High-throughput computations and machine learning for halide perovskite discovery

Jiaqi Yang and Arun Mannodi-Kanakkithodi*

Halide perovskites are materials of considerable interest for solar cells, photodiodes, LEDs, photocatalysis, and photorechargeable batteries. One of the most attractive features of this class of materials is the sheer tunability of their stability, electronic bandgaps, optical absorption behavior, and defect properties, via composition engineering, phase transformation, change in dimensionality, surface and interface engineering, and octahedral rotation and distortion. Due to the ease of simulating well-defined crystal structures and systematically investigating compositional and structural factors that affect their properties, first-principles-based density functional theory (DFT) computations are frequently used for studying halide perovskites, leading to high-throughput data sets, screening of promising materials, and training of machine learning (ML) models for accelerated prediction and optimization. In this article, we present an overview of computational data-driven discovery of novel halide perovskites using some examples from the literature we believe best represent success in this field. Specific methods used for prediction of properties, optimization and screening across large chemical spaces, and automated design of novel structures and compositions, are highlighted. DFT-ML-based design frameworks have been instrumental in expanding the pool of stable halide perovskites with desired optoelectronic properties and will continue to inform new discovery in close synergy with targeted experiments.

Introduction

The perovskite structure is ubiquitous and has been extensively investigated by materials scientists. In its simplest form, it is represented by a three-component ABX_3 formula unit where A and B are cations with different oxidation states while X is an anion, leading to 3D networks of BX_6 octahedral units held together in place by large A-site cations. While oxide perovskites were of great interest for much of the latter half of the 20th century, it's halide perovskites that incite the most curiosity in this day and age, having risen to prominence not two decades ago as materials of great promise for solar-cell absorption.[1–4] ABX_3 halide perovskites (HaPs) contain halogens such as I and Br as *X*, divalent cations such as Pb and Sn as *B*, and large monovalent cations as *A*, which can either be inorganic (e.g., Cs, Rb) or organic (e.g., methylammonium or MA, formamidinium or FA). $MAPbI_3$ and $FAPbI_3$ are two commonly studied hybrid organic–inorganic perovskites (HOIPs) and have shown power-conversion efficiencies between 20 and 25% when used as absorbers in single- or multi-junction solar cells.[5,6]

There is a tremendous body of published literature already on experimental and computational design of HaPs, both purely inorganic and hybrid, for a variety of optoelectronic applications, including photovoltaics (PV), photodiodes, lasers, and LEDs. Studies have also recommended HaPs for transistors, power electronics, photon-activated semiconductor catalysts, photorechargeable batteries, and even quantum information sciences.[7,8] Fundamentally, a perovskite structure is only considered stable if the A-site, B-site, and X-site ionic radii fulfill the well-known tolerance (t) and octahedral (μ) factors.[9] Not only may novel HaPs be designed so as to satisfy the stability factors, but metastable phases could also be isolated, which improve significantly upon properties. The broad range of electronic, optical, and defect behavior that can be achieved with HaPs is owed to the various ways in which they can be engineered:

Jiaqi Yang, Materials Engineering, Purdue University, West Lafayette, USA; yang1494@purdue.edu
Arun Mannodi-Kanakkithodi, Materials Engineering, Purdue University, West Lafayette, USA; amannodi@purdue.edu
*Corresponding author

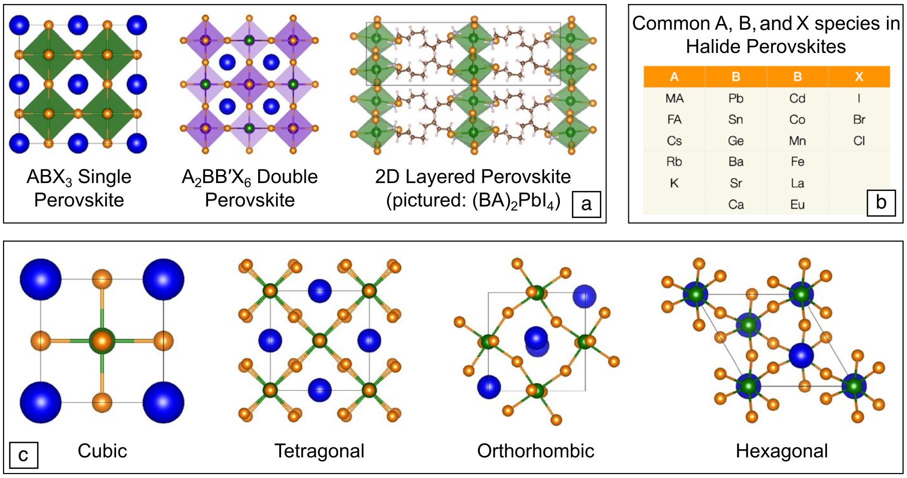
doi:10.1557/s43577-022-00414-2

- Composition: A majority of HaPs recommended as PV absorbers consist of some mix of MA, FA, and Cs at the A-site, primarily Pb at the B-site with minor fractions of other divalent cations such as Sn and Ge, and I or Br at the X-site often with little Cl. Since complex alloys can be generated within these chemical spaces with properties markedly different from "end-point" compositions, there has been extensive work on discovering and mixing novel organic molecular cations at the A-site, as well as replacing or partially substituting Pb at the B-site with other Group IV, Group II, or transition elements.[10–13] Whereas A-site mixing helps improve structural robustness and general stability to degradation, B-site and X-site mixing can help tune bandgaps and optical absorption.
- Structure/phase/dimensionality: Apart from the well-known cubic phase, ABX_3 perovskites exist in tetragonal, orthorhombic, or hexagonal phases,[14] or completely new structures discovered from evolutionary or minima hopping algorithms that still preserve corner-shared octahedral networks.[15] A-B-X perovskites may also manifest in the $A_2BB^{'}X_6$ double perovskite structure[16] or as layered 2D perovskites such as the $(L)_2A_{n-1}B_nX_{3n+1}$ Ruddlesden–Popper (RP) phase or the $(L)A_{n-1}B_nX_{3n+1}$ Dion-Jacobson (DJ) phase, where L is a large organic spacer molecule.[17] Different types of perovskite structures and phases are pictured in **Figure 1**, along with common A-, B-, and X-site species.
- Polymorphism: A surprisingly vast array of property behaviors can be accessed in the perovskite chemical space using (meta)stable structures resulting from rotations of organic molecules, octahedral distortions and tilting, short- or long-range ionic ordering in mixed compositions, and reoptimization of known structures in larger supercells.[18,19]
- Defects: Vacancies, self-interstitials, or impurities may be spontaneously present within HaPs, or intentionally introduced to change the equilibrium conductivity and solar absorption.[20,21]

First-principles-based density functional theory (DFT) computations have been systematically performed to examine each of the previously mentioned factors and how they contribute to desirable or undesirable optoelectronic properties of HaPs. DFT is now reliably applied for determining lattice parameters, heat of formation or decomposition, bandgaps, optical absorption spectra, and defect formation energies of a variety of HaPs, with mixed accuracy compared to experiments.[2,21] Although semi-local GGA-PBE reproduces stability and structure quite well, it is generally known to underpredict the bandgaps of solids as compared to nonlocal hybrid HSE06 functional or beyond-DFT GW approximation.[22] However, for HaPs, it is observed that although HSE06 works reasonably well for purely inorganic compounds, PBE bandgaps are often accidentally more accurate for many HOIPs containing Pb or Sn if spin–orbit coupling (SOC, important due to the relativistic effects of heavy atoms) is neglected.[21] This effect often holds true for defect formation energies as well.[20,23] HSE06 + SOC calculations are much more expensive, especially if involving full geometry optimization, but certainly more accurate for electronic and optical properties. Furthermore, PBE often needs to be replaced with methods like PBEsol (improved PBE for solids)[24] and PBE-D3 (to account for weak dispersion interactions when organic species are present)[25] for better structure optimization before being coupled with HSE06 to accurately reproduce the bandgaps.

Although high-throughput DFT (HT-DFT) for HaPs is often limited by the computational expense of a suitably accurate level of theory, it is the method of choice for materials scientists to generate data sets of the properties of HaPs, and mine them to uncover important chemical design rules, determine promising compositions for experiments, and perhaps more crucially, couple with state-of-the-art machine learning (ML) or artificial intelligence (AI) techniques that yield models for instant prediction and optimization, negating the need for endless computation. Indeed, the burgeoning field of materials informatics has seen many success stories to date with discoveries of new battery materials, capacitor dielectrics, high-entropy alloys, solid-state catalysts, etc.,[26,27] and has been instrumental in substantially accelerating the design of novel HaP compositions and structures as well.[28,29] Typically, such an approach would involve an ML regression, classification, or deep learning model trained on DFT data generated for representative compositions, using input descriptors that uniquely identify every data point, finally yielding as output the properties of interest with acceptable error bars. As discussed earlier, although DFT-level estimates will not always match experimentally measured values, predictions over various phases/polymorphs from multiple functionals will correlate well with experiments and help bridge the gap. DFT-ML will further accelerate such predictions and also provide pathways for inclusion of experimental data points within a multi-fidelity learning framework.

In the remainder of this article, we document some glittering examples of HT-DFT and ML used for screening and optimization of HaPs, including the authors' past and ongoing work. A general framework for DFT-ML-based perovskite screening is pictured in **Figure 2**a, while an outline of discovery via forward prediction and inverse design is shown in Figure 2b. We divide computational screening efforts into two broad categories based on the order of magnitude 10^n of HT-DFT data points being studied: small data set (n ~ 1–2) and large data set (n ~ 3–5). We emphasize the main DFT methods being used, chemical spaces being studied in terms of specific cation and anion species and perovskite structures/phases, and important trends and insights gained from (small


ABX_3 Single Perovskite
$A_2BB'X_6$ Double Perovskite
2D Layered Perovskite (pictured: $(BA)_2PbI_4$)
a
Common A, B, and X species in Halide Perovskites
b
c
Cubic
Tetragonal
Orthorhombic
Hexagonal

A	B	B	X
MA	Pb	Cd	I
FA	Sn	Co	Br
Cs	Ge	Mn	Cl
Rb	Ba	Fe	
K	Sr	La	
	Ca	Eu	

Figure 1. (a) Representative crystal structures of single, double, and 2D layered perovskites. (b) Common organic and inorganic species used at the A-, B-, and X-sites in halide perovskites (HaPs). (c) Prototypical perovskite phases: cubic, tetragonal, orthorhombic, and hexagonal.

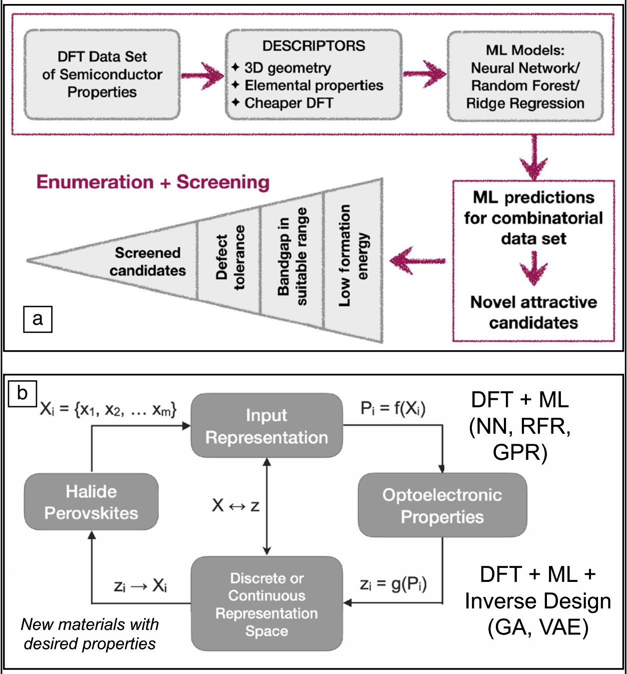

DFT Data Set of Semiconductor Properties
DESCRIPTORS
✦ 3D geometry
✦ Elemental properties
✦ Cheaper DFT
ML Models: Neural Network/ Random Forest/ Ridge Regression
Enumeration + Screening
Screened candidates
Defect tolerance
Bandgap in suitable range
Low formation energy
ML predictions for combinatorial data set
Novel attractive candidates
a
b
$X_i = \{x_1, x_2, \ldots x_m\}$
Input Representation
$P_i = f(X_i)$
DFT + ML (NN, RFR, GPR)
Halide Perovskites
$X \leftrightarrow z$
Optoelectronic Properties
$z_i \rightarrow X_i$
Discrete or Continuous Representation Space
$z_i = g(P_i)$
DFT + ML + Inverse Design (GA, VAE)
New materials with desired properties

Figure 2. (a) A typical density functional theory-machine learning (DFT-ML) training and screening framework, shown as an example for halide perovskites as solar-cell absorbers. This is a general approach followed in References 21, and 30–32. (b) An outline depicting forward property prediction and inverse design of new materials with desired optoelectronic properties; such an approach is followed in References 33 and 34. NN = neural network, RFR = random forest regression, GPR = Gaussian process regression, GA = genetic algorithm, and VAE = variational autoencoder.

or large) data, which inform subsequent experimental validation and/or ML models for enhanced discovery. Further, we present examples of "DFT-ML"-based prediction and screening across large HaP search spaces, again highlighting the main algorithms (e.g., random forests and neural networks) being applied. A more detailed discussion of our work in this area is presented, and some examples using more complex ML approaches such as deep learning and generative/inverse design are discussed. We end with an outlook on the importance of HT-DFT and ML in discovering better HaPs and how they should be synergistically performed with experimental synthesis and characterization.

High-throughput computational screening

Small data set

In this section, we focus on studies dealing with DFT data sets for a variety of HaPs that are of the order of 100 or less compounds. In such studies, the emphasis is on obtaining computational predictions for properties of some representative compounds and on unraveling some important trends that may have been missing from previous studies and could inform larger computational and experimental efforts. A classic example of screening across a small DFT data set is the work by Kar et al.,[14] wherein 30 $MABX_3$ compositions were studied in cubic, tetragonal, and orthorhombic phases. Screening across the 90 compounds using t and μ estimates along with bandgaps computed from both PBE and HSE06, including van der Waals (vdW) interactions and SOC, yielded nine suitable compounds; this process is pictured in **Figure 3**a, along with the chemical space being studied. The same authors followed up this study with replacing MA by Cs, Rb, and K[35] and determined some new purely inorganic perovskites with suitable bandgaps.

Similarly, Jiang et al.[36] used DFT to screen $MABI_3$ HOIPs in three phases with 27 options for cations B, leading to 10 compositions with nonzero bandgaps, some of which were validated via experimental synthesis and characterization. DFT computations involved geometry optimization using PBE followed by bandgap estimation from the GLLB-SC potential–another advanced functional for better electronic properties.[37] Further, for 90 HOIP iodides, bromides, and chlorides containing MA, FA, or EA (ethylammonium), Pu et al.[38] performed vdW-corrected PBE-D2 geometry optimization and calculated bandgaps from HSE06; decomposition enthalpies and bandgaps are shown in Figure 3c. For selected Sn-based perovskites, the effect of

Cl-doping and Cu-doping at X and B sites, respectively, was further investigated and a number of stable compositions with suitable bandgaps were identified–this shows the importance of not just modifying A/B/X components, but exploring doping of unexpected elements to achieve interesting properties.

Yamamoto et al.[39] used cluster expansion and DFT to study the stability and charge distribution of Cs, MA, and FA-based iodide perovskite alloys with two elements at the B-site, leading to identification of stable mixed compositions. Ray et al.[40] studied Cs-I perovskites with several B-site elements, in cubic, tetragonal, and orthorhombic phases, using multiple DFT functionals, including PBE-D3, PBEsol, and HSE06, with and without SOC. Investigation of bandgaps and formation energies from different methods allowed appropriate benchmarking for different B-site cations and enabled synthesis of Ba and Sr-based perovskites. Liu et al.[41] used DFT to study several organic molecules and Cs as A-site cations in mixed Si-Ge iodide HaPs, with structure optimization using optB86b-vdW and bandgaps from HSE06 + SOC, with HSE mixing fraction, usually kept fixed at 25%, changed to 50–60%–introducing another "parameter" that may be tuned for better property estimates. $DAGeI_3$ (DA = diamine) and $CsGe_{0.67}Si_{0.33}I_3$ were discovered as stable new compounds with suitable bandgaps.

Large data set

In this section, we discuss literature that involves extended DFT data sets, ranging from hundreds to hundreds of thousands. Such data sets are conducive to the application of ML, covering much wider structure-composition spaces, and gaining universal insights that can transform HaP design and discovery. There has been considerable HT-DFT effort in evaluating phase stability and determining metrics for general perovskite stability that improve upon t and μ. Sun et al.[42] used DFT to study 138 halide and chalcogenide single and double perovskites, and found that the quantity $(\mu + t)^{\eta}$, where η is the atomic packing fraction, correlates highly with decomposition energy and can be used as a proxy for perovskite stability: this correlation is pictured in Figure 3b. Similarly, Bartel et al.[9] used a data set of thousands of experimentally known perovskites and proposed a new stability criterion that considers ionic radii of A, B, and X species.

Castelli et al.[43] present a large DFT data set of 240 hybrid and inorganic ABX_3 HaPs considering Cs, MA, or FA as *A*, Pb, or Sn as *B*, and a mix of Cl, Br, and I at the X-site, in four phases. Structure optimization was performed using PBEsol, while bandgaps were calculated using GLLB-SC + SOC, including an electron–hole interaction term from a Bethe–Salpeter equation (BSE) calculation. This work revealed correct combinations of A/B cations, perovskite phase, and X-site mixing, which yields a PV-suitable bandgap and good stability. Mao et al.[44] used HT-DFT to study 260 purely inorganic ABX_3 HaPs, where A could be any alkali metal, B could be Pb, Sn, or Ge, and X could be F, Cl, Br, or I. Structure optimization was performed using PBE and all bandgaps were computed using GLLB-SC + SOC. This study has a partial chemical space overlap with the study by Castelli et al.,[43] but revealed novel stable Na, K, and Rb-based HaPs with bandgaps in the ~1.4-eV range. Both References 43 and 44 report accurate computations for a large data set of important HaP compositions and crucially, provide data that can be (and have been) utilized in follow-up studies for more computations and training ML models.

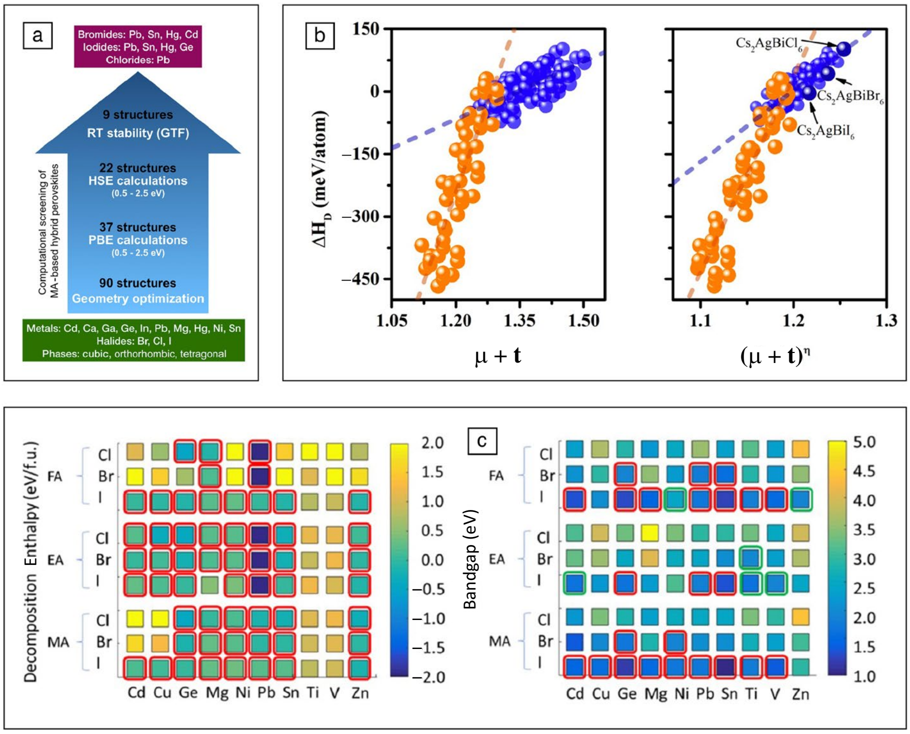


Figure 3. (a) Screening procedure using Perdew–Burke–Ernzerhof (PBE) and Heyd–Scuseria–Ernzerhof (HSE) bandgaps followed by tolerance factor to determine nine suitable halide perovskites (HaPs), applied by Reference 14. (b) Correlation between suggested perovsite stability factors and density functional theory (DFT) computed decomposition energies of halide and chalcogenide perovskites.[42] (c) Visualization of DFT computed decomposition energies and bandgaps across 90 HaPs.[38] RT, room temperature; GTF, Goldschmidt tolerance factor. Rights for reproducing figures have been obtained from the American Institute of Physics, American Chemical Society, and Elsevier.

In one of the most comprehensive HT-DFT studies of the last few years, Körbel et al.[45] swept through all electropositive elements in the periodic table and considered them as A- and B-site cations for halide, oxide, chalcogenide, and nitride perovskites. Spin-polarized PBE was used for geometry optimization and HSE06 (without SOC) for calculating band structure and electron/hole effective masses. From a set of >32,000 ABX_3 combinations, ~200 stable perovskites were found, many of which are completely novel; compounds with suitable bandgaps and low effective masses were screened for PV, and other materials with interesting ferroelectric properties were identified. Using the "K-computer," Nakajima et al.[46] studied >11,000 single and double HaPs where A is Cs, MA, or FA, X is I, Br, or Cl, and B cations are chosen from across Groups I, II, III, IV, V, and transition rows. Coarse-level bandgap screening from PBE was followed by HSE06+SOC calculations on ~2000 compounds for more accurate bandgaps; valence and conduction band edges were further estimated empirically, whereas hole and electron effective masses were obtained from PBE. The authors identified 51 Pb-free "low-toxic" HaPs with suitable properties for solar-cell absorption. Zhang et al.[47] used PBE computations to evaluate the energy above hull (with respect to halides collected from the Materials Project) and bandgaps of 980 double HaPs, and screened 112 stable compounds with narrow to wide bandgaps (from an HSE06-level estimate) that may be suitable for a range of applications.

Machine learning on computational data

DFT-ML-based prediction and screening

In this section, we describe how large DFT data sets such as those discussed in the previous section are used to train ML prediction or classification models, leading to screening across massive search spaces. In another example that exploits the universality of the perovskite octahedral arrangement, Park et al.[29] employed a DFT+ML approach to investigate the inherent (dis)similarity of various chalcogenide and halide perovskites. GGA-PBE calculations on 120 ABX_3 compounds, where A is an organic or Group I cation and B is a Group II or transition-metal cation, were performed to determine their energies above hull and to quantify their octahedral distortion, followed by GLLB-SC+SOC calculations for bandgaps. Random forest regression using t, μ, and tabulated elemental properties as inputs was able to predict bandgaps and octahedral distortion and reveal the descriptors of most importance to each property. Further, a reduced dimensional visualization of all compounds using octahedral and energy descriptors revealed ideal and distorted perovskite structures and their possible combinations, which could lead to stable mixed-ion perovskites with interesting properties.

Lee et al.[48] used a genetic algorithm approach to perform optimization in a search space containing >40,000 HOIPs in several phases containing a large variety of novel organic molecules, to determine the suitable candidates, which were simulated using PBE-D3 for structure, HSE06 for bandgap, and room-temperature *ab initio* molecular dynamics to test thermal stability. This study resulted in the discovery of 25 novel stable Pb-free HOIPs with suitable bandgaps and effective masses for light emission. In another classic example of DFT-ML applied to HaPs, Stanley et al.[30] used property density distribution functions ("property" referring to known quantities such as electronegativity and orbital radii) as inputs to train predictive models for bandgap and formation energy using a DFT data set of 344 compounds. Accurate predictions were made across thousands of new perovskite and perovskite-derived compounds and trends in the properties were examined.

In an example of using existing DFT data for ML, Gladkikh et al.[31] trained accurate predictive models for the bandgap of ABX_3 HaPs with the help of an ML technique suitable for small data sets called alternating conditional expectations (ACE) and known elemental properties as input. Models used the DFT data set reported by Körbel et al.[45] for training and were found to outperform traditional regression algorithms such as random forests and kernel ridge regression. Further, using a published DFT data set of 272 double HaPs[49] and four easily estimated quantities as descriptors, Yang et al.[50] trained an accurate bandgap prediction model from gradient boosting decision tree (GBDT) regression. Based on predictions across >16,000 compounds, 61 compositions were screened that satisfy perovskite stability criteria and have suitable bandgaps for solar absorption. PBE computations were further performed to examine the band structures, optical absorption, and thermal stability of many of these materials.

Gao et al.[28] trained XG-boost regression models using the double HaP DFT data set published by Zhang et al.[47] and supplemented with further computations, to predict the HSE06-level bandgap with higher accuracy as compared to methods such as support vector regression and neural networks: these results are captured in Figure 5a. Best models were used to screen across ~6000 possible compositions and perform further selected PBE and HSE calculations to determine some new materials, such as Na_2MgMnI_6 and K_2NaInI_6, with high thermal stability and PV-appropriate bandgap and absorption. Wu et al.[32] used ensemble ML models including gradient boosting, support vector, and kernel ridge regression, trained upon a DFT data set of 1300 HOIPs, to predict bandgaps of >38,000 ABX_3 compositions that fulfill charge neutrality and perovskite stability criteria. As shown in Figure 5b, this leads to 686 orthorhombic HOIPs with suitable bandgaps for solar absorption, 132 of which are stable and nontoxic and further investigated using extensive computations.

A case study: DFT-ML for HaP alloys

In our recently published and currently ongoing work, we performed high-throughput DFT computations for several important properties of mixed-composition HaPs and coupled this data set with several regression algorithms, including neural networks (NNs), random forests (RF), and Gaussian processes (GP), to train predictive models and screen across a search space two orders of magnitude larger than the size of the DFT data set. The DFT-ML screening framework employed in this work is captured in **Figure 4**, in terms of the number of compounds and A/B/X-site demographics at every level of screening. The motivation for this work comes from the fact that screening novel HaPs for PV absorption should involve complex alloys and not only the stability and bandgap, but other crucial objectives such as the likelihood of formation of point defects and a figure of merit or proxy for power-conversion efficiency (PCE) based on computed optical absorption. Keeping this in mind, we calculated a total of 16 different structural, energetic, electronic, optical, and defect (vacancy) properties, using both PBE and HSE06, of 229 pseudo-cubic ABX_3 HaPs with possible mixing at the B-site. Extensive visualization of this data set helped us understand (some expected, some novel) design rules such as having more Ba/Sr/Ca and more Cl increase the bandgap, having more MA and FA enhances resistance to decomposition, and many Cs or Rb-based HaPs may show high PV figures of merit but a lack of vacancy defect tolerance.

Based on this diverse and multi-objective DFT data set, we rigorously optimized NN models for every property, obtained prediction errors no greater than 5 to 10% for any property, and deployed them for predictions across ~18,000 HaP alloys, leading to screening of 392 candidates with sufficiently low decomposition energy, PBE bandgap between 1 and 2.5 eV, HSE06 bandgap between 1 and 3.5 eV, sufficiently high-vacancy formation energies with no charge transition levels in the bandgap (to avoid nonradiative recombination of charge carriers), and high PV figure of merit. This study helped identify desirable types of B-site mixing, and ongoing work involves experimental synthesis and characterization of screened compositions by experimental collaborators and a significant expansion of current DFT-ML capabilities to include A-site and X-site mixed alloys, completely new species at each site, other properties and phases, and inverse design models to automatically suggest new compositions with multiple desired objectives.

Deep learning, inverse design, and optimization

Now, we talk about the specific and challenging task of direct/automated recommendation of HaP compositions that are likely to possess attractive optoelectronic properties, typically predicted at the DFT-ML level. This process is often referred to as “inverse design” of materials and involves representation in a discrete or continuous space that allows efficient walks toward optimal data points, as depicted in Figure 2. Deep neural networks, such as convolutional neural networks (CNNs), generative adversarial networks (GANs), or variational autoencoders (VAEs), are utilized for such processes because they can work on large and complex data sets, and generally yield semi-continuous latent space representations that can be traversed for inverse design and optimization. Many non-NN approaches are also common for efficient search of materials, such as using Bayesian optimization, which would sequentially suggest new candidates and guide the search toward optimal materials that fulfill a single or multi-objective function.

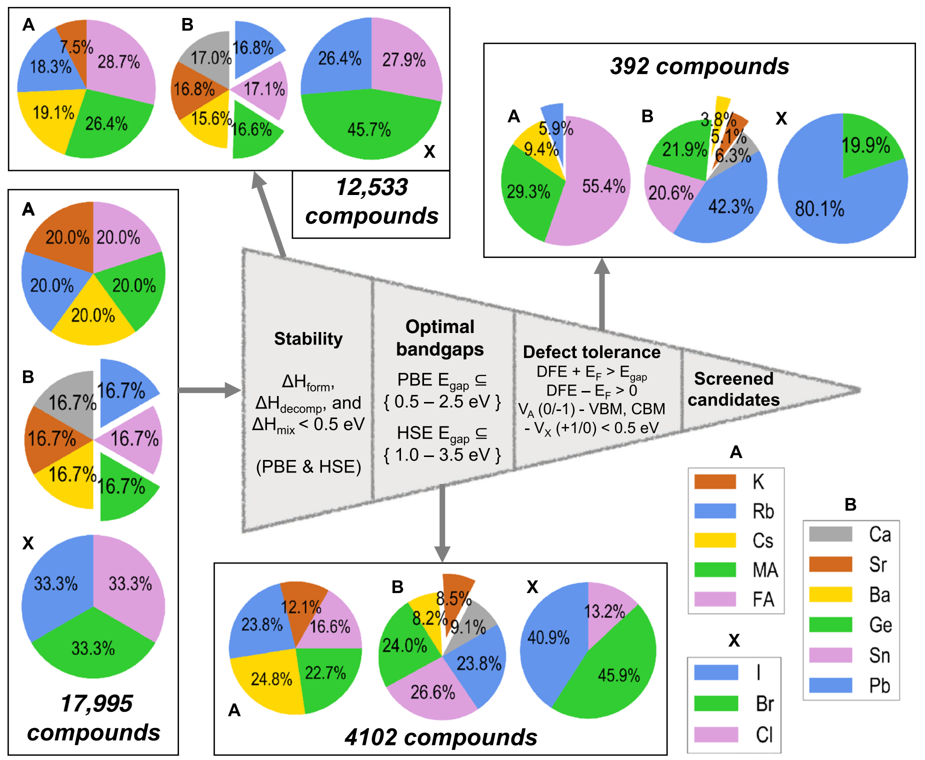


Figure 4. A density functional theory-machine learning screening framework applied in Reference 21 for obtaining 392 stable mixed-composition halide perovskites with suitable PBE and HSE06 bandgaps, defect tolerance, and high photovoltaic figure of merit, from a set of ~18,000 total compounds. Figure reproduced with permission from the Royal Society of Chemistry. DFE = defect formation energy, PBE = Perdew–Burke–Ernzerhof, HSE = Heyd–Scuseria–Ernzerhof, VBM = valence band maximum, and CBM = conduction band minimum.

Saidi et al.[33] used hierarchical CNNs—pictured in **Figure 5**c—to predict lattice constants, bandgaps, and octahedral tilt angles for HOIPs. The accuracy of bandgap prediction, in particular, is much higher than similar studies in the literature and is owed to careful and rigorous optimization of the deep CNN architecture. DFT data include PBE-optimized structures and bandgaps from

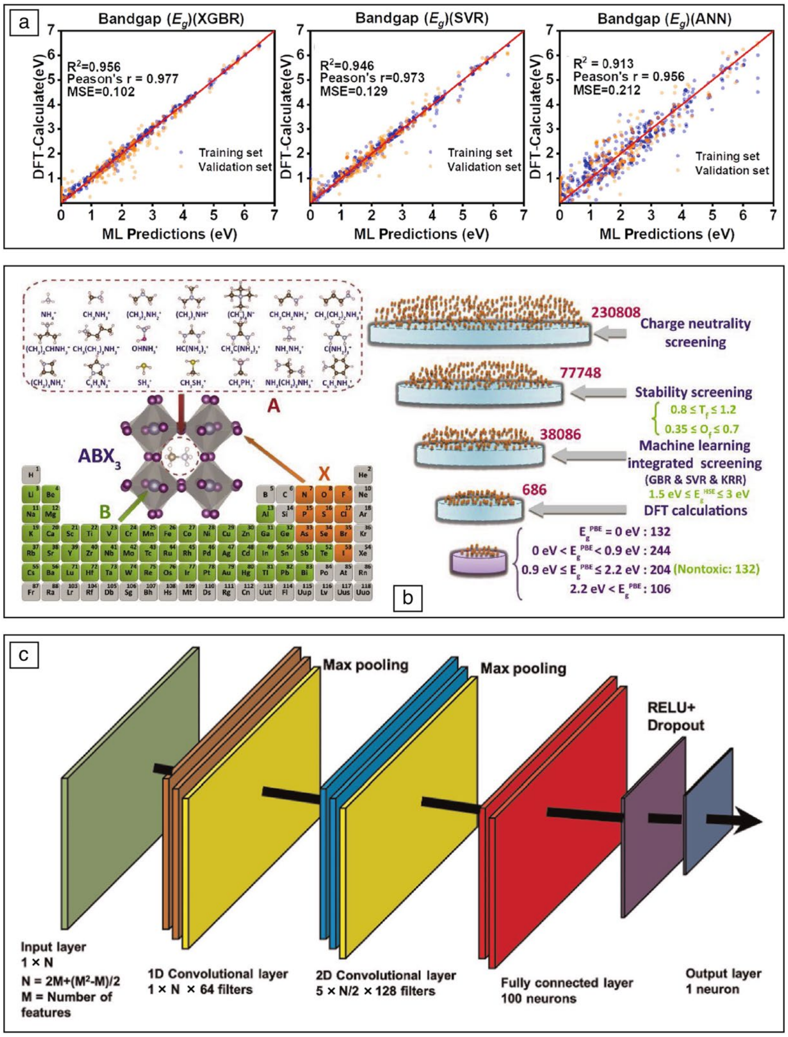


Figure 5. (a) Machine learning (ML) regression models trained for bandgaps of double halide perovskites (HaPs) by Gao et al.[28] (b) Density functional theory-machine learning (DFT-ML) screening methodology applied across a combinatorial ABX$_3$ space.[32] (c) Convolutional neural network architecture applied for HaPs by Saidi et al.[33] Figures reproduced with permission from Elsevier and Nature Publishing Group. XGBR = extreme gradient boosting regression, SVR = support vector regression, ANN = artificial neural network, MSE = mean squared error, KRR = kernel ridge regression, and RELU = rectified linear unit.

GLLB-SC + SOC, and other calculations are performed to determine the descriptors for ML. The automated design of periodic crystalline materials using deep NNs is still in its infancy; such automated structure design has been successfully carried out mostly for molecules.[51,52] However, Choubisa et al.[34] used a VAE, which contains multiple complex NN architectures for encoding and decoding, to design novel 2D and 3D HaP structures with desired stability and bandgap, based on an input representation called crystal site feature embedding (CSFE). Such a representation combined with DFT data enables highly accurate predictions as well as efficient exploration of a continuous latent space determined by the VAE to select novel points and invert them to entirely new mixed perovskite structures/compositions promising for UV and IR applications. Based on this study, small amounts of Cd-doping in HaPs were found to be desirable for solar absorption.

Based on a DFT data set of 438 single and double HaPs and tabulated elemental properties as descriptors, Chen et al.[53] used a Bayesian optimization model to efficiently search for new materials with a high unified figure of merit that combines decomposition energy and bandgap, to find new compositions with suitable properties much faster than random or brute-force searches. Herbol et al.[54] trained a Bayesian optimization model on DFT computed binding energies between (pure and mixed halide) ABX$_3$ HOIPs and various solvents, to efficiently search for perovskite–solvent combinations with optimal binding. Exhaustive computations were further performed to validate the findings for well-known perovskite compositions.

Outlook on synergistic discovery of halide perovskites

The overview of the DFT-ML HaP design literature presented in this article covers only a fraction of what is out there; naturally, we extend apologies to all the authors who were not acknowledged, and stress that this field is thriving at the moment with new publications every day addressing the data-driven optimization of HaPs. A table summarizing different ML techniques typically applied on perovskite DFT data sets is presented in the Supporting information, along with their advantages and disadvantages. The examples we covered show that there is already a gargantuan amount of DFT data that are (generally) available to the community and ready to be linked with rational experimental efforts. A DFT-ML perovskite recommendation engine can provide the impetus for automated synthesis and characterization of novel compositions, and their rigorous testing in devices or environments of interest, using robotic synthesis systems that are functional or currently being developed.[55,56] Such systems have the ability to quickly cycle through solvents, precursors, reference phases, etc., and measure important properties, which may also help improve the physical accuracy of DFT-ML predictions. DFT-ML predictions can further be integrated with experiments through active learning[57] and multi-fidelity learning[58] approaches; in the former, new experiments and computations will be performed for maximizing reward and minimizing uncertainty based on present ML models, whereas in the latter, properties from multiple DFT functionals and experiments will yield a combined data set for ML training and eventually making accurate predictions at the experimental level.

From the DFT-ML point of view, limitations that still need to be addressed include the exhaustive consideration of likely polymorphs and defects, and connecting DFT-level predictions with the ground truth, which includes experimental measurements affected by growth kinetics, temperature, etc. DFT computations ignore temperature, entropic contributions, finite size effects, and true randomness of ordering, but can be supplemented with *ab initio* thermodynamics and molecular dynamics (AIMD) simulations, which can further reveal the effects of molecular rattling and twisting, and octahedral rotations. Further, the quality of descriptors and choice of ML algorithms may yet provide some room for improvement. DFT data set sizes are also crucial: whereas small data sets ~10^2 points can provide good qualitative insights and screening, much larger data sets ranging from 10^3 to 10^4 points would generally be essential for accurate models based on composition and structure, applicable across wide chemical spaces. The accuracy of DFT-ML predictions will only ever be as good as the accuracy of the DFT data used for training and would need to be combined with experimental data within a multi-fidelity learning framework to achieve predictions at experimental fidelity.

Even in a field that appears to be saturated, there is tremendous interest and ongoing effort in further improving HaPs' stability against degradation, reducing toxicity, and enhancing efficiencies. We believe the discovery of next-generation HaPs hinges on the easy availability of all DFT-ML data described in this article and more, close connections between experiments, theory, and data science, and concentrated efforts toward optimal design.

Acknowledgments

A.M.K. acknowledges support from the School of Materials Engineering at Purdue University under Account No. F.10023800.05.002.

Data availability

No new data sets were generated or analyzed during this study.

Conflict of interest

On behalf of all authors, A.M.K. states that there is no conflict of interest.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1557/s43577-022-00414-2.

References

1. M.I.H. Ansari, A. Qurashi, M.K. Nazeeruddin, *J. Photochem. Photobiol. C* **35**, 1 (2018)
2. W.J. Yin, J.H. Yang, J. Kang, Y. Yan, S.H. Wei, *J. Mater. Chem. A* **3**, 8926 (2015). https://doi.org/10.1039/C4TA05033A
3. J.S. Manser, J.A. Christians, P.V. Kamat, *Chem. Rev.* **116**(21), 12956 (2016). https://doi.org/10.1021/acs.chemrev.6b00136
4. T.M. Brenner, D.A. Egger, L. Kronik, G. Hodes, D. Cahen, *Nat. Rev. Mater.* **1**(1), 15007 (2016). https://doi.org/10.1038/natrevmats.2015.7
5. P. Cui, D. Wei, J. Ji, H. Huang, E. Jia, S. Dou, T. Wang, W. Wang, M. Li, *Nat. Energy* **4**(2), 150 (2019)
6. M. Jeong, I.W. Choi, E.M. Go, Y. Cho, M. Kim, B. Lee, S. Jeong, Y. Jo, H.W. Choi, J. Lee, J.H. Bae, S.K. Kwak, D.S. Kim, C. Yang, *Science* **369**(6511), 1615 (2020). https://doi.org/10.1126/science.abb7167
7. S. Ahmad, C. George, D.J. Beesley, J.J. Baumberg, M. De Volder, *Nano Lett.* **18**(3), 1856(2018). https://doi.org/10.1021/acs.nanolett.7b05153
8. H. Huang, B. Pradhan, J. Hofkens, M.B.J. Roeffaers, J.A. Steele, *ACS Energy Lett.* **5**(4), 1107 (2020). https://doi.org/10.1021/acsenergylett.0c00058
9. C.J. Bartel, C. Sutton, B.R. Goldsmith, R. Ouyang, C.B. Musgrave, L.M. Ghiringhelli, M. Scheffler, *Sci. Adv.* **5**(2), eaav0693 (2019). https://doi.org/10.1126/sciadv.aav0693
10. S. Zhu, J. Ye, Y. Zhao, Y. Qiu, *J. Phys. Chem. C* **123**(33), 20476 (2019). https://doi.org/10.1021/acs.jpcc.9b04841
11. A. Banerjee, S. Chakraborty, R. Ahuja, *ACS Appl. Energy Mater.* **2**(10), 6990 (2019). https://doi.org/10.1021/acsaem.9b01479
12. J. Ding, S. Du, T. Zhou, Y. Yuan, X. Cheng, L. Jing, Q. Yao, J. Zhang, Q. He, H. Cui, X. Zhan, H. Sun, *J. Phys. Chem. C* **123**(24), 14969 (2019). https://doi.org/10.1021/acs.jpcc.9b03987
13. C. Greenland, A. Shnier, S.K. Rajendran, J.A. Smith, O.S. Game, D. Wamwangi, G.A. Turnbull, I.D.W. Samuel, D.G. Billing, D.G. Lidzey, *Adv. Energy Mater.* **10**(4), 1901350 (2020). https://doi.org/10.1002/aenm.201901350
14. M. Kar, T. Körzdörfer, *J. Chem. Phys.* **149**(21), 214701 (2018)
15. C. Kim, T.D. Huan, S. Krishnan, R. Ramprasad, *Sci. Data* **4**(1), 170057 (2017). https://doi.org/10.1038/sdata.2017.57
16. T.I. Al-Muhimeed, A. Shafique, A.A. AlObaid, M. Morsi, G. Nazir, M.M. AL-Anazy, Q. Mahmood, *Int. J. Energy Res.* **45**(13), 19645 (2021). https://doi.org/10.1002/er.7022
17. S. Yu, P. Liu, S. Xiao, *J. Mater. Sci.* **56**(20), 11656 (2021). https://doi.org/10.1007/s10853-021-06038-2
18. G.M. Dalpian, X.G. Zhao, L. Kazmerski, A. Zunger, *Chem. Mater.* **31**(7), 2497 (2019). https://doi.org/10.1021/acs.chemmater.8b05329
19. X.G. Zhao, G.M. Dalpian, Z. Wang, A. Zunger, *Phys. Rev. B* **101**, 155137 (2020). https://doi.org/10.1103/PhysRevB.101.155137
20. A. Mannodi-Kanakkithodi, J.S. Park, N. Jeon, D.H. Cao, D.J. Gosztola, A.B.F. Martinson, M.K.Y. Chan, *Chem. Mater.* **31**(10), 3599 (2019)
21. A. Mannodi-Kanakkithodi, M.K.Y. Chan, *Energy Environ. Sci.* **15**(5), 1930 (2022). https://doi.org/10.1039/D1EE02971A
22. M. Chan, G. Ceder, *Phys. Rev. Lett.* **105**(19), 403 (2010)
23. T. Shi, W.J. Yin, F. Hong, K. Zhu, Y. Yan, *Appl. Phys. Lett.* **106**(10), 103902 (2015). https://doi.org/10.1063/1.4914544
24. G.I. Csonka, J.P. Perdew, A. Ruzsinszky, P.H. Philipsen, S. Lebègue, J. Paier, O.A. Vydrov, J.G. Ángyán, *Phys. Rev. B* **79**(15), 155107 (2009)
25. S. Grimme, J. Antony, S. Ehrlich, H. Krieg, *J. Chem. Phys.* **132**(15), 154104 (2010). https://doi.org/10.1063/1.3382344
26. P. Schlexer-Lamoureux, K.T. Winther, J.A. Garrido-Torres, V. Streibel, M. Zhao, M. Bajdich, F. Abild-Pedersen, T. Bligaard, *ChemCatChem* **11**(16), 3581 (2019). https://doi.org/10.1002/cctc.201900595
27. J. Schmidt, M.R.G. Marques, S. Botti, M.A.L. Marques, *NPJ Comput Mater.* **5**(1), 83 (2019)
28. Z. Gao, H. Zhang, G. Mao, J. Ren, Z. Chen, C. Wu, I.D. Gates, W. Yang, X. Ding, J. Yao, *Appl. Surf. Sci.* **568**, 150916 (2021). https://doi.org/10.1016/j.apsusc.2021.150916
29. H. Park, R. Mall, F.H. Alharbi, S. Sanvito, N. Tabet, H. Bensmail, F. El-Mellouhi, *Phys. Chem. Chem. Phys.* **21**, 1078 (2019). https://doi.org/10.1039/C8CP06528D
30. J.C. Stanley, F. Mayr, A. Gagliardi, *Adv. Theory Simul.* **3**(1), 1900178 (2020). https://doi.org/10.1002/adts.201900178
31. V. Gladkikh, D.Y. Kim, A. Hajibabaei, A. Jana, C.W. Myung, K.S. Kim, *J. Phys. Chem. C* **124**(16), 8905 (2020). https://doi.org/10.1021/acs.jpcc.9b11768
32. T. Wu, J. Wang, *Nano Energy* **66**, 104070 (2019). https://doi.org/10.1016/j.nanoen.2019.104070
33. W.A. Saidi, W. Shadid, I.E. Castelli, *NPJ Comput. Mater.* **6**(1), 36 (2020). https://doi.org/10.1038/s41524-020-0307-8
34. H. Choubisa, M. Askerka, K. Ryczko, O. Voznyy, K. Mills, I. Tamblyn, E.H. Sargent, *Matter* **3**(2), 433 (2020). https://doi.org/10.1016/j.matt.2020.04.016
35. M. Kar, T. Körzdörfer, *Mater. Res. Express* **7**(5), 055502 (2020)
36. L. Jiang, T. Wu, L. Sun, Y.J. Li, A.L. Li, R.F. Lu, K. Zou, W.Q. Deng, *J. Phys. Chem. C* **121**(44), 24359 (2017). https://doi.org/10.1021/acs.jpcc.7b04685
37. F. Tran, S. Ehsan, P. Blaha, *Phys. Rev. Mater.* **2**, 023802 (2018). https://doi.org/10.1103/PhysRevMaterials.2.023802

38. W. Pu, W. Xiao, J. Wang, X. Li, L. Wang, *Mater. Des.* **198**, 109387 (2021). https://doi.org/10.1016/j.matdes.2020.109387

39. K. Yamamoto, S. Iikubo, J. Yamasaki, Y. Ogomi, S. Hayase, *J. Phys. Chem. C* **121**(50), 27797 (2017). https://doi.org/10.1021/acs.jpcc.7b07910

40. D. Ray, C. Clark, H.Q. Pham, J. Borycz, R.J. Holmes, E.S. Aydil, L. Gagliardi, *J. Phys. Chem. C* **122**(14), 7838 (2018). https://doi.org/10.1021/acs.jpcc.8b00226

41. D. Liu, Q. Li, H. Jing, K. Wu, *J. Phys. Chem. C* **123**(6), 3795 (2019). https://doi.org/10.1021/acs.jpcc.8b11695

42. Q. Sun, W.J. Yin, *J. Am. Chem. Soc.* **139**(42), 14905 (2017). https://doi.org/10.1021/jacs.7b09379

43. I.E. Castelli, J.M. García-Lastra, K.S. Thygesen, K.W. Jacobsen, *APL Mater.* **2**(8), 81514 (2014). https://doi.org/10.1063/1.4893495

44. X. Mao, L. Sun, T. Wu, T. Chu, W. Deng, K. Han, *J. Phys. Chem. C* **122**(14), 7670 (2018). https://doi.org/10.1021/acs.jpcc.8b02448

45. S. Körbel, M.A.L. Marques, S. Botti, *J. Mater. Chem. C* **4**, 3157 (2016). https://doi.org/10.1039/C5TC04172D

46. T. Nakajima, K. Sawada, *J. Phys. Chem. Lett.* **8**(19), 4826 (2017). https://doi.org/10.1021/acs.jpclett.7b02203

47. T. Zhang, Z. Cai, S. Chen, *ACS Appl. Mater. Interfaces* **12**(18), 20680 (2020). https://doi.org/10.1021/acsami.0c03622

48. B.D. Lee, W.B. Park, J.W. Lee, M. Kim, M. Pyo, K.S. Sohn, *Chem. Mater.* **33**(2), 782 (2021)

49. Y. Cai, W. Xie, Y.T. Teng, P.C. Harikesh, B. Ghosh, P. Huck, K.A. Persson, N. Mathews, S.G. Mhaisalkar, M. Sherburne, M. Asta, *Chem. Mater.* **31**(15), 5392 (2019). https://doi.org/10.1021/acs.chemmater.9b00116

50. Z. Yang, Y. Liu, Y. Zhang, L. Wang, C. Lin, Y. Lv, Y. Ma, C. Shao, *J. Phys. Chem. C* **125**(41), 22483 (2021). https://doi.org/10.1021/acs.jpcc.1c07262

51. A. Kadurin, S. Nikolenko, K. Khrabrov, A. Aliper, A. Zhavoronkov, *Mol. Pharm.* **14**(9), 3098 (2017). https://doi.org/10.1021/acs.molpharmaceut.7b00346

52. R. Gómez-Bombarelli, J.N. Wei, D. Duvenaud, J.M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T.D. Hirzel, R.P. Adams, A. Aspuru-Guzik, *ACS Cent. Sci.* **4**(2), 268 (2018). https://doi.org/10.1021/acscentsci.7b00572

53. X. Chen, C. Wang, Z. Li, Z. Hou, W.J. Yin, *Sci. China Mater.* **63**(6), 1024 (2020). https://doi.org/10.1007/s40843-019-1255-4

54. H.C. Herbol, W. Hu, P. Frazier, P. Clancy, M. Poloczek, *NPJ Comput. Mater.* **4**(1), 51 (2018). https://doi.org/10.1038/s41524-018-0106-7

55. R.E. Kumar, A. Tiihonen, S. Sun, D.P. Fenning, Z. Liu, T. Buonassisi (2021). *arXIV*:2110.03923 [cond.mat.mtrl-sci] (2021). https://doi.org/10.48550/ARXIV.2110.03923

56. Z. Li, M.A. Najeeb, L. Alves, A.Z. Sherman, V. Shekar, P. Cruz Parrilla, I.M. Pendleton, W. Wang, P.W. Nega, M. Zeller, J. Schrier, A.J. Norquist, E.M. Chan, *Chem. Mater.* **32**(13), 5650 (2020). https://doi.org/10.1021/acs.chemmater.0c01153

57. T. Lookman, P.V. Balachandran, D. Xue, R. Yuan, *NPJ Comput. Mater.* **5**, 21 (2019)

58. G. Pilania, J. Gubernatis, T. Lookman, *Comput. Mater. Sci.* **129**, 156 (2017). https://doi.org/10.1016/j.commatsci.2016.12.004 □

Jiaqi Yang is a senior PhD student in the Department of Materials Engineering at Purdue University. He received his BS degree in materials engineering from Shanghai Jiao Tong University, China. His research interests include interfacial electrocatalysts for fuel cell and semiconductors for optoelectronics. His current focus is on high-throughput screening and machine learning for discovery of halide perovskites. Yang can be reached by email at yang1494@purdue.edu.

Arun Mannodi-Kanakkithodi is an assistant professor in the Department of Materials Engineering at Purdue University. He received his PhD degree in materials science and engineering from the University of Connecticut in 2017, and worked as a postdoctoral researcher at the Center for Nanoscale Materials at Argonne National Laboratory from 2017 to 2020. His research involves using first-principles computational modeling, machine learning, and materials informatics to drive the design of new materials for energy-relevant applications. Mannodi-Kanakkithodi can be reached by email at amannodi@purdue.edu.