Big Data of Materials Science – Critical Role of the Descriptor

Luca M. Ghiringhelli1, Jan Vybiral2, Sergey V. Levchenko1, Claudia Draxl3, and Matthias Scheffler1
1 Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin-Dahlem, Germany
2 Technische Universität Berlin, Fachbereich Mathematik, Berlin, Germany
3 Humboldt-Universität zu Berlin, Institut für Physik and IRIS Adlershof, Berlin, Germany
(Dated: April 19, 2014)

Statistical learning of materials properties or functions so far starts with a largely silent, non-challenged step: the introduction of a descriptor. However, when the scientific connection between the descriptor and the actuating mechanisms is unclear, causality of the learned descriptor-property relation is uncertain. Thus, trustful prediction of new promising materials, identification of anomalies, and scientific advancement are doubtful. We analyze this issue and define requirements for a suited descriptor. For a classical example, the energy difference of zincblende/wurtzite and rocksalt semiconductors, we demonstrate how a meaningful descriptor can be found systematically.

PACS numbers: 02.60.Ed, 61.50.-f, 89.20.Ff, 31.15.E-

Using first-principles electronic-structure codes, a large number of known and hypothetical materials has been studied in recent years, and currently, the amount of calculated data increases exponentially with time. 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 (the causes) of a certain property or function is the desired science. A most impressive and influential example for the importance and impact of finding a physically meaningful descriptor is the development of the periodic table of the elements. The initial version had several “white spots”, i.e., elements that had not been found at that time. Below we will use another example 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 zincblende (ZB), wurtzite (WZ), or rocksalt (RS) structures. The structures and energies of ZB and WZ are very close and for the sake of simplicity 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 training (learning) 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 $\mathbf{d}_i$, with dimension $\Omega$. The generalization of the discrete data set \{${P_i, \mathbf{d}_i}$\} to a continuous function $P(\mathbf{d})$ has been traditionally achieved in terms of physical models, or mathematical fits. Scientific understanding of the descriptor $\mathbf{d}$ and of the $\mathbf{d} \rightarrow P$ relationship is needed for deciding with confidence what new materials should be studied next as most promising novel candidates and for identifying interesting anomalies.

In 1970 Phillips and van Vechten (Ph-vV) [1, 2] analyzed the classification challenge and came up with a two-dimensional (2D) 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/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/WZ. Thus, based on the ingenious descriptor $\mathbf{d} = (E_h, C)$ 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.

In recent years, the demand for finding the function

\[ P(\mathbf{d}) \]

relationship is needed for deciding with confidence what new materials should be studied next as most promising novel candidates and for identifying interesting anomalies.

FIG. 1: Ground-state structures of 68 octet binary compounds, arranged according to the two-dimensional descriptor introduced by Phillips and van Vechten [1, 2]. Because of visibility reasons only 10 materials are labeled for each structure. See the SI for more details.
$P(d)$ employed statistical learning theory, which is the focus of this paper. This strategy has been put forward by several authors in materials science [6-8], as well as in bio- and cheminformatics (see, e.g., Ref. [9] 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

$$
f(d) = \sum_{i=1}^{N} c_i \exp\left(-\|d - d_i\|^2/2\sigma^2\right),$$

where $N$ is the number of training data points. The coefficients $c_i$ are determined by minimizing $\sum_{i=1}^{N} (f(d_i) - P(d_i))^2 + \lambda \sum_{i,j=1}^{N} c_i c_j \exp\left(-\|d_i - d_j\|^2/2\sigma^2\right)$, where $\|d_i - d_j\|^2 = \sum_{\alpha=1}^{\Omega} (d_{i,\alpha} - d_{j,\alpha})^2$ is the squared $\ell_2$ norm of the difference of descriptors of different materials, i.e., their “similarity”. The regularization parameter $\lambda$ and $\sigma$ are chosen separately, usually with the help of leave-some-out cross validation [10], 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.

With very few exceptions [11, 12], in essentially all previous materials studies the possibly multidimensional descriptor was introduced ad hoc, i.e., without showing that it was the best (in some sense) within a certain broad class. In this Letter, we describe the challenge of finding suitable descriptors that classify the properties / materials systematically, restricting 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. However, in general the descriptor will be higher dimensional. Also the scientific question will be typically more complex than the structural classification. In this paper, we add the quantitative energy difference between ZB and RS ($\Delta E$) to the original Ph-vV challenge and describe a method that enables us to identify meaningful descriptors. We define the conditions 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* [13].

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 [14] with highly accurate basis sets, k-meshes, and integration grids. For the task discussed in this paper, 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 SI. 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 SI). However, the predictive power of KRR based on the descriptor $d = (Z_A, Z_B)$ is bad, as tested by leave-some-out cross validation (see Table I and SI). Obviously, the relation between $d = (Z_A, Z_B)$ and the property that we need to learn is by far too complex. If, instead of by $Z_A$ and $Z_B$, each atom of Fig. 2 (top) is identified by a string of three random numbers, KRR gives for the predicted $\Delta E$ a fit with a root mean square error (RMSE) of less than 0.01 eV. However, the leave-some-out cross validation uncovers the artifact: The RMSE is raised to 100% of the predicted quantity (see Table I).

We consider the following properties to be important, if not necessary, for a descriptor:

a) A descriptor $d$, uniquely characterizes the material $i$ as well as property-relevant elementary processes. 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) Materials with different properties $P_i$ should not cluster in the space of $d$. Fig. 1 servers as a good and Fig. 2 (top) as a bad example.

e) The dimension $\Omega$ of the descriptor should be as low as possible, but not lower.

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

In order to identify a good descriptor, we start with a large number $M$ of candidates for the components of $d$. We then look for the $\Omega$-dimensional ($\Omega = 1, 2, \ldots$) descriptor $d$ that gives the best linear fit of $P_d$: $P(d) = dc$, where $c$ is the $\Omega$-dimensional vector of coefficients. It is determined by minimizing the loss function $\|P - Dc\|^2$, where $D$ is a matrix with each of the $N$ rows being the descriptor $d_i$ for each training data point, and $P$ is the vector of the training values $P_i$. We emphasize that the choice of a linear fitting function for $P(d)$ is not restrictive since, as we will show below, non-linearities are included in a controlled way in the formation of the candidate components of $d$. The function $P(d)$ is then determined by only $\Omega$ parameters.

The task is now to find, among all the $\Omega$-tuples of
candidate features, the Ω-tuple which yields the smallest $\|P - Dc\|_2^2$. Unfortunately, a computational solution for such problem is infeasible (NP-hard) [15]. LASSO [13] provides sparse (i.e., low-dimensional) solutions by recasting the NP-hard problem into a convex minimization problem:

$$\text{argmin}_{c \in \mathbb{R}^M} \|P - Dc\|_2^2 + \lambda \|c\|_1$$

where the use of the \(\ell_1\)-norm (\(\|c\|_1 = \sum_{m=1}^M |c_m|\)) is crucial. The larger we choose \(\lambda > 0\), the smaller is the \(\ell_1\)-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 \(c\) become non-zero.

We are searching for a low-dimensional descriptor that minimizes the RMSE, given by \(\sqrt{1/N\|P - Dc\|_2^2}\), for our \(N = 82\) binary compounds. The property \(P\) that we want to predict is the difference in LDA energy (\(\Delta E\)) between RS and ZB for the given atom pair AB. The order is such that element A is the one with the smallest electronegativity \(EN\), defined according to Mulliken: 

$$EN = -(IP + EA)/2.$$ 

IP and EA are atomic ionization potential and electron affinity [17]. For systematically constructing the feature space, i.e., the component of the descriptor, and then selecting the most relevant of them, we implement an iterative approach. We start from 7 atomic features for atom A: IP(A) and EA(A), H(A) and L(A), i.e., the energies of the highest-occupied and lowest-unoccupied Kohn-Sham (KS) levels, as well as \(r_s(A)\), \(r_p(A)\), and \(d(A)\), i.e., the radii where the radial probability density of the valence \(s\), \(p\), and \(d\) orbitals are maximal. The same for atom B. Besides, information regarding the isolated AA, BB, and AB dimers was added to the list, namely their equilibrium distance, binding energy, and HOMO-LUMO KS gap (a total of 9 more features). Next, we define rules for linear and non-linear combinations of the just mentioned 23 features, which we call “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 10000 candidates grouped in subsets that are used in the different iterations, where we refined the feature space (see SI). A more detailed discussion will be given in Ref. [16]. In the language of KRR we design a kernel and we do it by using physical insight. In this way we can check new mechanisms that are tested one against each other. Not surprisingly, LASSO (and actually any other method) has difficulties in selecting among two highly correlated features [18]. In our case, for instance, \(r_s\) and \(r_p\) for the same atom have a large correlation (Pearson’s index larger than 0.95, in other words the two 82-dimensional vectors of the feature \(r_s\) and \(r_p\) are almost collinear). In these cases, it is not ensured that the first \(\Omega\) selected features form the best \(\Omega\)-dimensional descriptor. Although checking cor-

relations 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 are selected and a batch of least-square fits is performed, taking in turn as \(D\) each single feature, each pair, etc. We confirmed that this strategy always finds the best descriptor by running the mentioned extensive search for several subsets of hundreds of features.

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

$$\frac{\text{IP}(B) - \text{EA}(B)}{r_p(A)^2}, \frac{|r_s(A) - r_p(B)|}{\exp(r_s(A))}, \frac{|r_p(B) - r_s(B)|}{\exp(r_d(A) + r_s(B))}$$

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, physical intuition tells that a causal descriptor should build incrementally. The RMSE, the mean absolute error (\(\text{MAE} = (1/N)\sum_i^n |P_i - P(P_{dj})|\)), and the maxi-

![FIG. 2: 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 \(\Delta E > 0.5\) eV are outside the shown window (see SI).](image-url)
TABLE I: Statistical errors in eV for the least-square fit of all data (first three lines) and for the test set in a leave-10%-out cross validation (CV), averaged over 150 random selections of the training set (last three lines). In parentheses the corresponding errors for Gaussian-KRR at optimized \((\lambda, \sigma)\). The numbers are reported for selected descriptors, including Zunger’s descriptor \((r_s, r_n)\) [3]. For \(Z'_A, Z'_B\), each atom is identified by a string of three random numbers, the errors are averages over 10 random selections.

<table>
<thead>
<tr>
<th>Descriptor ((\lambda, \sigma)) of KRR</th>
<th>(Z'_A, Z'_B)</th>
<th>(r_s, r_n)</th>
<th>RMSE</th>
<th>MAE</th>
<th>MaxAE</th>
<th>RMSE, CV</th>
<th>MAE, CV</th>
<th>MaxAE, CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Z'_A, Z'_B) ((3\times10^{-4}, 0.1))</td>
<td>((0.01, \sqrt{0.3}))</td>
<td>((1\times10^{-4}, \sqrt{10}))</td>
<td>((3\times10^{-4}, \sqrt{3000}))</td>
<td>((1\times10^{-4}, \sqrt{300}))</td>
<td>((1\times10^{-4}, \sqrt{300}))</td>
<td>((1\times10^{-4}, \sqrt{300}))</td>
<td>((1\times10^{-4}, \sqrt{300}))</td>
<td>((1\times10^{-4}, \sqrt{300}))</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.41 (2\times10^{-4})</td>
<td>0.40 (5\times10^{-3})</td>
<td>0.31 (0.07)</td>
<td>0.14 (0.14)</td>
<td>0.10 (0.10)</td>
<td>0.11 (0.10)</td>
<td>0.18 (0.18)</td>
<td>0.12 (0.12)</td>
</tr>
<tr>
<td>MAE</td>
<td>0.28 (1\times10^{-4})</td>
<td>0.26 (3\times10^{-3})</td>
<td>0.20 (0.05)</td>
<td>0.12 (0.12)</td>
<td>0.08 (0.08)</td>
<td>0.09 (0.09)</td>
<td>0.16 (0.14)</td>
<td>0.12 (0.12)</td>
</tr>
<tr>
<td>MaxAE</td>
<td>0.22 (8\times10^{-4})</td>
<td>0.19 (0.03)</td>
<td>0.32 (0.25)</td>
<td>0.32 (0.31)</td>
<td>0.32 (0.28)</td>
<td>0.24 (0.22)</td>
<td>0.20 (0.19)</td>
<td></td>
</tr>
</tbody>
</table>

minimum absolute error (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. The errors for a famous 2D descriptor introduced by Zunger (Refs. [3, 5] and SI), based on sums and absolute differences of \(r_s\)’s and \(r_p\)’s, are also reported in Table I. Moreover, we list values for Gaussian-KRR fits, at \((\lambda, \sigma)\) where the RMSE for a leave-10%-out cross validation is minimized. Judging from the cross-validation errors, our 2D descriptor is as good as Zunger’s 2D descriptor. However, our descriptor bears the advantage that it is derived from a broad class of alternatives by a well-defined procedure, providing a basis for a systematic improvement (in particular, with increasing \(\Omega\)). 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 emphasize two interesting aspects. First, the components of the descriptors are not symmetric wrt 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 performing worse. Second, also \(d\) orbitals play a role in determining the predicted \(\Delta E\), but only as a minor correction, i.e., appearing only in the third or higher dimension. In Fig. 2 (bottom) we show the calculated and predicted \(\Delta E\), according to our best 2D descriptor. It is evident that our 2D descriptor fulfills conditions b) and d) given above while fulfillment of conditions a), c), and e) follows from its construction.

We tested the robustness of our descriptor by performing a leave-10%-out cross validation. Thereby, the overall procedure of selecting the descriptor is repeated from scratch on a learning set obtained by leaving 10% randomly selected materials out. The resulting fitted linear model is applied to the excluded materials and the RMSE on this set, averaged over 150 random selections, is 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 never changed while the 1D descriptor was the same in 90% of the cases.

Although classification between RS and ZB materials is achieved with a 2D descriptor, the quantitative description of \(\Delta E\) is not good enough (see “MaxAE, CV” in the table). We note that with the best 5D descriptor, the average lowers to a more acceptable value of 0.12 eV, while the 7D descriptor gives 0.11 eV. We are currently exploring an alternative LASSO-like approach by explicitly searching for the descriptor that yields the lowest MaxAE. This can be accomplished by substituting the squared \(\ell_2\) norm in the first term of Eq. 1 by the MaxAE (the so-called \(\ell_\infty\) norm).

Financial support from the Einstein Foundation Berlin is appreciated. We thank Krishna Rajan for bringing the relevance of the Ph-V analysis to our attention.

14. V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu,


[17] IP and EA are evaluated as the energy of the half-occupied Kohn-Sham orbital in the half positively and half negatively charged LDA atom. Atoms are calculated without spin polarization, with the same setting as that used for the compounds.

[18] Two columns of $\mathbf{D}$ are correlated if the absolute value of their Pearson’s correlation index is (about) 1, uncorrelated if this index is (around) 0.