

ARTICLE OPEN



Machine-learning structural and electronic properties of metal halide perovskites using a hierarchical convolutional neural network

Wissam A. Saidi¹✉, Waseem Shadid² and Ivano E. Castelli³

The development of statistical tools based on machine learning (ML) and deep networks is actively sought for materials design problems. While structure-property relationships can be accurately determined using quantum mechanical methods, these first-principles calculations are computationally demanding, limiting their use in screening a large set of candidate structures. Herein, we use convolutional neural networks to develop a predictive model for the electronic properties of metal halide perovskites (MHPs) that have a billions-range materials design space. We show that a well-designed hierarchical ML approach has a higher fidelity in predicting properties of the MHPs compared to straight-forward methods. In this architecture, each neural network element has a designated role in the estimation process from predicting complex features of the perovskites such as lattice constant and octahedral tilt angle to narrowing down possible ranges for the values of interest. Using the hierarchical ML scheme, the obtained root-mean-square errors for the lattice constants, octahedral angle and bandgap for the MHPs are 0.01 Å, 5°, and 0.02 eV, respectively. Our study underscores the importance of a careful network design and a hierarchical approach to alleviate issues associated with imbalanced dataset distributions, which is invariably common in materials datasets.

npj Computational Materials (2020)6:36; <https://doi.org/10.1038/s41524-020-0307-8>

INTRODUCTION

Realization of commercially viable solar electricity is critical for securing long-term economic growth and mitigating the impact of climate change. The metal halide perovskites (MHPs) and in particular MAPbI₃ (MA = CH₃NH₃) are currently the most investigated solar materials with ~25.2% power conversion efficiency (PCE) surpassing commercialized solar cells such as multicrystalline Si (c-Si, 21.3%), cadmium telluride (CdTe, 22.1%) and copper indium gallium selenide (CIGS, 22.3%)¹. However, in comparison to traditional solar absorbers, the main advantage of MHPs is their facile large-scale synthesis with relatively low cost². Several studies showed that the high PCE in MHP based solar cells results from a unique combination of features: large absorption coefficient > 3.0 × 10⁴ cm⁻¹ in the visible light region³, low exciton binding energy resulting in high quantum yield of free electrons and holes⁴, long electron-hole diffusion lengths^{5,6}, and electronically benign point^{7,8} and grain-boundary defects⁹. Another application of MHPs utilizes tandem solar cells, coupling a wide bandgap “top cell” with a narrow bandgap material like silicon as a “bottom cell”. Given that crystalline Si has a 1.1 eV bandgap, this requires materials with a 1.75 eV bandgap in order to current-match both junctions¹⁰. Current research focuses on finding optimum MHPs materials for either a single absorber or tandem solar cells that are cost-effective, stable and lead free¹¹. As an example, Li and Yang quite recently screened over 4500 hybrid compounds including those beyond the perovskite crystal structure, identifying 13 and 23 materials for solar energy conversion and light emitting diodes, respectively¹². Despite these and other efforts, the chemical space of materials design is too broad for “brute force” screening, and more efficient search methods are needed to find perovskites in different bandgap ranges.

There are vast opportunities for the discovery of new photovoltaic materials, but with considerable synthesis challenges, e.g., new materials can disproportionate into competing phases, and uniform films of the correct phase and composition for photovoltaics may prove difficult to form. “Computational experiments” or inverse-design approaches using predictive first-principles calculations offer a remedy, either narrowing down the number of potential solar absorbers or in guiding the experimental synthesis by determining the region of stability, if any, of the new compound. Further, previous studies show that useful compounds for different applications comprise only a small fraction of the total number of possible compounds^{13–16}. For example, Castelli et al. examined the performance of 19,000 oxide and oxynitride perovskite materials as solar water light splitters, finding only 20 interesting compounds for experimental follow-up^{14,17}. Also, DFT-based screening of 700 ternary alkali-transition metal borohydrides identified ~20 compounds with potential for reversible hydrogen storage¹⁵.

While ab initio investigations only require input of atomic numbers and ion positions to provide accurate materials properties, their relatively high computational cost limits their application on large datasets of millions of possible combinations. This is a serious limitation given that the material design space of hybrid MHPs is enormous. In contrast to the ~100 chemical elements available for all inorganic compounds, there are astronomical combinations yielding valid molecular species for the organic systems. The GDB-17 database¹⁸ contains 166.4 billion molecules composed of 17 atoms including H, C, N, O, S, and the halogens, and there are ~10 million in the FDB-17 database, which evenly covers a broad range of molecular sizes, polarities, and stereochemistries¹⁹. Even limiting the chemical space to ~10,000

¹Department of Mechanical Engineering and Materials Science, University of Pittsburgh, Pittsburgh, PA 15261, USA. ²Department of System and Information Systems, The University of North Carolina Charlotte, Charlotte, NC, USA. ³Department of Energy Conversion and Storage, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark.

✉email: alsaidi@pitt.edu

experimentally measured molecules^{20,21}, the number of potential candidates is still huge, precluding traditional screening methods.

Machine learning (ML) approaches can accelerate the materials search by orders of magnitude compared to traditional first-principles methods without significantly compromising their accuracy^{22–25}, contributing to a paradigm shift in materials design where big data, artificial intelligence and materials modeling are deeply entangled²⁶. For example, Pilania and collaborators used kernel-ridge regression to predict the bandgaps of double perovskites achieving a root-mean-square error (RMSE) of 0.37 eV²². Im and co-workers used gradient-boosted regression trees to identify lead-free double perovskites and achieve an RMSE of 0.3 eV²⁵. Weston and Stampfl compared different ML models in the prediction of bandgaps of kesterite I2-II-IV-V4 quaternary compounds, reporting an RMSE as low as 0.28 eV²⁴.

Herein, we systematically develop a dataset including structural and bandgap properties of 862 MHPs with the aim of developing a predictive ML model that captures the complex trends and correlations of this chemical space. We construct our materials design space by focusing mainly on variations in the organic molecule that has an enormous number of possibilities compared to those afforded by inorganic systems. For example, in comparison to inorganic compounds^{22–25}, organic compounds display a wider range of features like electrostatic moments, anisotropies, and a complex potential energy surface allowing the molecule to distort depending environment and other entities in the compound. Additionally, the atomic size of the constituent elements of the perovskites is particularly important since this relates to their stability according to the Goldschmidt tolerance factor²⁷. But simple features such as the Shannon radii for atomic size are not easily generalized to the molecular case^{22–25}. Furthermore, structural optimization of the MHPs is challenging due to the soft nature of their potential energy surface²⁸, which thus complicates building a consistent large database for ML training.

With the aim of addressing these challenges, we develop a ML approach for the MHPs that will (1) accurately estimate the bandgap using easy to calculate descriptors; (2) overcome the small size dataset problem with imbalanced distributions of target values, i.e., some range parts of a target value may have too many samples while other parts may have few or none; and (3) use simple ML approaches that are computationally not demanding, and are relatively simple to understand and control. We show that a hierarchical neural network architecture composed of a relatively small neural network elements can address all of these constraints. In this case, the ML elements are arranged such that each part plays a specific role in the prediction process. Each element is built using a convolutional neural network (CNN) framework and

independently trained for its designated role apart from the other elements. This simplifies the learning process for neural networks and avoids the need for more sophisticated network architectures with many hidden layers.

RESULTS

Database

We perform a high-throughput computational screening of ABX₃ MHPs, focusing mainly on variations at the “A-ion” site while keeping the lattice and the inorganic BX₃ network the same as in the parent MAPbI₃ structure. As a necessary constraint, we ensured selected A, B, and X species satisfy (a) the valence rule, requiring elements to have the correct oxidation state, namely A⁺¹, B⁺², and X⁻¹; and (b) the even-odd rule, requiring the material to have an even number of electrons in the unit cell to maintain a non-metallic behavior. Figure 1 summarizes the material design space. At the A site, we investigated Cs in addition to 18 different organic molecules: ammonium [NH₄]⁺, [AsH₄]⁺, [PH₄]⁺, [AsH₄]⁺, [PF₄]⁺, methylammonium [CH₃NH₃]⁺, [CH₃PH₃]⁺, [CH₃AsH₃]⁺, hydrazinium [(H₃N)(NH₂)]⁺, azetidinium [(CH₂)₃NH₂]⁺, formamidinium [NH₂(CH)NH₂]⁺, [NH₂(CH)PH₂]⁺, [NH₂(CH)AsH₂]⁺, imidazolium [C₃N₂H₅]⁺, dimethylammonium [(CH₃)₂NH₂]⁺, acetamidinium [NH₂-C-CH₃-NH₂]⁺, ethylammonium [(C₂H₅)NH₃]⁺, and hydroxylammonium [H₃NOH]⁺. Starting from A cations such as ammonium, methylammonium or dimethylammonium, other cations can be generated by replacing N with P, As, or Sb, or by replacing hydrogen in ammonium with halogen atoms such as [PF₄]⁺²⁸. In our database, we avoided using relatively large organic molecules such as tertiary methylammonium or other combinations such as [NF₄]⁺ and [PCl₄]⁺ since these were previously shown to destabilize the 3-dimensional perovskite network²⁸. At the B site, we considered the divalent metals Pb and Sn. At the halogen X site, we included 10 different combinations obtained from the three halogens (Cl, Br, I). Previous searches investigated a small portion of these compounds^{28–32}. In total, we have 380 different MHP compositions in the cubic phase, which was expanded to 862 to account for all permutations obtained by rearrangements of the tri-halide moiety.

The bandgap and structural parameters of all of the compounds in the dataset are computed using DFT. For each case, we start structural relaxation with the organic molecule chosen randomly at the center of the metal-halide octahedra. Previous studies show the orientation of the organic molecule has relatively small effect on electronic properties of the system and, on average, the bandgap changes less than 0.1 eV³³. MHPs are known to have a complex potential energy surface with many local minima^{28,34}. This adds a significant challenge to high-throughput

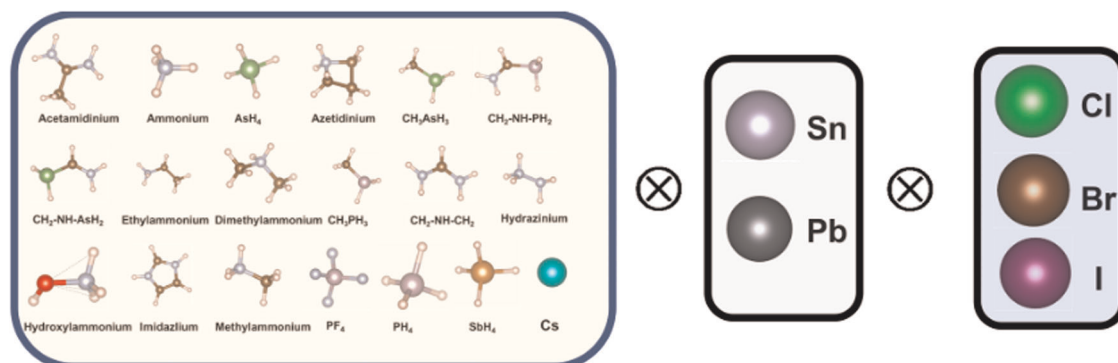


Fig. 1 Schematic representation of the selected material design space for the MHPs. The A site of ABX₃ can be occupied by Cs in addition to 18 different organic molecules: ammonium [NH₄]⁺, [AsH₄]⁺, [PH₄]⁺, [AsH₄]⁺, [PF₄]⁺, methylammonium [CH₃NH₃]⁺, [CH₃PH₃]⁺, [CH₃AsH₃]⁺, hydrazinium [(H₃N)(NH₂)]⁺, azetidinium [(CH₂)₃NH₂]⁺, formamidinium [NH₂(CH)NH₂]⁺, [NH₂(CH)PH₂]⁺, [NH₂(CH)AsH₂]⁺, imidazolium [C₃N₂H₅]⁺, dimethylammonium [(CH₃)₂NH₂]⁺, acetamidinium [NH₂-C-CH₃-NH₂]⁺, ethylammonium [(C₂H₅)NH₃]⁺, and hydroxylammonium [H₃NOH]⁺. The B site can be occupied by Sn or Pb, while as the X¹, X², and X³ halogen can be Cl, Br or I.

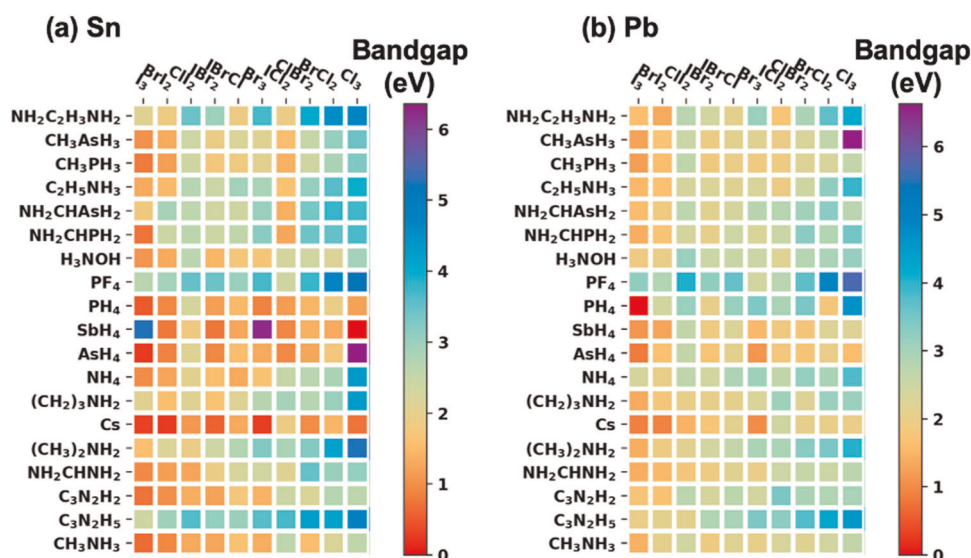


Fig. 2 Ground-truth bandgap values. Results obtained using GLLB-SC with SOC corrections for (a) Sn-based and (b) Pb-based perovskites. Each colored square corresponds to a different perovskite with the color indicating the bandgap value. The A-ion and the $X^1X^2X^3$ combination is identified by the labels on the vertical and horizontal axis, respectively.

computational studies because the outcome structure, and hence its properties, tend to be sensitive to the starting configuration. In our study, we fully relaxed the atomic coordinates and obtained the lattice constant using the Murnaghan equation of state assuming a perfect cubic structure. We find this approach to be important to have a consistent database given the sensitivity of the bandgap to lattice distortions. We ignored electron-phonon coupling effects that can affect the bandgap as these are relatively small^{35–37}. Figure 2 shows the bandgaps of the system obtained using GLLBC + SOC. Differences between the direct and indirect bandgaps were <0.1 eV, so this is ignored in our analysis. While the significant (0.2–6 eV) modulations of the bandgaps in the database ensure that the obtained correlations cover a wide range of systems, this complicates the development of an accurate ML model as we further discuss below.

DISCUSSION

Features engineering

Based on their complexity, we classify our descriptors into three main categories: elemental, precursor-based, and ABX₃-based features. We identified 29 elemental features that are obtained from the atoms and molecules forming the perovskite compound. These include first and second ionization energies and electron affinities, the electric dipole of the organic molecule, and atomic/molecular sizes. Also, we included the Goldschmidt tolerance factor,

$$t = \frac{(R_A + R_X)}{\sqrt{2}(R_B + R_X)}$$

often used to assess the stability of the perovskite structure and the recently introduced tolerance factor³⁸,

$$\tau = \frac{R_X}{R_B} - n_A \left(n_A - \frac{Y}{\ln(Y)} \right) \text{ with } Y = \frac{R_A}{R_B} \text{ and } n_A = 1.$$

We also identified six-precursor-based features derived from optimum structures of AX and BX_2 including lattice volumes, formation energies and bandgaps. The motivation for including these precursor-derived features is that MAPbI_3 can be synthesized from MAI and PbI_3 . While similar synthesis routes may not exist for different ABX_3 compounds, these features capture some of the chemistry of ABX_3 and hence are deemed important to

include. The last set of features considered are the lattice constant a and the metal-halide-metal octahedral tilt angle $\mathcal{O}$ obtained from the optimum ABX_3 structure, included because these geometric parameters affect the overlap of atomic orbitals and thus the bandgap²⁸. Obviously, a predictive model to be used for screening different ABX_3 systems should not depend on features obtained from the optimized ABX_3 structures. However, as we will discuss later, we will implement the ABX_3 -based features in a hierarchical neural network architecture composed of a relatively small number of neural network elements, each computationally tractable and relying on quickly calculated features.

Using the full set of features, we performed a statistical analysis and empirical validation to determine the features that have the most impact on the predictivity of the ML model, as quantified from the RMSE. This process results in 11 features out of the 37 original set of features. We note that this reduction of feature space is mainly important to speed up the calculations and to obtain a smoother training of the neural network. The identified important features include the size A , B and halides, first ionization energy and electron affinity of the organic molecule, bandgap, formation energy, and the volume of the AX precursor, in addition to the lattice constant and the octahedral angle of ABX_3 . As seen, the selected features are associated mainly with the A cation (organic molecule) rather than the B cation, which is rationalized given that our database encompasses mainly variations at the A site (19 different cases) rather than the B site (2 cases). The importance of some of these features for the bandgap in the MHPs has been reported before^{10,28,29,39}. For example, Filip and collaborators found this to be the case for the lattice constant and the octahedral angle²⁸. Further, the bandgaps of MHPs are found to be strongly dependent on electronegativities of the constituent species and the lattice constant, where the bandgaps increase (decrease) with an increase of the electronegativities (lattice constant)^{10,29}. Also, the electronegativity of the organic molecule can influence the bandgap indirectly: increasing the size of A tends to increase the lattice constant, and hence decrease the bandgap based on a one-dimensional quantum well description. Such a model applies to perovskites with $A = \text{FA}$, MA and Cs^{29} , as well as APbI_3 ($A = \text{NH}_4$, PH_4 , AsH_4 , and SbH_4)³⁹.

Our ML approach is based on a CNN since it can be equally applied for classification and regression models. Further, we found that convolutional layers capture important trends as multiple

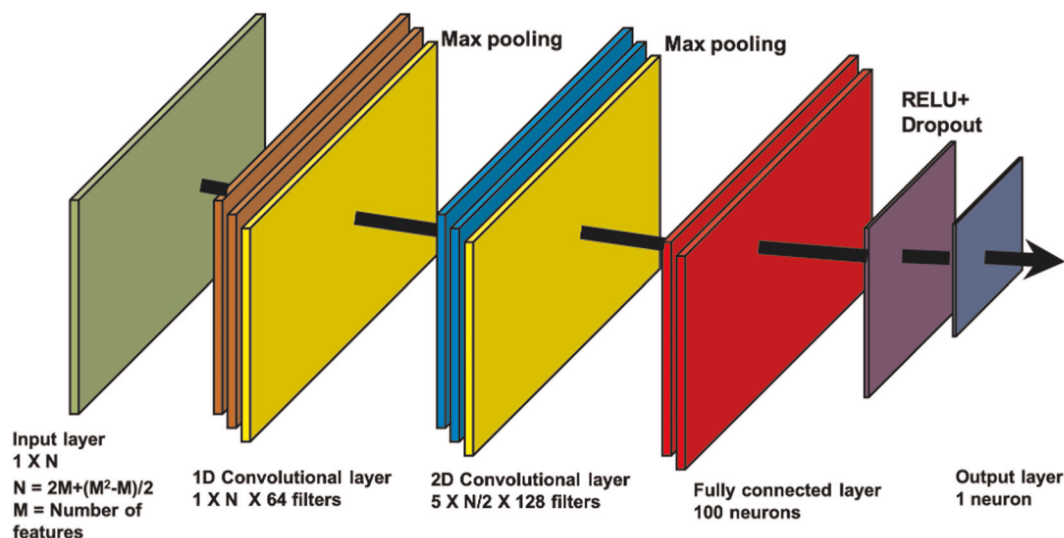


Fig. 3 Schematic representation of the convolutional neural network architecture. For the classifier network the element-wise rectified linear activation unit (relu) is replaced by an element-wise sigmoid unit.

perceptron layer was appreciably less accurate. Figure 3 shows pictorially the employed CNN architecture characterized by an input layer followed by two convolutional layers, one fully connected layer, and an output layer. This network design was implemented with TensorFlow Python API⁴⁰. The input layer is the set of predictors; The first convolutional layer consists of 64 filters, and employs a one dimensional (1D) convolutional kernel of size corresponding to the number of input features. The second convolutional layer consists of 128 filters, and employs a 2D convolutional kernel where the number of rows is equal to five and the number of columns is equal to the number of input features divided by two. The choice of 2D rather than 1D for the second layer is purposeful to accommodate the 2D output (filters and features) of the first layer. For both convolutional layers, the padding technique forces the output of each filter to have the same size as the input, followed by an element-wise rectified linear activation. Subsequent down sampling by a factor of two in a max pooling layer compresses the features between convolutional layers⁴¹. The fully connected layer consists of 100 hidden neurons connected to the down-sampled output of the second convolutional layer. This layer employs an element-wise rectified linear activation followed by a dropout layer with a drop rate of 0.2 to prevent overfitting. The last layer consists of one neuron with element-wise rectified linear activation that takes output of the fully connected layer as input and calculates the value of the bandgap. For the CNN for classification purposes, e.g., for estimating the range of the bandgap or the octahedral angle, the element-wise rectified linear activation is replaced by an element-wise sigmoid in all layers.

The CNN is trained using 80% of the data selected randomly from our dataset. For validation, we employed 12 different selections of the training set from the dataset, and the reported RMSE prediction estimation is an average over all of them. Prior to the training, we first standardized the input data by subtracting the mean of each feature from each value and divided by the standard deviation. The standardized features have a zero-mean and unit-variance, which ensures that all have equal weight without any distortions to the differences in the value ranges. Further, features standardization is important for optimizing the weights of the CNN during the training process given that inputs over different ranges may cause convergence problems manifesting as oscillatory behavior around the desired minimum. After standardization, we also included new features obtained from the

pair-wise interaction between the original descriptors to transform the linear regression model to a polynomial regression one. In this case, the pair-wise interactions between two different features capture the conjoint variation of any two independent predictors, and the self-interaction of features adds quadratic nonlinearity to the model. This feature augmentation step reduces the probability of missing important information that may improve the predictions of the CNN. Here, assuming that the original number of features is M , then after feature augmentation the total number of features will be increased by M due to the square terms and $M(M-1)/2$ from pair-wise interactions (see Fig. 3). After CNN training, we obtained an RMSE of 0.07 eV during bandgap prediction. Using all of the features also yielded similar RMSE for the bandgap, assuring that the dimensionality reduction does not affect the predictivity of the model.

One of the requirements to produce a reliable CNN is to have a balanced training dataset with homogenous distribution of samples in the output range. Imbalanced datasets may produce a model biased to learning parts of the range that have the most samples, and thus performs poorly on the other parts. As seen from the histogram in Fig. 4, the bandgap values vary between 0 and 6 eV with the majority of the values between 2 and 4 eV. Clearly, this imbalance in the representation of a physical property is a common problem in materials design, as there is no practical way to choose compounds that equally cover all the different ranges. To address the uneven nature of the dataset, we applied a hierarchical CNN (HCNN) that implements a classification algorithm to determine the bandgap range followed by another regressor CNN that estimates the bandgap value based on the outcome of the first CNN. Figure 5 shows schematically the dataflow contrast between the direct and hierarchical ML approaches.

We trained the HCNN by splitting the training set into six different classes based on the bandgap value. The first level structure of the hierarchical CNN design consists of one CNN model that outputs a vector of six binary values corresponding to the six ranges of the bandgaps. For each compound, all of these values should be zero except the value associated with the category this combination belongs to, which is one, a method referred to as one-hot representation⁴². Further, we implemented soft cutoffs for the range classification so that compounds that within ~ 0.2 eV of the boundary are included in both classes. Impressively, the HCNN approach reduces the error obtained

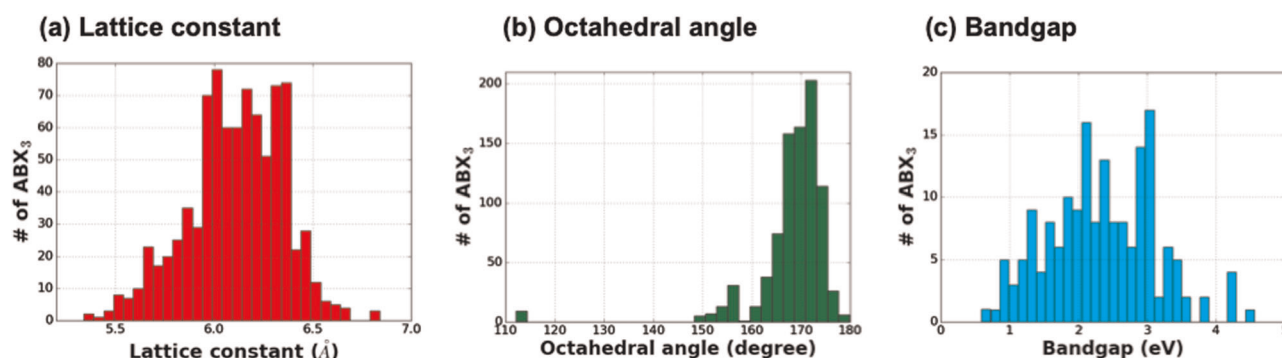


Fig. 4 Data distribution. Histogram analysis for (a) lattice constant, (b) octahedral angle, and (c) bandgap of the perovskites.

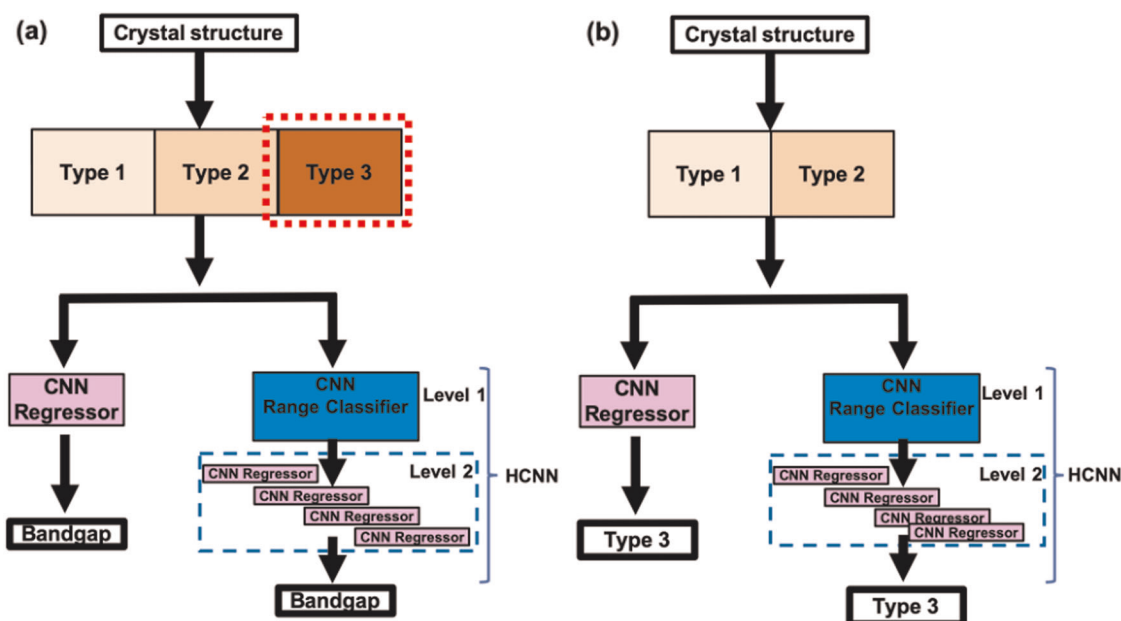


Fig. 5 HCNN architecture. Schematic representation of the convolutional neural network employed for (a) bandgap and (b) structural features of the perovskites. Type 1, 2, and 3 correspond respectively to elemental, precursor-derived and complex type features. In (a) and (b) we show two design schemes by either using a direct regressor network, or a combined hierarchical network that employs a classifier for the range followed by a regressor network for the output value. The dotted square in (a) indicates that Type 3 structural features can be obtained using the network design shown in (b).

before using standard CNN by a factor of three resulting in an RMSE of 0.02 eV or on average a 0.14 eV error in the bandgap.

While the HCNN has a high predictive accuracy, the feature list includes the lattice constant a and the octahedral angle ϕ of the hybrid ABX_3 that are not easily accessible without calculating the optimum structure of the ABX_3 using DFT, which we are trying to circumvent by using a CNN in the first place. Indeed, we find that both features are important to the high predictivity of the CNN model, as the RMSE increases from 0.07 to 0.16 eV when lattice constant and octahedral angle are removed from the features sets. It would be tempting to explore whether the a and ϕ can both be predicted using a CNN that employs only elemental and precursor-type features. We find that this is indeed the case, and both features can be predicted with a high accuracy, as can be seen in Fig. 5a, b. Using a CNN with an architecture similar to the one employed for the bandgap, the RMSE for the lattice constant and tilt angle are found to be 0.01 Å and 40°, respectively. The high RMSE for the octahedral angle is expected to be due to the imbalanced nature of the dataset, which is not the case for the lattice constants as seen from the histograms in Fig. 4b, c. To verify that this is the case, we employed an HCNN for

the octahedral angle as we did for the bandgap. The RMSE decreased in this case to 5°.

We employed the lattice constant CNN and the octahedral angle HCNN as two sub-nets in conjunction with an HCNN for the bandgap, all utilizing the same set of nine features identified before without the ABX_3 -type. Using this complex network architecture, we obtained an RMSE of 0.02 eV for the bandgap that is identical to the value obtained before the HCNN with the 11 feature sets including those of a and ϕ . Figure 6c compares the DFT-computed and CNN predicted values for the bandgap.

In summary, this study underscores the importance of careful design of the ML approaches to provide a predictive structure-property relationship for the MHPs. Particularly, we show that the lattice constant and the octahedral angle are key to obtain a predictive model for the bandgap, as the RMSE increases from 0.07 to 0.16 eV when these two features are removed from the dataset. Addressing this, we demonstrate that a CNN with easy-to-compute features can predict the lattice constant and octahedral angle with RMSE of 0.01 Å and 5°, respectively. Further, using the lattice constant and the tilt angle CNNs as two elements in the bandgap CNN would provide a predictive model with an RMSE of

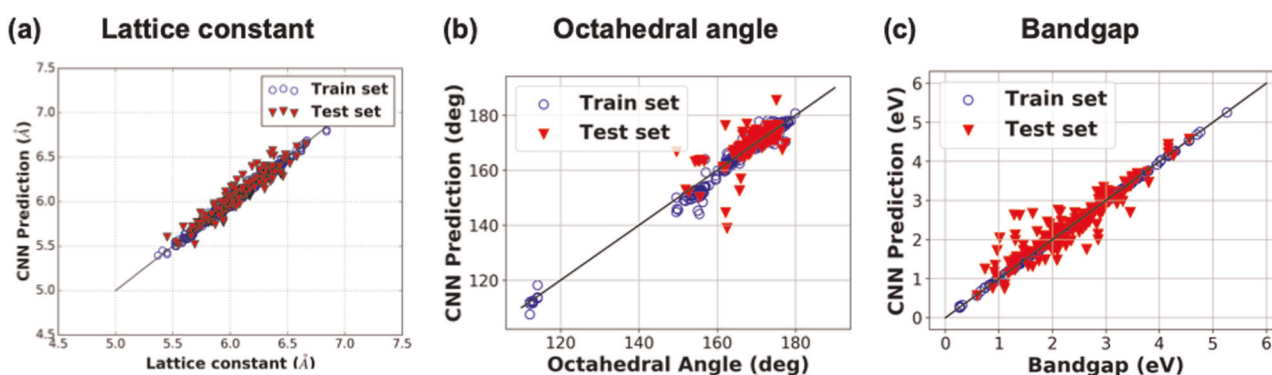


Fig. 6 HCNN performance. Representative plots for the prediction of the CNNs with 80% of the data for training and 20% for testing for (a) lattice constant, (b) octahedral angle and (c) bandgap.

0.02 eV, in agreement with the CNN that uses the DFT computed lattice constant and angle. Additionally, we show that a key to the success of this approach is in employing a hierarchical CNN to alleviate problems associated with the imbalanced nature of target values. Our findings are general and are likely to be applicable to other problems in materials design.

METHODS

The structural optimization of the MHPs is carried out using DFT calculations employing the Perdew-Burke-Ehrenzhof exchange-correlation functional revised for solids (PBEsol)⁴³ and projector augmented wave (PAW) pseudopotentials^{44,45} as implemented in the Vienna Ab initio Simulation Package (VASP) package. We expanded the electronic wavefunctions using plane-wave representation with a 400 eV cutoff. Sampling in the Brillouin zone, we used a $8 \times 8 \times 8$ shifted grids with 0.2 eV Gaussian smearing. All of the atomic coordinates are relaxed using a convergence threshold of 1 meV/Å on the atomic forces and 10^{-7} eV on the energies of the self-consistent electronic step. The band structure calculations are performed using GPAW⁴⁶ with spin-orbit coupling (SOC) corrections. To obtain accurate bandgaps at a reasonable computational cost, we employed the GLLB-SC functional⁴⁷ that can predict the bandgaps within an absolute accuracy of 0.5 eV⁴⁸. The GLLB-SC functional has been used in several studies to screen layered perovskites⁴⁹, perovskite oxides¹⁴, and also organic perovskites²⁹. For comparison, the GLLB-SC bandgaps of tetragonal MAPbI₃ and FAPbI₃ with SOC are 1.36 and 1.47 eV, comparable to the GW/SOC results^{29,50}. The SOC corrections decrease the bandgap on average by ~1.0 and 0.25 eV for Pb-based and Sn-based compounds. These results are consistent with previous studies^{29,50}. The reduction of the quasiparticle gap due to electron-hole interactions is relatively small ~0.1 eV and is hence ignored in this study²⁹.

For consistency, we calculated all features employed in the ML analysis rather than using values from literature. The electron affinity and ionization energy are obtained from energy differences of neutral and charged systems computed using Gaussian⁵¹ and the B3LYP functional. Such approach can be equally applied for the elements and the organic molecule. We have verified that the basis set is large and include diffuse basis functions to accurately calculate the electron affinity. For Pb, we used 0.365 eV for the electron-affinity as computed with relativistic effects⁵². To quantify the steric size, it is common to use Shannon radii of the elements. However, these do not exist for the organic cations. To be consistent for all elements/cations, we used as a size feature the radius of the sphere containing 90% of the electron density obtained from the B3LYP functional of the cation/anion that have the following nominal valence charges in ABX₃ perovskite structure: +1 for A cations, +2 for Pb and Sn, and -1 for the halides²⁸. Although the atom sizes of Pb/Sn and halide groups obtained from this method were different from the corresponding Shannon radii, it was comforting that both quantities showed similar trends. Further, it was reassuring that the less stringent constraint $R_A > R_B$ is also satisfied by all our systems using the new size feature. We also calculated the formation energies, lattice constant, and bandgap based on PBEsol for all AX and BX₂ compounds, under the assumption that these have the same crystal structure as in the parent MAI and PbI₂ systems.

DATA AVAILABILITY

A dataset of the 862 systems utilized in this study along with optimized atomic structures of the compounds are provided in the Supplementary Information.

Received: 8 December 2019; Accepted: 10 March 2020;

Published online: 14 April 2020

REFERENCES

1. National Renewable Energy Laboratory, Best Research-Cell Efficiencies Chart: <https://www.nrel.gov/pv/assets/pdfs/best-research-cell-efficiencies.20190802.pdf>.
2. Jena, A. K., Kulkarni, A. & Miyasaka, T. Halide perovskite photovoltaics: background, status, and future prospects. *Chem. Rev.* **119**, 3036–3103 (2019).
3. Green, M. A., Ho-Baillie, A. & Snaith, H. J. The emergence of perovskite solar cells. *Nat. Photon.* **8**, 506–514 (2014).
4. Miyata, A. et al. Direct measurement of the exciton binding energy and effective masses for charge carriers in organic-inorganic tri-halide perovskites. *Nat. Phys.* **11**, 582 (2015).
5. Xing, G. et al. Long-range balanced electron-and hole-transport lengths in organic-inorganic CH₃NH₃PbI₃. *Science* **342**, 344–347 (2013).
6. Stranks, S. D. et al. Electron-hole diffusion lengths exceeding 1 micrometer in an organometal trihalide perovskite absorber. *Science* **342**, 341–344 (2013).
7. Du, M. H. Efficient carrier transport in halide perovskites: theoretical perspectives. *J. Mater. Chem. A* **2**, 9091–9098 (2014).
8. Kim, J., Lee, S.-H., Lee, J. H. & Hong, K.-H. The role of intrinsic defects in methylammonium lead iodide perovskite. *J. Phys. Chem. Lett.* **5**, 1312–1317 (2014).
9. Shan, W. & Saidi, W. A. Segregation of native defects to the grain boundaries in methylammonium lead iodide perovskite. *J. Phys. Chem. Lett.* **8**, 5935–5942 (2017).
10. McMeekin, D. P. et al. A mixed-cation lead mixed-halide perovskite absorber for tandem solar cells. *Science* **351**, 151 (2016).
11. Giustino, F. & Snaith, H. J. Toward lead-free perovskite solar cells. *ACS Energy Lett.* **1**, 1233–1240 (2016).
12. Li, Y. & Yang, K. High-throughput computational design of organic-inorganic hybrid halide semiconductors beyond perovskites for optoelectronics. *Energy Environ. Sci.* **12**, 2233–2243 (2019).
13. Curtarolo, S. et al. The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
14. Castelli, I. E. et al. Computational screening of perovskite metal oxides for optimal solar light capture. *Energy Environ. Sci.* **5**, 5814–5819 (2012).
15. Hummelshøj, J. S. et al. Density functional theory based screening of ternary alkali-transition metal borohydrides: a computational material design project Bandgap calculations and trends of organometal halide perovskites. *J. Chem. Phys.* **131**, 014101 (2009).
16. Yu, L. & Zunger, A. Identification of potential photovoltaic absorbers based on first-principles spectroscopic screening of materials. *Phys. Rev. Lett.* **108**, 068701 (2012).
17. Castelli, I. E. et al. New cubic perovskites for one- and two-photon water splitting using the computational materials repository. *Energy Environ. Sci.* **5**, 9034–9043 (2012).
18. Reymond, J.-L. The chemical space project. *Acc. Chem. Res.* **48**, 722–730 (2015).
19. Visini, R., Awale, M. & Reymond, J.-L. Fragment Database FDB-17. *J. Chem. Inf. Modeling* **57**, 700–709 (2017).

20. Chase, M. W. et al. JANAF thermochemical tables, 1982 supplement. *J. Phys. Chem. Ref. Data* **11**, 695 (1982).
21. Rowley, R. L., W. W. V., Oscarson, J. L., Yang, Y. & Giles, N. F. *Data Compilation Tables of Properties of Pure Compounds* (American Institute for Chemical Engineering New York, 2012).
22. Pilania, G. et al. Machine learning bandgaps of double perovskites. *Sci. Rep.* **6**, 19375 (2016).
23. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *Npj Comput. Mater.* **2**, 16028 (2016).
24. Weston, L. & Stampfl, C. Machine learning the band gap properties of kesterite I2-II-IV-V4 quaternary compounds for photovoltaics applications. *Phys. Rev. Mater.* **2**, 085407 (2018).
25. Im, J. et al. Identifying Pb-free perovskites for solar cells by machine learning. *npj Comput. Mater.* **5**, 37 (2019).
26. Jose, R. & Ramakrishna, S. Materials 4.0: materials big data enabled materials discovery. *Appl. Mater. Today* **10**, 127–132 (2018).
27. Goldschmidt, V. M. Die Gesetze der Krystallochemie. *Naturwissenschaften* **14**, 477–485 (1926).
28. Filip, M. R., Eperon, G. E., Snaith, H. J. & Giustino, F. Steric engineering of metal-halide perovskites with tunable optical band gaps. *Nat. Commun.* **5**, <https://doi.org/10.1038/ncomms6757> (2014).
29. Castelli, I. E., Garcia-Lastra, J. M., Thygesen, K. S. & Jacobsen, K. W. Bandgap calculations and trends of organometal halide perovskites. *APL Mater.* **2**, 081514 (2014).
30. Filip, M. R. & Giustino, F. Computational screening of homovalent lead substitution in organic-inorganic halide perovskites. *J. Phys. Chem. C* **120**, 166–173 (2016).
31. Brandt, R. E., Stevanović, V., Ginley, D. S. & Buonassisi, T. Identifying defect-tolerant semiconductors with high minority-carrier lifetimes: beyond hybrid lead halide perovskites. *MRS Commun.* **5**, 265–275 (2015).
32. Yang, D. et al. Functionality-directed screening of Pb-free hybrid organic-inorganic perovskites with desired intrinsic photovoltaic functionalities. *Chem. Mater.* **29**, 524–538 (2017).
33. Castelli, I., Kristian Thygesen & Jacobsen, K. *Computational High-throughput Screening for Solar Energy Materials* (CRC Press, 2017).
34. Saidi, W. A. & Choi, J. J. Nature of the cubic to tetragonal phase transition in methylammonium lead iodide perovskite. *J. Chem. Phys.* **145**, 144702 (2016).
35. Foley, B. J. et al. Temperature dependent energy levels of methylammonium lead iodide perovskite. *Appl. Phys. Lett.* **106**, 243904 (2015).
36. Saidi, W. A., Poncé, S. & Monserat, B. Temperature dependence of the energy levels of methylammonium lead iodide perovskite from first-principles. *J. Phys. Chem. Lett.* **7**, 5247–5252 (2016).
37. Saidi, W. A. & Kachmar, A. Effects of electron-phonon coupling on electronic properties of methylammonium lead iodide perovskites. *J. Phys. Chem. Lett.* **7**, 7090–7097, <https://doi.org/10.1021/acs.jpclett.8b03164> (2018).
38. Bartel, C. J. et al. New tolerance factor to predict the stability of perovskite oxides and halides. *Sci. Adv.* **5**, eaav0693 (2019).
39. Filip, M. R., Verdi, C. & Giustino, F. GW band structures and carrier effective masses of $\text{CH}_3\text{NH}_3\text{PbI}_3$ and hypothetical perovskites of the type APbI_3 : $\text{A} = \text{NH}_4$, PH_4 , AsH_4 , and SbH_4 . *J. Phys. Chem. C* **119**, 25209–25219, (2015).
40. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, <http://tensorflow.org/> (2015).
41. Krohn, J. et al. *A Guide to Convolutional Neural Networks for Computer Vision*. (Morgan & Claypool Publishers, 2018).
42. Khan, S., Beyleveld, G. & Bassens, A. *Deep Learning Illustrated: A Visual, Interactive Guide to Artificial Intelligence*. (Pearson Education, 2019).
43. Perdew, J. P. et al. Restoring the density-gradient expansion for exchange in solids and surfaces. *Phys. Rev. Lett.* **100**, 136406 (2008).
44. Blochl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
45. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
46. Enkovaara, J. et al. Electronic structure calculations with GPAW: a real-space implementation of the projector augmented-wave method. *J. Phys.: Condens. Matter* **22**, 253202 (2010).
47. Kuisma, M., Ojanen, J., Enkovaara, J. & Rantala, T. T. Kohn-Sham potential with discontinuity for band gap materials. *Phys. Rev. B* **82**, 115106 (2010).
48. Castelli, I. E. et al. New light-harvesting materials using accurate and efficient bandgap calculations. *Adv. Energy Mater.* **5**, 1400915 (2015).
49. Ivano, E. C., Juan Maria, G.-L., Falco, H., Kristian, S. T. & Karsten, W. J. Stability and bandgaps of layered perovskites for one- and two-photon water splitting. *N. J. Phys.* **15**, 105026 (2013).
50. Umari, P., Mosconi, E. & De Angelis, F. Relativistic GW calculations on $\text{CH}_3\text{NH}_3\text{PbI}_3$ and $\text{CH}_3\text{NH}_3\text{SnI}_3$ perovskites for solar cell applications. *Sci. Rep.* **4**, 4467 (2014).
51. Gaussian 09 (Gaussian, Inc., Wallingford, CT, USA, 2009).
52. Tatewaki, H., Yamamoto, S., Moriyama, H. & Watanabe, Y. Electron affinity of lead: An ab initio four-component relativistic study. *Chem. Phys. Lett.* **470**, 158–161 (2009).

ACKNOWLEDGEMENTS

W.A.S. acknowledges the financial support from the National Science Foundation (Award No. DMR-1809085), useful discussions with Luca Ghiringhelli and Runhai Ouyang, and careful read by Richard Burke Garza. Calculations were performed in part at the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357, Extreme Science and Engineering Discovery Environment (XSEDE), and the University of Pittsburgh Center for Research Computing (CRC) through the computational resources provided. I.E.C. acknowledges support from the Department of Energy Conversion and Storage, Technical University of Denmark, through the Special Competence Initiative “Autonomous Materials Discovery (AiMade)”. [<http://www.aimade.org>].

AUTHOR CONTRIBUTIONS

W.A.S. designed the project. W.S. and W.A.S. carried out the machine learning design and calculations. W.A.S. and I.E.C. performed the high-throughput DFT calculations. W.A.S. wrote the manuscript. All authors discussed data and feature analysis, and reviewed the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at <https://doi.org/10.1038/s41524-020-0307-8>.

Correspondence and requests for materials should be addressed to W.A.S.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020