Please cite this article in press as: Ahmadi et al., Machine learning for high-throughput experimental exploration of metal halide perovskites, Joule (2021), https://doi.org/10.1016/j.joule.2021.10.001

Joule

CellPress

Perspective

Machine learning for high-throughput experimental exploration of metal halide perovskites

Mahshid Ahmadi,[1,*] Maxim Ziatdinov,[2,3] Yuanyuan Zhou,[4,5,*] Eric A. Lass,[1] and Sergei V. Kalinin[2,*]

SUMMARY

Metal halide perovskites (MHPs) have catapulted to the forefront of energy research due to the unique combination of high device performance, low materials cost, and facile solution processability. A remarkable merit of these materials is their compositional flexibility allowing for multiple substitutions at all crystallographic sites, and hence thousands of possible pure compounds and virtually a near-infinite number of multicomponent solid solutions. Harnessing the full potential of MHPs necessitates rapid exploration of multidimensional chemical space toward desired functionalities. Recent advances in laboratory automation, ranging from bespoke fully automated robotic labs to microfluidic systems and to pipetting robots, have enabled high-throughput experimental workflows for synthesizing MHPs. Here, we provide an overview of the state of the art in the automated MHP synthesis and existing methods for navigating multicomponent compositional space. We highlight the limitations and pitfalls of the existing strategies and formulate the requirements for necessary machine learning tools including causal and Bayesian methods, as well as strategies based on co-navigation of theoritical and experimental spaces. We argue that ultimately the goal of automated experiments is to simultaneously optimize the materials synthesis and refine the theoretical models that underpin target functionalities. Furthermore, the near-term development of automated experimentation will not lead to the full exclusion of human operator but rather automatization of repetitive operations, deferring human role to high-level slow decisions. We also discuss the emerging opportunities leveraging machine learning-guided automated synthesis to the development of high-performance perovskite optoelectronics.

INTRODUCTION

Metal halide perovskites (MHPs) are now one prominent focus of scientific interest due to their outstanding optoelectronic properties and low fabrication cost, offering tremendous promise for applications in photovoltaics (PVs),[1] light-emitting devices,[2] radiation sensors,[3–5] and many others. The standard three-dimensional (3D) MHPs possess the archetypical perovskite structure of ABX_3, where A, B, and X refer to monovalent organic/metal cations (e.g., methylammonium or MA^+, formamidinium or FA^+, and Cs^+), divalent metal cations (Pb^{2+} and Sn^{2+}), and halide anions (I^-, Br^-, and Cl^-), respectively. These compounds are among more than one thousand perovskites endmembers that have recently been predicted.[6–8]

Context & scale

Attainment of future energy sustainability calls for the development of large-scale electrical systems that can not only cheaply generate energy from renewable sources such as sun power but are also able to efficiently convert the electricity into desirable forms such as display applications. At the core of this prominent development is the discovery and optimization of semiconductor materials and devices. Metal halide perovskites (MHPs) are an emergent semiconductor family that has the potential to transform the current electrical and electronic systems. These materials exhibit a huge composition space, which offers an opportunity for exploring various properties and device applications, but meanwhile, this generates a grand challenge in the identification of suitable candidates for specific applications. The recent advances in machine learning-guided laboratory automation in materials synthesis and characterization have paved a promising way to address this challenge, but a further discussion and an in-depth thinking are required to enable more effective methods for high-throughput screening of MHP materials and to leverage these to develop high-performance energy devices such

Please cite this article in press as: Ahmadi et al., Machine learning for high-throughput experimental exploration of metal halide perovskites, Joule (2021), https://doi.org/10.1016/j.joule.2021.10.001

The characteristic aspect of these materials is the flexibility of the crystallographic structure, allowing for substitution on A, B, and X sites. For example, the simple combinatorial estimate by M. Saliba[9] shows combinations between 7 possible A site cations, 2 B site cations, and 3 halides, which can yield $(2^7 - 1) \cdot (2^2 - 1) \cdot (2^3 - 1) = 2{,}667$ possibilities. Beyond this, these materials form complex phase diagrams with varying limitations of solid solubility, multiple ferroic phases[10] with associated phase boundaries, etc. In addition to the 3D ABX_3 compounds, layered $A_2B^{IV}X_6$, and ordered $A_2B^{I}B^{III}X_6$ double perovskites have been explored.[11,12] Spatial confinement has been also discovered, as in 0D quantum dots (QDs) and 1D nanowires. Figure 1 illustrates an overview of the computationally calculated selection of desirable cation and anion substitutions in 3D ABX_3 MHPs with the Goldschmidt tolerance factor between 0.8 to 1,[13] and potential $B^{I}B^{III}$ metal cations in 3D $A_2B^{I}B^{III}X_6$ MHP structure.

However, despite extensive theoretical studies, only a handful of predicted compounds have been experimentally realized since doing so involves a complex and time consuming optimization cycle for synthesis and characterization. This has severely limited the discovery rate. For example, the toxicity of Pb cations in the MHPs can cause environmental problems, necessitating the search for alternative B cations (much like the search for Pb-free ferroelectrics for actuators[14,15]). The Pb-free MHPs including Sn based and Ge based MHPs have been extensively studied as alternatives.[16] Consequently, Pb-free halide double perovskites have been fabricated, such as Cs_2InAgX_6 and similar $A_2In^{(I)}M^{(III)}X_6$-based compounds.[17,18] Perhaps even more importantly, these vast compositional spaces contain tremendous opportunities for serendipitous discoveries of novel and improved functionalities.

More than one decade into the exploration of these materials, it has become obvious that the potential commercialization of MHP devices is limited by the long-term stability and responses to conditions such as light, humidity, and heat.[16,19] It has also been shown that alloying to form solid solutions offers a pathway to combat this issue and allow for development of more stable MHP photovoltaics.[20–23] However, the virtually infinite possible solid solutions, the fact that different aspects of figures of merit such as band gaps, chemical, structural, and thermal stabilities are optimized in the different regions of the phase diagrams, and the need for optimization of synthesis conditions for each specific composition make the classical search and optimization of these materials tedious and time consuming. To date, only a few tens of binary and ternary solid solution compositions have been explored[24–26] compared with over thousands of possible variants for 2- and 3-component systems[6–8] and hundreds of thousands of possible 4- and 5-component systems.

CATION AND ANION ALLOYING SUBSTANTIALLY EXPANDS THE COMPOSITION SPACE IN MHPs

While the exact mechanisms behind the stability improvements by alloying remain to be fully elucidated,[27] qualitative analogies can be drawn with other materials classes. From the thermodynamic viewpoint, doping on the multiple sites can increase stability of a solid solution if the enthalpy of mixing is negative, or by increasing the configurational entropy of the solid solution similar to the high entropy metal alloys and high entropy oxide perovskites.[28,29] In oxide materials, incorporation of multiple metal cations into single phase crystal structures has led to interesting novel and unexpected properties.[28] Correspondingly, it can be expected that mixed

as solar cells and light-emitting devices. From a fundamental viewpoint, the establishment of a solid link between machine learning-guided automated synthesis and device development will create a new paradigm of research, impacting the landscape of energy science.

[1]Joint Institute for Advanced Materials, Department of Materials Science and Engineering, The University of Tennessee Knoxville, Knoxville, TN 37996, USA

[2]Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

[3]Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

[4]Department of Physics, Hong Kong Baptist University, Kowloon, Hong Kong SAR, China

[5]Smart Society Lab, Hong Kong Baptist University, Kowloon, Hong Kong SAR, China

*Correspondence: mahmadi3@utk.edu (M.A.), yyzhou@hkbu.edu.hk (Y.Z.), sergei2@ornl.gov (S.V.K.)

https://doi.org/10.1016/j.joule.2021.10.001

Please cite this article in press as: Ahmadi et al., Machine learning for high-throughput experimental exploration of metal halide perovskites, Joule (2021), https://doi.org/10.1016/j.joule.2021.10.001

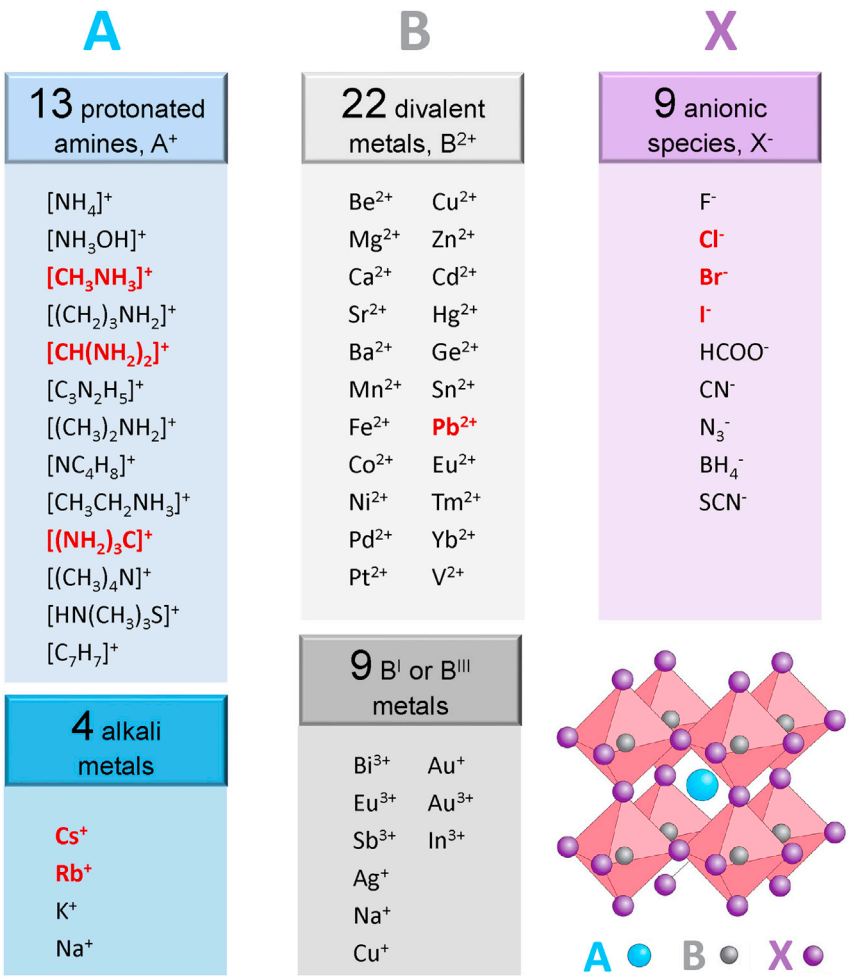

Figure 1. Compositional versatility of metal halide perovskites

An overview of possible calculated cation and anion substitutions in 3D ABX_3 MHPs with tolerance factor between 0.8 to 1[13] and potential $A_2B^IB^{III}X_6$ MHP structures. The red elements indicate the combination of cations and anions that are commonly explored.

cation effects in MHPs will lead to similar behaviors, with the additional effects not only on the crystallographic,[30] thermal,[30,31] moisture, and photostability[32,33] but also kinetic behaviors, phase stability, and ion migration. Some of these mechanisms have long been argued to underpin MHP functionalities, whereas others limit stabilities or performance.

From the viewpoint of kinetics, alloying can significantly reduce the ionic mobility, as is well known for the mixed alkali glass (MAE) effect in ionic conductors.[34] When more than two kinds of mobile ions are mixed in ionic conducting glasses and crystals, there is a non-linear decrease of the transport coefficients of either type of ions. While this effect created a major hurdle in ionic conductor materials, it can offer a pathway toward suppression of ion migration and phase segregation in MHPs. Recently, a combination of simulation and experimental studies revealed the suppression of iodide ion migration in $MAPbI_3$ by substitution of MA with a low concentration of seven different size cations including Rb^+, Cs^+, FA^+, Guanidium, dimethylammonium, and acetamidinium.[35] It was explained that cation substitution results in an increase in the activation energy of iodide diffusion. Furthermore,

Please cite this article in press as: Ahmadi et al., Machine learning for high-throughput experimental exploration of metal halide perovskites, Joule (2021), https://doi.org/10.1016/j.joule.2021.10.001

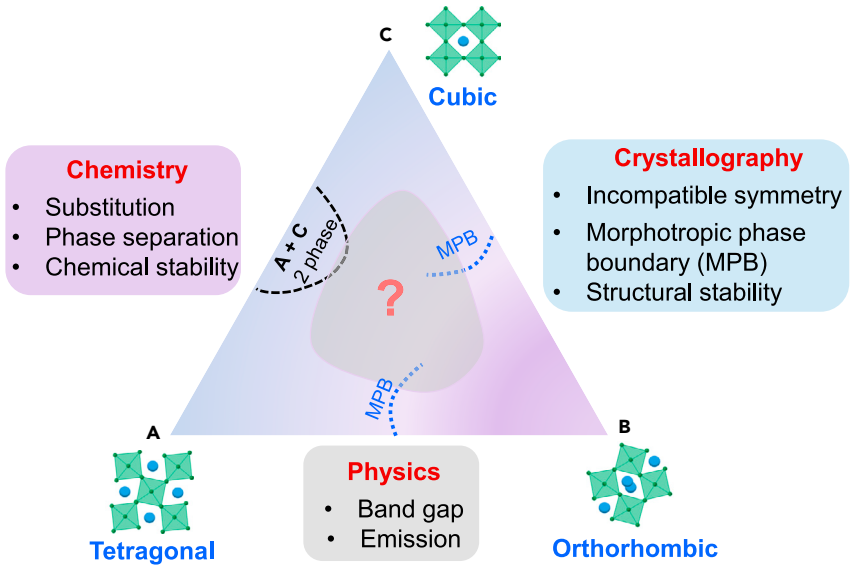

Figure 2. Structural and physicochemical phenomena in multicomponent MHPs

Schematic of unknown compositional regions in the phase diagram of a multicomponent MHP with possible two-phase regions, phase boundaries, and morphotropic phase boundaries (MPBs) with distinct structural, physical, and chemical behaviors from the endmembers which is difficult to predict via theoretical calculations, necessitating extensive experimental exploration. The additional degree of complexity is the potential capture of the solvent molecules by the lattice, leading to the emergence of the latent chemical variable.

chemical segregation and formation of the grain boundary phases, second phase inclusions, etc. can further lead to significant changes in macroscopic properties—either beneficial or detrimental to target functionalities.

These considerations have stimulated an extensive search for optimal compositions for specific applications, tailoring parameters such as stability, band gap, or photoluminescence (PL) quantum yield. Note that while several of these are (at least in principle) fundamental material properties, others are controlled by the disorder, defects density, and other complex to predict factors (see Figure 2). Similarly, in real materials additional and seldom recognized factors can be the presence of antisolvent molecules, complex local microstructures evolving due to phase separation or chemical instabilities, etc. Hence, materials optimization in these systems typically requires multiple experiments and optimization cycles even for a single composition. From the perspective of machine learning (ML), this creates hidden variables in the system (e.g., defect density) that can be affected by macroscopic processing conditions but are also sensitive or difficult to control by the environmental factors. As such, this creates latent variables and observational biases that need to be considered when applying ML models or need to be frozen when simple correlative models are used.

FACILE SOLUTION PROCESSABILITY OPENS THE DOOR TO AUTOMATED SYNTHESIS OF MHPs

Solution processing makes MHPs relatively easy to explore and offers the promise of low-cost, large-scale manufacturability. At the same time, it results in sensitivity toward the solvent choices and processing sequence when many initial components require different solvents and mixing sequence. Consequently, it can affect the composition, morphology, strain, and eventually the stability of the material.

Please cite this article in press as: Ahmadi et al., Machine learning for high-throughput experimental exploration of metal halide perovskites, Joule (2021), https://doi.org/10.1016/j.joule.2021.10.001

While simple in concept, experimental synthesis of the MHPs necessitates solving a very large number of engineering challenges, including stability with respect to solvents, solvent volatility, environmental sensitivity of the solutions and materials, and antisolvent and solvent addition pathways to avoid second phase precipitation. Commonly, antisolvent crystallization is an efficient solution-based method to produce these multicomponent MHP systems. To date, many studies have utilized manual trial-and-error approaches to determine which antisolvent is applicable for a particular perovskite system.[36–38] Automated experimentation (AE) has accelerated this process; however, few have investigated the effect of workflows regarding antisolvent engineering. Recently, Langner et al.,[39] have focused on the effect of a large number of antisolvents on single endmember systems; however, none have explored the effect of antisolvent engineering on the intrinsic stability of multicomponent MHPs in ambient conditions. In a recent study, we have utilized our previously developed synthesis workflow to examine how the choice of two antisolvents, toluene and chloroform, affect the stability of double cations and double halides perovskites over time in ambient condition.[40] Roughly 1,100 unique MHP compositions were synthesized by a pipetting robot and the stability was studied utilizing automated PL spectroscopy in ambient conditions for several hours. We developed an unsupervised ML technique using multivariate statistical analysis, specifically non-negative matrix factorization (NMF), to map the time- and compositional-dependent PL behavior of each combinatorial library using these specific antisolvents. Through the utilization of this workflow, we were able to effectively map the intrinsic stability of 1,100 unique MHP compositions depending on the specific antisolvent. This approach exemplifies how the automated synthesis workflow can be utilized to explore the materials processing with respect to the stability and can be further extended to explore possible dynamical processes, such as halide segregation, responsible for either the stability or eventual degradation as caused by the choice of antisolvent. The high-throughput study can finally demonstrate the vital role of antisolvents in the synthesis of high-quality multicomponent MHPs.

Overall, the solution-based synthesis of MHPs readily allows for the automated synthesis and discovery, enabled via fully automated chemical labs, microfluidic systems, or hybrid human-automated synthesis workflows. However, simple acceleration of synthesis and characterization is insufficient to compensate for the extreme dimensionality of the chemical and processing spaces, necessitating the development of capabilities for *in situ* and *ex situ* characterization and especially ML methods that can guide the synthesis process in the composition or processing parameter spaces. Next, we will overview the recent advances in automated synthesis and describe the opportunities and limitations of the current ML methods for navigation of synthesis spaces. We further discuss the emerging ML opportunities for guiding this process. Finally, we provide our perspectives on leveraging ML-guided automated synthesis to the development of perovskite device technologies.

PROMISING INSTRUMENTAL PARADIGMS HAVE BEEN DEVELOPED FOR AUTOMATED SYNTHESIS OF MHPs

The recent development in the experimental domains is the emergence of AE, where the artificial intelligence (AI)/ML methods are used both to enable automatization to reduce latency within a specific scientific domain application (i.e., make experiments faster) and to guide the discovery workflow (i.e., define the parameters and conditions of new experiments based on previous experimental results). Combination of these two elements gives rise to the concept of automated laboratories for accelerated discovery of new materials.

Please cite this article in press as: Ahmadi et al., Machine learning for high-throughput experimental exploration of metal halide perovskites, Joule (2021), https://doi.org/10.1016/j.joule.2021.10.001

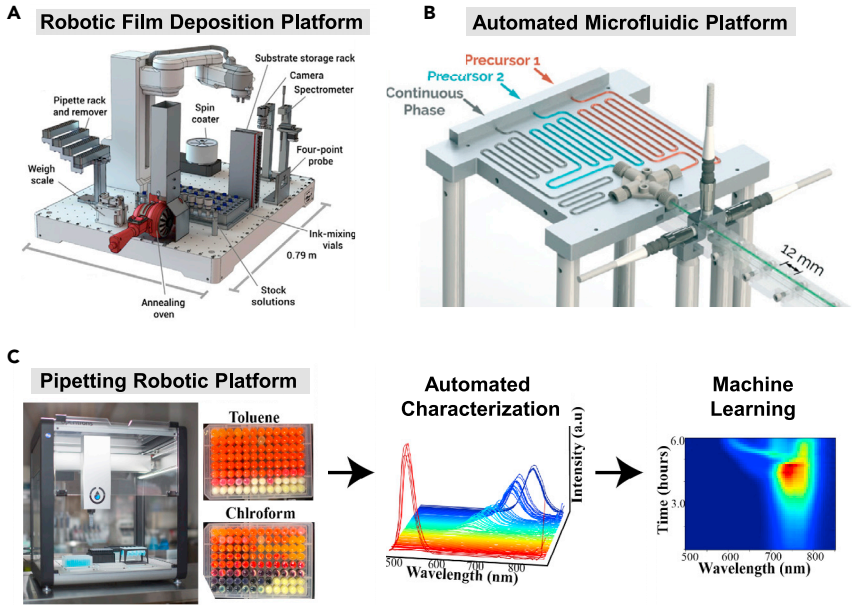

Figure 3. Paradigm of automated synthesis

(A–C) Schematics of (A) a self-driving laboratory for accelerated discovery of thin film materials,[49] (B) microfluidic system for accelerated synthesis of perovskite QDs,[52] and (C) pipetting robotic platform for high-throughput synthesis of MHP microcrystals, the PL spectra of 96 binary compositions and the ML analysis to effectively map the characteristic PL behaviors of the synthesized binary system.[40] (A) is adopted from Deeth et al.[49] with permission under the Creative Common licenses. (B) is adopted from Abolhasani et al.[52] with permission from the Royal Society of Chemistry. (C) is adopted from Ahmadi et al.[40]

Automated synthesis is rapidly becoming one of the fastest developing areas in materials science. This approach is rooted in the combinatorial synthesis that has broadly emerged in the late 90s[41–43] that allowed massive upscaling of synthesis capabilities. However, the broad propagation of this approach was limited by a characterization and physics discovery bottlenecks. While combinatorial libraries of many materials systems can be created, the studies of functional behaviors were performed one sample at a time. The exception was the materials with optical functionalities that allow for ready identification,[44] and to some extent magnetic materials.[41] A second direction was automated synthesis methods as e.g., developed by Cronin et al.[45,46] for solution-based organic materials, and Maruyama et al. for carbon nanotube growth.[47,48]

Currently, the development of automated experiment platforms in MHPs and other compounds for energy application follow three primary paradigms. One is the fully automated robotic labs as developed by e.g., Aspuru-Guzik et al. (Figure 3A)[49] and others.[39,50,51] Here, the human-based operations are substituted by the fully automated robotic handling. The synthesis process is controlled by Bayesian optimization (BO) workflow that can simultaneously optimize the optoelectronic properties by composition selection and processing parameters for thin film materials. The alternative is the microfluidic systems as e.g., developed by Abolhasani et al. (Figure 3B).[52,53,54] Here, using a modular microfluidic platform enables continuous manufacturing of inorganic MHP QDs guided by an ensemble neural network (ENN) exploration of the colloidal synthesis parameter space. Finally, significant acceleration of MHP microcrystals and QDs can be achieved by the combination of human and automated workflows, e.g., the recent studies based on

Please cite this article in press as: Ahmadi et al., Machine learning for high-throughput experimental exploration of metal halide perovskites, Joule (2021), https://doi.org/10.1016/j.joule.2021.10.001

micropipetting robots by us (Figure 3C)[55,56] and several other groups.[57,58–60] The experimental approach is much cheaper, still accelerated setup developed for synthesis of several hundred MHP samples per day and can be scaled up to 1–3 thousand samples/day throughput. This effort has recently shown the novelty of the autonomous research system and applicability of automated photoluminescent (PL) spectroscopy to characterize the ambient stability of MHP microcrystals[55] as well as QDs.[56] Compared with the previous studies, this approach not only develops a novel experimental workflow by utilizing a low-cost pipetting robot to create large combinatorial libraries of MHPs by an antisolvent method but also is one of the first to perform the systematic exploration of chemical and environmental stability on a material level.

Note that continuous combinatorial libraries like those used in pulsed laser deposition require composition characterization, since the local composition can significantly differ from the expected due to the specifics of the deposition process. The use of the well plate-based libraries addresses both issues. First, the compositions in this case are controlled by the pipetting/dilution process and endmember concentrations and as such are determined precisely (or at least as good as macroscopic synthesis). Second, using the well plate allows for characterization via optical spectrometry enabled by a multi-mode optical plate reader. Third, the properties of the synthesized microcrystals with the antisolvent precipitation approach have shown a good agreement with the properties of thin films deposited via antisolvent assisted crystallization.[25,55,58] Therefore, this approach can be reliably used to discover and optimize MHP thin films.

However, common to all three paradigms is the need for high-throughput characterization, including process monitoring and characterization of functionalities, and the need for feedback, where the composition and processing conditions are chosen based on characterization results and the iterative cycle is repeated toward the desired outcomes. In other words, it necessitates the active learning ML methods for navigation of composition and synthesis spaces. The questions we aim to address below include what these algorithms are expected to achieve, what are the existing ones and what are their limitation, and what are the future perspectives both from ML and workflow development perspectives.

MACHINE LEARNING ALGORITHMS BRIDGE THE GAP BETWEEN INSTRUMENTATION AND CONTROL, LEADING TOWARD EFFICIENT AUTOMATION IN MHPs SYNTHESIS AND CHARACTERIZATION

The proliferation of the automated synthesis methods and particularly the tremendous potential for the rapid growth in the field enabled by the availability of low-cost laboratory automation naturally requires development of control algorithms to navigate the composition and synthesis spaces. In some sense, these algorithms complement and potentially substitute the human-based decision-making process. However, human-based decision-making is a complex process that is based not only on the results of prior experiments but also on the general body of knowledge in the field available to an individual based on personal experience and communicated via scientific literature and interpersonal communication, and general physical principles (e.g., knowledge of basic thermodynamics and kinetics to name a few). Similarly, very often observation during the experiment (e.g., color change of the solution, etc.) are used to adjust synthesis condition or suggest early termination of an experiment. Hence, when discussing the ML algorithms for AE, we focus on similar considerations. Namely, (1) whether the decision-making process is based only on

Please cite this article in press as: Ahmadi et al., Machine learning for high-throughput experimental exploration of metal halide perovskites, Joule (2021), https://doi.org/10.1016/j.joule.2021.10.001

observed experimental results, or incorporates prior knowledge in some form, (2) whether the *in situ* observables, or more generally the hierarchy of observables are used, or the navigation is based on target functionalities only, and (3) whether the algorithm couples to a theoretical model, either known a priori or updated during the experiment. Next, we briefly overview some of the existing approaches and describe them in the light of the criteria above and formulate some of the related opportunities.

Bayesian optimization

Currently, the leading paradigm for the exploration of relatively low-dimensional parameter spaces is BO.[61–67] For MHP synthesis, the natural examples of such space will be the compositional space of the system, composition of antisolvent, or selection of ligands for nanoparticle growth. It is important to note that classical BO methods implicitly rely on the smooth changes of target functionalities within this parameter space (or the presence of only a small number of discontinuities) and requires relatively low-dimensional (< 6) parameter spaces to be effective. Hence, it can be readily adapted to exploration of compositional spaces, whereas adaptations to e.g., synthesis require additional dimensionality reduction steps.

The two key parts of the BO are the surrogate model that captures our prior beliefs about the (unknown) objective function and the acquisition function that trades off the exploration and exploitation. The most common choice for the surrogate model is the Gaussian process (GP). In general, GP refers to an approach for reconstructing a target function $f(\mathbf{x})$ over a certain parameter space $\mathbf{x}$ given the observations y_i at specific values $\mathbf{x}_i$.[62,64] For MHPs, the target function can be the band gap, PL intensity, or any other functionality of interest, whereas parameter space is composition of the material, antisolvent mixture, etc. Formally, this model is defined as $y = f(\mathbf{x}) + \varepsilon$, where $f \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$, with k as a covariance function (kernel), and ε represents Gaussian observation noise with variance s_n. This statement implies that the value of the function itself is unknown, but at each explored point the measured signal represents the true value of the function with noise.

The important element of the GP is the kernel function, defining the strength of connection between the points in the parameter space. The functional form of the kernel is chosen prior to the experiment and defines the physics of the explored phenomena. The numerical parameters of the kernel (e.g., length scale), as well as noise level, are inferred from the data during the GP model training, which is performed by maximizing the log marginal likelihood (more details about the Bayesian inference [BI] will be given in following sections).

Given a number of observations, the trained GP model seeks to reconstruct the values and uncertainties of the target function over the full parameter space. The predictive mean ($\overline{f_*}$) and variance ($\mathbb{V}[f_*]$) for a single new (test) point x_* are obtained with the trained GP model as.

$$\overline{f_*} = \mathbf{k}_*^\top \left(K + s_n^2 I\right)^{-1} \mathbf{y}, \qquad \text{(Equation 1)}$$

$$\mathbb{V}[f_*] = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top \left(K + s_n^2 I\right)^{-1} \mathbf{k}_* \qquad \text{(Equation 2)}$$

where $\mathbf{k}_*$ is the vector of covariance between the test point and n training points, and K is the n-by-n matrix of covariances evaluated at all pairs of training points. This step represents the basic GP prediction stage, sometimes also referred to as kriging.[68]

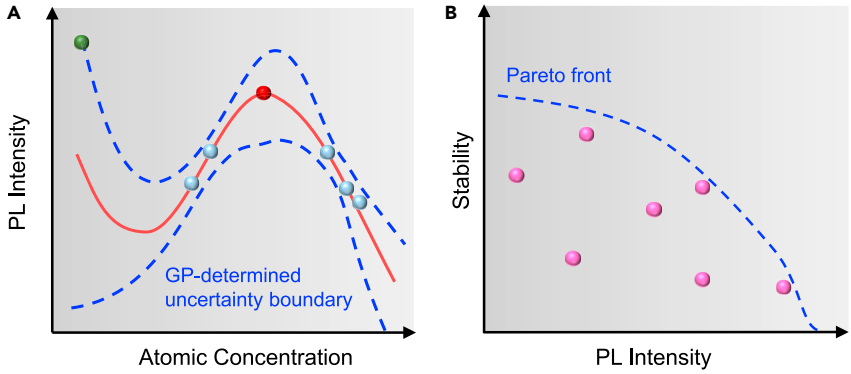

Figure 4. Framework for Bayesian optimization (BO)

(A) Schematic representation of the BO as applied to the optimization of the PL intensity. Here, the red curve illustrates the true (unknown) concentration dependence of PL intensity. The blue dots represent the points at which the measurements are taken, i.e., experimental data. The blue curves are the uncertainty boundaries determined by the GP. The red and green dots illustrate possible locations for subsequent experiment, chosen based on expected maximum and potential for unknown behaviors, respectively.

(B) Illustration of the Pareto front between the stability and PL intensity. Different points correspond for PL intensity-stability value pairs for different compositions. Pareto front shows the concentrations with the optimal balance between the two. The task for multiparameter BO is to map the Pareto front over the concentration space.

The GP-predicted mean and variance on all the test points are the interpolation between the training points and the measure of uncertainty, respectively. In other words, the GP process allows the estimation of the value of the target functionality across the whole parameter space, and how much it can be trusted. These parameters can serve as inputs into the acquisition function of BO which guides the exploration of the configuration space balancing the uncertainty-based exploration and the identification of the regions in this space where a desired property is maximized. In other words, the algorithm will explore the parameter space weighting the potential to discover behaviors of interest and degrees of uncertainty about this behavior. The common acquisition functions are the expected improvement (EI) over the current best results and the upper confidence bound (UCB) with provable cumulative regret bounds.[69]

To date, BO is the method of choice adopted in the automated synthesis[24,47,70,71] and other AE approaches, e.g., X-ray measurements or scanning probe microscopy,[72,73] as well as exploration of theoretical models and theory-experiment matching.[74] Note that the GP-based BO can also be readily adapted for exploring the vector function, i.e., multiple functionalities. In this case, the algorithm seeks to discover points offering an optimal balance between target functionalities, i.e., the Pareto front of the system (Figure 4). For example, exploring the concentration space for stability and PL intensity, the multi-objective BO will seek to find the regions where the pairs of these values lay on the outer shell of possible combinations.

AUTOMATED SYNTHESIS OF MHPs NECESSITATES THE DEVELOPMENT OF ADVANCED ML TOOLS BEYOND SIMPLE BAYESIAN OPTIMIZATION

The classical BO as described above has a number of significant limitations. First and foremost, this method does not utilize any prior information available about the

Please cite this article in press as: Ahmadi et al., Machine learning for high-throughput experimental exploration of metal halide perovskites, Joule (2021), https://doi.org/10.1016/j.joule.2021.10.001

system. Second, the experiment is optimized given only the final results and does not utilize the additional proxy information available during the experiment (e.g., color change during synthesis). Finally, and most importantly, it builds a fully non-parametric model, whereas materials properties tend to allow for (often simple) analytical functional approximations over the parameter space. Jointly, these considerations necessitate fairly low dimensionality of control parameter space. Next, we discuss some of the emergent trends in ML that allow one to relax some of these limitations, allowing for physics-based Bayesian modeling.

Bayesian optimization with "informed" kernels

The classical GP is completely defined in terms of its mean function and covariance function. The latter defines the strength of correlations between the properties in the adjacent points in compositional space, e.g., how close will be the PL intensity and the band gap be for adjacent compositions. Classical BO relies on the off-the-shelf kernel functions (radial basis function [RBF], Matern, etc.) that are the same across the whole composition space. This clearly is a significant limitation, since the properties tend to change slowly within the single solid solution region, and rapidly at the phase boundaries. Furthermore, GP with standard RBF kernels have inductive biases toward very simple solutions.

However, new opportunities emerge at the interface between deep learning and the BO via the deep kernel learning (DKL) approach.[75] In DKL, the kernel function is learned from the data using a neural network (hence, it is sometimes referred to as an "informed kernel"). Technically, the data (e.g., PL spectrum) are embedded into the (latent) feature space by a feedforward neural network, reducing it to a small number of descriptors. These descriptors are then used as input to a spectral mixture base kernel. The parameters of the base kernel and the weights of the neural network are trained jointly by maximizing the log marginal likelihood of the GP, yielding the non-parametric model of the system that can be used for BO. Note that this approach is reminiscent to a previously used approach combining the non-NMF and GP exploration[55]; however, in DKL the optimal features are discovered during the experiment, rather than engineered prior to the experiment. Due to this flexibility, the DKL GP could be well suited for non-stationary data and data with a complex hierarchical structure. Finally, whereas the classical GP-BO is limited to relatively low-dimensional spaces ($D \lesssim 6$), the DKL approach allows for performing optimization in a potentially high-dimensional space via learning the low-dimensional embeddings of the data.

In addition, the GP mean function can be used to capture trends in the data. It can be learned from the data with a neural network,[76] which is somewhat similar to the DKL approach. The learned mean function can be used for transfer learning and meta-learning with GP.[76]

GP/BO with phenomenological model

The distinctive aspect of the physics of many condensed matter systems is the existence of a large number of simple phenomenological relationships between the control parameters and target functionalities. Well-known examples include the compositional dependence of band gap and molar volume in solid solutions, thickness dependence of the switching voltage and domain size in ferroelectric materials,[77] or size dependence of plasticity. In some cases, these relationships can be derived from simple physical models; in others, they belie the complexity of underpinning mechanisms while still providing convenient approximations.

Please cite this article in press as: Ahmadi et al., Machine learning for high-throughput experimental exploration of metal halide perovskites, Joule (2021), https://doi.org/10.1016/j.joule.2021.10.001

The existence of such universal relationships opens the pathway to extend the BO approach by simultaneously exploring the parameter space of the system and discovering (or improving) the phenomenological model of materials behavior. From a technical standpoint, this approximate physical model of the system, e.g., compositional dependence of the band gap, becomes the GP mean function. In this manner, the prediction of the functionality across the composition space is based not only on the experimental data only, but also on expectation of what the physics of the system is. Such an augmentation of GP with a probabilistic model of the expected system's behavior would allow making "physics-informed" decisions about which points in the parameter space to evaluate next. In this case, one may start with a flexible GP (non-parametric regime) and gradually switch to a GP augmented with a structured probabilistic model (semi-parametric regime), although a more sophisticated interplay between the non-parametric and semi-parametric regimes is possible (e.g., explore the initial system behavior, discover the physical model, and then explore the deviations from such a model).

Causal experiments

The fundamental limitation of the classical BO methods, and in fact the vast majority of extant ML methods, is their correlative nature. For example, in classical supervised learning the neural network serves as a universal interpolator between the inputs (features) and outputs (targets or labels). However, in this approach the causal relationship between features and targets is ignored. For example, there is a strong correlation between the wetness of the grass and rain; however, it is the rain that makes the grass wet and not the wet grass causing rain to happen. While this example is obvious, for many physical systems the cause and effect relationships are less clear and at the same time represent an obvious interest for the scientist (Figure 5).

Notably, the issues of causality tend to be overlooked both by the physics and ML communities (Figure 6A).[78] In the theoretical physics world, causal relationships are often assumed to be known and intuitively understood. In some sense, in most cases, theory explores the specifics of a certain causal mechanism or proposes a new one. At the same time, in the ML community causality is often associated with domain expertise, e.g., the choice of the feature and target set. Yet in the experimental world, the causal relationships between the observables are often unknown or known partially (Figure 6B).[79] Given that the theoretical framework for experimental sciences often emerges from theory domain, it is unsurprising that corresponding issues often remain unexplored.

To illustrate a few examples of potentially non-trivial causal relationships that can emerge in experimental settings, the photovoltaic efficiency in thin films can be strongly affected by the presence of grain boundaries. However, whether the grain boundaries serve as a sink for the detrimental ions or instead degrade the carrier transport is unclear and suggests opposite strategies for materials optimization. Similarly, halide segregation and phase evolution in ionic systems can be driven by polarization system instabilities at morphotropic boundaries, or reciprocally pin polarization and result in complex domain structures due to frozen disorder effects. In addition to unknown causal relationships between observable variables and mechanisms, a significant factor can be the presence of possible unobserved confounding variables, i.e., the common source of observed functionalities. For example, if the improved defect density of a composition stems from the incorporation of antisolvent molecules during the synthesis that also changes the carrier mobility, that will result in a strong correlation between the defect density and carrier mobility (Figure 5). However, the direct change of composition via e.g., chemical

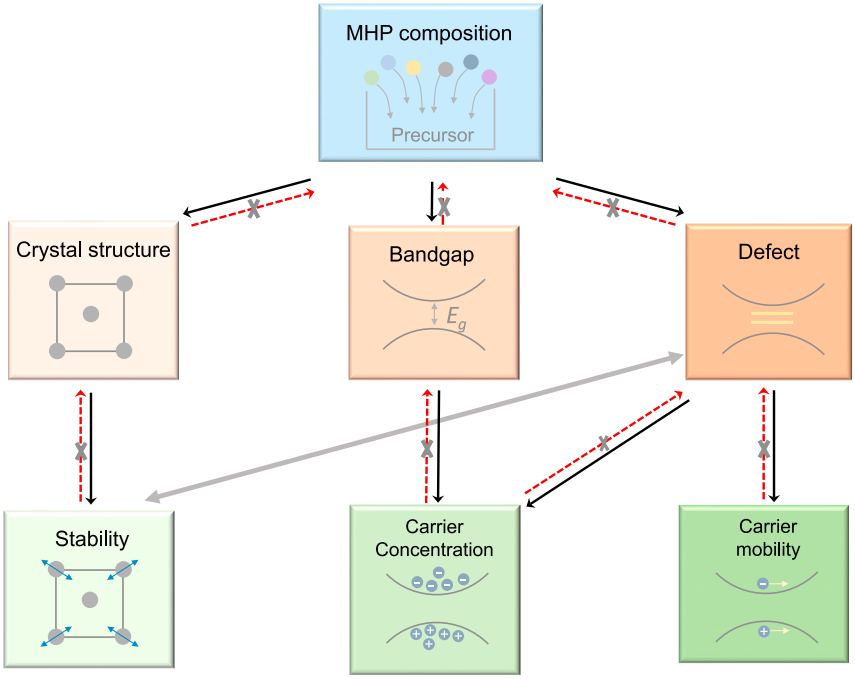

Figure 5. Causal relationship in MHPs

Physical learning: it is generally expected that processing variables (synthesis) will control composition, and composition will control parameters such as crystal structure, defect density, and band gap. Some of these can be controlled well (gross composition), some will be controlled only weakly (defect density). These materials parameters in turn control physical properties of interest (note that only some of the connections are shown). In classical physical discovery, the relationships are assumed to be known and their strength is explored using experiment. However, the "cause and effect" relationships (e.g., composition affects crystal structure, but crystal structure does not affect composition) are assumed to be known. Note that only a subset of causal relationships is shown.

doping will not affect carrier mobility (or rather will affect it in an unpredictable fashion, since the physical mechanisms involved will be completely different). This may result in suboptimal solutions. From these examples, it is clear that knowledge of causal mechanisms represents more than theoretical interest and in fact directly offers the pathways for interventions.

To overcome this limitation, one could use a "causal" BO recently introduced by Aglietti and co-workers.[80] In causal BO, the GP surrogate model integrates observational and interventional data through a causal prior distribution computed using Pearl's do-calculus.[81–86] It is important to note that causal BO does require actual "physical" interventions. However, it is not necessarily the case that causal intervention implies "human" intervention. The causal interventions are realized via a learned causal model that orders the control factors in the form of causal graphs. This both provides insight into the plausible mechanisms and allows to significantly reduce the volumes of the data necessary for the model training. The reason for the latter is that instead of high-dimensional probability density, p(outcome|X_1, X_2, X_3, …), the model now operates on much more low-dimensional space of conditional probabilities p(outcome|$X_{parents}$), p(X_1|$X_{1\_parents}$) and so on, with the control variables arranged as a corresponding directed acyclic graph.

In our case, the interventions can be performed by adjusting a particular growth/synthesis control variable. Correspondingly, the classical acquisition functions are

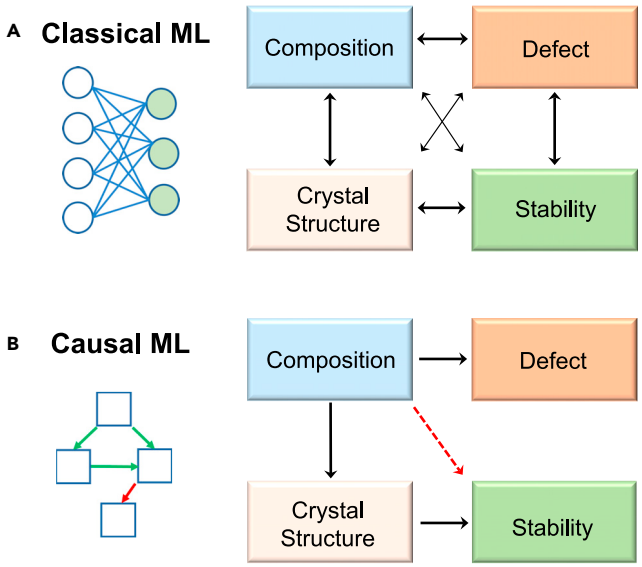

Figure 6. Classical versus causal machine learning (ML)
(A) Compared with physics learning, in (A) classical ML the correlative relationships between the variables are established and the causal links are ignored, often leading to paradoxical results.
(B) In causal ML, some (or none) of the relationships between observed and non-observed variables are assumed to be known, and discovery of the presence and functional form of the others is the goal of the analysis.

replaced with the causal ones to explore the possible interventions. Hence, the causal BO allows for balancing the observation-intervention trade-off in addition to the classical exploration-exploitation trade-off. Compared with the classical BO, the causal BO was demonstrated to reach the global optimum using a much smaller number of steps, both on synthetic and real examples. The causal BO method can be further extended to allow for multi-task causal GP which accounts for correlations between different intervention functions.[87] Also, it should be noted that while causal analysis for cases where causal structures have feedback and cannot be represented as directed acyclic graph (DAG) is still at infancy, we expect that it will be the directions that will become high-priority development in the general ML community. At the same time, existing DAG-based models are already capable of capturing many aspects of the physical behaviors—e.g., composition does affect the PL yield, but not vice versa. For other parameters, causal relationship can be discovered from observations and confirmed via intervention. This allows for knowledge transfer across different experimental setups and is critical in scenarios where experiments (interventions) on some variables cannot be performed.

Theory co-navigation

Over the last 15 years, there has been remarkable progress in applications of ML to theory and materials discovery.[88–91] However, this success has not yet been translated to experiment—exactly because not all-important functionalities can be predicted, and even for those that can be predicted theory often requires certain parameter refinement. In classical scientific domains, the models grew as a result of multiple decades of slow experiment-theory cycles. However, for automated synthesis some aspects of this discovery have to be accelerated. The purpose of this perspective is to illustrate the specific aspects, challenges, pitfalls, and opportunities of ML as applied to synthesis.

Please cite this article in press as: Ahmadi et al., Machine learning for high-throughput experimental exploration of metal halide perovskites, Joule (2021), https://doi.org/10.1016/j.joule.2021.10.001

It is a cornerstone belief in materials science that theory can offer the valuable guidance in search for new materials and optimization of their properties. Yet the fundamental problem in the theory-assisted materials synthesis is that the theoretical models have limited accuracy and precision, i.e., yield predictions with both epistemic (i.e., choice of model) and aleatoric (e.g., precision) errors. Similarly, many of the key properties are either difficult or impossible to calculate, for example, device- or even material-level stability (often due to the presence of latent factors or mechanisms not included in the model). Equivalently, experimentally synthesized and characterized materials properties often are accessible only with inherent uncertainties due to compositional and synthesis variability and can be affected by non-observable and non-controllable factors. Hence, the scientific discovery process often follows an iterative loop between materials synthesis, discovery, and theory development. A well-known example of this is semiconductor theory, where development of materials and device fabrication methods over decades[92] both enabled applications and enabled new physical discovery (e.g., quantum Hall effect that can be discovered only in very pure semiconductors). Hence, from the AE perspective, it is of interest to implement a co-navigation process to balance the functionality and uncertainties in the experimental discovery with the refinement of theoretical models for a specific materials system and synthesis route(s) and generalize it to broader materials classes.

Here, the relevant comparison may be that of the car driving. Theory provides the map, starting with a very low level of detail for simple models[93] (or complex materials), whereas more advanced theoretical methods provide higher details. The experiment in this analogy is driving the car along one of the roads indicated on the map. While the map provides a general direction and informs on the global topology, it cannot be used to stay on the road. At the same time, without the map choosing the direction is impossible. The causal ML methods allow the combination of the two paradigms, where the map is used to choose the initial direction of motion and inform the driver on new opportunities, while at the same time navigation is performed by the AE agent and derived information is used to refine the (local) map.

The idea of co-navigation of theory and experiment has been prevalent in materials science and engineering for several decades now, as outlined in the 2008 report by the National Academies of Sciences, Engineering, and Medicine entitled "Integrated Computational Materials Engineering: A Transformational Discipline for Improved Competitiveness and National Security,"[94] and reaffirmed by the announcement of the Materials Genome Initiative[95] in 2011. Similar ideas and efforts have been spawned across the globe over the last decade, including the so-called "Fourth Industrial Revolution," or Industry 4.0.[96] The idea of Integrated Computational Materials Engineering (ICME) is to bring computational modeling tools into the materials discovery and design process, with the goal of reducing the time and cost of materials development and deployment in commercial applications by 50% or more. A few examples of successful implementation of ICME to new materials development include references,[97–100] while QuesTek Innovations, LLC,[101] is a small business founded in Evanston, Illinois, focused on computational Materials by Design, which recently opened a new division in Europe.[102] However, to date, many of the ICME success stories have been in the realm of structural materials. Such a synergistic approach to materials discovery was recently reported by incorporation of density functional theory (DFT)-calculated free energies of formation, ΔG_f, for mixed cation MHP into a ML-informed experimental approach.[93] Using this co-navigation approach a MHP composition of $(Cs_{0.17}MA_{0.03}FA_{0.83})PbI_3$ with three times greater stability compared with a state of the art $MAPbI_3$ was discovered.

Please cite this article in press as: Ahmadi et al., Machine learning for high-throughput experimental exploration of metal halide perovskites, Joule (2021), https://doi.org/10.1016/j.joule.2021.10.001

Another case study was reported recently where simulation and AE were combined to accelerate research on mechanics of additively manufactured structures.[103]

Generative physical models

As an alternative to the availability of past information, we can consider the fact that phase evolution is indelibly linked to the generative physical models, e.g., system thermodynamics, molecular force fields, or DFT parameters. Note that it is important to separate the generative physical model (i.e., knowledge of mechanism that allows for representation in the formula, computational scheme, etc.) from the generative statistical models such as variational autoencoders or generative adversarial networks that offer black-box models capable of generating data from the same distribution as training data. In addition to fundamentally different operational principles, these two approaches come with opposite requirements for data volumes—whereas generative a physical model can be often selected, improved, or rejected with very few data points, the variational autoencoders and generative adversarial networks require large volumes of data to establish a statistical generative model.

Compared with classical BO methods, that means that we expect to have a (hidden from observation) global generative model that underpins observed phases. This is also different from the BO assisted by phenomenological models in that here past knowledge incorporated in thermodynamic potentials, molecular dynamic force fields, or DFT parameters as determined by the choice of the generative model, are used. Note that while the models rarely allow for exact calculations, they can often be tuned for specific materials systems and allow to generalize over similar systems.

It is also important to note that for parameters that cannot be calculated exactly (e.g., kinetics of degradation processes), there are often strong correlation (but not deterministic relationships) to the predictions of a simplified model, e.g., thermodynamic properties of the system. Hence, the thermodynamic model provides the global model for the systems, with some observed parameters directly linked to it and some strongly correlated. As such, it has the flexibility (akin to human intuition) to combine the quantitative knowledge and correlative trends into a single framework.

Such a co-navigation approach can be illustrated using simple models of chemical thermodynamic, ferroic,[77] or coupled chemical-ferroic behavior,[104] as examples. For chemical thermodynamics, the generative thermodynamic models based on calculation of phase diagrams (CALPHAD) approach can be used to co-navigate the experimental space (Figure 6). Here, an initial approximation of the thermodynamics of the system can be derived from the DFT or prior thermodynamic data to initiate the discovery process from the theory or experimental side, respectively.

The CALPHAD approach is a semi-empirical methodology to describe the thermodynamic properties of multicomponent alloys using composition- and temperature-dependent mathematical models of the Gibbs free energy of each possible phase in the system.[105,106] The general expression for the Gibbs free energy of a phase, φ, is

$$ {}^{\varphi}G = \sum_i x_i {}^{\varphi}G_i^{o} + RT\sum_i x_i \ln x_i + \sum_i \sum_{j\neq i} \left[x_i x_j \left({}^{\varphi}\Omega_{ij} + \sum_{k\neq i\neq j} x_k {}^{\varphi}\Omega_{ijk} \right) \right] \qquad \text{(Equation 3)} $$

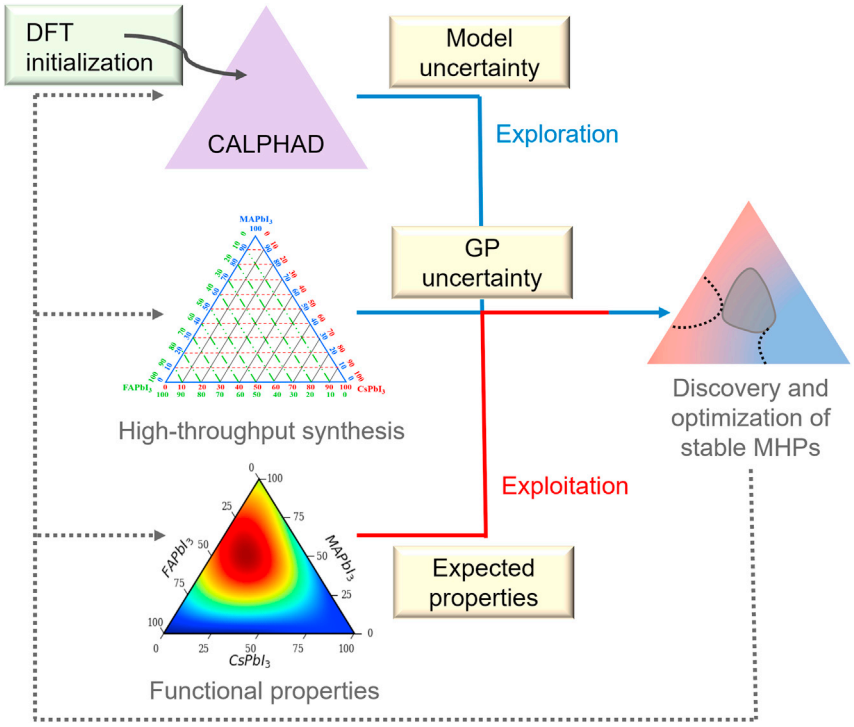

Figure 7. Co-navigation by theory and experiment

The flowchart of the co-navigation approach for discovery and optimization of vast compositional spaces in MHPs using thermodynamic generative model. Note that the workflow can implement more complex models including molecular dynamics or density functional theory, if required, number of calculations is feasible.

where x_i is the mole fraction of component i, ${}^{\varphi}G_i^{\circ}$ is the free energy of pure component i in the form of phase φ, R and T are the ideal gas constant and absolute temperature, respectively, and ${}^{\varphi}\Omega_{ij}$ and ${}^{\varphi}\Omega_{ijk}$ are binary and ternary interaction parameters, respectively, and can be functions of both composition and temperature. CALPHAD modeling starts with thermodynamic descriptions of the pure components, ${}^{\varphi}G_i^{\circ}$ and builds binary, ternary, and higher-order systems by starting with the lower-order subsystems, adding physics-based, but semi-empirical interaction parameters (${}^{\varphi}\Omega_{ij}$ and ${}^{\varphi}\Omega_{ijk}$) to describe nonideal multicomponent behavior. These interaction parameters can be optimized, or assessed, based on all available experimental or computational data. In recent years, CALPHAD modeling has been extended to describe other thermophysical properties (e.g., elastic moduli, and molar volume, electrical resistivity) and kinetic behavior.[100,106,107] The combined methodology has become a powerful tool for engineering materials and process optimization for predicting phase stability, time evolution, and material properties. As such, the CALPHAD approach offers an ideal theory-based modeling companion tool to AE and AI/ML (Figure 7).

An initial thermodynamic "map" can be constructed over an unknown composition space by first performing a select number of DFT calculations, typically of the pure "end-member" components, providing a foundation on which to build the CALPHAD model. Alternatively, the initial parameters can be derived via standard thermodynamic measurements. Limited DFT calculations can also be performed at intermediate compositions to provide an initial estimate of the nonideal solution behavior (i.e., the interaction parameters). Similar to the Gibbs free energy, a CALPHAD-type model can be

Please cite this article in press as: Ahmadi et al., Machine learning for high-throughput experimental exploration of metal halide perovskites, Joule (2021), https://doi.org/10.1016/j.joule.2021.10.001

constructed for any relevant material property, such as band gap, using available experimental information or DFT calculations as an initial calibration dataset. The preliminary CALPHAD models can then be used to identify initial regions in composition space where desirable phase stability and/or materials properties are expected.

With these, AE can then be targeted in these areas, providing new experimental information that can be used to improve model efficacy through refinement of the interaction parameters or inclusion of new parameters using classical least-squares (LS) or BI optimization.

Bayesian hierarchical modeling and model selection

In general, BI is based on the concepts of prior and posterior probabilities. The prior, $p(\theta_i)$, represents the level of knowledge about the system before the experiment. The experiment yields the data, D, based on which the posterior distribution is calculated via Bayes formula:[62,64].

$$p(\theta_i|D) = \frac{p(D|\theta_i)p(\theta_i)}{p(D)} \quad \text{(Equation 4)}$$

Here, $p(D|\theta_i)$ represents the likelihood that this data can be generated by the theory, e.g., given choice of model i, and model parameters θ. The $p(D)$ is the denominator, that defines the total space of possible outcomes.

It is instructive to compare the BI and the classical LS fitting approach. Here, we assume that the experimental observations are given in the form of measured scalar values y_i for the points x_i. In LS fitting, we assume that the observed behavior is described by the function $f(\theta)$ and seek to find the function parameter vector θ. For example, for the PL intensity this will take the form of fitting the observed PL peak to the chosen functional form, e.g., Lorentzian or Gaussian. The optimization is performed by minimizing the least square error between the data and prediction, defined as a calculated version of the function over the data points, $f(x_i)$. The minimum of mean square error in the space of parameters θ defines the point estimate of the function. Note that in a classical least square fits the parameter can be fixed, free, or rigidly constrained (e.g., non-negative).

The BI approach treats the prior information on parameters θ as a joint probability distribution. For convenience, the prior distributions are often marginalized, meaning that the distributions are considered to be independent and the probability density for one parameter is not affected by the others. For example, the probability distribution for the PL peak width is a priori independent on the peak height.

If the specific parameter is well known, the corresponding distribution is narrow and can be chosen based on prior experiments, available published data, or physical models. If the parameter is known poorly, the distribution is broad and for bound parameters is typically chosen as a uniform distribution and for the unbound parameter as a Gaussian (weakly informed priors). The result of the BI is then the posterior parameter distribution, reflecting the updated knowledge on the parameters. For example, before the experiment, we had no knowledge of the peak position and only know that peak should exist in a given spectral interval, corresponding to uniform prior. Given the experimental data, the peak position is localized at a certain interval, corresponding to a narrow posterior distribution.

Another strength of the BI approach is that it readily generalizes to distinguish models via hierarchical Bayesian modeling. In this case, a number of possible models

Please cite this article in press as: Ahmadi et al., Machine learning for high-throughput experimental exploration of metal halide perovskites, Joule (2021), https://doi.org/10.1016/j.joule.2021.10.001

of materials behavior can be selected, each with some prior probability so that total probability is one. For example, for the PL response of the MHP QDs, the models can be chosen to be Gaussian, Lorentzian, and double Gaussian, and double Lorentzian peaks.[56] As the simplest example, initially all models can have equal probability. The probability of the model then becomes one additional prior parameter, and the posterior probability distributions now include both the probability of the model and posterior distributions of model parameters.

Alternatively, the probabilities of the models can be estimated from the posterior densities using the widely applicable information criterion (*WAIC*)[108] as proposed by Gelman and co-workers.[109] The *WAIC* is defined as

$$WAIC = \sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p\left(y_i|\theta^S\right)\right) - \sum_{i=1}^{n} var_{post}\left(\log p(y_i|\theta)\right) \quad \text{(Equation 5)}$$

Here, S is the number of simulations draws and n is the number of available datapoints. The first term in (Equation 5) is the logarithm of predictive density defining the quality of the fit. The second term is the effective number of parameters, *pWAIC*, determined by the total variance of the log likelihood, $\log p(y_i|\theta)$. This term defines the complexity of the fitting function. Note that similar to the calculation of the Bayesian estimate and uncertainty, calculation of *WAIC* requires traces acquired during the sampling rather than just the point estimates. Subsequently, the probability of the model $p(M)$ is recovered via a Bayesian model averaging approach.[110] This approach has been previously applied for the analysis of the PL intensity in multicomponent $CsPbX_3$ (X: I, Br, Cl) QDs,[56] and analysis of domain wall structure in ferroelectrics[111] and system responses in scanning probe microscopy.[112]

Co-navigation of theory and experimental domains

The BO and BI with structured probabilistic models described above offer two limiting approaches for the exploration of the compositional spaces, based on the obtained data (model-free) and refinement of the previously known global model, respectively. We pose that the key step toward implementation of viable automated experiment workflows is the integration of the co-navigation approach, when both experimentally available data and the theoretical model are updated simultaneously, and the combined data-model uncertainty is minimized during the BO or active learning process (with the former performing the global optimization and the latter seeking to uncover the global distribution).

As an example, an initial global model can be chosen as an ideal solid solution (global prior) and several experimental compositions can be chosen as seed points. One can choose the class of prior functions that define the thermodynamics of the systems, i.e., define what the deviations from the ideal solid solution can be as parametrized by $G(X)$ functions in the CALPHAD model. Based on the initial experiment and model, the "unstructured" GP uncertainty and "structured" uncertainty of the generative models with a preset weight coefficient are combined to yield the total uncertainty of the system. In the exploratory mode, the next target of the AE is then chosen to minimize the combined total uncertainty. The strategies for the choice of the weight coefficient can be explored, but the obvious strategies for this can be either epsilon-greedy or "softmax" strategies favoring either model refinement or experimental exploration.

This approach allows a straightforward extension to the property optimization, or exploitation, mode. In this case, the choice of the next compositions to explore is defined by the maxima of the acquisition function that combines the predicted

Please cite this article in press as: Ahmadi et al., Machine learning for high-throughput experimental exploration of metal halide perovskites, Joule (2021), https://doi.org/10.1016/j.joule.2021.10.001

functionality (from GP) and total GP and probabilistic model uncertainties. Note that in this case the discovery and model development will both be driven by the choice of acquisition function, and the model derived this way can be expected to predict locations of these regions in the parameter space, and functional behaviors in these locations, but perform poorly outside.

The co-navigation strategy can minimize the global model and GP uncertainties while optimizing the target functionality. However, the key issue here is the connection between the experimentally detected functionalities and model prediction. For co-navigation, one can use the highly robust descriptors such as phase composition. We have recently demonstrated the framework for doing it via Bayesian methods[56] approximated by the number of PL peaks) and band gap (specifically, its deviation from the one expected from the average composition) as feedback signals, whereas stability will be approximated by the time dependence of PL intensity.

However, the additional flexibility of the co-navigation approach is that it in principle allows the exploration of composition-property relationships in the system and use of these as an additional search variable during the AE and when selecting targets for DFT calculations. As a simple example, consider the compositional dependence of the band gap determined from the experiment or DFT model. In the zero-order approximation, it will be a linear function of the composition. In a more realistic case, it will be an unknown function of the composition, $Eg = h(c)$. If the data, meaning the several sets of Eg for c values are available, the relationship between the two can be established using a functional fit, GP, or more complex approaches such as Bayesian neural networks[113] or variational autoencoders.[114] For simplicity, one can use the GP approach that is expected both to yield the best prediction of the Eg from the concentration, and the uncertainty of this prediction. This uncertainty can also be used for the exploration of composition space now.

We also note that in principle the co-navigation approach can be based on symbolic regression models. In this case, we utilize the fact that in many cases materials functionality follows certain functional relationships that provide a parametric model, and the co-navigation process seeks to refine the model and the experiment simultaneously.

LEVERAGING AUTOMATED SYNTHESIS TO THE CRITICAL DEVICE DEVELOPMENT CAN BE ONE OF THE FOUNDATIONAL COMPONENTS FOR POTENTIALLY TRANSFORMING THE CONVENTIONAL ENERGY TECHNOLOGIES

Finally, the ML-guided automated synthesis of MHPs may be leveraged to developing knowledge and understanding immediately valuable for improving the performance and stability of perovskite optoelectronics, imparting significant technological impacts. There have already been a handful of such studies that particularly concern solar cell applications. Zhao et al.[57] have employed a high-throughput robotic learning process to guide them to fabricate stable $FAPbI_3$ perovskite solar cells. In the first step of that study, the robotic deposition and screening points to a region of high phase stability in $FAPbI_3$ when incorporating at least 10 mol % MA and up to 5 mol % alkaline metal (Cs/Rb/K) in the A-cation sites of the $FAPbI_3$ perovskite. In the sequential device optimization step, only a few perovskite compositions are tried within this predetermined narrower composition window ($Cs_xMA_{0.15-x}FA_{0.85}PbI_3$), which leads to an optimal device stability ($Cs_{0.05}MA_{0.10}FA_{0.85}PbI_3$) at the operational stability. In another study, Buonassisi and co-workers[95] have showed a physical data-fusion approach to enable a human-free

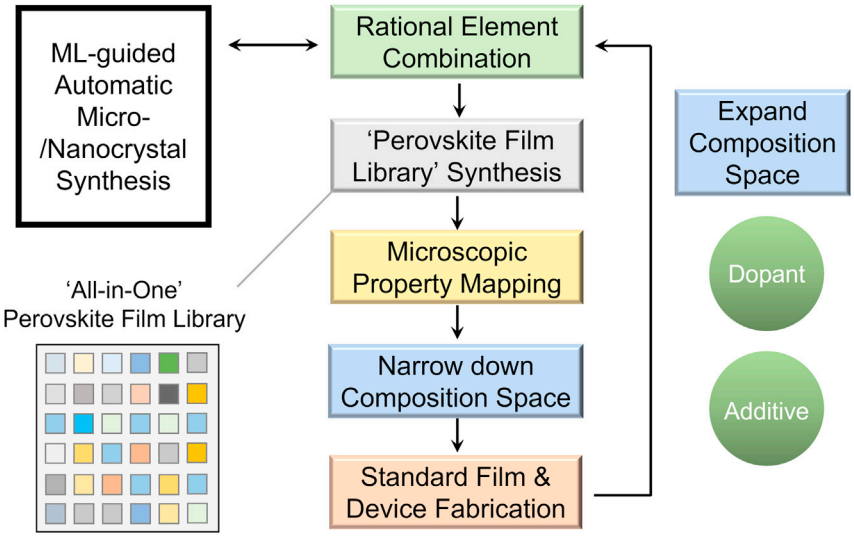

Figure 8. ML-guided thin film and device optimization

The flowchart for leveraging ML-guided automatic micro-/nanocrystal synthesis to MHP device optimization. The key device optimization loop entails exploring the "all-in-one" perovskite film library fabrication, narrowing down the interested candidate composition space, expanding the composition space for a fine screening via doping and additive.

decision-making process for screening $Cs_xMA_yFA_{1-x-y}PbI_3$ with the best stability, which is later attested by the device results. While both studies showcase the unprecedented power of ML-guided automated synthesis for accelerating the device development, there remains a huge research space for a further development of this promising research area.

Shown in Figure 8 is the flowchart that outlines the rational procedures connecting the ML-guided automatic micro-/nanocrystal synthesis to the optimization of perovskite optoelectronic devices. The key work loop iterates exploring the "all-in-one" perovskite film library fabrication, narrowing down the interesting candidate composition space for specific device applications, expanding the composition space for a fine screening via doping and additives.

The current studies are mostly limited to demonstrating effective composition screening based on solution-synthesized microcrystal or QD samples. However, the properties of bulk semiconductor crystals may differ substantially from those of thin films. Especially thin film materials contain various intra-grain and inter-grain interfaces which are usually absent in microcrystals or QDs,[115,116] while electronic or chemical tolerance to the interface defects is one crucial factor determining the suitability of an optoelectronic material to serve in a device. In this context, automated synthesis of perovskite samples in the form of thin or thick films needs to be further matured. Not only do the instrumental and ML methods of reliable robotic deposition of film libraries need to be improved, but various device-related factors should also be taken into consideration in the experimental design. For example, depending on the device applications (solar cells, light-emitting devices, radiation sensors, etc.), sample substrates such as mesoporous TiO_2-coated fluorine-doped tin oxide glasses and PEDOT:PSS-coated indium tin oxide glasses may be used for automatic film fabrication. The examination of as-synthesized film properties on these specific substrates can guide us to make a more accurate prediction regarding the device performance. Furthermore, the "all-in-one" design concept could be used for

Please cite this article in press as: Ahmadi et al., Machine learning for high-throughput experimental exploration of metal halide perovskites, Joule (2021), https://doi.org/10.1016/j.joule.2021.10.001

fabricating solid film sample libraries, which will involve the deposition of arrays of small-size thin films on single patterned substrates. This may unlock opportunities not only for a facile, automatic property investigation but also for a systematic screening of device performance. Based on the "all-in-one" film library, high spatial-temporal-resolution characterizations such as synchrotron (nanoscale X-ray diffraction, X-ray ptychography, and X-ray-induced current technique, etc.) and various microscopies (PL, AFM, TEM, etc.) will become immediately useful for rapidly characterizing these sample libraries, establishing a map of composition-property-performance correlation. This will help us to narrow down the interested composition space for specific applications and guide us to continue with standard film and device fabrications. It is also possible to fabricate "all-in-one" device libraries as needed. In this case, those device characterization techniques, previously established for large-scale device evaluation, can be used to compare the performance of different film samples in the library chip. For example, electroluminescence and lock-in thermography images can be used to quickly identify the compositions on the chip that make the best solar cells.[117]

Once the interesting perovskite composition space is narrowed down, it becomes feasible and promising to move to the device optimization process. Device optimization plays an important role in achieving the final performance of devices. For example, the efficiency of solar cells made using the same $MAPbI_3$ composition can easily vary from 0% to 20%, and their stability can also demonstrate substantial difference. The underlying factor is frequently related to the incorporation of certain additives and dopants. Especially those most recent achievements in reporting record perovskite solar cell devices have been invariably associated with the additive/dopant use in the processing of perovskite thin films.[118–121] These additives are found to serve for various functions, including crystallization control, defect passivation, carrier transport, anti-oxidation, ion-blocking, and their combinations.[116,122–124] Nevertheless, the discovery of these additives have been mostly driven by the chemistry intuition and trial-and-error processes. In fact, the types of additives that are claimed effective for perovskite devices are numerous, including small molecules, inorganic, polymer, and complex. Similar to perovskites, these additives or dopants can exhibit versatile compositions. For example, in an early study, Zong and co-workers[125] reported a complex additive of $SnF_2{\bullet}x$FACl ($x = 3$) to enhance the Sn-Pb mixed perovskite solar cells that are attractive for their low-band-gap characteristics. While this additive composition has contributed to decent improvement in the device performance, there are still a number of possible complexes based on the variation of Sn halide type, organic halide type, and their mole ratio. In this context, the interesting compositions screened by the initial device trials will be further integrated with the composition space of additives and dopants, expanding the composition space to study in the sequential automatic sample film library fabrication.

The optimization of the device preparation process via ML represents a considerably more complex task due to the higher dimensionality of control parameter space and much stronger role of the non-observable parameters. Here, the BO based methods discussed in previous sections can provide initial ML frameworks. For synthesis, more opportunities can be opened by methods such as reinforcement learning. However, these methods are known to be extremely data hungry, necessitating development of new strategies.

SUMMARY AND OUTLOOK

To summarize, the rapid development of AE over the last several years opens tremendous opportunities toward the low-cost automated synthesis of MHP

Please cite this article in press as: Ahmadi et al., Machine learning for high-throughput experimental exploration of metal halide perovskites, Joule (2021), https://doi.org/10.1016/j.joule.2021.10.001

materials and devices. However, harnessing these opportunities necessitates comparable developments in the ML methods controlling these instrumentations and allowing for navigation in multidimensional compositional and synthesis spaces. The simple combinatorial strategies are equivalent to grid search in these spaces and offer a several orders of magnitude advantage compared with sequential synthesis, completely insufficient to deal with the immensity of compositional and synthesis parameter spaces.

BO methods are rapidly becoming the paradigmatic methods underpinning AE to guide it toward specific functionalities. However, further developments require incorporation of prior knowledge and physical priors in the form of search candidates, addressing multidimensional and continuous spaces, incorporation of thermodynamic models, and known physical behaviors. Similarly, the existence of generative physics models but the partial knowledge of the latter, as well as out of distribution drifts require developing strategies for co-navigation of the experimental and theoretical domains when both are explored simultaneously, and incorporation of causal AE strategies.

Further progress in the field necessitates the rapid adoption of the fully autonomous, microfluidic, and combined synthesis workflows, the development now possible due to the low-cost of commercial tools and availability of Python interfaces. However, it also necessitates the development of ML methods and infrastructure optimized for automated experiment. This involves creation of publicly available "scientific" databases (as opposed to databases containing images of cats and dogs) for training and/or evaluation of the designed ML methods,[24,126] and better understanding of ML predictive behaviors under the dataset shifts[127]—i.e., when a model is applied outside the domain of training examples—which could include small changes in data acquisition parameters in the real-time and characterization systems. Note that the need for creation of the databases of materials properties and theoretical models is by now well recognized and is well reflected in a number of recent publications and programmatic documents.[128,129] The data repositories such as Citrination,[36] NoMAD,[37] and materials innovation network[38] are now becoming common. At the same time, the necessary algorithm and code base is now actively emerging in ML communities and are disseminated via repositories such as GitHub. Finally, cloud-based services such as Google Colab or Microsoft Azure now enable integration of the scientific publication with code and data as implemented in Jupyter papers,[57] and books and papers with code.[130]

The prospective development in ML-guided automated synthesis may stimulate the progress of numerous directions in the MHPs research. In particular, searching for Pb-free MHPs has been an active area of research that will influence the practical commercialization of MHP PVs. The automation will facilitate the establishment of a link between the experimental synthesis and theoretical predication, bringing new opportunities in accelerating the discovery of promising Pb-free candidates for perovskite PVs. In the context of structural versatility, MHPs can embrace members beyond these 3D ABX_3 perovskites or $A_2B^{I}B^{III}X_6$ double perovskites. The layered MHPs,[131] either in a Ruddlesden-Popper, Dion-Jacobson, or alternative-cation-interlayer (ACI) structure, have emerged as an attractive group of semiconductors with anisotropic electronic properties and enhanced stability, which are well suited for PVs and optoelectronic applications. More promising is that in layered MHPs, the size of organic cations is not a constraint for forming a perovskite structure. This makes MHPs a huge materials family with even more versatile functions, attesting the importance of ML-guided automated synthesis for accelerated materials screening and understanding. Furthermore, the methodology developed

based on MHPs as a model system can be extended to study various other solution-processed optical and electronic materials such as chalcohalides and organic-inorganic coordination complexes. Finally, owing to the solution processability of the whole perovskite device stack, ML-guided automated synthesis and characterization is also suitable for the investigation and optimization of other device layers including carrier-transporting, contact, and encapsulant layer, and their chemical/physical interactions with MHPs, transforming our conventional "black-box" synthesis/fabrication approach for co-optimization of various device components.

Finally, we believe that full automation of the synthesis process is actually unlikely (think of autonomous driving which—as realized by now—turned out to be a much harder problem than thought 3–5 years ago), and the purpose of ML is not to substitute human operator and decision-making but reduce it to the high-level high-latency decisions. In some sense, it is already the case for the techniques such as BO, where the acquisition function is selected based on the perceived (by human operator) target. It will also be the case for multi-objective BO methods, where the algorithm yields the compositions on the Pareto front of certain components of figures of merit, but it is up to humans (or a different expert system exploring the e.g., economic considerations) to define what the balance should be. This list can be continued, but ultimately, we believe that automated synthesis complements (or "augments") humans but does not substitute one.

ACKNOWLEDGMENTS

M.A. acknowledges support from National Science Foundation (NSF), award number # 2043205. This work was performed (S.V.K. and M.Z.) and partially supported at the Oak Ridge National Laboratory's Center for Nanophase Materials Sciences (CNMS), a US Department of Energy, Office of Science User Facility. Y.Z. acknowledges the start-up grants, Initiation Grant - Faculty Niche Research Areas (IG-FNRA) 2020/21 and Inter-disciplinary Matching Scheme 2020/21 of the Hong Kong Baptist University (HKBU) and the Early Career Scheme (No. 22300221) from the Hong Kong Research Grant Council.. The authors are very grateful to T. Buonassisi (MIT), K. Brown (Boston University), and E. Sargent (University of Toronto) for valuable comments and feedback.

AUTHOR CONTRIBUTIONS

M.A. led the project and contributed to writing the automated synthesis, characterization, and promising paradigm of materials discovery and design. M.Z. and S.V.K. contributed to writing the ML, pitfalls, and opportunities. Y.Z. contributed to the writing concerning the device aspects and the broad context of energy science. E.A.L. contributed to writing the theory and development of CALPHAD model for co-navigation. All authors contributed to the cross-disciplinary discussions and the editing of the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Park, N.-G., Grätzel, M., Miyasaka, T., Zhu, K., and Emery, K. (2016). Towards stable and commercially available perovskite solar cells. Nat. Energy *1*, 16152.

2. Zhao, B., Bai, S., Kim, V., Lamboll, R., Shivanna, R., Auras, F., Richter, J.M., Yang, L., Dai, L., Alsari, M., et al. (2018). High-efficiency perovskite–polymer bulk heterostructure light-emitting diodes. Nat. Photonics *12*, 783–789.

3. Lukosi, E., Smith, T., Tisdale, J., Hamm, D., Seal, C., Hu, B., and Ahmadi, M. (2019). Methylammonium lead tribromide semiconductors: ionizing radiation detection and electronic properties. Nucl. Instrum. Methods Phys. Res. A *927*, 401–406.

4. Ahmadi, M., Wu, T., and Hu, B. (2017). A review on organic–inorganic halide

perovskite photodetectors: device engineering and fundamental physics. Adv. Mater. *29*, 1605242.

5. Wei, H., and Huang, J. (2019). Halide lead perovskites for ionizing radiation detection. Nat. Commun. *10*, 1066.

6. Jain, A., Voznyy, O., and Sargent, E.H. (2017). High-throughput screening of lead-free perovskite-like materials for optoelectronic applications. J. Phys. Chem. C *121*, 7183–7187.

7. Nakajima, T., and Sawada, K. (2017). Discovery of Pb-free perovskite solar cells via high-throughput simulation on the K computer. J. Phys. Chem. Lett. *8*, 4826–4831.

8. Zhao, X.-G., Yang, D., Ren, J.-C., Sun, Y., Xiao, Z., and Zhang, L. (2018). Rational design of halide double perovskites for optoelectronic applications. Joule *2*, 1662–1673.

9. Saliba, M. (2019). Polyelemental, multicomponent perovskite semiconductor libraries through combinatorial screening. Adv. Energy Mater. *9*, 1803754.

10. Liu, Y., Kim, D., Ievlev, A.V., Kalinin, S.V., Ahmadi, M., and Ovchinnikova, O.S. (2021). Ferroic halide perovskite optoelectronics. Adv. Funct. Mater. *31*, 2102793.

11. Lee, L.C., Huq, T.N., MacManus-Driscoll, J.L., and Hoye, R.L.Z. (2018). Research update: bismuth-based perovskite-inspired photovoltaic materials. APL Mater *6*, 084502.

12. Park, B.-W., Philippe, B., Zhang, X., Rensmo, H., Boschloo, G., and Johansson, E.M.J. (2015). Bismuth based hybrid perovskites A3Bi2I9 (A: methylammonium or cesium) for solar cell application. Adv. Mater. *27*, 6806–6813.

13. Bartel, C.J., Sutton, C., Goldsmith, B.R., Ouyang, R., Musgrave, C.B., Ghiringhelli, L.M., and Scheffler, M. (2019). New tolerance factor to predict the stability of perovskite oxides and halides. Sci. Adv. *5*, eaav0693.

14. Rödel, J., Jo, W., Seifert, K.T.P., Anton, E.-M., Granzow, T., and Damjanovic, D. (2009). Perspective on the development of lead free piezoceramics. J. Am. Ceram. Soc. *92*, 1153–1177.

15. Hong, C.-H., Kim, H.-P., Choi, B.-Y., Han, H.-S., Son, J.S., Ahn, C.W., and Jo, W. (2016). Lead-free piezoceramics – where to move on? J. Materiomics *2*, 1–24.

16. Ju, M.-G., Chen, M., Zhou, Y., Dai, J., Ma, L., Padture, N.P., and Zeng, X.C. (2018). Toward eco-friendly and stable perovskite materials for photovoltaics. Joule *2*, 1231–1241.

17. Volonakis, G., Haghighirad, A.A., Milot, R.L., Sio, W.H., Filip, M.R., Wenger, B., Johnston, M.B., Herz, L.M., Snaith, H.J., and Giustino, F. (2017). Cs2InAgCl6: a new lead-free halide double perovskite with direct band gap. J. Phys. Chem. Lett. *8*, 772–778.

18. Xiao, Z., Du, K.-Z., Meng, W., Wang, J., Mitzi, D.B., and Yan, Y. (2017). Intrinsic instability of Cs2In(I)M(III)X6 (M = Bi, Sb; X = halogen) double perovskites: a combined density functional theory and experimental study. J. Am. Chem. Soc. *139*, 6054–6057.

19. Zhang, C.-X., Shen, T., Guo, D., Tang, L.-M., Yang, K., and Deng, H.-X. (2020). Reviewing and understanding the stability mechanism of halide perovskite solar cells. InfoMat *2*, 1034–1056.

20. Ono, L.K., Juarez-Perez, E.J., and Qi, Y. (2017). Progress on perovskite materials and solar cells with mixed cations and halide anions. ACS Appl. Mater. Interfaces *9*, 30197–30246.

21. Wang, Z., Shi, Z., Li, T., Chen, Y., and Huang, W. (2017). Stability of perovskite solar cells: a prospective on the substitution of the A cation and X anion. Angew. Chem. Int. Ed. *56*, 1190–1212.

22. Xu, F., Zhang, T., Li, G., and Zhao, Y. (2017). Mixed cation hybrid lead halide perovskites with enhanced performance and stability. J. Mater. Chem. A *5*, 11450–11461.

23. Bi, D., Tress, W., Dar, M.I., Gao, P., Luo, J., Renevier, C., Schenk, K., Abate, A., Giordano, F., Correa Baena, J.-P., et al. (2016). Efficient luminescent solar cells based on tailored mixed-cation perovskites. Sci. Adv. *2*, e1501170.

24. 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., et al. (2019). Accelerated development of perovskite-inspired materials via high-throughput synthesis and machine-learning diagnosis. Joule *3*, 1437–1451.

25. Chen, S., Hou, Y., Chen, H., Tang, X., Langner, S., Li, N., Stubhan, T., Levchuk, I., Gu, E., Osvet, A., and Brabec, C.J. (2018). Exploring the stability of novel wide bandgap perovskites by a robot based high throughput approach. Adv. Energy Mater. *8*, 1701543.

26. Jesper Jacobsson, T., Correa-Baena, J.-P., Pazoki, M., Saliba, M., Schenk, K., Grätzel, M., and Hagfeldt, A. (2016). Exploration of the compositional space for mixed lead halogen perovskites for high efficiency solar cells. Energy Environ. Sci. *9*, 1706–1724.

27. Schelhas, L.T., Li, Z., Christians, J.A., Goyal, A., Kairys, P., Harvey, S.P., Kim, D.H., Stone, K.H., Luther, J.M., Zhu, K., et al. (2019). Insights into operational stability and processing of halide perovskite active layers. Energy Environ. Sci. *12*, 1341–1348.

28. Sarkar, A., Velasco, L., Wang, D., Wang, Q., Talasila, G., de Biasi, L., Kübel, C., Brezesinski, T., Bhattacharya, S.S., Hahn, H., and Breitung, B. (2018). High entropy oxides for reversible energy storage. Nat. Commun. *9*, 3400.

29. Sarkar, A., Wang, Q., Schiele, A., Chellali, M.R., Bhattacharya, S.S., Wang, D., Brezesinski, T., Hahn, H., Velasco, L., and Breitung, B. (2019). High-entropy oxides: fundamental aspects and electrochemical properties. Adv. Mater. *31*, e1806236.

30. Wu, C., Guo, D., Li, P., Wang, S., Liu, A., and Wu, F. (2020). A study on the effects of mixed organic cations on the structure and properties in lead halide perovskites. Phys. Chem. Chem. Phys. *22*, 3105–3111.

31. Tan, W., Bowring, A.R., Meng, A.C., McGehee, M.D., and McIntyre, P.C. (2018). Thermal stability of mixed cation metal halide perovskites in air. ACS Appl. Mater. Interfaces *10*, 5485–5491.

32. Rehman, W., McMeekin, D.P., Patel, J.B., Milot, R.L., Johnston, M.B., Snaith, H.J., and Herz, L.M. (2017). Photovoltaic mixed-cation lead mixed-halide perovskites: links between crystallinity, photo-stability and electronic properties. Energy Environ. Sci. *10*, 361–369.

33. Donakowski, A., Miller, D.W., Anderson, N.C., Ruth, A., Sanehira, E.M., Berry, J.J., Irwin, M.D., Rockett, A., and Steirer, K.X. (2021). Improving photostability of cesium-doped formamidinium lead triiodide perovskite. ACS Energy Lett *6*, 574–580.

34. Habasaki, J., and Ngai, K.L. (2007). The mixed alkali effect in ionically conducting glasses revisited: A study by molecular dynamics simulation. Phys. Chem. Chem. Phys. *9*, 4673–4689.

35. Ferdani, D.W., Pering, S.R., Ghosh, D., Kubiak, P., Walker, A.B., Lewis, S.E., Johnson, A.L., Baker, P.J., Islam, M.S., and Cameron, P.J. (2019). Partial cation substitution reduces iodide ion transport in lead iodide perovskite solar cells. Energy Environ. Sci. *12*, 2264–2272.

36. Ghosh, S., Mishra, S., and Singh, T. (2020). Antisolvents in perovskite solar cells: importance, issues, and alternatives. Adv. Mater. Interfaces *7*, 2000950.

37. Taylor, A.D., Sun, Q., Goetz, K.P., An, Q., Schramm, T., Hofstetter, Y., Litterst, M., Paulus, F., and Vaynzof, Y. (2021). A general approach to high-efficiency perovskite solar cells by any antisolvent. Nat. Commun. *12*, 1878.

38. Xiao, M., Zhao, L., Geng, M., Li, Y., Dong, B., Xu, Z., Wan, L., Li, W., and Wang, S. (2018). Selection of an anti-solvent for efficient and stable cesium-containing triple cation planar perovskite solar cells. Nanoscale *10*, 12141–12148.

39. Langner, S., Häse, F., Perea, J.D., Stubhan, T., Hauch, J., Roch, L.M., Heumueller, T., Aspuru-Guzik, A., and Brabec, C.J. (2020). Beyond ternary OPV: high-throughput experimentation and self-driving laboratories optimize multicomponent systems. Adv. Mater. *32*, e1907801.

40. Higgins, K., Ziatdinov, M., Kalinin, S.V., and Ahmadi, M. (2021). High-throughput study of antisolvents on the stability of multicomponent metal halide perovskites through robotics-based synthesis and machine learning approaches. arXiv, arXiv:2106.03312.

41. Takeuchi, I., Famodu, O.O., Read, J.C., Aronova, M.A., Chang, K.-S., Craciunescu, C., Lofland, S.E., Wuttig, M., Wellstood, F.C., Knauss, L., and Orozco, A. (2003). Identification of novel compositions of ferromagnetic shape-memory alloys using composition spreads. Nat. Mater. *2*, 180–184.

42. Ohkubo, I., Christen, H.M., Kalinin, S.V., Jr., Jellison, G.E., Rouleau, C.M., and Lowndes, D.H. (2004). High-throughput growth temperature optimization of ferroelectric SrxBa1−xNb2O6 epitaxial thin films using a temperature gradient method. Appl. Phys. Lett. *84*, 1350–1352.

43. Christen, H.M., Rouleau, C.M., Ohkubo, I., Zhai, H.Y., Lee, H.N., Sathyamurthy, S., and Lowndes, D.H. (2003). An improved

continuous compositional-spread technique based on pulsed-laser deposition and applicable to large substrate areas. Rev. Sci. Instrum. *74*, 4058–4062.

44. Wang, J., Yoo, Y., Gao, C., Takeuchi, I., Sun, X., Chang, H., Xiang, X.-D., and Schultz, P.G. (1998). Identification of a blue photoluminescent composite material from a combinatorial library. Science *279*, 1712–1714.

45. Steiner, S., Wolf, J., Glatzel, S., Andreou, A., Granda, J.M., Keenan, G., Hinkley, T., Aragon-Camarasa, G., Kitson, P.J., Angelone, D., and Cronin, L. (2019). Organic synthesis in a modular robotic system driven by a chemical programming language. Science *363*, eaav2211.

46. Angelone, D., Hammer, A.J.S., Rohrbach, S., Krambeck, S., Granda, J.M., Wolf, J., Zalesskiy, S., Chisholm, G., and Cronin, L. (2021). Convergence of multiple synthetic paradigms in a universally programmable chemical synthesis machine. Nat. Chem. *13*, 63–69.

47. Nikolaev, P., Hooper, D., Perea-López, N., Terrones, M., and Maruyama, B. (2014). Discovery of wall-selective carbon nanotube growth conditions via automated experimentation. ACS Nano *8*, 10214–10222.

48. Nikolaev, P., Hooper, D., Webber, F., Rao, R., Decker, K., Krein, M., Poleski, J., Barto, R., and Maruyama, B. (2016). Autonomy in materials research: a case study in carbon nanotube growth. npj Comput. Mater. *2*, 16031.

49. MacLeod, B.P., Parlane, F.G.L., Morrissey, T.D., Häse, F., Roch, L.M., Dettelbach, K.E., Moreira, R., Yunker, L.P.E., Rooney, M.B., Deeth, J.R., et al. (2020). Self-driving laboratory for accelerated discovery of thin-film materials. Sci. Adv. *6*, eaaz8867.

50. Coley, C.W., Thomas, D.A., Lummiss, J.A.M., Jaworski, J.N., Breen, C.P., Schultz, V., Hart, T., Fishman, J.S., Rogers, L., Gao, H., et al. (2019). A robotic platform for flow synthesis of organic compounds informed by AI planning. Science *365*, eaax1566.

51. Burger, B., Maffettone, P.M., Gusev, V.V., Aitchison, C.M., Bai, Y., Wang, X., Li, X., Alston, B.M., Li, B., Clowes, R., et al. (2020). A mobile robotic chemist. Nature *583*, 237–241.

52. Epps, R.W., Felton, K.C., Coley, C.W., and Abolhasani, M. (2017). Automated microfluidic platform for systematic studies of colloidal perovskite nanocrystals: towards continuous nano-manufacturing. Lab Chip *17*, 4040–4047.

53. Abdel-Latif, K., Epps, R.W., Bateni, F., Han, S., Reyes, K.G., and Abolhasani, M. (2021). Self-driven multistep quantum dot synthesis enabled by autonomous robotic experimentation in flow. Adv. Intell. Syst. *3*, 2000245.

54. Epps, R.W., Bowen, M.S., Volk, A.A., Abdel-Latif, K., Han, S., Reyes, K.G., Amassian, A., and Abolhasani, M. (2020). Artificial chemist: an autonomous quantum dot synthesis bot. Adv. Mater. *32*, e2001626.

55. Higgins, K., Valleti, S.M., Ziatdinov, M., Kalinin, S.V., and Ahmadi, M. (2020). Chemical robotics enabled exploration of stability in multicomponent lead halide perovskites via machine learning. ACS Energy Lett *5*, 3426–3436.

56. Heimbrook, A., Higgins, K., Kalinin, S.V., and Ahmadi, M. (2021). Exploring the physics of cesium lead halide perovskite quantum dots via Bayesian inference of the photoluminescence spectra in automated experiment. Nanophotonics *10*, 1977–1989.

57. Zhao, Y., Zhang, J., Xu, Z., Sun, S., Langner, S., Hartono, N.T.P., Heumueller, T., Hou, Y., Elia, J., Li, N., et al. (2021). Discovery of temperature-induced stability reversal in perovskites using high-throughput robotic learning. Nat. Commun. *12*, 2191.

58. Gu, E., Tang, X., Langner, S., Duchstein, P., Zhao, Y., Levchuk, I., Kalancha, V., Stubhan, T., Hauch, J., Egelhaaf, H.J., et al. (2020). Robot-based high-throughput screening of antisolvents for lead halide perovskites. Joule *4*, 1806–1822.

59. Choubisa, H., Askerka, M., Ryczko, K., Voznyy, O., Mills, K., Tamblyn, I., and Sargent, E.H. (2020). Crystal site feature embedding enables exploration of large chemical spaces. Matter *3*, 433–448.

60. 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., et al. (2020). Robot-accelerated perovskite investigation and discovery. Chem. Mater. *32*, 5650–5663.

61. Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. Nature *521*, 452–459.

62. Lambert, B. (2018). A Student's Guide to Bayesian Statistics, First Edition (SAGE Publications).

63. Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L., and Zdeborová, L. (2019). Machine learning and the physical sciences. Rev. Mod. Phys. *91*, 045002.

64. Martin, O. (2018). Bayesian Analysis with Python: Introduction to Statistical Modeling and Probabilistic Programming Using PyMC3 and ArviZ, Second Edition (Packt Publishing).

65. Kruschke, J. (2014). Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan, Second Edition (Academic Press).

66. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., and Rubin, D.B. (2013). Bayesian Data Analysis, Third Edition (Chapman and Hall/CRC).

67. Gongora, A.E., Xu, B., Perry, W., Okoye, C., Riley, P., Reyes, K.G., Morgan, E.F., and Brown, K.A. (2020). A Bayesian experimental autonomous researcher for mechanical design. Sci. Adv. *6*, eaaz1708.

68. Matheron, G. (1963). Principles of geostatistics. Econ. Geol. *58*, 1246–1266.

69. Wilson, J.T., Hutter, F., and Deisenroth, M.P. (2018). Maximizing acquisition functions for Bayesian optimization. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 9906–9917.

70. Reyes, K.G., and Maruyama, B. (2019). The machine learning revolution in materials? MRS Bull *44*, 530–537.

71. Tabor, D.P., Roch, L.M., Saikin, S.K., Kreisbeck, C., Sheberla, D., Montoya, J.H., Dwaraknath, S., Aykol, M., Ortiz, C., Tribukait, H., et al. (2018). Accelerating the discovery of materials for clean energy in the era of smart automation. Nat. Rev. Mater. *3*, 5–20.

72. Noack, M.M., Doerk, G.S., Li, R.P., Fukuto, M., and Yager, K.G. (2020). Advances in Kriging-based autonomous X-ray scattering experiments. Sci. Rep. *10*, 1325.

73. Noack, M.M., Yager, K.G., Fukuto, M., Doerk, G.S., Li, R.P., and Sethian, J.A. (2019). A Kriging-based approach to autonomous experimentation with applications to X-ray scattering. Sci. Rep. *9*, 11809.

74. Kalinin, S.V., Ziatdinov, M., and Vasudevan, R.K. (2020). Guided search for desired functional responses via Bayesian optimization of generative model: hysteresis loop shape engineering in ferroelectrics. J. Appl. Phys. *128*, 024102.

75. Wilson, A.G., and Nickisch, H. (2015). Kernel interpolation for scalable structured Gaussian processes (KISS-GP). Proceedings of the 32nd International Conference on International Conference on Machine Learning *37*, 1775–1784.

76. Fortuin, V., and Rätsch, G. (2019). Meta-learning mean functions for gaussian processes. arXiv, arXiv:1901.08098.

77. Tagantsev, A.K., Cross, L.E., and Fousek, J. (2010). Domains in Ferroic Crystals and Thin Films (Springer).

78. Vasudevan, R.K., Ziatdinov, M., Vlcek, L., and Kalinin, S.V. (2021). Off-the-shelf deep learning is not enough, and requires parsimony, Bayesianity, and causality. npj Comput. Mater. *7*, 16.

79. Ziatdinov, M., Nelson, C.T., Zhang, X.H., Vasudevan, R.K., Eliseev, E., Morozovska, A.N., Takeuchi, I., and Kalinin, S.V. (2020). Causal analysis of competing atomistic mechanisms in ferroelectric materials from high-resolution scanning transmission electron microscopy data. npj Comp. Mater. *6*, 127.

80. Aglietti, V., Lu, X., Paleyes, A., and González, J. (2020). Causal Bayesian optimization. In Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, *108*, C. Silvia and C. Roberto, eds., pp. 3155–3164.

81. Pearl, J. (2009). Causality: Models, Reasoning and Inference (Cambridge University Press).

82. Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. Commun. ACM *62*, 54–60.

83. Pearl, J. (2019). On the Interpretation of do(x). J. Causal Inference *7*, 6.

84. Pearl, J. (2017). A linear "microscope" for interventions and counterfactuals. J. Causal Inference *5*, 15.

85. Bareinboim, E., and Pearl, J. (2016). Causal inference and the data-fusion problem. Proc. Natl. Acad. Sci. USA *113*, 7345–7352.

86. Galles, D., and Pearl, J. (1997). Axioms of causal relevance. Artif. Intell. *97*, 9–43.

87. Aglietti, V., Damoulas, T., Álvarez, M., and González, J. (2020). Multi-task causal learning with Gaussian processes. arXiv, arXiv:2009.12821.

88. Saidi, W.A., Shadid, W., and Castelli, I.E. (2020). Machine-learning structural and electronic properties of metal halide perovskites using a hierarchical convolutional neural network. npj Comput. Mater. *6*, 36.

89. Tao, Q., Xu, P., Li, M., and Lu, W. (2021). Machine learning for perovskite materials design and discovery. npj Comput. Mater. *7*, 23.

90. Li, Z., Xu, Q., Sun, Q., Hou, Z., and Yin, W.-J. (2019). Thermodynamic stability landscape of halide double perovskites via high-throughput computing and machine learning. Adv. Funct. Mater. *29*, 1807280.

91. Li, Z., Xu, Q., Sun, Q., Hou, Z., and Yin, W.-J. (2018). Stability engineering of halide perovskite via machine learning. arXiv, arXiv:1803.06042.

92. Hoddeson, L., and Riordan, M. (1998). Crystal Fire: The Invention of the Transistor and the Birth of the Information Age (W. W. Norton & Company).

93. Sun, S., Tiihonen, A., Oviedo, F., Liu, Z., Thapa, J., Zhao, Y., Hartono, N.T.P., Goyal, A., Heumueller, T., Batali, C., et al. (2021). A data fusion approach to optimize compositional stability of halide perovskites. Matter *4*, 1305–1322.

94. National Research Council (2008). Integrated Computational Materials Engineering: A Transformational Discipline for Improved Competitiveness and National Security (The National Academies Press).

95. National Science and Technology Council. (2011). Materials genome initiative for global competitiveness (Executive Office of the President)). https://www.mgi.gov/sites/default/files/documents/materials_genome_initiative-final.pdf.

96. Schwab, K. (2017). The Fourth Industrial Revolution (Crown Publishing Group Business).

97. Lass, E.A. (2017). Application of computational thermodynamics to the design of a Co-ni-based γ′-strengthened superalloy. Metall. Mater. Trans. A *48*, 2443–2459.

98. Montero-Chacón, F., Chiumenti, M., Segurado, J., and Doblaré, M. (2018). Integrated computational materials engineering in solar plants: the virtual materials design project. JOM *70*, 1659–1669.

99. Wong, T.T., and Paramsothy, M. (2018). ICME after one decade: success and challenges. JOM *70*, 1642–1643.

100. Lass, E.A., Stoudt, M.R., and Campbell, C.E. (2018). Systems design approach to low-cost coinage materials. Integr. Mater. Manuf. Innov. *7*, 52–69.

101. QUESTEK. A new age of materials design. www.questek.com.

102. QUESTEK EUROPE. Materials by design. www.questekeurope.com.

103. Gongora, A.E., Snapp, K.L., Whiting, E., Riley, P., Reyes, K.G., Morgan, E.F., and Brown, K.A. (2021). Using simulation to accelerate autonomous experimentation: a case study using mechanics. iScience *24*, 102262.

104. Morozovska, A.N., Eliseev, E.A., Svechnikov, G.S., and Kalinin, S.V. (2011). Nanoscale electromechanics of paraelectric materials with mobile charges: size effects and nonlinearity of electromechanical response of SrTiO3 films. Phys. Rev. B *84*, 045402.

105. Lukas, H., Fries, S.G., and Sundman, B. (2007). Computational Thermodynamics: the Calphad Method (Cambridge University Press).

106. Thermo-Calc, A.B. (2021). Calc Software 2021a. https://thermocalc.com/blog/thermo-calc-2021a-release-overview/.

107. Campbell, C.E., Boettinger, W.J., and Kattner, U.R. (2002). Development of a diffusion mobility database for Ni-based superalloys. Acta Mater *50*, 775–792.

108. Watanabe, S. (2013). A widely applicable Bayesian information criterion. J. Mach. Learn. Res. *14*, 867–897.

109. Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. Stat. Comput. *24*, 997–1016.

110. Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). Bayesian Anal *13*, 917–1007.

111. Nelson, C.T., Vasudevan, R.K., Zhang, X.H., Ziatdinov, M., Eliseev, E.A., Takeuchi, I., Morozovska, A.N., and Kalinin, S.V. (2020). Exploring physics of ferroelectric domain walls via Bayesian analysis of atomically resolved STEM data. Nat. Commun. *11*, 6361.

112. Vasudevan, R.K., Kelley, K.P., Eliseev, E., Jesse, S., Funakubo, H., Morozovska, A., and Kalinin, S.V. (2020). Bayesian inference in band excitation scanning probe microscopy for optimal dynamic model selection in imaging. J. Appl. Phys. *128*, 054105.

113. Wilson, A.G., and Izmailov, P. (2020). Bayesian deep learning and a probabilistic perspective of generalization. arXiv, arXiv:2002.08791.

114. Kingma, D.P., and Welling, M. (2019). An introduction to variational autoencoders. Foundations and Trends in Machine Learning *12*, 307–392.

115. Shao, S., and Loi, M.A. (2020). The role of the interfaces in perovskite solar cells. Adv. Mater. Interfaces *7*, 1901469.

116. Dunlap-Shohl, W.A., Zhou, Y., Padture, N.P., and Mitzi, D.B. (2019). Synthetic approaches for halide perovskite thin films. Chem. Rev. *119*, 3193–3295.

117. Li, Z., Klein, T.R., Kim, D.H., Yang, M., Berry, J.J., van Hest, M.F.A.M., and Zhu, K. (2018). Scalable fabrication of perovskite solar cells. Nat. Rev. Mater. *3*, 18017.

118. Min, H., Kim, M., Lee, S.U., Kim, H., Kim, G., Choi, K., Lee, J.H., and Seok, S.I. (2019). Efficient, stable solar cells by using inherent bandgap of α-phase formamidinium lead iodide. Science *366*, 749–753.

119. Kim, M., Kim, G.-H., Lee, T.K., Choi, I.W., Choi, H.W., Jo, Y., Yoon, Y.J., Kim, J.W., Lee, J., Huh, D., et al. (2019). Methylammonium chloride induces intermediate phase stabilization for efficient perovskite solar cells. Joule *3*, 2179–2192.

120. Jeong, J., Kim, M., Seo, J., Lu, H., Ahlawat, P., Mishra, A., Yang, Y., Hope, M.A., Eickemeyer, F.T., Kim, M., et al. (2021). Pseudo-halide anion engineering for α-FAPbI3 perovskite solar cells. Nature *592*, 381–385.

121. Bai, S., Da, P., Li, C., Wang, Z., Yuan, Z., Fu, F., Kawecki, M., Liu, X., Sakai, N., Wang, J.T.-W., et al. (2019). Planar perovskite solar cells with long-term stability using ionic liquid additives. Nature *571*, 245–250.

122. Heo, D.Y., Han, S.M., Woo, N.S., Kim, Y.J., Kim, T.-Y., Luo, Z., and Kim, S.Y. (2018). Role of additives on the performance of CsPbI3 solar cells. J. Phys. Chem. C *122*, 15903–15910.

123. Zhou, Y., Game, O.S., Pang, S., and Padture, N.P. (2015). Microstructures of organometal trihalide perovskites for solar cells: their evolution from solutions and characterization. J. Phys. Chem. Lett. *6*, 4827–4839.

124. Liu, S., Guan, Y., Sheng, Y., Hu, Y., Rong, Y., Mei, A., and Han, H. (2020). A review on additives for halide perovskite solar cells. Adv. Energy Mater. *10*, 1902492.

125. Zong, Y., Zhou, Z., Chen, M., Padture, N.P., and Zhou, Y. (2018). Lewis-adduct mediated Grain-Boundary functionalization for efficient ideal-bandgap perovskite solar cells with superior stability. Adv. Energy Mater. *8*, 1800997.

126. Hattrick-simpers, J.R., Gregoire, J.M., and Kusne, A.G. (2016). Perspective: composition–structure–property mapping in high-throughput experiments: turning data into knowledge. APL Mater *4*, 053211.

127. Green, M.L., Choi, C.L., Hattrick-simpers, J.R., Joshi, A.M., Takeuchi, I., Barron, S.C., Campo, E., Chiang, T., Empedocles, S., and Gregoire, J.M. (2017). Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies. Appl. Phys. Rev. *4*, 011105.

128. Young, S.R., Maksov, A., Ziatdinov, M., Cao, Y., Burch, M., Balachandran, J., Li, L.L., Somnath, S., Patton, R.M., Kalinin, S.V., and Vasudevan, R.K. (2018). Data mining for better material synthesis: the case of pulsed laser deposition of complex oxides. J. Appl. Phys. *123*, 11.

129. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp. 14003–14014.

130. https://paperswithcode.com/.

131. Li, X., Hoffman, J.M., and Kanatzidis, M.G. (2021). The 2D halide perovskite rulebook: how the spacer influences everything from the structure to optoelectronic device efficiency. Chem. Rev. *121*, 2230–2291.