JACS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Article

Cite This: *J. Am. Chem. Soc.* XXXX, XXX, XXX−XXX

pubs.acs.org/JACS

Learning-in-Templates Enables Accelerated Discovery and Synthesis of New Stable Double Perovskites

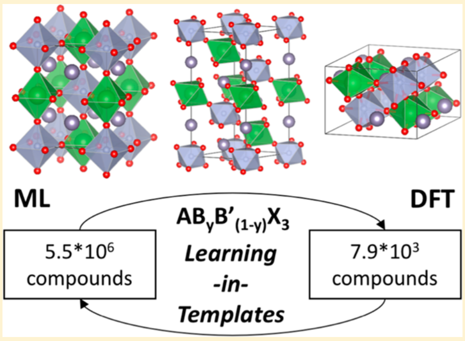
Mikhail Askerka,[†,§] Ziliang Li,[†,§] Mathieu Lempen,[†] Yanan Liu,[‡] Andrew Johnston,[†] Makhsud I. Saidaminov,[†] Zoltan Zajacz,[‡] and Edward H. Sargent*,[†]

[†] Department of Electrical and Computer Engineering, University of Toronto, 10 King's College Road, Toronto, ON, Canada M5S 3G4

[‡] Department of Earth Sciences, University of Toronto, 22 Russell Street, Toronto, ON, Canada M5S 3B1

Supporting Information

ABSTRACT: In the past three years, machine learning (ML) in combination with density functional theory (DFT) has enabled computational screening of compounds with the goal of accelerated materials discovery. Unfortunately, DFT+ML has, until now, either relied on knowledge of the atomic positions at DFT energy minima, which are *a priori* unknown, or been limited to chemical spaces of modest size. Here we report a strategy that we term learning-in-templates (LiT), wherein we first define a series of space group and stoichiometry templates corresponding to hypothesized compounds and, orthogonally, we allow any list of atoms to take on any template. The LiT approach is deployed in combination with previously established position-dependent representations and performs best with the representations that rely least on the atomic positions. Since the positions of the atoms in templates are known and do not change, LiT enables us to infer the properties of interest directly; additionally, LiT allows working with increased chemical spaces, since the same elements can take on a large number of templates. Only by using LiT were we able to span 5 × 10^6 double-perovskite compounds and achieve an acceleration factor of 700 compared to brute-force DFT, allowing us to predict never-before-screened compounds. Our findings motivated us to synthesize a new $BaCu_yTa_{(1-y)}S_3$ perovskite, which we show using an electron probe microanalyzer has a 5:3 molar ratio of Cu to Ta and, using powder X-ray diffraction (XRD) analysis combined with a DFT-based XRD simulation and fitting, indicate a new phase having an *I*4/*m* space group.



INTRODUCTION

Perovskites are a family of compounds with ABX_3 stoichiometry where the B site is situated in the center of BX_6 octahedra that can be arranged in various motifs and experience a variety of distortions/displacements.[1,2] Double perovskites with the chemical formula $AB_yB'_{(1-y)}X_3$ can accommodate two types of B cations that often differ in their oxidation states or ratios.[1] Due to the large number of cations/anions that can adopt the perovskite lattice,[3] perovskites became one of the first families of compounds to be the target of high-throughput simulations.[4−6]

It is within such large chemical spaces that accelerating high-throughput screening using machine learning is particularly applicable. The ultimate goal is to accelerate screening by several orders of magnitude when compared to the use of density functional theory (DFT) alone, while preserving the accuracy of DFT calculations.

In a joint machine learning (ML) and DFT pipeline, one first defines the chemical space to be considered, for instance, the chemical space of double perovskites. A subset of this chemical space is used to generate training data via DFT simulations. Then, using an appropriate atomistic representation of the compound, the DFT-calculated properties—often electronic (bandgaps, valence band minima, conduction band maxima, etc.) or thermodynamic (formation energies, energies above hull, etc.) properties—are machine learned.

The predictive ability of an ML method is assessed by its error on the test set (the portion of DFT-calculated compounds that is not used during training). The transferability of ML depends on the size of the predefined chemical space, i.e., how many compounds it can tackle. The efficiency of a DFT+ML pipeline is quantified by the acceleration factor, i.e., the speedup due to adding ML, which is the ratio of the total number of compounds in the chemical space to the number of DFT-computed calculations in the training set.

Designing a representation of atomistic systems[7] is a crucial decision in an ML study of materials. One family of representations that has been previously applied for perovskites[8−10] and other classes of compounds[11] is atomic feature-based representations. Here, each structure is encoded as a fixed-length vector containing the properties of individual atoms, ranging from solely their period and group[8,11] to a carefully selected set of over 10 features.[8] This has yielded high accuracy

Received: December 16, 2018
Published: January 31, 2019

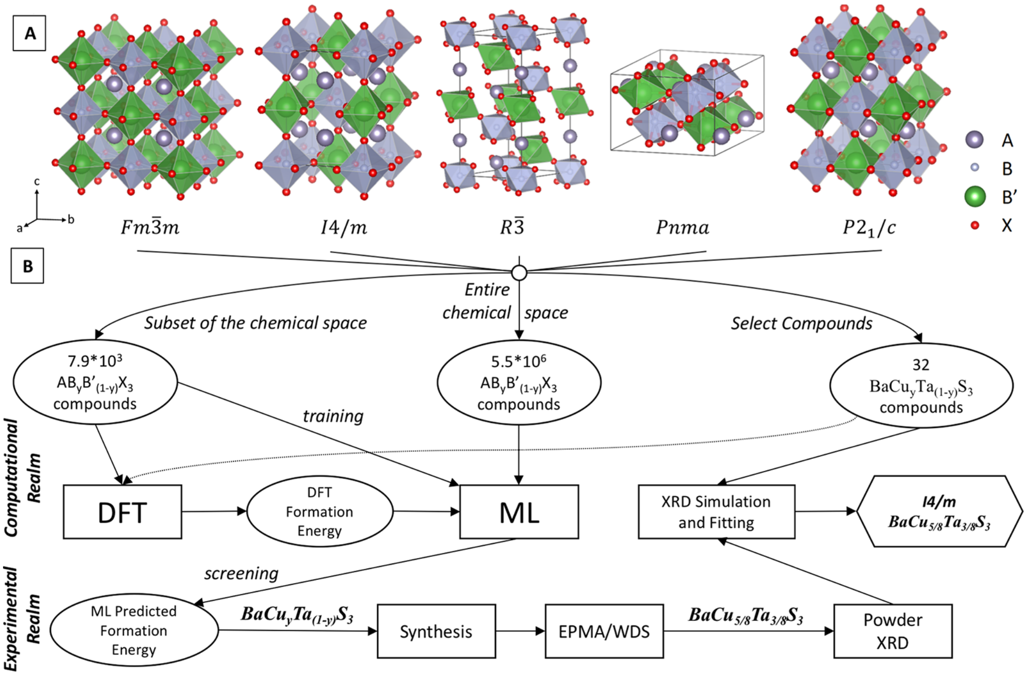


Figure 1. (A) Crystal structure of $AB_yB'_{(1-y)}X_3$ perovskites in different space groups (layered perovskite-like structure for $P2_1/c$ space group) projected along the [001] directions (up) and stereoscopic views (down). Coordination polyhedra of X atoms around B or B′ atoms are shown as octahedra. Bold line indicates the unit cell. Structures are constructed through VESTA.[10] (B) Workflow used in the present work. We define the chemical space spanned by perovskites of 5 space groups and 4 B:B′ ratios, which contains 5.5×10^6 compounds. We use a subset of that chemical space (7.9×10^3 compounds) to perform DFT calculations of perovskite formation energies and use them to train an ML algorithm. ML is then used to screen the entire chemical space and select a candidate for experimental synthesis. The experimentally synthesized compound is studied with EPMA to determine its molar composition, and powder XRD is used to get insight into its crystal structure. DFT is then performed on the possible space groups of the candidate compound, and the resulting optimized coordinates are used to simulate and fit powder XRD. The quality of this choice bears directly on the predictive ability, transferability, and efficiency of the DFT+ML approach.

toward the crystal formation energies (the highest accuracy of 0.10 eV/atom).[10]

This representation can be applied to structures that share an identical crystalline lattice and stoichiometry, but the approach fails to encode information regarding the arrangement of ions. This restriction to a single crystalline lattice and stoichiometry limits the transferability and acceleration of previously reported DFT+ML pipelines. In fact, our benchmark calculations suggest that if one is limited to the cubic chalcogenide oxide and sulfide double $A_2BB'X_6$ perovskites with 1:1 B:B′ ratio, where B and B′ have +3 and +5 oxidation states, the total size of the chemical space ($\sim 10^4$ compounds) can be explored with pure DFT at a moderate computational cost (6 CPU years compared to the common allocations of hundreds of CPU years on modern supercomputers) and therefore does not justify using ML.

Representations that typically achieve high ML accuracy (0.088 eV/atom[12]) rely on atomic positions: a sine matrix,[7] bond fractions,[13] density features,[13] radial distribution functions,[7] the Voronoi tessellation algorithm,[12,14] orbital field matrix,[15] distance-angular and element type distributions.[16] Since these representations take the positions of the atoms explicitly, they can be applied to compounds with an arbitrary structure. Unfortunately, prior deployments of such representations have required *a priori* knowledge of the optimal geometry of each structure.

In sum, prior approaches to ML have either

- been constrained to a single crystal space group, limiting the chemical space to order 10^4 already accessible using conventional full DFT; *or*
- spanned many space groups, but required *a priori* knowledge of the true crystal lattice

We took the view that this limitation could be overcome by

- first, defining a series of *templates* corresponding to the hypothesized different crystal structures a list of atoms could potentially adopt. It would be crucial that the set of templates fully spanned the set of known available space groups;
- orthogonally, allowing any list of atoms making up a hypothesized double perovskite to take on, in principle, any one of these templates;
- using DFT to train a machine learning algorithm to *predict* the formation energy for various templates that a given set of atoms would adopt;
- ultimately, using the predictive capacity of the trained DFT algorithm to capture the formation energy of a perovskite, in its template, without *a priori* knowledge nor assumption of the optimal space group of which the most stable form of the crystal would be a member

We term the method learning-in-templates (LiT) and describe herein its capacity to expand vastly the chemical space of perovskites explored computationally due to the ability of a given set of elements to take on any structural or stoichiometry template. Specifically, LiT allows us to span a 5.5×10^6 chemical space not practically accessible today using full DFT. The space spans, for the first time, structures beyond the simple cubic lattice and compositions beyond 1:1 B:B′ ratios.

This vast expansion of the chemical space is achieved while maintaining the ability to predict perovskite stability without

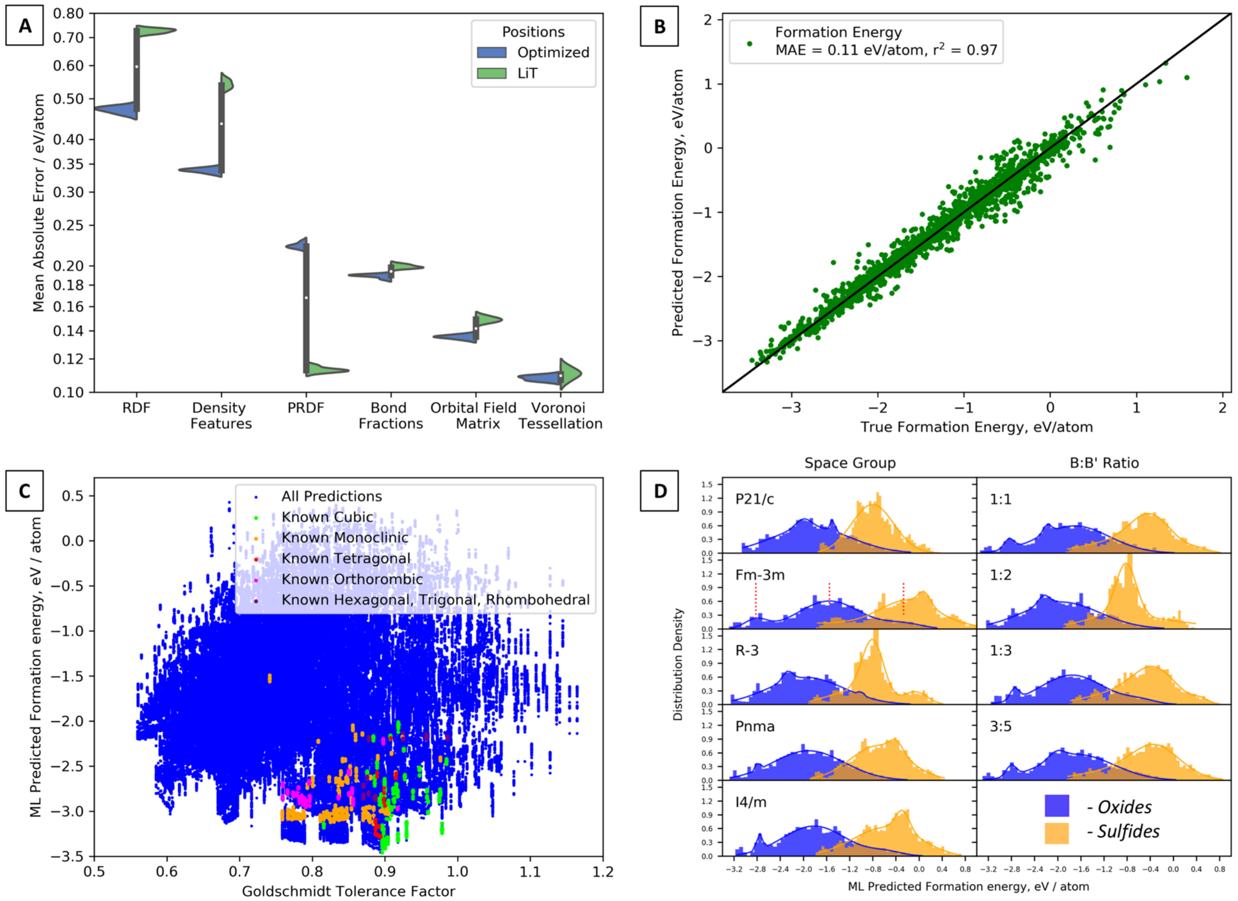


Figure 2. (A) Performance of LiT compared to the optimized positions when used with different representations. The distribution of the MAE is obtained through 5-fold cross validation. (B) Performance of the ML algorithm on test data on formation energies of perovskites with various space groups and B:B′ ratios. (C) ML predictions of the formation energies for 2.7 × 10^6 oxide perovskites. Points of colors other than blue show the formation energy ML predictions for the 1.1 × 10^3 known oxide perovskites with various space groups and 1:1 B:B′ ratio. (D) ML-predicted distribution of the formation energies across various space groups and B:B′ ratios along with the exponential fits to identify the stability regions.

prior knowledge of atomic positions. The broadened chemical space our approach allowed us to explore in turn enabled predictions of many new double perovskites, including $BaCu_{5/8}Ta_{3/8}S_3$, which we synthesized experimentally.

DISCUSSION

We consider crystal structures inspired by the most abundant structures of the 1.1× 10^4 experimentally known oxide double perovskites.[1] Figure 1A shows the space groups considered in this work: these include $Fm\bar{3}m$, $I4/m$, $R\bar{3}$, and *Pnma* for the rock-salt ordered perovskite and $P2_1/c$ for layered perovskite-like structure. Among these, $Fm\bar{3}m$ represents an aristotype cubic perovskite with the highest symmetry; the symmetry can be lowered by octahedra tilting and cation displacements.[1,17−19] A tilt around the (001) direction leaves a 4-fold axis intact while reducing the cubic symmetry to a tetragonal one. The most common tetragonal space group for double perovskites is $I4/m$ with the Glazer tilt system $a^0a^0c^-$, which indicates out-of-phase rotations of the octahedra. Considering the special case of $a^-a^-a^-$ tilting in which three axes are equally inclined to each other, a rhombic symmetry (space group $R\bar{3}$) is formed, as the tilt is equivalent to a single tilt about the (111) direction of the cubic perovskite. When considering cation displacements, the cubic symmetry could also be reduced to orthogonal symmetry with a space group *Pnma*. For these lattices, in addition to the most commonly encountered with 1:1 B^{3+}:B'^{+5} ratio, we also consider the 3:5 ratio along with previously reported 1:2[20−22] and 3:2[23] ratios.

An additional degree of freedom we consider is the arrangement of the B and B′ atoms within the same crystalline lattice and the B:B′ ratio. For example, the cubic $Fm\bar{3}m$ perovskite lattice has 40 atoms, out of which 8 are B/B′ atoms. To properly explore all possible arrangements of the 1:3 B:B′ ratio, the upper limit (not accounting for space group specific symmetry operations) for the number of structures to be considered is $C_8^2 = 28$. The number of combinations could be as large as 70 for 40-atom $Fm\bar{3}m$ and *Pnma* cells and therefore significantly increases the size of the chemical space. Thus, we construct the chemical space in the following manner:

- We considered the elements in the periodic table that exhibit +2 oxidation states for the A site, +3 oxidation state for the B site, +5 oxidation state for the B′ site, and oxygen and sulfur for the X site, resulting in 10 024 compositions.
- The following ratio/space groups combinations are considered:
- $P2_1/c$: 20-atom cell; 4 B/B′ sites; 2:2, 1:3, 3:1 ratios; total 14 combinations
- $Fm\bar{3}m$: 40-atom cell; 8 B/B′ sites; 4:4, 3:5, 5:3, 2:6, 6:2; total 238 combinations

DOI: 10.1021/jacs.8b13420

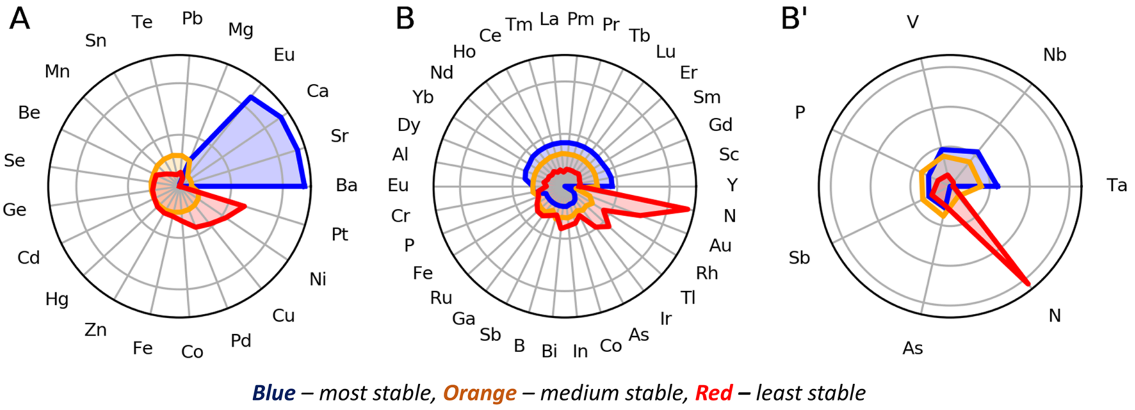


Figure 3. Regions of stability of oxide double perovskites based on the ML-predicted formation energy. Elements are colored based on the number of compounds containing that element in a given site. For the results on the sulfides, see Figure S8.

- $R\bar{3}$: 30 atoms; 6 B/B′ sites; 3:3, 2:4, 4:2; total 50 combinations
- *Pnma*: 40 atoms; 8 B/B′ sites; 4:4, 3:5, 5:3, 2:6, 6:2; total 238 combinations
- $I4/m$: 20 atoms; 4 B/B′ sites; 2:2, 1:3, 3:1; total 14 combinations
- overall 554 arrangements
- The whole space is 5.5× 10^6 compounds.
- The resulting chemical space is estimated to require 8.2 × 10^3 CPU years with brute-force DFT, which exceeds common annual supercomputer cluster allocations of 1000 CPU years by an order of magnitude.

We choose 7.9× 10^3 structures within the predefined chemical space to be calculated with DFT to form the training set. Our DFT calculations for the training set took approximately 12 CPU years (∼13 CPU hours per compound). The overall acceleration factor of our DFT+ML approach is therefore ∼700. The initial positions of each of the structures were transformed into an ML feature vector, and the formation energies were learned with a genetic algorithm-optimized[24] ML pipeline based on random forest regression (see the Methods for a more detailed description).

We investigated how the performance of LiT depends on the representation of choice. Figure 2A shows the distribution of the mean absolute error (MAE) for 5-fold cross validation for the fingerprints that were generated using the optimal positions as opposed to the templates. The difference in the performance of LiT compared to the optimized positions is most pronounced for the representations that only carry information derived from atomic coordinates, such as the radial distribution function (RDF, 0.26 eV/atom difference), the partial radial distrubution function (PRDF, 0.11 eV/atom difference), and density features (0.20 eV/atom difference). RDF does not have a notion of atomic types; therefore with templates that share identical atomic positions no learning is possible, which is reflected by a ∼0 r^2 (see Figure S3 and Table S2r^2 > 0 for the optimized positions since the information about chemical elements is indirectly encoded in the chemical bonds. On the other hand, the PRDF performs better with the templates (0.11 eV/atom) because as the atoms appear at the same positions the variations occur consistently at the same bins and the identities of the elements are encoded in different bin−element pairs.

The representations that are less sensitive to the positions, such as bond fractions, orbital field matrix, and Voronoi tessellation, show very similar performance for the optimized positions, and the templates with the latter are slightly inferior due to the loss of the structural information. Overall, in the study of double perovskites, the Voronoi tessellation fingerprint gives the lowest MAE (0.11 eV/atom) and the best performance with LiT compared to the optimized positions (0.003 eV/atom), and we thus use it in the results discussed below. The performance of our approach is in line with previously reported MAE on formation energies (Table S4); however, it offers a much higher acceleration factor due to templating. The performance of LiT is independent of the nature of the ions (Figure S6) and is consistent across the templates (Figure S7), as soon as they are equally represented in the training data. As we allowed for a symmetry change during structural relaxation, about 12% of our initial templates relaxed to a lower symmetry. We checked the performance of LiT on the structures that underwent the symmetry change and found an MAE of 0.10 eV/atom, pointing out that LiT can still be applied even if the optimized structure no longer belongs to the initial space group. We tested LiT on the literature data set of 5000 single perovskites[25] (Figures S9 and S10) and achieved a similar MAE. Additionally, we repeated this analysis for the energies above hull[26] in addition to the formation energies. As expected, the energies above hull are more challenging for an ML algorithm to learn since they have information about the additional competing phases, whereas no such information is contained in the input, which is reflected by a lower r^2 in 5-fold cross validation (0.86 for energies above hull compared to 0.97 for formation energies (Figures S3 and S13)). Despite this challenge, our model was able to achieve a low MAE of 0.091 eV/atom. The trends that we observe for the energies above hull regarding the performance of LiT compared to the optimized positions are completely analogous to those of the formation energies (Figure S11, S12, and S13). Our model performs well at smaller train set sizes, as reflected by a comparison of log−log learning curve plots (Figure S14).

We use the LiT with the Voronoi fingerprint to predict the stabilities of all of the oxide perovskites in our data set (Figure 2C). We marked our predictions green, orange, red, magenta, and purple if the $Fm\bar{3}m$, $P2_1/c$, $I4/m$, *Pnma*, or $R\bar{3}$ ML predictions corresponded to a cubic, monoclinic, tetragonal, orthorhombic, or hexagonal/trigonal/rhombohedral oxide double perovskite known experimentally.[1] The traditionally used Goldschmidt tolerance factor helps identify 79% by using a 0.8 to 1.0 cutoff rule. To achieve a similar level of confidence using the ML-predicted formation energies, −2.69 eV/atom can be used as a stability threshold since it corresponds to the 80th

DOI: 10.1021/jacs.8b13420

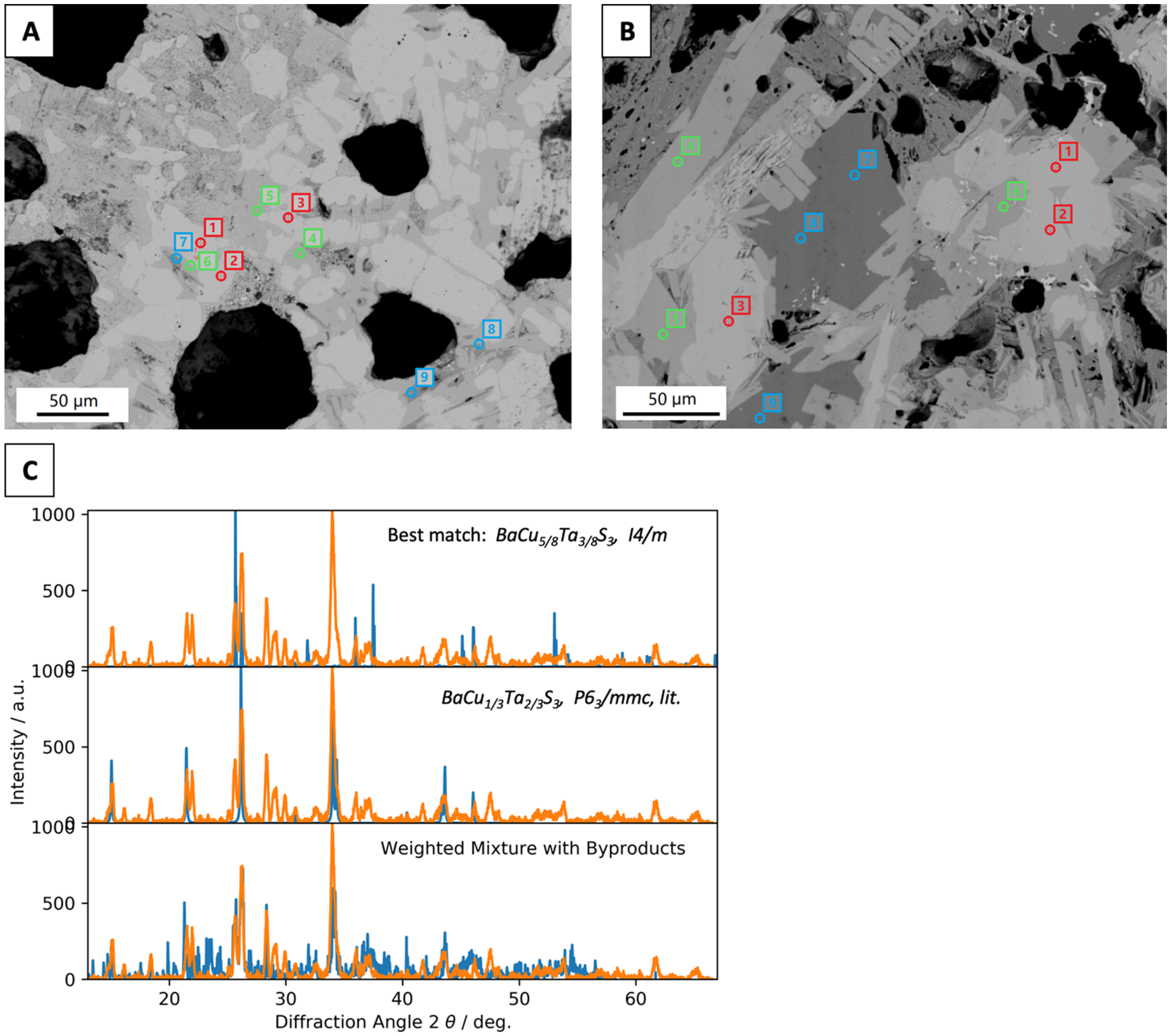


Figure 4. (A, B) Backscattered electron SEM image with spots for WDS analysis of the samples with the precursor ratio of (A) Cu:Ta = 1:1 and (B) Cu:Ta = 5:3. Spots were marked in different colors for discrimination of different phases. Red circle: $BaCu_{1/3}Ta_{2/3}S_3$ phase; green circle: new $BaCu_{5/8}Ta_{3/8}S_3$ phase; blue circle: amorphous phase containing Ba, Cu, and S. (C) Experimental XRD spectrum (orange) of the Ba−Cu−Ta−S sample along with the fitting based on the previously reported $BaCu_{1/3}Ta_{2/3}S_3$ and DFT-optimized $BaCu_{5/8}Ta_{3/8}S_3$ structure. The weighted mixture also includes the alternative DP space groups and the stable compounds on the Ba−Cu−Ta−S quaternary phase diagram.

percentile of the experimentally known oxide perovskites, which corresponds to a total of 6111 predicted stable compositions, which is about 6 times larger than the number of currently experimentally known oxide perovskites (for a similar analysis on energies above hull, see Figure S15).

We find that on average oxide perovskites are 1.01 eV/atom more stable than the sulfide ones (Figure 2D). This difference is most pronounced for the cubic perovskites ($Fm\bar{3}m$, −1.14 eV/atom) and least pronounced for the perovskites with a monoclinic structure (−0.91 eV/atom) (Figure 2, D). The most stable (5% percentile) oxide double perovskites have a hierarchy of stability, namely, $R\bar{3} < Pnma \cong I4/m < Fm\bar{3}m \cong P2_1/c$ for oxides and $R\bar{3} < Pnma \cong I4/m \cong P2_1/c < Fm\bar{3}m$ for sulfides. This indicates that for both sulfides and oxides additional relaxation of the cubic structure to lower symmetries can decrease the total energy of the system by up to 0.15 and 0.18 eV/atom for sulfides and oxides, respectively. Interestingly, the most stable oxides and sulfides behave differently with regard to the B:B′ ratio. On average, the oxides are less sensitive to the B:B′ ratio (standard deviation of 0.05 eV/atom) compared to the sulfides (0.09 eV/atom), with 1:2 being the most stable and 1:3 being the least stable for both oxides and sulfides.

Perovskites can be divided into islands of stability according to their formation energy (Figure 2C). In fact, fitting each of the distributions in Figure 2C with three Gaussian functions enables us to identify the centers and the width of stabilities. Using the oxide perovskites in the $I4/m$ space group as an example we define the region of most stable perovskites (−4 to −2.5 eV/atom), the perovskites with average stability (−2.5 to −0.8 eV/atom), and the unstable perovskites (−0.8 to 1.0 eV/atom). We then filtered out the number of compounds with particular elements that fall into each of the regions (Figure 3).

We find that the most stable perovskites are comprised of alkali earth elements (mostly Ba, Sr, and Ca) in the A site, rare

Table 1. Summary of WDS Spot Analysis Results[a]

	Cu/Ta precursor ratio	spots	Ba mol %	Cu mol %	Ta mol %	S mol %	O mol %
phase 1: $BaCu_{1/3}Ta_{2/3}S_3$	1:1	3	20.07	6.23	14.38	57.02	2.28
	5:3	37	19.90	6.08	14.49	56.33	3.20
phase 2: $BaCu_{5/8}Ta_{3/8}S_3$	1:1	3	20.12	13.26	8.01	56.81	1.77
	5:3	28	19.89	13.23	7.91	56.49	2.44
phase 3: Ba−Cu−S	1:1	4	16.28	29.97	1.40	51.47	0.88
	5:3	8	8.96	50.96	0.00	38.94	1.13

[a] Mole percentages of each element are average values of all effective spots.

earth elements in the B site, and Nb and Ta in the B′ site. In the average stability region, elements from groups 12 and 14 constitute the A site, Ce and group 13 constitute the B site, and V and group 15 are responsible for the B′ site. This region also shows the limitations of using the formation energy as a stability indicator since, for example, P (in the B′ site) oxide perovskites are predicted to have average stability, but there is no indication of such compounds in over 50 years of experimental literature. The least stable perovskites have mostly elements in their rare oxidation states (Se^{2+} and Te^{2+} in the A site) and nonmetals in the B′ site (nitrogen). These trends are in general agreement with a recent single perovskite study.[27]

The stability of the individual sulfides is linearly dependent on the stability of corresponding oxides (Figure S1): a simple linear fit yielded an r^2 of >0.81 for all space groups and B:B′ ratios. In all cases, the slope of the linear fit is less than 1 (varies from 0.81 for the *Pnma* group to 0.92 for the $Fm\bar{3}m$ group, 0.83 for 1:2 ratio to 0.90 to 3:5 ratio (Figure S2)), indicating that the gap between the oxides and sulfides shrinks as the stability decreases. In fact, certain sulfides in the region of average stability occasionally become more stable than corresponding oxides. An example is V and Sb (in the B′ site) based Ni−Fe, Ni−Co, and Co−Fe perovskites. This indicates that one might be able to synthesize these sulfide double perovskites from the corresponding oxides.

To verify the functionality of our ML+DFT approach in practical applications, we synthesized a new stable perovskite compound. We chose $BaCu_xTa_{(1-x)}S_3$ (Figure 2D, green points) since its average predicted formation energy is −1.35 eV/atom, which is within the 7.9% topmost stable sulfides, and it is composed of earth-abundant elements. We note that this perovskite has not been predicted previously: the noninteger B:B′ ratio combined with the lower symmetry space group this material possesses has resulted in this material hitherto being obscured. This is to our knowledge the first demonstrated synthesis of an ML-predicted perovskite, demonstrating the practicality of ML for use in next-generation materials discovery.

We performed a solid synthesis reaction[20] (see Methods for a description) and analyzed the resulting compound with an electron probe microanalyzer (EPMA) equipped with wavelength dispersive spectrometers (WDS) and powder X-ray diffraction (XRD) to confirm its composition and crystal structure, respectively. We initially performed the synthesis of the target compound $BaCu_xTa_{(1-x)}S_3$ using the 1:1 B:B′ stoichiometry of the precursors and used WDS to confirm its composition. We chose to use WDS because (a) its beam size and penetration depth (1 μm, and sub-μm to 20 μm correspondingly) matched well the size of phase regions that we obtained (20−50 μm) and (b) WDS has been widely used in the past for composition studies of sulfide compounds.[28,29] Additionally, WDS provided more reliable and consistent results compared to laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS) due to the limitations of the latter to determine stoichiometric amounts of sulfur given a commonly used internal standard (silicate glass doped with trace amounts of sulfur).

WDS spot analysis of sample 1 (Figure 4A and Table 1) showed three well-defined phases with different elemental mole ratios. Phase 1 is the hexagonal $BaCu_{1/3}Ta_{2/3}S_3$ (further confirmed with XRD) reported by Bu et al.[20] In this phase, the molar ratio of Cu and Ta is 1:2. Phase 2 contains all four precursor elements in the molar ratio of Ba:[Cu/Ta]:S = 1:1:3, which suggests a new perovskite phase. The Cu and Ta are off-stoichiometric with a calculated molar ratio of 5.007:2.993. This ratio is consistent across multiple electron beam shots and is likely 5:3, resulting in the chemical composition $BaCu_{5/8}Ta_{3/8}S_3$.

Phase 3 contains Ba, Cu, and S and is likely amorphous, as the molar ratio of Ba, Cu, and S varies with the Cu:Ta precursor ratio (Table 1). Each phase also contains a negligible portion of oxygen, which could be attributed to surface oxygen adsorption and partial substitution of S sites in the lattice. We then repeated the synthesis with a 5:3 Cu:Ta precursor ratio (sample 2), and the WDS spot analysis showed similar results with all three phases present.

To gain further insight on the crystal structure of the newly identified $BaCu_{5/8}Ta_{3/8}S_3$ phase, we performed powder XRD analysis of sample 2 (Figure 3C, orange). Due to a high complexity of getting structural information directly from the powder XRD pattern, we used DFT models to help identify the peaks. Our simulations (see method description for more information) showed that the experimentally observed peaks can be fitted using a combination of a previously reported $P6_3/mmc$ phase, a new double-perovskite phase, and minor portion of BaS phase. The $P2_1/c$ and $I4/m$ space groups give similar contributions (12% and 9%, respectively) to the overall spectrum and can be used to fit the double-perovskite portion. While both space groups are equally probable based on XRD fitting, DFT calculations indicate that the $I4/m$ phase is more (by 0.27 eV/atom) stable than the $P2_1/c$ phase, and therefore, we think that the newly synthesized compound forms the $I4/m$ lattice. Given that the compound has a perovskite structure, the Cu:Ta molar ratio of 5:3 suggests the average oxidation state of Cu would be +3.33, indicating that there is a mixture of Cu(III) and Cu(IV) or Cu(II) and Cu(IV). While these are rare oxidation states for Cu, mixed Cu(III)/Cu(IV) have previously been found in oxide perovskites.[30,31]

■ CONCLUSIONS

In sum, the unified DFT+ML strategy employed herein to find new perovskite compounds enabled a useful MAE of 0.11 eV/atom with an exceptionally high acceleration factor of 700. This enabled us to explore a never-before-tested chemical space of perovskites, expanding our search across five different space groups and four B:B′ stoichiometric ratios. Leveraging the wide array of new compounds predicted, we validated our approach by synthesizing a new phase of $BaCu_{5/8}Ta_{3/8}S_3$ double perovskite, with an $I4/m$ perovskite lattice, demonstrating the direct applicability of ML as a tool for the discovery of new optoelectronic materials.

■ METHODS

First-Principles Calculations. We used ground-state DFT with the Perdew−Burke−Ernzerhof (PBE)[32] GGA exchange−correlation functional as implemented in the Vienna ab Initio Simulation Package (VASP)[33] to perform structural optimization of the perovskites. All calculations allowed for spin polarization. We used a plane wave energy cutoff of 520 eV and Gaussian smearing (0.05 eV wide) to converge the electronic problem. The Monkhorst−Pack k-point mesh of 2 × 2 × 2 (density >900/atom) and the force convergence criterion of 0.0005 eV/atom × N atoms in the unit cell were used as implemented in the MPRelaxSet class of the Pymatgen python package.[26,34,35] The structures of the double perovskites were prepared using the ASE[36] python module, and the structure templates for various space groups were taken from the literature.[17,37,38] Structure optimizations allowed for a change in the crystal symmetry. The calculated total energies were then transformed into formation energies using the energies of the isolated atoms as a reference, as implemented in Pymatgen.

Machine Learning. To represent the atomistic systems, we employed the Voronoi tessellation algorithm as implemented in Magpie[12] through the Catlearn[39] python interface, resulting in a *1x271* feature vector for each data point (a discussion on the relative importance of structural vs Voronoi tessellation attributes can be found in ref 12). All other representations were prepared using the Matminer python package.[13] We optimized our ML algorithms using a genetic algorithm optimization of the architecture of a random forest based regressor as implemented in the TPOT[24] python module. We used 20 generations with a population size of 20 architectures for all representations. If the feature vector length exceeded 300 elements, which was the case with PRDF and the Orbital Field Matrix, we kept only the first 300 principal components to train the regression model. The MAE and the r^2 of 5-fold cross validation for various representations are presented in Tables S1 and S2, and Figures S2 and S3 show the performance on select samples. The performance of a number of classical regression algorithms[40] is given in Table S3 for comparison.

XRD Simulation and Fitting. In order to identify the peaks in the powder XRD spectrum, we used atomic positions from DFT calculations to simulate the XRD peak positions and intensities using the Xrayutilities[41] python module. For our analysis, we randomly sampled six structures for each of the space groups of the $BaCu_xTa_{(1-x)}S_3$ perovskite. Additionally, we augmented this list with 23 stable compounds on the quaternary Ba−Cu−Ta−S phase diagram. We then performed the least-squares fitting as implemented in the lmfit[42] python module of the experimental XRD spectrum using a linear combination of the simulated spectra as well as the experimental spectrum reported in ref 20. In the fitting the spectra were allowed to scale (by a factor of 0.01 to 100) and move as a whole (with the bounds of −0.2 to 0.2 to account for the error in the DFT-predicted lattice parameters). The results for the fit are $Pnma$: 0.07, $I4/m$: 0.09, $R\bar{3}$: 0.04, $Fm\bar{3}m$: 0.04, $P2_1/c$: 0.12, phase 1: 0.74.

Synthesis of Double Perovskites. To synthesize as-predicted on-stoichiometric compound Ba_2CuTaS_6, BaS powder (99.7%, Alfa Aesar), Cu powder (99.999%, Alfa Aesar), Ta powder (99.98%, Alfa Aesar), and S powder (99.5%, Alfa Aesar) were mixed in a fused-silica tube in a molar ratio of BaS/Cu/Ta/S = 2:1:1:4. The tube was evacuated to 2.7 Pa (20 mTorr), sealed, heated to 700 °C at 60 °C/h, and kept for 48 h. The tube was then cooled slowly (3 °C/h) to 400 °C before the furnace was turn off. For the off-stoichiometry compound $BaCu_{5/8}Ta_{3/8}S_3$, the molar ratio of precursors changed to BaS/Cu/Ta/S = 8:5:3:24, while the heating profile remained the same. Both products are black porous bulk.

WDS Spot Analysis. Compositions of the synthesized products were quantitatively evaluated with a JEOL JXA-8230 electron probe microanalyzer, equipped with five wavelength dispersive spectrometers, housed in the Earth Sciences Department, University of Toronto. The as-prepared bulk products were first embedded in epoxy pucks, polished to 1 μm or better with a monocrystalline diamond suspension, and then carbon coated before characterization.

An accelerating voltage of 15 kV, a beam current of 10 nA, and a focused beam were used to analyze Ba, Cu, S, Sb, Ta, Ce, Y, and O in all the synthesized products (Sb, Ce, and Y are not included in the Ba−Cu−Ta−S samples but are analyzed due to other samples unrelated to this work). A counting time of 20 s on peak and 10 s on each side of the background were used for all the analyses. Spot positions were chosen such that the chemical composition was identical at a region with a radius of at least 20 μm. According to such criteria, we chose three main phases for WDS spot analysis, and each phase contains at least three spots. For each spot, mass percentages of each element were collected and translated to molar percentages if the total mass percentage falls into 100 ± 1.5%. Table 1 summarizes the average molar ratio of each detected element in each phase. Apart from the four precursor elements, each phase also contains a small portion of oxygen, which could be attributed to surface oxygen adsorption and partial substitution of S sites in the lattice.

XRD Analysis. Crystal structures of the products were investigated with a Phillips PW1830 powder X-ray diffractometer at the University of Toronto. Cu Kα radiation (λ = 1.5406 Å), coupled with a Ni filter between the X-ray source and the sample, was employed as the incident beam, and a xenon gas proportional detector was used for all the XRD scans. Scanning range was set to be between 13° and 67° (2θ), which covers all the major feature peaks for the possible phase members, with a step size of 0.02° (2θ) and a dwell time of 1.25 s per step.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/jacs.8b13420.

> Supplementary figures as well as the instructions on the details of the EPMA and XRD analyses as well as data to reproduce the figures and the demo-code. (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*ted.sargent@utoronto.ca

ORCID

Mikhail Askerka: 0000-0003-3134-6496

Edward H. Sargent: 0000-0003-0396-6495

Author Contributions

§M. Askerka and Z. Li contributed equally to the work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

E.H.S. and all coauthors from the Department of Electrical and Computer Engineering at the University of Toronto acknowledge the financial support from the Ontario Research Fund−Research Excellence Program, the Natural Sciences and Engineering Research Council of Canada (NSERC), and the support from the Global Research Outreach program of Samsung Advanced Institute of Technology. DFT computations were performed on the Niagara supercomputer at the SciNet HPC Consortium[38] and the Blue Gene SOSCIP/IBM (Southern Ontario Smart Computing Innovation Platform, funded by the Federal Economic Development Agency of Southern Ontario, the Province of Ontario, IBM Canada Ltd., Ontario Centres of Excellence, Mitacs and 15 Ontario academic member institutions). SciNet is funded by the Canada Foundation for

Innovation under the auspices of Compute Canada, the Government of Ontario. M.I.S. acknowledges the support of Banting Postdoctoral Fellowship Program, administered by the Government of Canada. We thank Prof. Oleksandr Voznyy for helpful discussions.

■ REFERENCES

(1) Vasala, S.; Karppinen, M. A2B′ B ″O6 perovskites: a review. *Prog. Solid State Chem.* **2015**, *43*, 1−36.

(2) Woodward, P. M. Octahedral tilting in perovskites. I. Geometrical considerations. *Acta Crystallogr., Sect. B: Struct. Sci.* **1997**, *53*, 32−43.

(3) Reller, A.; Williams, T. Perovskites: Chemical Chameleons. *Chem. Br.* **1989**, *25*, 1227−1230.

(4) Castelli, I. E.; Olsen, T.; Datta, S.; Landis, D. D.; Dahl, S.; Thygesen, K. S.; Jacobsen, K. W. Computational screening of perovskite metal oxides for optimal solar light capture. *Energy Environ. Sci.* **2012**, *5*, 5814−5819.

(5) Armiento, R., Kozinsky, B., Hautier, G., Fornari, M., Ceder, G. High-throughput screening of perovskite alloys for piezoelectric performance and thermodynamic stability, *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89*, DOI: 10.1103/PhysRevB.89.134103.

(6) Castelli, I. E., Thygesen, K. S., Jacobsen, K. W. Bandgap engineering of double perovskites for one-and two-photon water splitting, *MRS Online Proc. Libr.* **2013**, *1523*, DOI: 10.1557/opl.2013.450.

(7) Schutt, K. T., Glawe, H., Brockherde, F., Sanna, A., Muller, K. R., Gross, E. K. U. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties, *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89*, DOI: 10.1103/PhysRevB.89.205118.

(8) Schmidt, J.; Shi, J.; Borlido, P.; Chen, L.; Botti, S.; Marques, M.A. Predicting the Thermodynamic Stability of Solids Combining Density Functional Theory and Machine Learning. *Chem. Mater.* **2017**, *29*, 5090−5103.

(9) Pilania, G.; Mannodi-Kanakkithodi, A.; Uberuaga, B. P.; Ramprasad, R.; Gubernatis, J. E.; Lookman, T. Machine learning bandgaps of double perovskites. *Sci. Rep.* **2016**, *6*, 19375.

(10) Pilania, G.; Gubernatis, J. E.; Lookman, T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput. Mater. Sci.* **2017**, *129*, 156−163.

(11) Faber, F. A.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Machine Learning Energies of 2 Million Elpasolite (ABC(2)D(6)) Crystals. *Phys. Rev. Lett.* **2016**, *117*, DOI: 10.1103/PhysRevLett.117.135502.

(12) Ward, L.; Liu, R.; Krishna, A.; Hegde, V. I.; Agrawal, A.; Choudhary, A.; Wolverton, C. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2017**, *96*, 024104.

(13) Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J.; Bystrom, K.; Dylla, M.; Chard, K. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **2018**, *152*, 60−69.

(14) Ward, L., Agrawal, A., Choudhary, A., Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials, *Npj Comput. Mater.* **2016**, *2*, DOI: 10.1038/npjcompumats.2016.28.

(15) Lam Pham, T.; Kino, H.; Terakura, K.; Miyake, T.; Tsuda, K.; Takigawa, I.; Chi Dam, H. *Sci. Technol. Adv. Mater.* **2017**, *18*, 756−765.

(16) Faber, F. A.; Christensen, A. S.; Huang, B.; von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.* **2018**, *148*, 241717.

(17) Barnes, P. W.; Lufaso, M. W.; Woodward, P. M. Structure determination of A(2)M(3+)TaO(6) and A(2)M(3+)NbO(6) ordered perovskites: octahedral tilting and pseudosymmetry. *Acta Crystallogr., Sect. B: Struct. Sci.* **2006**, *62*, 384−396.

(18) Glazer, A. The classification of tilted octahedra in perovskites. *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **1972**, *28*, 3384−3392.

(19) Lufaso, M. W.; Barnes, P. W.; Woodward, P. M. Structure prediction of ordered and disordered multiple octahedral cation perovskites using SPuDS. *Acta Crystallogr., Sect. B: Struct. Sci.* **2006**, *62*, 397−410.

(20) Bu, K.; He, J.; Wang, D.; Zheng, C.; Huang, F. Crystal structure of the mixed-metal tris-ulfide BaCu1/3Ta2/3S3. *Acta Crystallogr. E Crystallogr. Commun.* **2017**, *73*, 713−715.

(21) Park, H. M.; Lee, H. J.; Cho, Y. K.; Nahm, S. Crystal structures of (Ba 1− x La x)[Mg (1+ x)/3 Nb (2− x)/3] O 3 with 0.9≤ x≤ 1.0. *J. Mater. Res.* **2003**, *18*, 1003−1010.

(22) Ivanov, S.; Eriksson, S.-G.; Tellgren, R.; Rundlof, H. Evolution of the atomic and magnetic structure of Sr3Fe2WO9: A temperature dependent neutron powder diffraction study. *Mater. Res. Bull.* **2001**, *36*, 2585−2596.

(23) Zeng, Z.; Fawcett, I. D.; Greenblatt, M.; Croft, M. Large magnetoresistance in double perovskite Sr2Cr1.2Mo0.8O6-*δ*. *Mater. Res. Bull.* **2001**, *36*, 705−715.

(24) (a) Olson, R. S.; Moore, J. H. In *Workshop on Automatic Machine Learning*; 2016; pp 66−74. (b) Olson, R. S.; Moore, J. H. In *Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science*; Proceedings of GECCO, 2016; pp 485−492.

(25) Emery, A. A.; Wolverton, C. High-throughput DFT calculations of formation energy, stability and oxygen vacancy formation energy of ABO3 perovskites. *Sci. Data* **2017**, *4*, 170153.

(26) Balachandran, P. V., Emery, A. A., Gubernatis, J. E., Lookman, T., Wolverton, C., Zunger, A. Predictions of new ABO(3) perovskite compounds by combining machine learning and density functional theory. *Phys. Rev. Mater.***2018**, *2*, DOI: 10.1103/PhysRevMaterials.2.043802.

(27) Versavel, M. Y.; Haber, J. A. Thin-film growth of low temperature lead antimony sulfide plagionite phases. *Chem. Commun. (Cambridge, U. K.)* **2006**, 3543−3545.

(28) Feng, K.; Zhang, X.; Yin, W.; Shi, Y.; Yao, J.; Wu, Y. New Quaternary Rare-Earth Chalcogenides Ba Ln Sn2Q6 (Ln= Ce, Pr, Nd, Q= S; Ln= Ce, Q= Se): Synthesis, Structure, and Magnetic Properties. *Inorg. Chem.* **2014**, *53*, 2248−2253.

(29) Darracq, S.; Kang, S.; Choy, J.; Demazeau, G. Stabilization of the Mixed Valence Cu (III)/Cu (IV) in the Perovskite Lattice of La1-xSrxCuO3 under High Oxygen Pressure. *J. Solid State Chem.* **1995**, *114*, 88−94.

(30) Demazeau, G.; Darracq, S.; Choy, J. High oxygen pressures and the stabilization of a new mixed valence Cu (III)/Cu (IV). *High Pressure Res.* **1994**, *12*, 323−328.

(31) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.

(32) Kresse, G.; Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1996**, *54*, 11169−11186.

(33) Monkhorst, H. J.; Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **1976**, *13*, 5188−5192.

(34) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, 011002.

(35) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **2013**, *68*, 314−319.

(36) Larsen, A. H., Mortensen, J. J., Blomqvist, J., Castelli, I. E., Christensen, R., Dulak, M., Friis, J., Groves, M. N., Hammer, B., Hargus, C., Hermes, E. D., Jennings, P. C., Jensen, P. B., Kermode, J., Kitchin, J. R., Kolsbjerg, E. L., Kubal, J., Kaasbjerg, K., Lysgaard, S., Maronsson, J. B., Maxson, T., Olsen, T., Pastewka, L., Peterson, A., Rostgaard, C., Schiotz, J., Schutt, O., Strange, M., Thygesen, K. S., Vegge, T., Vilhelmsen, L., Walter, M., Zeng, Z. H., Jacobsen, K. W. The atomic

simulation environment-a Python library for working with atoms, *J. Phys.: Condens. Matter* **2017**, *29*, 273002.

(37) Saines, P. J.; Kennedy, B. J.; Elcombe, M. M. Structural phase transitions and crystal chemistry of the series Ba(2)LnB ’ O-6 (Ln = lanthanide and B ’ = Nb5+ or Sb5+). *J. Solid State Chem.* **2007**, *180*, 401−409.

(38) Radaelli, P. G.; Iannone, G.; Marezio, M.; Hwang, H. Y.; Cheong, S. W.; Jorgensen, J. D.; Argyriou, D. N. Structural effects on the magnetic and transport properties of perovskite A(1-x)A(x)’MnO3 (x = 0.25, 0.30). *Phys. Rev. B: Condens. Matter Mater. Phys.* **1997**, *56*, 8265−8276.

(39) CatLearn.

(40) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Machine Learning Res.* **2011**, *12*, 2825−2830.

(41) Kriegner, D.; Wintersberger, E.; Stangl, J. Xrayutilities: a versatile tool for reciprocal space conversion of scattering data recorded with linear and area detectors. *J. Appl. Crystallogr.* **2013**, *46*, 1162−1170.

(42) Newville, M., Stensitzki, T., Allen, D. B., Rawlik, M., Ingargiola, A., Nelson, A. LMFIT: non-linear least-square minimization and curve-fitting for Python. *Astrophysics Source Code Library* **2016**, https://zenodo.org/record/11813#.XEuJ1s9KhTZ.

DOI: 10.1021/jacs.8b13420