Mendeleev Communications

Mendeleev Commun., 2025, **35**, 383–385

Topological representation of layered hybrid lead halides for machine learning using universal clusters

Ekaterina I. Marchenko,*[a,b] Maria G. Khrenova,[c] Vadim V. Korolev,[d] Eugene A. Goodilin[a,c] and Alexey B. Tarasov[a,c]

[a] *Department of Materials Science, M. V. Lomonosov Moscow State University, 119991 Moscow, Russian Federation. E-mail: marchenko-ekaterina@bk.ru*

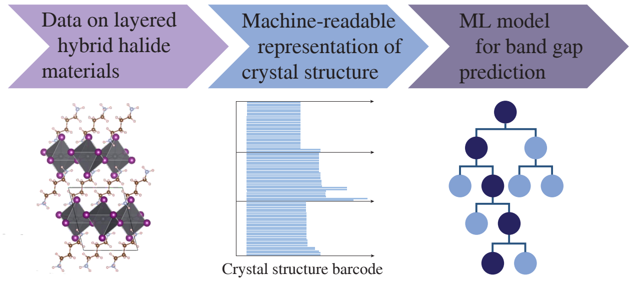
[b] *Department of Geology, M. V. Lomonosov Moscow State University, 119991 Moscow, Russian Federation*

[c] *Department of Chemistry, M. V. Lomonosov Moscow State University, 119991 Moscow, Russian Federation*

[d] *Institute for Artificial Intelligence, M. V. Lomonosov Moscow State University, 119192 Moscow, Russian Federation*

DOI: 10.71267/mencom.7653

Prediction of band gaps in layered hybrid halide compounds promising for photovoltaic and optoelectronic applications was performed using a machine learning approach. In order to facilitate the discovery and design of new hybrid halide materials with tailored electronic properties, machine learning models were enhanced with invariant topological representations of these materials using the atom-specific persistent homology method.


Data on layered hybrid halide materials
Machine-readable representation of crystal structure
ML model for band gap prediction
Crystal structure barcode

Keywords: hybrid halide perovskites, topological representations, band gaps, structure–property relationships, machine learning.

The design and discovery of novel materials with tailored electronic properties are crucial to the advancement of fields such as photovoltaics, optoelectronics and energy storage. One class of such materials involves layered hybrid lead halide compounds with perovskite-derived crystal structures, or lead halide perovskites (LHPs). They have attracted considerable attention due to their tunable band gaps, which are essential for optimizing their performance in electronic and energy-related applications.[1–5] However, accurate theoretical prediction of the band gaps in these materials remains challenging due to the complex interplay of their atomic and electronic structures.[6,7]

In recent years, machine learning has emerged as a powerful tool to accelerate materials discovery by predicting key properties such as the band gap in LHPs from structural data.[8–10] The band gap is known to depend on a number of geometric descriptors in LHPs, including metal–halogen bond lengths, bond angles between atoms in the inorganic substructure, layer shift factor and some others.[6,11–13] However, modern machine learning algorithms are plausible to input crystal structure information written universally, such as a multidimensional vector. Topological representation methods that capture the spatial arrangement and connectivity of atoms in a material have shown promise in enhancing the accuracy of machine learning models. In this context, developing efficient topological descriptors for hybrid halide compounds could significantly improve our ability to predict their electronic behavior. This article explores the topological representation[14] of layered hybrid lead halide compounds and its application to machine learning models for band gap prediction.

In this work, we utilized a dataset comprising 140 two-dimensional perovskite-related crystal structures exhibiting the (100) structural type, characterized by a perovskite block thickness of $n = 1$. This dataset[†] was sourced from the published work,[8] wherein the band gap values had previously been calculated with high precision using density functional theory (DFT) methods. Each material in the database is encoded in the Crystallographic Information File (CIF) format and is available for unrestricted access. The website provides comprehensive descriptions of the materials' properties, including DFT-derived and experimental band gap values, chemical formulas, space groups and other relevant data.

Figure 1 illustrates the construction of barcodes for two-dimensional perovskite materials. Our methodology encompasses several essential steps. Initially, from the crystallographic information files (.cif) contained in the dataset, we systematically extract various types of atoms occupying different crystallographic sites and their combinations within the unit cell. Around each atom in the unit cell, a sphere is constructed within a cutoff radius, which contains a cloud of points (atoms). Then, for each sphere, the number of bonds and the distances between atoms are calculated, which are reflected in the barcode as lines.[‡] The number of lines

[†] The complete dataset on the two-dimensional perovskites employed in this investigation is available on the official website of the NMSE database: http://www.pdb.nmse-lab.ru/.

[‡] To generate the barcodes, we set a cutoff radius of 10 Å for the point cloud region surrounding each atom. For LHP structures, this radius is optimal since interactions between adjacent layers of lead–halogen octahedra are possible up to such distances. By calculating the bond lengths from the central atom to its neighboring atoms, the resulting barcode encapsulates the geometric and topological characteristics of the structure. In Figure 1, the solid lines indicate contacts in pairs of atoms within the designated sphere, the number of lines corresponds to the frequency of these contacts in the unit cell, while the line lengths represent the distances between atoms. The order of the lines in the barcode can be chosen arbitrarily. In addition

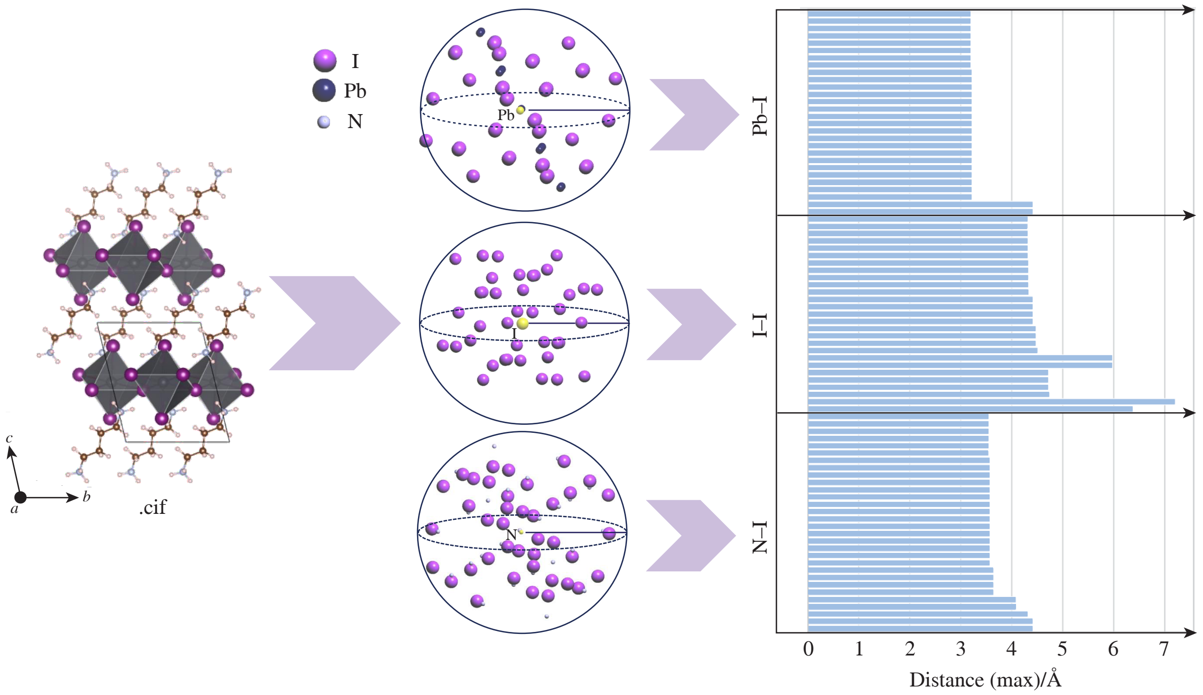


Figure 1 The main stages of constructing a barcode of the LHP crystal structure. Crystal structures were visualized using the Vesta[16] and TOPOSpro[17] programs.

means the number of bonds, and their lengths represent the interatomic distances. In this way, we obtain an atom-specific topological fingerprint representation[14] for the LHP to extract detailed crystal information pertinent to machine learning applications. The topology of these structures is encoded as specific barcodes. Utilizing this topological representation, we applied the gradient boosting tree model[15] to predict the band gaps of the materials.

The methodology and algorithm for constructing specific barcodes in Python were developed as outlined previously.[14] The fundamental concept underlying this approach is that there are a limited number of atoms in a unit cell, each characterized by a distinctive structural environment that defines its unique topological fingerprints. This approach is universal and invariant since heterogeneous parameters such as unit cell parameters, angles and atomic coordinates are not involved in representing crystal structure descriptors. All geometric parameters in this representation are presented homogeneously as interatomic distances in the local environment of each atom. As an example, for the (100) layered hybrid lead halide crystal structure with $n = 1$, combinations of three significant atomic pairs Pb–X, X–X and N–X (X is a halogen atom) are identified as having a substantial impact on the band structure of the material.[12,18] The changes in interatomic distances in these three pairs make a significant contribution to the change in the LHP band gaps compared to the changes in distances in other atomic pairs.

Thus, as a result of converting the classical CIF of the crystal structure into a barcode, we obtain a data set that is easy to represent in a machine-readable form as a multidimensional vector due to the homogeneity of the data representation. Such information is easy to process using modern libraries for machine learning. Compared with other machine-readable crystal structure representations based on structure graphs and Coulomb matrices,[19] topological descriptors using persistent homology have the advantage of uniquely encoding structures at both local and global levels without requiring assumptions about the underlying physics. We chose gradient boosted regression trees (GBRT)[15] as the machine learning algorithm to evaluate the accuracy, robustness and efficiency of the topological-based features.[§] The performance metrics for the model predicting band gaps using topological feature vectors were as follows: $R^2 = 0.8$, RMSE = 0.17 eV and MAE = 0.12 eV (Figure 2). These results are consistent with contemporary machine learning models aimed at predicting the band gaps in hybrid perovskites.[8,10] Furthermore, a commendable MAE was achieved despite the limited size of the dataset. Thus, the representation of crystal LHP structures as barcodes is a good general-purpose machine-readable representation for the targeted design of this class of materials.

Beyond LHP materials, this approach presents opportunities for addressing both the direct problem (predicting the physical properties of materials from their crystal structure) and the inverse problem (predicting crystal structures with desired properties) for other hybrid materials related to the group of hybrid lead halides, including those with 3D, 1D and 0D inorganic substructures. Future advancements in this methodology will focus on predicting and decoding barcodes into potential sets of promising crystal structures.

to the topological information, composition-based features were incorporated, which include stoichiometric attributes reflecting elemental fractions, elemental property statistics derived from all atoms in the crystal and electronic structure attributes.

[§] GBRT effectively integrates multiple weak predictors to formulate a robust model. The training process involves sequentially adding trees to diminish the loss function of the current model. To mitigate overfitting, each model update utilizes various randomly selected subsets of both training data and features. Hyperparameter optimization was performed through cross-validation, evaluated using the R^2 metric. The hyperparameters used in GBRT include: *n_estimators* = 300 000, *learning_rate* = 0.001, *max_depth* = 7, *min_samples_split* = 5, *subsample* = 0.85 and *max_features* = sqrt. The machine learning models were constructed using scikit-learn software (version 0.19.2) as indicated in the previous work.[20] A ten-fold cross-validation approach was utilized to validate the proposed methodology, with random splitting of the data repeated 20 times to assess the robustness of the model. The median performance metrics and standard deviation across these repeated experiments were documented. Voronoi tessellations and Coulomb matrices were replicated using Magpie, which is freely available under an open-source license.[21]

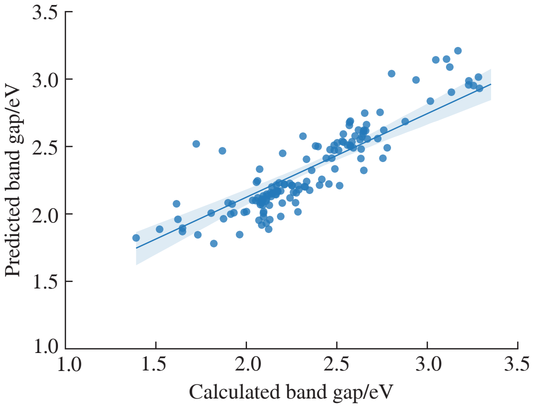


Figure 2 Comparison of the band gap values calculated by DFT with those predicted by the machine learning algorithm for 2D hybrid lead halide materials.

In summary, we have shown that the topological representation of crystal structures is suitable as a machine-readable representation of the organic–inorganic periodic structures for machine learning algorithms. Using this invariant representation, machine learning algorithms successfully predict composition–structure–property relationships for LHP materials.

This work was supported by the Interdisciplinary Scientific and Educational Schools of Lomonosov Moscow State University (grant no. 23-Sh03-04).

References

1 J.-C. Blancon, J. Even, C. C. Stoumpos, M. G. Kanatzidis and A. D. Mohite, *Nat. Nanotechnol.*, 2020, **15**, 969; https://doi.org/10.1038/s41565-020-00811-1.

2 J. Huang, Y. Yuan, Y. Shao and Y. Yan, *Nat. Rev. Mater.*, 2017, **2**, 17042; https://doi.org/10.1038/natrevmats.2017.42.

3 W. Li, Z. Wang, F. Deschler, S. Gao, R. H. Friend and A. K. Cheetham, *Nat. Rev. Mater.*, 2017, **2**, 16099; https://doi.org/10.1038/natrevmats.2016.99.

4 G. Grancini and M. K. Nazeeruddin, *Nat. Rev. Mater.*, 2019, **4**, 4; https://doi.org/10.1038/s41578-018-0065-0.

5 L. Mao, C. C. Stoumpos and M. G. Kanatzidis, *J. Am. Chem. Soc.*, 2019, **141**, 1171; https://doi.org/10.1021/jacs.8b10851.

6 E. I. Marchenko, V. V. Korolev, S. A. Fateev, A. Mitrofanov, N. N. Eremin, E. A. Goodilin and A. B. Tarasov, *Chem. Mater.*, 2021, **33**, 7518; https://doi.org/10.1021/acs.chemmater.1c02467.

7 Z. Wan, Q.-D. Wang, D. Liu and J. Liang, *New J. Chem.*, 2021, **45**, 9427; https://doi.org/10.1039/D1NJ01518D.

8 E. I. Marchenko, S. A. Fateev, A. A. Petrov, V. V. Korolev, A. Mitrofanov, A. V. Petrov, E. A. Goodilin and A. B. Tarasov, *Chem. Mater.*, 2020, **32**, 7383; https://doi.org/10.1021/acs.chemmater.0c02290.

9 W. B. Park, J. Chung, J. Jung, K. Sohn, S. P. Singh, M. Pyo, N. Shin and K.-S. Sohn, *IUCrJ*, 2017, **4**, 486; https://doi.org/10.1107/S205225251700714X.

10 C.-S. Hu, R. Mayengbam, M.-C. Wu, K. Xia and T. C. Sum, *Commun. Mater.*, 2024, **5**, 106; https://doi.org/10.1038/s43246-024-00545-w.

11 E. I. Marchenko, V. V. Korolev, A. Mitrofanov, S. A. Fateev, E. A. Goodilin and A. B. Tarasov, *Chem. Mater.*, 2021, **33**, 1213; https://doi.org/10.1021/acs.chemmater.0c03935.

12 E. I. Marchenko, S. A. Fateev, A. A. Ordinartsev, P. A. Ivlev, E. A. Goodilin and A. B. Tarasov, *Mendeleev Commun.*, 2022, **32**, 315; https://doi.org/10.1016/j.mencom.2022.05.007.

13 E. I. Marchenko, E. A. Kobeleva, N. N. Eremin, E. A. Goodilin and A. B. Tarasov, *Mendeleev Commun.*, 2024, **34**, 650; https://doi.org/10.1016/j.mencom.2024.09.008.

14 Y. Jiang, D. Chen, X. Chen, T. Li, G.-W. Wei and F. Pan, *npj Comput. Mater.*, 2021, **7**, 28; https://doi.org/10.1038/s41524-021-00493-w.

15 B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary and C. Wolverton, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2014, **89**, 094104; https://doi.org/10.1103/PhysRevB.89.094104.

16 K. Momma and F. Izumi, *J. Appl. Crystallogr.*, 2011, **44**, 1272; https://doi.org/10.1107/S0021889811038970.

17 V. A. Blatov, A. P. Shevchenko and D. M. Proserpio, *Cryst. Growth Des.*, 2014, **14**, 3576; https://doi.org/10.1021/cg500498k.

18 E. I. Marchenko, S. A. Fateev, V. V. Korolev, V. Buchinskiy, N. N. Eremin, E. A. Goodilin and A. B. Tarasov, *J. Mater. Chem. C*, 2022, **10**, 16838; https://doi.org/10.1039/D2TC03202C.

19 S. Li, Y. Liu, D. Chen, Y. Jiang, Z. Nie and F. Pan, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2022, **12**, e1558; https://doi.org/10.1002/wcms.1558.

20 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 and É. Duchesnay, *The Journal of Machine Learning Research*, 2011, **12**, 2825; https://dl.acm.org/doi/10.5555/1953048.2078195.

21 L. Ward, A. Agrawal, A. Choudhary and C. Wolverton, *npj Comput. Mater.*, 2016, **2**, 16028; https://doi.org/10.1038/npjcompumats.2016.28.

Received: 16th October 2024; Com. 24/7653