Machine Learning Accelerated Insights of Perovskite Materials

Shuaihua Lu, Yilei Wu, Ming-Gang Ju, and Jinlan Wang

1 Introduction

Conventionally, perovskite as mineral name is applied to the class of materials possessing the same type of crystal structure as $CaTiO_3$, which initially was discovered in 1839 by the Prussian mineralogist Gustav Rose in the Ural Mountains and was named after the Russian mineralogist Count Lev Aleksevich von Perovski. Perovskites have a general formula with or derived from composition ABX_3, which exhibit many fantastic chemical and physical properties and is one of the most intensely studied material in material field. Generally, perovskites are composed with a large cation at A site and an octahedral BX_6. A corner-shared network is formed with the BX_6 octahedras and the cation A is filled in the caves between the octahedras. Nonideal ionic size ratios and electronic instabilities are compensated by tilting and distorting of BX_6 octahedras. Except these general perovskite structures, many perovskite variants also attract widespread attention, such as double perovskite and layered perovskite. Furthermore, substitution of perovskite A, B, or X sites is allowed for tailoring of properties to meet particular application. Due to the structure and composition flexibility, perovskites can vary from insulating to metallicity, with a wide range of possible applications such as electronic device and sensor [1], magnetic memory components [2], and solar cell [3].

Shuaihua Lu and Yilei Wu contributed equally with all other contributors.

S. Lu · Y. Wu · M.-G. Ju (✉) · J. Wang (✉)
School of Physics, Southeast University, Nanjing, China
e-mail: shlu@seu.edu.cn; ylwu@seu.edu.cn; juming@seu.edu.cn; jlwang@seu.edu.cn

Y. Cheng et al. (eds.), *Artificial Intelligence for Materials Science*, Springer Series in Materials Science 312, https://doi.org/10.1007/978-3-030-68310-8_8

In recent decades, lead halide perovskites have made tremendous progress in photovoltaic and optoelectronic field. High visible absorption, long carrier-diffusions lengths, and fantastic defect tolerance have led to solar cells with certified efficiency of 25.5% [4]. However, chemical stability, mechanical reliability, and toxicity still are three critical obstacles in the path of eventual commercialization of the emerging perovskite solar cells. This prompts a research focus in halide perovskites to predict new perovskites with targeted properties, especially those composed of abundant, nontoxic elements and with thermal, chemical, and dynamic stability. The latter objectives have been traditionally met through performing density functional theory (DFT) calculations of electronic properties, optical absorption properties, defect properties, and performing ab initio molecular dynamics simulations for materials at various given compositions [3, 5]. Recently, with advances in the descriptor-based modeling techniques, researchers are able to perform high-throughput (HT) screening to rapidly estimate certain targeted material properties [6, 7]. Moreover, machine learning (ML), a modeling approach that has received growing attention, has been employed to accelerate the discovery of new perovskite materials [8]. In brief, the ML method can unveil hidden physical properties of materials, if given abundant data and a learning rule, thereby mapping between inputs and output data [9]. So far, most ML studies on perovskites have been focused on all-inorganic perovskites, double perovskites, and anti-perovskites, which all possess a particular type of crystal structure. Due to their simpler and particular crystal structures compared to the prevailing hybrid organic-inorganic perovskites (HOIPs), various ML methods with different choices of descriptors have been benchmark tested for predicting new and stable perovskites [10–12]. This is because of the difficulty of representing organic cations in a fixed length vector to be compatible with many ML algorithms. To match these challenges, developing flexible, transferrable, and reasonable representations becomes one of the important areas of research in ML for HOIPs. As an alternative to learning from first-principle computational data, ML techniques are also optimal for predicting targeted properties through training with numerous experimental data, mapping between the high performance of devices and the various physical and chemical origins, such as bandgap, absorption, and defect properties.

Herein, we bring a brief and in-depth review of ML-guided design and discovery of perovskite materials for photovoltaic application, a field where LHPs with superior performance and low cost are promising candidate for Next Gen PVs. Our review begins with a discussion of construction of data sets, alongside the challenges of the various collections of material data sets. The next section will provide a review of the material representations including descriptors and feature engineering. The final section reviews the ML applications in recent studies such as the ML techniques accelerate the discovery and design of new perovskites with desire stabilities and bandgaps and discovery of factors in experimental processes, which are significantly related to performance of devices.

2 Learning with Perovskite Databases

The cornerstone of ML material discovery is high-quality material data set, and enough material data will ensure the performance of ML models. For perovskite-based photovoltaic materials, abundant data have been generated through the high-throughput calculations and experiments. Besides, some databases containing the properties of perovskites also provide considerable data. We will discuss these three data sources in detail.

In recent years, due to the continuing development of computing power, HT computational material discovery strategy has become an effective and efficient way to discover new functional materials, especially perovskite materials. Among them, tens of thousands of new perovskite-based materials have been predicted for photovoltaic applications. The HT computational method uses the first-principle calculations to build a large-scale material database, which includes existing and hypothetical materials. To facilitate such large-scale computational tasks and data analysis, a number of well-developed software frameworks are developed, including AFLOW [13], pymatgen [14], the Atomic Simulation Environment [15], MatCloud [16], and so on.

The material properties directly determine the applications of materials [17]. As shown in Fig. 1, for the design of perovskite-based photovoltaic materials, evaluating the stability of perovskites is the first step, which is also one of the challenges restricting the practical application of perovskites. The stability of perovskite is mainly evaluated by three different aspects: (1) structural stability (or formability), (2) thermodynamic stability, and (3) dynamic stability. The formability of perovskite is mainly judged by simple structure descriptors, which will be described in detail in the material representation sect. 3.1. In general, the formation energy ΔH_f and the energy above convex hull E_{hull} are utilized to evaluate the thermodynamic stability of perovskites. The formation energy ΔH_f describes the energy change of a material from an elemental component to a compound, and negative values indicate stable compounds. The energy above convex hull E_{hull} describes whether a compound tends to decompose into various elemental, binary, ternary, or more complex components, while negative values indicate unstable compounds. Thermal and dynamic stability represents a more realistic evaluation of material stability in the operating environment. Computationally, phonon calculations are main methods to assess the dynamic stability of materials and ab initio molecular dynamics is adopted to estimate thermal stability. Due to the complexity and time-consuming of these calculations, it is usually performed only for selected promising candidates in HT screening processes. Secondly, the optical and electronic properties determine the applications of perovskites. In HT calculations, the bandgap is one of the most commonly physical parameters to evaluate the photovoltaic performance of a material, because it directly affects the photovoltaic performance of perovskite materials. The effective masses of electron and hole are directly related to the mobility of the material. The small and balanced effective mass is beneficial for carrier mobility in the solar cell materials [18].

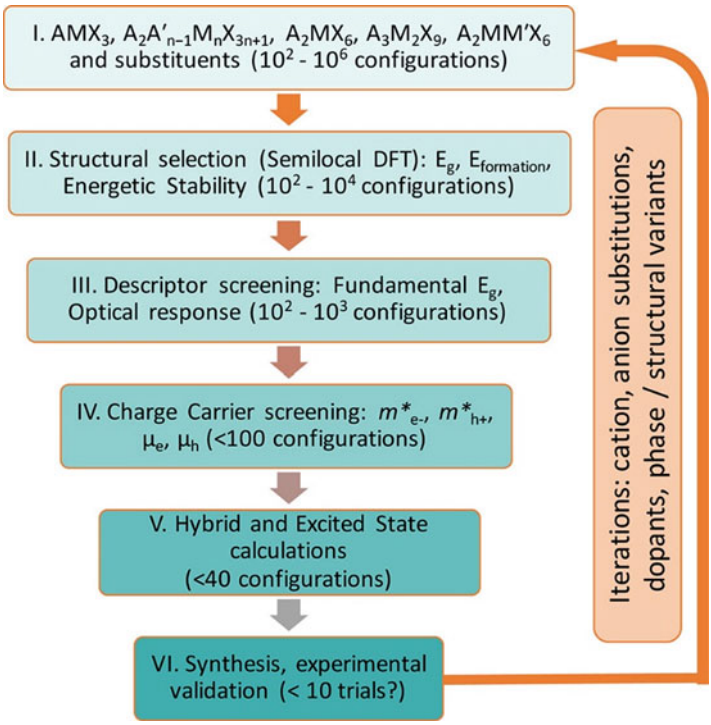

Fig. 1 Rational flowchart for perovskites discovery using HT calculations. (Reprinted with permission from Ref. 17. Copyright © 2017 American Chemical Society)

Inorganic perovskites have firstly been screened by HT first-principle calculations because of the simple crystal structures and abundant candidate materials. In 2012, Castelli et al. [23, 24] explored the bandgap of around 19,000 compounds (perovskite oxides with one or more replacements for oxygen neighbors in the periodic table) using sophisticated semi-local functional called GLLB-SC. Korbel et al. [19] extensively studied the stability and electronic properties of the possible ABX_3 perovskites, where X is a nonmetal and A and B cover a large part of the periodic table. One hundred and ninety-nine perovskites were screened out from more than 32,000 compounds after thermodynamic stability evaluation, and the selected perovskites were characterized by calculating a variety of electronic properties, such as electronic bandgap, average hole effective mass, and so on. Emery and Wolverton [20] presented an exhaustive dataset of 5329 cubic and distorted inorganic perovskites in terms of formation energies, bandgap, and some other properties, which were calculated using density functional theory (the calculation workflow is shown in Fig. 2a).

In addition to the simple inorganic perovskite materials with formula ABX_3, inorganic double perovskite materials have also received significant attention due to the phase space of possible compounds is substantially larger, which increases

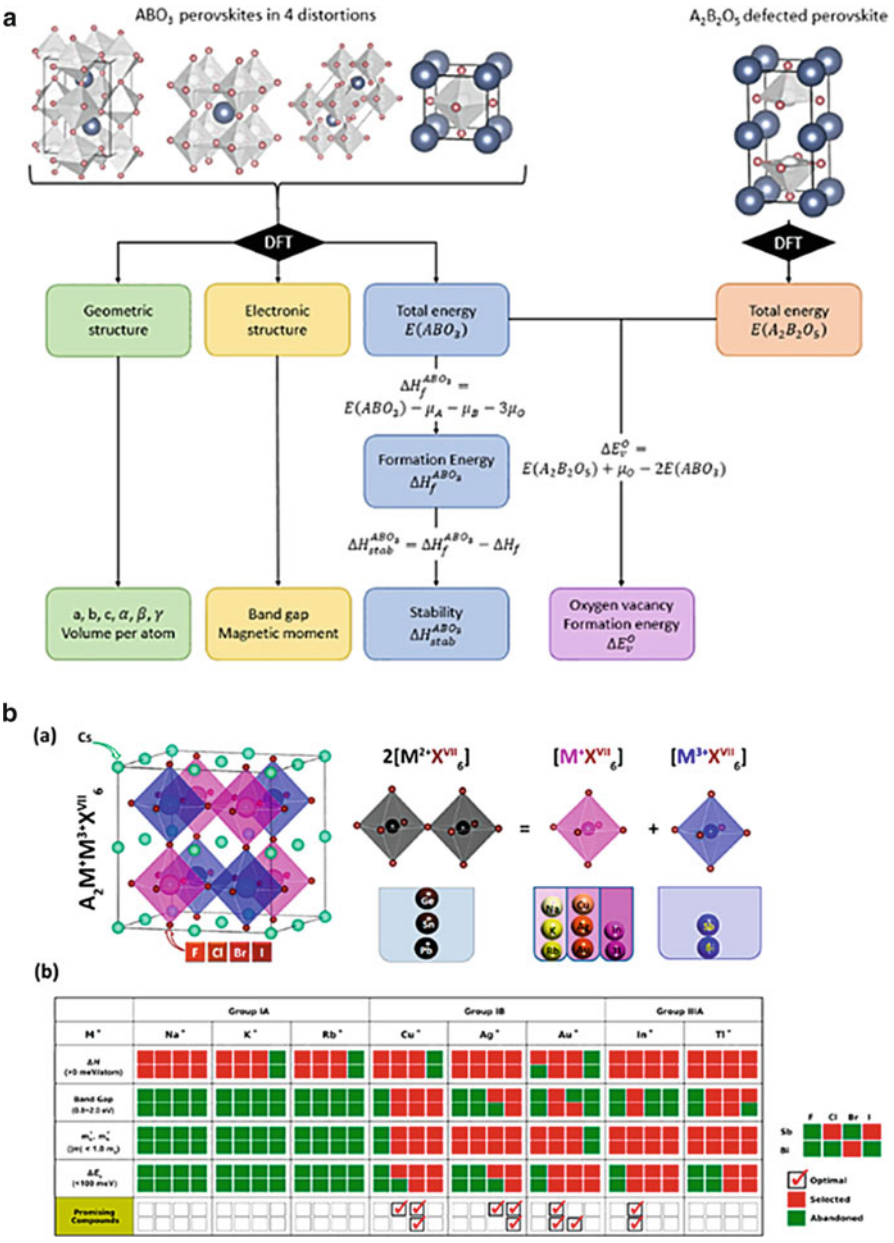

Fig. 2 (**a**) Workflow to calculate all the properties in the dataset. (Reprinted from Ref. 20). (**b**) Space of candidate perovskites for materials screening and materials screening process by considering gradually the properties relevant to photovoltaic performance. (Reprinted with permission from Ref. 21. Copyright 2017 American Chemical Society)

considerably the probability of finding promising candidates with the desired properties [21, 22, 25–27]. As shown in Fig. 2b, Zhao et al. [21] constructed a rich class of double perovskites without Pb^{2+} ions to solve the toxicity of perovskites. After gradually considering the properties relevant to photovoltaic performance, i.e., decomposition enthalpy, bandgap, carrier effective masses, and exciton binding energy, 11 optimal materials were identified as candidates in photovoltaic field. Subsequently, Cai et al. [22] computed structural, electronic, and transport properties of around 1000 double perovskite halides using high-throughput first-principles calculations to aid the discovery of photovoltaic materials (see Fig. 2d).

Compared with inorganic perovskite, hybrid organic-inorganic perovskites (HOIPs), as one of the most promising photovoltaic materials, have attracted tremendous interest recently. The most distinguished virtues of HOIPs include high power conversion efficiency (PCE), low-cost experimental synthesis, and tunable bandgaps. In order to find more stable hybrid perovskites with higher PCE, a lot of HT computing works have emerged [17, 28–33]. An HT computational screening study [28] for 11,025 compositions of HOIP compounds in ABX_3 and $A_2B'B''X_6$ forms has been reported, where A is an organic or inorganic component, B'/B'' is a metal cation, and X is a halogen anion. The computational results contain bandgap values at the scalar relativistic PBE level of all compositions. Besides, the hole and electron effective masses of 1923 candidate semiconductors with bandgaps smaller than 3.5 eV were also estimated. Another effort on computational screening of possible replacements for methylammonium or lead was shown in Fig. 3a, in which 11 different molecular organic cations and 29 different divalent cations were considered [29]. All thermodynamically stable hybrid perovskites were then further characterized by their bandgaps and effective masses. Moreover, Jacobs et al. [34] focused on finding materials that comprise nontoxic elements, stable in a humid operating environment, and have an optimal bandgap for single junction. From a set of 1845 materials, 15 materials passed all screening criteria for single junction cell applications. Notably, these efforts primarily focused on the single perovskite or double perovskite structure. Besides perovskite structures, there exist in principles other organic-inorganic hybrid ternary metal halide compounds with appropriate metal elements and the stoichiometry of component elements that are more stable and even show better optoelectronic properties than the typical perovskite structures. Li and Yang [30] carried out HT calculations on 4507 hypothetical compounds. The chemical formulas of selected candidates include A_2BX_4, $A_3B_2X_9$, and A_2BX_6, in which A = MA (CH_3NH_3), FA ($CH(NH_2)_2$), AD ($(CH_2)_2NH_2$), and X = Cl, Br, or I. As shown in Fig. 3b, the bandgap and electron/hole effective masses of all these candidates were calculated and used to screen appropriate candidates, thereby the formation enthalpy and decomposition enthalpy of those were computed to evaluate the stabilities.

HT calculations have produced considerable data on perovskite materials, especially thermodynamic stability and electronic properties, while theoretical predicted materials are often difficult to experimental synthesis. The fabrication process based on non-vacuum solution has obvious advantages, such as being suitable for scale-

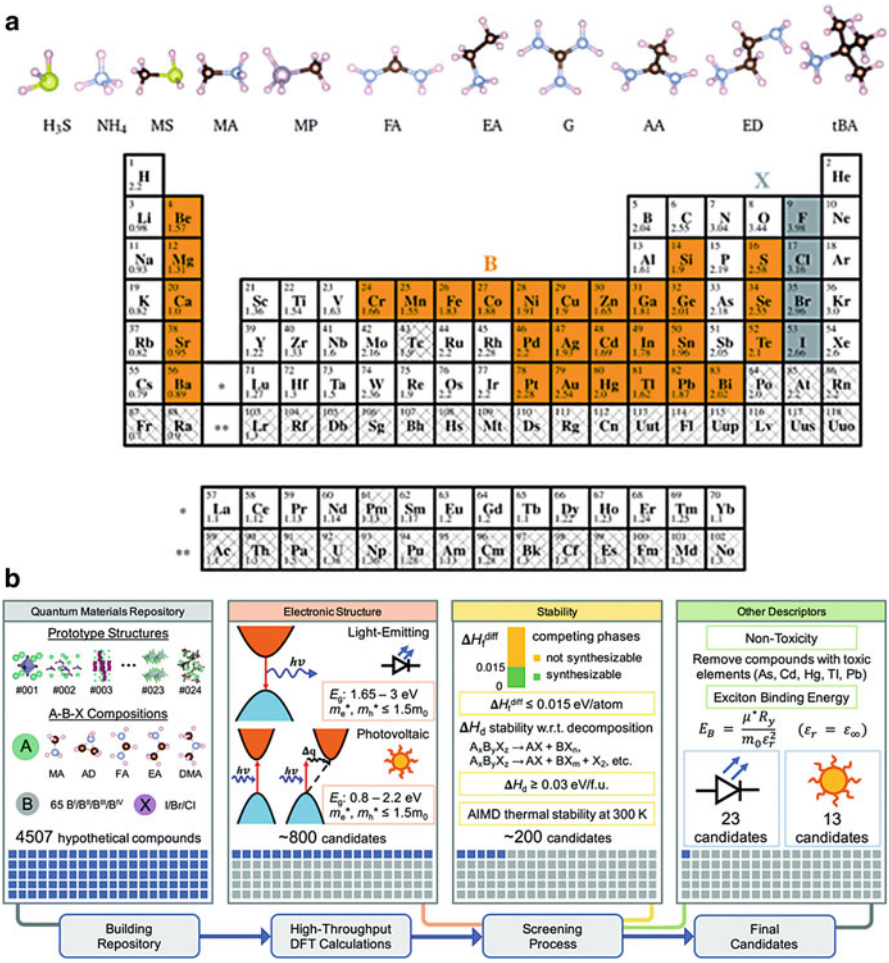

Fig. 3 (**a**) Molecular cations and periodic system of the elements considered for candidate perovskites. (Reprinted with permission from Ref. 29. Copyright 2018 The Royal Society of Chemistry). (**b**) Schematic diagram of the HT screening process for a total number of 4507 compounds, which were generated from 24 different crystal structures. (Reprinted with permission from Ref. 30. Copyright 2019 The Royal Society of Chemistry)

up production, lowering process temperature, lowering energy consumption, and lowering costs, thereby receiving increasing attention in the photovoltaic field. In addition, solvent-based methods can be implemented flexibly in automated HT experimentation, allowing rapid screening of perovskites [37–45]. Chen et al. [46] built an automatic HT experimentation platform for synthesis and characterization of HOIPs with suitable wide bandgap. This platform automatically and efficiently synthesized 95 perovskite polycrystalline samples derived from binary mixtures of five common perovskite precursors and then measured the

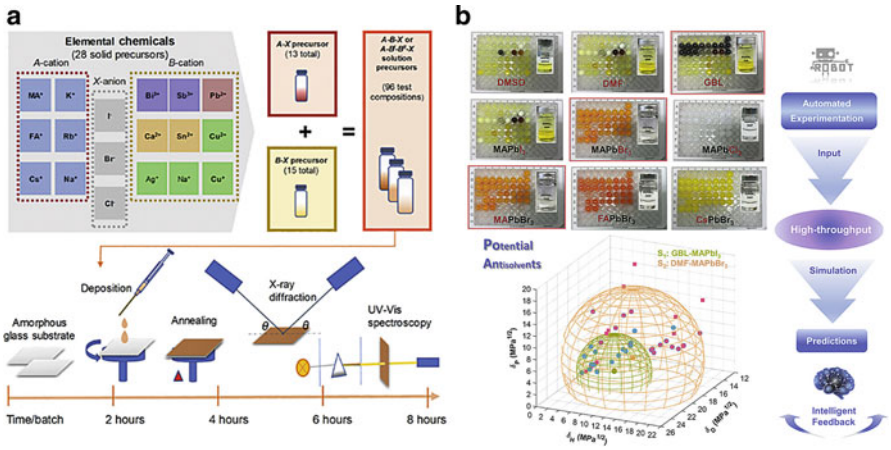

Fig. 4 (**a**) Sketch of the optimized experimental workflow. (Reprinted with permission from Ref. 35. Copyright 2019 Elsevier Inc.). (**b**) A robotic platform was adopted to conduct a comprehensive solvent engineering for making lead halide perovskites in a high-throughput manner. Deeper insights into the working mechanisms and selection criteria of antisolvents were investigated and summarized. (Reprinted with permission from Ref. 36. Copyright 2020 Elsevier Inc.)

corresponding photoluminescence and absorption, yielding six compositing perovskite sample with an optical bandgap of ≈1.75 eV. Apart from exploring stable HOIPs with wide bandgap, the discovery of new perovskite compounds has also been attempted with HT experimentations. Figure 4a illustrates the sketch of the optimized experimental workflow, which enables the realization of rapid search for new lead-free perovskites in the multi-parameter chemical space [35]. Moreover, a self-assembled semi-automated platform based on a standard pipetting robot was utilized to screen the efficient antisolvents for different solvent-perovskite systems and study the influence of interactions among the solvent molecules, cations, metal-halides, and antisolvents (Fig. 4b). In this work, 336 combinations of perovskite-solvent-antisolvent could be prepared and characterized by the platform in 2 days [36]. Although HT experimentations have made remarkable achievements on the discovery and evolution of perovskite materials, they are still in their infancy due to their higher cost and complexity with respect to HT calculations.

Besides HT computations and experimentations, some databases also provide considerable perovskite data for ML after years of development. The HybriD^3 material database [47], jointly created by Duke University and others, comprehensively collects experimental and computational material data of crystalline organic-inorganic compounds. The database contains existing, predicted, and newly synthesized materials. Researchers in the Chemical Material Solution Center of Korea Research Institute of Chemical Technology collected data on the detailed characteristics, structure, and performance of each layer of perovskite solar cells from the literatures, and established a perovskite solar cell database (Perovskite Solar Cells DB) [48]. The database collected a total of 688 documents, 2711

structures, and 17,098 properties, and readers can search independently for different properties and structures, and the database also provides corresponding links to data literatures for reference. The Computational Material Repository (CMR) [49], led by the Center for Atomic Materials Physics of the Technical University of Denmark, uses effective methods to represent and analyze the electronic structure of materials. Among them, there are a number of different CMR projects that cover different types of perovskites. The analysis shows that these perovskite projects include electronic structure, spectrum, and some other different properties. Marchenko et al. provided an open-access database of experimentally investigated hybrid organic-inorganic two-dimensional perovskite-like crystal structure, which contains various properties of 515 compounds from published literatures [50]. In addition, many comprehensive online material databases built from first-principles calculations also contain a large amount of perovskite data, including AFLOWLIB [13], Materials Project [51], Open Quantum Materials Database (OQMD) [52], and Atomly [53]. These databases not only provide a large amount of perovskite data, but also have become an important carrier of information circulation and an important link of data analysis in materials science.

3 Materials Representations

The process of converting the material system into an accurate numerical representation is the key for ML model building to achieve great performance [54–56]. In this process, the relationship between microstructure and target properties (quantitative structure property relationships (QSPR)) enables to be established by using descriptive parameters (defined as descriptors or features) [57, 58]. In general, different problems need to choose specific material descriptors, which heavily rely on the characteristics and target properties of materials. Therefore, to accurately and comprehensively describe the QSPR of materials, the construction of material descriptors usually requires the prior knowledge of the fundamental chemistry and physics [54].

The construction process of the material descriptors is actually to integrate the physical and chemical knowledge related to the target properties into the ML model, which controls the performance of a ML approach. In addition to satisfying desired accuracy of the predictions, any good material descriptor should satisfy the following conditions: (1) descriptors can uniquely describe materials and basic processes related to target properties; (2) materials with large differences (similarities) should be represented by descriptors with the same large differences (similarities); (3) the descriptors should be determined in such a way as to avoid extensive calculations to make a preliminary assessment of the material properties; and (4) the dimensions of descriptors should be kept as low as possible while ensuring model accuracy [7]. In the following content, we will give a concise summary of descriptors for perovskites in photovoltaic applications.

3.1 Descriptors for Perovskites in Photovoltaic Applications

In the past few years, numerous studies of perovskite material design based on ML techniques have emerged that target stability, bandgap, PCE, and other photovoltaic properties. Accordingly, a variety of material descriptors for perovskite have developed and provided an effective way to describe the QSPR between structures and photovoltaic properties for perovskites. These descriptors that can be obtained directly without calculations or experiments mainly fall into three categories: element properties, crystal structure, and experimental parameter. The element property descriptors are mainly used to provide the elemental information of perovskite composition, including the atomic number, Mendeleyev number, orbital radii of atoms, ionic radius of ions, electronegativity, and so on. Crystal structure descriptors contain tolerance factor, octahedral factor, Smooth Overlap of Atomic Positions (SOAP) [59], Crystal Graph Convolutional Neural Networks (CGCNN) [10], and so on. Experimental parameter descriptors, such as precursor concentration, pK_a values, and so on, are usually applied to accelerated experimental synthesis or characterization of perovskites. In addition to the three main categories mentioned above, some other descriptors are also utilized for perovskites, such as binary element descriptors (a set of binary digits representing the presence of chemical elements) [58]. After the descriptor is selected, Fig. 5 schematically illustrates the procedure to generate such descriptors for compounds.

Among these three types of descriptors, the development of crystal structure descriptors plays a significant role in perovskite design. The general chemical formula of perovskites is ABX_3, and the crystal structure of cubic perovskites is shown in Fig. 6a, respectively. To describe the formability of cubic perovskites, Goldschmidt [62] proposed an empirical formula named tolerance factor based on crystal structure in 1926, defined as $t = (r_A + r_X)/\sqrt{2}(r_B + r_X)$, in which the r_A, r_B, and r_X represent ionic radii of A-, B-, and X-site ions, respectively. According to the rigid sphere model, the length of A-X bonds and B-X bonds can be assumed as $r_A + r_X$ and $r_B + r_X$, respectively. Instead of tolerance factor, Li et al. [63]

Fig. 5 Schematic illustration of how to generate compound descriptors. In representation matrix, x_{ij} denotes the representation of feature j in compound i. Here, x_{ij} is a scalar or vector

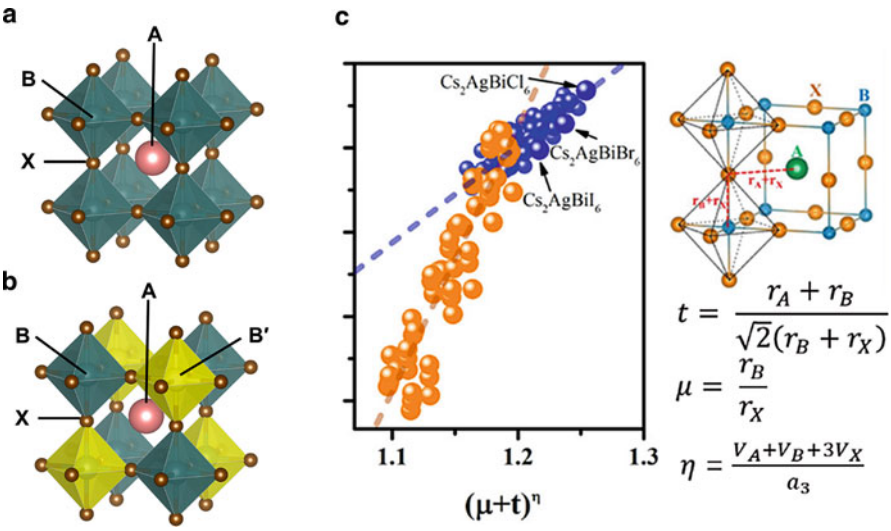

Fig. 6 Structure of (**a**) cubic perovskites ABX_3 and (**b**) double perovskites $A_2BB'X_6$. (**c**) Representation of tolerance factor (t), octahedral factor (μ), and atomic packing fraction (η) for cubic perovskites. r_A, r_B, and r_X represent the ionic radii of A-, B-, and X-site ions, respectively. According to the rigid sphere model, V_A, V_B, and V_X represent the atomic volume of A-, B-, and X-site atoms, respectively. a represents the lattice constant of the cubic cell of perovskites. (Reprinted with permission from Ref. 60. Copyright 2017 American Chemical Society)

proposed a binary descriptor (t, μ) to further clarify the formability of perovskites, which μ is the octahedral factor defined as $\mu = r_B/r_X$ [2]. Based on the analysis of existing perovskites, the stable region in (t, μ) map for halide perovskites is $0.813 < t < 1.107$ and $0.377 < \mu < 0.895$ [64]. These empirical rules successfully guide the discovery of numerous stable perovskites; however, both tolerance factor and octahedron factor are developed based on the inorganic cubic perovskite structure, resulting in the great limitation of application and low predictive accuracy for the formability of other perovskite structures, such as HOIPs and double perovskites. To improve the tolerance factor reliability for HOIPs, Kieslich et al. [65] extended the Goldschmidt tolerance factor by considering the effective radii of organic ions in HOIPs. According to the results, HOIPs were expected to form for tolerance factor between 0.8 and 1, as in the case of solid-state perovskites. To expand the application range of tolerance factor to double perovskites (structure is show in Fig. 6b), Sun and Yin [66] combined the atomic packing fraction (η) with t and μ, and developed a geometric structure descriptor $(t + \mu)^\eta$, which was linearly related to the decomposition energies of perovskites (Fig. 6c). For cubic halide and chalcogenide perovskites, the accuracy of thermodynamic stability prediction was over 86%. Filip et al. [60] developed a generalized tolerance factor $t = (r_A/r_X + 1)/[2(\overline{\mu} + 1)^2 + \Delta\mu^2]^{1/2}$ by analyzing crystal structure of double perovskites. In contrast to the traditional tolerance factor, the generalized tolerance factor was taken into account two octahedral parameters related to the B- and B′-

site cations, the average octahedral factor $\overline{\mu} = (r_B + r_{B'})/2r_X$ and the octahedral mismatch $\Delta\mu = |r_B - r_{B'}|/2r_X$. Benefitting from these careful considerations, the predictive accuracy of generalized tolerance factor for double perovskites reached 80%. These developed crystal structure descriptors have laid the foundation for subsequent ML studies of perovskite materials.

Designing descriptors based on physical and chemical intuition might introduce deviation, resulting in ignoring the best descriptor and hidden structure-property relationship. Fortunately, big-data analysis and symbolic regression technology can quickly and intelligently construct ideal descriptors for target property. One of the attractive methods is the sure independence screening and sparsifying operator (SISSO) [61]. Base on this algorithm, Bartel et al. [12] developed an improved tolerance factor $\tau = \frac{r_X}{r_B} - n_A\left(n_A - \frac{r_A/r_B}{\ln(r_A/r_B)}\right)$, where n_A represents the oxidation state of A-site ions. The new descriptor τ exposes a high prediction accuracy of perovskite stability (92%) on the dataset containing 576 experimentally existing ABX_3 compounds, while the Goldschmidt tolerance factor t only correctly classifies 74% compounds on the same dataset. In particular, the Goldschmidt tolerance factor t can correctly distinguish 49% of non-perovskites, and τ achieves 89% accuracy for non-perovskites, leading to the great improvement of predictive capability. Besides, the new descriptor τ exhibits the high accuracy for $A_2B'B''X_6$ compounds (91% accuracy), suggesting the strong generalization ability on perovskites.

3.2 Feature Engineering

For any ML method that targets toward a desired material property, it usually depends on certain number of features (descriptors). Although there may be many factors that affect the target property of materials, the number of features must be reasonable. The best strategy is to choose features that perfectly represent the corresponding property, and the number of features should be less than the number of materials in input dataset to avoid the curse of dimensionality and model overfitting [67]. Especially for material simulation, the amount of data available may be only about 10^3 or less. For such small-scale dataset, how to reasonably choose the material descriptor is crucial [68].

In the process of ML, we usually preliminary perform a relatively rough descriptor screening process. First of all, some features based on the prior knowledge of the physic and chemistry are chosen to build the initial feature set. Then, feature selection, basically a ranking procedure, is applied to pick out the best features by evaluating the model performance. For small-scale data sets, Lu et al. [69, 70] employed a "last-place elimination" feature selection procedure in a ML algorithm to optimize the most relevant features. One can also do feature screening through batch processing, such as principal component analysis, clustering, and so on. For large-scale data sets, because the scale of data set itself is very large, features can be

extracted automatically through a deep learning algorithm without artificial feature set construction. For example, Ziletti et al. [71] constructed a deep learning neural network model based on diffraction images for automatic classification of crystal structures.

In addition to some of the conditions mentioned above for descriptor construction, some specific conditions according to physics and chemistry should also be satisfied. Regarding the design of the descriptor, no matter which form is used, it should be invariant to certain transformations-spatial translational symmetry and rotational symmetry. Therefore, we cannot simply turn the descriptor into a pure "data problem." It should contain some physical and chemical origins.

4 Machine Learning in Perovskite-Based Material Discovery and Study

4.1 Stability

The first step in design of new perovskites is to evaluate the stability, which is usually assessed by the tolerance factor and the octahedron factor. Although these two factors provide a quantitative range for the formability of stable perovskites, their predictions are not accurate enough. The ML-based approach can describe the materials more detailed by constructing appropriate descriptors, and thereby more reliable prediction results can be obtained theoretically. To predict the formability of ABO_3 perovskite, Pilania et al. [75] trained a ML model based on 354 ABO_3 compounds, and created a high-dimensional feature space relating to perovskite structure formability. The approach achieves 95% accuracy in the prediction of perovskite formability. Subsequently, the authors utilized this ML-based approach to search new perovskite halides. In this work, a ML model based on 185 experimentally known perovskites was built to evaluate the formability of perovskite halides. After exploring a number of initial features, ionic radii, tolerance factor, and octahedron factor were determined as the three effective features affecting perovskite formability, demonstrating the great importance of geometric factor on perovskite formability. The trained model achieved an accuracy of 92% for the test set [76]. In addition to formability, some ML models made good performance in predicting the thermodynamic stability of perovskites [77, 78]. For example, ML model was applied to predict the thermodynamic stability of all possible perovskite and antiperovskite crystals that can be generated with elements from hydrogen to bismuth (excluding rare gases and lanthanides) according to the energy above the convex hull. ML algorithm gives the mean absolute error (MAE) of the energy above the convex hull (121 meV/atom) in the test set of 230,000 perovskites, after being trained in 20,000 samples (Fig. 7a) [79]. In addition to cubic perovskites, ML-based approach also makes a progress in identifications of diverse phases of perovskites. Balachandran et al. [72] developed a two-step framework to search for

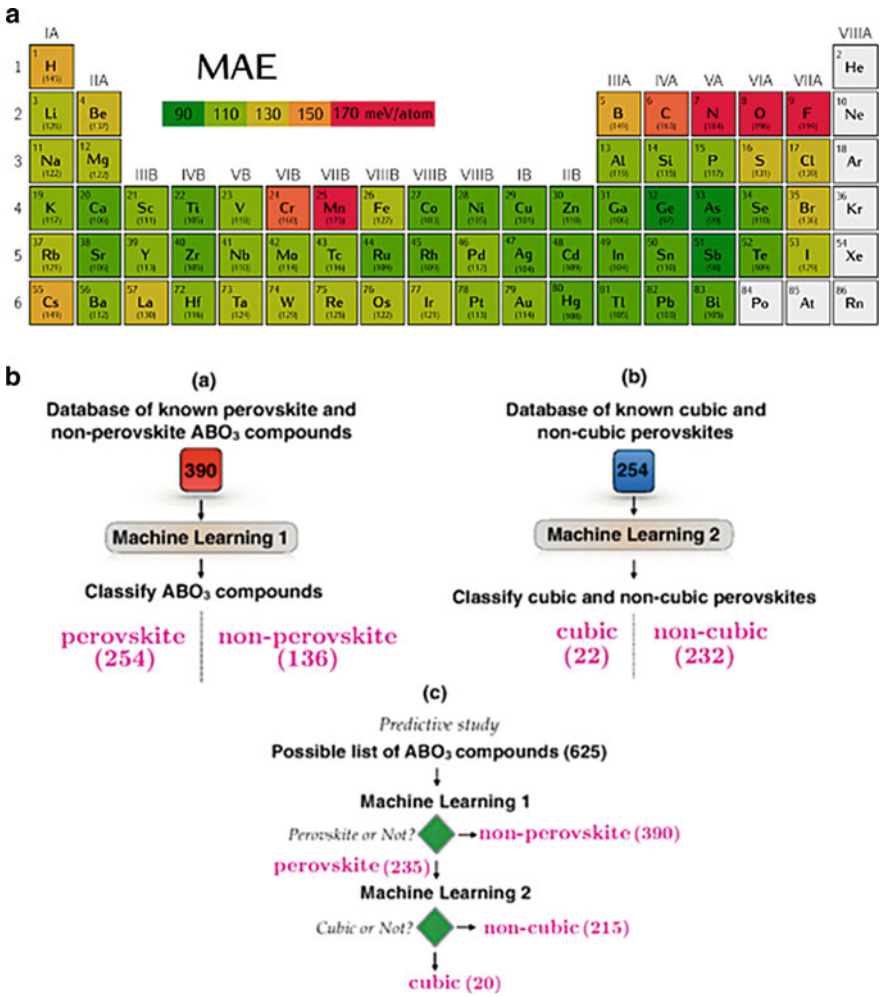

Fig. 7 (**a**) Mean absolute error (MAE, meV/atom) of the test set for AdaBoost used with extremely random tress averaged over all compounds containing each element of the periodic table. The numbers in parentheses are the actual MAE for each element. (Reprinted with permission from Ref. 72. Copyright 2017 American Chemical Society). (**b**) The ML workflow for the prediction of new ABO_3 cubic perovskites. Two independent ML models for the classification of ABO_3 into perovskites or not (machine learning 1) and cubic or noncubic perovskites (machine learning 2). (Reprinted with permission from Ref. 73. Copyright 2018 American Physical Society)

cubic perovskites in ABO_3 compounds (Fig. 7b). Firstly, a ML model was utilized for distinguishing perovskites and non-perovskites with an average cross-validation accuracy of 90%. Then, another ML model was applied for screening out cubic perovskites, and the average cross-validation accuracy was over 94%. Ye et al. [73] introduced deep neural network into predicting the formation energy of perovskite

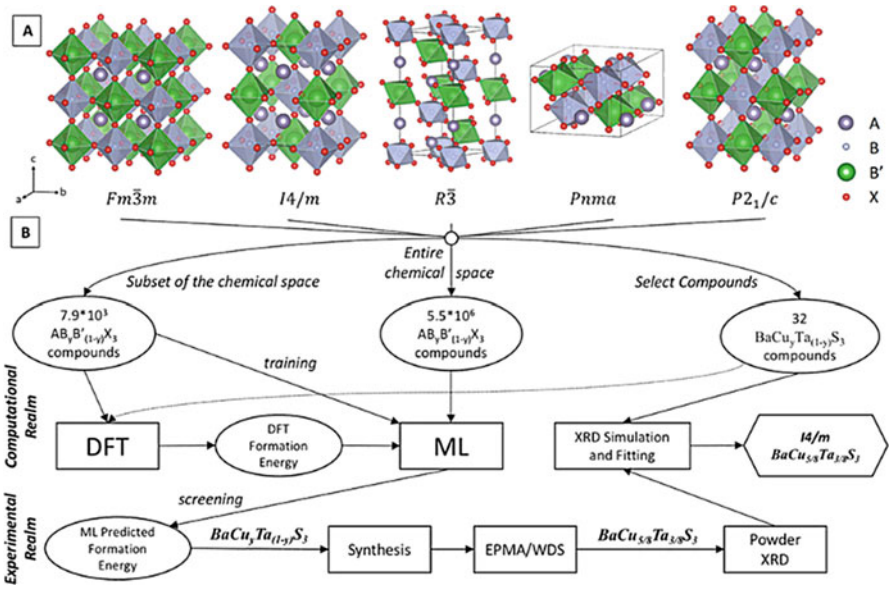

Fig. 8 (**a**) Crystal structure of $AB_yB'_{(1-y)}X_3$ perovskites in different space groups. The ML workflow for prediction of stability perovskite for experimental synthesis. (Reprinted with permission from Ref. 11. Copyright 2019 American Chemical Society)

oxides. Based on only the two descriptors of electronegativity and ionic radius, the trained ML model obtained a high accuracy and a MAE of 20–34 meV/atom. Furthermore, a new binary encoding scheme was introduced to, including the effect of cation orderings, extend ML models to mixed perovskites with low MAE (20–39 meV/atom).

The success of ML in evaluating the thermodynamic stability of single perovskites has inspired more application of ML-based approaches for other more complex perovskites, such as double perovskites, mixed perovskites, and HOIPs [11, 70, 74, 81–83]. Askerka et al. [11] proposed a learning-in-template strategy to rapidly select out double perovskites from 5×10^6 candidates. As displayed in Fig. 8, a series of possible templates corresponding to different crystal structures and stoichiometries were defined. In principle, any $A_2B'B''X_6$ compound belongs to one of these templates. The training and test sets contain formation energy data of 7.9×10^3 compounds in defined chemical space, and the accuracy of the trained ML model is up to 97%. Considering the difference of ionic radii of X-site ions in mixed X-site inorganic perovskites, Lu et al. proposed a modified Goldschmidt tolerance factor and octahedron factor using ML feature engineering. By applying the optimal feature set contained two new descriptors, the accuracy of the gradient boosting classification (GBC) model for perovskite formability is up to 89% [70]. Ali et al. [83] built a deep neural network model to study the cubic phase stability of mixed-cation perovskites. The predicted cubic phase-

stability diagram reveals that with increasing Cs proportion, perovskites possess higher cubic phase stability. This stems from that the large ionic radii of organic molecules in HOIPs bring the internal stress, and the small ionic radii of Cs might offset this internal stress. Moreover, under the guidance of ML-predicted results, $MA_{0.85}DMA_{0.15}PbI_3$ (dimethylammonium (DMA)) can be recovered to the cubic phase at room temperature by adding <10 mol% of cesium cation additives. This suggests that the established ML model can effectively guide further experimental synthesis, avoiding plenty of trial-and-error processes.

The stability of perovskite devices in the operation environment is very important for the practical application of perovskites [80, 84]. Sun et al. [35] utilized a fully connected deep neural network to classify compounds based on experimental X-ray diffraction data into 0D, 2D, and 3D structures with 90% accuracy, more than ten times faster than human analysis. Kirman et al. [85] constructed a framework by combining HT experiments and convolution neural networks to effectively guide unexplored perovskite single crystals experiments. With 7000 graphs from 96 perovskite single crystal growth experiments with different experimental parameters, the ML model was trained to recognize whether crystals could be possibly grown. In addition to distinguishing perovskite crystals, exploring the impact of experimental parameters on the crystallization of perovskite crystals can effectively guide the sequence experiments. Accordingly, a ML regression model was utilized to establish the map between experimental parameters and the probabilities of crystallization, and returned optimal experimental parameters for crystallization.

The poor environmental stability of perovskites severely hinders their practical applications. Various works have discovered that posttreatment with small molecules by dip-coating or spin-coating can effectively improve the stability of perovskites in the humid environment [80, 84]. However, the addition of some molecules (such as amines) might destroy the perovskites film structures. Therefore, it is of practical importance for improving the environmental stability of perovskites through finding suitable molecules possessing compatibility for the perovskite film. Yu et al. [86] established a ML model to study the relationship between properties of amines and their reactivity, and achieved 86% accuracy on predicting the outcomes for whether the qualities of perovskite films are maintained after posttreatment. The results show that amine compounds and pyridine derivatives with a few hydrogen bond donors, large space volume, and large number of substitutions on nitrogen atoms have high compatibility with perovskite films, which can effectively guide further experimental synthesis.

4.2 Photovoltaic Property

The most important electronic property for a solar absorber is bandgap. According to the Shockley-Queisser limit, perovskites with bandgap in the optimal range of 0.9–1.6 eV are promising for single-junction solar cells [87]. Therefore, selecting perovskites with appropriate bandgaps is a vital step in solar cell design. It is

well known that DFT calculations based on PBE functional seriously underestimates bandgaps for semiconductors and insulators. However, advanced theoretical methods (such as hybrid functional or GW) are computationally expensive and time consuming making a high-throughput search inefficient, not to mention experiments. An effective strategy is to combine HT calculation or experimentation with ML to minimize the high cost. In recent years, bandgap prediction has been attempted across a wide range of materials, especially perovskites, using different ML methods such as neural networks, support vector regression, and gradient boosting regression (GBR) [88–97]. Pilania et al. [98] applied kernel ridge regression algorithm to predict the bandgap of double perovskites at the GLLB-SC-level, in which a systematic feature-engineering approach was utilized to identify the optimal feature set from a set of more than 1.2 million candidate features. The final ML model achieved a high prediction accuracy on bandgap (about 0.947). In order to obtain more accurate bandgap values, the researchers developed a multi-fidelity framework combining first-principles calculations and ML techniques, which can estimate high-fidelity data based on low-fidelity data [99]. In this work, PBE-level bandgap values of 599 double perovskites were treated as low-fidelity data, while bandgap values at the Heyd-Scuseria-Ernzerhof (HSE06) level of the same perovskites were treated as the high-fidelity data. By utilizing the framework, high-fidelity HSE06-level bandgap values were approximated from low-fidelity PBE-level bandgap values. Besides inorganic perovskite, Lu et al. [69] developed a framework combining ML techniques and DFT calculations to rapidly predict bandgaps of HOIPs. The GBR model was trained based on PBE-level bandgap values of 212 HOIPs, and achieved high coefficient of determination (R^2) of 97% (Fig. 9c). As is shown in Fig. 9a, the feature importance reveals that, in structure features, the tolerance factor has the most significant impact on bandgap. Besides, the ionization energy, electronegativity, and electron affinity energy of B-site ions are more related to bandgap than those of A and X-site ions. Subsequently, the trained model was applied to predict the bandgap of 5158 unexplored HOIPs and the prediction result is shown in Fig. 9c. Finally, six HOIPs were picked out and validated using DFT calculations. Results in Fig. 9d show that the accuracy of ML-predicted bandgaps is comparable to that of DFT calculations. Similarly, Marchenko developed a ML model using GBR basing on the open-access database of experimentally investigated HOIPs with a 2D perovskite-like crystal structure for the prediction of a bandgap with accuracy within 0.1 eV [50]. The SOAP kernel was used to describe the local atomic environment of each atom and the trained model achieved R^2 as high as 0.9.

Rational chemical mixing is an effective approach to appropriately tune properties of perovskites. For example, mixing halogen elements can adjust the bandgap of halide perovskites, leading to higher performance as solar cells materials [101]. Choubisa et al. developed a descriptor related to the atomic arrangement for mixed perovskites, called as crystal site feature embedding (CSFE, see Fig. 10) [102]. Based on the CSFE representation, the ML model for total energies achieves an MAE of 3.5 meV/atom, and the ML model for bandgaps possesses an MAE of 0.069 eV. The trained model was applied to the predicted bandgap of triple B-site

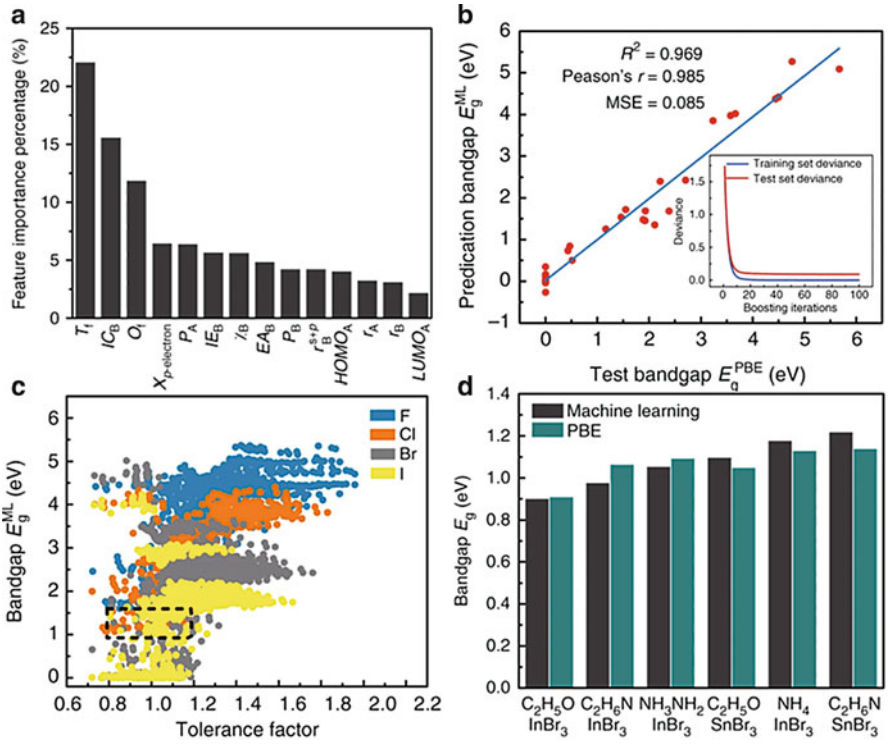

Fig. 9 (**a**) Importance of the selected features. The 14 selected features are ranked using GBR algorithm. (**b**) The fitting results of test bandgaps E_g^{PBE} and predicted bandgaps E_g^{ML}. The subplot is the convergence of model accuracy for five cross-validation split of the data. (**c**) Data visualization of predicted bandgaps for all possible HOIPs (one color represents a class of halogen perovskites) with tolerance factor. (**d**) A comparison between ML-predicted and DFT-calculated results of six selected HOIPs. (Reprinted with permission from Ref. 70)

$MAPb_xSn_yCd_zI_3$ perovskites. ML-predicted results revealed that a small proportion of Cd can tune the bandgap of perovskites to the optimal range for photovoltaic applications. Furthermore, ML models for total energies and bandgaps based on CSFE representation are also suitable for two-dimensional perovskites, with a MAE of 7 meV/atom and 0.13 eV, respectively. Moreover, a variational autoencoder was employed to realize inverse design for perovskites with target properties. Besides searching for potential solar cells materials (bandgap between 1.1 and 1.3 eV), perovskites for infrared sensors (bandgap ~1 eV) and ultraviolet lasers (bandgap ~3.2 eV) were also screened, and selected perovskites were validated by DFT calculations based on HSE06 functional.

The power conversion efficiency (PCE) is a standard parameter to assess the ability of light-electron energy conversion for photovoltaic devices, relating to the optical absorption performance of absorber materials, defect structures, energy-level mismatch, etc. [100]. Searching for high-performance PSCs generally based on the

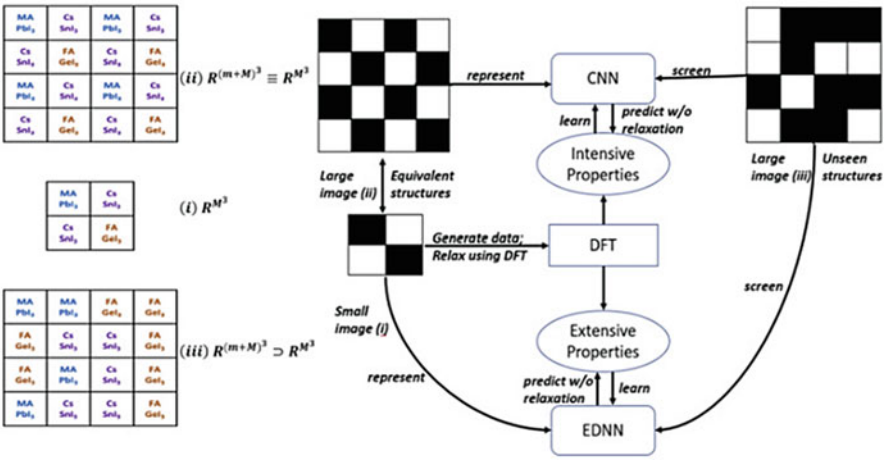

Fig. 10 Workflow for materials discovery using crystal site feature embedding. (Reprinted with permission from Ref. 100. Copyright 2020 Elsevier Inc.)

exhaustive search method, which brings expensive cost of time, materials, equipment, and man power. ML techniques could provide some guidelines and accelerate the discovery of high-performance PSCs without numerous experiments. Odabasi and Yildirim [104] systematically reviewed publications related to perovskites solar cells, and collected 1921 data from 800 publications. Constituent materials and preparation methods of perovskites solar cells were selected as the input variables of the random forest regression model, and PCE was taken as the output variable. Since the input variables of n-i-p and p-i-n perovskites based solar cells were different, two models were trained for each type of solar cells, respectively. For ML model of n-i-p solar cells, the root mean squared error (RMSE) of training set and test set is 1.70 and 3.29, respectively. The model of p-i-n solar cells achieves the RMSE of 1.51 and 2.91 for training set and test set, respectively (Fig. 11). For perovskite solar cells with PCE in the range of 18–23.3%, the association rule mining techniques results exhibit that mixing cations is an effective approach to obtain the solar cells with stabilized PCE higher than 18%. Li et al. [103] collected 333 PSC data from 2000 peer-reviewed publications, and proposed a two-step framework to study the performance of PSCs. At the first step, the experimental bandgap values of perovskites are predicted. Then an ML model to predict the PCE of PSCs was established with considering experimental bandgaps, the difference of HOMO energy level between hole transporting layers and perovskites, and the difference of LUMO energy level between electron transporting layers and perovskites. The RMSE of ML models for bandgaps and PCE is 0.06 eV and 3.23%, respectively. The optimal bandgap range corresponding to the highest PCE is from 1.15 to 1.35 eV, demonstrating high consistency with the Shockley-Queisser limit.

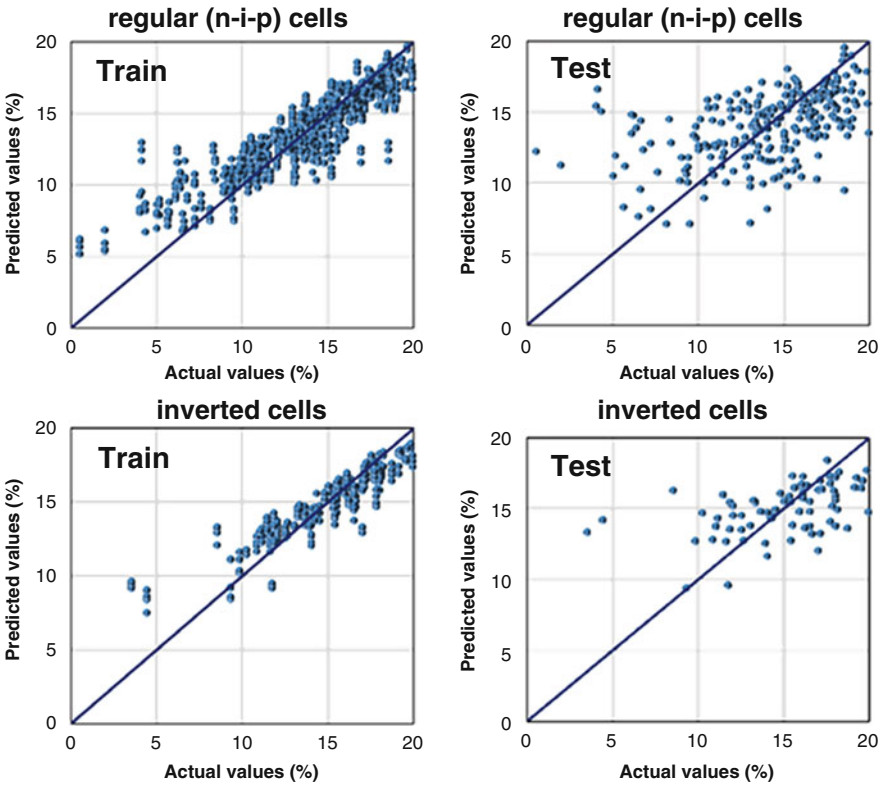

Fig. 11 Actual versus predicted performances by random forest model for training and testing for regular (n-i-p) cells; training and testing for inverted cells. (Reprinted with permission from Ref. 103. Copyright 2018 Elsevier Ltd.)

5 Conclusion and Prospects

To summarize, the rapid development of ML techniques has accelerated novel perovskite material discovery and mechanism exploration. Considering the continuous developments in experimental and computational tools, as well as the ML and data management technologies, ML-aided perovskite researches will increase dramatically. However, there are also some challenges that need to be overcome for the more effective utilization of ML in perovskite researches.

First of all, the importance of data issues in ML researches cannot be overemphasized. Perovskite data generated from the first-principle calculations are rich, especially some material properties like bandgap and formation energy. However, high-quality calculation or experimental data are still lacking due to the high cost and time consuming. Consequently, the models trained basing on low-quality datasets contain bias and are not very suitable for practical application. When high-quality datasets tend to be small, models constructed from these datasets have

limited generalization ability while they are sensitive to the outliers, noise and imbalanced data structure. The strategies to solve the above contradictions and challenges are as follows. (1) Transfer learning may be beneficial for the ML application in small-scale datasets of perovskite-based photovoltaic materials. A ML model (usually with a small data set) can be built from different (but similar) structures' data with larger-scale. In detail, the ML models are firstly utilized to analyze the large datasets created with low-cost computational methods, then the experimental or high accuracy computational data as supplements are utilized to correct the internal bias of the model. This technique can not only fix the problem of data lack, but also reduce the gap between theory and experiment. (2) Using active learning algorithm. Active learning is an iterative procedure, where the initial model is trained on a small dataset, and in each step, the model is re-trained on data expanded with new samples, which are added based on results from the previous steps in order to maximize the learning rate. (3) In fact, the overwhelming majority of scientific knowledge is published as text, so scientific literature is in fact served as data sources as well. In addition to the material data, the literatures contain valuable knowledge about the connections and relationships between the data items interpreted by the authors. Therefore, extracting data and QSPR from the literatures through techniques such as natural language processing will facilitate the development of ML in perovskite research. (4) The high-quality perovskite database should be built according to FAIR (Findable, Accessible, Interoperable, Reusable) data sharing principle [105].

There are also some issues need to be addressed for ML models. A good ML model mainly depends on two factors: first, the predictive performance of the ML model; and second, the interpretability potential of the model. On one hand, the generalization ability of the model is not always verified. A lot of work has achieved excellent performance on the training and test sets, but the prediction results of extended dataset are often not validated using first-principle calculations or experiments. On the other hand, it is often challenging to provide a physical and chemical interpretation of complex ML models, as the goal of the learning process is to find a model that maximizes prediction performance, which may require (possibly non-linear) combinations of hundreds of features. But if the model can be explained based on physical and chemical principle, it will help researchers to insight the structure-property relationship of materials more deeply. In this aspect, feature importance analysis, model visualization, and SHAP analysis [106] will help the development of interpretable models.

Acknowledgments This work is supported by the National Key Research and Development Program of China (2017YFA0204800), Natural Science Foundation of China (21525311, 21773027, 22033002), the National Natural Science Foundation of Jiangsu (BK20180353), the Fundamental Research Funds for the Central Universities of China, and Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX20_0075) in China.

References

1. Vasala, S., & Karppinen, M. (2015). A2B′B″O6 perovskites: A review. *Progress in Solid State Chemistry, 43*, 1.
2. Kobayashi, K. I., Kimura, T., Sawada, H., Terakura, K., & Tokura, Y. (1998). Room-temperature magnetoresistance in an oxide material with an ordered double-perovskite structure. *Nature, 395*, 677.
3. Ju, M.-G., Chen, M., Zhou, Y., Dai, J., Ma, L., Padture, N. P., & Zeng, X. C. (2018). Toward eco-friendly and stable perovskite materials for photovoltaics. *Joule, 2*, 1231.
4. PCE. Retrieved from https://www.nrel.gov/pv/cell-efficiency.html.
5. Zhao, X.-G., Yang, D., Ren, J.-C., Sun, Y., Xiao, Z., & Zhang, L. (2018). Rational design of halide double perovskites for optoelectronic applications. *Joule, 2*, 1662.
6. Fuelling discovery by sharing. 2013. *Nature Materials*, *12*, 173.
7. Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C., & Scheffler, M. (2015). Big data of materials science: Critical role of the descriptor. *Physical Review Letters, 114*, 105503.
8. Sarmiento-Pérez, R., Cerqueira, T. F. T., Körbel, S., Botti, S., & Marques, M. A. L. (2015). Prediction of stable nitride perovskites. *Chemistry of Materials, 27*, 5957.
9. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature, 559*, 547.
10. Xie, T., & Grossman, J. C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters, 120*, 145301.
11. Askerka, M., Li, Z., Lempen, M., Liu, Y., Johnston, A., Saidaminov, M. I., Zajacz, Z., & Sargent, E. H. (2019). Learning-in-templates enables accelerated discovery and synthesis of new stable double perovskites. *Journal of the American Chemical Society, 141*, 3682.
12. Bartel, C. J., Sutton, C., Goldsmith, B. R., Ouyang, R., Musgrave, C. B., Ghiringhelli, L. M., & Scheffler, M. (2019). New tolerance factor to predict the stability of perovskite oxides and halides. *Science Advances, 5*, eaav0693.
13. Curtarolo, S., Setyawan, W., Wang, S., Xue, J., Yang, K., Taylor, R. H., Nelson, L. J., Hart, G. L. W., Sanvito, S., Buongiorno-Nardelli, M., Mingo, N., & Levy, O. (2012). AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Computational Materials Science, 58*, 227.
14. Ong, S. P., Richards, W. D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V. L., Persson, K. A., & Ceder, G. (2013). Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science, 68*, 314.
15. Hjorth Larsen, A., Jorgen Mortensen, J., Blomqvist, J., Castelli, I. E., Christensen, R., Dulak, M., Friis, J., Groves, M. N., Hammer, B., Hargus, C., Hermes, E. D., Jennings, P. C., Bjerre Jensen, P., Kermode, J., Kitchin, J. R., Leonhard Kolsbjerg, E., Kubal, J., Kaasbjerg, K., Lysgaard, S., Bergmann Maronsson, J., Maxson, T., Olsen, T., Pastewka, L., Peterson, A., Rostgaard, C., Schiotz, J., Schutt, O., Strange, M., Thygesen, K. S., Vegge, T., Vilhelmsen, L., Walter, M., Zeng, Z., & Jacobsen, K. W. (2017). The atomic simulation environment—A Python library for working with atoms. *Journal of Physics. Condensed Matter, 29*, 273002.
16. Yang, X., Wang, Z., Zhao, X., Song, J., Zhang, M., & Liu, H. (2018). MatCloud: A high-throughput computational infrastructure for integrated management of materials simulation, data and resources. *Computational Materials Science, 146*, 319.
17. Chakraborty, S., Xie, W., Mathews, N., Sherburne, M., Ahuja, R., Asta, M., & Mhaisalkar, S. G. (2017). Rational design: A high-throughput computational screening and experimental validation methodology for lead-free and emergent hybrid perovskites. *ACS Energy Letters, 2*, 837.
18. Li, Y., & Yang, K. (2020). High-throughput computational design of halide perovskites and beyond for optoelectronics. *WIREs Computational Molecular Science*. https://doi.org/10.1002/wcms.1500.

19. Körbel, S., Marques, M. A. L., & Botti, S. (2016). Stability and electronic properties of new inorganic perovskites from high-throughput ab initio calculations. *Journal of Materials Chemistry C, 4*, 3157.
20. Emery, A. A., & Wolverton, C. (2017). High-throughput DFT calculations of formation energy, stability and oxygen vacancy formation energy of ABO3 perovskites. *Scientific Data, 4*, 170153.
21. Zhao, X. G., Yang, J. H., Fu, Y., Yang, D., Xu, Q., Yu, L., Wei, S. H., & Zhang, L. (2017). Design of lead-free inorganic halide perovskites for solar cells via cation-transmutation. *Journal of the American Chemical Society, 139*, 2630.
22. Cai, Y., Xie, W., Teng, Y. T., Harikesh, P. C., Ghosh, B., Huck, P., Persson, K. A., Mathews, N., Mhaisalkar, S. G., Sherburne, M., & Asta, M. (2019). High-throughput computational study of halide double perovskite inorganic compounds. *Chemistry of Materials, 31*, 5392.
23. Castelli, I. E., Olsen, T., Datta, S., Landis, D. D., Dahl, S., Thygesen, K. S., & Jacobsen, K. W. (2012). Computational screening of perovskite metal oxides for optimal solar light capture. *Energy & Environmental Science, 5*, 5814.
24. Castelli, I. E., Landis, D. D., Thygesen, K. S., Dahl, S., Chorkendorff, I., Jaramillo, T. F., & Jacobsen, K. W. (2012). New cubic perovskites for one- and two-photon water splitting using the computational materials repository. *Energy & Environmental Science, 5*, 9034.
25. Wang, H.-C., Pistor, P., Marques, M. A. L., & Botti, S. (2019). Double perovskites as p-type conducting transparent semiconductors: A high-throughput search. *Journal of Materials Chemistry A, 7*, 14705.
26. Jiang, X., & Yin, W.-J. (2021). High-throughput computational screening of oxide double perovskites for optoelectronic and photocatalysis applications. *Journal of Energy Chemistry, 57*, 351–358.
27. Zhang, T., Cai, Z., & Chen, S. (2020). Chemical trends in the thermodynamic stability and band gaps of 980 halide double perovskites: A high-throughput first-principles study. *ACS Applied Materials & Interfaces, 12*, 20680.
28. Nakajima, T., & Sawada, K. (2017). Discovery of Pb-free perovskite solar cells via high-throughput simulation on the K computer. *Journal of Physical Chemistry Letters, 8*, 4826.
29. Körbel, S., Marques, M. A. L., & Botti, S. (2018). Stable hybrid organic–inorganic halide perovskites for photovoltaics from ab initio high-throughput calculations. *Journal of Materials Chemistry A, 6*, 6463.
30. Li, Y., & Yang, K. (2019). High-throughput computational design of organic–inorganic hybrid halide semiconductors beyond perovskites for optoelectronics. *Energy & Environmental Science, 12*, 2233.
31. Filip, M. R., & Giustino, F. (2015). Computational screening of homovalent lead substitution in organic–inorganic halide perovskites. *The Journal of Physical Chemistry C, 120*, 166.
32. Unger, E. L., Kegelmann, L., Suchan, K., Sörell, D., Korte, L., & Albrecht, S. (2017). Roadmap and roadblocks for the band gap tunability of metal halide perovskites. *Journal of Materials Chemistry A, 5*, 11401.
33. Kim, C., Huan, T. D., Krishnan, S., & Ramprasad, R. (2017). A hybrid organic-inorganic perovskite dataset. *Scientific Data, 4*, 170057.
34. Jacobs, R., Luo, G., & Morgan, D. (2019). Materials discovery of stable and nontoxic halide perovskite materials for high-efficiency solar cells. *Advanced Functional Materials, 29*, 1804354.
35. Sun, S., Hartono, N. T. P., Ren, Z. D., Oviedo, F., Buscemi, A. M., Layurova, M., Chen, D. X., Ogunfunmi, T., Thapa, J., Ramasamy, S., Settens, C., DeCost, B. L., Kusne, A. G., Liu, Z., Tian, S. I. P., Peters, I. M., Correa-Baena, J.-P., & Buonassisi, T. (2019). Accelerated development of perovskite-inspired materials via high-throughput synthesis and machine-learning diagnosis. *Joule, 3*, 1437.
36. Gu, E., Tang, X., Langner, S., Duchstein, P., Zhao, Y., Levchuk, I., Kalancha, V., Stubhan, T., Hauch, J., Egelhaaf, H. J., Zahn, D., Osvet, A., & Brabec, C. J. (2020). Robot-based high-throughput screening of antisolvents for lead halide perovskites. *Joule, 4*, 1806.

37. Ishihara, H., Sarang, S., Chen, Y.-C., Lin, O., Phummirat, P., Thung, L., Hernandez, J., Ghosh, S., & Tung, V. (2016). Nature inspiring processing route toward high throughput production of perovskite photovoltaics. *Journal of Materials Chemistry A, 4*, 6989.
38. Baker, J., Hooper, K., Meroni, S., Pockett, A., McGettrick, J., Wei, Z., Escalante, R., Oskam, G., Carnie, M., & Watson, T. (2017). High throughput fabrication of mesoporous carbon perovskite solar cells. *Journal of Materials Chemistry A, 5*, 18643.
39. Chen, S., Zhang, L., Yan, L., Xiang, X., Zhao, X., Yang, S., & Xu, B. (2019). Accelerating the screening of perovskite compositions for photovoltaic applications through high-throughput inkjet printing. *Advanced Functional Materials, 29*, 1905487.
40. Jeong, D.-N., Lee, D.-K., Seo, S., Lim, S. Y., Zhang, Y., Shin, H., Cheong, H., & Park, N.-G. (2019). Perovskite cluster-containing solution for scalable D-Bar coating toward high-throughput perovskite solar cells. *ACS Energy Letters, 4*, 1189.
41. Li, J., Du, P., Li, S., Liu, J., Zhu, M., Tan, Z., Hu, M., Luo, J., Guo, D., Ma, L., Nie, Z., Ma, Y., Gao, L., Niu, G., & Tang, J. (2019). High-throughput combinatorial optimizations of perovskite light-emitting diodes based on all-vacuum deposition. *Advanced Functional Materials, 29*, 1903607.
42. Dahl, J. C., Wang, X., Huang, X., Chan, E. M., & Alivisatos, A. P. (2020). Elucidating the weakly reversible Cs-Pb-Br perovskite nanocrystal reaction network with high-throughput maps and transformations. *Journal of the American Chemical Society, 142*, 11915.
43. Li, Z., Najeeb, M. A., Alves, L., Sherman, A. Z., Shekar, V., Cruz Parrilla, P., Pendleton, I. M., Wang, W., Nega, P. W., Zeller, M., Schrier, J., Norquist, A. J., & Chan, E. M. (2020). Robot-accelerated perovskite investigation and discovery. *Chemistry of Materials, 13*, 5650–5663.
44. Reinhardt, E., Salaheldin, A. M., Distaso, M., Segets, D., & Peukert, W. (2020). Rapid characterization and parameter space exploration of perovskites using an automated routine. *ACS Combinatorial Science, 22*, 6.
45. Surmiak, M. A., Zhang, T., Lu, J., Rietwyk, K. J., Raga, S. R., McMeekin, D. P., & Bach, U. (2020). High-throughput characterization of perovskite solar cells for rapid combinatorial screening. *Solar RRL, 4*, 2000097.
46. Chen, S., Hou, Y., Chen, H., Tang, X., Langner, S., Li, N., Stubhan, T., Levchuk, I., Gu, E., Osvet, A., & Brabec, C. J. (2018). Exploring the stability of novel wide bandgap perovskites by a robot based high throughput approach. *Advanced Energy Materials, 8*, 1701543. https://doi.org/10.1002/aenm.201701543.
47. The HybriD^3 materials database. Retrieved from https://materials.hybrid3.duke.edu/.
48. Perovskite Solar Cells DB. Retrieved from http://www.perovskite.info/perovskite/perovSearch.
49. The Computational Material Repository. Retrieved from https://cmr.fysik.dtu.dk/#.
50. Marchenko, E. I., Fateev, S. A., Petrov, A. A., Korolev, V. V., Mitrofanov, A., Petrov, A. V., Goodilin, E. A., & Tarasov, A. B. (2020). Database of two-dimensional hybrid perovskite materials: open-access collection of crystal structures, band gaps, and atomic partial charges predicted by machine learning. *Chemistry of Materials, 32*(17), 7383–7388.
51. Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., & Persson, K. A. (2013). Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials, 1*, 011002.
52. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B., & Wolverton, C. (2013). Materials design and discovery with high-throughput density functional theory: The Open Quantum Materials Database (OQMD). *JOM, 65*, 1501.
53. Atomly. Retrieved from https://atomly.net/.
54. Pilania, G., Balachandran, P. V., Gubernatis, J. E., & Lookman, T. (2020). *Data-based methods for materials design and discovery: Basic ideas and general methods* (Vol. 1, p. 1). San Rafael, CA: Morgan & Claypool Publishers.
55. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A., & Kim, C. (2017). Machine learning in materials informatics: Recent applications and prospects. *NPJ Computational Materials, 3*, 54.

56. Mueller, T., Kusne, A. G., & Ramprasad, R. (2016). Machine learning in materials science. In A. L. Parrill & K. B. Lipkowitz (Eds.), *Reviews in computational chemistry* (p. 186). Hoboken, NJ: Wiley.
57. Muratov, E. N., Bajorath, J., Sheridan, R. P., Tetko, I. V., Filimonov, D., Poroikov, V., Oprea, T. I., Baskin, I. I., Varnek, A., Roitberg, A., Isayev, O., Curtarolo, S., Fourches, D., Cohen, Y., Aspuru-Guzik, A., Winkler, D. A., Agrafiotis, D., Cherkasov, A., & Tropsha, A. (2020). QSAR without borders. *Chemical Society Reviews, 49*, 3525.
58. Seko, A., Togo, A., & Tanaka, I. (2018). Descriptors for machine learning of materials data. In *Nanoinformatics* (p. 3). Singapore: Springer.
59. Bartók, A. P., Kondor, R., & Csányi, G. (2013). On representing chemical environments. *Physical Review B, 87*, 184115.
60. Filip, M. R., & Giustino, F. (2018). The geometric blueprint of perovskites. *Proceedings of the National Academy of Sciences of the United States of America, 115*, 5397.
61. Ouyang, R., Curtarolo, S., Ahmetcik, E., Scheffler, M., & Ghiringhelli, L. M. (2018). SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Physical Review Materials, 2*, 083802.
62. Goldschmidt, V. M. (1926). Die gesetze der krystallochemie. *Naturwissenschaften, 14*, 477.
63. Li, C., Soh, K. C. K., & Wu, P. (2004). Formability of ABO3 perovskites. *Journal of Alloys and Compounds, 372*, 40.
64. Li, C., Lu, X., Ding, W., Feng, L., Gao, Y., & Guo, Z. (2008). Formability of ABX3 (X = F, Cl, Br, I) halide perovskites. *Acta Crystallographica. Section B, 64*, 702.
65. Kieslich, G., Sun, S., & Cheetham, A. K. (2015). An extended Tolerance Factor approach for organic-inorganic perovskites. *Chemical Science, 6*, 3430.
66. Sun, Q., & Yin, W. J. (2017). Thermodynamic stability trend of cubic perovskites. *Journal of the American Chemical Society, 139*, 14905.
67. Zhou, Q., Lu, S., Wu, Y., & Wang, J. (2020). Property-oriented material design based on a data-driven machine learning technique. *Journal of Physical Chemistry Letters, 11*, 3920.
68. Balachandran, P. V., Xue, D., Theiler, J., Hogden, J., Gubernatis, J. E., & Lookman, T. (2018). Importance of feature selection in machine learning and adaptive design for materials. In *Materials discovery and design* (p. 59). Cham: Springer.
69. Lu, S. H., Zhou, Q. H., Ouyang, Y. X., Guo, Y. L., Li, Q., & Wang, J. L. (2018). Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nature Communications, 9*, 3405.
70. Lu, S., Zhou, Q., Ma, L., Guo, Y., & Wang, J. (2019). Rapid discovery of ferroelectric photovoltaic perovskites and material descriptors via machine learning. *Small Methods*. https://doi.org/10.1002/smtd.201900360.
71. Ziletti, A., Kumar, D., Scheffler, M., & Ghiringhelli, L. M. (2018). Insightful classification of crystal structures using deep learning. *Nature Communications, 9*, 2775.
72. Balachandran, P. V., Emery, A. A., Gubernatis, J. E., Lookman, T., Wolverton, C., & Zunger, A. (2018). Predictions of new ABO(3) perovskite compounds by combining machine learning and density functional theory. *Physical Review Materials, 2*, 043802.
73. Ye, W., Chen, C., Wang, Z., Chu, I.-H., & Ong, S. P. (2018). Deep neural networks for accurate predictions of crystal stability. *Nature Communications, 9*, 3800.
74. Im, J., Lee, S., Ko, T. W., Kim, H. W., Hyon, Y., & Chang, H. (2019). Identifying Pb-free perovskites for solar cells by machine learning. *NPJ Computational Materials, 5*, 37.
75. Pilania, G., Balachandran, P. V., Gubernatis, J. E., & Lookman, T. (2015). Classification of ABO3 perovskite solids: A machine learning study. *Acta Crystallographica B Structural Science, Crystal Engineering and Material, 71*, 507.
76. Pilania, G., Balachandran, P. V., Kim, C., & Lookman, T. (2016). Finding new perovskite halides via machine learning. *Frontiers in Materials, 3*, 19.
77. Li, W., Jacobs, R., & Morgan, D. (2018). Predicting the thermodynamic stability of perovskite oxides using machine learning models. *Computational Materials Science, 150*, 454.
78. Xu, Q., Li, Z., Liu, M., & Yin, W. J. (2018). Rationalizing perovskite data for machine learning and materials design. *Journal of Physical Chemistry Letters, 9*, 6948.

79. Schmidt, J., Shi, J., Borlido, P., Chen, L., Botti, S., & Marques, M. A. L. (2017). Predicting the thermodynamic stability of solids combining density functional theory and machine learning. *Chemistry of Materials, 29*, 5090.
80. Yang, S., Wang, Y., Liu, P., Cheng, Y.-B., Zhao, H. J., & Yang, H. G. (2016). Functionalization of perovskite thin films with moisture-tolerant molecules. *Nature Energy, 1*, 15016.
81. Li, Z., Xu, Q., Sun, Q., Hou, Z., & Yin, W.-J. (2019). Thermodynamic stability landscape of halide double perovskites via high-throughput computing and machine learning. *Advanced Functional Materials, 29*, 1807280.
82. Mazaheri, T., Sun, B., Scher-Zagier, J., Thind, A. S., Magee, D., Ronhovde, P., Lookman, T., Mishra, R., & Nussinov, Z. (2019). Stochastic replica voting machine prediction of stable cubic and double perovskite materials and binary alloys. *Physical Review Materials, 3*, 063802.
83. Ali, A., Park, H., Mall, R., Aissa, B., Sanvito, S., Bensmail, H., Belaidi, A., & El-Mellouhi, F. (2020). Machine learning accelerated recovery of the cubic structure in mixed-cation perovskite thin films. *Chemistry of Materials, 32*, 2998.
84. Zhang, H., Ren, X., Chen, X., Mao, J., Cheng, J., Zhao, Y., Liu, Y., Milic, J., Yin, W.-J., Grätzel, M., & Choy, W. C. H. (2018). Improving the stability and performance of perovskite solar cells via off-the-shelf post-device ligand treatment. *Energy & Environmental Science, 11*, 2253.
85. Kirman, J., Johnston, A., Kuntz, D. A., Askerka, M., Gao, Y., Todorovic, P., Ma, D., Prive, G. G., & Sargent, E. H. (2020). Machine-learning-accelerated perovskite crystallization. *Matter, 2*, 938.
86. Yu, Y., Tan, X., Ning, S., & Wu, Y. (2019). Machine learning for understanding compatibility of organic–inorganic hybrid perovskites with post-treatment amines. *ACS Energy Letters, 4*, 397.
87. Shockley, W., & Queisser, H. J. (1961). Detailed balance limit of efficiency of p-n junction solar cells. *Journal of Applied Physics, 32*, 510.
88. Zhang, L., He, M., & Shao, S. (2020). Machine learning for halide perovskite materials. *Nano Energy, 78*, 105380.
89. Yılmaz, B., & Yıldırım, R. (2021). Critical review of machine learning applications in perovskite solar research. *Nano Energy, 80*, 105546.
90. Takahashi, K., Takahashi, L., Miyazato, I., & Tanaka, Y. (2018). Searching for hidden perovskite materials for photovoltaic systems by combining data science and first principle calculations. *ACS Photonics, 5*, 771.
91. Agiorgousis, L. M., Sun, Y. Y., Choe, D. H., West, D., & Zhang, S. (2019). Machine learning augmented discovery of chalcogenide double perovskites for photovoltaics. *Advanced Theory and Simulations*. https://doi.org/10.1002/adts.201800173.
92. Wu, T., & Wang, J. (2019). 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, 66*, 104070.
93. Chaube, S., Khullar, P., Srinivasan, S. G., & Rai, B. (2020). A statistical learning framework for accelerated bandgap prediction of inorganic compounds. *Journal of Electronic Materials, 49*, 752.
94. Gladkikh, V., Kim, D. Y., Hajibabaei, A., Jana, A., Myung, C. W., & Kim, K. S. (2020). Machine learning for predicting the band gaps of ABX(3) perovskites from elemental properties. *Journal of Physical Chemistry C, 124*, 8905.
95. Jao, M. H., Chan, S. H., Wu, M. C., & Lai, C. S. (2020). Element code from pseudopotential as efficient descriptors for a machine learning model to explore potential lead-free halide perovskites. *Journal of Physical Chemistry Letters, 11*, 8914.
96. Park, H., Mall, R., Ali, A., Sanvito, S., Bensmail, H., & El-Mellouhi, F. (2020). Importance of structural deformation features in the prediction of hybrid perovskite bandgaps. *Computational Materials Science, 184*, 109858.

97. Saidi, W. A., Shadid, W., & Castelli, I. E. (2020). Machine-learning structural and electronic properties of metal halide perovskites using a hierarchical convolutional neural network. *NPJ Computational Materials, 6*, 36.
98. Pilania, G., Mannodi-Kanakkithodi, A., Uberuaga, B. P., Ramprasad, R., Gubernatis, J. E., & Lookman, T. (2016). Machine learning bandgaps of double perovskites. *Scientific Reports, 6*, 19375.
99. Pilania, G., Gubernatis, J. E., & Lookman, T. (2017). Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Computational Materials Science, 129*, 156.
100. Li, W., Wang, Z., Deschler, F., Gao, S., Friend, R. H., & Cheetham, A. K. (2017). Chemically diverse and multifunctional hybrid organic–inorganic perovskites. *Nature Reviews Materials, 2*, 16099.
101. Zhang, X., Li, L., Sun, Z., & Luo, J. (2019). Rational chemical doping of metal halide perovskites. *Chemical Society Reviews, 48*, 517.
102. Choubisa, H., Askerka, M., Ryczko, K., Voznyy, O., Mills, K., Tamblyn, I., & Sargent, E. H. (2020). Crystal site feature embedding enables exploration of large chemical spaces. *Matter, 3*, 433–448.
103. Li, J., Pradhan, B., Gaur, S., & Thomas, J. (2019). Predictions and strategies learned from machine learning to develop high-performing perovskite solar cells. *Advanced Energy Materials, 9*, 16099.
104. Odabasi, C., & Yildirim, R. (2019). Performance analysis of perovskite solar cells in 2013–2018 using machine-learning tools. *Nano Energy, 56*, 770.
105. Milkinson, M., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data, 3*, 160018.
106. Lundberg, S. M., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell. 2*, 56–67.