Revealing the Role of Local Octahedral Distortions in Hybrid Halide Perovskites Through Physical-Informed Data-Driven Machine Learning

Xian Chen [1], Zhe-Ning Chen [2], Tianmin Wu [2, 3, *], Jun Chen [2 *]

1   *College of Information Engineering, Fujian Business University, Fuzhou 350015, P. R. China*

2   *Fujian Provincial Key Laboratory of Theoretical and Computational Chemistry, Xiamen 361005, P. R. China*

3   *Key Laboratory of Opto-Electronic Science and Technology for Medicine of Ministry of Education, Fujian Provincial Key Laboratory of Photonics Technology, College of Photonic and Electronic Engineering, Fujian Normal University, Fuzhou 350117, P. R. China*

* To whom correspondence should be addressed. E-mails: wtm@fjnu.edu.cn (T.W.); chen@xmu.edu.cn (J.C.)

Keywords: Machine learning; Structure-property relationship; Local octahedral distortion; First-principles calculation; Hybrid halide perovskites

Abstract

Data-driven materials design often encounters challenges from small and imbalanced datasets. The complex structural and physicochemical properties of hybrid halide perovskites, coupled with these limitations, create obstacles for performing feature engineering and extracting key fingerprints. Herein, we employed a physical-informed data-driven modeling approach to identify lattice geometric fingerprints, such as the distortion index (DI) and effective coordination number (ECoN), and to establish a robust structure-properties relationship mapping the electronic bandgap, resulting in improved model performance. Lattice compression simulations across multiple phases of MAPbI$_3$ further confirmed a strong correlation between DI and ECoN with electronic bandgap, validating the robustness of the selected octahedra geometrical fingerprints. By adjusting the *s-p* antibonding coupling, the pressure-driven reduction in local octahedral distortion, induced by the anisotropic hydrogen bonding between the inorganic framework and organic cation, narrows the electronic bandgap and facilitates the *p-p* transitions, thereby boosting the transition dipole moment and band-edge absorption. Combining data mining with physical analysis, we have successfully clarified the significant impact of lattice geometry on the electronic properties and identified key octahedral geometric fingerprints for effectively describing the electronic bandgap, while also revealed the microphysical mechanisms of local octahedral distortion on the optoelectronic properties of hybrid halide perovskites.

1. Introduction

The efficient harvesting, storage, and utilization of clean energy are fundamental to combating climate change and achieving sustainable development[1-4]. Solar energy, due to its wide availability and low environmental impact, ranks among the most significant renewable energy sources and drives the growing demand for photovoltaic technologies[4]. Exploring prospective photovoltaic materials with superior performance and excellent stability for energy production is currently one of the biggest challenges and has attracted significant attention[5, 6]. Hybrid halide perovskites have emerged as a class of promising candidates for various optoelectronic applications with unique virtues[7-9], including excellent optical absorption, highly adjustable bandgap, long carrier diffusion lengths, and low exciton binding energy. Besides, the ease of fabrication, the low-cost of compositional precursors, and the solution processability have dramatically stimulated the development of perovskites solar cells[10, 11], with power conversion efficiency (PCEs) had boosted from 3.8%[12] to a certified value of 26.7% in a short time[13, 14]. Despite unprecedented progress and promising applications, designing and evaluating new photovoltaic materials for further performance enhancement remains challenging, due to the complex diversity of hybrid halide perovskites and limited understanding of the structure-composition-property relationship[10, 11, 15].

Recently, data-driven material design methodologies based on machine learning (ML), an interdisciplinary field that integrates computer science, mathematics, and material science, have been envisioned as a decisive enabler for accelerating the discovery of various functional materials[16]. Machine learning approaches typically adopt a framework that extracts descriptors from the atomic, geometric, and physicochemical properties, and identified several key target properties (e. g., electronic bandgap[17-19], carrier mobility[20], and formation energy[21], etc.) through statistical analysis or learning-based importance analysis as criteria for screening candidates for various applications. These descriptors are subsequently employed in machine learning models to establish a rigorous mapping to the target properties, enabling the rapid traversal of the vast chemical space and facilitating the screening of

potential candidates. Given the vast chemical space and complex structural, physicochemical properties of hybrid halide perovskites, machine learning has become a powerful tool favored by researchers for accelerating material design[20, 22, 23], and it is even directly integrated with high-throughput experiments to optimize the performance and manufacturing of perovskite solar cells. Although the integration of big data, artificial intelligence, and materials modeling aims to achieve a paradigm shift in materials design by fast deriving rules and trends from existing data to guide material development without understanding the underlying physical mechanisms, the common issue of small and imbalanced datasets often results in significant sacrifices in accuracy. Feature engineering, which leverages experience to construct appropriate and suitable descriptors or fingerprints instead of relying solely on traditional atomic physicochemical descriptors[22], promises to address the prevalent issues. However, implementing feature engineering itself presents a significant challenge.

Feature engineering has been employed to extract key descriptors for hybrid halide perovskites, including the electronegativities of constituent elements[19, 20], Goldschmidt tolerance factor (TF)[17], and packing factor (PF)[18, 19], to effectively establish structure-property relationships and accelerate material design. However, the significant impact of lattice geometry on atomic orbital overlap, and consequently on the electronic bandgap[24], emphasizes the need for novel and effective geometrical fingerprints. Given this, a physical-informed data-driven modeling approach is employed to assess the impact of local octahedra geometric fingerprints on the electronic properties, with the goal of establishing a robust structure-property relationship mapping the electronic bandgap and improving machine learning performance. Through feature engineering, we identified key octahedral geometric fingerprints, such as the distortion index (DI) and effective coordination number (ECoN), to capture complex trends and correlations within the chemical space of hybrid halide perovskites. Further lattice compression simulations across multiple phases (cubic, tetragonal and orthorhombic) of MAPbI$_3$ demonstrate that the pressure-driven non-monotonic variation of DI and ECoN closely align with the electronic bandgap, highlighting the strong correlation between local octahedral distortion and electronic properties, and validating the interpretability and

explainability of machine learning models. By modulating the *s-p* antibonding coupling, the local O_h symmetry of PbI_6 octahedra, reflected in DI and ECoN, shows a strong correlation with the electronic bandgap. Reducing local octahedral distortion facilitates *p-p* transitions, thereby enhancing both the transition dipole moment, carrier mobility, and band-edge absorption. Moreover, the local octahedral distortion primarily originates from anisotropic hydrogen bonding between the inorganic framework and NH_3 group in MA^+ in hybrid halide perovskites. Combining data mining with physical analysis, current study clarified the significant impact of lattice geometry on electronic properties, identified key octahedral geometric fingerprints for describing the electronic bandgap, while also revealed the microphysical mechanisms of local octahedral distortion on the optoelectronic properties of hybrid halide perovskites.

2. Methodology

To systemically validate the capability of the selected configuration feature characteristics for describing the electronic bandgap of hybrid halide perovskites, a variety of classical and advanced machine learning algorithms (13 models) are introduced. These algorithms include Gradient Boosting Regression (GBR), Supporting Vector Regression (SVR), Decision Tree (DT), Kernel Ridge Regression (KRR), K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), Random Forest (RF), LASSO, extremely Randomized Trees (ExtraTree), Gradient Boosting Decision Tree (GBDT), eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Categorical Boosting (CatBoost). To evaluate the accuracy and generalization ability of the trained model for all these machine learning algorithms, a 10-fold cross-validation was employed. Specifically, to assess the performance of these machine learning models, common metrics: mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (R^2 score) are unitized in the current study. These metrics are defined as follows: $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|$, $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$, and $R^2 = \sum_{i=1}^{n}(y_i - \widehat{y_i})^2 / \sum_{i=1}^{n}(y_i - \bar{y})^2$, where *n* is the number of observations, y_i is the actual observed value of the response variable for the *i*-th observation, $\widehat{y_i}$ is the predicted value of the response variable for the *i*-th observation according to the model,

and $\bar{y}$ is the mean of the observed values. The SHapley Additive exPlanations (SHAP) method was used to extract and analyze important features of atomic physicochemical and geometric descriptors, revealing their relationship with electronic bandgap.

Furthermore, the formula for electronic bandgap of hybrid halide perovskites was identified using the sure independence screening and sparsifying operator (SISSO) method[25], a machine learning method with an exceptional performance on huge-dimensional space. Initially, 10 features were selected from 36 physical chemistry quantities (**Table S1**) through Least Absolute Shrinkage and Selection Operator Regression (LASSO) algorithm as the primary features. The feature space was denoted as ϕ_0 and served as the starting point for the SISSO process. A set of algebraic and functional operations (**Table S2**) were iteratively applied to the initial feature space ϕ_0 to construct augmented feature spaces with increased dimensionalities. After one, two, and three iterations, three feature spaces ϕ_1, ϕ_2, and ϕ_3 were generated, containing 370, 411092, and 7.28×10^{10} features, respectively. To identify the formula, an efficient dimensionality reduction method, sequential importance sampling (SIS), was unitized to screen the features and identify a subset with the largest correlation with the response. A key hyperparameter to tune was the dimensionality of the solution, i.e., the dimensionality of the identified formula represented by D. The fitted regression formula was written as: $y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_D x_D$, where y is the response, β_i and x_i are the solved coefficient and feature value for the i-th dimension, respectively. To evaluate the predictive performance of different dimensionality, cross-validation techniques are implemented. In detail, the training data was divided into ten subsets, and the formula was then identified using nine of these subsets, while the remaining subset was reserved to assess the prediction errors ($RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}$), in which n is the number of observations, y_i and $\widehat{y_i}$ are the actual observed value and the predicted value of the response variable according to the model for the i-th observation, respectively.

The first-principles calculations were performed using the density-functional theory formalism implemented in the Vienna Ab-initio Simulation Package (VASP)[26-28]. The frozen-core projected augmented wave (PAW) pseudopotentials were employed to

describe the electron-core interactions[29]. The generalized gradient approximation (GGA) formulated by Perdew, Burke, and Ernzerhof (PBE) was utilized as the exchange-correlation functional[30]. Electrons taken to be valence are $1s^1$ of H, $2s^2 2p^2$ of C, $2s^2 2p^3$ of N, $5d^{10} 6s^2 6p^2$ of Pb, and $5s^2 5p^5$ of I. The effect of van der Waals (vdW) dispersion interactions was estimated with the pairwise D3 three-body dispersion correction proposed by Grimme (DFT-D3)[31]. The electronic wavefunctions were expanded using a plane-wave basis with a cutoff energy of 520 eV, and the sampling of the first Brillouin zone was adopted by applying the Monkhorst-Pack scheme. The cell parameters and atomic positions were fully relaxed until the energy and residual forces on atoms converged to 1×10^{-7} eV and 0.001 eV/Å by applying the conjugate-gradient algorithm, respectively. A reciprocal space sampling with a Γ center of $7 \times 7 \times 7$ $\boldsymbol{k}$-point mesh of the Brillouin zone was applied in the cubic phase, $5 \times 5 \times 3$ k-mesh for tetragonal and orthorhombic structural optimization. Finer $\boldsymbol{k}$-point meshes were adopted for the calculations of electronic band structure and projected density of states (DOS) for each configuration. The absorption coefficient was defined as $\alpha(\omega) = \frac{\sqrt{2}e}{\hbar c} \left[(\varepsilon_1^2 + \varepsilon_2^2)^{\frac{1}{2}} - \varepsilon_1 \right]^{\frac{1}{2}}$, in which ε_1 and ε_2 represent the real and imaginary parts of the dielectric function, respectively. The crystal lattice free volume was calculated by using the Multiwfn 3.3.9 software[32]. The transition dipole moments were calculated based on the wavefunction from VASP using VASPKIT code[33].

3. Results and discussion

3.1 Data-Driven Revelation of the Relationship between Local Octahedral Geometric Fingerprints and Electronic Bandgaps

Although geometric parameters significantly influence the overlap of atomic orbitals and, consequently, the electronic bandgap of hybrid halide perovskites, visualization of dataset achieved from previous high throughput first-principles calculations (**Figure 1a**) reveals unclear trends and periodicities between the extract geometric fingerprints and the electronic bandgap. In detail, the definitions of the selected geometric fingerprints are given in **Table S1**. To ensure the generalizability of the closed structure-property relationship between the geometric features and electronic

1 bandgap, we selected all types of crystal structures (cubic, tetragonal, and orthorhombic)

2 with electronic bandgaps calculated using the PBE functional. Pearson correlation

3 coefficient matrix between pairs of 10 selected fingerprints (e. g. the electronegativity

4 of halide and geometric features), along with the matrix for all features, as illustrated

5 in **Figure 1b** and **S1**, respectively. The evaluated low linear correlation coefficients

6 clearly indicate that the nonredundant descriptors are constructed and robust feature

7 engineering is applied. Then, multiple machine learning algorithms are employed to

8 evaluate the impact of the geometrical fingerprints of the local MX_6 octahedra on the

9 electronic bandgap, and four indices are introduced to estimate the prediction errors:

10 coefficient of determination (R^2), mean absolute error (MAE), mean squared error

11 (MSE). The results shown in **Figure 1c** and **S2**, clearly indicates the robustness of the

12 newly proposed geometrical fingerprints in predicting electronic properties, with a

13 notable improvement in prediction accuracy across all machine learning models once

14 the geometric features are incorporated.
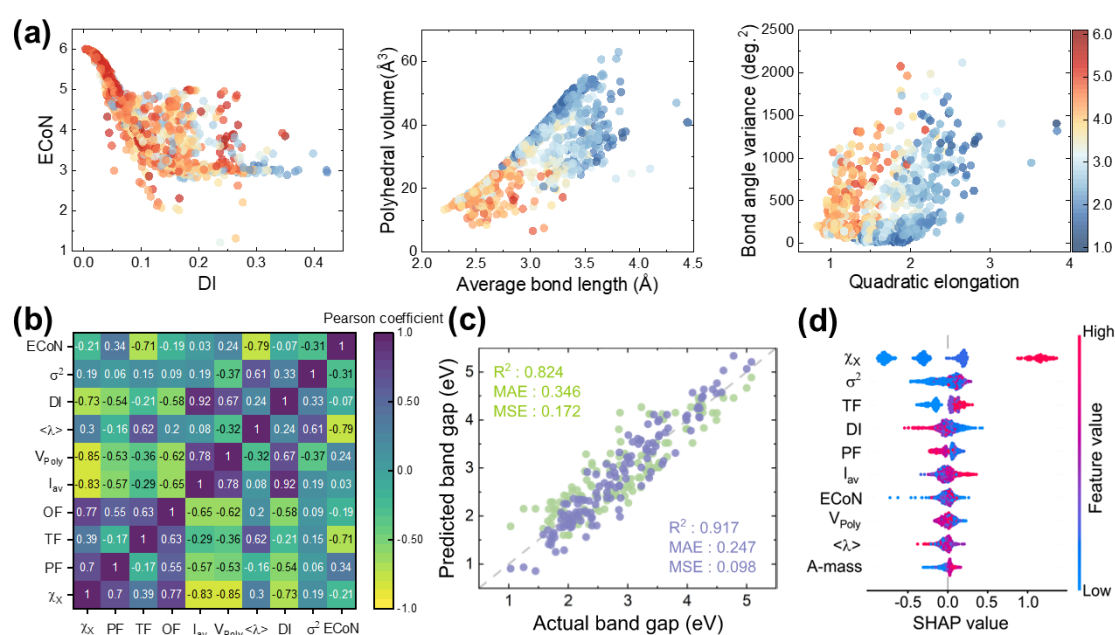
16 **Figure. 1 (a)** Data visualization of the electronic bandgap for hybrid halide perovskites based on

17 octahedral geometric parameters. **(b)** Heat map of the Pearson correlation coefficient matrix among

18 10 selected features. **(c)** Fitting results comparing actual and predicted electronic bandgap using

19 XGBoost with (Green) and without (Violet) geometrical fingerprints, along with R^2, MAE, and

20 MSE to estimate the prediction errors. **(d)** SHAP value plots highlighting the top 10 features

21 identified by XGBoost.

The SHAP method is adopted to interpret the predictions of complex machine learning models and to assess the importance and contributions of various fingerprints in predicting the bandgap of hybrid halide perovskites. The SHAP value plots of top 10 features (XGBoost), as shown in **Figure 1d**, highlight that the electronegativity of halide (χ_X), TF, and PF significantly correlate with the electronic properties. Additionally, bond-angle variance (σ^2) and DI, average bond length (l_{av}), effective coordination number (ECoN), and bond length quadratic elongation ($<\lambda>$) of MX_6 octahedra also play a crucial role. χ_X plays the most important role in predicting the electronic bandgap, with the increase in the electronegativity from I to F positively affecting the expected bandgap. Higher TF and OF are also associated with a widening of bandgap output, while an increase in the PF shortens the average M-X bond length (l_{av}) and reduces the polyhedral volume (V_{Poly}), thereby narrowing the bandgap. Interestingly, although both reflect the local distortion of the MX_6 octahedra, the σ^2 has a favorable influence, while the DI has a somewhat adverse effect on the bandgap prediction. ECoN of MX_6 octahedra positively contributes to bandgap prediction, while $<\lambda>$ has a negative effect.

To address the issue of multiple descriptors influencing the electronic bandgap and to enhance the interpretability of the machine learning algorithm, the SISSO approach is employed to identify reliable, low-cost formulas through mathematical operations between these features. These formulas describe the relationship between ten primary descriptors and the electronic bandgap, expressed in a nonlinear functional form due to the compressed sensing methodology with different feature dimensions (**Figure S3**). The identified SISSO formula for the electronic bandgap demonstrates promising predictive accuracy compared to DFT calculations, with an RMSE as low as 0.38 eV (**Figure S4**). Several top 3D descriptor identified by SISSO of ϕ_3 (**Table S3**):

$$\frac{\chi_X \times TF}{e^{TF} \times (\chi_X{}^3 - e^{\chi_X})}, \frac{(OF \times DI) - \chi_X}{\chi_X{}^3 - e^{\chi_X}}, \frac{TF^3 - e^{TF}}{\chi_X{}^3 - e^{\chi_X}}, |(\chi_X \times DI) - e^{-PF}| - |e^{-PF} - (DI \times ECoN)|.$$

In addition to halide electronegativity (χ_X), TF, and octahedral factor (OF), and local octahedral geometric parameters: DI, and ECoN, are identified as key descriptors. While machine learning algorithm has highlighted the significant influence of local

octahedral geometrical parameters on the electronic structure, the unclear physical understanding drives further exploration of how changes in the local octahedral geometrical parameters affect their electronic structure and optoelectronic properties, particularly in relation to the interaction between the organic cations and metal halide framework.

## 3.2 Physics-Driven Validation of the Impact of Local Octahedral Geometric Fingerprints on the Electronic Structure
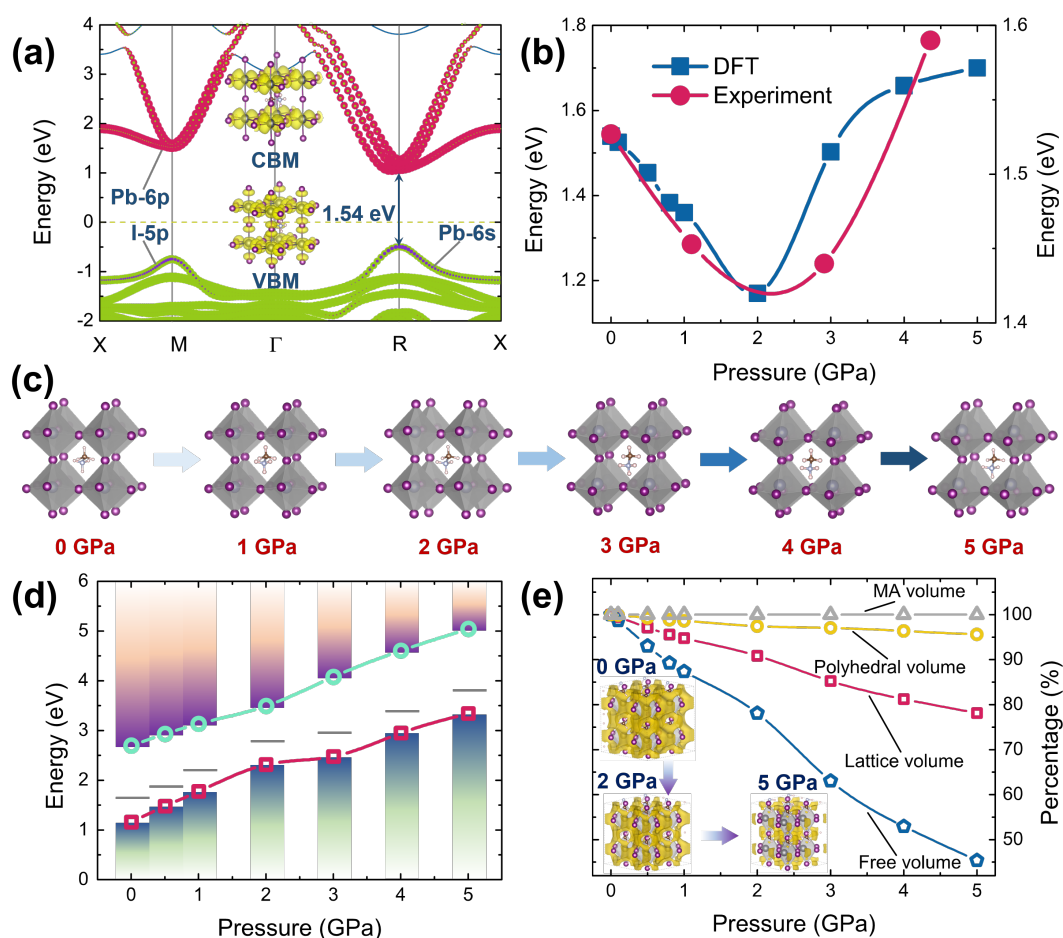
Figure. 2 (a) Orbital projected electronic band structure of cubic MAPbI$_3$. Pressure-induced **(b)** variation of the electronic bandgap with experimental values from Ref. 26, **(c)** changes in crystal structure, **(d)** changes in VBM and CBM, and **(e)** variation in lattice volume, free volume, polyhedral volume, and MA cation volume. Yellow color in the inset represents the free volume at different pressures.

The hybrid nature and soft inorganic framework in organic-inorganic hybrid halide perovskites make their structural alterations and optoelectronic properties susceptible

to external stimuli, thus rendering the lattice compression a powerful tool for investigating the impact of local octahedral geometrical parameters on their electronic structure. To further clarify the structure-property relationship, the representative material MAPbI$_3$ is selected for additional study. The calculated unpressurized orbital-projected electronic band structure of cubic MAPbI$_3$ is shown in **Figure 2a**, where the valence band maximum (VBM) is dominated by an antibonding hybrid state between Pb-*6s* and I-*5p* orbitals, while the conduction band minimum (CBM) primarily originates from Pb-*6p* orbitals, with an electronic bandgap of 1.51 eV. Under pressure, although VBM and CBM remain dominated by a hybrid state between Pb-*6s* and I-*5p* orbitals and Pb-*6p* orbital (**Figure S5-S7**), respectively, both the crystal structure and electronic structures undergo significant changes, as shown in **Figure 2b-c** and **S5-S7**. Notably, the electronic bandgap exhibits a significant red shift under pressure, followed by significant blue shifts. This nonmonotonic variation, where the bandgap initially decreases and then increases with external pressure, well aligns with previous experimental observations. Despite that such pressure-induced nonmonotonic behavior is commonly observed in other organic-inorganic hybrid halide perovskites, its atomic-level mechanism remains unclear. CBM, primarily originating from the Pb-*6p* orbital, increases linearly with rising pressure. In contrast, VBM, influenced by an antibonding hybrid state between Pb-*6s* and I-*5p* orbitals, initially increases steeply but then slows, resulting in a nonmonotonic variation of the electronic bandgap as function of pressure. Interestingly, the electronic bandgap of MAPbI$_3$ reaches an optimal value of 1.34 eV at ~1.5 GPa, enabling the highest power-to-electricity conversion efficiency (PCE) for solar cells, as predicted by the Shockley-Queisser limit.

Regarding structural changes (**Figure 2e**), the significant reduction in lattice volume with increasing pressure is mainly attribute to a notable decrease in free volume, while the relatively rigid MA$^+$ and metal halide octahedral structures experience smaller changes. Initially, as shown in **Figure 3a**, the lattice volume reduction is driven by the shortening of Pb-I bond lengths (< 2 GPa), followed by a decrease in Pb-I-Pb bond angles (> 2 GPa). The sharp decline of Pb-I-Pb bond angle (> 2 GPa), reflecting increased global distortion between adjacent octahedra, suggests the occurrence of

significant octahedral tilting or a pressure-induced phase transition. Additionally, the shortening of the Pb-I bond with pressure enhances the coupling between Pb-*6s* and I-*5p* antibonding orbitals, leading to an increase of VBM (**Figure 3b**). In contrast, the localized nature of the Pb-*6p* orbitals makes CBM relatively insensitive to the shortening of the Pb-I bond lengths, resulting in a narrowing of the electronic bandgap. The sharp decrease in the Pb-I-Pb bond angle with pressure (> 2 GPa) causes the partially broken of Pb-I bonding[24, 34] and the weakening of Pb-*6p* and I-*5s* orbital coupling, resulting in the widening the electronic bandgap (**Figure 2b**). The synergy between Pb-I bonds and Pb-I-Pb bond angles provides a detailed atomic-scale explanation for the nonmonotonic variation of the electronic bandgap with pressure. Given the critical role of the inorganic framework in determining the band-edge electronic structure, understanding its microstructural physical parameters is essential and warrants significant attention[35-38].
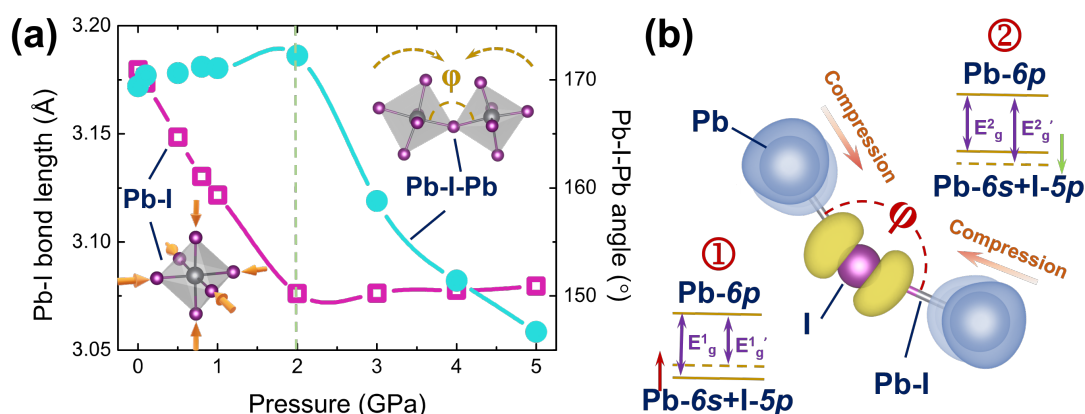
Figure. 3 (a) Variation of average Pb-I bond length (purple) and average Pb-I-Pb bond angle (cyan) as function of pressure. **(b)** Schematic representation of the bandgap **(1)** red- and **(2)** blue-shift as pressure below and exceeds 2 GPa, respectively.

The local octahedral geometric parameters of unpressurized MAPbI$_3$ (**Figure 4a**), including $\langle\lambda\rangle$, σ^2, DI, and ECoN, show that the PbI$_6$ octahedra experience considerable distortion (**Figure 4b**). In addition to the shape of MA$^+$ and hydrogen bonding between NH$_3$ group with iodine atoms, the lone pair of nonbonding electrons in Pb^{2+} itself will also lower the coordination symmetry of PbI$_6$ octahedron. However, the lack of consistent correlation between the pressure-driven changes in $\langle\lambda\rangle$ and σ^2 with the

1 electronic bandgap suggests that they do not strongly correlate with the electronic

2 properties. Fortunately, DI and ECoN of local octahedra are sensitive indicators of

3 microstructure modifications. The non-monotonic variation of the DI with increasing

4 pressure, depicted by the standard deviation of Pb-I bond lengths (**Figure S8**), closely

5 mirrors the pressure-induced changes in the electronic bandgap. Under mild pressure

6 (< 2 GPa), the enhancement of local O_h symmetry in PbI_6 octahedra (**Figure 4a**)

7 strengthens the *s-p* antibonding coupling between the lone-pair Pb-$6s^2$ with I-$5p$ orbitals,

8 as well as the unoccupied Pb-$6p$ orbitals coupling[39], leading to a narrowing of electronic

9 bandgap. However, as the pressure increases beyond 2 GPa, the local O_h symmetry

10 deteriorates (**Figure 4b**), resulting in the breaking of some Pb-I bond and weakening of

11 the *s-p* antibonding coupling[24, 40] and causing an electronic bandgap widening.

12 Moreover, **Figure 4b** shows that the pressure-induced lattice modifications primarily

13 occur along the *ac*-lattice plane.
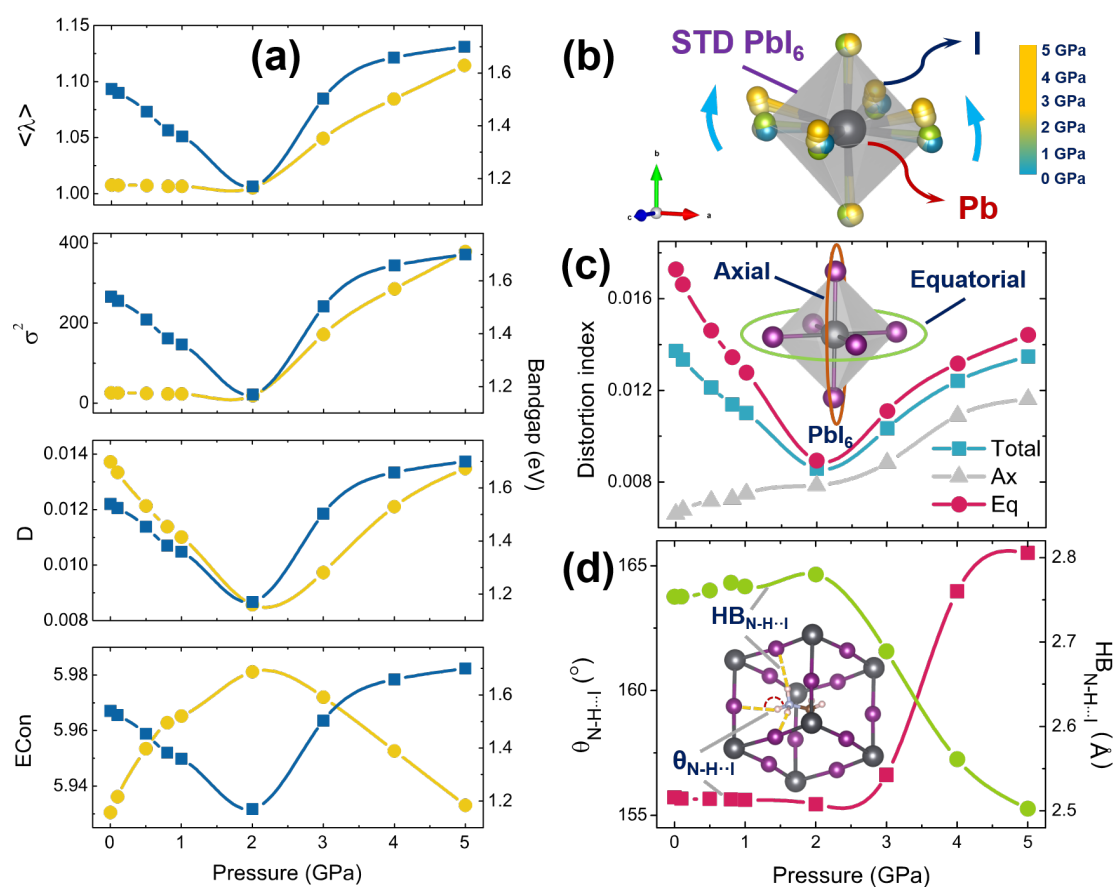
15 **Figure. 4 (a)** Variation of <λ>, σ^2, DI, and ECoN (yellow circles) as compared to bandgap (blue

16 square) under pressure. **(b)** Shape of PbI_6 octahedra under pressure, with the standard PbI_6 octahedra

(green) serving as a benchmark. **(c)** Variation of total DI, along with axial (DI_{ax}) and equatorial (DI_{eq}) components of PbI_6 octahedra, and **(d)** average hydrogen bond length ($HB_{N-H\cdots I}$, green circle) and hydrogen bond angle ($\theta_{N-H\cdots I}$, red square) as a function of pressure, respectively.

We define the DI contribution from the PbI_6 octahedron in the equatorial direction (*ac*-lattice plane) as DI_{eq}, while the contribution along the axial direction (*b*-axis) as DI_{ax}. The matching trends of the overall DI and DI_{eq} under pressure further confirm that the distortion of local octahedron is primarily driven by lattice structure changes along the *ac*-lattice plane. Under moderate pressure (< 2 GPa), the external force equalizes Pb-I bond lengths in the equatorial direction ($Pb-I_{eq}$), significantly reducing DI and, especially DI_{eq}. Meanwhile, stretching or compression of Pb-I bond lengths in the axial direction ($Pb-I_{ax}$) lead to a noticeable reduction in DI (**Figure 4c**). As pressure increase exceeds 2 GPa, the Pb atom deviates from the center of PbI_6 octahedron, causing a substantial distortion in all Pb-I bond lengths and resulting in a significant increase in both DI_{eq} and DI_{ax}. The uneven variation of local octahedral distortion with increasing pressure is closely linked to the interaction between the MA^+ and the PbI_6 inorganic framework through hydrogen bonding, as shown in the insert of **Figure 4d**. Among the hydrogen bonds in the Pb-I cage ($HB_{N-H\cdots I}$ < 3 Å), while each iodine atom along the axial direction forms a hydrogen bond with NH_3 group in MA^+, only some iodine atoms along the equatorial direction participates in hydrogen bond formation. Consequently, this hydrogen bonding inhibits the distortion along the axial direction (DI_{ax}) but exacerbates it along the equatorial direction (DI_{eq}). The slight elongation of $HB_{N-H\cdots I}$ (used as a proxy for hydrogen bond strength) and the stable hydrogen bond angle ($\theta_{N-H\cdots I}$) under pressure below 2 GPa suggest a weakening interaction between MA^+ and the PbI_6 inorganic framework, contributing to a reduction in DI and enhancement of local octahedral symmetry, consistent with previous study[41]. Beyond 2 GPa, the distortion and shrinkage of Pb-I cage lead to a sharp decrease in $HB_{N-H\cdots I}$ and a corresponding increase in $\theta_{N-H\cdots I}$. Hydrogen bonding between MA^+ and the PbI_6 inorganic framework plays a crucial role not only in structural stabilization but also in influencing the octahedral distortion of the inorganic framework.

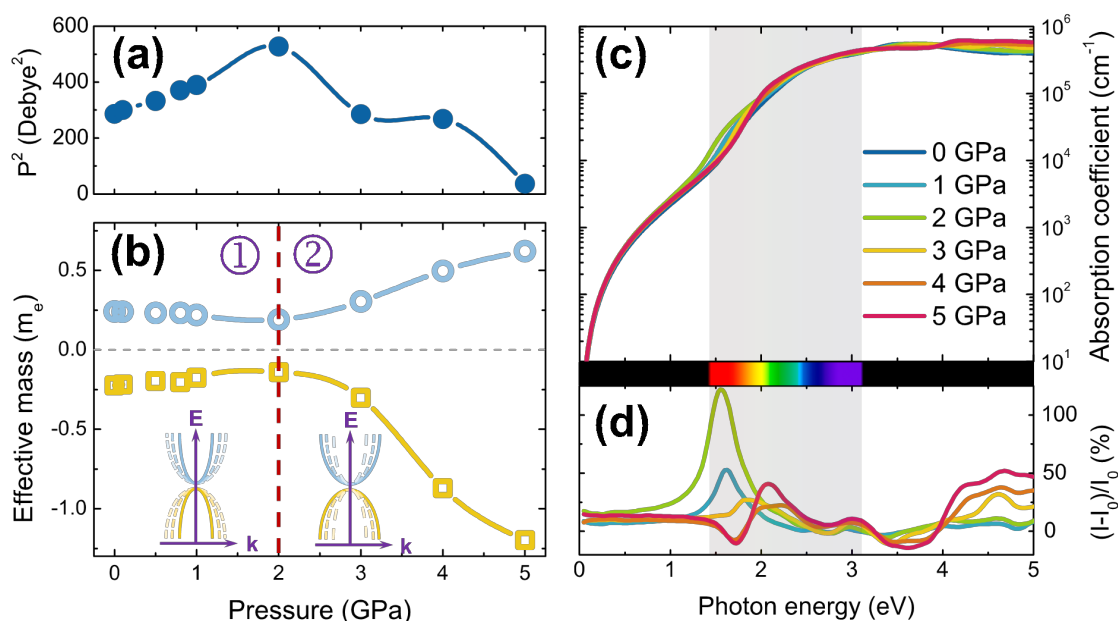## 3.3 Influence of Local Octahedral Distortion on Optoelectronic Properties

Figure. 5 Variation of **(a)** dipole matrix element at the R point and **(b)** effective mass of electrons (blue) and holes (yellow) under pressure. **(c)** Variation of linear absorption spectra, and **(d)** corresponding differential absorption spectrum ($(I-I_0)/I_0$, percentage) as function of pressure. I_0 is linear absorption spectra under 0 GPa pressure.

In addition to affecting the electronic bandgap, the improvement in local symmetry of PbI$_6$ octahedra has a significant impact on the optical transitions. According to the Fermi Golden rule[42], the optical absorption of a semiconductor at photonic energy $\hbar\omega$ is directly correlated with $\frac{2\pi}{\hbar}\int|\langle v|\widehat{H}|c\rangle|^2\frac{2}{8\pi^3}\delta\big(E_c(\vec{k})-E_v(\vec{k})-\hbar\omega\big)d^3k$, where $\langle v|\widehat{H}|c\rangle$ is the transition dipole matrix between the valance band (VB) and conduction band (CB), and the second term corresponds to the joint density of states (JDOS) at $\hbar\omega$. The transition dipole matrix at the R point increases significantly under pressure below 2 GPa (**Figure 5a** and **S5-S7**), with the transition dipole moment (P^2) at 2 GPa reaching more than twice that at 0 GPa, highlighting the importance of band edges in photon-electron interaction. Strong optical transitions at the band edge are known to be dominated by *p-p* transitions[39], as the lower CBs primarily originate from unoccupied Pb-*6p* orbitals, while the upper VBs are dominated by hybrid states of I-*5p* orbitals and Pb-*6s* orbitals (**Figure 2a**). Enhanced local O_h symmetry strengthens the antibonding coupling between Pb-6*s* and I-5*p* orbitals and promotes the *p-p* transitions, leading to

an increase in the transition dipole moment (**Figure 5a**). However, the deterioration of local O_h symmetry (>2 GPa) weakens the antibonding coupling, resulting in an indirect bandgap and a sharp decrease in the transition dipole moment.

Additionally, the improvement of local octahedral symmetry enhances the dispersion of CBM and VBM (**Figure 5b**), which slightly reduces the effective masses for electron (m_e^*) and hole (m_h^*), resulting in the improvement of carrier mobility. Conversely, as pressure increases (>2 GPa), the breaking of local symmetry weakens band dispersion, leading to heavier effective masses for electrons and holes. Furthermore, the pressure-induced modifications in linear absorption spectra primarily occur in the visible light range, which accounts for the most usable portion of the full solar spectrum (**Figure 5c-d**). The enhancement of transition dipole matrix increases optical absorption in the visible light range (< 2 GPa), while the reduction in bandgap causes a significant red-shift in the differential absorption spectrum. However, as pressure exceeds 2 GPa, a broad indirect bandgap and a weaker transition dipole matrix leads to an obvious blue-shift, or even a negative peak in the differential absorption spectrum (**Figure S7**). Consequently, under mild pressure (< 2 GPa), the reduction in octahedral distortion promotes the *s-p* orbital overlaps and *p-p* transitions, narrowing the bandgap and enhancing band dispersion, thus increasing visible light absorption. When pressure exceeds 2 GPa, the weakening of *s-p* orbital overlap due to increased octahedral distortion causes a decline in visible light absorption.

While the pressure-driven nonmonotonic variation of the electronic bandgap in cubic MAPbI$_3$ has been well explained, and the structure-property relationships between local octahedral distortion and optoelectronic properties have been established, further verification is needed to confirm its universal applicability. The low-temperature tetragonal and orthorhombic MAPbI$_3$, which also feature the PbI$_6$ inorganic framework with voids occupied by MA$^+$, serve as suitable candidates for such verification. The calculated lattice parameters (**Table S4**) and electronic bandgap for the tetragonal (**Figure S9**) and orthorhombic (**Figure S10**) MAPbI$_3$ are in good agreement with previous studies[41, 43-45]. The orbital-projected band structure and projected density of states (PDOS) for these phases under various pressures are displayed in **Figure S11-**

S14. Both VBM and CBM of tetragonal (**Figure S9**) and orthorhombic (**Figure S10**)

phases MAPbI$_3$ are similarly composed by a hybrid state between Pb-$6s$ and I-$5p$ orbital

coupling and Pb-$6p$ orbitals, respectively. Furthermore, the crystal lattice free volume

in both tetragonal and orthorhombic phases MAPbI$_3$ decreases significantly, making a

major contribution to the reduction in lattice volume (**Figure S15-S16**).

3.4 Robustness of Local Octahedral Geometric Fingerprints across Multiple

Phases
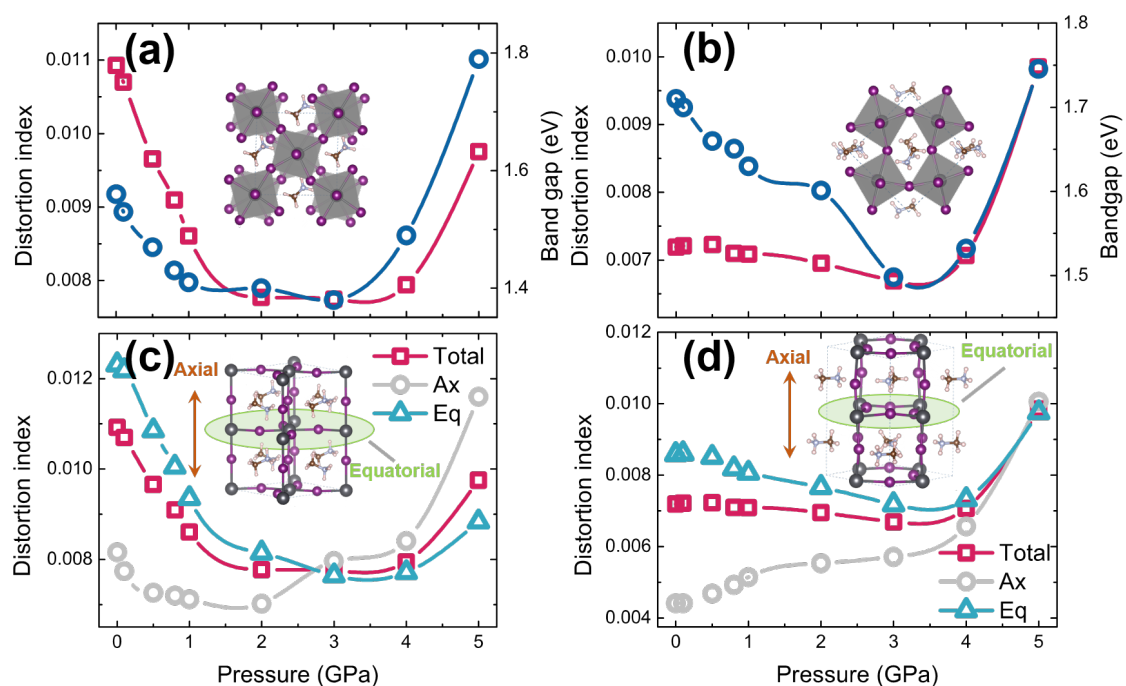
Figure. 6 (a) Variation of the electronic bandgap (blue circle) and DI (red square) of **(a)** tetragonal

and **(b)** orthorhombic MAPbI$_3$ under pressure, respectively. Total, axial (DI$_{ax}$), and equatorial (DI$_{eq}$)

distortion index of **(c)** tetragonal and **(d)** orthorhombic MAPbI$_3$, as function of pressure, respectively.

Interestingly, both tetragonal and orthorhombic phases MAPbI$_3$ exhibit a

nonmonotonic variation of the bandgap with external pressure, as shown in **Figure 6a-**

b. Unlike the monotonic variation of Pb-I bond length, Pb-I-Pb bond angle, and $\langle\lambda\rangle$

and σ^2 with pressure (**Figure S17-S18**), the nonmonotonic changes in DI (**Figure 6a-**

b) and ECoN (**Figure S19-S20**) for the PbI$_6$ octahedra closely align with the pressure-

driven evolution of the electronic bandgap. The intrinsic local distortion of the PbI$_6$

octahedra causes uneven Pb-I bond lengths, resulting in a wide bandgap. Mild external

pressure (< 3 GPa) improves the local octahedral symmetry in both tetragonal and

orthorhombic MAPbI$_3$, as reflected by the decrease in DI (**Figure 6a-b**), leading to a narrowing of the electronic bandgap. However, as the local octahedral distortion increases under further compression (> 3 GPa), the electronic bandgap increases rapidly. The lower the local octahedral distortion, the smaller the electronic bandgap. Moreover, as shown in **Figure S21-S22**, the reduction in local octahedral distortion under pressure in both phases MAPbI$_3$ enhances the optical absorption in the visible light range.

Similar to cubic MAPbI$_3$, as shown in **Figure 6c-d**, variation of local octahedral distortion in the tetragonal and orthorhombic phases exhibit strong anisotropy under pressure. Specifically, DI$_{ax}$ in both phases follows a monotonic or quasi-monotonic trend with pressure, while the nonmonotonic pressure-driven changes in DI$_{eq}$ correspond to the overall DI trend. Mild external pressure (< 3 GPa) equalizes Pb-I bond lengths along the equatorial direction (Pb-I$_{eq}$), significantly reducing the total and equatorial octahedral distortion (DI and DI$_{eq}$), and narrowing the electronic bandgap. However, pressure-induced deviations of Pb-I bond lengths along the axial direction (Pb-I$_{ax}$) increase the axial distortion (DI$_{ax}$). As Pb atom shifts from the center of PbI$_6$ octahedron (indicated by σ^2) under further compression, both DI$_{eq}$ and DI$_{ax}$ rise sharply, and the resulting breakdown in local octahedral symmetry leads to the electronic bandgap widening. Therefore, irrespective of crystalline phase, the local distortion of PbI$_6$ octahedra, characterized by deviations in Pb-I bond lengths, DI and ECoN, plays a pivotal role in determining the optoelectronic properties of hybrid halide perovskites.

4. Conclusions

To summarize, utilizing physical-informed data-driven machine learning techniques, we have evaluated the impact of lattice geometric fingerprints on the electronic properties in hybrid halide perovskites, extracting robust geometrical fingerprints to establish a robust structure-property relationship mapping the electronic bandgap and improve the performance of machine learning models dealing with small and imbalanced datasets. In addition to atomic and physicochemical properties descriptors like halide electronegativity and TF, feature engineering identified key octahedral geometric fingerprints, such as the DI and ECoN, to effectively capture

complex trends and correlations within chemical space. Furthermore, lattice compression simulations across multiple phases of MAPbI$_3$ (cubic, tetragonal, and orthorhombic) demonstrate that the pressure-driven non-monotonic variations of DI and ECoN closely aligns with the electronic bandgap, validating the effectiveness of the selected local octahedral geometric fingerprints and confirming the reliability of ML's interpretability and explainability. By adjusting the *s-p* antibonding coupling, the pressure-driven reduction in local octahedral distortion (enhanced local O_h symmetry), induced by the anisotropic hydrogen bonding between the inorganic framework and organic cation, narrows the electronic bandgap and facilitates the *p-p* transitions, thereby boosting the transition dipole moment, carrier mobility, and band-edge absorption. Combining data mining with physical analysis, our theoretical investigations has thoroughly examined the significant impact of lattice geometry on electronic properties, successfully identified effective octahedral geometric fingerprints for describing the electronic bandgap, while also revealed the microphysical mechanisms of local octahedral distortion on the optoelectronic properties of hybrid halide perovskites.

Acknowledgements

References

1. Liu, C.; Yang, Y.; Chen, H.; Xu, J.; Liu, A.; Bati, A. S. R.; Zhu, H.; Grater, L.; Hadke, S. S.; Huang, C.; Sangwan, V. K.; Cai, T.; Shin, D.; Chen, L. X.; Hersam, M. C.; Mirkin, C. A.; Chen, B.; Kanatzidis, M. G.; Sargent, E. H., Bimolecularly Passivated Interface Enables Efficient and Stable Inverted Perovskite Solar Cells. *Science* **2023**, *382*, 810-815.
2. Liu, C.; Yuan, J.; Masse, R.; Jia, X.; Bi, W.; Neale, Z.; Shen, T.; Xu, M.; Tian, M.; Zheng, J.; Tian, J.; Cao, G., Interphases, Interfaces, and Surfaces of Active Materials in Rechargeable Batteries and Perovskite Solar Cells. *Adv. Mater.* **2021**, *33*, e1905245.
3. Zhou, Z.; Qiao, H. W.; Hou, Y.; Yang, H. G.; Yang, S., Epitaxial Halide Perovskite-Based Materials for Photoelectric Energy Conversion. *Energy & Environ. Sci.* **2021**, *14* (1), 127-157.
4. Xia, J.; Sohail, M.; Nazeeruddin, M. K., Efficient and Stable Perovskite Solar Cells

by Tailoring of Interfaces. *Adv. Mater.* **2023**, *35*, e2211324.

5. Kim, J. Y.; Lee, J. W.; Jung, H. S.; Shin, H.; Park, N. G., High-Efficiency Perovskite Solar Cells. *Chem. Rev.* **2020**, *120* (15), 7867-7918.

6. Zhan, X.; Chen, X.; Li, C.; Jin, T.; Wang, Y.; Chen, Z. N.; Wu, T.; Chen, J.; Zhuang, W., Can Lead-Free Double Halide Perovskites Serve as Proper Photovoltaic Absorber? *J. Phys. Chem. Lett.* **2023**, 10784-10793.

7. Zhang, H.; Pfeifer, L.; Zakeeruddin, S. M.; Chu, J.; Grätzel, M., Tailoring Passivators for Highly Efficient and Stable Perovskite Solar Cells. *Nat. Rev. Chem.* **2023**, *7*, 632-652.

8. Dai, Z.; Padture, N. P., Challenges and Opportunities for the Mechanical Reliability of Metal Halide Perovskites and Photovoltaics. *Nat. Energy* **2023**, *8*, 1319-1327.

9. Nie, T.; Fang, Z.; Ren, X.; Duan, Y.; Liu, S., Recent Advances in Wide-Bandgap Organic-Inorganic Halide Perovskite Solar Cells and Tandem Application. *Nano-Micro Lett.* **2023,** *15*, 70.

10. Huang, J.; Yuan, Y.; Shao, Y.; Yan, Y., Understanding the Physical Properties of Hybrid Perovskites for Photovoltaic Applications. *Nat. Rev. Mater.* **2017**, *2* (7), 17042.

11. Li, W.; Wang, Z.; Deschler, F.; Gao, S.; Friend, R. H.; Cheetham, A. K., Chemically Diverse and Multifunctional Hybrid Organic-Inorganic Perovskites. *Nat. Rev. Mater.* **2017**, *2*, 16099.

12. Kojima, A.; Teshima, K.; Shirai, Y.; Miyasaka, T., Organometal Halide Perovskites as Visible-Light Sensitizers for Photovoltaic Cells. *J. Am. Chem. Soc.* **2009**, *131*, 6050-6051.

13. NREL, Best Research-Cell Efficiencies, https://www.nrel.gov/pv/cell-efficiency.html (accessed: Sep 23 2024).

14. Meng, H.; Mao, K.; Cai, F.; Zhang, K.; Yuan, S.; Li, T.; Cao, F.; Su, Z.; Zhu, Z.; Feng, X.; Peng, W.; Xu, J.; Gao, Y.; Chen, W.; Xiao, C.; Wu, X.; McGehee, M. D.; Xu, J., Inhibition of Halide Oxidation and Deprotonation of Organic Cations with Dimethylammonium Formate for Air-Processed p-i-n Perovskite Solar Cells. *Nat. Energy* **2024**, *9*, 536-547.

15. Li, C.; Chen, X.; Jin, T.; Wu, T.; Chen, J.; Zhuang, W., Impact of Functional Groups in Spacer Cations on the Properties of PEA-based 2D Monolayer Halide Perovskites. *Nano Mater. Sci.* **2024**. DOI: https://doi.org/10.1016/j.nanoms.2024.02.005.

16. De Angelis, F., The Impact of Machine Learning in Energy Materials Research: The Case of Halide Perovskites. *ACS Energy Lett.* **2023**, *8*, 1270-1272.

17. Lu, S.; Zhou, Q.; Ouyang, Y.; Guo, Y.; Li, Q.; Wang, J., Accelerated Discovery of Stable Lead-free Hybrid Organic-Inorganic Perovskites via Machine Learning. *Nat. Commun* **2018**, *9*, 3405.

18. Wu, T.; Wang, J., Global Discovery of Stable and Non-Toxic Hybrid Organic-Inorganic Perovskites for Photovoltaic Systems by Combining Machine Learning Method with First Principle Calculations. *Nano Energy* **2019**, *66*, 104070.

19. Wu, T.; Wang, J., Deep Mining Stable and Nontoxic Hybrid Organic-Inorganic Perovskites for Photovoltaics via Progressive Machine Learning. *ACS Appl. Mater. Inter.* **2020**, *12*, 57821-57831.

20. Chen, X.; Lu, S.; Chen, Q.; Zhou, Q.; Wang, J., From Bulk Effective Mass to 2D

Carrier Mobility Accurate Prediction via Adversarial Transfer Learning. *Nat. Commun.* **2024**, *15*, 5391.

21. Im, J.; Lee, S.; Ko, T.-W.; Kim, H. W.; Hyon, Y.; Chang, H., Identifying Pb-free Perovskites for Solar cells by Machine Learning. *npj Comput. Mater.* **2019**, *5*, 37.

22. Anand, D. V.; Xu, Q.; Wee, J.; Xia, K.; Sum, T. C., Topological Feature Engineering for Machine Learning Based Halide Perovskite Materials Design. *npj Comput. Mater.* **2022**, *8*, 203.

23. Wu, Y.; Wang, C. F.; Ju, M. G.; Jia, Q.; Zhou, Q.; Lu, S.; Gao, X.; Zhang, Y.; Wang, J., Universal Machine Learning Aided Synthesis Approach of Two-Dimensional Perovskites in a Typical Laboratory. *Nat. Commun.* **2024**, *15*, 138.

24. Kong, L.; Liu, G.; Gong, J.; Hu, Q.; Schaller, R. D.; Dera, P.; Zhang, D.; Liu, Z.; Yang, W.; Zhu, K.; Tang, Y.; Wang, C.; Wei, S. H.; Xu, T.; Mao, H. K., Simultaneous Band-Gap Narrowing and Carrier-Lifetime Prolongation of Organic-Inorganic Trihalide Perovskites. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 8910-8915.

25. Ouyang, R.; Curtarolo, S.; Ahmetcik, E.; Scheffler, M.; Ghiringhelli, L. M., SISSO: A Compressed-Sensing Method for Identifying the Best Low-Dimensional Descriptor in an Immensity of Offered Candidates. *Phys. Rev. Mater.* **2018**, *2*, 083802.

26. Kresse, G.; Furthmuller, J., Efficiency of ab-Initio Total Energy Calculations for Metals and Semiconductors using a Plane-Wave Basis Set. *Comput. Mater. Sci.* **1996**, *6*, 15-50.

27. Kresse, G.; Hafner, J., Abinitio Molecular-Dynamics for Liquid-Metals. *Phys. Rev. B* **1993**, *47*, 558-561.

28. Kresse, G.; Furthmuller, J., Efficient Iterative Schemes for ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B* **1996**, *54*, 11169-11186.

29. Blochl, P. E., Projector Augmented-Wave Method. *Phys. Rev. B* **1994**, *50*, 17953-17979.

30. Perdew, J. P.; Burke, K.; Ernzerhof, M., Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865-3868.

31. Lee, K.; Murray, E. D.; Kong, L. Z.; Lundqvist, B. I.; Langreth, D. C., Higher-Accuracy van der Waals Density Functional. *Phys. Rev. B* **2010**, *82*, 4.

32. Lu, T.; Chen, F., Multiwfn: a Multifunctional Wavefunction Analyzer. *J. Comput. Chem.* **2012**, *33*, 580-592.

33. Wang, V.; Xu, N.; Liu, J.-C.; Tang, G.; Geng, W.-T., VASPKIT: A User-Friendly Interface Facilitating High-Throughput Computing and Analysis Using VASP code. *Comput. Phys. Commun.* **2021**, *267*, 108033.

34. Liu, G.; Kong, L.; Yang, W.; Mao, H.-k., Pressure Engineering of Photovoltaic Perovskites. *Mater. Today* **2019**, *27*, 91-106.

35. Li, X.; Fu, Y.; Pedesseau, L.; Guo, P.; Cuthriell, S.; Hadar, I.; Even, J.; Katan, C.; Stoumpos, C. C.; Schaller, R. D.; Harel, E.; Kanatzidis, M. G., Negative Pressure Engineering with Large Cage Cations in 2D Halide Perovskites Causes Lattice Softening. *J. Am. Chem. Soc.* **2020**, *142*, 11486-11496.

36. Lee, J.-H.; Jaffe, A.; Lin, Y.; Karunadasa, H. I.; Neaton, J. B., Origins of the Pressure-Induced Phase Transition and Metallization in the Halide Perovskite $(CH_3NH_3)PbI_3$. *ACS Energy Lett.* **2020**, *5*, 2174-2181.

37. Fu, Y.; Jiang, X.; Li, X.; Traore, B.; Spanopoulos, I.; Katan, C.; Even, J.; Kanatzidis, M. G.; Harel, E., Cation Engineering in Two-Dimensional Ruddlesden-Popper Lead Iodide Perovskites with Mixed Large A-Site Cations in the Cages. *J. Am. Chem. Soc.* **2020**, *142*, 4008-4021.

38. Coduri, M.; Strobel, T. A.; Szafranski, M.; Katrusiak, A.; Mahata, A.; Cova, F.; Bonomi, S.; Mosconi, E.; De Angelis, F.; Malavasi, L., Band Gap Engineering in MASnBr$_3$ and CsSnBr$_3$ Perovskites: Mechanistic Insights through the Application of Pressure. *J. Phys. Chem. Lett.* **2019**, *10*, 7398-7405.

39. Yin, W. J.; Shi, T.; Yan, Y., Unique Properties of Halide Perovskites as Possible Origins of the Superior Solar Cell Performance. *Adv. Mater.* **2014**, *26*, 4653-4658.

40. Lu, X.; Yang, W.; Jia, Q.; Xu, H., Pressure-Induced Dramatic Changes in Organic-Inorganic Halide Perovskites. *Chem. Sci.* **2017**, *8*, 6764-6776.

41. Lee, J. H.; Bristowe, N. C.; Bristowe, P. D.; Cheetham, A. K., Role of Hydrogen-Bonding and its Interplay with Octahedral Tilting in CH$_3$NH$_3$PbI$_3$. *Chem. Commun.* **2015**, *51*, 6434-6437.

42. Dresselhaus, S., Lecture Note: Solid State Physics Part II Optical Properties of Solids.

43. Octavio de Araujo, L.; Sabino, F. P.; Rego, C. R. C.; Guedes-Sobrinho, D., Bulk Rashba Effect Splitting and Suppression in Polymorphs of Metal Iodine Perovskites. *J. Phys. Chem. Lett.* **2021**, *12*, 7245-7251.

44. Baikie, T.; Fang, Y.; Kadro, J. M.; Schreyer, M.; Wei, F.; Mhaisalkar, S. G.; Graetzel, M.; White, T. J., Synthesis and Crystal Chemistry of the Hybrid Perovskite (CH$_3$NH$_3$)PbI$_3$ for Solid-State Sensitised Solar Cell Applications. *J. Mater. Chem. A* **2013**, *1*, 5628.

45. Weller, M. T.; Weber, O. J.; Henry, P. F.; Di Pumpo, A. M.; Hansen, T. C., Complete Structure and Cation Orientation in the Perovskite Photovoltaic Methylammonium Lead Iodide between 100 and 352 K. *Chem. Commun.* **2015**, *51*, 4180-4183.