Hierarchical Symbolic Regression for Identifying Key Physical Parameters Correlated with Bulk Properties of Perovskites

Lucas Foppa,[1,2] Thomas A. R. Purcell,[1,2] Sergey V. Levchenko,[3] Matthias Scheffler,[1,2] and Luca M. Ghiringhelli[1,2]
[1]The NOMAD Laboratory at Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, D-14195 Berlin, Germany
[2]The NOMAD Laboratory at Humboldt-Universität zu Berlin, Zum Großen Windkanal 6, D-12489 Berlin, Germany
[3]Skolkovo Institute of Science and Technology, Bolshoy Boulevard 30/1, 121205 Moscow, Russia

(Received 23 February 2022; revised 27 April 2022; accepted 10 June 2022; published 25 July 2022)

Symbolic regression identifies nonlinear, analytical expressions relating materials properties and key physical parameters. However, the pool of expressions grows rapidly with complexity, compromising its efficiency. We tackle this challenge hierarchically: identified expressions are used as inputs for further obtaining more complex expressions. Crucially, this framework can transfer knowledge among properties, as demonstrated using the sure-independence-screening-and-sparsifying-operator approach to identify expressions for lattice constant and cohesive energy, which are then used to model the bulk modulus of ABO_3 perovskites.

DOI: 10.1103/PhysRevLett.129.055301

The identification of physical parameters that are correlated with materials properties or functions is a key step for understanding the underlying processes and accelerating the discovery of improved or even novel materials [1]. Ideally, one would use physical models to describe the materials properties of interest [2]. However, due to the intricate interplay of processes that might be responsible for a certain materials property, the explicit physical modeling might be unfeasible, or even inappropriate. An alternative approach is to use artificial intelligence (AI) to uncover complex relationships. Nevertheless, most widely used AI approaches require datasets that are much larger than those that are typically available in materials science, and only a few AI methods are well suited for small datasets [3–5]. Furthermore, conventional AI produces blackbox models that make it difficult to disentangle the contributions from the various input parameters and determine which underlying processes are the most important to optimize. These problems are exacerbated for the typical scenario in which one is interested in finding materials that exhibit an exceptional performance, for which only a few data points are available.

A possible avenue for linking physical reasoning and data-centric approaches is symbolic regression (SR) [6–8], which identifies nonlinear analytical expressions relating a target property to the key input parameters, even for small datasets. These input parameters are typically physical quantities that are possibly related to the underlying processes governing the property. Traditionally, SR uses genetic-programming techniques to optimize the analytic expressions, which are combinations of the input parameters using mathematical operators such as addition, multiplication, exponentiation, etc., for a given problem [6,7,9–11]. These approaches randomly generate an initial population of possible expressions, and then stochastically apply genetic operators (e.g., mutation and crossover) until some optimal solution is found. More recently, the sure-independence-screening-and-sparsifying-operator (SISSO) [12,13] approach was introduced for the identification of analytical expressions by applying the compressed sensing methodology [14,15] to SR. The SISSO approach starts with the collection of physical input parameters, termed *primary features*. Then, a more expansive pool of expressions is iteratively built by exhaustively applying a set of mathematical operators to both the primary features and previously generated expressions (feature-creation step). The number of recursive applications of the operators used to construct the pool of expressions is called the *rung* (q). Finally, compressed sensing is used to identify the best D-dimensional linear model by performing an ℓ_0 regularization on a subspace S of the all generated expressions, where S is selected using sure-independence screening [16], with the Pearson correlation as the projection score. The outcome of the SISSO analysis is a low D-dimensional descriptor vector containing, as components, the expressions selected from the pool of expressions. A SISSO-derived model for a property P has the form

$$P^{\mathrm{SISSO}} = \sum_{i=0}^{D} c_i d_i, \tag{1}$$

where c_i are fitting coefficients and d_i are the descriptor components. We also label the model components, $\alpha_i = c_i d_i$, which will be used for the construction of more complex models.

SR has already been used to model several materials properties and functions [7,17–23]. However, the combinatorial growth of the pool of possible expressions with respect to the number of primary features and to the number of times that the mathematical operators are applied can make an exhaustive search for the optimal descriptors impractical. This is problematic because in the (initial) absence of understanding of the underlying processes, one would like to offer an extensive set of input parameters to avoid missing the important ones. For addressing this challenge, we introduce a *hierarchical* SR approach that enables an efficient identification of complex descriptors by keeping the number of expressions considered in the analysis at a manageable level. The foundation of this approach is the systematic refeeding of expressions identified in one step as inputs for the identification of more complex expressions in subsequent steps. A crucial implication of this hierarchical framework is that it can be extended to transfer knowledge learned for one property to another one, thus also highlighting physical relationships between materials properties.

We demonstrate the hierarchical SR approach in the context of SISSO. Hierarchical SISSO (HI-SISSO) starts with an initial set of primary features, which is used to obtain an initial model for the property of interest. Then, the obtained model and its components (P^{SISSO} and α_i, respectively), are evaluated for all the materials in the dataset and added to the initial primary feature set. Finally, using this extended primary feature set, new, more complex models are obtained by applying SISSO for a second time. Models and components obtained for one (or more) property (properties) with SISSO can also be used to model a second, related property.

Additionally, in this Letter, we also introduce a new concept into SISSO [24], hereafter called "multiple residuals," which increases the algorithm's efficiency with respect to the size of subspaces needed for ℓ_0 regularization. This procedure updates the SISSO algorithm to use the residuals of the r best models during the sure-independence screening step of SISSO, instead of using the residual of only the best model as done previously [25]. By using the multiple-residual scheme with HI-SISSO, we are able to expedite the search for the best models and considerably reduce (optimize) not only the overall size of the pool of expressions to be considered in the analysis, but also the size of the subspaces of expressions needed for the identification of the best descriptors.

We demonstrate the capabilities of HI-SISSO with two examples, i.e., by modeling the lattice constant (a_0) and bulk modulus (B_0) of ABO_3 cubic perovskites. First, we identify models for the lattice constant (a_0) of each material. Then, we exploit the expressions identified for a_0 and cohesive energies (E_0, defined as the energy per atom required to atomize the crystal) to improve the learning of the bulk moduli (B_0) of the perovskites. We consider 504 ABO_3 materials formed by the A and B elements indicated in Fig. 1. The lattice constants, cohesive energies, and bulk moduli are calculated using density functional theory (DFT) with the PBEsol [40] exchange-correlation functional [25].

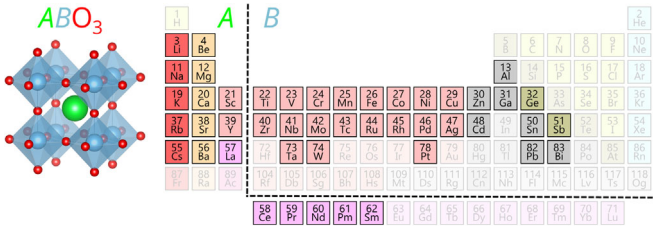

FIG. 1. Materials space of A and B elements corresponding to the 504 cubic ABO_3 perovskites considered in the dataset.

Perovskites display a remarkable diversity of compositions and properties that make them interesting for very different functions and devices (see, e.g., Refs. [41,42]). We focus here on perovskite mechanical properties, specifically the equilibrium lattice constant a_0 and the bulk modulus B_0, the second derivative of the cohesive energy E_0 at a_0. Both quantities are correlated [43–45], which has been described by Verma and Kummar (VK) [46] for cubic perovskites:

$$B_0^{\mathrm{VK}} = C_0 + C_1 \frac{(n_A n_B)^{C_2}}{(a_0)^{3.5}}. \tag{2}$$

Here, C_0, C_1, and C_2 are fitted constants, and n_A and n_B are, respectively, the expected oxidation state of the A and B species in the ABO_3 compound, as approximated by their group number on the periodic table. The approximation implies that all alkali and alkaline earth metals will have an oxidation state of one and two, respectively, and all other A elements will have an oxidation state of three. The oxidation state of the B atom is then set to ensure all materials are charge neutral, i.e., $n_B = 6 - n_A$. The exponent for a_0 in Eq. (2) comes from physical arguments, as described in Ref. [44].

As primary features, we use 23 properties related to the A and B elements of the ABO_3 perovskites, hereafter *atomic features*. These features represent information about the radius, charge, electronic energy levels, and oxidation number for free atoms. The complete list of primary features and operators used in this problem is provided in Supplemental Material [25].

In order to evaluate the performance of our models, we randomly split the dataset of 504 materials into five subsets. Four subsets are combined and used to train the models (training set) and the remaining subset is used to assess the

performance (test set). The training set is used to determine the optimal model complexity with respect to its predictability via a fivefold cross-validation (CV) scheme. Within the fivefold CV scheme, the training set is further split into five subsets. Then, four of these subsets are combined and used to train the model, while the remaining subset is used as validation set. This process is repeated until all five subsets are used as validation sets once and the average of validation root-mean-squared error across the five CV iterations (CV RMSE) is used as the performance metric. The parameters that provide the optimal complexity are considered those associated to the lowest CV RMSE [25]. Within SISSO, the model complexity is controlled by the rung q used to construct the pool of expressions and by the descriptor dimension D. Here we consider descriptors with $D = 1$ up to $D = 5$. Once the model complexity is determined by CV, a model is trained using all the materials of the training set at the optimal complexity. This model is used to predict the properties of the materials in the test set. Finally, the whole procedure is repeated five times, i.e., so that each of the five subsets is considered once as test set. We discuss the performance of the SISSO-derived models based on the distribution of absolute test errors across the 504 materials. We note that more complex models lead to lower *training* errors, but they do not necessarily improve the performance in terms of *test* or *prediction* errors.

The absolute-test-error distributions associated to the lattice constant models obtained with SISSO using rung $q = 1$ and $q = 2$ are shown as gray and red violin plots in Fig. 2(a). These violin plots display the density of data points as a function of their prediction errors. The width of the violin body reflects the number of test data points with that error. The distribution of the absolute test errors is shifted toward lower values when the rung increases from 1 to 2. This shows that the models become more accurate as the mathematical operators are applied for a second time in order to generate more complex expressions. With our primary features and set of operators, rung 1 and 2 pools of features contain on the order of thousands and millions of elements, respectively. To demonstrate how complex descriptors can be found while keeping the number of considered expressions small, we collect the a_0, $q = 1$ model and its components, and use them as new primary features, along with the atomic features, in a second step of SISSO application. In this second step, we also used $q = 1$. We refer to the resulting models as HI-SISSO(a_0) in Fig. 2(a), the parentheses indicating that the expressions describing a_0 identified in the first step are added to the primary feature set, along with the atomic features, in the second step.

The absolute test errors associated to the HI-SISSO(a_0) models [Fig. 2(a), in blue] are lower compared to the errors of the $q = 1$ models obtained with one-step application of SISSO [Fig. 2(a), in gray]. Additionally, the performance of the HI-SISSO(a_0) models is superior compared with the SISSO approach with $q = 2$ [Fig. 2(a), in red]. These results show that HI-SISSO provides a tractable way of increasing the effective rung—and thus the complexity—of a model at a tiny fraction of the computational cost required for a higher rung, since the pool of expressions that needs to be treated is 3 orders of magnitude smaller. Furthermore, by refeeding the models themselves into the primary feature space, we are able to increase the effective dimension of the descriptors without the combinatorial explosion associated with ℓ_0 regularization at higher dimensions.

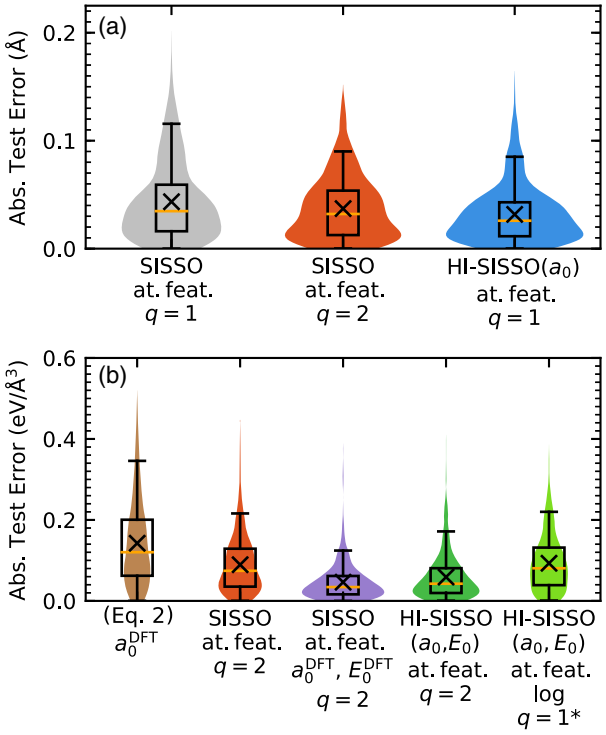

FIG. 2. (a) Distribution of a_0 absolute test-set errors for various sets of hyperparameters (x-axis labels). (b) Distribution of B_0 absolute test-set errors for various sets of hyperparameters and models (x-axis labels). The black "cross mark" represent the mean absolute error, the orange lines are the median absolute error, the boxes are the quartiles, and the whiskers are the minimum and 95% absolute error. In the figure labels, "at. feat." stands for atomic features and the star in $q = 1^*$ indicates the reduced set of operators used in the log-regression approach.

In Fig. 2, we note the presence of outliers for which the absolute test errors are high with respect to the distribution average. These data points are associated to materials with A and/or B elements which are significantly different compared to the A and B elements in the training sets. The detailed analysis of test errors is presented in Supplemental Material along with the discussion of a test set composed by materials containing chemical elements which were unseen during training [25].

We next address the bulk modulus of the perovskites. The distribution of absolute test errors associated to Eq. (2) (with C_0, C_1, and C_2 fitted to the training sets) is shown in

brown in Fig. 2(b) as a baseline for evaluating the performance of the models derived by SR. For the SISSO analysis of bulk modulus, we consider rung $q = 2$. The absolute-test-error distribution corresponding to the SISSO models obtained with the atomic features [Fig. 2(b), in red] shows that this approach has an improved performance compared to Eq. (2). The model's accuracy improves significantly if the DFT-calculated cohesive energy E_0^{DFT}, lattice constant a_0^{DFT}; and $(a_0^{\mathrm{DFT}})^{-3.5}$ are also included as primary features [Fig. 2(b), in violet]. This shows that a_0 and E_0 are both key parameters to describe the bulk modulus. Note that $(a_0^{\mathrm{DFT}})^{-3.5}$ was explicitly included as primary feature because it is suggested as an important parameter by Eq. (2). It would be obtained automatically using $q = 3$, but this would be numerically expensive.

The lattice constant and the cohesive energy provide necessary information to model the bulk modulus. However, the use of a_0^{DFT}, $(a_0^{\mathrm{DFT}})^{-3.5}$, and E_0^{DFT} as primary features is inconvenient. In order to calculate a_0 and E_0 in DFT, one must perform a geometry relaxation, which is already the majority of the work needed to calculate B_0 itself. To circumvent this issue, we offered, as primary features, the SISSO and HI-SISSO models for a_0 and E_0 [25]—as well as their components and the rescaled quantity $(a_0)^{-3.5}$—instead of the DFT-calculated quantities. In this analysis, the atomic features are kept in the primary feature set. We indicate the resulting B_0 models by HI-SISSO(a_0, E_0) in Fig. 2(b) (dark green). By using the HI-SISSO(a_0, E_0) approach, the test errors are significantly reduced compared to the one-step application of SISSO to the atomic features. Indeed, the model performance gets closer to that of the models obtained using the DFT-calculated parameters a_0^{DFT} and E_0^{DFT}, even though the HI-SISSO(a_0, E_0) models depend only on the atomic features, which makes it useful to search for new materials. These results demonstrate the potential of HI-SISSO to transfer information among materials properties, thus circumventing the use of resource-consuming primary features.

We then exploited the B_0 model obtained by the HI-SISSO(a_0, E_0) approach, trained using the entire dataset of 504 ABO_3 materials, for the screening of new materials [25]. We evaluated 7308 single (ABO_3) and double perovskite compositions of the type $A_2BB'O_6$ constructed from all the A and B elements in the initial dataset (Fig. 1). Then, we looked at the materials with the lowest predicted B_0 values, since they are scarce in the training set. This situation corresponds to the typical scenario in materials discovery, in which the behavior of interest is associated to only few of the available observations. Among the 10 materials with the lowest B_0 predicted by the HI-SISSO approach, we identify the double perovskites Cs_2ZnBiO_6, Cs_2CdBiO_6, Cs_2CdPbO_6, Cs_2ZnPbO_6, Cs_2ZnCdO_6, Rb_2ZnBiO_6, and Rb_2CdBiO_6, with predicted B_0 in the range 0.49–0.53 eV/Å^3. The properties of these materials were evaluated explicitly by further DFT calculations and they were confirmed as highly compressible perovskites, with DFT-calculated B_0 of 0.45, 0.45, 0.46, 0.45, 0.46, 0.60, and 0.41 eV/Å^3, respectively. The root-mean-squared error calculated on the 10 materials with the lowest predicted B_0 is 0.081 eV/Å^3. By recalling that the model was trained on simpler single perovskites, its predictive ability beyond the training region is remarkable. Moreover, only 8 materials, out of the 504 used for training, present $B_0 < 0.50$ eV/Å^3. We note that our (HI-)SISSO approach is learning results of the DFT PBEsol theory. Thus, when estimating the experimental properties of the perovskites, in addition to the error of the (HI-)SISSO models to predict the DFT-calculated perovskite properties, the errors resulting from the DFT PBEsol approach should be taken into account [25].

Finally, we identified with HI-SISSO a power-law-type expression for B_0, in the spirit of Eq. (2). For this purpose, we applied a logarithm transformation to the property vector and candidate expressions, and then ran SISSO in this transformed space. We then backtransformed the resulting expression in the form of Eq. (1) using exponentiation to get the power-law model shown in Eq. (3). We offered the atomic features and the SISSO $q = 2$ models for a_0 and E_0, denoted $a_0^{\mathrm{SISSO}(q=2)}$ and $E_0^{\mathrm{SISSO}(q=2)}$, respectively, as primary features. The components of these models are also included in the primary feature set. Here, we used $q = 1$ with a reduced mathematical operator set containing only the operators addition and subtraction. The rescaled lattice $(a_0^{\mathrm{SISSO}(q=2)})^{-3.5}$ was not included as primary feature because by this approach such term will be automatically considered. The best model identified using the entire dataset of 504 materials at the optimal dimensionality identified via CV ($D = 3$ [25]) is

$$B_0^{\mathrm{HI\text{-}SISSO}} = 2.99 \frac{(IP_B - EA_B)^{0.419} (E_0^{\mathrm{SISSO}(q=2)})^{0.964}}{(a_0^{\mathrm{SISSO}(q=2)} - 5.09 \times 10^{-4} \frac{EA_B n_A}{|r_{s,B}^{\mathrm{cat}} - r_{s,B}|})^{2.75}}, \quad (3)$$

where IP_B is the ionization potential of the B atom, EA_B is the electron affinity of the B atom, n_A is the oxidation number of the A atom, $r_{s,B}$ is the radius of the valence-s orbital of the neutral atom, and $r_{s,B}^{\mathrm{cat}}$ is the radius of the valence-s orbital of the 1+ cation. The equations for $a_0^{\mathrm{SISSO}(q=2)}$ and $E_0^{\mathrm{SISSO}(q=2)}$ as well as for other SISSO-derived models are shown in Supplemental Material [25]. SISSO selects IP_B, EA_B, $r_{s,B}^{\mathrm{cat}}$, $r_{s,B}$, a_0^{SISSO}, and E_0^{SISSO} as the key parameters correlated with B_0. Therefore, SISSO recovers the parameters a_0 and n_A, which also enter Eq. (2). However, the description of B_0 provided by Eq. (3) goes beyond the empirical model, since the log-regression models provide a significantly better

performance [light green in Fig. 2(b)] compared to Eq. (2)—even though they do not outperform the models obtained with the linear-regression approach, dark green in Fig. 2(b). Equation (3) highlights that B_0 is directly proportional to E_0 and to $(IP_B - EA_B)$. This might reflect the ionic interaction contributions to B_0. Indeed, the latter term indicates the relevance of the ionization of B species, which is present as a cation in the perovskite. In the denominator of Eq. (3), a_0 and $(EA_B n_A/|r_{s,B}^{\mathrm{cat}} - r_{s,B}|)$ appear. The latter term possibly captures the charge-dependent effective size of the B species in the perovskite, analogous to the Shannon effective radii [47], and it might be related to covalent contributions to B_0, since these interactions depend on the overlap (and thus distance) between the interacting orbitals of B cations and O^{-2} species. Despite this analysis, we like to stress that by assigning a specific, physical meaning to each term of the equations derived by (HI-)SISSO, one might overlook the possibly intricate interplay of processes governing the properties. Furthermore, the physical relationship between the identified parameters and the underlying physics might be indirect, as the correlations do not necessarily reflect direct causality.

In this Letter, we introduced a hierarchical SR framework to efficiently address complex materials properties and functions. This approach provides the key physical parameters reflecting the underlying processes responsible for the behavior of interest, while increasing the performance of SR models. The analysis described in this Letter can be reproduced and modified at the Novel-Materials Discovery (NOMAD) AI Toolkit [48].

The dataset of calculated perovskite properties as well as the input and output files of the DFT calculations are available at the NOMAD Repository and Archive [49].

This work was funded by the NOMAD Center of Excellence (European Union's Horizon 2020 research and innovation program, Grant Agreement No. 951786), the ERC Advanced Grant TEC1p (European Research Council, Grant Agreement No. 740233), and the project FAIRmat (FAIR Data Infrastructure for Condensed-Matter Physics and the Chemical Physics of Solids, German Research Foundation, Project No. 460197019). L. F. acknowledges the funding from the Swiss National Science Foundation, postdoc mobility Grant No. P2EZP2_181617. T. A. R. P. would like to thank the Alexander von Humboldt Foundation for their support through the Alexander von Humboldt Postdoctoral Fellowship Program.

[1] Luca M. Ghiringhelli, Jan Vybiral, Sergey V. Levchenko, Claudia Draxl, and Matthias Scheffler, Big Data of Materials Science: Critical Role of the Descriptor, Phys. Rev. Lett. **114**, 105503 (2015).

[2] Karsten Reuter, Catherine Stampf, and Matthias Scheffler, *Ab Initio* atomistic thermodynamics and statistical mechanics of surface properties and functions, in *Handbook of Materials Modeling: Methods*, edited by S. Yip (Springer, Dordrecht, Netherlands, 2005), pp. 149–194.

[3] Shuo Feng, Huiyu Zhou, and Hongbiao Dong, Using deep neural network with small dataset to predict material defects, Mater. Des. **162**, 300 (2019).

[4] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky, E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, arXiv:2101.03164.

[5] Pierre-Paul De Breuck, Geoffroy Hautier, and Gian-Marco Rigagnese, Materials property prediction for limited datasets enabled by feature selection and joint learning with modnet, npj Comput. Mater. **7**, 83 (2021).

[6] John R. Koza, Genetic programming as a means for programming computers by natural selection, Stat. Comput. **4**, 87 (1994).

[7] Yiqun Wang, Nicholas Wagner, and James M. Rondinelli, Symbolic regression in materials science, MRS Commun. **9**, 793 (2019).

[8] Michael Schmidt and Hod Lipson, Distilling free-form natural laws from experimental data, Science **324**, 81 (2009).

[9] Tim Mueller, Eric Johlin, and Jeffrey C. Grossman, Origins of hole traps in hydrogenated nanocrystalline and amorphous silicon revealed through machine learning, Phys. Rev. B **89**, 115202 (2014).

[10] Fenglin Yuan and Tim Mueller, Identifying models of dielectric breakdown strength from high-throughput data via genetic programming, Sci. Rep. **7**, 17594 (2017).

[11] Silviu-Marian Udrescu and Max Tegmark, AI Feynman: A physics-inspired method for symbolic regression, arXiv:1905.11481.

[12] Runhai Ouyang, Stefano Curtarolo, Emre Ahmetcik, Matthias Scheffler, and Luca 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).

[13] Runhai Ouyang, Emre Ahmetcik, Christian Carbogno, Matthias Scheffler, and Luca M. Ghiringhelli, Simultaneous learning of several materials properties from incomplete databases with multi-task SISSO, J. Phys. Mater. **2**, 024002 (2019).

[14] Lance J. Nelson, Gus L. W. Hart, Fei Zhou, and Vidvuds Ozoliņš, Compressive sensing as a paradigm for building physics models, Phys. Rev. B **87**, 035125 (2013).

[15] E. J. Candes and M. B. Wakin, An introduction to compressive sampling, IEEE Signal Process. Mag. **25**, 21 (2008).

[16] Jianqing Fan and Jinchi Lv, Sure independence screening for ultrahigh dimensional feature space, J. R. Stat. Soc. Ser. B **70**, 849 (2008).

[17] Christopher J. Bartel, Samantha L. Millican, Ann M. Deml, John R. Rumptz, William Tumas, Alan W. Weimer, Stephan Lany, Vladan Stevanović, Charles B. Musgrave, and Aaron M. Holder, Physical descriptor for the Gibbs energy of inorganic crystalline solids and temperature-dependent materials chemistry, Nat. Commun. **9**, 4168 (2018).

[18] S. R. Xie, G. R. Stewart, J. J. Hamlin, P. J. Hirschfeld, and R. G. Hennig, Functional form of the superconducting critical temperature from machine learning, Phys. Rev. B **100**, 174513 (2019).

[19] Runhai Ouyang, Exploiting ionic radii for rational design of halide perovskites, Chem. Mater. **32**, 595 (2020).

[20] Guohua Cao, Runhai Ouyang, Luca M. Ghiringhelli, Matthias Scheffler, Huijun Liu, Christian Carbogno, and Zhenyu Zhang, Artificial intelligence for high-throughput discovery of topological insulators: The example of alloyed tetradymites, Phys. Rev. Mater. **4**, 034204 (2020).

[21] Lucas Foppa, Luca M. Ghiringhelli, Frank Girgsdies, Maike Hashagen, Pierre Kube, Michael Hävecker, Spencer J. Carey, Andrey Tarasov, Peter Kraus, Frank Rosowski, Robert Schlögl, Annette Trunschke, and Matthias Scheffler, Materials genes of heterogeneous catalysis from clean experiments and artificial intelligence, MRS Bull. **46**, 1016 (2021).

[22] Zhong-Kang Han, Debalaya Sarker, Runhai Ouyang, Aliaksei Mazheika, Yi Gao, and Sergey V Levchenko, Single-atom alloy catalysts designed by first-principles calculations and artificial intelligence, Nat. Commun. **12**, 1833 (2021).

[23] Ning He, Runhai Ouyang, and Quan Qian, Learning interpretable descriptors for the fatigue strength of steels, AIP Adv. **11**, 035018 (2021).

[24] Thomas A. R. Purcell, Matthias Scheffler, Christian Carbogno, and Luca M. Ghiringhelli, SISSO++: A C++ implementation of the sure independence screening and sparsifying operator approach, J. Open Source Softwaare **7**, 3960 (2022).

[25] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.129.055301 for details on the DFT calculations, the multiple-residuals approach, the primary features and operators used, the cross-validation strategy, the screening of double perovskites, as well as the expressions of the SISSO and HI-SISSO models, the analysis of the selected-element test set, the analysis of test errors, and the comparison of HI-SISSO with the kernel-ridge regression approach. The supplemental material includes Refs. [26–39].

[26] Volker Blum, Ralf Gehrke, Felix Hanke, Paula Havu, Ville Havu, Xinguo Ren, Karsten Reuter, and Matthias Scheffler, *Ab initio* molecular simulations with numeric atom-centered orbitals, Comput. Phys. Commun. **180**, 2175 (2009).

[27] Maja-Olivia Lenz, Thomas A. R. Purcell, David Hicks, Stefano Curtarolo, Matthias Scheffler, and Christian Carbogno, Parametrically constrained geometry relaxations for high-throughput materials science, npj Comput. Mater. **5**, 123 (2019).

[28] Simuck F. Yuk, Krishna Chaitanya Pitike, Serge M. Nakhmanson, Markus Eisenbach, Ying Wai Li, and Valentino R. Cooper, Towards an accurate description of perovskite ferroelectrics: Exchange and correlation effects, Sci. Rep. **7**, 43482 (2017).

[29] Guo-Xu Zhang, Anthony M. Reilly, Alexandre Tkatchenko, and Matthias Scheffler, Performance of various density-functional approximations for cohesive properties of 64 bulk solids, New J. Phys. **20**, 063020 (2018).

[30] A. S. Verma and V. K. Jindal, Lattice constant of cubic perovskites, J. Alloys Compd. **485**, 514 (2009).

[31] H. Landolt, R. Börnstein, D. Bimberg, M. Schulz, H. Weiss, and O. Madelung, in *Magnetic and Other Properties of Oxides and Related Compounds*, edited by K. H. Hellwege, Landolt-Börnstein, New Series, Group III: Crystal and Solid State Physics Vol. 12 (Springer-Verlag, Berlin, 1982).

[32] A. J. Smith and A. J. E. Welch, Some mixed metal oxides of perovskite structure, Acta Crystallogr. **13**, 653 (1960).

[33] Jiangang He and Cesare Franchini, Screened hybrid functional applied to $3d^0 \rightarrow 3d^8$ transition-metal perovskites $LaMO_3$ (M = Sc–Cu): Influence of the exchange mixing parameter on the structural, electronic, and magnetic properties, Phys. Rev. B **86**, 235117 (2012).

[34] L. Q. Jiang, J. K. Guo, H. B. Liu, M. Zhu, X. Zhou, P. Wu, and C. H. Li, Prediction of lattice constant in cubic perovskites, J. Phys. Chem. Solids **67**, 1531 (2006).

[35] Aakash Naik and Thomas A. R. Purcell, https://gitlab.mpcdf.mpg.de/nomad-lab/atomic-features-package.

[36] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder, Python Materials Genomics (pymatgen): A robust, open-source Python library for materials analysis, Comput. Mater. Sci. **68**, 314 (2013).

[37] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, npj Comput. Mater. **2**, 16028 (2016).

[38] Christopher J. Bartel, Christopher Sutton, Bryan R. Goldsmith, Runhai Ouyang, Charles B. Musgrave, Luca M. Ghiringhelli, and Matthias Scheffler, New tolerance factor to predict the stability of perovskite oxides and halides, Sci. Adv. **5**, eaav0693 (2019).

[39] Sami Vasala and Maarit Karppinen, $A_2B'B''O_6$ perovskites: A review, Prog. Solid State Chem. **43**, 1 (2015).

[40] Gábor I. Csonka, John P. Perdew, Adrienn Ruzsinszky, Pier H. T. Philipsen, Sébastien Lebègue, Joachim Paier, Oleg A. Vydrov, and János G. Ángyán, Assessing the performance of recent density functionals for bulk solids, Phys. Rev. B **79**, 155107 (2009).

[41] Ajay Kumar Jena, Ashish Kulkarni, and Tsutomu Miyasaka, Halide perovskite photovoltaics: Background, status, and future prospects, Chem. Rev. **119**, 3036 (2019).

[42] Jonathan Hwang, Reshma R. Rao, Livia Giordano, Yu Katayama, Yang Yu, and Yang Shao-Horn, Perovskites in catalysis and electrocatalysis, Science **358**, 751 (2017).

[43] Marvin L. Cohen, Theory of bulk moduli of hard solids, Mater. Sci. Eng. A **105–106**, 11 (1988).

[44] Marvin L. Cohen, Calculation of bulk moduli of diamond and zinc-blende solids, Phys. Rev. B **32**, 7988 (1985).

[45] George J. Fischer, Zichao Wang, and Shun-ichiro Karato, Elasticity of $CaTiO_3$, $SrTiO_3$ and $BaTiO_3$ perovskites up to 3.0 Gpa: The effect of crystallographic structure, Phys. Chem. Miner. **20**, 97 (1993).

[46] A. S. Verma and A. Kumar, Bulk modulus of cubic perovskites, J. Alloys Compd. **541**, 210 (2012).

[47] R. D. Shannon and C. T. Prewitt, Effective ionic radii in oxides and fluorides, Acta Crystallogr. Sect. B **25**, 925 (1969).

[48] https://nomad-lab.eu/aitoolkit/hierarchical_sisso.

[49] 10.17172/NOMAD/2022.02.21-3.