Role of Hydrogen Bonding on the Design of New Hybrid Perovskites Unraveled by Machine Learning

Rachid Laref, Florian Massuyeau, Romain Gautier

Role of Hydrogen Bonding on the Design of New Hybrid Perovskites Unraveled by Machine Learning

Rachid Laref, Florian Massuyeau, Romain Gautier**

Centre National de la Recherche Scientifique (CNRS), IMN, 2 rue de la Houssinière, 44322 Nantes, France
E-mail: florian.massuyeau@cnrs.fr, romain.gautier@cnrs-imn.fr

Keywords: Hybrid lead halide perovskite, Hydrogen bonding, Machine learning, Classification, Descriptors.

Abstract

Selecting a set of reactants to accurately design a new low dimensional hybrid perovskite could greatly accelerate the discovery of materials with great potential in photovoltaics, or solid-state lighting. However, this design is challenging as most hybrid metal halides are not perovskites and no feature has been clearly associated to the structural characteristics of the inorganic metal halide network. In this work, we firstly demonstrated that the organic molecules are key parameters to determine the structure type of the inorganic network (i.e. perovskite vs. non-perovskite). Then, machine learning (ML) algorithms were used to identify the key features of the organic cations leading to the perovskite structure type. Using a large dataset of hybrid metal halides, we extracted the organic molecules of all hybrid lead halide compounds, calculated 2756 molecular descriptors and fingerprints for each of these organic molecules, and were able to predict through ML techniques if a specific organic amine will lead to the perovskite type with an accuracy up to 88.65%. Descriptors related to hydrogen bonding were identified as important features. Thus, a simple but reliable design principle could be demonstrated: the presence of primary ammonium cation is the primary condition to prepare hybrid lead halide perovskites regardless of their dimensionalities.

1. Introduction

In the past decade, hybrid lead halide perovskites, have gained a lot of attention, owing to their remarkable properties that are suitable for diverse applications.[1–5] Thus, these materials have been largely investigated for photovoltaics applications due to their direct band gap, large absorption coefficients, and long diffusion lengths of the exciton.[6–8] Moreover, their tunable band gap along with high photoluminescence quantum yields make them very promising materials for the next generation of light devices.[9–11] More recently, hybrid lead halides perovskites have also been investigated for ferro- and piezo-electric, magnetoresistive and catalysis applications.[12–14] Thus, predicting the design of a new hybrid metal halide perovskite prior its synthesis could accelerate the discovery of these functional materials.

Predicting the structure type of a new hybrid metal halide compound from its constituents is challenging. With the exception of 3D hybrid perovskites which can be reliably predicted using the Goldschmidt tolerance factor,[15] the key structural features of the organic cation allowing the prediction of hybrid lead halide perovskites with lower dimensionalities (2D, 1D) remain unclear. A general intuition of some chemists with a long experience in the synthesis of new hybrid lead halides is that primary amines lead to 2D hybrid perovskites. However, to our knowledge, this intuition has never been clearly stated in the literature because counterexamples exist (some primary amines lead to non-perovskites and some secondary amines lead to perovskites). In addition, contrary information have been reported in the literature. For example, Zhang et al. reported that some main factors of the spacer used to obtain 2D perovskites are the net positive charge, the degree of substitution of the perovskite anchoring site, the hydrogen-bonding ability, the space-filling ability and the stereo chemical configuration.[16] In addition, prior studies analyzing the relationships between organic cations and the lead halide networks are all based on the hypothesis that the size and shape of the organic cations are important parameters.[17–19] It is also important to note that previous works only analyzed a limited numbers (typically from three to ten) of structures with a limited number of features selected based on the expertise and intuition of the chemists.

In this article, we use machine learning (ML) algorithms to explore a very large number of features without relying on human intuition or hypothesis. We apply these algorithms to a large number of structures (626 structures from the CSD), in order to investigate the effects of the organic cation on the structural characteristics of the inorganic network of hybrid lead halide materials. 2756 features (molecular descriptors and fingerprints) were calculated to describe, for each organic cations, all structural properties aspects (see the methods section for details). These molecular descriptors have been used for decades in pharmaceutical chemistry

to guide the design of new drugs.[20] In contrary, the use of such descriptors in materials science remains rare.[21] After building and analyzing machine learning models, we identified a strong correlation between the perovskite structure type of hybrid lead halide materials and the descriptors related to hydrogen bonding. Our results suggest that the number of primary ammonium on the organic cations can be used as a simple and efficient descriptor to predict the formation of hybrid lead halide perovskite materials.

2. Results and Discussion

2.1. Data analysis

In our dataset, 385 different organic cations involved in 626 different crystal structures of hybrid lead halides were extracted (see SI for the list of all SMILES and CSD Refcodes). The histograms represented in Figure 1a summarizes the number of distinct structures that could be obtained using a same organic cation. It is interesting to note that 365 organic cations lead exclusively to perovskite or non-perovskite compounds while only 20 organic cations lead to both perovskite and non-perovskite compounds. This statistic shows that the organic cation is the determining parameter linked to the structure type of the inorganic network. Moreover, organic cations that lead to more than one phase very likely lead to a single type of hybrid lead halide network (either perovskite or non-perovskite). This observation further confirms the crucial role of the organic cation on the crystallization of perovskite materials.

The review of the dataset also reveals that some descriptors and fingerprints had null values for all structures. For instance, the descriptor counting the number of boron atoms is equal to zero for all structures because none of these structures contain boron atoms. In total, we removed 268 null descriptors and 421 null fingerprints because they could interfere with the model rather than provide useful information.

2.2. Prediction of the structure type by ML.

Using the built dataset, two analyses were carried out independently: (i) Identification of *the best combination of ten descriptors* which aims at reaching the highest possible accuracy for the ML prediction, and (ii) Identification of *the ten best individual descriptors* which aims at identifying the key parameters to discriminate between perovskite and non-perovskite.

For analysis (i), feature selection process was adopted to select the best combination of ten descriptors that provided the best accuracy. This process is called sequential feature selection (SFS) and is provided by scikit-learn.[22] It is used in forward direction, meaning that the algorithm start without feature, choose the best one (best accuracy) and select one

supplementary feature at each stage to maximize the score until the number of features to select is obtained (in our case, 10). The selected descriptors obtained through this process were: C1SP2 (Doubly bound carbon bound to one other carbon), nHsNH3p (Count of atom-type H E-State: $-NH_3^+$), C1SP3 (Singly bound carbon bound to one other carbon), MDEN-22 (Molecular distance edge between all secondary nitrogens), MPC5 (Molecular path count of order 5), topoRadius (Topological radius (minimum atom eccentricity), PubchemFP441 (C(-C)(=C)), PubchemFP697 (C-C-C-C-C-C(C)-C), PubchemFP737 (Cc1cc(N)ccc1), PubchemFP528 ([#1]-N-C-[#1]) (see *list of descriptors* and *list of fingerprints* in SI).

To evaluate the importance of each selected descriptor on the classification task, we used classes of functions with random forest classifier from Scikit-Learn (Figure 1b). The plot suggests that the descriptor nHsNH3p (number of primary amine groups in the organic cation) is the most informative descriptor followed by the fingerprints PubchemFP737 and PubchemFP528. These two fingerprints indicate the presence of substructures containing nitrogen atom in the organic cation (see *list of fingerprints* in SI).

During the learning and validation steps, a k-fold cross validation procedure with k=10 was adopted. The dataset was split into ten subsets, Nine subsets were used for training while one was used for testing. This process was repeated ten times, using a different subset for testing, so that each subset was used for testing once. The Optuna library from Python was used to optimize the hyper-parameters of each machine learning algorithm (See table S1).[23]

The performances of several machine learning algorithms were compared in terms of classification accuracy. It is defined as the number of structures correctly classified over the number of total structure presented in the test set. At the end of the cross-validation process, the average accuracy was calculated across the ten rounds. Figure 1c represents the boxplot of the classification accuracies obtained with different machine learning algorithms available in Scikit-learn: K-Nearest Neighbors (KNN), Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM) and Gradient Boosting Decision Tree (GBDT). One can observe that the GBDT and RF exhibit the best performances in terms of accuracy, with a mean of 88.65% across 10 rounds of cross-validation. It is important to note that this accuracy is relatively high and enables to guide efficiently the design of hybrid perovskites, especially as ca. 5 % of reported organic molecules lead to both perovskite and non-perovskite structures.

As the analysis of the feature importances (analysis (i)) showed that some descriptors are very informative while others are not, an extensive analysis including all descriptors and fingerprints was carried out (analysis (ii)). The objective of analysis (ii) was to better identify

the important descriptors and the relationships between them. Therefore, a decision tree model was trained using one descriptor at a time, and this process was repeated for all descriptors and fingerprints. Figure 1d shows the accuracy for each of the 2756 descriptors and fingerprints. Only a few descriptors have an accuracy above 80%. The full list of accuracies for each descriptor and fingerprint is provided in the supplementary materials. One can note that the best descriptors in terms of accuracy are related to the number of $-NH_3^+$ in the organic cation, with an accuracy of 84.99%. It is interesting to note that this accuracy is close to the accuracy of the ML prediction using the best combination of ten descriptors (88.65% - see analysis (i)). Thus, this demonstrates that the individual descriptor related to the number of $-NH_3^+$ in the organic cation are making most of the prediction. This top descriptor is followed by the H-bond Lipinski descriptor, defined as the sum of the number of N-H bonds and O-H bonds with an accuracy of 82.27%. Additionally, the top 10 best descriptors (whose accuracies are depicted as red dots in Figure 1d) are correlated to each other and all of them are related to hydrogen bonding (Figure 1e – See *list of descriptors* in SI for the definition of descriptors). One can notice that the WTPT-5- descriptor, which indicates the sum of path lengths starting from nitrogen atoms, is not correlated positively to the other nine descriptors. However, the WTPT-5- descriptor is related to the other nine descriptors since the hydrogen bond donors is related to the nitrogen atom. Therefore, the sum of path lengths starting from nitrogen atom (WTPT-5-) is lower for higher numbers of hydrogen bonding donors. Such relationship could also explain a weak inverse correlation with the other nine descriptors.

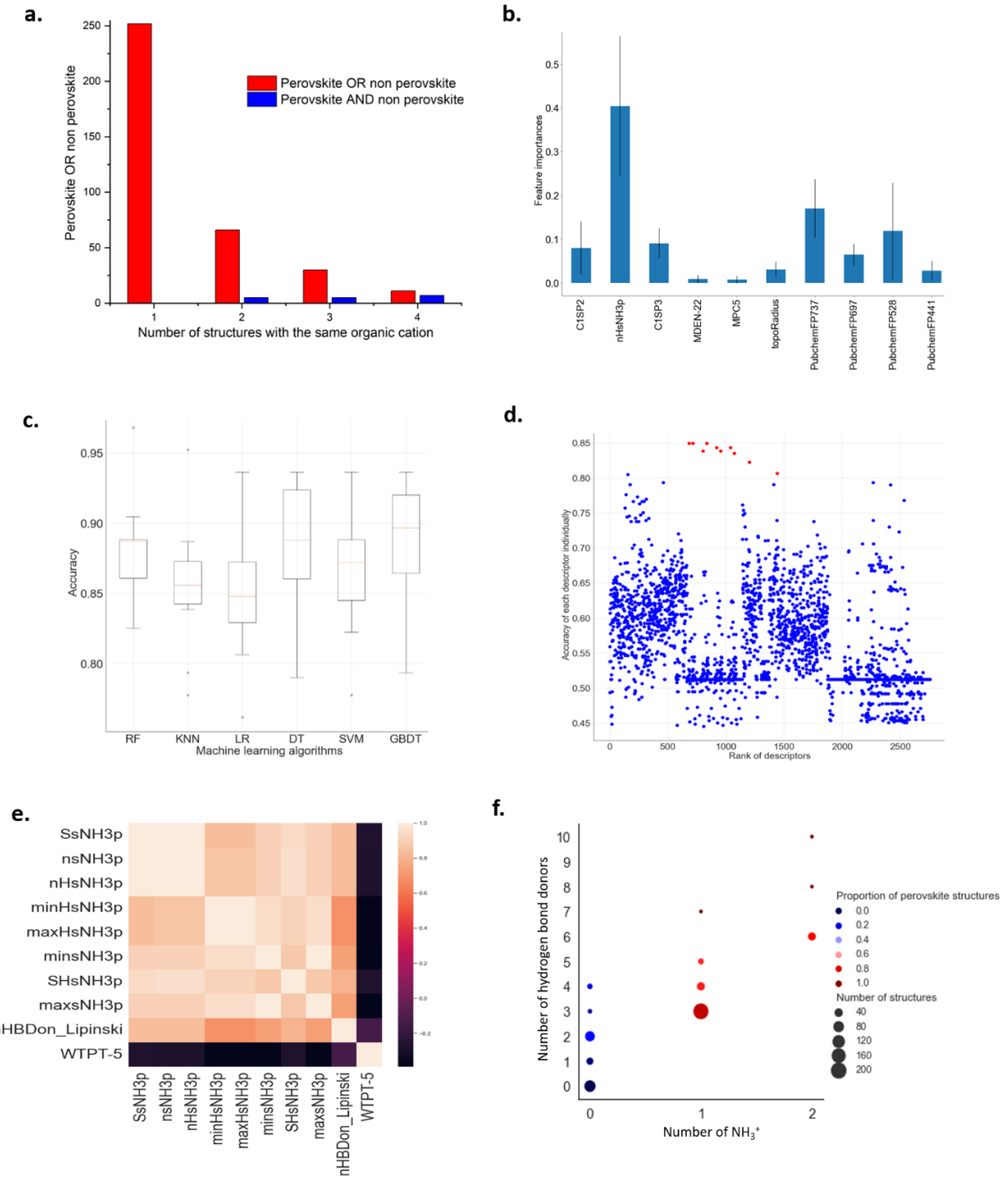

Figure 1. Statistical analysis and identification of key descriptors by machine learning. **a.** Repartition of the organic molecules according to the structure type (perovskite vs. non-perovskite) of the corresponding hybrid lead halide compound. The red bars are associated to organic molecules exclusively leading to either a perovskite or a non-perovskite compound whereas the blue bars are associated to organic molecules leading to both perovskite and non-perovskite compounds. **b.** Feature importances of the ten selected descriptors (blue bars are the feature importances of the forest, along with their inter-trees variability represented by the error bars). **c.** Comparison of the performances of different ML algorithms represented by box plots for the prediction of the structure type of hybrid lead halides using the ten selected descriptors and fingerprints (10 rounds of cross-validation). **d.** Classification accuracy using decision tree algorithm with each individual descriptors and fingerprints (the ones represented by dots in red exhibit the best accuracies). **e.** Correlation matrix of the top ten descriptors (SsNH3p, nsNH3p, nHsNH3p, minHsNH3p, maxHsNH3p, minsNH3p, SHsNH3p, MaxsNH3P, nHBDon_Lipinski, WTPT-5,). **f.** Statistical analysis of the structure type in terms of the number of hydrogen bonding donors and the number of -NH_3^+.

2.3. Analysis of the top descriptors

To better understand how the top ten descriptors could predict the perovskite structures regardless of their dimensionalities, one-level decision trees were plotted (Figure 2). Figure 2a shows that the presence of -NH_3^+ groups in the organic cation leads to the perovskite structure type: of the 321 perovskite structures, 291 are built from an organic cation with at least one primary ammonium group. In the absence of -NH_3^+ group, non perovskite structures are more likely obtained. Thus, among the 305 non perovskite structures, 241 are built of organic molecules with no primary ammonium group. These conclusions are in agreement with the results obtained from the decision tree using the nHBDon_Lipinski descriptor (Figure 2b). Decision trees of the other eight top descriptors are provided in the supplementary information (See Figure S1).

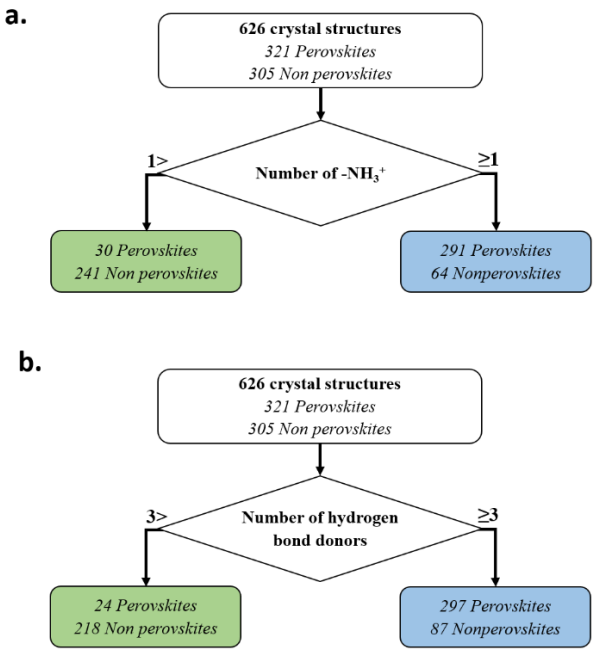

Figure 2. One-level decision tree using the individual descriptor: **a.** nHsNH3p (number of -NH_3^+) and **b.** nHBDon_Lipinski (number of hydrogen bond donors).

The coordination of $[PbX_6]^{2-}$ anions to the extended bond network can be analyzed in the light of the ML results. The nucleophilic X^- ligands of $[PbX_6]^{2-}$ must interact with the ammonium group of the organic molecules through hydrogen bonding. However, the coordination of these lead halide anions into a perovskite type structure creates a strong nucleophilic site between four $[PbX_6]^{2-}$ corner-sharing octahedra (Figure 3). The analysis of the crystal structures of different compounds shows that the -NH_3^+ groups are localized in these strong nucleophilic sites through the creation of a dense hydrogen bonding network with three or more halide ligands. As secondary and tertiary ammonium ions cannot easily create such

dense hydrogen bonding, they are not likely to lead to perovskite networks. The investigation of the outliers (the secondary or/and tertiary ammonium cations leading to perovskite compounds and the primary ammonium cations leading to non perovskite compounds) is also informative. Thus, the presence of at least one $-NH_2^+$ (e.g. $(C_6H_9N_3)PbBr_4$ [24]) or at least two $-NH^+$ (e.g. $(C_8H_8N_4)PbI_4$ [25]) on the organic cations play, in some cases, the same role as the presence of one $-NH_3^+$ as a dense hydrogen bonding network can be created with the halide ions of the four $[PbX_6]^{2-}$ corner-sharing octahedra of a perovskite structure. This observation is also confirmed by statistical analysis (Figure 1f) as, in the absence of primary amine group, one can observe that it is possible to obtain perovskites structures if the number of hydrogen bond donors is sufficiently high (i.e. ≥ 2). In the cases where primary ammonium cations do not lead to perovskite type structures (30 cases reported in the CCDC), primary ammonium cations very likely lead to 1D networks of face sharing metal halide octahedra (e.g. $C_6H_5CH(CH_3)NH_3PbX_3$[26]). In such cases, the $-NH_3^+$ cations are localized between two adjacent chains to create a dense network of hydrogen bonding with halide ions (Figure S4). It is also interesting to note that the nucleophilicities of the halide ions do not modify this trend as the ML analysis shows that the descriptor "Nature of the halide" (i.e. Cl, Br or I) does not play a key role in the determination of the structure type of the metal halide compounds (i.e. perovskite or non-perovskite).

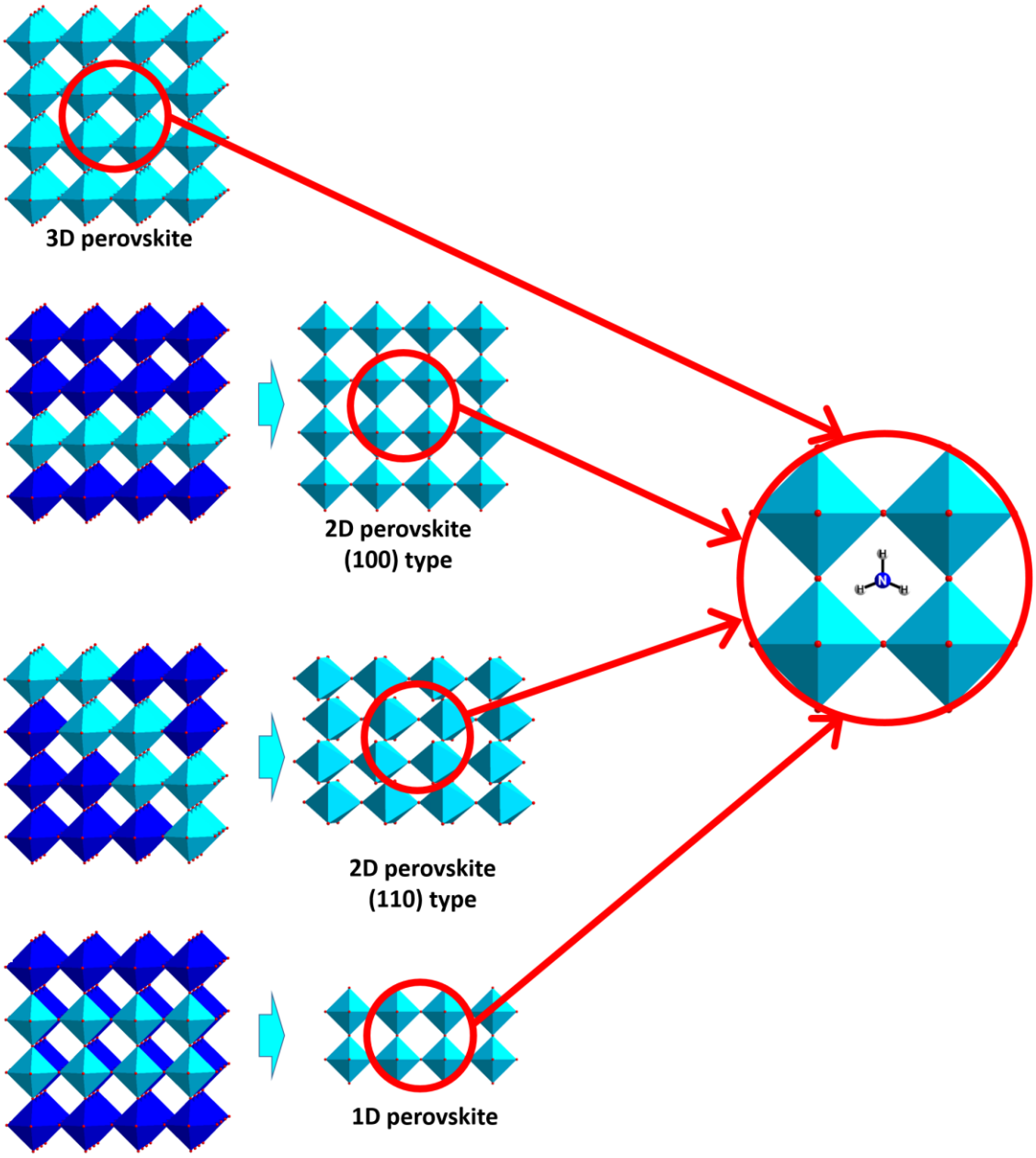

Figure 3. Representation of the nucleophilic site for -NH_3^+ primary ammonium cation between four $PbBr_6$ octahedra of the common 3D, 2D (100), 2D (110) and 1D perovskite frameworks. The position of –NH_3^+ in this nucleophilic site enables the stabilization of perovskite type structures regardless of the dimensionalities of the inorganic lead halide network.

3. Conclusion

Machine learning can play a key role in extracting useful insights and understanding physicochemical phenomena that can accelerate and automate the discovery of new materials. In this work, machine learning algorithms clearly identified the effect of organic cations on the perovskite structure type of hybrid lead halides. The number of -NH_3^+ groups was

identified as a descriptor highly correlated with the architecture of the inorganic network. The presence of primary ammonium cation creates a dense hydrogen bonding network around the nucleophilic halide ions favoring the crystallization of hybrid lead halide perovskites. Other descriptors leading to high accuracies of the classification perovskite vs. non perovskite are also related to the hydrogen bonding and the environment around the nitrogen atoms of the organic molecules, confirming the key role of primary amines on the design of perovskite type structures.

In the future, we believe that such ML algorithms will become very important tools to identify how the nature of the building units are linked to the structural characteristics of the resulting materials. In addition of providing key insights to design new materials with specific functionalities, this approach can also statistically confirm or reject previous hypotheses which are very often based on human intuition after analyses of small amounts of data.

4. Methods

4.1. Data collection and preparation.

The data used in this study were extracted from the Cambridge Structural Database (CSD) as of December 2021 (Figure S2). Despite the CSD being a rich and reliable database, the collected data required a process of cleaning and preprocessing (see codes *calculating SMILES and similarity test* and *descriptors and fingerprint calculation using padelpy* in SI). From this database, 1472 CIFs of hybrid lead halides were collected, and the organic cations were extracted from each CIF as a graph. This graph was converted to a Simplified Molecular Input Line Entry Specification (SMILES) using the CSD API. Structures containing more than one organic cation were omitted, as it is impossible to determine reliably which cation is responsible of a specific inorganic network. After this filtration, 1073 structures remained. To avoid overfitting, a similarity test algorithm was used to exclude similar or quasi-similar structures from the dataset. This algorithm compared lattice parameters using predetermined similarity factors. If two structures had the same lattice parameters within the given tolerance, they were considered to be the same under different names and one was excluded from the dataset. After applying this filter, 817 structures remained. Finally, the following steps enabled to build the input matrix containing descriptors values and the output vector indicating if the structure is of perovskite or non-perovskite type: (i) From each SMILES, 1875 descriptors and 881 fingerprints were calculated using Padelpy API.[27] (ii) The structures for which it was not possible to calculate all the descriptors and fingerprints were

removed (e.g. for some inorganic cations such as Cs^+). Our dataset was reduced to 626 structures. (iii) Each of these 626 structures was automatically labeled as perovskite vs. non perovskite by using a class of functions from Pymatgen API.[28] These functions are based on testing if the PbX_6 octahedra share exclusively corners (perovskite) or not (non perovskite).[29] After verification, 33 structures were incorrectly labeled and had to be corrected. (iv) Finally, we obtained our input matrix consisting of 626 rows (structures) and 2756 columns (descriptors + fingerprints). The output vector containing the structure type for each structure was also constructed.

Supporting Information

This material is available free of charge via the Internet from the Wiley Online Library. The material includes the *Supporting Information* file (Figure S1-S4, Table S1), *List of descriptors, List of fingerprints, datasets and results* (*database, Accuracy_of_each_descriptor, best_descriptors, list_of_SMILES, structures_names*), and *python codes* (*Calculating SMILES and similarity test* code - calculates the smiles for each structure and test the similarity between structures in order to keep only different structures, *Descriptors and fingerprint calculation code using Padelpy, Statistics on perovskite and non-perovskite database, Feature selection by deleting descriptors with null values followed by sequential feature selection, Features importances, Hyperparameters optimization using optuna, Comparison of descriptor-based machine learning models for perovskite and non-perovskite materials, Classification accuracy for the prediction of perovskite type using individual descriptors and fingerprints, Decision trees to analyze the performance of each of the best descriptors, Correlation of the best ten descriptors, Proportion of perovskite structures based on the number of hydrogen bonds and NH3 molecules*). A detailed description of how the datasets and python codes are related to each others is represented in Figure S3 of the *Supporting Information* file.

Funding Sources

This work benefited from the French government aid managed by the Agence Nationale de la Recherche under the France 2030 programme with the reference ANR 22 PEXD 0008 and ANR-21-ERCC-0009-01. Calculations were partly conducted at Centre de Calcul Intensif des Pays de la Loire (CCIPL), Université de Nantes.

References

[1] M. A. Green, A. Ho-Baillie, H. J. Snaith, *Nature Photonics* **2014**, *8*, 506.
[2] S. D. Stranks, H. J. Snaith, *Nature Nanotech* **2015**, *10*, 391.
[3] B. Saparov, D. B. Mitzi, *Chem. Rev.* **2016**, *116*, 4558.
[4] N.-G. Park, *J. Phys. Chem. Lett.* **2013**, *4*, 2423.
[5] H. J. Snaith, *J. Phys. Chem. Lett.* **2013**, *4*, 3623.
[6] J.-P. Correa-Baena, M. Saliba, T. Buonassisi, M. Grätzel, A. Abate, W. Tress, A. Hagfeldt, *Science* **2017**, *358*, 739.
[7] J. Tong, Z. Song, D. H. Kim, X. Chen, C. Chen, A. F. Palmstrom, P. F. Ndione, M. O. Reese, S. P. Dunfield, O. G. Reid, J. Liu, F. Zhang, S. P. Harvey, Z. Li, S. T. Christensen, G. Teeter, D. Zhao, M. M. Al-Jassim, M. F. A. M. van Hest, M. C. Beard, S. E. Shaheen, J. J. Berry, Y. Yan, K. Zhu, *Science* **2019**, *364*, 475.
[8] D. Bi, C. Yi, J. Luo, J.-D. Décoppet, F. Zhang, S. M. Zakeeruddin, X. Li, A. Hagfeldt, M. Grätzel, *Nat Energy* **2016**, *1*, 1.
[9] Z.-X. Zhang, C.-Y. Su, J.-X. Gao, T. Zhang, D.-W. Fu, *Sci. China Mater.* **2021**, *64*, 706.
[10] S. A. Veldhuis, P. P. Boix, N. Yantara, M. Li, T. C. Sum, N. Mathews, S. G. Mhaisalkar, *Advanced Materials* **2016**, *28*, 6804.
[11] H.-C. Wang, Z. Bao, H.-Y. Tsai, A.-C. Tang, R.-S. Liu, *Small* **2018**, *14*, 1702433.
[12] Z.-X. Zhang, H.-Y. Zhang, W. Zhang, X.-G. Chen, H. Wang, R.-G. Xiong, *J. Am. Chem. Soc.* **2020**, *142*, 17787.
[13] D. Hao, J. Zou, J. Huang, *InfoMat* **2020**, *2*, 139.
[14] H.-P. Wang, S. Li, X. Liu, Z. Shi, X. Fang, J.-H. He, *Advanced Materials* **2021**, *33*, 2003309.
[15] C. J. Bartel, C. Sutton, B. R. Goldsmith, R. Ouyang, C. B. Musgrave, L. M. Ghiringhelli, M. Scheffler, *Science Advances* **2019**, *5*, eaav0693.
[16] F. Zhang, H. Lu, J. Tong, J. J. Berry, M. C. Beard, K. Zhu, *Energy Environ. Sci.* **2020**, *13*, 1154.
[17] D. G. Billing, A. Lemmerer, *CrystEngComm* **2007**, *9*, 236.
[18] R. Lyu, C. E. Moore, T. Liu, Y. Yu, Y. Wu, *J. Am. Chem. Soc.* **2021**, *143*, 12766.
[19] O. J. Weber, K. L. Marshall, L. M. Dyson, M. T. Weller, *Acta Cryst B* **2015**, *71*, 668.
[20] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Advanced Drug Delivery Reviews* **2012**, *64*, 4.
[21] Q. Ai, D. M. Williams, M. Danielson, L. G. Spooner, J. A. Engler, Z. Ding, M. Zeller, A. J. Norquist, J. Schrier, *J. Chem. Phys.* **2021**, *154*, 184708.
[22] 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, *J. Machine Learning Res.* 2011, 12, 2825.
[23] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Association For Computing Machinery, New York, NY, USA, **2019**, pp. 2623–2631.
[24] Y. Li, G. Zheng, C. Lin, J. Lin, *Crystal Growth & Design* **2008**, *8*, 1990.
[25] I. Zimmermann, S. Aghazada, M. K. Nazeeruddin, *Angewandte Chemie International Edition* **2019**, *58*, 1072.
[26] D. G. Billing, A. Lemmerer, *CrystEngComm* **2006**, *8*, 686.
[28] PaDELPy: A Python Wrapper for PaDEL-Descriptor Software, 2022. https://github.com/ecrl/padelpy (accessed 2022-09-16).
[28] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, *APL Materials* **2013**, *1*, 011002.
[29] F. Massuyeau, T. Broux, F. Coulet, A. Demessence, A. Mesbah, R. Gautier, *Advanced Materials* **2022**, *34*, 2203879.

Table of Contents