The generalization of the discrete data set \( \{ P, \mathbf{d}_i \} \) 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) \cite{1, 2} analyzed the classification challenge of ZB/WZ vs RS structures. They 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 \cite{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 \cite{3, 4}. We will come back to this below.

In recent years, the demand for finding the function

![Experimental ground-state structures of 68 octet binary compounds](image-url)
\( P(d) \) employed statistical learning theory, which is the focus of this paper. This strategy has been put forward by several authors in materials science \cite{bib10,bib11} as well as in bio- and cheminformatics (see, e.g., Ref. \cite{bib11} 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: 
\[
P(d) = \sum_{i=1}^{N} c_i \exp \left( -\|d_i - d\|^2 / 2\sigma^2 \right),
\]
where \( N \) is the number of training data points. The coefficients \( c_i \) are determined by minimizing 
\[
\sum_{i=1}^{N} \left( P(d_i) - P_i \right)^2 + \lambda \sum_{i,j=1}^{N} c_i c_j \exp \left( -\|d_i - d_j\|^2 / 2\sigma^2 \right),
\]
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 \cite{bib12}, 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. In this Letter, we describe the challenge of finding physically meaningful descriptors for the accurate prediction of a given property of a class of materials, where we restrict ourselves to \textit{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 quantitive 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 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 \cite{bib13}.

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 FHIaims \cite{bib14} 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 and all in- and output files can be downloaded from \url{http://nomad-repository.eu}. 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 \textit{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) when the whole set is used for learning. 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.

The nuclear numbers may be felt as a too simplistic choice, therefore we encode into a candidate descriptor the electronic aufbau principle, by using \( (R_A, R_B, V_A, V_B) \), i.e., the principal quantum number (the same as the row, \( R \), in the periodic table) of both elements and the number of valence electrons, \( V \), of A (note that for the octet binaries \( V_B = 8 - V_A \)). Also in this case, the cross-validation errors for the KRR fit are very large (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) The dimension \( \Omega \) of the descriptor should be as low as possible, but not lower.

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

In order to identify a good descriptor, we start with a large number \( M \) of candidates (the “feature space”) 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 \( P(d) \): 
\[
P(d) = P(d_c) = \|P - Dc\|^2_2,
\]
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. 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 for-
noma 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 $\Omega$-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:

$$\arg\min_{c \in \mathbb{R}^M} \|P - Dc\|_2^2 + \lambda \|c\|_1 \quad (1)$$

where the use of the $\ell_1$-norm ($\|c\|_1 = \sum_{\alpha=1}^M |c_\alpha|$) 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 $\lambda > 0$, such that the solution of Eq.\(^{[1]}\) is zero. If $\lambda < \lambda_i$, one or more coordinates of $c$ become non-zero.

| Descriptor | $\mathbf{Z}_A, \mathbf{Z}_B$ | $\mathbf{V}, \mathbf{R}|r_{\mathbf{r}_A, r_{\mathbf{r}_B}}$ | 1D | 2D | 3D | 5D |
|-----------|-----------------|-----------------|----|----|----|----|
| RMSE      | $2 \times 10^{-4}$ | 0.16             | 0.07 | 0.14 | 0.10 | 0.08 | 0.06 |
| MaxAE     | $2 \times 10^{-4}$ | 0.42             | 0.25 | 0.32 | 0.32 | 0.24 | 0.20 |
| RMSE, CV  | 0.19             | 0.22             | 0.09 | 0.14 | 0.11 | 0.08 | 0.07 |
| MaxAE, CV | 0.43             | 0.46             | 0.17 | 0.27 | 0.18 | 0.16 | 0.12 |

TABLE I: Root mean square error (RMSE) and maximum absolute error (MaxAE) in eV for the least-square 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 $(\mathbf{Z}_A, \mathbf{Z}_B), (\mathbf{V}, \mathbf{R}), (r_{\mathbf{r}_A, r_{\mathbf{r}_B}})$\(^{[3]}\) are for a KRR fit at hyperparameters ($\lambda, \sigma$) that minimize the RMSE for the L-10%-OCV (see SI). The errors for the $\Omega = 1, 2, 3, 5$ (noted as 1D, 2D, 3D, 5D) descriptors are for a (linear) least square 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.

We note that the so-called “feature selection” is a widespread set of techniques that are used in statistical analysis in different fields\(^{[10]}\), and LASSO is a feature selection technique. In this paper, we add the constraint of looking for physically motivated features for constructing our descriptor, by means of a construction of the feature space based on scientific insight and by 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. For systematically constructing the feature space, i.e., the candidate components 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 $r_{\mathbf{r}_A}$, $r_{\mathbf{r}_B}$, and $r_{\mathbf{r}_d}(A)$, 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). These 23 features, we call “primary features”. Next, we define rules for linear and non-linear 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, where we refined the feature space (see SI). A more detailed discussion will be given in Ref.\(^{[17]}\). 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 these cases, it is not ensured that the first $\Omega$ selected features form the best $\Omega$-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 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, the first, the first two, and all three of the following features:

$$\frac{IP(B) - EA(B)}{r_p(B)^2} \cdot \frac{|r_{\mathbf{r}_A}(B) - r_{\mathbf{r}_B}(B)|}{\exp(r_{\mathbf{r}_A}(B))}, \frac{|r_{\mathbf{r}_B}(B) - r_{\mathbf{r}_A}(B)|}{\exp(r_{\mathbf{r}_d}(A))} \quad (2)$$

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, intuition tells that a physically meaningful descriptor should build incrementally. 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 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. The cross-validation error of the linear fit with our 2D descriptor is as small as the highly non-linear KRR-fit with Zunger’s 2D descriptor. However, our descriptor bears the advantage that it is derived from a broad class of options 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. Note that the symmetry was explicitly broken in the construction of the test set, as the order AB in the compound is such that \( \text{EN}(A) < \text{EN}(B) \). 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) given above, while fulfillment of conditions a), c), and d) follows from its construction.

In order to further test the robustness and the physical meaningfulness of our descriptor we performed one test by perturbing the value of the property \( \Delta E \). A uniform noise in the interval \([-0.1, 0.1] \) eV was added to each value of \( \Delta E \) and the 2D descriptor of Eq. 2 was identified 93% of the times, with an increase of RMSE only by 10%. More details are reported in Ref. [17]. This test shows that the model allows for a relatively large uncertainty in the measure property, as long as the physical meaning does not change (i.e. materials deep into the ZB or RS region are not completely misplaced). In other words, the automatically identified descriptor contains the important physically meaningful ingredients for the prediction of \( \Delta E \), even though a physical model that justifies the \( P(d) \) mapping is not transparent.

We now comment on the causality of the learned descriptor-property relationship. We start by saying that understanding of causality is different from that in the statistical inference community, namely in the sense of the PC algorithm [19] or counterfactual reasoning [20] because we do not deal with random variables extracted from a known or unknown distribution. Our materials have definite values of property and primary features and there is a one to one mapping between the set of primary features and the material. Therefore, we use the following observations as an indication of having identified a causal (which, in our understanding, is equivalent to saying “physically meaningful”) descriptor for the property \( \Delta E \): a) the stability of the selection of the descriptor upon both L-10%-OCV and perturbation of the values of the property, b) the simple form of the \( P(d) \) dependence (small number of fit parameters and simple functional form, see Eq. 2 and SI). Aspect b) was actually explicitly built in for the selection of the model, but it was not given a priori that this choice would imply its stability, in the sense described at point a). We note that the simplicity of our model is in sharp contrast with what is yielded by, for instance, KRR, where as many fit parameters as number of observed points are in principle necessary. Physical intuition tells that the dimensionality of the model is intrinsic to the physical property and system, not depending on the number of observations.

Financial support from the Einstein Foundation Berlin is appreciated. JV acknowledges the financial support of the ERC CZ grant LL1203. 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.
[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] C. C. Fischer, K. J. Tibbetts, D. Morgan, and G. Ceder, Nature Materials 5, 641 (2006).
[7] K. Rajan, Annu. Rev. Mater. Res. 38, 299 (2008).
[8] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, Scientific Reports 3, 2810 (2013).
[9] L. J. Nelson, G. L. W. Hart, F. Zhou, and V. Ozoliņš, Phys. Rev. B 87, 035125 (2013).
[10] T. Mueller, E. Johlin, and J. C. Grossman, Phys. Rev. B 89, 115202 (2014).
[11] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. 108, 058301 (2012).
[12] T. Hastie, R. Tibshirani, and J. Friedman, The elements of statistical learning (Springer, New York, 2009).
[13] R. Tibshirani, J. Royal Statist. Soc. B 58, 267 (1996).
[14] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, Comp. Phys. Comm. 180, 2175 (2009).
[15] S. Arora and B. Barak, Computational Complexity: A Modern Approach (University Press, Cambridge, 2009).
[16] I. Guyon and A. Elisséeff, J Mach Learn Res. 3, 1157 (2003).
[17] L. M. Ghiringhelli, J. Vybiral, S. Levchenko, C. Draxl, and M. Scheffler (2014), in preparation.
[18] Two columns of $\mathbf{D}$ are correlated if the absolute value of their Pearson’s correlation index is (about) 1.
[19] P. Spirtes, C. Glymour, and R. Scheines, Causation, Prediction and Search (Springer Lecture Notes in Statistics no. 81, Springer-Verlag, New York, 1993).
[20] J. Pearl, Causality: models, reasoning and inference (MIT press, Cambridge, 2000).