Machine Learning: Science and Technology

Data-driven enhancement of cubic phase stability in mixed-cation perovskites

IOP Publishing Mach. Learn.: Sci. Technol. 2 (2021) 025030 https://doi.org/10.1088/2632-2153/abdaf9

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
25 June 2020

REVISED
18 December 2020

ACCEPTED FOR PUBLICATION
12 January 2021

PUBLISHED
14 April 2021

PAPER

Data-driven enhancement of cubic phase stability in mixed-cation perovskites

Heesoo Park[1], Adnan Ali[1], Raghvendra Mall[2], Halima Bensmail[2], Stefano Sanvito[3] and Fedwa El-Mellouhi[1]

[1] Qatar Environment and Energy Research Institute, Hamad Bin Khalifa University, PO Box 34110 Doha, Qatar
[2] Qatar Computing Research Institute, Hamad Bin Khalifa University, PO Box 34110 Doha, Qatar
[3] School of Physics, AMBER and CRANN Institute, Trinity College, Dublin 2, Ireland

E-mail: hpark@hbku.edu.qa and felmellouhi@hbku.edu.qa

Keywords: deep learning, density functional theory, data-driven materials discovery, hybrid organic inorganic perovskites

Supplementary material for this article is available online

Abstract

Mixing cations has been a successful strategy in perovskite synthesis by solution-processing, delivering improvements in the thermodynamic stability as well as in the lattice parameter control. Unfortunately, the relation between a given cation mixture and the associated structural deformation is not well-established, a fact that hinders an adequate identification of the optimum chemical compositions. Such difficulty arises since local distortion and microscopic disorder influence structural stability and also determine phase segregation. Hence, the search for an optimum composition is currently based on experimental trial and error, a tedious and high-cost process. Here, we report on a machine-learning-reinforced cubic-phase-perovskite stability predictor that has been constructed over an extensive dataset of first-principles calculations. Such a predictor allows us to determine the cubic phase stability at a given cation mixture regardless of the various cations' pair and concentration, even assessing very dilute concentrations, a notoriously challenging task for first-principles calculations. In particular, we construct machine learning models, predicting multiple target quantities such as the enthalpy of mixing and various octahedral distortions. It is then the combination of these targets that guide the laboratory synthesis. Our theoretical analysis is also validated by the experimental synthesis and characterization of methylammonium–dimethylammonium-mixed perovskite thin films, demonstrating the ability of the stability predictor to drive the chemical design of this class of materials.

1. Introduction

Methylammonium lead iodide, $MAPbI_3$, is the most notable member among the hybrid organic-inorganic ABX_3 perovskites, where A and B are, respectively, an organic ($MA = CH_3NH_3$) and a metal cation (Pb), and X is a halide (I). Since $MAPbI_3$ was proposed as absorber in solar cells [1–3], these compounds have emerged as the attractive materials in the field of photovoltaics. Today, $MAPbI_3$ is not the only member of this class, which comprises compounds incorporating several organic cations such as ammonium (AM), hydrazinium (HZ), formamidinium (FA), and guanidinium (GUA). Most interestingly, there are also several examples of mixed-cation perovskites, where MA is mixed together with other organic cations at different concentrations, forming $MA_xA'_{1-x}PbI_3$ [$A' =$ AM, HZ, FA, or GUA] two-cation perovskites. The ability of synthesizing perovskites with organic-cation mixtures provides a powerful mean for tuning their electronic properties (for example, bandgap), hence it offers a new design tool for applications in photovoltaics [4–9] and light-emitting diodes [10–13].

Recently, the importance of maintaining the perovskite cubic phase was identified by Aydin *et al*, who demonstrated that the outdoor performance of perovskite/silicon tandem solar cells could be improved by a careful choice of the perovskite layer composition [14]. The stability of the cubic phase is often a problem for outdoor operation since lattice distortions may significantly alter the bandgap, but unfortunately may be

present under various temperature and illumination conditions. For this reason, it is crucial to optimize the chemical composition of the perovskites so to enhance the robustness of the cubic phase, while delivering an optimal bandgap.

When preparing a new mixed-cation perovskite, a rough guideline for its stability is provided by the Goldschmidt tolerance factor, τ. This descriptor writes $\tau = (r_A + r_X)/\sqrt{2}(r_B + r_X)$, where r_α ($\alpha = A, B$ and X) is the effective ionic radius of the α species. A compound is likely to crystallize in the cubic phase (space group $Pm\bar{3}m$) when τ lies in between 0.8 and 1.0. Beyond these ranges, for $\tau > 1$, a hexagonal phase is expected, while for $\tau < 0.8$, the crystal candidates can adopt either an orthorhombic structure or even a non-perovskite one. Hence, the tolerance factor represents a useful tool to understand the phase stability of hybrid organic/inorganic perovskites, and it has guided their synthesis for the past decade.

Despite this success with homogeneous compounds, the tolerance factor appears less useful often for engineering poly-elemental perovskites. In fact, τ is determined only by the ionic radii of the constituent elements, so that it does not describe any detail associated with local strain and information concerning the enthalpy of mixing [15, 16]. A possible step forward in our ability to predict the stability of mixed-cation perovskites is represented by machine-learning (ML) models. These, for instance, have enabled us to forecast the physical properties of novel stable perovskite compositions by interpolating chemical features of the constituent ions [17–23]. Progress has also been made in the stability prediction of compounds. For example, Bartel *et al* [23] proposed a modified tolerance factor, taking into account the oxidation states of the various ions, which was used to identify perovskites likely to be synthesized. In addition, more stringent descriptors for the stability of the perovskite cubic phase have also been suggested. These have enabled us to find the optimal A-site cation at a given BX_3 inorganic network [22, 24].

The experimental synthesis of hybrid perovskites often utilizes tiny concentrations (∼2%–5%) of an additive cation to achieve critical structural stability. Such optimal concentration is challenging to control in experiments and difficult to simulate for theory. At the current computational power, density functional theory (DFT) can calculate the formation energy of a cation-mixed compound almost at any desired concentration. However, for low concentrations, the computational task becomes costly since the low concentration translates into large supercells. Phase segregation makes the problem even more severe, requiring multiple runs for different conformations at the same concentration. One way to overcome this limitation is to use approximations such as *Cluster Expansion* to extrapolate trends at the macroscale [25, 26]. Unfortunately, the accuracy of cluster expansion approaches for predicting the trends of mixed-cation hybrid perovskites is still to be proven.

Here we aim at building a single universal machine learning (ML) model capable of predicting the properties of perovskites made of Group 14 elements (Ge, Sn, Pb), halides (I, Br, Cl), and a range of cations. This task is achieved by performing enthalpy-of-mixing DFT calculations for only eight concentrations for each pair of cations across all compositions. The resulting dataset enables us to derive trends and train a cubic-phase stability predictor capable of mapping the entire concentration range. This chemical diversity exploration is a task highly costly and practically unreachable by DFT alone. Since the structural stability and deformation is related to all the interactions between the elements at play, we have taken into account all the element's available chemical features. Then, the selection of the suitable ML algorithm has been obtained by carefully identifying the main descriptors [27], while focusing on predicting a range of target quantities, all associated with the crystal stability and structure.

In this work, by using a DFT dataset for a limited number of concentrations, we have constructed ML models that enable us to browse the full concentration range at no additional cost, offering an acceleration in the time for materials discovery of at least two orders of magnitude. Our computational strategy is then put to the test with the successful lab-scale synthesis and characterization of a representative mixed-cation perovskite, while the collective indicator of predictive variables guided the laboratory synthesis.

2. Methods

2.1. Construction of the dataset

The A-cation of an ABX_3 perovskite, which locates within the cavity of the BX_6 network, can be chosen from the alkali metals, Li^+, Na^+, K^+, Rb^+ and Cs^+, or among monovalent extended molecules. In this work, for instance, we have investigated some of the most popular organic cation choices, namely, hydronium (HY^+), ammonium (AM^+), sulfonium (SF^+), phosphonium (PH^+), hydroxylammonium (HA^+), methylammonium (MA^+), hydrazinium (HZ^+), formamidinium (FA^+), formamide (FO^+), ethylammonium (EA^+), dimethylammonium (DMA^+), and guanidinium (GUA^+). In general the A-cation is expected to induce a local deformation through the interaction with the ions of BX_6 octahedra. The deformations are specific to a particular cation, especially when it is non-spherical, and determine both the stability and the equilibrium structure at a given composition. As such, it is important that the dataset used

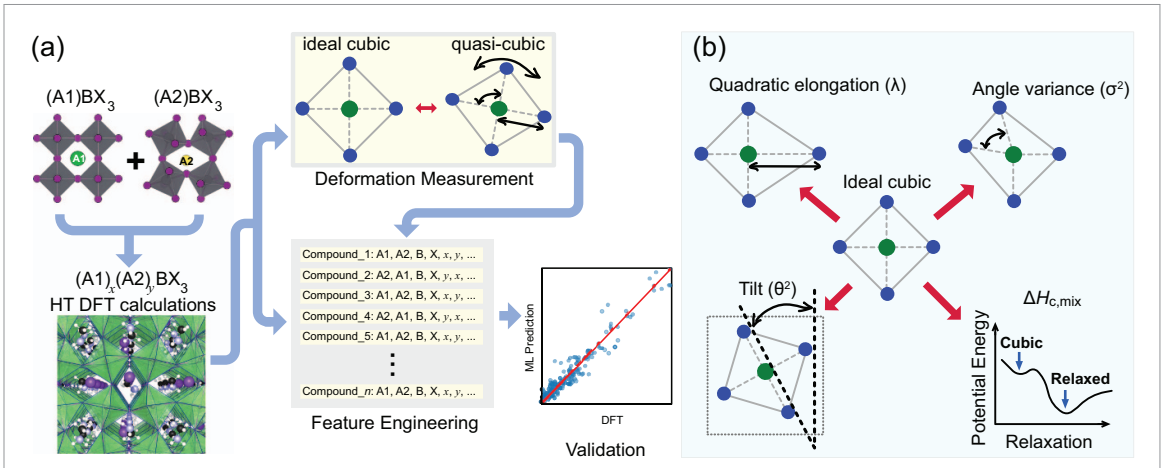

Figure 1. (a) The workflow of our high-throughput DFT calculations and the construction of the ML models. The equilibrium structure of single- and two-cation perovskites is firstly calculated by DFT and both structural and thermodynamics quantities are extracted. These represent the features of the ML models. (b) Schematic list of the target variables: octahedral quadratic elongation, λ, angle variance, σ^2, tilt, θ^2, and the distortion enthalpy of the mixture, $\Delta H_{c,mix}$.

to construct the ML models contains a sufficiently large number of structures presenting distortions. For this reason, we have designed a rigorous workflow for our DFT calculations (see figure 1(a)), which enables us to cover as much distortion space as possible as a function of the A/A'-cation content.

First, we prepare unit cells, where the organic cation is randomly oriented. These initial coordinates are fully relaxed and the various parameters describing the structure are extracted [21, 22]. Then we construct prototypes of two-cation perovskites, namely, $A_xA'_{1-x}BX_3$ where $0 \leqslant x \leqslant 1$, by using $2 \times 2 \times 2$ supercells containing 40 ionic sites (see table S1 in the supporting information (SI) for a list of such two-cation pairs) (available online at stacks.iop.org/MLST/2/025030/mmedia). These initial structures are also relaxed, and their structural information extracted, as described below.

Such a procedure provided us a total of 1864 single- and two-cation entries. However, we found that some compounds relaxed far from the perovskite crystal structure. Thus, we excluded these non-perovskite compounds from the train and test dataset. The 1780 remaining halide perovskites, which kept the 3-dimensional (3D) connectivity of BX_6 octahedra, were included in the ML train/test dataset. Among the 1780 perovskites, there were 504 mixed-cation perovskites and 1276 pure perovskite; we carried out the perovskites' optimization with six different initial orientation to mimic the cations random arrangements.

All calculations are performed at the generalized-gradient approximation level using the Perdew–Burke–Ernzerhof (PBE) [28, 29] functional, including Tkatchenko-Scheffler dispersive-interaction corrections [30]. A plane-waves energy cutoff of 520 eV is employed in the projector-augmented wave (PAW) [31, 32] method implemented in the VASP code [33–35], alongside with a $2 \times 2 \times 2$ k-points Monkhorst–Pack mesh. All compounds are relaxed until the energies and forces are converged within 10^{-7} eV/atom and 0.01 eV/Å, respectively. We quantify the octahedral deformations of the BX_6 octahedra by measuring the octahedral quadratic elongation, λ, the angle variance, σ^2, and the tilt, θ^2, of the relaxed structures, as the schematic list is shown in figure 1(b). One can find the equations as well as the additional information on how these quantities are extracted in the section of structural description and figure S1, in the SI.

In addition to the structural parameters describing the local lattice deformations, we have introduced a related thermodynamical quantity, named the *distortion enthalpy of the mixture*, $\Delta H_{c,mix}$. We define $\Delta H_{c,mix}$ as

$$\Delta H_{c,mix}(A_xA'_{1-x}BX_3) = x\Delta H_c(ABX_3) + (1-x)\Delta H_c(A'BX_3) + \Delta H_{mix}(A_xA'_{1-x}BX_3)\,, \quad (1)$$

where $\Delta H_c(ABX_3)$ and $\Delta H_c(A'BX_3)$ are the relaxation enthalpies of the single-cation perovskites ABX_3 and $AB'X_3$, respectively, where ΔH_c denotes the enthalpy difference between the cubic structure and the fully relaxed one [22]:

$$\Delta H_c = H_{cubic} - H_{relax}\,. \quad (2)$$

Finally, the enthalpy of mixing of ABX_3 and $A'BX_3$ in their equilibrium (distorted) structure, ΔH_{mix}, takes the usual definition:

$$\Delta H_{mix}\left(A_xA'_{1-x}BX_3\right) = \Delta H\left(A_xA'_{1-x}BX_3\right) - \left\{x\Delta H\left(A_xBX_3\right) + (1-x)\Delta H\left(A'_{1-x}BX_3\right)\right\}\,, \quad (3)$$

where $\Delta H(\alpha)$ is enthalpy of formation for the α compound. Thus, $\Delta H_{\rm c,mix}$ can be explicitly written as

$$\begin{aligned}&\Delta H_{\rm c,mix}(A_xA'_{1-x}BX_3) = \\ &= H_{\rm relax}(A_xA'_{1-x}BX_3) - \{xH_{\rm relax}(ABX_3) + (1-x)H_{\rm relax}(A'BX_3)\} + \\ &+ x\{H_{\rm cubic}(ABX_3) - H_{\rm relax}(ABX_3)\} + (1-x)\{H_{\rm cubic}(A'BX_3) - H_{\rm relax}(A'BX_3)\}\,,\end{aligned}$$

namely, it is the sum of the enthalpy of mixing of the two perovskites ABX_3 and $A'BX_3$, and the weighted average of the relaxation enthalpies of their single cation form. Clearly if both ABX_3 and $A'BX_3$ have a cubic structure, then $\Delta H_{\rm c,mix} = \Delta H_{\rm mix}$. We note that $\Delta H_{\rm c,mix}$ is minimized (becomes more negative) when the two-cation mixture has a lower enthalpy than the components and there is no enthalpy gain in relaxing the pure structure from the cubic ones. If one assumes that there is no elastic interaction between the octahedra of the two-cation perovskite, then $\Delta H_{\rm c,mix}$ will describe the tendency of forming a stable cubic mixture. We then expect that a small $\Delta H_{\rm c,mix}$ will characterize homogeneous two-cation cubic perovskites, while large values will be indicative of possible phase segregation, as usually observed for multi-cation compounds with high Cs concentrations [16, 36]. Throughout this work, we follow the rather standard procedure in high-throughput electronic structure theory to replace the enthalpy with the total DFT energy [37].

2.2. Machine learning

We built the ML models, now widely used in materials science [38–40], by using the *deeplearning* and *xgboost* modules implemented in the 'h2o' library for **R** [41]. 'h2o.deeplearning' is based on a multi-layer *feed-forward artificial neural network*, and it is denoted as DL (deep learning) within this article. DL is able to learn through non-linear mapping functions that take as input a set of features. These are then combined via a fully-connected network. In contrast, XGBoost (*h2o.xgboost*) is based on the tree-boosting method. The learning procedure is determined by the split points and the corresponding ensemble of the classification labels, while this algorithm creates the varying decision trees. A regularized model formalization of XGBoost controls the problem of over-fitting, helping to generate a better predictive model.

The choice of selecting DL and XGBoost is motivated by our previous experience and published work on a dataset of pure (single cation) perovskites. In that work, we compared the performances of various ML algorithms, such as a generalized linear model (GLM), random forest (RF), gradient-boosting machine (GBM), XGBoost, deep learning (DL; feed-forward neural network), in predicting the structural deformation parameters and the material stability [22]. In that case the data set consisted of ABQ_3 chalcogenide (I–V–VI_3) and ABX_3 halide (I–II–VII_3) perovskites, where charge neutrality was achieved by choosing the following elements: $B = V^{5+}$, Nb^{5+}, Ta^{5+}, Ge^{3+}, Sn^{3+} and Pb^{3+}; $Q = O^{2-}$, S^{2-}, Se^{2-}, Te^{2-}; and $X = F^-$, Cl^-, Br^- and I^-. We found that the linear regression method results in weak models compared to neuron- and tree-based methods due to the non-linear dependence of the chemical compositions' target properties.

In the present work, the model inputs consist of 24 chemical and structural features extracted from tabulated elemental properties, while the target quantities are λ, σ^2, θ^2, and $\Delta H_{\rm c,mix}$. Only these target variables are obtained from the DFT calculations for each composition. Some additional features describe the organic cations, namely their effective ionic radius and the number of lone pairs. Finally, in the case of two-cation perovskites, $A_xA'_{1-x}BX_3$, we also consider the relative A/A' concentration, the effective-ionic-radius difference between the A and A' cations, and the weight-averaged effective ionic radius, $r^*_{A,\rm avg} = xr_A + (1-x)r_{A'}$. In contrast, each elemental ion is represented by the period/group number, the ionization energy, the electron affinity of B and X, and the electronegativity difference between B and X. Also, we include as structural features the supercell size, the octahedral and the tolerance factor. As a consequence of our selection, all the input features are costless elemental features obtained from the chemical elements' tabulated properties. These properties have been assessed in terms of their correlation and importance for the machine learning training and the impact of any less important features have been verified (see figures S5 and S6 in the SI).

Clearly, the three features concerning only the two-cation perovskites are not defined for the single-cation ones. In order to maintain the same feature-vector size for both single- and two-cation structures, in the former case, we set the minority cation concentration to zero ($x = 1$), while assigning a randomly chosen cation to A'. In practice we represent the ABX_3 composition as $A_1A'_0BX_3$. Moreover, note also that the same structure can be represented in two equivalent ways. For instance, the mixture $Cs_{3/4}MA_{1/4}PbI_3$ is identical to the mixture $MA_{1/4}Cs_{3/4}PbI_3$. Thus, to reduce the bias on the specific definition for the cation occupation, we balance our data set in such a way to have an equal number of $A_xA'_{1-x}BX_3$ structures with $x > 0.5$ and $x < 0.5$.

We performed a careful hyper-parameter optimization to construct reliable models, and we systematically vary the hyper-parameters over a regular grid. A grid search on a wide range of hyper-parameters was used for DL and XGBoost training by combining several hyper-parameters. For the hyper-parameters of DL, we

varied the number of hidden layers to be either 2 or 3, and the number of neurons in each hidden layer was either 128 or 64. The used activation function is the Rectifier activation function with the learning rate in the set of {0.0001, 0.001, 0.01}. Similarly, for XGBoost, we have varied the number of trees from the set {100, 350, 1000}, the maximum tree depth within {3, 5, 7}, and the learning rate in the range {0.001, 0.01, 0.1, 0.5}.

In general, 85% of the calculated single-cation and two-cation perovskite structures are part of the training dataset, while the remaining 15% constitute the test dataset. In order to select the optimized hyper-parameters, we perform 5-fold cross-validation for each hyper-parameter combination until the model is optimized to have deviance of less than 10^{-10} or there is no improvement in performance for three consecutive epochs.

Before setting the training set size to 85%, we have inspected the ML performance so to evaluate the bias-variance trade-off problem. We split the dataset into two groups by varying the train/test split ratio and optimize the hyper-parameters at every inspection. We find that for the training set larger than 80% of the data, the mean absolute error (MAE) has to be considered converged, as the ML performance convergence is presented in figure S2 in the SI. (Note that the size of the test set was kept constant throughout the test). The optimal training/test split was found to be 85/15.

As a further test, we have also built a more standard regression model using the GLM method. GLM is found to severely underperform ($R^2 < 0.6$, MAE > 35 meV ion^{-1} for $\Delta H_{\mathrm{c,mix}}$) due to the non-linear relation between the input features and the target quantities. As one can see in figure S3 of the SI, DL and XGBoost always outperform GLM for all prepared training/test datasets in several training trials. Hence, as we move forward, we will only discuss the best-performing DL and XGBoost models and analyze their predictions.

Finally, in closing this section, we wish to revisit the general philosophy of this work. Our aim is to generate a single universal model capable of predicting structure and stability of perovskites made of Group 14 elements (Ge, Sn, Pb), halides (Cl, Br, I), and multiple cations. The variety of these building blocks makes it difficult to perform efficient training using a linear model. In particular, we aim at constructing a model capable of predicting unseen concentrations and compositions, which are unreachable by direct DFT calculations. These concentrations and compositions will be then validated by experiments. As a result, the strategy used here to build continuously improved ML predictors comprises three tightly interconnected phases, namely,

- *Phase 1*: Build a DFT database and train the ML models.
- *Phase 2*: Validate the ML predictions via lab-synthesis/characterization of perovskite films.
- *Phase 3*: Reinforce the DFT database and retrain the ML models by taking into account the feedback from *Phase 2*.

An early version of the ML models (*Phase 1*) has been already built where we used only pure perovskites' DFT results [22, 42], while in follow-up work, we have identified, via lab synthesis, where the models could be improved [24]. This current work reports the findings of multiple months of *Phase 3* work, enabling us to revise and improve our earlier model after the prediction-to-lab validation of *Phase 2*. Therefore, we have performed the DFT calculations for perovskites, where the cations are mixed explicitly at different concentrations. This extended dataset enables us to predict the enthalpy of mixing and addresses the issues identified in *Phase 2*. We systematically improve the models and provide systematic lab-scale validation and testing.

2.3. Preparation and deposition of perovskite films

To incorporate dimethylammonium (DMA) with methylammonium (MA) at the cation site of the perovskite structure, we prepare two different solutions inside the glove box, namely, DMAI (0.1 g, Greatcell)/PbI_2 (0.265 g, TCI)/DMSO (450 µl, Sigma Aldrich), leading to $DMAPbI_3$, and MAI (0.09 g, Greatcell)/PbI_2 (0.265 g, TCI)/DMSO (450 µl, Sigma Aldrich), leading to $MAPbI_3$. The solutions are then mixed at different concentrations so to yield $DMA_xMA_{1-x}PbI_3$, at the following stoichiometric fractions $x = 0$, 0.005, 0.01, 0.02, 0.05, 0.1 and 0.15. Then the glass substrates are inserted into the glove box for spin coating. Two pipettes are kept ready with 100 µl of the perovskite solution and 25 µl of chlorobenzene, used as an anti-solvent. The spin coater is programmed for a single run at 1000 and 4000 rpm for 10 and 30 s, respectively. Chlorobenzene is then dropped while the coater is running at a time comprised between 22 and 25 s during the second step. All the samples are then annealed at 100 °C on a hot plate inside the glove box for 1 h.

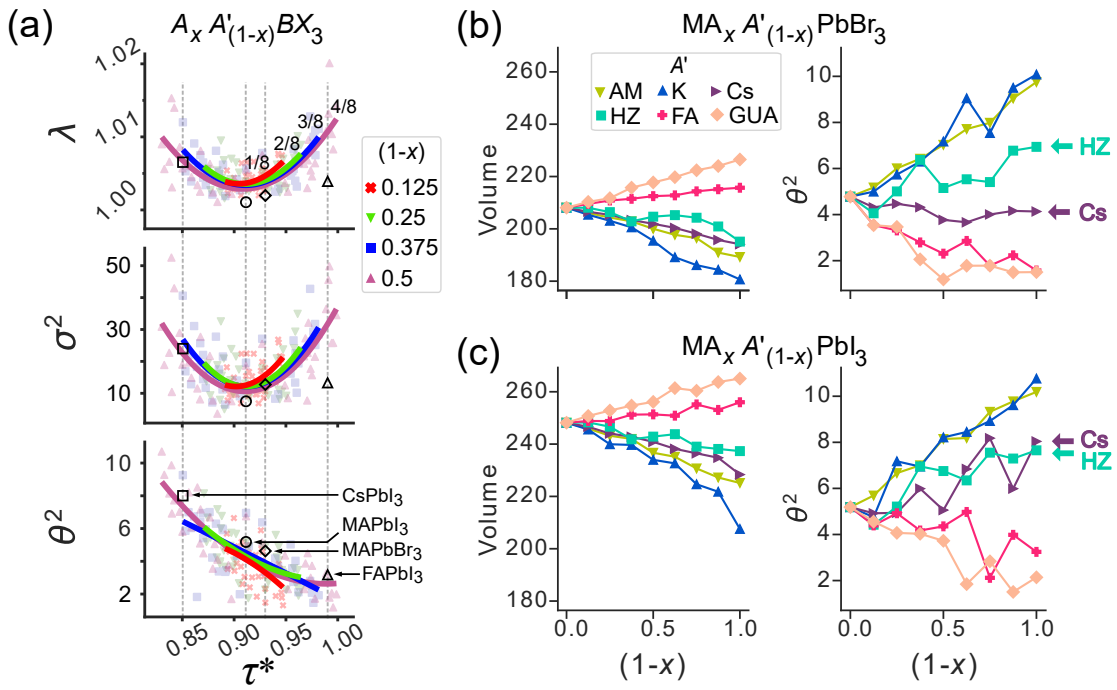

Figure 2. (a) Analysis of the octahedral deformation parameters as a function of τ^* over the entire DFT dataset. The octahedral elongation, λ, the angle variance, σ^2, and the tilt, θ^2, are presented as a function of the extended tolerance factor, τ^*. The A' concentrations $(1-x)$ investigated are 0.125, 0.25, 0.375 and 0.5. The solid lines are fits corresponding to each concentration (color code for both data and fit: 0.125 red, 0.25 green, 0.375 blue, 0.5 purple). Data in panel (a) are for compounds where the inorganic BX_3 component is $SnBr_3$, SnI_3, $PbBr_3$ and PbI_3. In (b) and (c), we report the equilibrium volume (Å^3/f.u.) and θ^2 as a function of the A'-cation concentration for $MA_xA'_{(1-x)}PbBr_3$ and $MA_xA'_{(1-x)}PbI_3$, respectively.

3. Results and discussion

3.1. Structural deformation: high-throughput calculations

The octahedral distortion and tilting in cubic perovskites are, in general, determined by the relative size of the cations with respect to the available void position and by the interaction with the surrounding ions in the BX_6 octahedral network [13, 22, 43]. Thus, displacement and tilt of the cations within the voids are necessary to stabilize the perovskite structure. While an ideal cation preserves the cubic perovskite symmetry, smaller and larger than ideal cations cause the void space to shrink or stretch.

We may relate this local structural change to the extended tolerance factor, which is defined as $\tau^* = (r^*_{A,\mathrm{avg}} + r_X)/\sqrt{2}(r_B + r_X)$. This extended definition is similar to the conventional definition of tolerance factor, with the exception that, for mixed-cation compounds, the ionic radius of the A-site is replaced by the corresponding weighted-average value taken over all the cations in the mixture, namely, $r^*_{A,\mathrm{avg}} = xr_A + (1-x)r_{A'}$. Then, τ^* is used as a descriptor in the conventional way so that a composition is likely to crystallize in the cubic phase for $\tau^* = 1$, in a distorted-cubic phase for $0.8 < \tau^* < 1$, while orthorhombic and hexagonal systems are expected for $\tau^* < 0.8$ and $\tau^* > 1$, respectively. As such, the extended tolerance factor can be used as a driving map in the synthesis of mixed-cation compounds [44], where the designing rule is simply that of fine-tuning $r^*_{A,\mathrm{avg}}$ and hence τ^*.

However, unfortunately, the dependence of the final compound's stability and structure over the specific cation mixture appears more complicated than what is expected by merely averaging the cation radii, particularly for organic cations due to the non-spherical and polyatomic features. In contrast to monatomic cation, the organic cations interact with the surrounding halides by the non-spherically steric effect and the hydrogen-bonds in addition to their charge balancing role. We investigate this resulting structural deformation in detail while analyzing our DFT dataset.

Figure 2(a) helps us in understanding the influence of the cation size on the octahedral distortions by showing the dependence of λ, σ^2 and θ^2 on the extended tolerance factor, τ^*. In the figure, we present results for all available A/A' mixtures and the BX_3 inorganic sublattices $SnBr_3$, SnI_3, $PbBr_3$ and PbI_3. It is clear that the octahedral deformation parameters, λ and σ^2, depend quadratically on τ^*, regardless of the relative concentration of A and A'. However, we also note that the minimum of such parabolic curves is found at different τ^*'s for different compositions; accordingly, we find $\tau^* = 0.90$ for the low-A' concentration

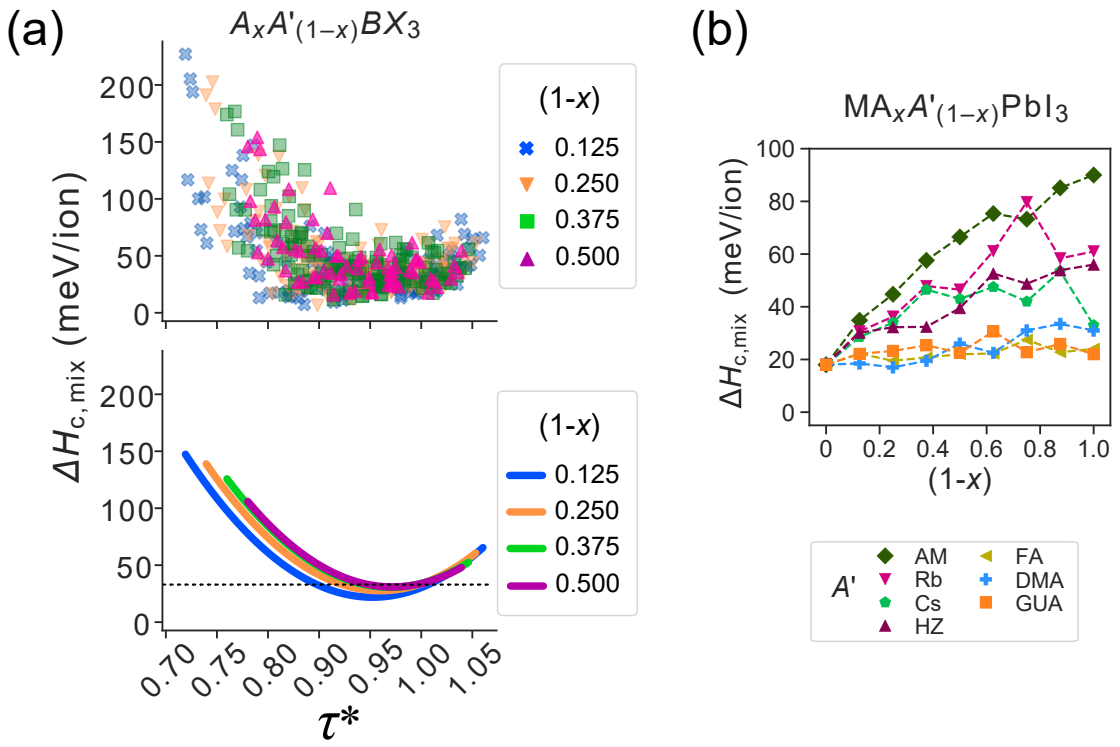

Figure 3. (a) $\Delta H_{c,mix}$ as a function of τ^* for different cation mixtures. The DFT-calculated values are presented in the upper panel, while the corresponding best fits are in the lower one. The colors encode the different cation concentrations investigated. (b) $\Delta H_{c,mix}$ for $MA_xA'_{1-x}PbI_3$ and different A' cations (see legend) as a function of $(1-x)$.

perovskites (perovskite with a significantly more abundant cation type) and $\tau^* = 0.92$ for the high-A' concentration ones (perovskites with a similar cations abundance). This observation means that the minimum of the distribution moves to higher τ^* as the concentration of the second cation A' gets larger (more A' content is mixed). At the same time, the dependence of the tilt parameter θ^2 is more straightforward. It increases as τ^* is reduced from $\tau^* = 1$ (the value attributed to cubic perovskites), displaying an effect associated with the effective size of the A-site cation [45].

More details can be extracted from figures 2(b) and (c), where we present both the cell volume and the octahedral tilt, θ^2, as a function of the A' cation fraction for MA-based perovskites, where the inorganic part is either PbI_3 or $PbBr_3$ and for different A' cations. In general, both the volume and θ^2 depend linearly on $(1-x)$, although the sign of the linear slope is opposite for these two quantities, namely for those cations where the volume increases (decreases) with $(1-x)$, θ^2 gets smaller (larger).

The only exception is for $MA_xCs_{1-x}PbBr_3$, for which both the volume and θ^2 decreases with $(1-x)$, although the variations are actually small. Such anti-correlation between the behavior of the volume and that of θ^2 indicates that the BX_6 framework can substitute MA with a second cation by either expanding the volume and reducing the octahedral tilt or by contracting the volume and becoming more distorted. The fact reveals that the behavior of molecular cations is not solely determined by their effective ionic radius. For instance, one can note that in $MA_xA'_{1-x}PbI_3$ the two cations HZ and Cs contract the cell volume and rotate the PbI_6 octahedra by approximately the same amount at any concentration, despite having significantly different ionic radii ($r_{HZ} = 217$ pm and $r_{Cs} = 188$ pm). Incidentally the ionic radius of HZ is essentially the same as that of MA ($r_{MA} = 217$ pm), meaning that both the volume change and the distortions induced by the substitution of MA with HZ take place regardless of the fact that τ^* remains constant. This behavior is attributed to the more complex non-covalent bonding and repulsion exerted by a given organic cation on the inorganic octahedral network. This fact is associated with the number of lone pairs as earlier presented regarding the organic cation's input features [22].

3.2. Phase stability: DFT calculations

We now turn our attention to the analysis of the phase stability of the mixed-cation perovskites. Figure 3(a) reports the DFT-calculated $\Delta H_{c,mix}$ (top panel) as a function of τ^* for a broad range of cation compositions. The figure includes all the cation mixtures considered and the data are presented with the convention that $x > (1-x)$, thus that, for example, the perovskite $MA_{0.25}Cs_{0.75}PbI_3$ is represented as $Cs_{0.75}MA_{0.25}PbI_3$ in the

plot. The lower panel of figure 3(a) displays the best-fit to the DFT data. We note that all the $\Delta H_{c,mix}(\tau^*)$ curves have a parabolic shape with the minima located in the interval $0.9 < \tau^* < 1$, regardless of the relative cation concentration. Such minimum moves to higher τ^* as the relative cation concentration approaches 50:50. Interestingly, $\Delta H_{c,mix}$ appears clear that structures with an equal cation mixture tend to be less favorable, nevertheless τ^* indicates a cubic structure for a given perovskite yet.

We also observe that the minimum value of $\Delta H_{c,mix}$ is always below the 34 meV ion^{-1} line (dashed line in the figure) within the range of $0.85 < \tau^* < 1.00$. Furthermore, it gets lower as the relative concentration of the cations moves away from 50:50 (the minima are distributed over a range of about 10 meV ion^{-1}). Recalling the definition of $\Delta H_{c,mix}$, the behavior observed here indicates that structures in which one cation type occupies the majority of the available sites are more enthalpically favorable but likely to be more distorted than the equal-proportion mixtures. These, in contrast, seem to prefer a cubic structure. At this point, however, it is worth noting that the spread of the DFT values is pretty large (top panel), much larger than the difference between the best-fit curves. It is then clear that many other factors, in addition to τ^*, determine the stability of a given composition.

It is important to note that our dataset may include some local minima structures because we have not explored all possible molecular orientations and orders in our DFT calculations. Similar to Frost *et al* who performed large-scale calculations employing a $25 \times 25 \times 1$ supercell [46], here we also find distortions of the edge-sharing BX_6 octahedra due to the electric dipole between the neighboring A-sites. However, given the relatively limited size of our supercells, our calculations cannot explore long-range order and may include local minima instead of the global ones. Note that, in general, we observe relatively small energy differences between the total energy of structures relaxed from different initial conditions, indicating that the potential energy surface as a function of the local molecule orientation is relatively flat. This fact gives us confidence that, even if local minima are considered, our data are still suitable for constructing the ML model.

As MA is the most notable organic cation in solution-processed perovskite synthesis, we now focus our analysis on $MA_xA'_{1-x}PbI_3$ (where $0 \leqslant x \leqslant 1$) compounds as a function of the type and concentration of A'. Figure 3(b) presents $\Delta H_{c,mix}$ as a function of the concentration of the A' cation for all the cations investigated. Note that $\Delta H_{c,mix}$ for $(1-x) = 0$ (beginning of the plot) corresponds to the deviation of $MAPbI_3$ from its cubic structure, while for $(1-x) = 1$ (end of the plot) one finds the same quantity for $A'PbI_3$. In all cases $\Delta H_{c,mix}$ at the $(1-x) = 1$ endpoint is higher than at $(1-x) = 0$, meaning that the deviation energy from the cubic structure of MA is less pronounced than that of all the other cations investigated. Because of this quantity definition, $\Delta H_{c,mix}$ generally increases with $(1-x)$, although the rate of increase is significantly different for different A'.

For FA, DMA, and GUA, $\Delta H_{c,mix}$ fluctuates only moderately as their concentration is ramped up, suggesting that such cations can be easily mixed with MA with arbitrary relative abundance. As $\Delta H_{c,mix}$ remains close to its endpoint across the entire composition, one can conclude that the possible mixed perovskites may have a structure close to cubic. Interestingly, the ionic radii of these three cations are all significantly larger than that of MA.

For two other cations, AM and HZ, the situation is similar, namely $\Delta H_{c,mix}$ increases linearly with $(1-x)$, although the relaxation from the cubic phase has much higher energy gain in $AMPbI_3$ and $HZPbI_3$ than in $MAPbI_3$, so that the curves are much steeper. This outcome means that as one moves away from $MAPbI_3$ such two-cation compositions are likely to form heavily distorted structures. Finally, for Cs and Rb $\Delta H_{c,mix}$ is not monotonic with $(1-x)$, presenting a maximum at around $1-x = 0.8$. Cs is particularly interesting because the endpoints have similar $\Delta H_{c,mix}$. This is significantly lower than the values taken for mixed compositions, meaning that Cs-rich $MA_xCs_{1-x}PbI_3$ are unlikely to be formed.

However, it should be noted that this figure displays only the part of the $MA_xA'_{1-x}PbI_3$ dataset of Ge-, Sn-, Pb-halide perovskites. In fact, when one looks at all perovskites made of Group 14 heavy elements and halides, the trend no longer follows a monotonous dependence. It is, in fact, significantly more complex, showing clear non-linearities caused by relationships between the various A, B, and X elements. For this reason, a single universal predictor, capable of learning and describing several $AA'BX_3$ element combinations, cannot be formulated by using linear regression models. Note that a non-linear dependence in the formation energy of cubic-perovskite alloys has been reported previously in literature [47], indicating that the observed deviation of $\Delta H_{c,mix}$ from what expected from a composition-weighted average should not be surprising.

Let us now turn our attention to the enthalpy of mixing, which is presented for $MA_xA'_{1-x}PbI_3$ as a function of $(1-x)$ in figures 4(a) and (b), for A' cations having an effective ionic radius, respectively, smaller and larger than that of MA. In the lower panels, we show the corresponding free energy of mixing also, ΔG_{mix}, at 300 K. This is defined as $\Delta G_{mix} = \Delta H_{mix} - T\Delta S$, where the entropy of mixing is approximated with that of a solid state solution of A and A' cations, namely $\Delta S = k_B[x\ln(x) + (1-x)\ln(1-x)]$, with k_B being the Boltzmann constant.

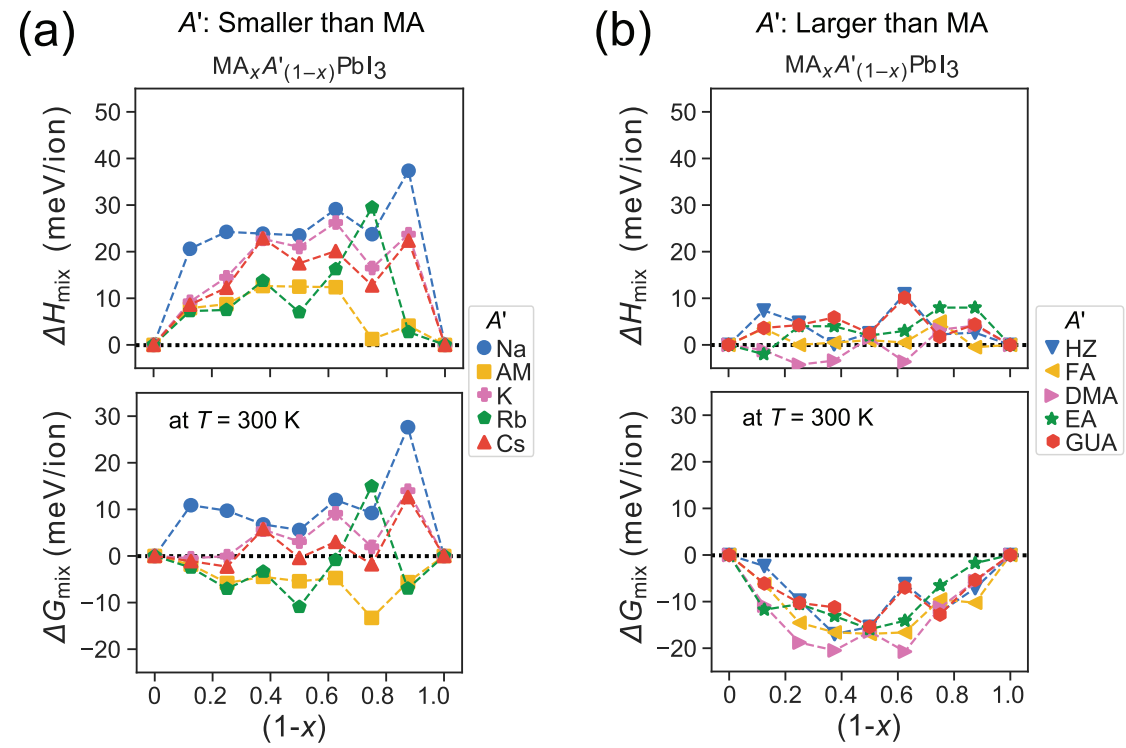

Figure 4. In panel (a) and (b) we plot the enthalpy of mixing, $\Delta H_{\rm mix}$ and the corresponding free energy at 300 K, $\Delta G_{\rm mix} = \Delta H_{\rm mix} - T\Delta S$, for $MA_xA'_{1-x}PbI_3$ perovskites. Panel (a) is for A' cations with ionic radius smaller than that of MA, while panel (b) is for those with a larger one. In evaluating the free energy ΔS is the mixing entropy contribution, which reads $\Delta S = k_B[x\ln(x) + (1-x)\ln(1-x)]$.

For cations smaller than MA (figure 4(a)) the enthalpy of mixing is always positive, meaning that there is never an enthalpy gain in forming two-cation structures. Since the cations interact with the surrounding anions mainly by steric forces, the positive enthalpy of mixing is attributed to the local octahedral deformation introduced by the mixtures. When also entropy is considered (see $\Delta G_{\rm mix}$), among all the investigated A' cations, only AM presents negative free energy of mixing across the entire composition range, a fact that suggests the possibility of forming various $MA_xAM_{1-x}PbI_3$ perovskites, as demonstrated experimentally [48, 49]. There are two other cations displaying a negative $\Delta G_{\rm mix}$ at 300 K, namely Cs^+ and Rb^+, although for only a restricted number of concentrations around MA-rich compositions.

In contrast, when MA is mixed together with another ion, presenting an ionic radius larger than that of MA, $\Delta H_{\rm mix}$ appears in general small, it never exceeds 10 meV ion^{-1}. And, for a few cations at some concentrations, it turns negative (see figure 4(b)). As a result, the corresponding values for the free energy of mixing at 300 K are always negative, indicating that such cations mixtures are likely to be accessible by the synthesis. Notably, MA-based perovskites incorporating a second cation with an ionic radius larger than MA often crystallize into a hexagonal phase [7, 44, 50, 51].

Note that the cubic phase stability criterion, as well as $\Delta H_{\rm mix}$, are expressed as total energy differences and, as such, they are sensitive to the possible uncertainty over the total energy. Assessing the error is not simple because enthalpy data are available only for a handful of compounds. At the same time, a benchmark of our calculations obtained by re-calculating the entire dataset with a different functional is simply not practical, as the numerical cost is high. Here we note that typically PBE total energies are accurate for this class of compounds. The error on total-energy differences is small since the error on total energies is not particularly bias against certain compounds. Thus, in general, we expect that our DFT results provide a quantitative description of the materials trends investigated here.

3.3. ML models: performance

We now discuss the performance of our ML models in predicting the structural properties of the two-cation-mixed perovskites, namely, in predicting λ, θ^2, σ^2, and $\Delta H_{\rm c,mix}$. Figures 5(a) and (b) compare the values of the target variables predicted by both DL and XGBoost against the DFT-calculated results, while we apply the best model for the test dataset.

A visual inspection of figures 5(a) and (b) already suggests that $\Delta H_{\rm c,mix}$ is the target variable better predicted by the two ML models. In fact, we find an MAE of 13.0 and 7.5 meV ion^{-1} for DL and XGBoost,

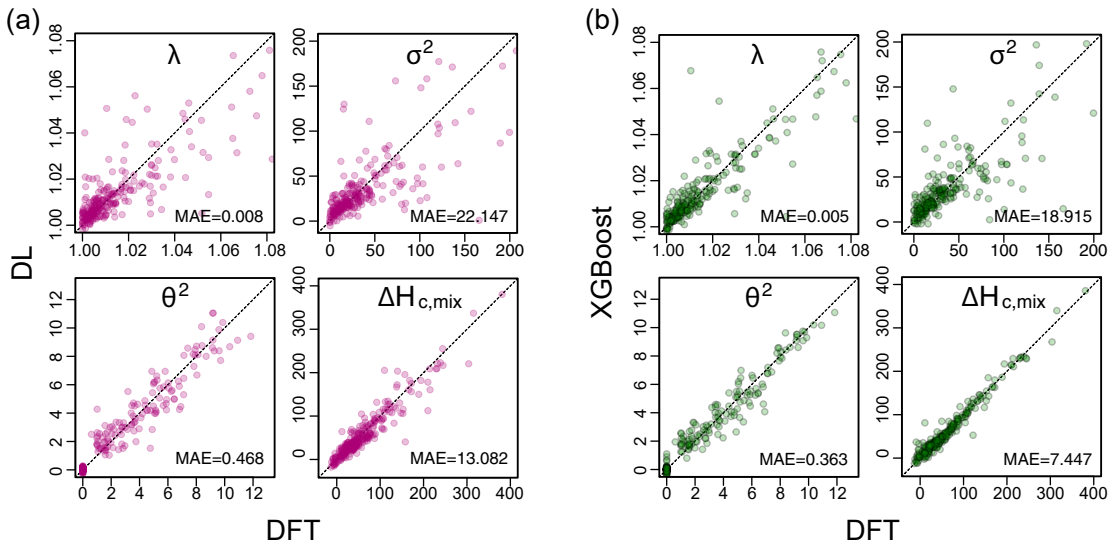

Figure 5. Performance of the two ML models in predicting the perovskites structural properties. (a) DL and (b) XGBoost predictions against the DFT data for λ, σ^2, θ^2 and $\Delta H_{c,mix}$ (meV ion^{-1}).

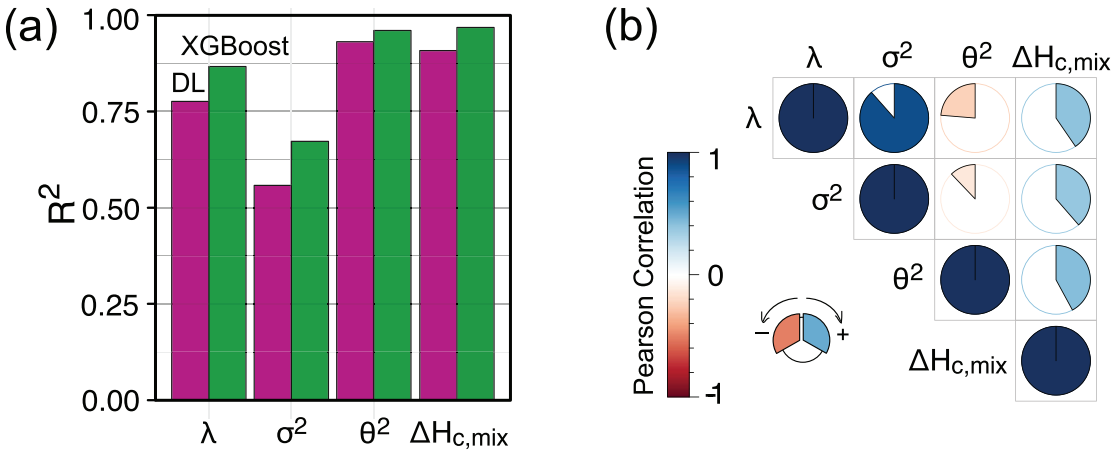

Figure 6. (a) Comparison of the R^2 of the different target variables for the two ML models investigated (DL in purple, XGBoost in green). (b) Pearson correlation coefficients between the various target variables.

respectively, against a quantity that can take values as large as 400 meV ion^{-1}. Such attribution is confirmed by the R^2 coefficient, which is found to be 0.967 for XGBoost and slightly lower for DL. Interestingly, the second-best predicted variable is θ^2, for which the XGBoost R^2 is also pretty high, 0.960 (the MAE is of the order of 0.4 for a quantity that takes values as large as 12).

In order to examine the contribution of the input features to both $\Delta H_{c,mix}$ and θ^2, we have estimated the feature importance of the models (see figure S5 in the SI). Although there are differences in the ranks depending on the target variable and algorithms, the feature importance estimation shows that the weight-averaged, effective ionic, and tolerance factors are commonly high in the ranks. In contrast, the difference in effective ionic radii between A and A' is low. This analysis supports the empirical approaches by which the approximate properties of mixed-ion perovskite are assessed by the weight-average effective ionic radius of the mixed ions instead of the individual radius.

The performance of the two ML models in predicting λ, and σ^2 is less satisfactory, and we were not able to train models with R^2 higher than 0.8 (0.7 for σ^2). In the meantime, we are probably overlooking additional features of relevance for λ, and σ^2. These are likely related to the non-spherical symmetry of the organic cations, a fact which is not taken into full account in the chemical input features of the models.

In general, the two models show a similar prediction capability, as confirmed by the similar R^2 coefficients over the entire target-variable range (see figure 6(a)), with XGBoost slightly outperforming DL. Although the difference between the two algorithms is small, we attribute the slightly better results of the XGBoost model to its bagging property because it preserves the raw dataset's configuration rather than

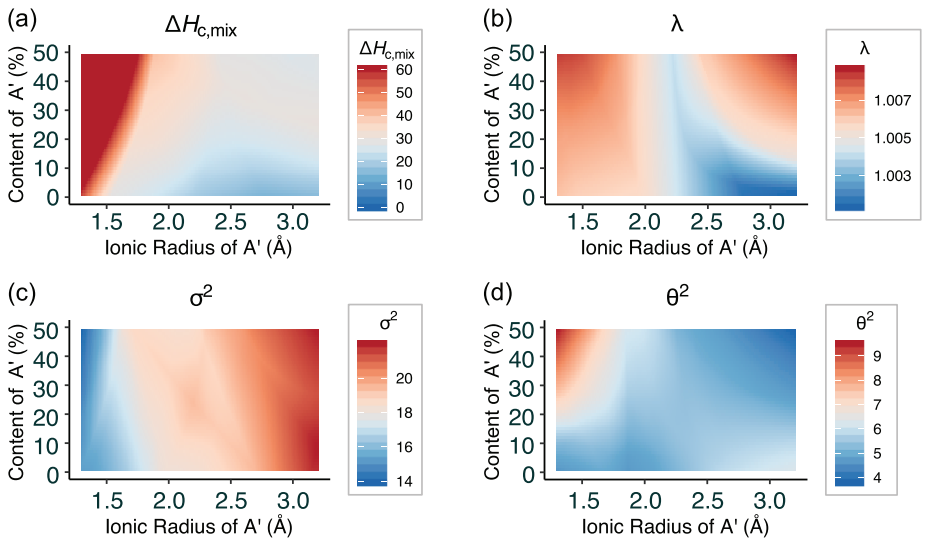

Figure 7. ML predictions for both phase stability and structural properties of $MA_xA'_{1-x}PbI_3$ perovskites as a function of the ionic radius and the A' concentration. In the various panels we present results obtained from the DL model for: (a) $\Delta H_{c,mix}$, (b) λ, (c) σ^2, and (d) θ^2.

interpolating its trend so that its predictions are discrete values (see also figure S7 in the SI). Thus, we speculate that this XGBoost's accuracy can be attributed to the presentation of the dataset values. In contrast, the DL model would be advantageous in accessing the intermediate and dilute concentrations that are inaccessible to DFT calculation due to the large supercells needed.

A better understanding of the possible correlation between the various target variables can be obtained by looking at the Pearson linear correlation coefficients presented in figure 6(b). It is notable that the most correlated variables are the octahedral quadratic elongation, λ, and the angle variance, σ^2, indicating that these types of distortions are always simultaneously present in two-cation perovskites. Besides, we find that all the structural features, λ, σ^2 and θ^2, have a positive linear correlation to the distortion enthalpy of the mixture, $\Delta H_{c,mix}$. As the radius-related features are the main contributor to the target variable rather than each element's chemical properties, the ionic size relative to the surrounding ions plays a crucial role in the structural deformation. As expected, this relationship implies that the most distorted perovskites display the largest enthalpy gain, originating from relaxing the cubic structure into a lower symmetry one.

3.4. ML models: composition and structural predictions

In our previous work, we have used the relaxation enthalpy, defined in equation (2) [22], as a descriptor for the phase stability of mixed-cation perovskites and their tendency to form cubic structures. This descriptor was constructed by interpolating the DFT data obtained for only single-cation perovskites. Taking advantage of our much larger DFT dataset, including several two-cation structures, we now evaluate whether $\Delta H_{c,mix}$ is a better quantity to describe the stability of mixed-ion perovskites. Most importantly, $\Delta H_{c,mix}$ can now be correlated to the various structural target properties defining the compound. Our results, as predicted by DL, are presented in figure 7(a), where we show a map of $\Delta H_{c,mix}$ as a function of the cation radius and of the A' concentration for $MA_xA'_{1-x}PbI_3$.

Such $\Delta H_{c,mix}$ map presents a very distinctive feature. It is a relatively smooth function for $r_{A'} > 2$ Å, while it drastically increases as a function of the A' cation concentration for $r_{A'} < 2$ Å. The boundary is, thus, set close to the ionic radius of MA. This essentially suggests that the incorporation of a second cation together with MA is thermodynamically allowed only at low concentrations and mostly for cations with an ionic radius larger than MA. It is then expected that relatively small cations such as Cs, may incorporate in the perovskite structure, but without forming a solid-state cation solution, namely, they will segregate.

The structural characteristics of mixed MA-based perovskites can then be extracted from the remaining panels, as shown in figures 7(b)–(d). In the small $r_{A'}$ region, one finds that all structures are predicted having a relatively large octahedral quadratic elongation, λ, regardless of the secondary cation concentration. One then expects that, if the small-radius cations can be incorporated in a diluted mixture, the resulting structure will display some tetrahedral distortion. In the same region of parameters, the angle variance, σ^2, remains moderate, while the angle tilt parameter, θ^2, increases as the concentration increases, following a pattern similar to that of $\Delta H_{c,mix}$. In contrast, for $r_{A'} > 2$ Å, one encounters a region extending to concentrations up to 10%, where the octahedral quadratic elongation remains rather close to unity. In the same region, θ^2 is

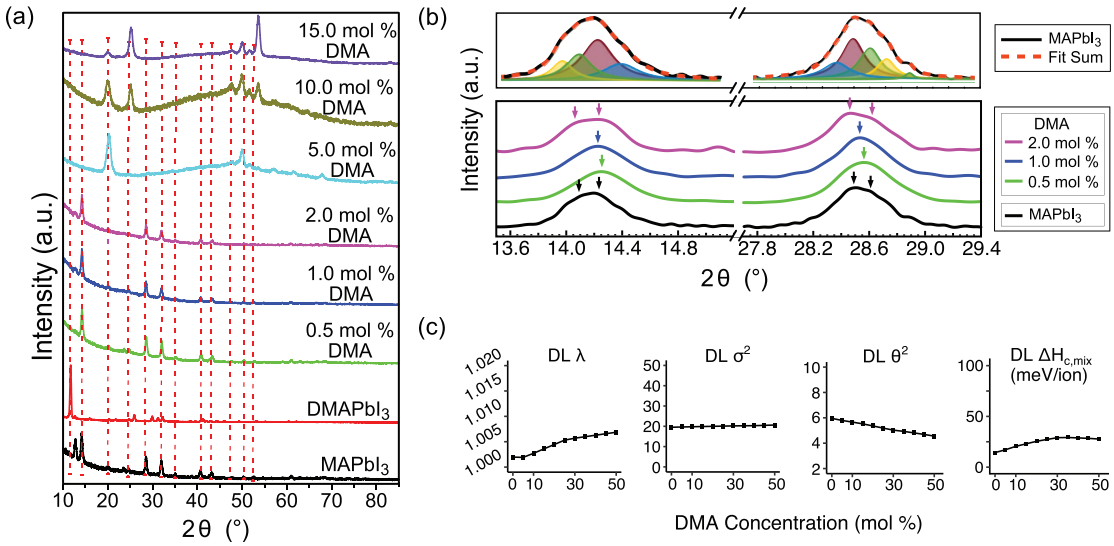

Figure 8. (a) XRD (x-ray diffraction) spectra of the synthesized $MA_xDMA_{1-x}PbI_3$ thin films as a function of the DMA concentration. (b) High-resolution XRD spectra at selected peaks and the corresponding fit for $MAPbI_3$ (top panel) and for perovskites incorporating 0, 0.5, 1.0, and 2.0 mol% DMA (bottom panel). The arrows indicate the location of the main peaks. (c) DL predictions of the structural deformations in $MA_xDMA_{1-x}PbI_3$.

small, while σ^2 is in general large. This representation means that for cations with an ionic radius larger than MA, $MA_xA'_{1-x}PbI_3$ perovskites can be stabilized in an almost cubic structure, where the strain associated with the large cations can be accommodated into local distortions of the PbI_3 octahedra.

To validate our analysis, we have decided to synthesize a family mixed-cation perovskites, where MA was combined with a cation presenting a significantly larger ionic radius. In particular, we have found DMA to be particularly attractive ($r_{A'}$ for DMA is about 270 pm). In fact, for small DMA concentrations, $MA_xDMA_{1-x}PbI_3$ appears to lie in a region of high stability and small distortions.

3.5. $MA_xDMA_{1-x}PbI_3$ thin films at room temperature

We have then explored our predictions on the structural deformation and stability of mixed-cation perovskites by fabricating $MA_xDMA_{1-x}PbI_3$ thin films with $(1-x) = 0.0$ to 0.15 (0–15 mol% DMA cation). Our thin-film fabrication results are presented in figure 8(a), where we present the XRD analysis across the composition range. According to the XRD spectra, the integrity of the 3D perovskite structure is maintained only up to MA concentrations of 2.0 mol%. Beyond such value, the broadening of the main XRD peaks indicates the presence of significant structural distortion induced by the large DMA cation (note the appearance of a robust peak at 20°, which is absent for both pure $MAPbI_3$ and $DMAPbI_3$). Since $MA_{0.85}DMA_{0.15}PbI_3$ has been successfully synthesized in powder form before [50], we speculate that the lack of crystallization of thin films with DMA content above 2 mol% is attributed to the use of a glass substrate.

Despite the restricted range of DMA content for thin-film growth on glass, high-resolution XRD spectra enable us to trace possible structural phase transitions as a function of the DMA concentration (see figure 8(b)). In particular, we analyze in detail the fine structure of the 2θ peaks at 14.2° and 28.6°. Their peak decomposition in the case of $MAPbI_3$ is illustrated in the top panel of figure 8(b), and it is characteristic of the tetragonal phase. As DMA is incorporated into the structure (lower panel of figure 8(b)), the two main peaks forming the signal at 14.2° and 28.6° first merge in a single peak for 0.5 and 1 mol% DMA and then separate again for 2 mol% DMA content. Moreover, as only the minor diffraction at $2\theta = 23.5°$ can be used to distinguish between those phases [52, 53], the evident disappearance of the pattern at 2 mol% DMA incorporation, as shown in figure S11 in the SI. Thus, we observe a tetragonal to cubic transition at room temperature for small DMA concentrations, followed by a new distorted phase for $x > 0.02$.

This behavior is in agreement with the predictions from our ML models (DL), presented in figure 8(c). The model forecasts that the octahedral quadratic elongation, λ, first decreases at moderate DMA contents, and then increases again, thus confirming our experimental findings. Simultaneously, the angle variance, σ^2, and tilt, θ^2, slightly decreases and increases, respectively. More DMA incorporation results in the octahedral tilt reduction as the concentration $(1-x)$ gets larger. The behavior of $\Delta H_{c,mix}$ with $(1-x)$ seems to confirm the trend described by the structural parameters. Namely, it increases by about 5 meV ion^{-1} across an extensive composition range, with such increase being more pronounced at small x. Clearly, we do not expect the models to be quantitative capable of describing our experimental finding because they do not include any

effects arising from the substrate. Note that if the extended tolerance factor is used to analyze the structure evolution as a function of x, one will predict a monotonic increase from the cubic structure as the DMA concentration is increased, in contrast with experiments. This fact highlights the need to use a range of descriptors, instead of a single one, to predict the structural stability of hybrid compounds incorporating molecular cations.

Taking the example of the $MA_xDMA_{(1-x)}PbI_3$, the lowest concentration we could reasonably achieve with DFT correspond to $(1-x) = 12.5\%$ or 1/8 of DMA. On the experimental side, the concentration of interest could be as small as $(1-x) = 2\%$ or 1/50 requiring a simulation cell at least 6.25 times larger. Carrying the DFT calculations on such large systems is highly impractical and poses obstacles in the computing resources since DFT calculation that often scales as $\mathcal{O}(N^3)$, (computational complexity and cost for DFT calculations as a function of the number of electrons N). Hence, $\mathcal{O}(N^3)$ would mean 6.25^3 ($\sim$244) folds increase in required computing resources to obtain the enthalpies for low concentration such as 2 mol% DMA incorporation.

After a careful analysis of the correlation in our dataset, we found that the A-site cation description could be effective if one considers the ionic radius and the number of lone electron pairs as features. Certainly, this choice is not exhaustive. For instance, additional features describing the cation alignment and its dipole moment may be relevant in driving phase transitions [54] and possibly should be included. A detailed assessment of the importance of such additional descriptors will be the subject of future work.

4. Conclusions

The crystal structure of hybrid organic-inorganic perovskites is a decisive factor in determining their stability and electronic properties, so that its control is critical for several applications. With the aim of providing a quantitative map of the structural properties of mixed-cation halide perovskites and hence to guide their synthesis, we have trained machine learning models over an extensive density-functional-theory database to predict the octahedral quadratic elongation, angle variance, and tilt, and the distortion enthalpy of the mixture. These models provide a much more complete view of the structure across vast compositional space and correct the oversimplifications associated with single-parameter empirical descriptors, such as the Goldschmidt tolerance factor.

As an example of our strategy's success, we have analyzed the experimental case of $MA_xDMA_{1-x}PbI_3$. Since the effective ionic radius of DMA is significantly larger than that of MA, the weighted-average radius expects a monotonic increase by the tolerance factor as a function of the DMA concentration. This inference, in turn, implies an increasingly more pronounced deviation from the cubic structure as DMA is incorporated. Our experiments contrast this expectation, and we were able to detect the formation of a cubic MA-DMA mixed phase for < 2 mol% DMA incorporation at room temperature. Our ML models correctly predict such a phase. The ML models show a non-monotonic behavior of the octahedral elongation at small DMA concentrations. Our results highlight the importance of modeling multiple structural descriptors in the study of hybrid perovskites, providing insight into the deformation over a range of chemical compositions in various aspects. The organic cations' alignment order may contribute to predicting the octahedral quadratic elongation and angle variance. Nonetheless, such an alignment is unlikely to persist, being assumed that a solar-cell device operates at room temperature or higher.

Data availability statement

The authors declare that the relevant data supporting the findings of this study are available within the paper and its supporting information files. The supporting information file includes: Details of the mixing combination of the cations; Definition of structural descriptors; SEM images of the representative perovskites; Comparisons of ML prediction. The raw/processed data can be obtained by contacting the authors.

Acknowledgments

This work is sponsored by the Qatar National Research Fund (QNRF) through the National Priorities Research Program (NPRP8-090-2-047) and by the Qatar Environment and Energy Research Institute, through the Novel Materials for Energy program (FE). Computational resources have been provided by the research computing group at Texas A&M University at Qatar. We are grateful to QEERI core labs for the XRD and SEM characterizations. S S thanks the AMBER center (Grant No. 12/RC/2278_P2) for financial support.

ORCID iDs

Heesoo Park https://orcid.org/0000-0002-4276-6843
Adnan Ali https://orcid.org/0000-0003-3976-326X
Raghvendra Mall https://orcid.org/0000-0003-1779-3150
Stefano Sanvito https://orcid.org/0000-0002-0291-715X
Fedwa El-Mellouhi https://orcid.org/0000-0003-4338-9290

References

[1] Mitzi D B, Dimitrakopoulos C D and Kosbar L L 2001 Structurally tailored organic–inorganic perovskites: optical properties and solution-processed channel materials for thin-film transistors *Chem. Mater.* **13** 3728–740
[2] Kojima A, Teshima K, Shirai Y and Miyasaka T 2009 Organometal halide perovskites as visible-light sensitizers for photovoltaic cells *J. Am. Chem. Soc.* **131** 6050–51
[3] Kim H-S *et al* 2012 Lead iodide perovskite sensitized all-solid-state submicron thin film mesoscopic solar cell with efficiency exceeding 9% *Sci. Rep.* **2** 591
[4] Ono L K, Juarez-Perez E J and Qi Y 2017 Progress on perovskite materials and solar cells with mixed cations and halide anions *ACS Appl. Mater. Interfaces* **9** 30197–246
[5] Yi C, Luo J, Meloni S, Boziki A, Ashari-Astani N, Grätzel C, Zakeeruddin S M, Röthlisberger U and Grätzel M 2016 Entropic stabilization of mixed A-cation ABX_3 metal halide perovskites for high performance perovskite solar cells *Energy Environ. Sci.* **9** 656–62
[6] McMeekin D P *et al* 2016 A mixed-cation lead mixed-halide perovskite absorber for tandem solar cells *Science* **351** 151–5
[7] Jodlowski A D *et al* 2017 Large guanidinium cation mixed with methylammonium in lead iodide perovskites for 19% efficient solar cells *Nat. Energy* **2** 972–9
[8] Correa-Baena J-P *et al* 2019 Homogenized halides and alkali cation segregation in alloyed organic–inorganic perovskites *Science* **363** 627–31
[9] Palmstrom A F *et al* 2019 Enabling flexible all-perovskite tandem solar cells *Joule* **3** 2193–204
[10] Xu B *et al* 2017 Bright and efficient light-emitting diodes based on MA/Cs double cation perovskite nanocrystals *J. Mater. Chem.* C **5** 6123–8
[11] Adjokatse S, Fang H-H and Loi M A 2017 Broadly tunable metal halide perovskites for solid-state light-emission applications *Mater. Today* **20** 413–24
[12] Lozano G 2018 The role of metal halide perovskites in next-generation lighting devices *J. Phys. Chem. Lett.* **9** 3987–97
[13] Robinson K, Gibbs G V and Ribbe P H 1971 Quadratic elongation: a quantitative measure of distortion in coordination polyhedra *Science* **172** 567–70
[14] Aydin E, Allen T G, Bastiani M D, Xu L, Ávila J, Salvador M, Kerschaver E V and Wolf S D 2020 Interplay between temperature and bandgap energies on the outdoor performance of perovskite/silicon tandem solar cells *Nat. Energy* **5** 851–9
[15] Motta C, El-Mellouhi F and Sanvito S 2015 Charge carrier mobility in hybrid halide perovskites *Sci. Rep.* **5** 12746
[16] Kubicki D J, Prochowicz D, Hofstetter A, Zakeeruddin S M, Grätzel M and Emsley L 2017 Phase segregation in Cs-, Rb- and K-doped mixed-cation $(MA)_x(FA)_{1-x}PbI_3$ hybrid perovskites from solid-state NMR *J. Am. Chem. Soc.* **139** 14173–80
[17] Pilania G, Balachandran P V, Kim C and Lookman T 2016 Finding new perovskite halides via machine learning *Front. Mater.* **3** 3285
[18] Lu S, Zhou Q, Ouyang Y, Guo Y, Li Q and Wang J 2018 Accelerated discovery of stable lead-free hybrid organic–inorganic perovskites via machine learning *Nat. Commun.* **9** 3405
[19] Graser J, Kauwe S K and Sparks T D 2018 Machine learning and energy minimization approaches for crystal structure predictions: a review and new horizons *Chem. Mater.* **30** 3601–12
[20] Takahashi K, Takahashi L, Miyazato I and Tanaka Y 2018 Searching for hidden perovskite materials for photovoltaic systems by combining data science and first principle calculations *ACS Photonics* **5** 771–5
[21] Park H, Mall R, Alharbi F H, Sanvito S, Tabet N, Bensmail H and El-Mellouhi F 2018 Exploring new approaches towards the formability of mixed-ion perovskite by DFT and machine learning *Phys. Chem. Chem. Phys.* **21** 1078–88
[22] Park H, Mall R, Alharbi F H, Sanvito S, Tabet N, Bensmail H and El-Mellouhi F 2019 Learn-and-match molecular cations for perovskites *J. Phys. Chem.* A **123** 7323–34
[23] Bartel C J, Sutton C, Goldsmith B R, Ouyang R, Musgrave C B, Ghiringhelli L M and Scheffler M 2019 New tolerance factor to predict the stability of perovskite oxides and halides *Sci. Adv.* **5** eaav0693
[24] Ali A, Park H, Mall R, Aïssa B, Sanvito S, Bensmail H, Belaidi A and El-Mellouhi F 2020 Machine learning accelerated recovery of the cubic structure in mixed-cation perovskite thin films *Chem. Mater.* **32** 2998–3006
[25] Yamamoto K, Iikubo S, Yamasaki J, Ogomi Y and Hayase S 2017 Structural stability of iodide perovskite: a combined cluster expansion method and first-principles study *J. Phys. Chem.* C **121** 27797–804
[26] Guedes-Sobrinho D, Guilhon I, Marques M and Teles L K 2019 Thermodynamic stability and structural insights for $CH_3NH_3Pb_{1-x}Si_xI_3$, $CH_3NH_3Pb_{1-x}Ge_xI_3$ and $CH_3NH_3Pb_{1-x}Sn_xI_3$ hybrid perovskite alloys: a statistical approach from first principles calculations *Sci. Rep.* **9** 11061
[27] Chen C, Zuo Y, Ye W, Li X, Deng Z and Ong S P 2020 A critical review of machine learning of energy materials *Adv. Energy Mater.* **10** 1903242
[28] Perdew J P, Burke K and Wang Y 1996 Generalized gradient approximation for the exchange-correlation hole of a many-electron system *Phys. Rev.* B **54** 16533–9
[29] Perdew J P, Burke K and Ernzerhof M 1996 Generalized gradient approximation made simple *Phys. Rev. Lett.* **77** 3865–8
[30] Tkatchenko A and Scheffler M 2009 Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data *Phys. Rev. Lett.* **102** 073005
[31] Blöchl P E 1994 Projector augmented-wave method *Phys. Rev.* B **50** 17953–79
[32] Kresse G and Joubert D 1999 From ultrasoft pseudopotentials to the projector augmented-wave method *Phys. Rev.* B **59** 1758–75
[33] Kresse G and Hafner J 1993 Ab initio molecular dynamics for liquid metals *Phys. Rev.* B **47** 558–61
[34] Kresse G and Furthmüller J 1996 Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set *Phys. Rev.* B **54** 11169–86

[35] Kresse G and Furthmüller J 1996 Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set *Comput. Mater. Sci.* **6** 15–50
[36] Schelhas L T *et al* 2019 Insights into operational stability and processing of halide perovskite active layers *Energy Environ. Sci.* **12** 1341–8
[37] Curtarolo S, Hart G L W, Nardelli M B, Mingo N, Sanvito S and Levy O 2013 The high-throughput highway to computational materials design *Nat. Mater.* **12** 191
[38] Schmidt J, Marques M R G, Botti S and Marques M A L 2019 Recent advances and applications of machine learning in solid-state materials science *npj Comput. Mater.* **5** 83
[39] Chen T and Guestrin C 2016 XGBoost: a scalable tree boosting system *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining* (*New York*) pp 785–94
[40] Chen C, Zuo Y, Ye W, Li X, Deng Z and Ong S P 2020 A critical review of machine learning of energy materials *Adv. Energy Mater.* **10** 1903242
[41] The H2O.ai Team. H2O: Scalable Machine Learning Version 3.25.0.4629 (available at: www.h2o.ai) (Accessed 10 September 2019)
[42] Park B-W and Seok S I 2018 Intrinsic instability of inorganic–organic hybrid halide perovskite materials *Adv. Mater.* **31** 1805337
[43] Bechtel J S and Van der Ven A 2018 Octahedral tilting instabilities in inorganic halide perovskites *Phys. Rev. Mater.* **2** 025401
[44] Li Z, Yang M, Park J-S, Wei S-H, Berry J J and Zhu K 2016 Stabilizing perovskite structures by tuning tolerance factor: formation of formamidinium and cesium lead iodide solid-state alloys *Chem. Mater.* **28** 284–92
[45] Lee J-H, Bristowe N C, Bristowe P D and Cheetham A K 2015 Role of hydrogen-bonding and its interplay with octahedral tilting in $CH_3NH_3PbI_3$ *Chem. Commun.* **51** 6434–7
[46] Frost J M, Butler K T and Walsh A 2014 Molecular ferroelectric contributions to anomalous hysteresis in hybrid perovskite solar cells *APL Mater.* **2** 081506
[47] Dalpian G M, Zhao X-G, Kazmerski L and Zunger A 2019 Formation and composition-dependent properties of alloys of cubic halide perovskites *Chem. Mater.* **31** 2497–506
[48] Oku T, Ohishi Y,and Ueoka N 2018 Highly (100)-oriented $CH_3NH_3PbI_3$(Cl) perovskite solar cells prepared with NH_4Cl using an air blow method *RSC Adv.* **8** 10389–95
[49] Si H, Zhang Z, Liao Q, Zhang G, Ou Y, Zhang S, Wu H, Wu J, Kang Z and Zhang Y 2019 A-Site management for highly crystalline perovskites *Adv. Mater.* **32** 1904702
[50] Franssen W M J, Bruijnaers B J, Portengen V H L and Kentgens A P M 2018 Dimethylammonium incorporation in lead acetate based $MAPbI_3$ perovskite solar cells *ChemPhysChem* **19** 3107–15
[51] Ke W, Spanopoulos I, Stoumpos C C and Kanatzidis M G 2018 Myths and reality of $HPbI_3$ in halide perovskite solar cells *Nat. Commun.* **9** 4785
[52] Baikie T, Fang Y, Kadro J M, Schreyer M, Wei F, Mhaisalkar S G, Graetzel M and White T J 2013 Synthesis and crystal chemistry of the hybrid perovskite $(CH_3NH_3)PbI_3$ for solid-state sensitised solar cell applications *J. Mater. Chem.* A **1** 5628–41
[53] Song Z, Watthage S C, Phillips A B, Tompkins B L, Ellingson R J and Heben M J 2015 Impact of processing temperature and composition on the formation of methylammonium lead iodide perovskites *Chem. Mater.* **27** 4612–9
[54] Maheshwari S, Fridriksson M B, Seal S, Meyer J and Grozema F C 2019 The relation between rotational dynamics of the organic cation and phase transitions in hybrid halide perovskites *J. Phys. Chem.* C **123** 14652–61