

## MATERIALS SCIENCE

## New tolerance factor to predict the stability of perovskite oxides and halides

Christopher J. Bartel<sup>1\*</sup>, Christopher Sutton<sup>2</sup>, Bryan R. Goldsmith<sup>3</sup>, Runhai Ouyang<sup>2</sup>, Charles B. Musgrave<sup>1,4,5</sup>, Luca M. Ghiringhelli<sup>2\*</sup>, Matthias Scheffler<sup>2</sup>

Predicting the stability of the perovskite structure remains a long-standing challenge for the discovery of new functional materials for many applications including photovoltaics and electrocatalysts. We developed an accurate, physically interpretable, and one-dimensional tolerance factor,  $\tau$ , that correctly predicts 92% of compounds as perovskite or nonperovskite for an experimental dataset of 576  $ABX_3$  materials ( $X = O^{2-}, F^-, Cl^-, Br^-, I^-$ ) using a novel data analytics approach based on SISSO (sure independence screening and sparsifying operator).  $\tau$  is shown to generalize outside the training set for 1034 experimentally realized single and double perovskites (91% accuracy) and is applied to identify 23,314 new double perovskites ( $A_2BB'X_6$ ) ranked by their probability of being stable as perovskite. This work guides experimentalists and theorists toward which perovskites are most likely to be successfully synthesized and demonstrates an approach to descriptor identification that can be extended to arbitrary applications beyond perovskite stability predictions.

## INTRODUCTION

Crystal structure prediction from chemical composition continues as a persistent challenge to accelerated materials discovery (1, 2). Most approaches capable of addressing this challenge require several computationally demanding electronic-structure calculations for each material composition, limiting their use to a small set of materials (3–6). Alternatively, descriptor-based approaches enable high-throughput screening applications because they provide rapid estimates of material properties (7, 8). Notably, the Goldschmidt tolerance factor,  $t$  (9), has been used extensively to predict the stability of the perovskite structure based only on the chemical formula,  $ABX_3$ , and the ionic radii,  $r_i$ , of each ion ( $A, B, X$ )

$$t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)} \quad (1)$$

The perovskite crystal structure, as shown in Fig. 1A, is defined as any  $ABX_3$  compound with a network of corner-sharing  $BX_6$  octahedra surrounding a larger  $A$ -site cation ( $r_A > r_B$ ), where the cations,  $A$  and  $B$ , can span the periodic table and the anion,  $X$ , is typically a chalcogen or halogen. Distortions from the cubic structure can arise from size mismatch of the cations and anion, which results in additional perovskite structures and nonperovskite structures. The  $B$  cation can also be replaced by two different ions, resulting in the double perovskite formula,  $A_2BB'X_6$  (Fig. 1B). Single and double perovskite materials have exceptional properties for a variety of applications such as electrocatalysis (10), proton conduction (11), ferroelectrics (12) (using oxides,  $X = O^{2-}$ ), battery materials (13) (using fluorides,  $X = F^-$ ), as well as photovoltaics (14) and optoelectronics (15) (using the heavier halides,  $X = Cl^-, Br^-, I^-$ ).

The first step in designing new perovskites for these applications is typically the assessment of stability using  $t$ , which has informed the design of perovskites for over 90 years. However, as reported in recent

studies, its accuracy is often insufficient (16). Considering 576  $ABX_3$  solids experimentally characterized at ambient conditions and reported in (17–19) (see Fig. 1C for the  $A, B$ , and  $X$  elements in this set),  $t$  correctly distinguishes between perovskite and nonperovskite for only 74% of materials and performs considerably worse for compounds containing heavier halides [chlorides (51% accuracy), bromides (56%), and iodides (33%)] than for oxides (83%) and fluorides (83%) (Fig. 2A, fig. S1, and table S1). This deficiency in generalization to halide perovskites severely limits the applicability of  $t$  for materials discovery.

In this work, we present a new tolerance factor ( $\tau$ ), which has the form

$$\tau = \frac{r_X}{r_B} - n_A \left( n_A - \frac{r_A/r_B}{\ln(r_A/r_B)} \right) \quad (2)$$

where  $n_A$  is the oxidation state of  $A$ ,  $r_i$  is the ionic radius of ion  $i$ ,  $r_A > r_B$  by definition, and  $\tau < 4.18$  indicates perovskite. A high overall accuracy of 92% for the experimental set (94% for a randomly chosen test set of 116 compounds) and nearly uniform performance across the five anions evaluated [oxides (92% accuracy), fluorides (92%), chlorides (90%), bromides (93%), and iodides (91%)] is achieved with  $\tau$  (Fig. 2B, fig. S1, and table S1). Like  $t$ , the prediction of perovskite stability using  $\tau$  requires only the chemical composition, allowing the tolerance factor to be agnostic to the many structures that are considered perovskite. In addition to predicting if a material is stable as perovskite,  $\tau$  also provides a monotonic estimate of the probability that a material is stable in the perovskite structure. The accurate and probabilistic nature of  $\tau$ , as well as its generalizability over a broad range of single and double perovskites, allows new physical insights into the stability of the perovskite structure and the prediction of thousands of new double perovskite oxides and halides, 23,314 of which are provided here and ranked by their probability of being stable in the perovskite structure.

## RESULTS AND DISCUSSION

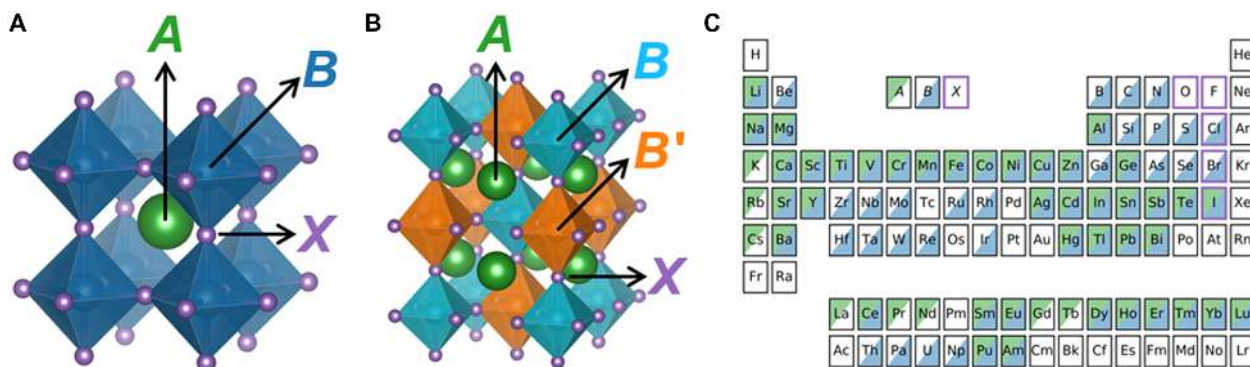
## Finding an improved tolerance factor to predict perovskite stability

One key aspect of the performance of  $t$  is how well the sum of ionic radii estimates the interatomic bond distances for a given structure.

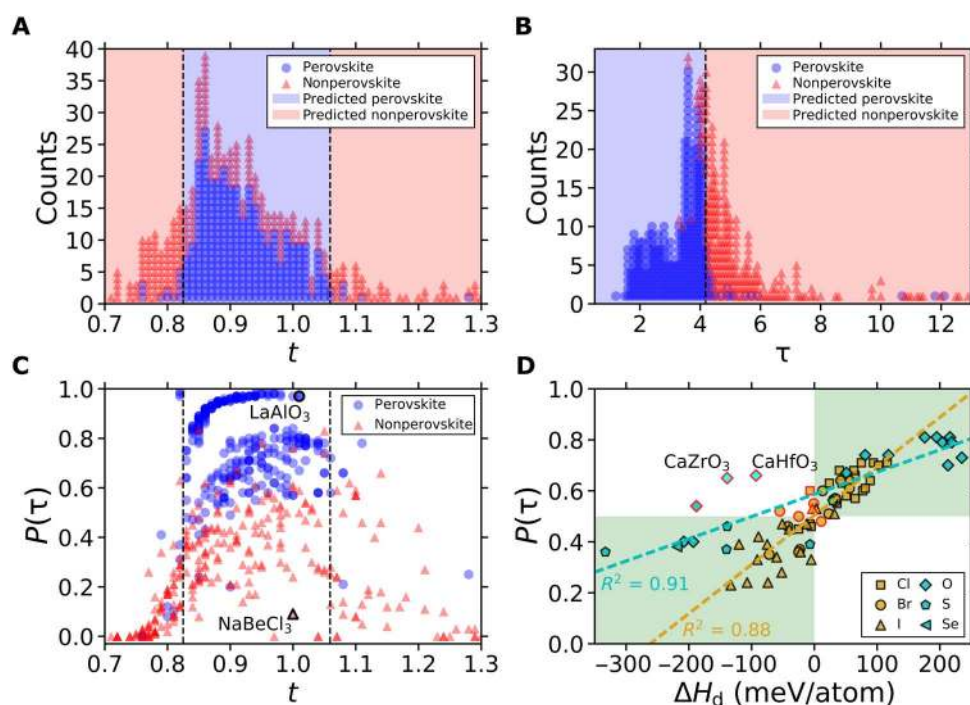
Copyright © 2019  
The Authors, some  
rights reserved;  
exclusive licensee  
American Association  
for the Advancement  
of Science. No claim to  
original U.S. Government  
Works. Distributed  
under a Creative  
Commons Attribution  
License 4.0 (CC BY).

<sup>1</sup>Department of Chemical and Biological Engineering, University of Colorado Boulder, Boulder, CO 80309, USA. <sup>2</sup>Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, D-14195 Berlin, Germany. <sup>3</sup>Department of Chemical Engineering, University of Michigan, Ann Arbor, MI 48109-2136, USA. <sup>4</sup>Department of Chemistry, University of Colorado Boulder, Boulder, CO 80309, USA. <sup>5</sup>Materials and Chemical Science and Technology Center, National Renewable Energy Laboratory, Golden, CO 80401, USA.

\*Corresponding author. Email: ghiringhelli@fhi-berlin.mpg.de (L.M.G.); christopher.bartel@colorado.edu (C.J.B.)



**Fig. 1. Perovskite structure and composition.** (A)  $ABX_3$ , in the cubic single perovskite structure ( $Pm\bar{3}m$ ), where the A cation is surrounded by a network of corner-sharing  $BX_6$  octahedra. (B)  $A_2BB'X_6$ , in the rock salt double perovskite structure ( $Fm\bar{3}m$ ), where the A cations are surrounded by an alternating network of  $BX_6$  and  $B'X_6$  octahedra. In this structure, inverting the B and B' cations results in an equivalent structure. While the ideal cubic structures are shown here, perovskites may also adopt various noncubic structures. (C) Map of the elements that occupy the A, B, and/or X sites within the 576 compounds experimentally characterized as perovskite or nonperovskite at ambient conditions and reported in (17–19).



**Fig. 2. Assessing the performance of the improved tolerance factor,  $\tau$ .** (A) A decision tree classifier determines that the optimal bounds for perovskite formability using the Goldschmidt tolerance factor ( $t$ ) are  $0.825 < t < 1.059$ , which yields a classification accuracy of 74% for 576 experimentally characterized  $ABX_3$  solids. (B)  $\tau$  achieves a classification accuracy of 92% on the set of 576  $ABX_3$  solids based on perovskite classification for  $\tau < 4.18$ , with this decision boundary identified using a one-node decision tree. All classifications made by  $t$  and  $\tau$  on the experimental dataset are provided in table S1. The largest value of  $\tau$  in the experimental set of 576 compounds is 181.5; however, all points with  $\tau > 13$  are correctly labeled as nonperovskite and are not shown to highlight the decision boundary. The outlying compounds at  $\tau > 10$  that are labeled perovskite yet have large  $\tau$  are  $PuVO_3$ ,  $AmVO_3$ , and  $PuCrO_3$ , which may indicate poorly defined radii or incorrect experimental characterization. (C) Comparison of Platt-scaled classification probabilities,  $P(\tau)$ , versus  $t$ .  $LaAlO_3$  and  $NaBeCl_3$  are labeled to highlight the variation in  $P(\tau)$  at nearly constant  $t$ . (D) Comparison between  $P(\tau)$  and the decomposition enthalpy ( $\Delta H_d$ ) for 36 double perovskite halides calculated using density functional theory (DFT) in the  $Fm\bar{3}m$  structure in (32) and 37 single and double perovskite chalcogenides and halides in the  $Pm\bar{3}m$  structure in (33). The legend corresponds with the anion, X. Positive decomposition enthalpy ( $\Delta H_d > 0$ ) indicates that the structure is stable with respect to decomposition into competing compounds. The green and white shaded regions correspond with agreement and disagreement between the calculated  $\Delta H_d$  and the classification by  $\tau$ . Points of disagreement are outlined in red.  $CaZrO_3$  and  $CaHfO_3$  are labeled because they are known to be stable in the perovskite structure, although they are unstable in the cubic structure (34, 35). For this reason, the best-fit line for the chalcogenides ( $X = O^{2-}, S^{2-}, Se^{2-}$ ) excludes these two points.

Shannon's revised effective ionic radii (20) based on a systematic empirical assessment of interatomic distances in nearly 1000 compounds are the typical choice for radii because they provide ionic radius as a function of ion, oxidation state, and coordination number

for the majority of elements. Most efforts to improve  $t$  have focused on refining the input radii (17, 19, 21, 22) or increasing the dimensionality of the descriptor through two-dimensional (2D) structure maps (18, 23, 24) or high-dimensional machine-learned models (25–27).

However, all hitherto applied approaches for improving the Goldschmidt tolerance factor are only effective over a limited range of  $ABX_3$  compositions. Despite its modest classification accuracy,  $t$  remains the primary descriptor used by experimentalists and theorists to predict the stability of perovskites.

The SISSO (sure independence screening and sparsifying operator) approach (28) was used to identify an improved tolerance factor for predicting whether a given compound is perovskite [determined by experimental realization of any structure with corner-sharing  $BX_6$  octahedra (21) at ambient conditions] or nonperovskite [determined by experimental realization of any structure(s) without corner-sharing  $BX_6$  octahedra, including, in some cases, failed synthesis of any  $ABX_3$  compound]. Of the 576 experimentally characterized  $ABX_3$  solids, 80% were used to train and 20% were used to test the SISSO-learned descriptor. Several alternative atomic properties were considered as candidate features, and among them, SISSO determined that the best performing descriptor,  $\tau$  (Eq. 2 and Fig. 2B), depends only on oxidation states and Shannon ionic radii (see Materials and Methods for an explanation of the approach used for descriptor identification and a discussion of alternative approaches). For the set of 576  $ABX_3$  compositions,  $\tau$  correctly labels 94% of the perovskites and 89% of the nonperovskites compared with 94 and 49%, respectively, using  $t$ . The primary advantage of  $\tau$  over  $t$  is the remarkable reduction in compounds that are predicted to be perovskite but are not experimentally identified as stable perovskites, with false-positive rates for  $\tau$  and  $t$  of 11 and 51%, respectively. Full confusion matrices along with additional performance metrics for  $\tau$  and  $t$  are provided in table S2. The large decrease in false-positive rate (from 51% to 11%) while substantially increasing the overall classification accuracy (from 74% to 92%) demonstrates that  $\tau$  improves significantly upon  $t$  as a reliable tool to guide experimentalists toward which compounds can be synthesized in perovskite structures.

Beyond the improved accuracy, a crucial advantage of  $\tau$  is the monotonic (continuous) dependence of perovskite stability on  $\tau$ . As  $\tau$  decreases, the  $\tau$ -based probability of being perovskite,  $P(\tau)$ , increases, where perovskites are expected for an empirically determined range of  $\tau < 4.18$  (Fig. 2B; Materials and Methods for details). Probabilities are obtained using Platt's scaling (29), where the binary classification of perovskite/nonperovskite is transformed into a continuous probability estimate of perovskite stability,  $P(\tau)$ , by training a logistic regression model on the  $\tau$ -derived binary classification. Probabilities cannot similarly be obtained with  $t$  because the stability of the perovskite structure does not increase or decrease monotonically with  $t$ , where  $0.825 < t < 1.059$  results in a classification as perovskite (this range maximizes the classification accuracy of  $t$  on the set of 576 compounds). While  $P(\tau)$  is sigmoidal with respect to  $\tau$  because of the logistic fit (fig. S2), a bell-shaped behavior of  $P(\tau)$  with respect to  $t$  is observed because of the multiple decision boundaries required for  $t$  (Fig. 2C). This relationship leads to an increase in  $P(\tau)$  (i.e., probability of perovskite stability using  $\tau$ ), with an increase in  $t$  until a value of  $t \sim 0.9$ . Beyond this range, the probabilities level out or decrease as  $t$  increases further.

The disparity between the  $\tau$ -derived perovskite probability,  $P(\tau)$ , and the assignment by  $t$  can be significant, especially in the range where  $t$  predicts a stable perovskite ( $0.825 < t < 1.059$ ). A comparison of the perovskite ( $\text{LaAlO}_3$ ) and the nonperovskite ( $\text{NaBeCl}_3$ ) illustrates the discrepancy between these two approaches.  $t$  incorrectly predicts both compounds to be perovskite ( $t = 1.0$ ), whereas  $P(\tau)$  varies from  $<10\%$  for  $\text{NaBeCl}_3$  to  $>97\%$  for  $\text{LaAlO}_3$ , in agreement with the experimental results. For  $\text{NaBeCl}_3$ , instability in the perovskite structure arises

from an insufficiently large  $\text{Be}^{2+}$  cation on the  $B$  site, which leads to unstable  $\text{BeCl}_6$  octahedra. This contribution to perovskite stability is accounted for in the first term of  $\tau$  (Eq. 2,  $r_X/r_B = \mu^{-1}$ , where  $\mu$  is the octahedral factor).

$\mu$  is the typical choice for a second feature used in combination with  $t$  (18, 19, 23) and was recently used to assess the predictive accuracy of Goldschmidt's "no-rattling" principle. In this analysis, six inequalities dependent on  $t$  and  $\mu$  were derived and used to predict the formability of single and double perovskites with a reported accuracy of  $\sim 80\%$  (30). Notably, training a decision tree algorithm on the bounds of  $t$  and  $\mu$  that optimally separate perovskite from nonperovskite leads to a classification accuracy of 85% for this dataset (fig. S3). In contrast to these 2D descriptors based on  $(t, \mu)$ ,  $\tau$  incorporates  $\mu$  as a 1D descriptor yet still achieves a higher accuracy of 92%, demonstrating the capability of the SISSO algorithm to identify a highly accurate tolerance factor composed of intuitively meaningful parameters.

The nature of geometrical descriptors, such as  $t$  or  $\mu$ , is fundamentally different than that of data-driven descriptors, such as  $\tau$ .  $t$  and  $\mu$  are derived from geometric constraints that indicate when the perovskite structure is a possible structure that can form. However, these constraints do not necessarily indicate when the perovskite structure is the ground-state structure and does form. For instance, if  $t = 1$  and the ionic limit on which  $t$  was derived is applicable (the interatomic distances are sums of the ionic radii), these criteria do not suggest that perovskite is the ground-state structure, only that the interatomic distances are such that the lattice constants in the  $A$ - $X$  and  $B$ - $X$  directions can be commensurate with the perovskite structure. The fact that  $t$  does not guarantee the formation of the perovskite structure is evident by the high false-positive rate (51%) in the region of  $t$  where perovskite is expected ( $0.825 < t < 1.059$ ). Similarly, although  $\mu$  may fall within the range where  $BX_6$  octahedra are expected based on geometric considerations ( $0.414 < \mu < 0.732$ ), the octahedra that form may be edge or face sharing, and therefore, the observed structure is nonperovskite. In this work, SISSO searches a massive space of potential descriptors to identify the one that most successfully detects when a given chemical formula will or will not crystallize in the perovskite structure, and because this is the target property,  $\tau$  emerges as a much more predictive descriptor than  $t$  or  $\mu$ .

Although the classification by  $\tau$  disagrees with the experimental label for 8% of the 576 compounds, the agreement increases to 99% outside the range  $3.31 < \tau < 5.92$  (200 compounds) and 100% outside the range  $3.31 < \tau < 12.08$  (152 compounds). The experimental dataset may also be imperfect as compounds can manifest different crystal structures as a function of the synthesis conditions due to, e.g., defects in the experimental samples (impurities, vacancies, etc.). These considerations emphasize the usefulness of  $\tau$ -derived probabilities, in addition to the binary classification of perovskite/nonperovskite, which address these uncertainties in the experimental data and corresponding classification by  $\tau$ .

### Comparing $\tau$ to calculated perovskite stabilities

The precise and probabilistic nature of  $\tau$ , as well as its simple functional form—depending only on widely available Shannon radii (and the oxidation states required to determine the radii)—enables the rapid search across composition space for stable perovskite materials. Before attempting synthesis, it is common for new materials to be examined using computational approaches; therefore, it is useful to compare the predictions from  $\tau$  with those obtained using density functional theory (DFT). The stabilities (decomposition enthalpies,  $\Delta H_d$ ) of 73 single and

double perovskite chalcogenides and halides were recently examined with DFT using the Perdew-Burke-Ernzerhof (31) exchange-correlation functional (DFT) (32, 33).  $\tau$  is found to agree with the calculated stability for 64 of 73 calculated materials. Importantly, the probabilities that result from classification with  $\tau$  linearly correlate with  $\Delta H_d$ , demonstrating the value of the monotonic behavior of  $\tau$  and  $P(\tau)$  (Fig. 2D and table S3).

Although  $\tau$  appears to disagree with these DFT calculations for nine compounds, six disagreements lie near the decision boundaries [ $P(\tau) = 0.5$ ,  $\Delta H_d = 0$  meV/atom], suggesting that they cannot be confidently classified as stable or unstable perovskites using  $\tau$  or DFT calculations of the cubic structure. Of the remaining disagreements,  $\text{CaZrO}_3$  and  $\text{CaHfO}_3$  reveal the power of  $\tau$  compared with DFT calculations of the cubic structure, as these two oxides are known to be isostructural with the orthorhombic perovskite  $\text{CaTiO}_3$ , from which the name perovskite originates (34, 35).  $\Delta H_d < -90$  meV/atom for these two compounds in the cubic structure, indicating that they are nonperovskites. In contrast,  $\tau$  predicts both compounds to be stable perovskites with ~65% probability, which agrees with the experimental results. These results show that a key challenge in the prediction of perovskite stability from quantum chemical calculations is the requirement of a specific structure as an input, as there are more than a dozen unique structures classified as perovskite (i.e., those having corner-sharing  $\text{BX}_6$  octahedra) and many more that are nonperovskite.

Several recent machine-learned descriptors for perovskite stability have been trained or tested on DFT-calculated stabilities of only the cubic perovskite structure (33, 36–38). However, less than 10% of perovskites are observed experimentally in this structure (21), leading to an inherent disagreement between the descriptor predictions and experimental observations. Recently, it was shown that of 254 synthesized perovskite oxides ( $\text{ABO}_3$ ), DFT calculations in the Open Quantum Materials Database (39) predict only 186 (70%) to be stable or even moderately unstable (within 100 meV/atom of the convex hull) (27). The discrepancy is likely associated with the difference in energy between the true perovskite ground state and the calculated high-symmetry structure(s). Because  $\tau$  was trained exclusively on the experimental characterization of  $\text{ABX}_3$  compounds,  $\tau$  is informed by the true ground-state (or metastable but observed) structure of each  $\text{ABX}_3$  and the potential for these compounds to decompose into any compound(s) in the  $A$ - $B$ - $X$  composition space. A principal advantage of  $\tau$  over many existing descriptors is that its identification and validation were based on experimentally observed stability or instability of a structurally diverse dataset.

### Extension to double perovskite oxides and halides

Double perovskites are particularly intriguing as an emerging class of semiconductors that offer a lead-free alternative to traditional perovskite photoabsorbers and an increased compositional tunability for enhancing desired properties such as catalytic activity (10, 16, 40). Still, the experimentally realized composition space of double perovskites is relatively unexplored compared with the number of possible  $A$ ,  $B$ ,  $B'$ , and  $X$  combinations that can form  $\text{A}_2\text{BB}'\text{X}_6$  compounds. The set of 576 compounds used for training and testing  $\tau$  is composed of 49  $A$  cations, 67  $B$  cations, and 5  $X$  anions, from which >500,000 double perovskite formulas,  $\text{A}_2\text{BB}'\text{X}_6$ , can be constructed. Comparison with the Inorganic Crystal Structure Database (ICSD) (30, 41) reveals only 918 compounds (<0.2%) with known crystal structures, 868 of which are perovskite.

Although  $\tau$  was only trained on  $\text{ABX}_3$  compounds, it is readily adaptable to double perovskites because it depends only on composition and not structure. To extend  $\tau$  to  $\text{A}_2\text{BB}'\text{X}_6$  formulas,  $r_B$  is approximated as the arithmetic mean of the two  $B$ -site radii ( $r_B$ ,  $r_{B'}$ ).  $\tau$  correctly clas-

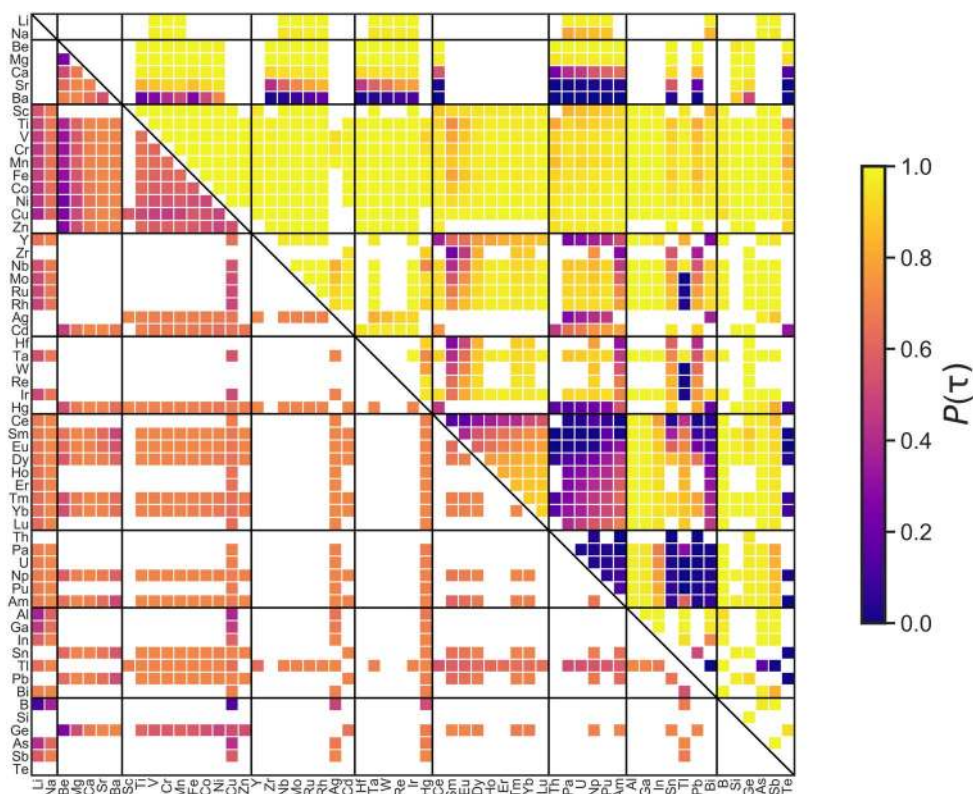
sifies 91% of these 918  $\text{A}_2\text{BB}'\text{X}_6$  compounds in the ICSD (compared with 92% on 576  $\text{ABX}_3$  compounds), recovering 806 of 868 known double perovskites (table S4). The geometric mean has also been used to approximate the radius of a site with two ions (42). We find that this has little effect on classification with  $\tau$ , as 91% of the 918  $\text{A}_2\text{BB}'\text{X}_6$  compounds are also correctly classified using the geometric mean for  $r_B$ , and the classification label differs for only 14 of 918 compounds using the arithmetic or geometric mean. Although  $\tau$  was identified using 460  $\text{ABX}_3$  compounds, the agreement with experiment on these compounds (92%) is comparable to that on the 1034 compounds (91%) that span  $\text{ABX}_3$  (116 compounds) and  $\text{A}_2\text{BB}'\text{X}_6$  (918 compounds) formulas and was completely excluded from the development of  $\tau$  (i.e., test set compounds). This result indicates pronounced generalizability to predicting experimental realization for single and double perovskites that are yet to be discovered. With  $\tau$  thoroughly validated as being predictive of experimental stability, the space of yet-undiscovered double perovskites was explored to identify 23,314 charge-balanced double perovskites that  $\tau$  predicts to be stable at ambient conditions (of >500,000 candidates). These compounds are provided in table S4 including assigned oxidation states and radii along with  $t$  and  $\tau$ , predictions made using each tolerance factor, and classification in the ICSD where available. There are thousands of additional compounds with substitutions on the  $A$  and/or  $X$  sites,  $\text{AA}'\text{BB}'(\text{XX}')_3$ , that are expected to be similarly rich in yet-undiscovered perovskite compounds.

Two particularly attractive classes of materials within this set of  $\text{A}_2\text{BB}'\text{X}_6$  compounds are double perovskites with  $A = \text{Cs}^+$ ,  $X = \text{Cl}^-$  and  $A = \text{La}^{3+}$ ,  $X = \text{O}^{2-}$ , which have garnered substantial interest in a number of applications including photovoltaics, electrocatalysis, and ferroelectricity. The ICSD contains 45 compounds (42 perovskites) with the formula  $\text{CsBB}'\text{Cl}_6$ , 43 of which are correctly classified as perovskite or nonperovskite by  $\tau$ . From the high-throughput analysis using  $\tau$ , we predict an additional 420 perovskites to be stable with 164 having at least the probability of perovskite formation as the recently synthesized perovskite,  $\text{Cs}_2\text{AgBiCl}_6$  [ $P(\tau) = 69.6\%$ ] (43). A map of perovskite probabilities for charge-balanced  $\text{Cs}_2\text{BB}'\text{Cl}_6$  compounds is shown in Fig. 3 (lower triangle). Within this set of 164 probable perovskites, there is an opportunity to synthesize double perovskite chlorides that contain  $3d$  transition metals substituted on one or both  $B$  sites, as 83 new compounds of this type are predicted to be stable as perovskite with high probability.

While double perovskite oxides have been explored extensively for a number of applications, the small radius and favorable charge of  $\text{O}^{2-}$  yields a massive design space for the discovery of new compounds. For  $\text{La}_2\text{BB}'\text{O}_6$ , ~63% of candidate compositions are found to be charge-balanced compared with only ~24% of candidate  $\text{Cs}_2\text{BB}'\text{Cl}_6$  compounds. The ICSD contains 85  $\text{La}_2\text{BB}'\text{O}_6$  compounds, all of which are predicted to be perovskite by  $\tau$  in agreement with the experiment. We predict an additional 1128 perovskites to be discoverable in this space, with a remarkable 990 having  $P(\tau) \geq 85\%$  (Fig. 3, upper triangle). All 128  $\text{ABX}_3$  compounds in the experimental set that meet this threshold are experimentally realized as perovskite, suggesting that there is ample opportunity for perovskite discovery in lanthanum oxides.

### Compositional mapping of perovskite stability

In addition to enabling the rapid exploration of stoichiometric perovskite compositions,  $\tau$  provides the probability of perovskite stability,  $P(\tau)$ , for an arbitrary combination of  $n_A$ ,  $r_A$ ,  $r_B$ , and  $r_X$ , which is shown in Fig. 4. For each grouping shown in Fig. 4, experimentally realized perovskites and nonperovskites are shown as single



**Fig. 3. Map of predicted double perovskite oxides and halides.** Lower triangle: Probability of forming a stable perovskite with the formula  $Cs_2BB'Cl_6$  as predicted by  $\tau$ . Upper triangle: Probability of forming a stable perovskite with the formula  $La_2BB'O_6$  as predicted by  $\tau$ . White spaces indicate  $B/B'$  combinations that do not result in charge-balanced compounds with  $r_A > r_B$ . The colors indicate the Platt-scaled classification probabilities,  $P(\tau)$ , with higher  $P(\tau)$  indicating a higher probability of forming a stable perovskite.  $B/B'$  sites are restricted to ions that are labeled as  $B$  sites in the experimental set of 576  $ABX_3$  compounds.

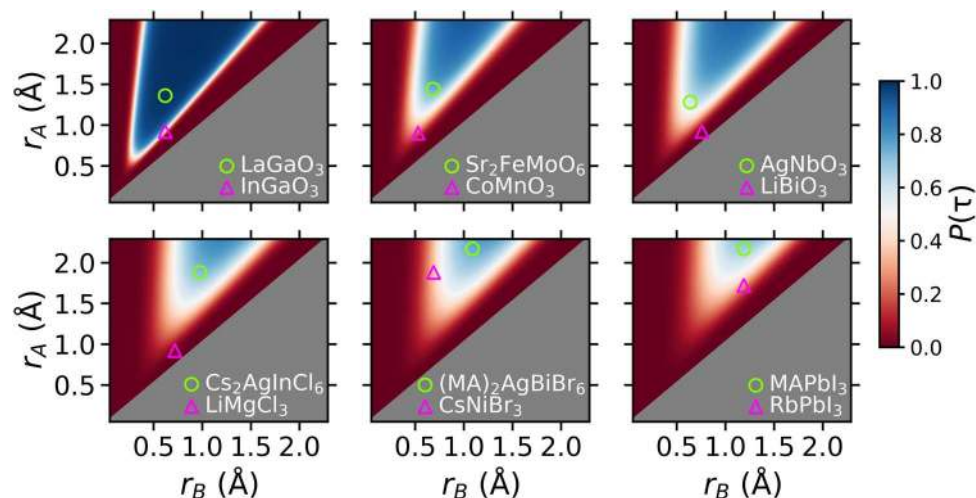
points to compare with the range of values in the predictions made from  $\tau$ . Doping at various concentrations presents a nearly infinite number of  $A_{1-x}A'_x B_{1-y}B'_y (X_{1-z}X'_z)_3$  compositions that allows the tuning of technologically useful properties.  $\tau$  suggests the size and concentration of dopants on the  $A$ ,  $B$ , or  $X$  sites that likely lead to improved stability in the perovskite structure. Conversely, compounds that lie in the high-probability region are likely amenable to ionic substitutions that decrease the probability of forming a perovskite but may improve a desired property for another application. For example,  $LaCoO_3$ , with  $P(\tau) = 98.9\%$ , should accommodate reasonable ionic substitutions (i.e.,  $A$  sites of comparable size to  $La$  or  $B$  sites of comparable size to  $Co$ ) and was recently shown to have enhanced oxygen exchange capacity and nitric oxide oxidation kinetics with stable substitutions of  $Sr$  on the  $A$  site (44).

The probability maps in Fig. 4 arise from the functional form of  $\tau$  (Eq. 2) and provide insights into the stability of the perovskite structure as the size of each ion is varied. The perovskite structure requires that the  $A$  and  $B$  cations occupy distinct sites in the  $ABX_3$  lattice, with  $A$  12-fold and  $B$  6-fold coordinated by  $X$ . When  $r_A$  and  $r_B$  are too similar, nonperovskite lattices that have similarly coordinated  $A$  and  $B$  sites, such as cubic bixbyite, become preferred over the perovskite structure. On the basis of the construct of  $\tau$ , as  $r_A/r_B \rightarrow 1$ ,  $P(\tau) \rightarrow 0$ , which arises from the  $+x/\ln(x)$  ( $x = r_A/r_B$ ) term, where  $\lim_{x \rightarrow 1} \frac{x}{\ln(x)} = +\infty$  and larger values of  $\tau$  lead to lower probabilities of forming perovskites. When  $r_A = r_B$ ,  $\tau$  is undefined, yet compounds where  $A$  and  $B$  have identical radii are rare and not expected to adopt perovskite structures ( $t = 0.71$ ).

The octahedral term in  $\tau (r_X/r_B)$  also manifests itself in the probability maps, particularly in the lower bound on  $r_B$  where perovskites are expected as  $r_X$  is varied. As  $r_X$  increases,  $r_B$  must similarly increase to enable the formation of stable  $BX_6$  octahedra. This effect is noticeable when separately comparing compounds containing  $Cl^-$  (left),  $Br^-$  (center), and  $\Gamma^-$  (right) (bottom row of Fig. 4), where the range of allowed cation radii decreases as the anion radius increases. For  $r_B \ll r_X$ ,  $r_X/r_B$  becomes large, which increases  $\tau$  and therefore decreases the probability of stability in the perovskite structure. This accounts for the inability of small  $B$ -site ions to sufficiently separate  $X$  anions in  $BX_6$  octahedra, where geometric arguments suggest that  $B$  is sufficiently large to form  $BX_6$  octahedra only for  $r_B/r_X > 0.414$ . Because the cation radii ratios strongly affect the probability of perovskite, as discussed in the context of  $x/\ln(x)$ ,  $r_X$  also has a noticeable indirect effect on the lower bound of  $r_A$ , which increases as  $r_X$  increases.

The role of  $n_A$  in  $\tau$  is more difficult to parse, but its placement dictates two effects on stability—as  $A$  is more oxidized (increasing  $n_A$ ),  $-n_A^2$  increases the probability of forming the perovskite structure, but  $n_A$  also magnifies the effect of the  $x/\ln(x)$  term, increasing the importance of the cation radii ratio. Notably,  $n_A = 1$  for most halides and some oxides (245 of the 576 compounds in our set), and in these cases,  $\tau = \frac{r_X}{r_B} + \frac{r_A/r_B}{\ln(r_A/r_B)} - 1$  for all combinations of  $A$ ,  $B$ , and  $X$  and  $n_A$  plays no role as the composition is varied.

This analysis illustrates how data-driven approaches not only can be used to maximize the predictive accuracy of new descriptors but also can be leveraged to understand the actuating mechanisms of a target property—in this case, perovskite stability. This attribute distinguishes



**Fig. 4. The effects of ionic radii and oxidation states on the stability of single and double perovskite oxides and halides.** Top row:  $X = O^{2-}$  (left to right:  $n_A = 3^+, 2^+, 1^+$ ). Bottom row:  $n_A = 1^+$  (left to right:  $X = Cl^-, Br^-, I^-$ ). The experimentally realized perovskites  $LaGaO_3$ ,  $Sr_2FeMoO_6$ ,  $AgNbO_3$ ,  $Cs_2AgInCl_6$ ,  $(MA)_2AgBiBr_6$ , and  $MAPbI_3$  are shown as open circles in the corresponding plot, which are all predicted to be stable by  $\tau$ . The experimentally realized nonperovskites  $InGaO_3$ ,  $CoMnO_3$ ,  $LiBiO_3$ ,  $LiMgCl_3$ ,  $CsNiBr_3$ , and  $RbPbI_3$  are shown as open triangles and predicted to be unstable in the perovskite structure by  $\tau$ . The organic molecule, methylammonium (MA), is shown in the last two panels. While  $(MA)_2AgBiBr_6$  and  $MAPbI_3$  are correctly classified with  $\tau$ , only inorganic cations were used for descriptor identification; therefore,  $r_A = 1.88 \text{ \AA}$  ( $Cs^+$ ) is the largest cation considered. The gray region where  $r_B > r_A$  is not classified because, when this occurs,  $A$  becomes  $B$  and vice versa based on our selection rule  $r_A > r_B$ .

$\tau$  from other descriptors for perovskite stability that have emerged in recent years. For instance, three recent works have shown that the experimental formability of perovskite oxides and halides can be separately predicted with high accuracy using kernel support vector machines (26), gradient boosted decision trees (25), or a random forest of decision trees (27). While these approaches can yield highly accurate models, the resulting descriptors are not documented analytically, and therefore, the mechanism by which they make the perovskite/nonperovskite classification is opaque.

## CONCLUSIONS

We report a new tolerance factor,  $\tau$ , that enables the prediction of experimentally observed perovskite stability significantly better than the widely used Goldschmidt tolerance factor,  $t$ , and the 2D structure map using  $t$  and the octahedral factor,  $\mu$ . For 576  $ABX_3$  and 918  $A_2BB'X_6$  compounds, the prediction by  $\tau$  agrees with the experimentally observed stability for >90% of compounds, with >1000 of these compounds reserved for testing generalizability (prediction accuracy). The deficiency of  $t$  arises from its functional form and not the input features, as the calculation of  $\tau$  requires the same inputs as  $t$  (composition, oxidation states, and Shannon ionic radii). Thus,  $\tau$  enables a superior prediction of perovskite stability with negligible computational cost. The monotonic and 1D nature of  $\tau$  allows the determination of perovskite probability as a continuous function of the radii and oxidation states of  $A$ ,  $B$ , and  $X$ . These probabilities are shown to linearly correlate with DFT-computed decomposition enthalpies and help clarify how chemical substitutions at each of the sites modulate the tendency for perovskite formation. Using  $\tau$ , we predict the probability of double perovskite formation for thousands of unexplored compounds, resulting in a library of stable perovskites ordered by their likelihood of forming perovskites. Because of the simplicity and accuracy of  $\tau$ , we expect its use to accelerate the discovery and design of state-of-the-art perovskite materials for applications ranging from photovoltaics to electrocatalysis.

## MATERIALS AND METHODS

### Radii assignment

To develop a descriptor that takes as input the chemical composition and outputs a prediction of perovskite stability, the features that comprise the descriptor must also be based only on composition. However, it is not known a priori which cation will occupy the  $A$  or  $B$  site given only a chemical composition,  $CC'X_3$  ( $C$  and  $C'$  being cations). Therefore, we developed a systematic method for determining which cation is  $A$  or  $B$  to enable  $\tau$  to be applied to an arbitrary new material. First, a list of allowed oxidation states is defined for each cation based on Shannon's radii (20). All pairs of oxidation states for  $C$  and  $C'$  that charge-balance  $X_3$  are considered. If more than one charge-balanced pair exists, a single pair is chosen on the basis of the electronegativity ratio of the two cations ( $\chi_C/\chi_{C'}$ ). If  $0.9 < \chi_C/\chi_{C'} < 1.1$ , the pair that minimizes  $|n_C - n_{C'}|$  is chosen, where  $n_C$  is the oxidation state for  $C$ . Otherwise, the pair that maximizes  $|n_C - n_{C'}|$  is chosen. With the oxidation states of  $C$  and  $C'$  assigned, the values of the Shannon radii for the cations occupying the  $A$  and  $B$  sites are chosen to be closest to the coordination number of 12 and 6, which are consistent with the coordination environments of the  $A$  and  $B$  cations in the perovskite structure. Last, the radii of the  $C$  and  $C'$  cations were compared, and the larger cation is assigned as the  $A$ -site cation. This strategy reproduced the assignment of the  $A$  and  $B$  cations for 100% of 313 experimentally labeled perovskites.

### Selection of $\tau$

For the identification of  $\tau$  among the offered candidates, the oxidation states ( $n_A, n_B, n_X$ ), ionic radii ( $r_A, r_B, r_X$ ), and radii ratios ( $r_A/r_B, r_A/r_X, r_B/r_X$ ) comprise the primary features,  $\Phi_0$ , where  $\Phi_n$  refers to the descriptor space with  $n$  iterations of complexity as defined in (28). For example,  $\Phi_1$  refers to the primary features ( $\Phi_0$ ), together with one iteration of algebraic/functional operations applied to each feature in  $\Phi_0$ .  $\Phi_2$  then refers to the application of algebraic/functional operations to all potential descriptors in  $\Phi_1$ , and so forth.

Note that  $\Phi_m$  contains all potential descriptors within  $\Phi_{n < m}$ , with a filter to remove redundant potential descriptors. For the discovery of  $\tau$ , complexity up to  $\Phi_3$  is considered, yielding  $\sim 3 \times 10^9$  potential descriptors. An alternative would be to exclude the radii ratios from  $\Phi_0$  and construct potential descriptors with complexity up to  $\Phi_4$ . However, given the minimal  $\Phi_0 = [n_A, n_B, n_X, r_A, r_B, r_X]$ , there are  $\sim 10^8$  potential descriptors in  $\Phi_3$ , so  $\sim 10^{16}$  potential descriptors would be expected in  $\Phi_4$  (based on  $\sim 10^2$  being present in  $\Phi_1$  and  $\sim 1 \times 10^4$  in  $\Phi_2$ ), and this number is impractical to screen using available computing resources.

The dataset of 576  $ABX_3$  compositions was partitioned randomly into an 80% training set for identifying candidate descriptors and a 20% test set for analyzing the predictive ability of each descriptor. The top 100,000 potential descriptors most applicable to the perovskite classification problem were identified using one iteration of SISSO with a subspace size of 100,000. Each descriptor in the set of  $\sim 3 \times 10^9$  was ranked according to domain overlap, as described by Ouyang *et al.* (28). To identify a decision boundary for classification, a decision tree classifier with a maximum depth of two was fit to the top 100,000 candidate descriptors ranked based on domain overlap. Domain overlap (and not decision tree performance) was used as the SISSO ranking metric because of the much lower computational expense associated with applying this metric. Notably,  $\tau$  was the 14,467th highest ranked descriptor by SISSO using the domain overlap metric, and hence, this defines the minimum subspace required to identify  $\tau$  using this approach. Without evaluating a decision tree model for each descriptor in the set of  $\sim 3 \times 10^9$  potential descriptors, we cannot be certain that a subspace size of 100,000 is sufficient to find the best descriptor. However, the identification of  $\tau$  within a subspace as small as 15,000 suggests that a subspace size of 100,000 is sufficiently large to efficiently screen the much larger descriptor space. We have also conducted a test on this primary feature space ( $\Phi_0 = [n_A, n_B, n_X, r_A, r_B, r_X, r_A/r_B, r_A/r_X, r_B/r_X]$ ) with a subspace size of 500,000. Even after increasing the subspace size by 5 $\times$ ,  $\tau$  remains the highest performing descriptor (a classification accuracy of 92% on the 576-compound set). An important distinction between the SISSO approach described here and by Ouyang *et al.* (28) is the choice of sparsifying operator (SO). In this work, domain overlap was used to rank the features in SISSO, but a decision tree with a maximum depth of two was used as the SO (instead of domain overlap) to identify the best descriptor of those selected by SISSO. This alternative SO was used to decrease the leverage of individual data points, as the experimental labeling of perovskite/nonperovskite is prone to some ambiguity based on synthesis conditions, defects, and other experimental considerations.

The benefit of including the radii ratios in  $\Phi_0$  was made clear by comparing the performance of  $\tau$  to the best descriptor obtained using the minimal primary feature space with  $\Phi_0 = [n_A, n_B, n_X, r_A, r_B, r_X]$ . Repeating the procedure used to identify  $\tau$  yields a  $\Phi_3$  with  $\sim 1 \times 10^8$  potential descriptors. The best 1D descriptor was found to be  $\frac{r_B}{n_X(r_A - r_B)} + \frac{r_B}{r_A} - \frac{r_X}{r_B}$ , with a classification accuracy of 89%.

### Alternative features

We also considered the effects of including properties outside of those required to compute  $t$  or  $\tau$ . Beginning with  $\Phi_0 = [n_A, n_B, n_X, r_A, r_B, r_X, r_{cov,A}, r_{cov,B}, r_{cov,X}, IE_A, IE_B, IE_X, \chi_A, \chi_B, \chi_X]$ , where  $r_{cov,i}$  is the empirical covalent radius of neutral element  $i$ ,  $IE_i$  is the empirical first ionization energy of neutral element  $i$ , and  $\chi_i$  is the Pauling electronegativity of element  $i$ , all taken from WebElements (45), an aggregation of a number of references that are available within. Repeating the procedure used to

identify  $\tau$  results in  $\sim 6 \times 10^{10}$  potential descriptors in  $\Phi_3$ . The best performing 1D descriptor was found to be  $\frac{r_A/r_B - \sqrt{\chi_X}}{r_{cov,X}/r_B - r_{cov,A}/r_{cov,X}}$  with a classification accuracy of 90%, lower than  $\tau$  that makes use of only the oxidation states and ionic radii and is only slightly higher than the accuracy of the descriptor obtained using the minimal feature set.

### Increasing dimensionality

To assess the performance of descriptors with increased dimensionality, following the approach to higher dimensional descriptor identification using SISSO described in (28), the residuals from classification by  $\tau$  (those misclassified by the decision tree, Fig. 2B) were used as the target property in the search for a second dimension to include with  $\tau$ . From the same set of  $\sim 3 \times 10^9$  potential descriptors constructed to identify  $\tau$ , the 100,000 1D descriptors that best classify the 41 training set compounds misclassified by  $\tau$  were identified on the basis of domain overlap. Each of these 100,000 descriptors was paired with  $\tau$ , and the performance of each 2D descriptor was assessed using a decision tree with a maximum depth of two. The best performing 2D descriptor was found to be  $\left( \tau, \frac{|r_A r_X / r_B^2 - n_B r_A / r_B|}{|r_A r_B / r_X^2 - r_A / r_B + n_B|} \right)$ , with a classification accuracy of 95% on the 576-compound set. Improvements are expected to diminish as the dimensionality increases further due to the iterative nature of SISSO and the higher-order residuals used for subspace selection. Although the second dimension leads to slightly improved classification performance on the experimental set compared with  $\tau$ , the simplicity and monotonicity of  $\tau$ , which enables physical interpretation and the extraction of meaningful probabilities, support its selection instead of the more complex 2D descriptor. The benefits and capabilities of having a meaningfully probabilistic 1D tolerance factor, such as  $\tau$ , are described in detail within the main text.

### Potential for overfitting

The SISSO algorithm as implemented here selects  $\tau$  from a space of  $\sim 3 \times 10^9$  candidate descriptors, and the only parameter that is fit is the optimum value of  $\tau$  that defines the decision boundary for classification as perovskite or nonperovskite,  $\tau = 4.18$ . This decision boundary was optimized using a decision tree to maximize the classification accuracy on the training set of 460 compounds. In this case, Gini impurity was minimized to optimize the decision boundary, but alternative cost functions based on Kullback-Leibler divergence or classification accuracy (e.g.,  $l_2$ ) would find the same decision boundary. The SISSO descriptor identification is done from billions of candidates, but these functions comprise a discrete set, i.e., they form a basis in a large dimensional space where the number of training points is the dimensionality of the space, which is not densely covered by the functions. Therefore, the selection of only one function,  $\tau$ , cannot overfit the data. However, if some physical mechanism determining the stability of perovskites is not represented in the training set, it might be missed by the learned formula (here,  $\tau$ ), and therefore, the generalizability of the model would be hampered. However, the 94% accuracy achieved by  $\tau$  on the excluded set of 116 compounds shows that  $\tau$  can generalize outside of the training data.

### Alternative radii for more covalent compounds

Ionic radii are required inputs for  $\tau$  (and  $t$ ), and although the Shannon effective ionic radii are ubiquitous in solid-state materials research, a new set of  $B^{2+}$  radii was recently proposed for 18 cations to account for how their effective cationic radii vary as a function of increased

covalency with the heavier halides (19). These revised radii apply to 129 of the 576 experimentally characterized compounds compiled in this dataset (62% of halides). Using these revised radii results in a 5% decrease in the accuracy of  $\tau$  to 86% for these 129 compounds compared to a classification accuracy of 91% using the Shannon radii for these same compounds. The application of  $\tau$  using Shannon radii for presumably covalent compounds was further validated by noting that  $\tau$  correctly classifies 37 of 40 compounds that contain Sn or Pb and achieves an accuracy of 91% for 141 compounds with  $X = \text{Cl}^-$ ,  $\text{Br}^-$ , or  $\text{I}^-$ . In addition to the higher accuracy achieved by  $\tau$  when using Shannon radii, we note that the Shannon radii are more comprehensive than the revised radii in (19), applying to more ions, oxidation states, and coordination environments, and are thus recommended for the calculation of  $\tau$ .

### Computer packages used

SISSO was performed using Fortran 90. Platt's scaling (29) was used to extract classification probabilities for  $\tau$  by fitting a logistic regression model on the decision tree classifications using threefold cross-validation. Decision tree fitting and Platt scaling were performed within the Python package scikit-learn. Data visualizations were generated within the Python packages Matplotlib and Seaborn.

### SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at <http://advances.sciencemag.org/cgi/content/full/5/2/eaav0693/DC1>

Table S1. The 576  $\text{ABX}_3$  used for training and testing  $\tau$ .

Table S2. Confusion matrices for  $\tau$  (above) and  $t$  (below).

Table S3. Additional information associated with Fig. 2D.

Table S4. Double perovskite oxides and halides.

Fig. S1. Comparing the performance of  $t$  and  $\tau$  by composition.

Fig. S2. Sigmoidal relationship between  $P(t)$  and  $\tau$ .

Fig. S3.  $(t, \mu)$  structure map for 576  $\text{ABX}_3$  solids.

### REFERENCES AND NOTES

1. L. Pauling, The principles determining the structure of complex ionic crystals. *J. Am. Chem. Soc.* **51**, 1010–1026 (1929).
2. S. M. Woodley, R. Catlow, Crystal structure prediction from first principles. *Nat. Mater.* **7**, 937–946 (2008).
3. S. Kirkpatrick, C. D. Gelatt Jr., M. P. Vecchi, Optimization by simulated annealing. *Science* **220**, 671–680 (1983).
4. J. P. K. Doye, D. J. Wales, Thermodynamics of global optimization. *Phys. Rev. Lett.* **80**, 1357–1360 (1998).
5. S. Goedecker, Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems. *J. Chem. Phys.* **120**, 9911–9917 (2004).
6. A. R. Oganov, A. O. Lyakhov, M. Valle, How evolutionary crystal structure prediction works—And why. *Acc. Chem. Res.* **44**, 227–237 (2011).
7. S. Curtarolo, G. L. W. Hart, M. Buongiorno Nardelli, N. Mingo, S. Sanvito, O. Levy, The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
8. L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, M. Scheffler, Big data of materials science: Critical role of the descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).
9. V. M. Goldschmidt, Die gesetze der kristallochemie. *Naturwissenschaften* **14**, 477–485 (1926).
10. J. Hwang, R. R. Rao, L. Giordano, Y. Katayama, Y. Yu, Y. Shao-Horn, Perovskites in catalysis and electrocatalysis. *Science* **358**, 751–756 (2017).
11. C. Duan, J. Tong, M. Shang, S. Nikodemski, M. Sanders, S. Ricote, A. Almansoori, R. O'Hayre, Readily processed protonic ceramic fuel cells with high performance at low temperatures. *Science* **349**, 1321–1326 (2015).
12. R. E. Cohen, Origin of ferroelectricity in perovskite oxides. *Nature* **358**, 136–138 (1992).
13. T. Yi, W. Chen, L. Cheng, R. D. Bayliss, F. Lin, M. R. Plews, D. Nordlund, M. M. Doeff, K. A. Persson, J. Cabana, Investigating the intercalation chemistry of alkali ions in fluorine perovskites. *Chem. Mater.* **29**, 1561–1568 (2017).
14. J.-P. Correa-Baena, M. Saliba, T. Buonassisi, M. Grätzel, A. Abate, W. Tress, A. Hagfeldt, Promises and challenges of perovskite solar cells. *Science* **358**, 739–744 (2017).
15. M. V. Kovalenko, L. Protesescu, M. I. Bodnarchuk, Properties and potential optoelectronic applications of lead halide perovskite nanocrystals. *Science* **358**, 745–750 (2017).
16. W. Li, Z. Wang, F. Deschler, S. Gao, R. H. Friend, A. K. Cheetham, Chemically diverse and multifunctional hybrid organic–inorganic perovskites. *Nat. Rev. Mater.* **2**, 16099 (2017).
17. H. Zhang, N. Li, K. Li, D. Xue, Structural stability and formability of  $\text{ABO}_3$ -type perovskite compounds. *Acta Crystallogr. B* **63**, 812–818 (2007).
18. C. Li, X. Lu, W. Ding, L. Feng, Y. Gao, Z. Guo, Formability of  $\text{ABX}_3$  ( $X = \text{F}, \text{Cl}, \text{Br}, \text{I}$ ) halide perovskites. *Acta Crystallogr. B* **64**, 702–707 (2008).
19. W. Travis, E. N. K. Glover, H. Bronstein, D. O. Scanlon, R. G. Palgrave, On the application of the tolerance factor to inorganic and hybrid halide perovskites: A revised system. *Chem. Sci.* **7**, 4548–4556 (2016).
20. R. D. Shannon, Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Crystallogr. A* **32**, 751–767 (1976).
21. M. W. Lufaso, P. M. Woodward, Prediction of the crystal structures of perovskites using the software program *SPuDS*. *Acta Crystallogr. B* **57**, 725–738 (2001).
22. G. Kieslich, S. Sun, A. K. Cheetham, Solid-state principles applied to organic–inorganic perovskites: New tricks for an old dog. *Chem. Sci.* **5**, 4712–4715 (2014).
23. C. Li, K. C. K. Soh, P. Wu, Formability of  $\text{ABO}_3$  perovskites. *J. Alloys Compd.* **372**, 40–48 (2004).
24. M. Becker, T. Klüner, M. Wark, Formation of hybrid  $\text{ABX}_3$  perovskite compounds for solar cell application: First-principles calculations of effective ionic radii and determination of tolerance factors. *Dalton Trans.* **46**, 3500–3509 (2017).
25. G. Pilania, P. V. Balachandran, J. E. Gubernatis, T. Lookman, Classification of  $\text{ABO}_3$  perovskite solids: A machine learning study. *Acta Crystallogr. B* **71**, 507–513 (2015).
26. G. Pilania, P. V. Balachandran, C. Kim, T. Lookman, Finding new perovskite halides via machine learning. *Front. Mater.* **3**, 19 (2016).
27. P. V. Balachandran, A. A. Emery, J. E. Gubernatis, T. Lookman, C. Wolverton, A. Zunger, Predictions of new  $\text{ABO}_3$  perovskite compounds by combining machine learning and density functional theory. *Phys. Rev. Mater.* **2**, 043802 (2018).
28. R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, L. M. Ghiringhelli, SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Phys. Rev. Mater.* **2**, 083802 (2018).
29. J. C. Platt, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, in *Advances in Large Margin Classifiers*, A. J. Smola, P. B. Bartlett, B. Schölkopf, D. Schuurmans, Eds. (MIT Press, 1999), vol. 10, 61–74.
30. M. R. Filip, F. Giustino, The geometric blueprint of perovskites. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 5397–5402 (2018).
31. J. P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
32. X.-G. Zhao, D. Yang, Y. Sun, T. Li, L. Zhang, L. Yu, A. Zunger, Cu–In halide perovskite solar absorbers. *J. Am. Chem. Soc.* **139**, 6718–6725 (2017).
33. Q. Sun, W.-J. Yin, Thermodynamic stability trend of cubic perovskites. *J. Am. Chem. Soc.* **139**, 14905–14908 (2017).
34. H. D. Megaw, Crystal structure of double oxides of the perovskite type. *Proc. Phys. Soc.* **58**, 133 (1946).
35. A. Feteira, D. C. Sinclair, K. Z. Rajab, M. T. Lanagan, Crystal structure and microwave dielectric properties of alkaline-earth hafnates,  $\text{AHfO}_3$  ( $A = \text{Ba}, \text{Sr}, \text{Ca}$ ). *J. Am. Ceram. Soc.* **91**, 893–901 (2008).
36. J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, M. A. L. Marques, Predicting the thermodynamic stability of solids combining density functional theory and machine learning. *Chem. Mater.* **29**, 5090–5103 (2017).
37. F. A. Faber, A. Lindmaa, O. A. von Lilienfeld, R. Armiento, Machine learning energies of 2 million elpasolite ( $\text{ABC}_2\text{D}_6$ ) crystals. *Phys. Rev. Lett.* **117**, 135502 (2016).
38. T. Xie, J. C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
39. S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, C. Wolverton, The open quantum materials database (OQMD): Assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **1**, 15010 (2015).
40. P. V. Kamat, J. Bisquert, J. Buriak, Lead-free perovskite solar cells. *ACS Energy Lett.* **2**, 904–905 (2017).
41. M. Hellenbrandt, The Inorganic Crystal Structure Database (ICSD)—Present and future. *Crystallogr. Rev.* **10**, 17–22 (2004).
42. W. Li, E. Ionescu, R. Riedel, A. Gurlo, Can we predict the formability of perovskite oxynitrides from tolerance and octahedral factors? *J. Mater. Chem. A* **1**, 12239–12245 (2013).
43. E. T. McClure, M. R. Ball, W. Windl, P. M. Woodward,  $\text{Cs}_2\text{AgBiX}_6$  ( $X = \text{Br}, \text{Cl}$ ): New visible light absorbing, lead-free halide perovskite semiconductors. *Chem. Mater.* **28**, 1348–1354 (2016).



44. S. O. Choi, M. Penninger, C. H. Kim, W. F. Schneider, L. T. Thompson, Experimental and computational investigation of effect of Sr on NO oxidation and oxygen exchange for  $\text{La}_{1-x}\text{Sr}_x\text{CoO}_3$  perovskite catalysts. *ACS Catal.* **3**, 2719–2728 (2013).
45. Source: WebElements, [www.webelements.com/](http://www.webelements.com/).

**Acknowledgments:** We thank A. Holder for helpful discussions regarding the manuscript. **Funding:** This project has received funding from the European Union's Horizon 2020 research and innovation program (#676580: The NOMAD Laboratory—A European Center of Excellence and #740233: TEC1p), the Berlin Big-Data Center (BBDC, #01IS14013E), and BiGmax, the Max Planck Society's Research Network on Big-Data-Driven Materials-Science. C.J.B. acknowledges support from a U.S. Department of Education Graduate Assistantship in Areas of National Need. C.S. acknowledges funding by the Alexander von Humboldt Foundation. C.B.M. acknowledges support from NSF award CBET-1433521, which was cosponsored by the NSF and the U.S. Department of Energy (DOE), Office of Energy Efficiency and Renewable Energy (EERE), Fuel Cell Technologies Office and from DOE award EERE DE-EE0008088. Part of this research was performed using computational resources sponsored by the U.S. DOE, Office of EERE and located at the National Renewable Energy Laboratory. **Author contributions:** M.S. and C.J.B. conceived the idea. C.J.B., C.S., and B.R.G. designed the studies. C.J.B. performed the studies. C.J.B., C.S., and B.R.G. analyzed the results and wrote the manuscript. R.O. provided the SISSO algorithm and facilitated its

implementation. C.B.M., L.M.G., and M.S. supervised the project. All the authors discussed the results and implications and edited the manuscript. **Competing interests:** The authors declare that they have no competing financial interests. **Data and materials availability:** A repository containing all files necessary for classifying  $\text{ABX}_3$  and  $\text{AA}'\text{BB}'(\text{XX}')_3$  compositions as perovskite or nonperovskite using  $\tau$  is available at <https://github.com/CJBartel/perovskite-stability>. A graphical interface allowing users to classify compounds with  $\tau$  is also available at <https://analytics-toolkit.nomad-coe.eu>. The classification of all compounds shown in the manuscript is available in the Supplementary Materials. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 10 August 2018  
Accepted 21 December 2018  
Published 8 February 2019  
10.1126/sciadv.aav0693

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

## New tolerance factor to predict the stability of perovskite oxides and halides

Christopher J. Bartel, Christopher Sutton, Bryan R. Goldsmith, Runhai Ouyang, Charles B. Musgrave, Luca M. Ghiringhelli and Matthias Scheffler

*Sci Adv* 5 (2), eaav0693.  
DOI: 10.1126/sciadv.aav0693

### ARTICLE TOOLS

<http://advances.sciencemag.org/content/5/2/eaav0693>

### SUPPLEMENTARY MATERIALS

<http://advances.sciencemag.org/content/suppl/2019/02/04/5.2.eaav0693.DC1>

### REFERENCES

This article cites 43 articles, 6 of which you can access for free  
<http://advances.sciencemag.org/content/5/2/eaav0693#BIBL>

### PERMISSIONS

<http://www.sciencemag.org/help/reprints-and-permissions>

Use of this article is subject to the [Terms of Service](#)

---

*Science Advances* (ISSN 2375-2548) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. 2017 © The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. The title *Science Advances* is a registered trademark of AAAS.