SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates

Runhai Ouyang,[1] Stefano Curtarolo,[1,2] Emre Ahmetcik,[1] Matthias Scheffler,[1] and Luca M. Ghiringhelli[1,*]

[1]*Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin-Dahlem, Germany*

[2]*Center for Materials Genomics and Department of Mechanical Engineering and Materials Science, Duke University, Durham, North Carolina 27708, USA*

(Received 20 May 2018; published 7 August 2018)

The lack of reliable methods for identifying descriptors—the sets of parameters capturing the underlying mechanisms of a material's property—is one of the key factors hindering efficient materials development. Here, we propose a systematic approach for discovering descriptors for materials' properties, within the framework of compressed-sensing-based dimensionality reduction. The sure independence screening and sparsifying operator (SISSO) tackles immense and correlated features spaces, and converges to the optimal solution from a combination of features relevant to the materials' property of interest. In addition, SISSO gives stable results also with small training sets. The methodology is benchmarked with the quantitative prediction of the ground-state enthalpies of octet binary materials (using *ab initio* data) and applied to the showcase example of predicting the metal/insulator classification of binaries (with experimental data). Accurate, predictive models are found in both cases. For the metal-insulator classification model, the predictive capability is tested beyond the training data: It rediscovers the available pressure-induced insulator-to-metal transitions and it allows for the prediction of yet unknown transition candidates, ripe for experimental validation. As a step forward with respect to previous model-identification methods, SISSO can become an effective tool for automatic materials development.

DOI: 10.1103/PhysRevMaterials.2.083802

I. INTRODUCTION

The materials-genome initiative [1] has fostered high-throughput calculations and experiments. Correspondingly, computational initiatives (e.g., Refs. [2–5]), have already tackled many thousands of different systems (see Refs. [6–16]). Much of the data of this field is available in the FAIR Repository and Archive of the NOMAD Centre of Excellence [17,18]. On close inspection, one realizes that such data collections are so far inefficiently exploited, and only a tiny amount of the contained information is actually used. Despite the number of possible materials being infinite, the request for specific properties, e.g., a material that is stable, nontoxic, with an optical band gap between 0.8 and 3.2 eV, drastically reduces the set of candidates. This implies that, in terms of functional materials, the structural and chemical space of compounds is sparsely populated. Identifying these few materials—known materials as well as materials that have not been created to date—requires an accurate, predictive approach.

Several methods, falling under the umbrella names of artificial intelligence or (big-)data analytics (including data mining, machine/statistical learning, compressed sensing, etc.) have being developed and applied to the wealth of materials-science data [19–28], but so far, no general and systematic approach has been established and demonstrated. The challenge here is that many different processes and phenomena exist, controlled by atomic structure, electron charge, spin, phonons, polarons and other quasiparticles, and tiny changes in structure or composition can cause a qualitative change of the materials property (phase transitions). For example, less than 0.001% impurities can turn an insulator into a conductor. This type of complexity is a significant element of the fourth paradigm in materials science [18,29,30], which recognizes that it may not be possible to describe many properties of functional materials by a single, physically founded model, i.e., via a closed, analytical expression. The reason is that such properties are determined by several multilevel, intricate theoretical concepts. Thus, insight is obtained by searching for structure and patterns in the data, which arise from functional relationships (including but not limited to linear correlations) with different processes and functions. Finding a descriptor, the set of parameters capturing the underlying mechanism of a given materials property or function, that reveals these relationships is the key, intelligent step. Once the descriptor has been identified, essentially every learning approach (e.g., regressions, including kernel-based ones, artificial neural networks, etc.) can be applied straightforwardly. These issues and in particular the central role of the descriptor were implicitly assumed in many seminal machine-learning works applied to materials science, but it was only later explicitly identified in the works of Ghiringhelli *et al.* [7,31]. These authors recast the descriptor-search challenge into a compressed-sensing (CS) formulation. The CS approach has been shown to be effective for reproducing a high-quality reconstructed signal starting from a very small set of observations [32,33]. Mathematically, given a set of samples measured incoherently, $\boldsymbol{P}$, CS finds the sparse solution $\boldsymbol{c}$ of an underdetermined system of linear equations $\boldsymbol{Dc} = \boldsymbol{P}$ ($\boldsymbol{D}$ is called the sensing matrix with columns $\gg$ rows). If the number of nonzero entries in $\boldsymbol{c}$

*ghiringhelli@fhi-berlin.mpg.de

is smaller than the size of $\boldsymbol{P}$, then CS effectively reduces the dimensionality of the problem [32,34,35]. In the specific case treated in Refs. [7,31], given a set of materials m_i with observable properties listed in vector $\boldsymbol{P}_i$ and a huge list of possible test features d_j (forming the features space), the linear projection of each i material into the j feature forms the i, j components of the sensing matrix $\boldsymbol{D}$. The sparse solution of $\arg\min_{\boldsymbol{c}} (\|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|_2^2 + \lambda\|\boldsymbol{c}\|_0)$, where $\|\boldsymbol{c}\|_0$ is the number of nonzero components of $\boldsymbol{c}$, gives the optimum n-dimensional descriptor, i.e., the set of features selected by the n nonzero components of the solution vector $\boldsymbol{c}$.

In Refs. [7,31], a modification of the least absolute shrinkage and selection operator (LASSO) [36] was introduced for finding the optimal solution. However, moving beyond the showcase application demonstrated in those papers (predicting the ground-state crystal structure of octet binaries semiconductors), it turns out that the method is unable to deal with large feature spaces, i.e., with situations where knowledge about the underlying processes is not well developed and when in addition to the atomic properties, also collective properties, e.g., the electronic band structure, play a role. When the space of candidate features (the feature space) gets large (larger than few thousand elements) and/or when features are correlated, the approach breaks down.

In the present paper, we provide a strong and efficient solution of these problems, i.e., we present a new method, called sure independent screening and sparsifying operator (SISSO), which can deal with an immensity of candidate features (billions or more) and does not suffer when features are correlated. The outcome of SISSO is a mathematical model, in the form of explicit, analytic functions of basic, input physical quantities. This aspect gives the opportunity to inspect the equations and suggest means to test the generalization ability of the model.

II. RESULTS AND DISCUSSION

Features space construction. All quantities that are hypothesized to be relevant for describing the target property (the so-called primary features [7,31]) are used as a starting point for the construction of the space [37,38]. Features are of atomic (species *per se*) and collective origin (atoms embedded in the environment). Then, a combination of algebraic/functional operations is recursively performed for extending the space. For instance, the starting point $\boldsymbol{\Phi}_0$ may comprise readily available and relevant properties, such as atomic radii, ionization energies, valences, bond distances, and so on. The operators set is defined as

$$\hat{\boldsymbol{H}}^{(\mathrm{m})} \equiv \left\{I, +, -, \times, /, \exp, \log, |-|, \sqrt{}, ^{-1}, ^{2}, ^{3}\right\}[\phi_1, \phi_2],$$

where ϕ_1 and ϕ_2 are objects in Φ (for unary operators only ϕ_1 is considered) and the superscript $^{(\mathrm{m})}$ indicates that dimensional analysis is performed to retain only meaningful combinations (e.g., no unphysical items such as $size + energy$ or $size + size^2$). The intrinsically linear relationship observables $\leftrightarrow$ descriptor in the CS formalism is made nonlinear by equipping the features space with nonlinear operators in $\hat{\boldsymbol{H}}^{(\mathrm{m})}$. At each iteration, $\hat{\boldsymbol{H}}^{(\mathrm{m})}$ operates on all available combinations, and the features space grows recursively as:

$$\boldsymbol{\Phi}_n \equiv \bigcup_{i=1}^{n} \hat{\boldsymbol{H}}^{(\mathrm{m})}[\phi_1, \phi_2], \quad \forall \phi_1, \phi_2 \in \Phi_{i-1}. \tag{1}$$

The number of elements in $\boldsymbol{\Phi}_n$ grows very rapidly with n. It is roughly of the order of $\sim (\#\boldsymbol{\Phi}_0)^{2^n} \times (\#\hat{\boldsymbol{H}}_{\boldsymbol{2}})^{2^n-1}$ where $\#\boldsymbol{\Phi}_0$ and $\#\hat{\boldsymbol{H}}_{\boldsymbol{2}}$ are the numbers of elements and binary operators in $\boldsymbol{\Phi}_0$ and $\hat{\boldsymbol{H}}$, respectively. For example, $\#\boldsymbol{\Phi}_3 \sim 10^{11}$ with $\#\hat{\boldsymbol{H}}_{\boldsymbol{2}} = 5$ and $\#\boldsymbol{\Phi}_0 = 10$. To avoid *a priori* bias and contrary to previous works [37], no features were disregarded despite the size of the resulting features space. Instead, we extend the sparse-solution algorithm (using sparsifying operators (SO) [39]) and tackle huge sensing matrices representative of features spaces containing coherent elements overcoming the limitations of LASSO-based methods [7,31].

Solution algorithm. The ℓ_0-norm regularized minimization [42] is the obvious path for finding the best sparse solution of linear equations. It is performed through combinatorial optimization by penalizing the number of nonzero coefficients. The algorithm is NP hard and thus infeasible when the features space becomes very large. Efficient methods can be employed to approximate the correct ℓ_0 solution [43] with ideal features space (e.g., having uncorrelated basis sets). Among them are the convex optimization by ℓ_1-norm [44] regularization LASSO [36]) and the various greedy algorithms such as the matching pursuit (MP) [45] and orthogonal matching pursuit (OMP) [46,47]. Unfortunately, with correlated features spaces, approximated results can largely deviate from the ideal ℓ_0 solutions [43,48]. Corrections have been proposed, for example the LASSO+ℓ_0 scheme comprising LASSO prescreening and subsequent ℓ_0 optimization [7,31], and the the ℓ_1 analysis and ℓ_1 synthesis [49]. However, when the features space size becomes of the order of 10^6–10^9, ℓ_1-based methods also become computationally infeasible. As previously mentioned, here we overcome the huge size of the problem by combining SO with sure independence screening (SIS) [50,51], which has been shown to be effective for dimensionality reduction of ultra-high-dimensional features spaces [50]. SIS scores each feature (standardized) with a metric (correlation magnitude, i.e., the absolute of inner product between the target property and a feature) and keeps only the top ranked [50]. After the reduction, SO is used to pinpoint the optimal n-dimensional descriptor. The smaller the dimensionality, the better the outcome: progressively larger n are tested until the leftover residual error is within quality expectation. The combination of SIS and SO is called SISSO. Figure 1 illustrates the idea.

SISSO. Out of the huge features space ($\sim 10^{10}$ elements or more), SIS selects the subspace $\boldsymbol{S}_{1\mathrm{D}}$ containing the features having the largest correlation with the response $\boldsymbol{P}$ (target material property). Generally, the larger the subspace $\cup\boldsymbol{S}_{i\mathrm{D}}$, the higher the probability it contains the optimal descriptor. However, the chosen size of $\cup\boldsymbol{S}_{i\mathrm{D}}$, depends on (i) which type of SO is later used, (ii) the dimensionality n requested, and (iii) the available computational resources. With SO (LASSO), $\cup\boldsymbol{S}_{i\mathrm{D}}$ can contain as much as $10^5 \sim 10^6$ elements, depending on $\#\boldsymbol{P}$. With SO(ℓ_0), the largest obtainable size is typically 10^5 for $n = 2$, 10^3 for $n = 3$, 10^2 for $n = 4$, etc. (because the number of needed evaluation grows combinatorially with n). If n is large, e.g., > 10, then the maximum possible $\#\boldsymbol{S}_{i\mathrm{D}}$ converge

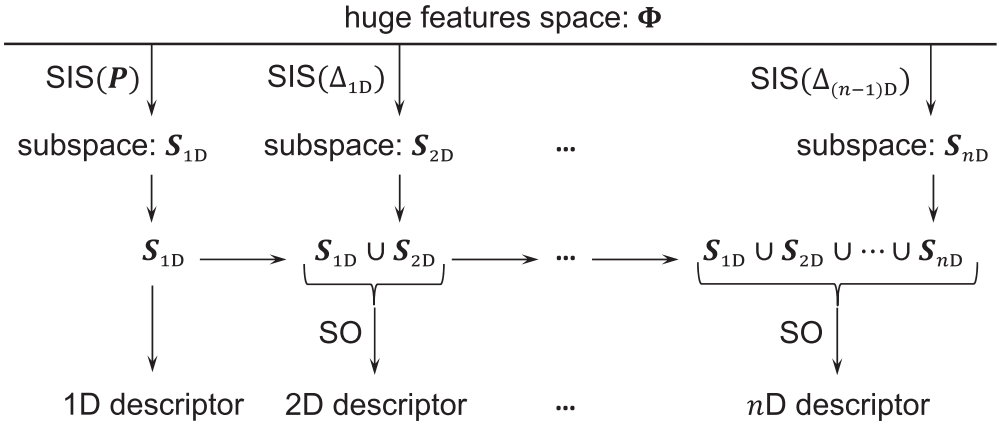

FIG. 1. The method SISSO combines unified subspaces having the largest correlation with residual errors $\boldsymbol{\Delta}$ (or $\boldsymbol{P}$) generated by sure independence screening (SIS) with sparsifying operator (SO) to further extract the best descriptor.

to 1: SISSO becomes OMP. From inside $\boldsymbol{S}_{1D}$, SO(ℓ_0) finds the best 1D descriptor, which is trivially the first ranked feature. In other words, the SIS solution in one dimension is already the SISSO solution. The residual error for an n-dimensional model is defined as $\boldsymbol{\Delta}_{nD} \equiv \boldsymbol{P} - \boldsymbol{d}_{nD}\boldsymbol{c}_{nD}$, where $\boldsymbol{d}_{nD}$ is the matrix with columns being the selected features from the whole features space, and the $\boldsymbol{c}_{nD} = (\boldsymbol{d}_{nD}{}^T\boldsymbol{d}_{nD})^{-1}\boldsymbol{d}_{nD}{}^T\boldsymbol{P}$ is the least-square solution of fitting $\boldsymbol{d}_{nD}$ to $\boldsymbol{P}$. If the error, the root-mean-square of the residual $\rho_{RMS}(\boldsymbol{\Delta}_{nD})$, is below a certain threshold then descriptor is considered fit. Otherwise the method recursively considers a higher-dimensional solution. In general, for a n-dimensional descriptor, SIS selects the subspace $\boldsymbol{S}_{nD}$ with response $\boldsymbol{\Delta}_{(n-1)D}$. Then SO extracts the best nD descriptor, with response $\boldsymbol{P}$, from the union of all the previously selected subspaces $\boldsymbol{S}_{nD} \cup \boldsymbol{S}_{(n-1)D} \cup \cdots \cup \boldsymbol{S}_{1D}$. Candès and Romberg [52] have shown that to identify the best n-dimensional descriptor with overwhelming probability the size of the response—in our case the number of materials observations $\boldsymbol{P}$—needs to satisfy the relationship $\#\boldsymbol{P} \gtrsim k \cdot n \cdot \log(\#\Phi)$, where k is a constant (around $1 \sim 10$ [31]) and $\#\Phi$ is the size of the features space [32]. Differently from the typical CS scenario, here $\#\boldsymbol{P}$ is fixed [31]; then, when $\#\Phi$ increases, the maximum n decreases in order to satisfy the relationship [52]. In practice, features spaces of growing sizes (Φ_0, Φ_i, …) and different n are tested until a model with required accuracy [$\rho_{RMS}(\boldsymbol{\Delta}_{nD}) <$ threshold] is obtained.

SISSO has advantages over MP [45] and OMP [46]. MP searches a linear model reproducing $\boldsymbol{P}$ by adding dimensionality to a descriptor while preserving selected features and corresponding coefficients. OMP improves MP by reoptimizing the coefficients every time a new component is introduced, $n \to n+1$, but still preserving previously selected features. SISSO both reselects features and reoptimizes coefficients at each dimensional increment. SISSO reduces to OMP when each subspace in the union has unit size ($\#\boldsymbol{S}_{iD} = 1, \forall i$). Still, it differs from iterative SIS [50], which reduces to simple MP when all $\#\boldsymbol{S}_{iD} = 1$.

Benchmark: Quantitative prediction. SISSO is benchmarked by comparing the relative stability of octet binary materials between rock-salt (RS) and zinc-blende (ZB) configurations. The reference data is taken from Ref. [7], including the target calculated *ab initio* enthalpy difference, RS and ZB for 82 materials and the 23 primary features related to material compositions forming Φ_0 (see Supplemental Material [40] for a list of the primary features considered in this study). All quantities are calculated with density-functional theory in the local-density approximation. Details are given in Refs. [7,31]. Then, with a combination of the previously defined operator set, $\hat{\boldsymbol{H}}^{(m)}$, and Eq. (1), the features spaces $\boldsymbol{\Phi}_1$ (small, $\#\boldsymbol{\Phi}_1 = 556$), $\boldsymbol{\Phi}_2$ (large, $\#\boldsymbol{\Phi}_2 \sim 10^5$), and $\boldsymbol{\Phi}_3$ (huge, $\#\boldsymbol{\Phi}_3 \sim 10^{11}$) are constructed.

Figure 2(a) shows the training errors (ρ_{RMS}) of different SO: LASSO, LASSO+ℓ_0, OMP, and ℓ_0 are compared while operating on the small features space $\boldsymbol{\Phi}_1$. LASSO suffers because of the correlations existing inside $\boldsymbol{\Phi}_1$ (see Fig. S1 in the Supplemental Material [40] for a figure showing the correlation between features); LASSO+ℓ_0 and OMP both surpass LASSO; ℓ_0 is the reference: it gives the exact global minimum solution for descriptors of any dimension. However, even with ℓ_0 the error is still too large for many thermodynamical predictions, $\rho_{RMS}(\boldsymbol{\Delta}_{nD}) \gtrsim 40$ meV/atom, and this is due to the too-small size of $\boldsymbol{\Phi}_1$.

Figure 2(b) shows, for the larger $\boldsymbol{\Phi}_2$, SIS combined with LASSO+ℓ_0 as SO [SISSO (LASSO+ℓ_0)], SISSO(ℓ_0), and OMP are compared for generating a 3D descriptor: SISSO(ℓ_0) is the only approach improving consistently with subspace size $\#\cup\boldsymbol{S}_{iD}$ and it always surpasses OMP when each $\#\boldsymbol{S}_{iD} \gg 1$; SISSO (LASSO+$\ell_0$) does not improve over OMP because of the failure of LASSO in dealing with correlated features [43]. Obviously, the larger the features space, the better the obtainable model (at least equal). When exhaustive searches become computationally impossible, SISSO can still find the optimal solution if the subspace produced by SIS is big enough.

Figure 2(c) shows the errors for one- to five-dimensional descriptors calculated by SISSO(ℓ_0) while operating in the large $\boldsymbol{\Phi}_2$ and huge $\boldsymbol{\Phi}_3$ spaces. For $n = 1$, SIS reduces to the best 1D descriptor, so no ℓ_0 is needed. For $n = 2, 3, 4, 5$ the size of the SIS subspace is chosen to follow the previously mentioned relationship [52] applied to the subspace $\#\boldsymbol{S} \sim \exp(\#\boldsymbol{P}/kn)$. With $\#\boldsymbol{P} = 82$ and $k = 3.125$, the total size of all the selected subspaces is $\#\cup\boldsymbol{S}_{iD} = 5 \times 10^5, 6 \times 10^3, 7 \times 10^2, 2 \times 10^2$ for $n = 2, 3, 4, 5$, respectively. For all these sizes, the application of ℓ_0 regularization as SO involves 10^{10}–10^{11} independent least-square-regression evaluations.

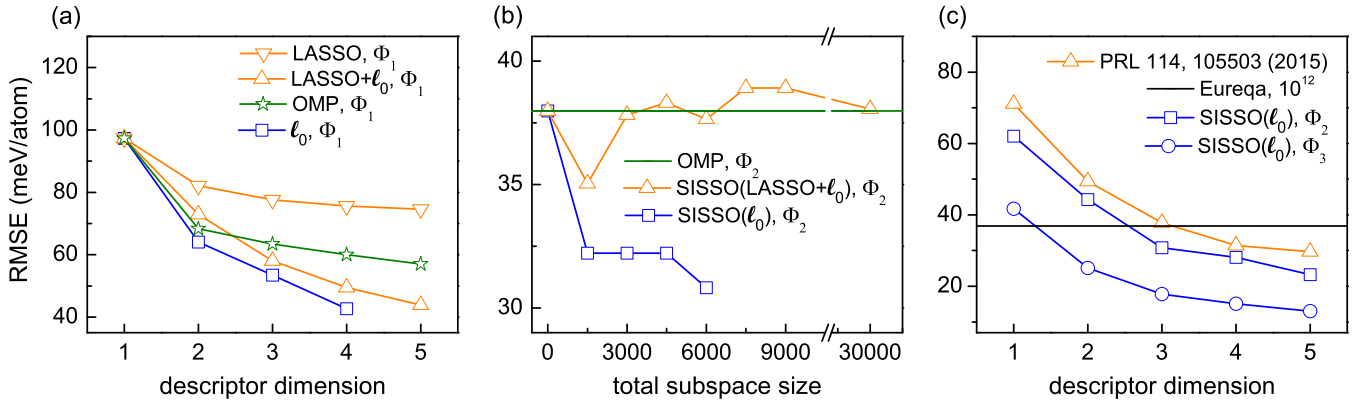

FIG. 2. Benchmark of algorithms. (a) Training error: RMSE versus descriptor dimension for different SOs operating on the smallest $\mathbf{\Phi}_1$. (b) Training error: RMSE versus subspace size in the SIS step to find a three-dimensional (3D) descriptor by OMP or SISSO with the same large features space $\mathbf{\Phi}_2$ (see Supplemental Material [40] for a similar picture for a 2D descriptor). (c) Training error: RMSE by SISSO(ℓ_0) with $\mathbf{\Phi}_2$ and $\mathbf{\Phi}_3$ compared with previous work [7] (features space size $\sim$ 4500) and with the EUREQA software [41] (evaluated functions 10^{12}, larger than $\#\mathbf{\Phi}_3$).

This is computationally feasible due to our (trivially) parallel implementation of SISSO (for instance, for this application, the production calculations were run on 64 cores). The training errors for the descriptors identified from $\mathbf{\Phi}_3$ are systematically better than those coming from $\mathbf{\Phi}_2$, thanks to the higher complexity (see Supplemental Material [40] for the functional forms of the descriptors). SISSO(ℓ_0) with $\mathbf{\Phi}_2$ is systematically better than the previous work by Ghiringhelli *et al.* [7,31], due to the allowed larger features spaces. Note that when SISSO(ℓ_0) is applied to the same features space as in Ref. [7], it also finds the same descriptor: SISSO combined with the features space of Ref. [7] has the same results of the yellow line of Fig. 2(c). Performance is also compared with the commercial software EUREQA [41] by using the same operator set and primary features ($\mathbf{\Phi}_0$), and 10^{12} evaluated functions, a number comparable to $\#\mathbf{\Phi}_3$. SISSO(ℓ_0) in $\mathbf{\Phi}_3$ with $n \geqslant 2$ and SISSO(ℓ_0) in $\mathbf{\Phi}_2$ with $n \geqslant 3$ have both lower training error than EUREQA.

In order to directly compare, over the same dataset, the ability of different approaches to find optimal or close-to-optimal solutions of the CS problem, in Figs. 2(a)–2(c) we illustrated training errors. With practical applications in mind, it is imperative to determine the performance of the obtained model on data that are not used for the training. In statistical learning [53,54], this is performed via cross validation (CV), a class of techniques that, by splitting the data set into a training and a test set in various ways, aims at detecting underfitting and overfitting, i.e., when the complexity of the fitted model is too small or too large, respectively. In CS, dedicated CV techniques have been proposed [55,56]. Specifically, in a CS-based iterative technique such as SISSO, the only source of overfitting can come from a too large dimensionality of the descriptor [note that there is only one fitting coefficient per dimension, i.e., features recursively built via Eq. (1) do not contain fitting parameters]. For this benchmark application, we applied the CS-CV scheme proposed in Ref. [55] with leave-10%-out (LTO) CV (the data set is split 40 times in a training set containing 90% randomly selected data points and a test set with the remaining 10%) and leave-one-out (LOO) CV (one data point constitutes the test set, and the procedure is iterated $\#P$ times). The model is trained on the training set (the whole SISSO procedure, i.e., including the selection of the descriptor) and the error is measured on the test set. In such framework, the CV error decreases with the descriptor dimension, until the approximate descriptor will try to fit the data (containing possible errors) starting from primary features having intrinsic limitations, thus causing a subsequent increase in the CV error. The descriptor dimension at which the CV error starts increasing identifies the maximum dimensionality of that particular model. This is determined by the features space, in turn determined by set of primary features, operators set, and number of iterations of the features space construction, and the training set. CS-CV is performed for Φ_3 with the subspace sizes reported in the description of Fig. 2(c), and for subspace of unit size (for which SISSO becomes OMP). It is found that the dimensionality minimizing the error is two for both the CV schemes of SISSO(ℓ_0). In order to achieve a smaller prediction error, one would then need to add new primary features, possibly substituting features that are never selected in a descriptor, or increase the complexity of the features space, or both. OMP finds the same dimensionality of the problem ($2 \sim 3$), has a lower computational cost but a cost of worse performance in terms of prediction error.

Figure 3(b) depicts the box plots for the distribution of errors as function of the dimensionality for SISSO(ℓ_0)-LTOCV results with features space Φ_3 [RMSE shown in Fig. 3(a)]. The 1st and 99th percentiles (extrema of the error bar), the 25th and 75th percentiles (lower and upper limits of the rectangle), and the median (50th percentile, intermediate horizontal line) are marked. The maximum absolute errors are also indicated by crosses. The worsening of the RMSE beyond two dimensions is mainly determined by an increase in the largest errors (the 99th percentile), while most of the errors remain small (median and lower percentiles $\sim$ constant).

LOOCV is also used to inspect how often the same descriptor is selected. The test operates in $\#\Phi_3$ with SISSO(ℓ_0). The LOOCV descriptor agrees with the one found over all data 79, 73, 58 times out of 82 iterations. It is remarkable, as the size of Φ_3 is of the order 10^{11} features and there

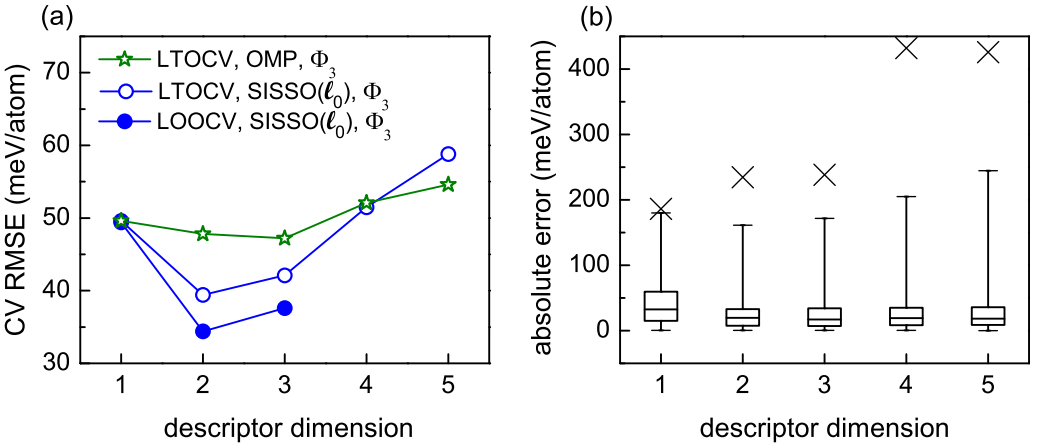

FIG. 3. Benchmark of algorithms. (a) Cross validation: LTOCV and LOOCV results for the features space Φ_3 with OMP and SISSO (ℓ_0). (b) Cross validation: Box plots of the absolute errors for the SISSO(ℓ_0)-LTOCV results with features space Φ_3. The upper and lower limits of the rectangles mark the 75th and 25th percentiles of the distribution, the internal horizontal line indicates the median (50th percentile), and the upper and lower limits of the error bars depict the 99th and 1st percentiles. The crosses represent the maximum absolute errors.

are only 82 data points. This means that the 1D, 2D, 3D descriptor is selected from 10^{11}, 10^{22}, 10^{32} combinations, respectively. We note that descriptors that are selected using the reduced training data set need be correlated with the full data-set descriptors, implying the existence of a hidden correlation between the functional forms. Hence, selecting different descriptors does not imply overfitting (this is independently determined via CS-CV), but choosing different existing approximate functional relationship among the primary features.

Application: Classification models. The SISSO framework can be readily adapted to predict categorical properties (as opposed to continuous properties such as an energy difference), i.e., it can be applied for classification. In the space of descriptors, each category's domain is approximated as the region of space (area, in two dimensions) within the convex hull of the corresponding training data. SISSO finds the low-dimensional descriptor yielding the minimum overlap (or maximum separation) between convex regions. Formally, given a property with M categories, the norm for classification is defined as:

$$\hat{\boldsymbol{c}} \equiv \arg\min_{\boldsymbol{c}} \left(\sum_{i=1}^{M-1} \sum_{j=i+1}^{M} O_{ij} + \lambda \|\boldsymbol{c}\|_0 \right), \tag{2}$$

where O_{ij} is the number of data in the overlap region between the i and j domain, $\boldsymbol{c}$ is a sparse vector (0/1 elements) so that a feature k is selected (deselected) when $c_k = 1(0)$, and λ is a parameter controlling the number of nonzero elements in $\boldsymbol{c}$. Of all the possible solutions of Eq. (2) having the same dimension and overlap, we chose the one with minimum n-dimensional overlap volume [57]:

$$\Omega \equiv \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \frac{\Omega_{ij}}{\min(\Omega_i, \Omega_j)}, \tag{3}$$

where Ω_i, Ω_j, and Ω_{ij} are the n-dimensional volumes of the i, j, and overlap ij domains. Finally, the correlation between property and feature for SIS is defined as $\left(\sum_{i=1}^{M-1}\sum_{j=i+1}^{M} O_{ij} + 1\right)^{-1}$: high correlation $\Leftrightarrow$ low overlap.

SISSO for classification is tested on a simple metal/nonmetal classification of binary systems. The training systems are far from creating an exhaustive list and, as such, the test is strictly meant for benchmarking the validity and implementation of Eqs. (2)–(3). All essential atomic and structural parameters are included as primary features in $\boldsymbol{\Phi}_0$. They originate from the WebElements [58] (atomic) and SpringerMaterials [59] (structural) databases (see Supplemental Material [40] for a list of the features considered in this study). Among them are the Pauling electronegativity χ, ionization energy IE, covalent radius r_{cov}, electron affinity, valence [number of valence electrons for A and (eight-valence) for B], coordination number, interatomic distance between A and B in crystal, atomic composition x_A, and a packing fraction, here the normalized ratio between the volume of spherical atoms and the unit cell: $\sum V_{\text{atom}}/V_{\text{cell}}$ with $V_{\text{atom}} = 4\pi r_{\text{cov}}^3/3$. The operator set $\hat{\boldsymbol{H}}^{(\text{m})}$ and Eq. (1) are then used to generate $\boldsymbol{\Phi}_3$ ($\sim 10^8$ elements). Note that SISSO finds its optimal descriptor based on combinations of the input physical quantities (features): nonoptimal outcomes indicate that the target property depends on features not yet considered in $\boldsymbol{\Phi}_0$. As such, to avoid garbage in, garbage out, SISSO requires physical intuition in the choice of features to add: conveniently, important and nonimportant features will be automatically promoted or neglected. Here, since metallicity also depends on interstitial charge, the inclusion of a packing fraction related to superpositions of orbitals is advantageous. Given a set of features, SISSO finds their best combination leading to the optimum descriptor. If the packing fraction were removed from the primary list, SISSO would autonomously select the combination of features trying to replicate as much as possible the lost descriptive power, in this case the AB atomic distances (see Table I in the Appendix). The experimental binary data set, extracted from the SpringerMaterials database [59] and used for training the SISSO model, contains A_xB_{1-x} materials having: (i) every possible A species; (ii) B as p-block element (plus H and with the condition $A \neq B$, i.e.,

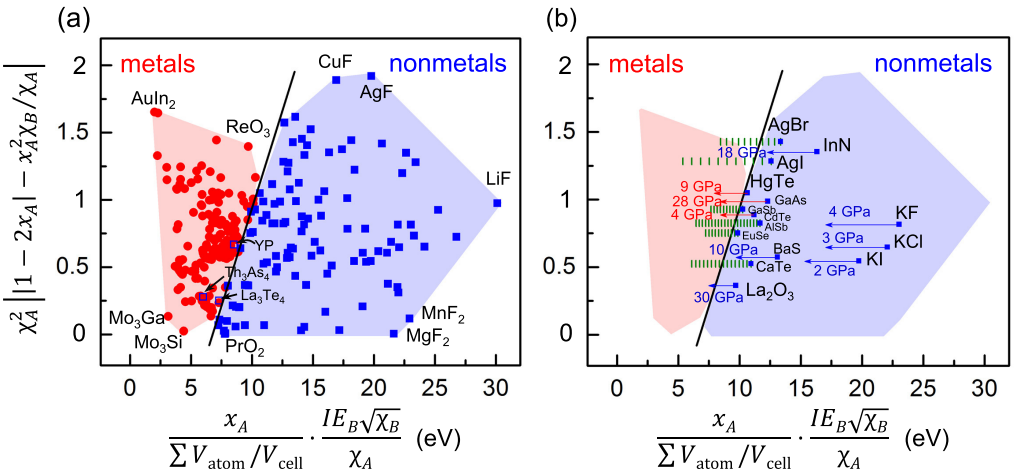

FIG. 4. SISSO for classification. (a) An almost perfect classification (99%) of metal/nonmetal for 299 materials. Symbols: χ, Pauling electronegativity; IE, ionization energy; x, atomic composition; $\sum V_{\text{atom}}/V_{\text{cell}}$, packing fraction. Red circles, blue squares, and open blue squares represent metals, nonmetals, and the three erroneously characterized nonmetals, respectively. (c) Reproduction of pressure-induced insulatormetals transitions (red arrows), of materials that remain insulators upon compression (blue arrows), and computational predictions at step of 1 GPa (green bars).

elemental solids, such as carbon diamond, are not tackled); (iii) nonlayered structure and without AA and BB dimers (the coordination polyhedron of A comprises only B atoms, and vice versa); (iv) good experimental characterization and without large distortions (we do not have any distortion feature). A total of 299 binaries in 15 prototypes (NaCl, CsCl, ZnS, CaF_2, Cr_3Si, SiC, TiO_2, ZnO, FeAs, NiAs, Al_2O_3, La_2O_3, Th_3P_4, ReO_3, ThH_2) are then used (see Supplemental Material [40] for a list of the training materials). Details on the feature-space construction and model identification are given in the Appendix. Out of $\mathbf{\Phi}_3$, SISSO(ℓ_0) identifies a two-dimensional descriptor with a training accuracy of $\sim 99.0\%$. The convex domains, indicating metallic and nonmetallic materials, are shown in Fig. 4. The figure also includes a line calculated with a support-vector machine [60], to help visualizing the separation between convex domains. These plots are called material-properties maps (or charts [7,61–64]) and SISSO has been specifically designed to identify low-dimensional regions, possibly nonoverlapping.

Figure 4(a) shows the three incorrectly classified systems (blue empty squares). YP (NaCl prototype) might have a slightly erroneous position in the figure: the covalent radius r_{cov}(Y) (controlling the packing fraction) suffers from large intrinsic errors (see Fig. 2 of Ref. [65]) and therefore the compound position might be misrepresented. La_3Te_4 and Th_3As_4 (Th_3P_4 prototype) are different. In this case, SISSO indicates that the primary features are not enough or that the compounds have been experimentally misclassified (due to defects or impurities [66–68]). Inspection of the found descriptor suggests a justification of the involved primary features. The x projection [x axis in Fig. 4(a)] indicates that the higher the packing fraction $\sum V_{\text{atom}}/V_{\text{cell}}$, i.e., the higher the interstitial charge, the higher the propensity of a material to be a metal. This is not surprising. The merit of the descriptor found by SISSO is to (i) provide a quantitative account of the dependence of metallicity on the packing fraction, allowing for predictions (see below) and (ii) reveal the functional form packing fraction metallicity: It is not trivial that the descriptor is linear with the inverse packing fraction. Metallicity also correlates with the electronegativity of the A species, often the main electron donor, by competing against the B species, a p element trying to complete its covalent/ionic bonds by filling the unoccupied orbitals and thus removing interstitial charge. Thus it is not surprising that the material with largest x projection is LiF, a purely ionic compound with closed electron shells: the ratio among the two extreme electronegativities, (Li has the lowest, F the highest), pushes the compound toward the rightmost corner of the nonmetals domain. On the other side, $AuIn_2$ is the compound farthest from the nonmetals region: Au has the highest χ among transition-metals and In has one of the smallest χ of the considered p elements. Available experimental band gaps were also extracted (see Supplemental Material [40] for a figure showing distribution of band gaps). The robustness of the descriptor is corroborated by leave-one-out cross validation. In 97.6% of the times, LOOCV reproduces the same functional solution obtained from the whole data. In the few cases where the descriptor differs from the all-data one, the packing fraction always remains; even more: the packing fraction is present in all features selected by SIS at the first iteration.

Beyond the training: Prediction of metallization by compression. Although pressure is neither included in the features space nor in the training data, its effect can be tested by reducing V_{cell}. Among the training data, we have three systems experiencing pressure-induced insulator-to-metal transition: HgTe, GaAs, and CdTe. HgTe, CdTe, and GaAs go from insulating zinc blende to metallic rock salt (or an orthorhombic oI4 phase for GaAs) at ~ 9, 4, and 28 GPa, respectively (see red arrows). Geometrical parameters (cell volumes) at normal and high pressure are taken from the experimental databases and used to modify the x coordinate of the descriptor. Concurrently, we have also looked for materials that do not become metallic with high-pressure structural transitions (indicated by the blue arrows). In this case our model again makes a correct prediction. Figure 4(b) shows that the descriptor is perfectly capable of reproducing the correct metallic state. The idea can

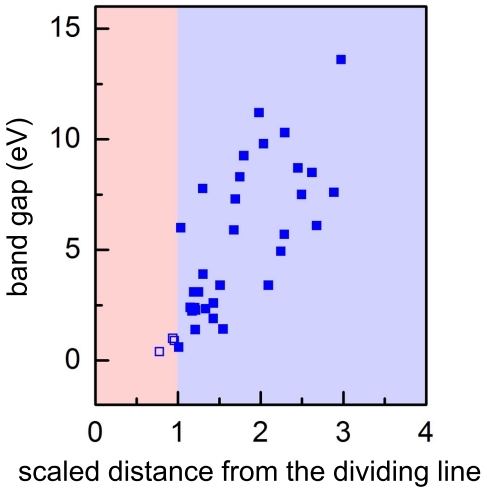

FIG. 5. SISSO for classification. Correlation between the band gap of the nonmetals and the scaled coordinate from the dividing line.

be extended to systems that have not yet been fully characterized to predict potential insulator-to-metal transitions. The subset of prototypes which are reasonably close to the domain convex hull and have a fully characterized *ab initio* elastic tensors [69] are computationally compressed by having their V_{cell} reduced following the first-order linearized bulk modulus relation: $(V_{\text{cell}}(p) - V_{\text{cell}}(0))/V_{\text{cell}}(0) \sim -p/B_T$, where p is the pressure and B_T is the isothermal bulk modulus extracted from the entries in the AFLOW.org repository [69] (see SI for the entries data). The panel shows a set of compounds for which the descriptor predicts the transition to metallic. The green marks are positioned at 1 GPa steps to allow an informed guess of the pressure. Within this approximation, some compounds are predicted to become metallic at pressure between 5 and 15 GPa: AgBr, AgI, GaSb, AlSb, EuSe, and CaTe. Pressure-induced structural phase transitions are also not considered in such analysis and thus, the insulator-to-metal transition pressure might be overestimated facilitating experimental validations.

Beyond the training: Significance of the distance from the dividing line. Figure 5 depicts the experimental band gap of the insulators vs. the scaled distance from the dividing line, i.e., the dimensionless ratio between the x projection of its descriptor versus the x projection of the dividing line corresponding to the y projection of its descriptor value. With this rescaling, the dividing line corresponds to the vertical line $x = 1$. The trend of the data points reveals that the descriptor found by SISSO, trained only on a categorical property, includes a quantitative, albeit approximate, account of how strongly an insulator is far from being a metal, by locating materials with large band gaps further from the line than small-gap materials.

General remarks on the descriptor-property relationship identified by SISSO. As clear from the two application cases presented here, the equations found by SISSO are not necessarily unique and all components of the descriptors may change at each added dimension. This reflects the approximate nature of the equations and the unavoidable relationships among features (one or more primary features may be accurately described by nonlinear functions of a subset of the remaining features). We also note that the mathematical constraints imposed in order to obtain solutions efficiently (linear combination of nonlinear functions for the continuous-property case and minimally overlapping convex hulls in the classification case), are very flexible but not complete. That is, the found descriptor-property relationship is intrinsically approximate.

III. CONCLUSIONS

We have presented an efficient approach for extracting effective materials descriptors out of huge and possibly strongly correlated features spaces. This algorithm, called SISSO (sure independence screening and sparsifying operators) tackles huge spaces while retaining the effectiveness of compressed sensing. Specifically, SISSO is built to work also (but not limited to) when only relatively small training sets are available. SISSO autonomously finds the best descriptor from a combination of features (physical properties), and it is capable of determining the ones not relevant to the problem, so that the features space can be further optimized. SISSO identifies the descriptor-property relationship in terms of an analytical equation. It does not need to be exact—a simple, analytical descriptor-property function may not even exist—but it is the most accurate expression given the available features space. If an exact, analytic expression does indeed exist, SISSO is expected to find it if included in the features space.

SISSO shows superior advantages with respect to other established methods, e.g., OMP and LASSO as well as the software EUREQA, especially when dealing with a correlated features spaces. SISSO does not have the limitation of LASSO, which suffers with large and highly correlated features spaces. Currently, the only issue of SISSO is the required computer memory needed to handle the features space, and efforts are underway for more efficient implementations. Our approach is benchmarked on the quantitative modeling of enthalpy differences for a set of zinc-blende and rock-salt prototypes and applied to the metal/insulator classification of binaries. The robustness of the classification is corroborated by the proper reproduced insulator-metal transitions, which allows us to predict a set of systems for further experimental analyses.

ACKNOWLEDGMENTS

The authors thank Daria M. Tomecka, Cormac Toher, and Corey Oses for their valuable help in collecting the data for the metal/insulator application. Angelo Ziletti, Ankit Kariryaa, and Fawzi Mohamed are gratefully acknowledged for helping setting up the Python notebooks and the overall Analytics-Toolkit infrastructure (see section "Reproducibility"). This project has received funding from the European Unions Horizon 2020 research and innovation program (No. 676580: The NOMAD Laboratory - a European Centre of Excellence and No. 740233: TEC1p), and the Berlin Big-Data Center (BBDC, No. 01IS14013E). S.C. acknowledges DOD-ONR (N00014-13-1-0635, N00014-11-1-0136, N00014-15-1-2863) and the Alexander von Humboldt Foundation for financial support.

APPENDIX

In this Appendix, we present details on the metal/insulator-classification application.

TABLE I. Dependence of the metal-insulator classification descriptors on the prototypes of training binary materials.

prototypes	#materials	primary features	descriptor	classification accuracy
NaCl	132	IE_A, IE_B, χ_A, χ_B, r_{covA}, r_{covB}, EA_A, EA_B, v_A, v_B, d_{AB}	$d_1 := \frac{IE_A IE_B (d_{AB} - r_{\text{covA}})}{\exp(\chi_A)\sqrt{r_{\text{covB}}}}$	100%
NaCl, CsCl, ZnS, CaF_2, Cr_3Si	217	IE_A, IE_B, χ_A, χ_B, r_{covA}, r_{covB}, d_{AB}, CN_A, CN_B	$d_1 := \frac{IE_B d_{AB}^2}{\chi_A r_{\text{covA}}^2 \sqrt{CN_B}}$, $d_2 := \frac{IE_A^2 r_{\text{covB}} \log(IE_A) \lvert r_{\text{covA}} - r_{\text{covB}} \rvert}{CN_B}$	100%
NaCl, CsCl, ZnS, CaF_2, Cr_3Si, SiC, TiO_2, ZnO, FeAs, NiAs	260	IE_A, IE_B, χ_A, χ_B, r_{covA}, r_{covB}, d_{AB}, CN_A, CN_B	$d_1 := \frac{d_{AB}/r_{\text{covA}} - \chi_A/\chi_B}{\exp(CN_B/IE_B)}$, $d_2 := \frac{r_{\text{covA}}^3 d_{AB} IE_B}{\lvert \chi_B/\chi_A - \lvert CN_B - CN_A \rvert \rvert}$	99.6%[a]
NaCl, CsCl, ZnS, CaF_2, Cr_3Si, SiC, TiO_2, ZnO, FeAs, NiAs	260	IE_A, IE_B, χ_A, χ_B, x_A, x_B, $V_{\text{cell}}/\sum V_{\text{atom}}$	$d_1 := \frac{V_{\text{cell}}}{\sum V_{\text{atom}}} \frac{\sqrt{\chi_B}}{\chi_A}$, $d_2 := \frac{IE_A IE_B}{\exp(V_{\text{cell}}/\sum V_{\text{atom}})}$	99.6%[a]
NaCl, CsCl, ZnS, CaF_2, Cr_3Si, SiC, TiO_2, ZnO, FeAs, NiAs, Al_2O_3, La_2O_3, Th_3P_4, ReO_3, ThH_2	299	IE_A, IE_B, χ_A, χ_B, x_A, x_B, $V_{\text{cell}}/\sum V_{\text{atom}}$	$d_1 := \frac{x_B}{\sum V_{\text{atom}}/V_{\text{cell}}} \frac{IE_B\sqrt{\chi_B}}{\chi_A}$, $d_2 := \chi_A^2 \lvert \lvert 1 - 2x_A \rvert - x_A^2 \frac{\chi_B}{\chi_A} \rvert$	99.0%[b]

[a]One entry misclassified: YP compound in NaCl prototype.

[b]Three entry misclassified: YP compound in NaCl prototype; Th_3As_4 and La_3Te_4 compounds in Th_3P_4 prototype.

Primary features. Descriptors are to be identified by SISSO from a systematically constructed large/huge features space in which components are generated by recursively transforming a set of input primary features, Φ_0, via algebraic operations, $\hat{\boldsymbol{H}} \equiv \{I, +, -, \times, /, \exp, \log, |-|, \sqrt{}, ^{-1}, ^2, ^3\}$. Primary features usually comprise properties of isolated atoms (atomic features) and properties of the materials (composition and geometry). For the test on binaries' metal/nonmetal classification, the following is the full list of considered primary features:. (i) first ionization energy, IE_A (A species) and IE_B (B species); (ii) electron affinity, EA_A and EA_B; (iii) atom covalent radius, r_{covA} and r_{covB}; (iv) Pauling electronegativity, χ_A and χ_B; (v) valence, v_A (number of valence electrons) and v_B (8 number of valence electrons); (vi) coordination number, CN_A (number nearest-neighbor B of A) and CN_B; (vii) interatomic distance between A and B in crystal, d_{AB}; (viii) atomic composition x_A (or $x_B = 1 - x_A$); and (ix) the ratio of the cell volume to the total atom volume in the unit cell of the crystal, $V_{\text{cell}}/\sum V_{\text{atom}}$ ($V_{\text{atom}} = 4\pi r_{\text{cov}}^3/3$).

It is critical to limit the redundant and unnecessary primary features in Φ_0 to enhance computational performance (the size of features space Φ_n increases very fast with $\#\Phi_0$) and to increase SIS success rate: the higher #subspace/#Φ, the higher the probability that SIS subspaces contain the best models. Starting from an empty Φ_0, few primary features are added. SISSO is then applied to identify the best model, with $\hat{\boldsymbol{H}}$ as operators space. If an appropriate quality of the model is not achieved (e.g., the number of correctly classified materials is lower than a desired threshold), other primary features are added in Φ_0 to check for improvements. Primary features preserved in Φ_0 may become redundant or unnecessary on a later stage, e.g., when new ones are added. To retain computationally manageable sizes of the features space, tests are performed to remove those primary features that either are never appearing in the identified descriptor or that do not improve the performance of the model (in this specific case, when the number of correctly classified materials does not increase). Eventually, Φ_0 will converge to the best possible small set of primary features, along with the best models that can be generated from it.

Data variety. The influence of data variety on the descriptors is investigated and Table I shows how the metal-insulator classification descriptors depend on the prototypes of training materials.

The first calculation starts with a data set of all the available materials (132) in NaCl prototype. The initial features space, Φ_0, contains the primary features of all the 10 atomic parameters (Table I), and one structural parameter of interatomic distance d_{AB} to capture the geometrical differences between the training rock-salt materials. SISSO is then applied: (i) Φ_3 is constructed; (ii) the best descriptor is identified from Φ_3 for classifying the metals and insulators with 100% accuracy. The simple descriptor is shown in Table I. It indicates that a rock-salt compound tends to become nonmetal when the large interatomic distance is decreased with the radius of species A.

Next, the number of prototypes is increased to five, for a total of 217 materials. However, with the previous Φ_0 and calculation settings, SISSO fails to identify a descriptor having perfect classification (there are seven points in the overlap region between the metal and nonmetal domains). The nonoptimal outcome indicates that the classification depends on primary features not yet considered. First, Φ_0 is slimmed by reducing its size to 7—EA_A, EA_B, v_A, and v_B are removed—without affecting the quality of the predictions (eight points in the overlap region). Second, two new features CN_A and CN_B are added ($\#\Phi_0 \rightarrow 9$) to describe the different coordination environments of the prototypes. SISSO finds a 2D descriptor from the constructed Φ_3 with 100% classification, shown in Table I. From the descriptor, the geometrical differences between training materials are captured by the two features of d_{AB} and CN_B: systems belonging to such five prototypes with large d_{AB} and small CN_B tend to be nonmetals.

The number of prototypes is increased to 10, for a total of 260 materials. As shown in Table I, with the previous $\#\Phi_0 = 9$, the identified best descriptors are 2D and have 99.6% classification (only one point, YP compound in NaCl prototype, is misclassified). Although the classification is excellent, the descriptor is complicated. Searching for a

simplification, new primary features of atomic composition x_A, x_B, and $V_{\text{cell}}/\sum V_{\text{atom}}$ are introduced to replace r_{covA}, r_{covB}, d_{AB}, CN_A, and CN_B, leading to $\#\Phi_0 \rightarrow 7$. With the same training materials, SISSO finds a much simpler descriptor having the same accuracy of 99.6% (YP compound remains misclassified). This result shows that the choice of proper primary features leads to descriptors' simplification.

Finally, all the available 15 prototypes of binary materials (299) are considered and used with the seven primary features in Φ_0. With a constructed Φ_3 of size 10^8, SISSO identifies the best 2D descriptor with a classification accuracy of 99.0% (three misclassified compounds: YP compound in NaCl prototype, Th_3As_4 and La_3Te_4 in Th_3P_4 prototype). When new information (compounds and/or prototypes) is added, the functional form of the descriptors adapts. For predictive models, the data set requires all necessary information, e.g., by uniform sampling of the whole chemical and configurational space of the property of interest. The above 15 prototypes are not all the available prototypes for binary materials, and the layered materials (e.g., MoS_2, and those materials having A-A or B-B dimers, e.g., FeS_2, are not included) as the presented model is strictly illustrative of the method.

Reproducibility. To enable reproducibility, online tutorials where results can be interactively reproduced (and extended) are presented within the framework of the NOMAD Analytics-Toolkit (analytics-toolkit.nomad-coe.eu). For the RS/ZB benchmark application: analytics-toolkit.nomad-coe.eu/tutorial-SIS. For the metal-nonmetal classification: analytics-toolkit.nomad-coe.eu/tutorial-metal-nonmetal. The SISSO code, as used for the work presented here, but ready for broader applications is open source and can be found at github.com/rouyang2017/SISSO.

[1] Office of Science and Technology Policy, White House, *Materials Genome Initiative for Global Competitiveness*, https://obamawhitehouse.archives.gov/mgi, 2011.

[2] S. Curtarolo, G. L. W. Hart, W. Setyawan, M. J. Mehl, M. Jahnátek, R. V. Chepulskii, O. Levy, and D. Morgan, *AFLOW: Software for High-Throughput Calculation of Material Properties,* http://materials.duke.edu/aflow.html, 2010.

[3] A. Jain, G. Hautier, C. J. Moore, S. P. Ong, C. C. Fischer, T. Mueller, K. A. Persson, and G. Ceder, A high-throughput infrastructure for density functional theory calculations, Comput. Mater. Sci. **50**, 2295 (2011).

[4] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD), JOM **65**, 1501 (2013).

[5] D. D. Landis, J. Hummelshøj, S. Nestorov, J. Greeley, M. Dułak, T. Bligaard, J. K. Nørskov, and K. W. Jacobsen, The computational materials repository, Comput. Sci. Eng. **14**, 51 (2012).

[6] A. A. White, Big data are shaping the future of materials science, MRS Bull. **38**, 594 (2013).

[7] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, Big Data of Materials Science: Critical Role of the Descriptor, Phys. Rev. Lett. **114**, 105503 (2015).

[8] S. R. Kalidindi and M. De Graef, Materials data science: Current status and future outlook, Annu. Rev. Mater. Res. **45**, 171 (2015).

[9] W. Sun, S. T. Dacek, S. P. Ong, G. Hautier, A. Jain, W. D. Richards, A. C. Gamst, K. A. Persson, and C. Gerbrand, The thermodynamic scale of inorganic crystalline metastability, Sci. Adv. **2**, e1600225 (2016).

[10] E. Perim, D. Lee, Y. Liu, C. Toher, P. Gong, Y. Li, W. N. Simmons, O. Levy, J. J. Vlassak, J. Schroers, and S. Curtarolo, Spectral descriptors for bulk metallic glasses based on the thermodynamics of competing crystalline phases, Nat. Commun. **7**, 12315 (2016).

[11] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, NPJ Comput. Mater. **2**, 16028 (2016).

[12] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, and A. Tropsha, Universal fragment descriptors for predicting electronic properties of inorganic crystals, Nat. Commun. **8**, 15679 (2017).

[13] K. Fujimura, A. Seko, Y. Koyama, A. Kuwabara, I. Kishida, K. Shitara, C. Fisher, H. Moriwake, and I. Tanaka, Accelerated materials design of lithium superionic conductors based on first-principles calculations and machine learning algorithms, Adv. Energy Mater. **3**, 980 (2013).

[14] S. Curtarolo, G. L. W. Hart, M. Buongiorno Nardelli, N. Mingo, S. Sanvito, and O. Levy, The high-throughput highway to computational materials design, Nat. Mater. **12**, 191 (2013).

[15] B. Meredig and C. Wolverton, A hybrid computational-experimental approach for automated crystal structure solution, Nat. Mater. **12**, 123 (2013).

[16] C. C. Fischer, K. J. Tibbetts, D. Morgan, and G. Ceder, Predicting crystal structure by merging data mining with quantum mechanics, Nat. Mater. **5**, 641 (2006).

[17] NOMAD: Novel Materials Discovery, https://www.nomad-coe.eu, 2015.

[18] C. Draxl and M. Scheffler, NOMAD: The FAIR concept for Big-Data-driven materials science, MRS Bull. (to be published) [arXiv:1805.05039].

[19] A. Bartók, P. Albert, M. C. Payne, R. Kondor, and G. Csányi, Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, Without the Electrons, Phys. Rev. Lett. **104**, 136403 (2010).

[20] J. Carrete, N. Mingo, S. Wang, and S. Curtarolo, Nanograined half-heusler semiconductors as advanced thermoelectrics: An ab initio high-throughput statistical study, Adv. Func. Mater. **24**, 7427 (2014).

[21] K. Rajan, Materials informatics: The materials "gene" and big data, Annu. Rev. Mater. Res. **45**, 153 (2015).

[22] T. Mueller, A. G. Kusne, and R. Ramprasad, *Machine Learning in Materials Science* (John Wiley & Sons, New York, 2016), pp. 186–273.

[23] C. Kim, G. Pilania, and R. Ramprasad, From organized high-throughput data to phenomenological theory using machine learning: The example of dielectric breakdown, Chem. Mater. **28**, 1304 (2016).

[24] F. A. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, Machine Learning Energies of 2 Million Elpasolite (ABC_2D_6) Crystals, Phys. Rev. Lett. **117**, 135502 (2016).

[25] K. Takahashi and Y. Tanaka, Materials informatics: A journey towards material design and synthesis, Dalton Trans. **45**, 10497 (2016).

[26] A. Bartók, S. De, C. Poelking, N. Bernstein, J. Kermode, G. Csányi, and M. Ceriotti, Machine learning unifies the modeling of materials and molecules, Sci. Adv. **3**, 1701816 (2017).

[27] B. R. Goldsmith, M. Boley, J. Vreeken, M. Scheffler, and L. M. Ghiringhelli, Uncovering structure-property relationships of materials by subgroup discovery, New J. Phys. **19**, 013031 (2017).

[28] T. L. Pham, N. D. Nguyen, V. D. Nguyen, H. Kino, T. Miyake, and H. C. Dam, Learning structure-property relationship in crystalline materials: A study of lanthanide-transition metal alloys, J. Chem. Phys. **148**, 204106 (2018).

[29] T. Hey, S. Tansley, and K. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery* (Microsoft Research, Seattle, 2009).

[30] A. Agrawal and A. Choudhary, Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science, APL Mater. **4**, 053208 (2016).

[31] L. M. Ghiringhelli, J. Vybiral, E. Ahmetcik, R. Ouyang, S. V. Levchenko, C. Draxl, and M. Scheffler, Learning physical descriptors for materials science by compressed sensing, New J. Phys. **19**, 023017 (2017).

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

[33] L. J. Nelson, G. L. W. Hart, F. Zhou, and V. Ozolins, Compressive sensing as a paradigm for building physics models, Phys. Rev. B **87**, 035125 (2013).

[34] E. J. Candès, J. Romberg, and T. Tao, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information, IEEE Trans. Inf. Theory **52**, 489 (2006).

[35] D. L. Donoho, Compressed sensing, IEEE Trans. Inform. Theory **52**, 1289 (2006).

[36] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. R. Stat. Soc. Ser. B **58**, 267 (1996).

[37] P. Sondhi, *Feature Construction Methods: A Survey*, Tech. Rep., sifaka.cs.uiuc.edu, 2009.

[38] I. Guyon and A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. **3**, 1157 (2003).

[39] P. Breen, *Algorithms for Sparse Approximation*, Tech. Rep., School of Mathematics, University of Edinburgh, 4 Year Project Report, 2009.

[40] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevMaterials.2.083802 for the list of primary features considered in both applications, the functional forms of the descriptors for the RS/ZB example, the list of training materials – and distribution of their band gaps – for the metal/insulator application.

[41] M. Schmidt and H. Lipson, Distilling free-form natural laws from experimental data, Science **324**, 81 (2009).

[42] The ℓ_0 norm of a vector is the number of its nonzero components.

[43] D. L. Donoho and M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization, Proc. Natl. Acad. Sci. U.S.A. **100**, 2197 (2003).

[44] The ℓ_1 norm is the sum of the absolute values of the components of a vector.

[45] S. G. Mallat and Z. Zhang, Matching pursuits with time-frequency dictionaries, IEEE Trans. Signal Process **41**, 3397 (1993).

[46] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition, in *The Twenty-Seventh Asilomar Conf.: Signals, Systems and Computers* (IEEE, Pacific Grove, 1993), Vol. 1, pp. 40–44.

[47] J. A. Tropp and A. C. Gilbert, Signal recovery from random measurements via orthogonal matching pursuit, IEEE Trans. Inform. Theory **53**, 4655 (2007).

[48] J. A. Tropp, Greed is good: Algorithmic results for sparse approximation, IEEE Trans. Inform. Theory **50**, 2231 (2004).

[49] E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall, Compressed sensing with coherent and redundant dictionaries, Appl. Comput. Harmon. Anal. **31**, 59 (2011).

[50] J. Fan and J. Lv, Sure independence screening for ultra-high dimensional feature space, J. R. Statist. Soc. B **70**, 849 (2008).

[51] J. Fan, R. Samworth, and Y. Wu, Ultrahigh dimensional feature selection: beyond the linear model, J. Mach. Learn. Res. **10**, 2013 (2009).

[52] E. J. Candès and J. Romberg, Sparsity and incoherence in compressive sampling, Inverse Prob. **23**, 969 (2007).

[53] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Vol. 1 (Springer Series in Statistics, Berlin, 2009), 2nd ed.

[54] J. Pearl, *Causality: Models, Reasoning and Inference* (Cambridge University Press, New York, 2009), 2nd ed.

[55] P. Boufounos, M. F. Duarte, and R. G. Baraniuk, Sparse signal reconstruction from noisy compressive measurements using cross validation, in *2007 IEEE/SP 14th Workshop on Statistical Signal Processing* (IEEE, Piscataway, 2007), pp. 299–303.

[56] R. Ward, Compressed sensing with cross validation, IEEE Trans. Inf. Theory **55**, 5773 (2009).

[57] A. F. Bialon, T. Hammerschmidt, and R. Drautz, Three-parameter crystal-structure prediction for sp-d-valent compounds, Chem. Mater. **28**, 2550 (2016).

[58] https://www.webelements.com

[59] https://materials.springer.com

[60] T. Joachims, *Making Large-Scale SVM Learning Practical*, Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund, 1998.

[61] M. F. Ashby, A first report on deformation-mechanism maps, Acta Mater. **20**, 887 (1972).

[62] D. G. Pettifor, A chemical scale for crystal-structure maps, Solid State Commun. **51**, 31 (1984).

[63] D. G. Pettifor, The structures of binary compounds. I. Phenomenological structure maps, J. Phys. C **19**, 285 (1986).

[64] O. Isayev, D. Fourches, E. N. Muratov, C. Oses, K. Rasch, A. Tropsha, and S. Curtarolo, Materials cartography: Representing and mining materials space using structural and electronic fingerprints, Chem. Mater. **27**, 735 (2015).

[65] B. Cordero, V. Gómez, A. E. Platero-Prats, M. Revés, J. Echeverría, E. Cremades, F. Barragán, and S. Alvarez, Covalent radii revisited, Dalton Trans. **21**, 2832 (2008).

[66] M. Pardo and J. Flahaut, Les systémes CaTe – $LaTe_3$ formés avec les éléments des terres rares et l'yttrium, Bull. Soc. Chim. Fr. **1969**, 6 (1969).

[67] A. F. May, J.-P. Fleurial, and G. J. Snyder, Thermoelectric performance of lanthanum telluride produced via mechanical alloying, Phys. Rev. B **78**, 125205 (2008).

[68] P. J. Markowski, Z. Henkie, and A. Wojakowski, Electronic properties of Th_3As_4-U_3As_4 solid solutions, Solid State Commun. **32**, 1119 (1979).

[69] C. Toher, C. Oses, J. J. Plata, D. Hicks, F. Rose, O. Levy, M. de Jong, M. D. Asta, M. Fornari, M. Buongiorno Nardelli, and S. Curtarolo, Combining the AFLOW GIBBS and elastic libraries to efficiently and robustly screen thermomechanical properties of solids, Phys. Rev. Mater. **1**, 015401 (2017).