This is the accepted manuscript made available via CHORUS. The article has been published as:

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

Runhai Ouyang, Stefano Curtarolo, Emre Ahmetcik, Matthias Scheffler, and Luca M. Ghiringhelli

Phys. Rev. Materials **2**, 083802 — Published 7 August 2018

DOI: [10.1103/PhysRevMaterials.2.083802](10.1103/PhysRevMaterials.2.083802)
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. Ghiringhelli1,*

1Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin-Dahlem, Germany
2Materials Science, Duke University, Durham, 27708, NC, USA
(Dated: July 16, 2018)

The lack of reliable methods for identifying descriptors — the sets of parameters capturing the underlying mechanisms of a materials 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. SISSO (sure independence screening and sparsifying operator) 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 are tested beyond the training data: It rediscovers the metal/insulator classification of binaries (with experimental data). Accurate, predictive models are found in both cases.

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 [6–16]). Much of the data of this field is available in the FAIR Repository and Archive of the NOMAD Center 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, non-toxic, 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 multi-level, 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 was 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, $P$, CS finds the sparse solution $c$ of an underdetermined system of linear equations $Dc = P$ ($D$ is called the sensing matrix with columns $\gg$ rows). If the number of nonzero entries in $c$ is smaller than the size of $P$, then CS effectively reduces the dimensionality of the problem [32, 34, 35]. In the specific case treated in [7, 31], given a set of materials $m_i$ with observable properties listed in vector $P_i$ 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 \( D \). The sparse solution of “arg min \(|\|P - Dc\|_2^2 + \lambda|c|_0\)”, where \(|c|_0\) is the number of nonzero components of \( c \), gives the optimum \( n \)-dimensional descriptor, i.e., the set of features “selected” by the the \( n \) non-zero components of the solution vector \( c \).

In Refs. [7, 31], a modification of LASSO (least absolute shrinkage and selection operator, [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 descriptors (the feature space) gets large (larger than few thousands 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 SISSO (sure independence screening and sparsifying operator), which can deal with an immensity of offered candidate descriptors (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.

**RESULTS AND DISCUSSION**

**Features e 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 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 \( \Phi_0 \) 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

\[
\vec{H}^{(m)} = \{1, +, -, \times, /, \exp, \log, |-|, \sqrt{\cdot}^{-1}, 2, 3\} [\phi_1, \phi_2],
\]

where \( \phi_1 \) and \( \phi_2 \) are objects in \( \Phi \) (for unary operators only \( \phi_1 \) is considered) and the superscript \( (m) \) indicates that dimensional analysis is performed to retain only “meaningful” combinations (e.g., no unphysical items like ‘size + energy’ or ‘size + size\(^2\)’). The intrinsically linear relationship observables \( \leftrightarrow \) descriptor in the CS formalism is made non-linear by equipping the features space with non-linear operators in \( \vec{H}^{(m)} \). At each iteration, \( \vec{H}^{(m)} \) operates on all available combinations, and the features space grows recursively as:

\[
\Phi_n \equiv \bigcup_{i=1}^n \vec{H}^{(m)} [\phi_1, \phi_2], \quad \forall \phi_1, \phi_2 \in \Phi_{i-1}. \tag{1}
\]

The number of elements in \( \Phi_n \) grows very rapidly with \( n \). It is roughly of the order of \( \sim \#(\Phi_0)^2 \times (\#\hat{H}_2)^{2^n-1} \) where \( \#\Phi_0 \) and \( \#\hat{H}_2 \) are the numbers of elements and binary operators in \( \Phi_0 \) and \( \hat{H} \), respectively. For example, \( \#\Phi_0 \approx 10^{11} \) with \( \#\hat{H}_2 = 5 \) and \( \#\Phi_0 = 10 \). 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 \( \ell_0 \)-norm regularized minimization [41] is the obvious path for finding the best sparse solution of linear equations. It is performed by SIS (sure independence screening) with SO (sparsifying operator) to further extract the best descriptor.

![Diagram](image-url)
through combinatorial optimization by penalizing
the number of non-zero coefficients. The algorithm is NP-
hard and thus infeasible when the features space becomes
very large. Efficient methods can be employed to appox-
imate the correct \( \ell_0 \) solution [42] with ideal features space
e.g., having uncorrelated basis sets). Amongst them are
the convex optimization by \( \ell_1 \)-norm [43] regularization
LASSO [36]) and the various greedy algorithms such as
the matching pursuit (MP) [44] and orthogonal matching
pursuit (OMP) [45, 46]. Unfortunately, with correlated
features spaces, approximated results can largely deviate
from the ideal \( \ell_0 \) solutions [42, 47]. Corrections have
been proposed, for example the LASSO+\( \ell_0 \) scheme
comprising LASSO prescreening and subsequent \( \ell_0 \) optimization
[7, 31], and the the \( \ell_1 \)-analysis and \( \ell_1 \)-synthesis [48].
However, when the features space size becomes of the
order of \( 10^6 \) to \( 10^9 \), \( \ell_1 \) 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) [49, 50], which
has been shown to be effective for dimensionality reduc-
tion of ultra-high dimensional features spaces [49]. SIS
scores each feature (standardized) with a metric (cor-
relation magnitude, i.e., the absolute of inner product
between the target property and a feature) and keeps
only the top ranked [49]. After the reduction, SO is used
to pinpoint the optimal \( n \)-dimensional descriptor. The
smaller the dimensionality, the better the outcome: pro-
gressively larger \( n \) are tested until the “left-over” 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 \( S_{\text{id}} \) containing the
features having the largest correlation with the response
\( P \) (target material property). Generally, the larger
the subspace \( \cup S_{\text{id}} \), the higher the probability it con-
tains the optimal descriptor. However, the chosen size
of \( \cup S_{\text{id}} \), depends on \( i \) which type of SO is later used,
\( ii \) the dimensionality \( n \) requested, and \( iii \) the avail-
able computational resources. With SO(LASSO), \( \cup S_{\text{id}} \)
can contain as much as \( 10^5 \) to \( 10^6 \) elements, depending
on \#\( P \). With SO(\( \ell_0 \)), the largest obtainable size is typ-
ically \( 10^5 \) for \( n = 2, 10^3 \) for \( n = 3, 10^2 \) for \( n = 4 \), etc.
(because the number of needed evaluation grows com-
binatorially with \( n \)). If \( n \) is large, e.g., \( >10 \), then the
maximum possible \#\( S_{\text{id}} \) converge to 1: SISSO becomes
OMP. From inside \( S_{\text{id}} \), SO(\( \ell_0 \)) finds the best 1D descrip-
tor, which is trivially the first ranked feature. In other
words, the SIS solution in 1D is already the SISSO solu-
tion. The residual error for a \( n \)-dimensional model is
defined as \( \Delta_{\text{id}} = P - d_{\text{id}}c_{\text{id}} \), where \( d_{\text{id}} \)
is the matrix with columns being the selected features from the
whole features space, and the \( c_{\text{id}} = (d_{\text{id}}^T d_{\text{id}})^{-1} d_{\text{id}}^T P \) is
the least square solution of fitting \( d_{\text{id}} \) to \( P \). If the error, the
root-mean-square of the residual \( \rho_{\text{RMS}}(\Delta_{\text{id}}) \), is below a
certain threshold then descriptor is considered fit. Oth-
ewise the method recursively considers a higher dimen-
sion solution. In general, for a \( n \)-dimensional descrip-
tor, SIS selects the subspace \( S_{\text{id}} \) with response \( \Delta_{(n, 1)\text{id}} \). Then SO extracts the best \( n \)D descriptor, with response
\( P \), from the union of all the previously selected subspaces
\( S_{\text{id}} \) \( \cup S_{(n-1)\text{id}} \) \( \cup \cdots \cup S_{1\text{id}} \). Candès and Romberg [51] have
shown that to identify the best \( n \)-dimensional descriptor
with “overwhelming probability” the size of the response
in the case the number of materials observations \( P \)
needs to satisfy the relationship \#\( P \geq k \cdot n \cdot \log(\#\Phi) \),
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 \#\( P \) is fixed [31]; then, when
\#\( \Phi \) increases, the maximum \( n \) decreases in order to sat-
ify the relationship [51]. In practice, features spaces of
growing sizes (\( \Phi_0, \Phi_1, \cdots \)) and different \( n \) are tested until

FIG. 2. Benchmark of algorithms. (a) Training error: RMSE versus descriptor dimension for different SOs operating on
the smallest \( \Phi_1 \). (b) Training error: RMSE versus subspace size in the SIS step to find a 3D descriptor by OMP or SISSO with
the same large features space \( \Phi_1 \) (see Supplemental Material at [URL will be inserted by publisher] for a similar picture for a
2D descriptor). (c) Training error: RMSE by SISSO(\( \ell_0 \)) with \( \Phi_2 \) and \( \Phi_1 \) compared with previous work [7] (features space size
\( \sim 4500 \)) and with the Eureqa software [40] (evaluated functions \( 10^{12} \), larger than \#\( \Phi_1 \)).
a model with required accuracy \(\rho_{\text{DMS}}(\Delta_{ab}) < \text{threshold}\) is obtained.

SISSO has advantages over MP [44] and OMP [45]. MP searches a linear model reproducing \(P\) by adding dimensionality to a descriptor while preserving selected features and corresponding coefficients. OMP improves MP by re-optimizing the coefficients every time a new component is introduced, \(n \rightarrow n + 1\), 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 \((#S_0 = 1, \forall i)\). Still, it differs from iterative SIS [49] which reduces to simple MP when all \(#S_0 = 1\).

**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 \(ab\) \textit{initio} enthalpy difference, RS and ZB for 82 materials and the 23 primary features related to material compositions forming \(\Phi_0\) (see Supplemental Material at [URL will be inserted by publisher] 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, \(H^{(m)}\), and Eq. (1), the features spaces \(\Phi_1\) (small, \#\(\Phi_1 = 556\)), \(\Phi_2\) (large, \#\(\Phi_2 \sim 10^5\)), and \(\Phi_3\) (huge, \#\(\Phi_3 \sim 10^{11}\)) are constructed.

**Figure 2(a).** The training errors \(\rho_{\text{DMS}}\) of different SO: LASSO, \(\text{LASSO} + \ell_0\), OMP, and \(\ell_0\) are compared while operating on the small features space \(\Phi_1\). LASSO suffers because of the correlations existing inside \(\Phi_1\) (see Figure S1 in the Supplemental Material at [URL will be inserted by publisher] for a figure showing the correlation between features): \(\text{LASSO} + \ell_0\) and OMP both surpass LASSO; \(\ell_0\) is the reference: it gives the exact global minimum solution for descriptors of any dimension. However, even with \(\ell_0\) the error is still too large for many thermodynamical predictions — \(\rho_{\text{DMS}}(\Delta_{ab}) \gtrsim 40\) meV/atom — and this is due to the too-small size of \(\Phi_1\).

**Figure 2(b).** For the larger \(\Phi_2\), SIS combined with \(\text{LASSO} + \ell_0\) as SO — SISSO(\(\text{LASSO} + \ell_0\)) —, SISSO(\(\ell_0\)), and OMP are compared for generating a 3D descriptor: SISSO(\(\ell_0\)) is the only approach improving consistently with subspace size \(\# \cup S_0\) and it always surpasses OMP when each \(#S_0 \gg 1\); SISSO(\(\text{LASSO} + \ell_0\)) does not improve over OMP because of the failure of LASSO in dealing with correlated features [42]. Obviously, the larger the features space and 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).** The errors for 1- to 5-dimensional descriptors are calculated by SISSO(\(\ell_0\)) while operating in the large \(\Phi_2\) and huge \(\Phi_3\) spaces. For \(n = 1\), SIS reduces to the best 1D descriptor, so no \(\ell_0\) is needed. For \(n = 2, 3, 4, 5\) the size of the SIS subspace is chosen to follow the previously mentioned relationship [51] applied to the subspace \(\#S \sim \exp(\#P/\#P]\). With \(\#P = 82\) and \(k = 3, 125\), the total size of all the selected subspaces is \(# \cup S_0 = 5 \cdot 10^5, 6 \cdot 10^5, 7 \cdot 10^5, 2 \cdot 10^2\) for \(n = 2, 3, 4, 5\), respectively. For all these sizes, the application of \(\ell_0\) regularization as SO involves \(10^{10} \sim 10^{11}\) independent least-square-regression evaluations. 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 \(\Phi_3\) are systematically better than those coming from \(\Phi_2\), thanks to the higher complexity (see Supplemental Material at [URL will be inserted by publisher] for the functional forms of the descriptors). SISSO(\(\ell_0\)) with \(\Phi_3\) is systematically better than the previous work by Ghiringhelli et al. [7, 31], due to the allowed larger features spaces. Note that when SISSO(\(\ell_0\)) 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 Figure 2(c). Performance is also compared with the commercial software Eureqa [40] by using the same operator set and primary features \(\Phi_0\), and \(10^{12}\) evaluated functions, a number comparable to \#\(\Phi_3\). SISSO(\(\ell_0\)) in \(\Phi_1\) with \(n \geq 2\) and SISSO(\(\ell_0\)) in \(\Phi_2\) with \(n \geq 3\) have both lower training error than Eureqa.

**Figure 3(a).** Training errors were illustrated in Figure 2(a-c), 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. 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 [52, 53], this is performed via \textit{cross validation} (CV), a class of techniques that, by splitting the dataset 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 [54, 55]. Specifically, in a CS-based iterative technique like 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. [54] with leave-10%-out (LTO) CV (the dataset is split in 40 training set containing 90% randomly selected data points and a test set with the remaining 10%) and leave-one-out (LOO) CV (one data points constitutes the test set, and the procedure is iterated \#\(P\) times). The model
The error is two for both the CV and for subspace of unit size (for which SISSO subspace sizes reported in the description of Figure 2(c), and the training set. The number of iterations of the features space construction — determined by set of primary features, operators set, and this is determined by the features space — in turns determines the maximum dimensionality of that particular model. This is 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 \( \Phi_3 \) with the subspace sizes reported in the description of Figure 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(\( \ell_0 \)). 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 (2~3), 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(\( \ell_0 \))-LTOCV results with features space \( \Phi_3 \) (RMSE shown in (a)). The 1% and 99% (extrema of the “error bar”), the 25% and 75% (lower and upper limits of the rectangle) and the median (intermediate horizontal line) percentiles are marked. The maximum absolute errors are also indicated by crosses. The worsening of the RMSE beyond 2D is mainly determined by an increase in the largest errors (the 99%-percentile), while most of the errors remain small (median/lower percentiles \( \sim \) constant).

LOOCV is also used to inspected how often the same descriptor is selected. The test operates in \( \#\Phi_3 \) with SISSO(\( \ell_0 \)). 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 \( \Phi_3 \) is of the order \( 10^{11} \) features and there 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 over-fitting (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 like 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 2D) 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:

\[
\hat{c} = \arg \min_c \left( \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} O_{ij} + \lambda \|c\|_0 \right),
\]

where \( O_{ij} \) is the number of data in the overlap-region between the \( i \)– and \( j \)–domain, \( c \) is a sparse vector (0/1 elements) so that a feature \( k \) is selected/deselected when \( c_k = 1(0) \), and \( \lambda \) is a parameter controlling the number of nonzero elements in \( c \). Of all the possible solutions of Eq. (2) having the same dimension and overlap, we
chose the one with minimum n-dimensional overlap volume [56]:

\[
\Omega = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \frac{\Omega_{ij}}{\min(\Omega_{i}, \Omega_{j})},
\]

(3)

where \( \Omega_{i} \), \( \Omega_{j} \), and \( \Omega_{ij} \) are the n-dimensional volumes of the \( i \)-, \( j \)-, and overlap \( ij \)-domains. Finally, the SIS correlation “property+feature” is defined as \( \left( \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \Omega_{ij} + 1 \right)^{-1} \): high correlation ⇔ 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 Equations (2-3). All essential atomic and structural parameters are included as primary features in \( \Phi_0 \). They originate from the WebElements [57] (atomic) and SpringerMaterials [58] (structural) databases (see Supplemental Material at [URL will be inserted by publisher] for a list of the features considered in this study). Amongst them are the Pauling electronegativity \( \chi \), ionization energy \( IE \), covalent radius \( r_{\text{cov}} \), electron affinity, valence (number of valence electrons for \( A \) and (8-valence) for \( B \)), coordination number, interatomic distance between \( A \) and \( B \) in crystal, atomic composition \( x_A \), and a “packing parameter”, 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 \( \mathbf{H}^{(m)} \) and Eq. (1) are then used to generate \( \Phi_3 \) (~ 10^8 elements).

Note that SISSO finds its optimal descriptor based on combinations of the input physical quantities (features): non-optimal outcomes indicate that the target property depends on features not yet-considered in \( \Phi_0 \). As such, to avoid “garbage in, garbage out”, SISSO requires physical intuition in the choice of features to add: conveniently, important and non-important features will be automatically promoted or neglected. Here, since metallicity also depends on “interstitial charge”, the inclusion of a packing parameter 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 parameter 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 [59]. The experimental binary data set, extracted from the SpringerMaterials database [58] 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., elemental solids, such as carbon diamond, are not tackled); iii. non-layered structure and without 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 299 binaries in 15 prototypes (\( \text{NaCl}, \text{CsCl}, \text{ZnS}, \text{CaF}_2, \text{Cr}_3\text{Si}, \text{SiC}, \text{TiO}_2, \text{ZnO}, \text{FeAs}, \text{NiAs}, \text{Al}_2\text{O}_3, \text{La}_2\text{O}_3, \text{Th}_3\text{P}_4, \text{ReO}_3, \text{ThH}_2 \))
are then used (see Supplemental Material at [URL will be inserted by publisher] for a list of the training materials). Details on the feature-space construction and model identification are given in Appendix. Out of \( \Phi_3 \), SISSO\((\ell_0) \) identifies a 2-dimensional descriptor with a training accuracy of \( \sim 99.0\% \). The convex domains, indicating metallic and non-metallic materials, are shown in Figure 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 non overlapping.

**Figure 4(a)** shows the three incorrectly classified systems (blue empty squares). YP (NaCl prototype) might have slightly erroneous position in the figure: the covalent radius \( r_{\text{co}} \) (Y) (controlling the packing parameter) suffers of large intrinsic errors (see Figure 2 of Ref. [65]) and therefore the compound position might be misrepresented. La\(_3\)Te\(_4\) and Th\(_4\)As\(_4\) (Th\(_3\)P\(_4\) prototype) are different. In this case, SISSO indicates that the primary feature 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 Figure 4(a) — indicates that the higher the packing factor \( \sum V_{\text{atom}}/V_{\text{cell}} \), 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 factor, allowing for predictions (see below) and ii) reveal the functional form packing factor \( \rightarrow \) metallicity: It is not trivial that the descriptor is linear with the inverse packing factor. 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 amongst the two extreme electronegativities, (Li has the lowest, F the highest), pushes the compound toward the rightmost corner of the non-metals domain. On the other side, AuIn\(_2\) is the compound furthest from the non-metals region: Au has the highest \( \chi \) amongst transition metals and In has one of the smallest \( \chi \) of the considered \( p \)-elements. Available experimental band gaps were also extracted (see Supplemental Material at [URL will be inserted by publisher] 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 metalization by compression.** Although pressure is neither included in the features space nor in the training data, its effect can be tested by reducing \( V_{\text{cell}} \). Amongst the training data, we have 3 systems experiencing pressure-induced insulator\( \rightarrow \)metal transition: HgTe, GaAs and CdTe. HgTe, CdTe and GaAs go from insulating zinc blende to metallic rock salt (or an orthorhombic ol4 phase for GaAs) at \( \sim 9, 4, \) 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 be extended to systems which have not yet been fully characterized to predict potential insulator\( \rightarrow \)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 \( V_{\text{cell}} \) 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\( \rightarrow \)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
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 \(\rightarrow\) property relationship in terms of an analytical equation. It does not need to be exact — a simple, analytical descriptor \(\rightarrow\) 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\(\leftrightarrow\)metal transitions, which allows 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. This project has received funding from the European Unions Horizon 2020 research and innovation program (#765680: The NOMAD Laboratory — an European Center of Excellence and #740233: TEC1p), and the Berlin Big-Data Center (BBDC, #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.

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, \(\Phi_0\), via algebraic operations, \(\vec{H} \equiv \{I, +, -, \times, /, \exp, \log, | - |, \sqrt{\cdot}, ^{-1, 2, 3}\}\). Primary features usually comprise of 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: (1) first ionization energy, \(IE_A\) (\(A\)-species) and \(IE_B\) (\(B\)-species); (2) electron affinity, \(EA_A\) and \(EA_B\); (3) atom covalent radius, \(r_{\text{cov}}A\) and \(r_{\text{cov}}B\); (4) Pauling electronegativity, \(\chi_A\) and \(\chi_B\); (5) valence, \(v_A\) (#valence electrons) and \(v_B\) (#valence electrons); (6) coordination number, \(CN_A\).
(7) interatomic distance between $A$ and $B$ in crystal, $d_{AB};$ (8) atomic composition $x_A$ (or $x_B = 1 - x_A$); and (9) the ratio of the cell volume to the total atom volume in the unit cell of the crystal, $V_{cell} / \sum V_{atom}$ ($V_{atom} = 4\pi r_{cov}^3 / 3$).

It is critical to limit the redundant and unnecessary primary features in $\Phi_0$ to enhance computational performance (the size of features space $\Phi_n$ increases very fast with $\#\Phi_0$) and to increase SIS success rate: the higher $\#subspace/\#\Phi$ the higher the probability that SIS subspaces contain the best models. Starting from an empty $\Phi_0$, few primary features are added. SISSO is then applied to identify the best model, with $H$ 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 $\Phi_0$ to check for improvements. Primary features preserved in $\Phi_0$ 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, $\Phi_0$ 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, $\Phi_0$, 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: (1) $\Phi_3$ is constructed; (2) the best descriptor is identified from $\Phi_3$ 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 non metal when the large interatomic distance is decreased with the radius of species $A$.

Next, the number of prototypes is increased to 5, for a total of 217 materials. However, with the previous $\Phi_0$ and calculation-settings, SISSO fails to identify a descriptor having perfect classification (there are 7 points in the overlap-region between the metal and non metal domains). The non-optimal outcome indicates that the classification depends on primary features not yet considered. First, $\Phi_0$ is slimmed by reducing its size to $7 - E_{A}, E_{B}, v_{A}$, and $v_{B}$ are removed — without affecting the quality of the predictions (8 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 $\Phi_3$ 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 5 prototypes with large $d_{AB}$ and small $CN_B$ tend to be non metals.

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 is 2D have 99.6% classification (only point, YP-compound in NaCl-prototype, is misclassified). Although the classification is excellent, the descriptor is complicate. Searching for a simplification, new primary features of atomic composition $x_A$, $x_B$, and $V_{cell} / \sum V_{atom}$ 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 simple 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 7 primary features in $\Phi_0$. With a constructed $\Phi_3$ of size $10^6$, SISSO identifies the best 2D descriptor with a classification accuracy of 99.0% (three misclassified compounds: YP-compound in NaCl-prototype, Th$_3$As$_4$ and La$_3$Te$_4$ in Th$_3$P$_4$-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.
### Table I. Dependence of the metal-insulator classification descriptors on the prototypes of training binary materials.

<table>
<thead>
<tr>
<th>prototypes</th>
<th>#materials</th>
<th>primary features</th>
<th>descriptor</th>
<th>class.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NaCl, CsCl, ZnS, CaF$_2$, Cr$_3$Si</td>
<td>132</td>
<td>$IE_A$, $IE_B$, $\chi_A$, $\chi_B$, $CN_A$, $CN_B$</td>
<td>$A := \frac{IE_B d_{AB} - IE_A d_{AB}}{\exp(</td>
<td>\chi_A</td>
</tr>
<tr>
<td>NaCl, CsCl, ZnS, CaF$_2$, Cr$_3$Si, SiC, TiO$_2$, ZnO, FeAs, NiAs</td>
<td>217</td>
<td>$IE_A$, $IE_B$, $\chi_A$, $\chi_B$, $CN_A$, $CN_B$</td>
<td>$A := \frac{IE_B d_{AB} - IE_A d_{AB}}{\exp(</td>
<td>\chi_A</td>
</tr>
<tr>
<td>NaCl, CsCl, ZnS, CaF$_2$, Cr$_3$Si, SiC, TiO$_2$, ZnO, FeAs, NiAs</td>
<td>260</td>
<td>$IE_A$, $IE_B$, $\chi_A$, $\chi_B$, $CN_A$, $CN_B$</td>
<td>$A := \frac{IE_B d_{AB} - IE_A d_{AB}}{\exp(</td>
<td>\chi_A</td>
</tr>
<tr>
<td>NaCl, CsCl, ZnS, CaF$_2$, Cr$_3$Si, SiC, TiO$_2$, ZnO, FeAs, NiAs</td>
<td>260</td>
<td>$IE_A$, $IE_B$, $\chi_A$, $\chi_B$, $CN_A$, $CN_B$</td>
<td>$A := \frac{IE_B d_{AB} - IE_A d_{AB}}{\exp(</td>
<td>\chi_A</td>
</tr>
<tr>
<td>NaCl, CsCl, ZnS, CaF$_2$, Cr$_3$Si, SiC, TiO$_2$, ZnO, FeAs, NiAs, Al$_2$O$_3$, La$_2$O$_3$, Th$_3$P$_4$, ReO$_3$, ThH$_2$</td>
<td>299</td>
<td>$IE_A$, $IE_B$, $\chi_A$, $\chi_B$, $CN_A$, $CN_B$</td>
<td>$A := \frac{IE_B d_{AB} - IE_A d_{AB}}{\exp(</td>
<td>\chi_A</td>
</tr>
</tbody>
</table>

* One entry misclassified: YP-compound in NaCl-prototype.

b Three entry misclassified: YP-compound in NaCl-prototype; Th$_3$As$_4$ and La$_3$Te$_4$-compounds in Th$_3$P$_4$-prototype.

---

* ghiringhelli@fhi-berlin.mpg.de


41. The $\ell_0$-norm of a vector is the number of its non-zero components.
43. The $\ell_1$-norm is the sum of the absolute values of the components of a vector.
57. https://www.webelements.com
58. https://materials.springer.com
62. D. G. Pettifor, *A chemical scale for crystal-structure


