Big Data of Materials Science: Critical Role of the Descriptor

Luca M. Ghiringhelli,[1,*] Jan Vybiral,[2] Sergey V. Levchenko,[1] Claudia Draxl,[3] and Matthias Scheffler[1]
[1]*Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin-Dahlem, Germany*
[2]*Department of Mathematical Analysis, Charles University, 18675 Prague, Czech Republic*
[3]*Humboldt-Universität zu Berlin, Institut für Physik and IRIS Adlershof, 12489 Berlin, Germany*
(Received 14 April 2014; revised manuscript received 20 October 2014; published 10 March 2015)

Statistical learning of materials properties or functions so far starts with a largely silent, nonchallenged step: the choice of the set of descriptive parameters (termed *descriptor*). However, when the scientific connection between the descriptor and the actuating mechanisms is unclear, the causality of the learned descriptor-property relation is uncertain. Thus, a trustful prediction of new promising materials, identification of anomalies, and scientific advancement are doubtful. We analyze this issue and define requirements for a suitable descriptor. For a classic example, the energy difference of zinc blende or wurtzite and rocksalt semiconductors, we demonstrate how a meaningful descriptor can be found systematically.

DOI: 10.1103/PhysRevLett.114.105503 PACS numbers: 61.50.-f, 02.60.Ed, 71.15.Mb, 89.20.Ff

Using first-principles electronic-structure codes, a large number of known and hypothetical materials have been studied in recent years, and currently, the amount of calculated data increase exponentially with time. The targets of these studies are, for example, the stable structure of solids or the efficiency of potential photovoltaic, thermoelectric, battery, or catalytic materials. Utilizing such data like a reference book (query and read out what was stored) is an avail. Finding the actuating mechanisms of a certain property or function and describing it in terms of a set of physically meaningful parameters (henceforth termed *descriptor*) is the desired science. A most impressive and influential example for the importance and impact of finding a descriptor is the periodic table of elements, where the elements are categorized (described) by two numbers, the table row and column. The initial version had several "white spots," i.e., elements that had not been found at that time, but the chemical properties of these elements were roughly known already from their position in the table. Interestingly, the physical meaning of this two-dimensional descriptor became clear only later. Below we will use an example from materials science to discuss and demonstrate the challenge of finding meaningful descriptors: the prediction of the crystal structure of binary compound semiconductors, which are known to crystallize in zinc blende (ZB), wurtzite (WZ), or rocksalt (RS) structures. The structures and energies of ZB and WZ are very close and for the sake of clarity we will not distinguish them here. The energy difference between ZB and RS is larger, though still very small, namely just 0.001% or less of the total energy of a single atom. Thus, high accuracy is required to predict this difference. This refers to both steps, the explicit calculations and the identification process of the appropriate descriptor (see below). The latter includes the representation of the descriptor-property relation.

For brevity, we only write "property," characterized by a number P_i in the following, with i denoting the actually calculated material, but we mean the materials function(s) as well. In general, the property will be characterized by a string of numbers, but here we like to keep the discussion simple. Analogously, the multidimensional descriptor is denoted as a vector $\boldsymbol{d}_i$, with dimension Ω. The generalization of the discrete data set $\{P_i, \boldsymbol{d}_i\}$ to a continuous function $P(\boldsymbol{d})$ has been traditionally achieved in terms of physical models, or mathematical fits. Scientific understanding of the descriptor $\boldsymbol{d}$ and of the relationship between $\boldsymbol{d}$ and P is needed for deciding with confidence what new materials should be studied next as the most promising novel candidates and for identifying interesting anomalies.

In 1970, Phillips and van Vechten (Ph-vV) [1,2] analyzed the classification challenge of ZB or WZ versus RS structures. They came up with a two-dimensional descriptor, i.e., two numbers that are related to the experimental dielectric constant and nearest-neighbor distance in the crystal [1,2]. Figure 1 shows their conclusion. Clearly, in this representation ZB or WZ and RS structures separate nicely: Materials in the upper left part crystallize in the RS structure, those in the lower right part are ZB or WZ. Thus, based on the ingenious descriptor $\boldsymbol{d} = (E_h, C)$ one can predict the structure of unknown compounds without the need of performing explicit experiments or calculations. Several authors have taken up the Ph-vV challenge and identified alternative descriptors [3–5]. We will come back to this below.

Published by the American Physical Society under the terms of the Creative Commons Attribution 3.0 License. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

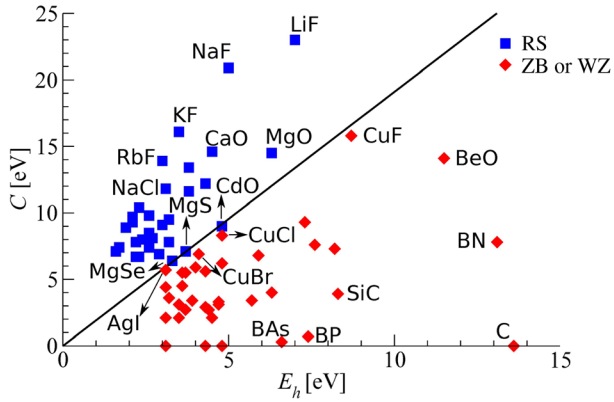

FIG. 1 (color online). Experimental ground-state structures of 68 octet binary compounds, arranged according to the two-dimensional descriptor introduced by Phillips [2] and van Vechten [1]. Because of visibility reasons, only 10 materials are labeled for each structure. See the Supplemental Material for more details [6].

In recent years, the demand for finding the function $P(\boldsymbol{d})$ employed statistical learning theory, which is the focus of this Letter. This strategy has been put forward by several authors in materials science [14–18], as well as in bio- and cheminformatics (see, e.g., Ref. [19] and references therein). Most of these works employed the kernel ridge regression (KRR) approach. For a Gaussian kernel, the fitted property is expressed as a weighted sum of Gaussians: $P(\boldsymbol{d}) = \sum_{i=1}^{N} c_i \exp(-\|\boldsymbol{d}_i - \boldsymbol{d}\|_2^2/2\sigma^2)$, where N is the number of training data points. The coefficients c_i are determined by minimizing $\sum_{i=1}^{N}[P(\boldsymbol{d}_i) - P_i]^2 + \lambda \sum_{i,j=1}^{N,N} c_i c_j \exp(-\|\boldsymbol{d}_i - \boldsymbol{d}_j\|_2^2/2\sigma^2)$, where $\|\boldsymbol{d}_i - \boldsymbol{d}_j\|_2^2 = \sum_{\alpha=1}^{\Omega}(d_{i,\alpha} - d_{j,\alpha})^2$ is the squared ℓ_2 norm of the difference of descriptors of different materials, i.e., their "similarity." The regularization parameter λ and σ are chosen separately, usually with the help of leave-some-out cross validation [20], i.e., by leaving some of the calculated materials out in the training process and testing how the predicted values for them agree with the actually calculated ones.

In essentially all previous materials studies the possibly multidimensional descriptor was introduced *ad hoc*, i.e., without demonstrating that it was the best (in some sense) within a certain broad class (see Ref. [17] for an impressive exception). In this Letter, we describe an approach for finding descriptors for the accurate prediction of a given property of a class of materials, where we restrict ourselves to *ab initio* data.

For the example shown in Fig. 1, statistical learning is unnecessary, because one can determine the classification by visual inspection of the 2D plot. In this Letter, we add the quantitative energy difference between ZB and RS to the original Ph-vV challenge. In general, the descriptor will be higher dimensional. Additionally, the scientific question will be typically more complex than the structural classification. We define the conditions that a proper descriptor must fulfill in order to be suitable for causal "learning" of materials properties, and we demonstrate how the descriptor with the lowest possible dimensionality can be identified. Specifically, we will use the least absolute shrinkage and selection operator (LASSO) for feature selection [21].

All data shown in this study have been obtained with density-functional theory using the local-density approximation (LDA) for the exchange-correlation interaction. Calculations were performed using the all-electron full-potential code FHI-aims [7] with highly accurate basis sets, **k** meshes, and integration grids. For the task discussed in this Letter, the quality of the exchange-correlation functional is irrelevant. Nevertheless, we stress that the LDA provides a good description of the studied materials. In particular, we have computed equilibrium lattice constants and total energies for all three considered lattices (ZB, WZ, RS) of a set of 82 binary materials. The full list of these materials and all calculated properties can be found in the Supplemental Material [6], and all input and output files can be downloaded from the NoMaD repository [22]. Furthermore, we calculated several properties of the isolated neutral atoms and dimer molecules (see below).

Let us start with a simple example that demonstrates the necessity of *validation* in the search for descriptors. The nuclear numbers of a binary semiconductor AB, Z_A and Z_B, unambiguously identify the lowest energy structure: They define the many-body Hamiltonian, and its total energies for the different structures give the stable and metastable structures. Figure 2 (top) displays the total-energy differences of the ZB and RS structures as function of Z_A and Z_B. When using the KRR approach, the data can be fitted well (see Supplemental Material [6]) when the whole set is used for learning. However, the predictive power of KRR based on the descriptor $\boldsymbol{d} = (Z_A, Z_B)$ is bad, as tested by leave-some-out cross validation (see Table I and Supplemental Material [6]). Obviously, the relation between $\boldsymbol{d} = (Z_A, Z_B)$ and the property that we need to learn is by far too complex.

For a descriptor, we consider the following properties to be important, if not necessary.

(a) A descriptor $\boldsymbol{d}_i$ uniquely characterizes the material i as well as property-relevant elementary processes.
(b) (b) Materials that are very different (similar) should be characterized by very different (similar) descriptor values.
(c) The determination of the descriptor must not involve calculations as intensive as those needed for the evaluation of the property to be predicted.
(d) The dimension Ω of the descriptor should be as low as possible (for a certain accuracy request).

Although the Ph-vV descriptor $\boldsymbol{d} = (E_h, C)$ fulfills conditions (a), (b), and (d), it falls short on condition (c). In contrast, $\boldsymbol{d} = (Z_A, Z_B)$ fails for conditions (b) and (d).

In order to identify a good descriptor, we start with a large number M of candidates (the "feature space") for the

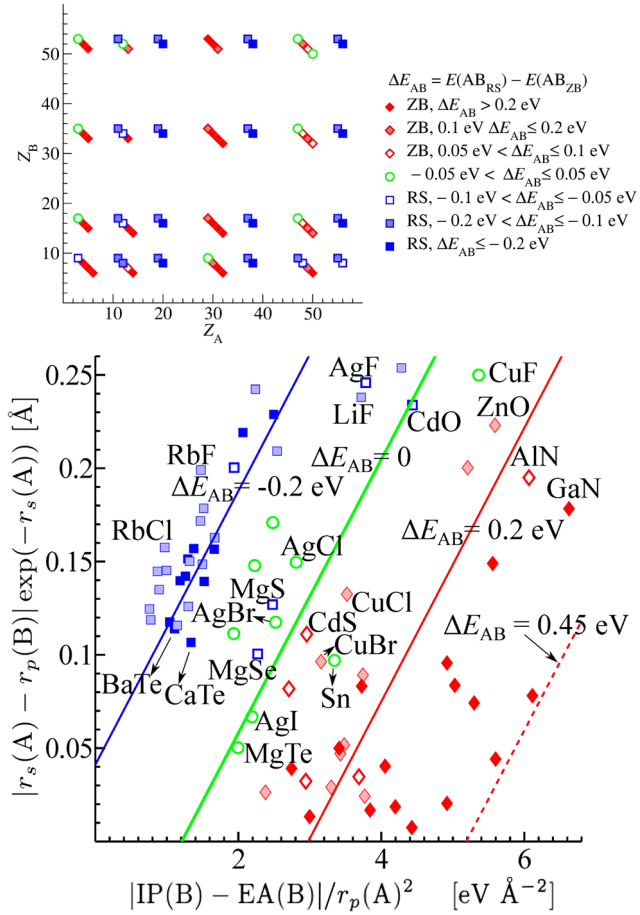

FIG. 2 (color online). Calculated energy differences between RS and ZB structures of the 82 octet binary AB materials, arranged by using the nuclear numbers (Z_A, Z_B) as descriptor (top) and according to our optimal two-dimensional descriptor (bottom). In the bottom panel, seven ZB materials with predicted $\Delta E_{AB} > 0.5$ eV are outside the shown window (see Supplemental Material [6]).

components of $\boldsymbol{d}$. We then look for the Ω-dimensional ($\Omega = 1, 2, \ldots$) descriptor $\boldsymbol{d}$ that gives the best linear fit of $P(\boldsymbol{d})$: $P(\boldsymbol{d}) = \boldsymbol{d}\boldsymbol{c}$, where $\boldsymbol{c}$ is the Ω-dimensional vector of coefficients. It is determined by minimizing the *loss* function $\|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|_2^2$, where $\boldsymbol{D}$ is a matrix with each of the N rows being the descriptor $\boldsymbol{d}_i$ for each training data point, and $\boldsymbol{P}$ is the vector of the training values P_i. We emphasize that the choice of a linear fitting function for $P(\boldsymbol{d})$ is not restrictive since, as we will show below, nonlinearities are included in a controlled way in the formation of the candidate components of $\boldsymbol{d}$. The function $P(\boldsymbol{d})$ is then determined by only Ω parameters.

The task is now to find, among all the Ω-tuples of candidate features, the Ω-tuple that yields the smallest $\|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|_2^2$. Unfortunately, a computational solution for such a problem is infeasible (NP-hard) [23]. LASSO [21] provides sparse (i.e., low-dimensional) solutions by recasting the NP-hard problem into a convex minimization problem

$$\underset{\boldsymbol{c} \in \mathbb{R}^M}{\operatorname{argmin}} \|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|_2^2 + \lambda \|\boldsymbol{c}\|_1, \tag{1}$$

where the use of the ℓ_1-norm ($\|\boldsymbol{c}\|_1 = \sum_{\alpha=1}^{M} |c_\alpha|$) is crucial. The larger we choose $\lambda > 0$, the smaller the ℓ_1-norm of the solution of Eq. (1) and vice versa. There is actually a smallest $\bar{\lambda} > 0$, such that the solution of Eq. (1) is zero. If $\lambda < \bar{\lambda}$, one or more coordinates of $\boldsymbol{c}$ become nonzero.

We note that the so-called "feature selection" is a widespread set of techniques that are used in statistical analysis in different fields [24], and LASSO is one of them. LASSO was successfully demonstrated in Ref. [17], for identifying the low-dimensional representation of the formation energy of an alloy, within the cluster expansion of the Hamiltonian. Obviously, when a well-identified basis set, such as the cluster expansion, is not available for the property to be modeled, the feature space must be constructed differently. In this Letter, we start from scientific insight, i.e., defining physically motivated primary features that form the basis for a large feature space. We then search for a low-dimensional descriptor that minimizes the RMSE, given by $\sqrt{(1/N)\|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|_2^2}$, for our $N = 82$ binary compounds. The property P that we aim to predict is the difference in the LDA energies between RS and ZB for the given atom pair AB, ΔE_{AB}. The order of the two atoms is such that element A has the smallest Mulliken electronegativity: $\mathrm{EN} = -(\mathrm{IP} + \mathrm{EA})/2$. IP and EA are atomic ionization potential and electron affinity.

For constructing the feature space, i.e., the candidate components of the descriptor, and then selecting the most

TABLE I. Root-mean-square error (RMSE) and maximum absolute error (MaxAE) in eV for the least-squares fit of all data (first two lines) and for the test set in a leave-10%-out cross validation (L-10%-OCV), averaged over 150 random selections of the training set (last two lines). The errors for (Z_A, Z_B) and (r_σ, r_π) [3] are for a KRR fit at hyperparameters (λ, σ) that minimize the RMSE for the L-10%-OCV (see Supplemental Material [6]). The errors for the $\Omega = 1, 2, 3, 5$ (noted as 1D, 2D, 3D, 5D) descriptors are for the LASSO fit. In the L-10%-OCV for the latter descriptors, the overall LASSO-based selection procedure of the descriptor (see text) is repeated at each random selection of the test set.

Descriptor	Z_A, Z_B	r_σ, r_π	1D	2D	3D	5D
RMSE	2×10^{-4}	0.07	0.14	0.10	0.08	0.06
MaxAE	8×10^{-4}	0.25	0.32	0.32	0.24	0.20
RMSE, CV	0.19	0.09	0.14	0.11	0.08	0.07
MaxAE, CV	0.43	0.17	0.27	0.18	0.16	0.12

relevant of them, we implemented an *iterative* approach. At first we defined primary features. These are (for atom A): $\mathrm{IP}(A)$ and $\mathrm{EA}(A)$, $\mathrm{H}(A)$, and $\mathrm{L}(A)$, i.e., the energies of the highest-occupied and lowest-unoccupied Kohn-Sham levels, as well as $r_s(A)$, $r_p(A)$, and $r_d(A)$, i.e., the radii where the radial probability density of the valence s, p, and d orbitals are maximal. The same was done for atom B. In addition to these atomic data, we offered information on AA, BB, and AB dimers, namely, their equilibrium distance, binding energy, and HOMO-LUMO Kohn-Sham gap. Altogether, these are 23 primary features.

Next, we define rules for linear and nonlinear combinations of the primary features. One can easily generate a huge number of candidate descriptors, e.g., all thinkable but not violating basic physical rules. In the present study, we used about 10 000 candidates grouped in subsets that are used in the different iterations (see Supplemental Material [6]). A more detailed discussion will be given in Ref. [25]. In the language of KRR, this approach designs a kernel, done here by using physical insight. Not surprisingly, LASSO (and actually any other method) has difficulties in selecting among highly correlated features [26]. In these cases, it is not ensured that the first Ω selected features form the best Ω-dimensional descriptor. Although checking correlations between pairs is straightforward and computationally reasonably inexpensive, discovering correlations between triples and more-tuples is computationally prohibitive. Therefore, we adopted a different strategy: The first 25–30 features proposed by LASSO were selected and a batch of least-squares fits performed [when the descriptor is fixed, i.e., the nonzero components of $\boldsymbol{c}$ are fixed, Eq. (1) reduces to a linear, least-squares, fit], taking in turn as $\boldsymbol{D}$ each single feature, each pair, etc. We confirmed that this strategy always finds the best descriptor by running the mentioned extensive search for several different subsets of hundreds of features.

Our procedure identifies as best (i.e., lowest RMSE) 1D, 2D, and 3D descriptors, the first, the first two, and all three of the following features:

$$\frac{\mathrm{IP}(B)-\mathrm{EA}(B)}{r_p(A)^2}, \qquad \frac{|r_s(A)-r_p(B)|}{\exp[r_s(A)]}, \qquad \frac{|r_p(B)-r_s(B)|}{\exp[r_d(A)]}. \tag{2}$$

Note that, mathematically, the descriptor does not necessarily need to build up incrementally in this way; e.g., the 1D one may not be a component of the 2D one. However, in our study, it does. The RMSE and MaxAE for the 1D, 2D, 3D descriptors are reported in Table I. By adding further dimensions to the descriptor, the decrease of the RMSE becomes smaller and smaller.

We tested the robustness of our descriptor by performing a leave-10%-out cross validation (L-10%-OCV). Thereby, the overall procedure of selecting the descriptor is repeated from scratch on a learning set obtained by randomly selecting 90% of the materials. The resulting fitted linear model is applied to the excluded materials and the prediction errors on this set, averaged over 150 random selections, are recorded. The results are shown in Table I. Not only the RMSE, but also the selection of the descriptor, proved very stable. In fact, the 2D descriptor was selected 100% of the times, while the 1D descriptor was the same in 90% of the cases.

The errors for the 2D descriptor introduced by Zunger (Refs. [3,5] and Supplemental Material [6]), based on sums and absolute differences of r_s's and r_p's, are also reported in Table I. The cross-validation error of the linear fit with our 2D descriptor is as small as the highly nonlinear KRR fit with Zunger's 2D descriptor. However, our descriptor bears the advantage that it was derived from a broad class of options by a well-defined procedure, providing a basis for a systematic improvement (with increasing Ω). Our LASSO-derived descriptor contains physically meaningful quantities, like the band gap of B in the numerator of the first component and the size mismatch between valence s and p orbitals (numerators of the second and third component). We note that the components of the descriptors are not symmetric with respect to exchange between A and B. Symmetric features were included in the feature space, but never emerged as prominent, and, for the selected descriptors, symmetrized versions were explicitly constructed, tested, and systematically found to perform worse. This reflects that the symmetry was explicitly broken in the construction of the test set, as the order AB in the compound is such that $\mathrm{EN}(A) < \mathrm{EN}(B)$. Furthermore, we find that d orbitals appear only in the third or higher dimension. In Fig. 2 (bottom) we show the calculated and predicted ΔE_{AB}, according to our best 2D descriptor. It is evident that our 2D descriptor fulfills all above noted conditions, where conditions (a), (c), and (d) are in fact ensured by construction.

In order to further test the robustness and the physical meaningfulness of the identified descriptor, we performed tests by perturbing the value of the property ΔE_{AB} by adding uniform noise in the interval $[-0.1, 0.1]$ eV. The 2D descriptor of Eq. (2) was identified 93% of the times, with an increase of the RMSE by 10% only. More details are reported in Ref. [25]. This test shows that the model allows for some uncertainty in the measured property. Larger noise terms, however, destroy the reliable identification of the descriptor (see Ref. [25]). This analysis implies that the descriptor identified by LASSO contains the important physically meaningful ingredients for the prediction of ΔE_{AB}, even though a physical model that justifies the $P(\boldsymbol{d})$ mapping is not transparent.

Finally, we comment on the causality of the learned descriptor-property relationship. The simplicity of our model is in sharp contrast with what is yielded by, for instance, KRR, where as many fit parameters as observed points are, in principle, necessary. As an indication of

having identified a causal (physically meaningful) descriptor for the property ΔE_{AB}, we use the stability of the selection of the descriptor upon both L-10%-OCV and perturbation of the values of the property, under the condition that the $P(\boldsymbol{d})$ dependence has a small number of fit parameters and a simple functional form [see Eq. (2) and Supplemental Material [6]).

Financial support from the Einstein Foundation Berlin is appreciated. J. V. acknowledges financial support by the ERC-CZ Project No. LL1203 of the Ministry of Education, Youth and Sports of the Czech Republic. We thank Krishna Rajan for bringing the relevance of the Ph-vV analysis to our attention and Judea Pearl for inspiring discussions on the concept of causality in the context of statistical inference.

*ghiringhelli@fhi-berlin.mpg.de

[1] J. A. van Vechten, Phys. Rev. **182**, 891 (1969).
[2] J. C. Phillips, Rev. Mod. Phys. **42**, 317 (1970).
[3] A. Zunger, Phys. Rev. B **22**, 5839 (1980).
[4] D. G. Pettifor, Solid State Commun. **51**, 31 (1984).
[5] Y. Saad, D. Gao, T. Ngo, S. Bobbitt, J. R. Chelikowsky, and W. Andreoni, Phys. Rev. B **85**, 104104 (2012).
[6] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.114.105503, which includes Refs. [3,5,7–13], for details on the iterative LASSO procedure, the performance of several descriptors with KRR and linear-least-square regression, details on the cross-validation strategy, details on the LDA calculations, list of calculated atomic and primary features for all considered materials.
[7] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, Comput. Phys. Commun. **180**, 2175 (2009).
[8] D. Donoho, IEEE Trans. Inf. Theory **52**, 1289 (2006).
[9] E. J. Candés, J. Romberg, and T. Tao, IEEE Trans. Inf. Theory **52**, 489 (2006).
[10] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing* (Springer, New York, 2013).
[11] D. M. Ceperley and B. J. Alder, Phys. Rev. Lett. **45**, 566 (1980).
[12] J. P. Perdew and Y. Wang, Phys. Rev. B **45**, 13244 (1992).
[13] E. van Lenthe, E. J. Baerends, and J. G. Snijders, J. Chem. Phys. **101**, 9783 (1994).
[14] C. C. Fischer, K. J. Tibbetts, D. Morgan, and G. Ceder, Nat. Mater. **5**, 641 (2006).
[15] K. Rajan, Annu. Rev. Mater. Res. **38**, 299 (2008).
[16] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, Sci. Rep. **3**, 2810 (2013).
[17] L. J. Nelson, G. L. W. Hart, F. Zhou, and V. Ozoliņš, Phys. Rev. B **87**, 035125 (2013).
[18] T. Mueller, E. Johlin, and J. C. Grossman, Phys. Rev. B **89**, 115202 (2014).
[19] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).
[20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer, New York, 2009).
[21] R. Tibshirani, J. R. Stat. Soc. Ser. B **58**, 267 (1996).
[22] The NoMaD (Novel Materials Discovery) repository contains full input and output files of calculations in materials science, http://nomad-repository.eu.
[23] S. Arora and B. Barak, *Computational Complexity: A Modern Approach* (Cambridge University Press, Cambridge, England, 2009).
[24] I. Guyon and A. Elisseeff, J. Mach. Learn. Res. **3**, 1157 (2003).
[25] L. M. Ghiringhelli, J. Vybiral, S. Levchenko, C. Draxl, and M. Scheffler (to be published).
[26] Two columns of $\boldsymbol{D}$ are correlated if the absolute value of their Pearson's correlation index is (about) 1.