type of structure (30 classes)

	accuracy	balanced accuracy	precision	recall	F1 score	MCC
DecisionTreeClassifier	0.633±0.018	0.59±0.08	0.64±0.04	0.633±0.018	0.626±0.026	0.603±0.019
RandomForestClassifier	0.701±0.031	0.63±0.06	0.66±0.04	0.701±0.031	0.670±0.035	0.671±0.034
ExtraTreesClassifier	0.72±0.04	0.65±0.07	0.70±0.04	0.72±0.04	0.70±0.04	0.70±0.04
XGBClassifier	0.697±0.025	0.63±0.06	0.704±0.035	0.697±0.025	0.689±0.030	0.671±0.026
CatBoostClassifier	0.673±0.032	0.67±0.06	0.72±0.07	0.703±0.032	0.70±0.05	0.68±0.04

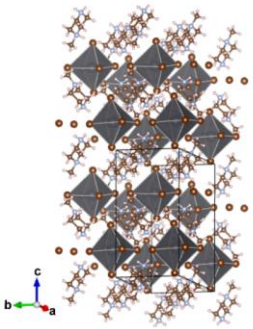
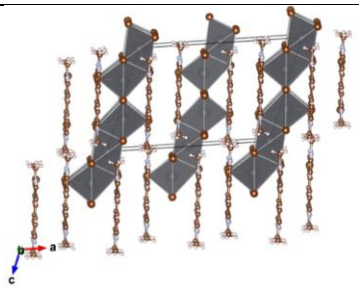
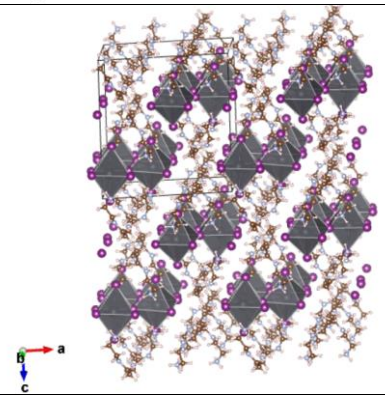
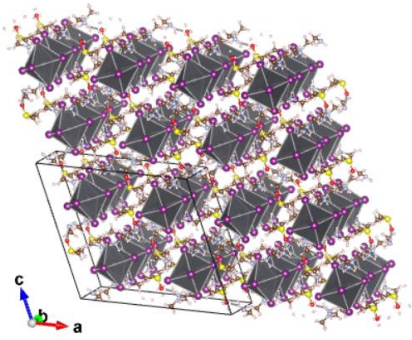
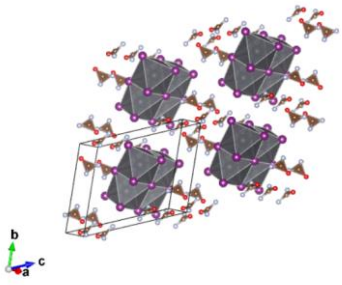
Connection of octahedra (6 classes)

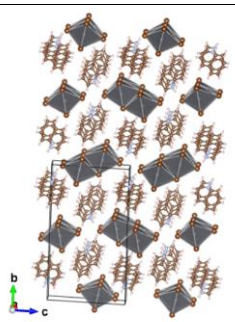
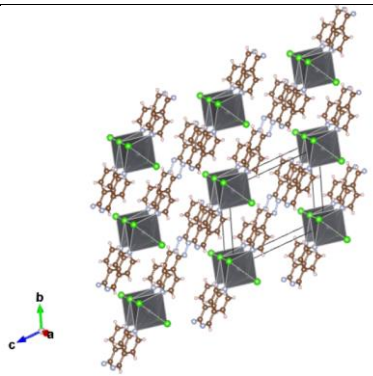
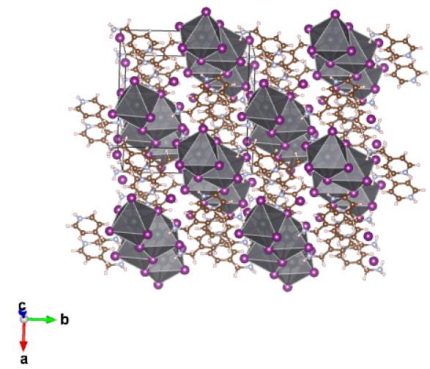
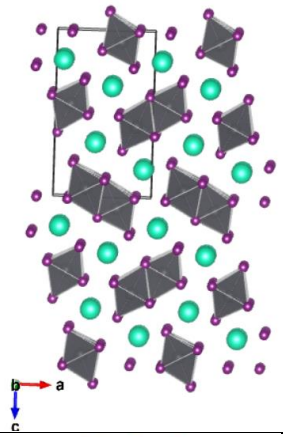
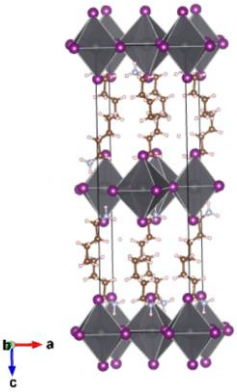
	accuracy	balanced accuracy	precision	recall	F1 score	MCC
DecisionTreeClassifier	0.74±0.05	0.54±0.07	0.76±0.04	0.74±0.05	0.74±0.04	0.59±0.06
RandomForestClassifier	0.791±0.027	0.50±0.05	0.75±0.04	0.791±0.027	0.760±0.027	0.65±0.04
ExtraTreesClassifier	0.799±0.018	0.56±0.06	0.764±0.019	0.799±0.018	0.772±0.014	0.664±0.027
XGBClassifier	0.77±0.04	0.54±0.05	0.762±0.033	0.77±0.04	0.764±0.034	0.64±0.06
CatBoostClassifier	0.807±0.027	0.61±0.04	0.804±0.028	0.807±0.027	0.803±0.025	0.69±0.04

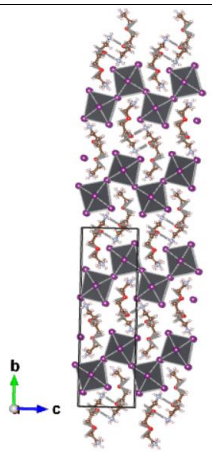
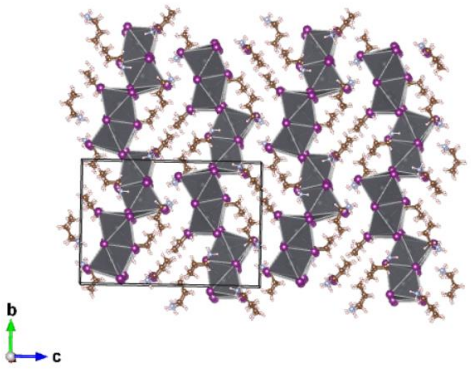
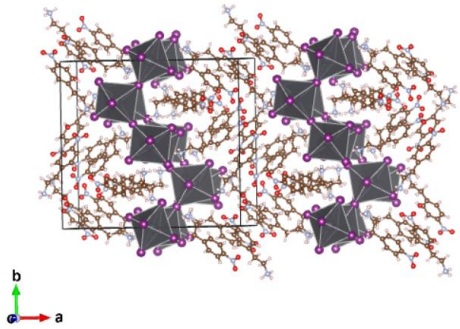
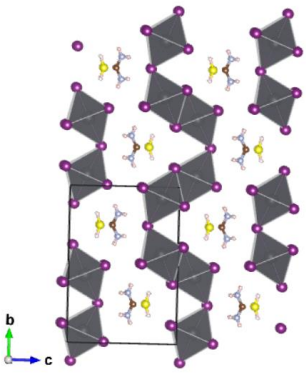
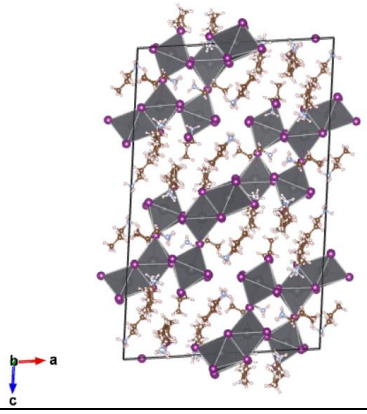
Dimensionality (4 classes)

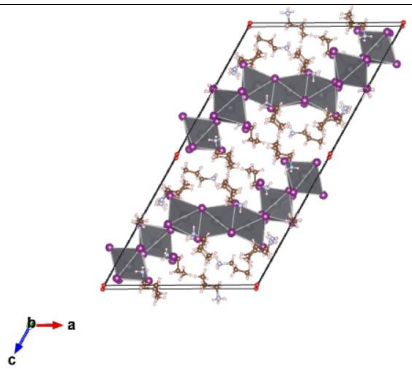
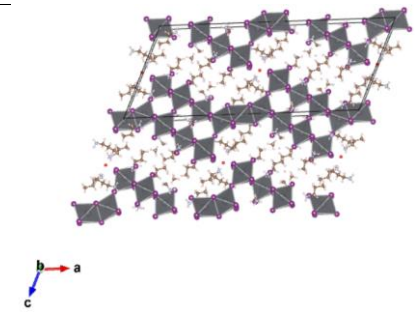
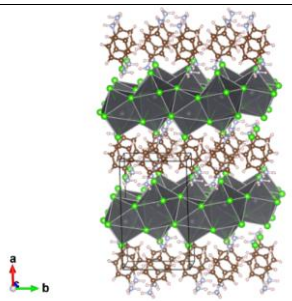
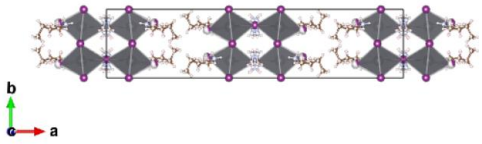
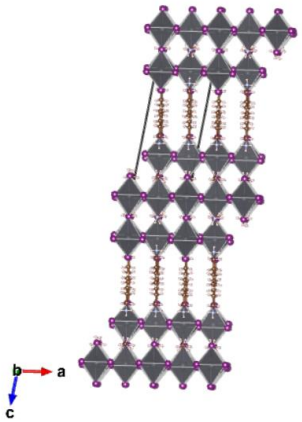
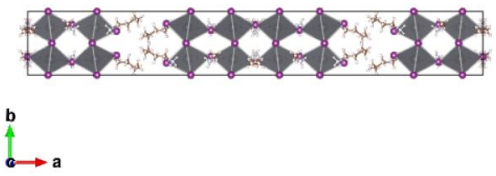
	accuracy	balanced accuracy	precision	recall	F1 score	MCC
DecisionTreeClassifier	0.820±0.022	0.72±0.07	0.835±0.021	0.820±0.022	0.825±0.020	0.712±0.034
RandomForestClassifier	0.861±0.024	0.71±0.05	0.843±0.028	0.861±0.024	0.846±0.023	0.77±0.04
ExtraTreesClassifier	0.857±0.033	0.68±0.04	0.83±0.04	0.857±0.033	0.840±0.034	0.76±0.05
XGBClassifier	0.863±0.024	0.75±0.04	0.860±0.020	0.863±0.024	0.859±0.021	0.78±0.04
CatBoostClassifier	0.853±0.029	0.72±0.04	0.855±0.022	0.853±0.029	0.851±0.025	0.76±0.05

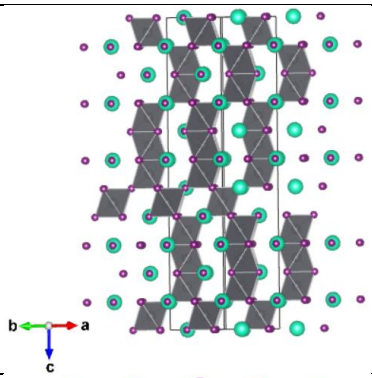
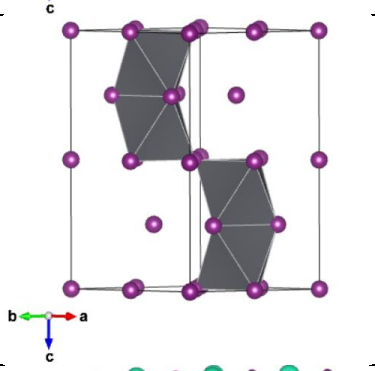
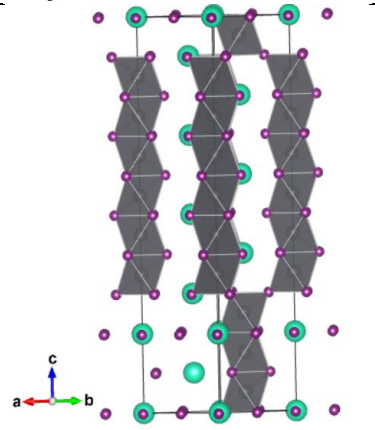
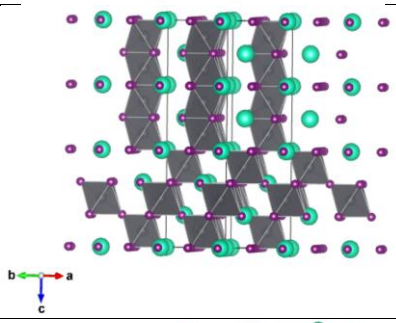
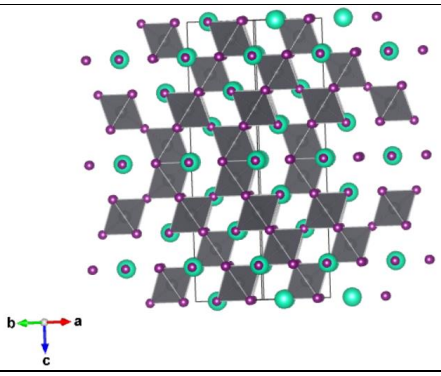
Table S3. Structure types and their number containing in dataset 2.

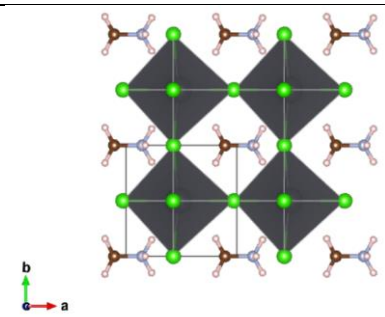
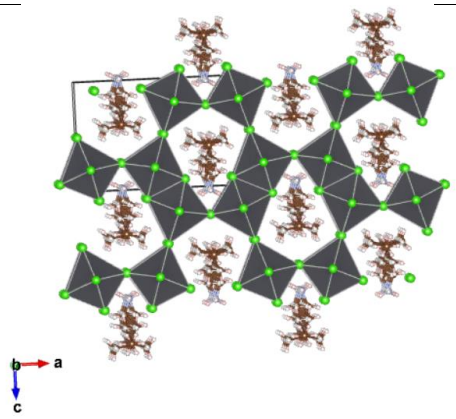
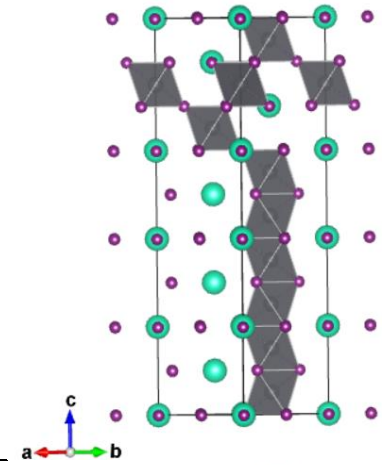
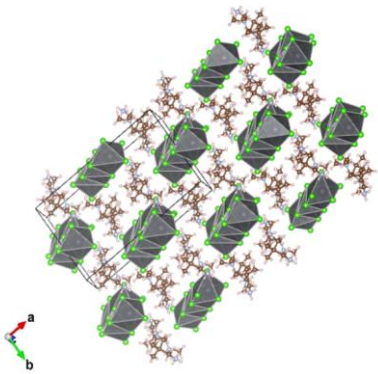
№	Structure type	number of structures in the dataset 2
1		21
2		8
3		61
4		5
5		4

6		10
7		8
8		8
9		7
14		91

15		52
16		6
18		7
19		3
20		3

21		3
22		4
24		3
26		49
27		37
28		11

29		21
30		2
31		16
33		3
34		6

35		18
36		3
37		7
38		8

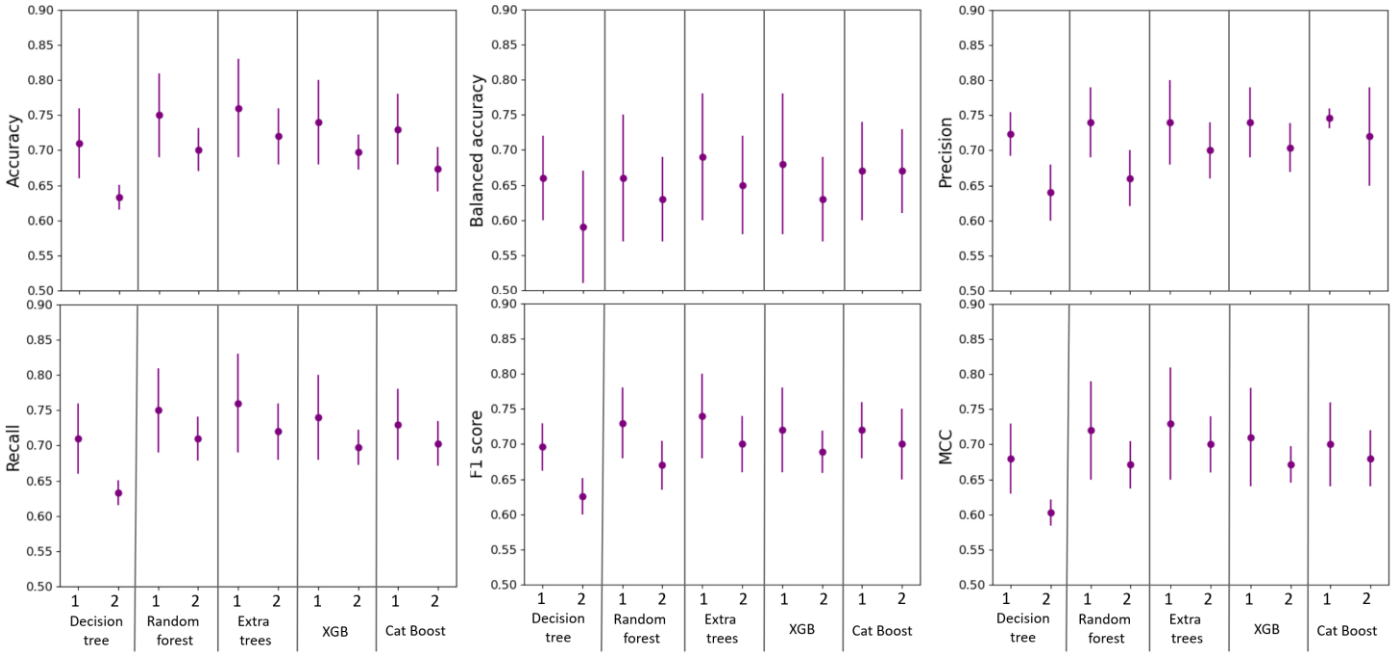


Figure S5. Comparison of model performance of different ML algorithms of classifications XRD patterns by structure type on 14 classes (1) and 30 classes (2). For classification into 14 classes, the model used dataset 1 consisting of 272 structures; for classification into 30 classes, the enlarged dataset 2 containing 485 structures was used. The numbers 1 and 2 indicate dataset 1 and dataset 2 and ML model performance for classification on 14 and 30 classes respectively.

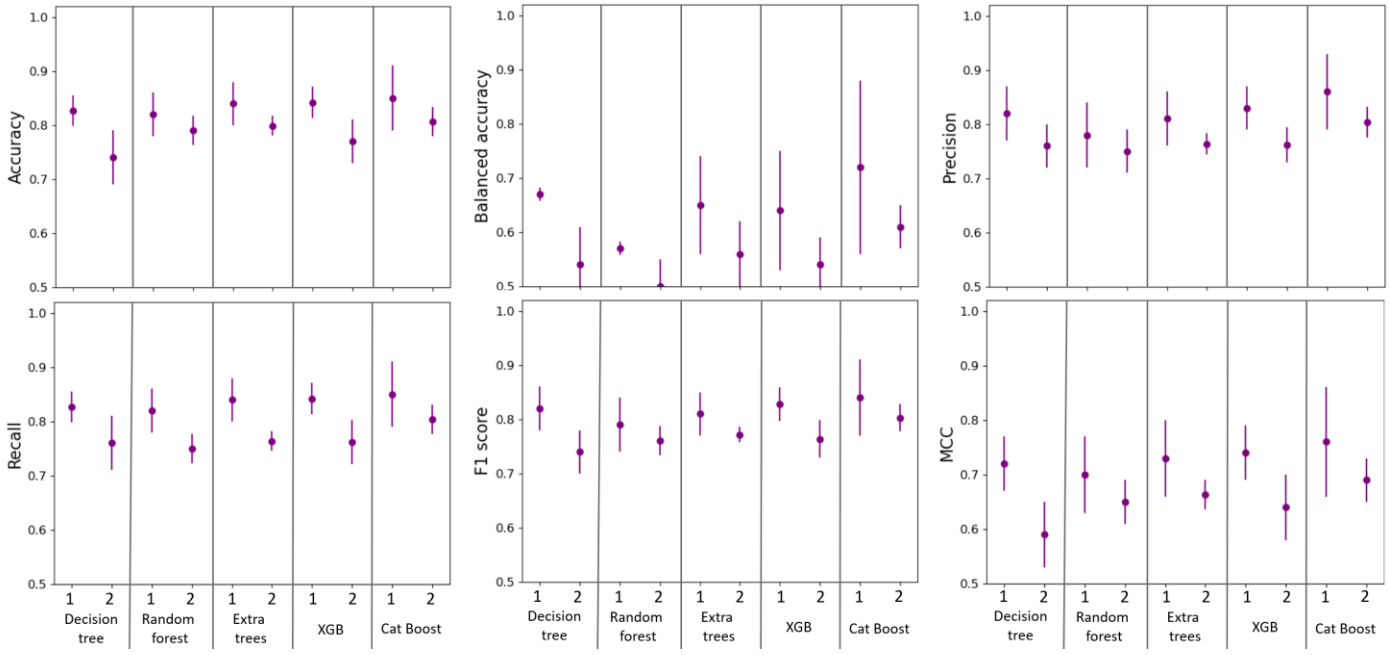


Figure S6. Comparison of the performance of different ML classification algorithms in identifying octahedral connectivity types from XRD patterns across six classes. The numbers 1 and 2 represent dataset 1 and dataset 2, respectively

References

1. *Mitzi D.B.* // 1996. № 4. P. 791.

2. *Tremblay M.H., Bacsa J., Zhao B. et al.* // Chem. Mater. 2019. V. 31. № 16. P. 6145.

3. *Nazarenko O., Kotyrba M.R., Wörle M. et al.* // Inorg. Chem. 2017. V. 56. № 19. P. 11552.

4. López C.A., Alvarez-Galván M.C., Martínez-Huerta M.V. et al. // CrystEngComm 2020. V. 22. № 4. P. 767.