# Towards machine learning for microscopic mechanisms:
# a formula search for crystal structure stability based on atomic properties

Udaykumar Gajera,[1, 2] Loriano Storchi,[3] Danila Amoroso,[1, 4] Francesco Delodovici,[1] and Silvia Picozzi[1]

[1)]*Consiglio Nazionale delle Ricerche, CNR-SPIN c/o Università "G. D'Annunzio", 66100 Chieti,*
*Italy*

[2)]*Chemistry Department, University of Turin, via Pietro Giuria, 7, 10125, Torino, Italy*

[3)]*Dipartimento di Farmacia, Universitá degli Studi G. D'Annunzio, 66100 Chieti, Italy*

[4)]*NanoMat/Q-mat/CESAM,Universite de Liege, B-4000 Liege, Belgium*

Machine Learning (ML) techniques are revolutionizing the way to perform efficient materials modeling. Nevertheless, not all the ML approaches allow for the understanding of microscopic mechanisms at play in different phenomena. To address the latter aspect, we propose a combinatorial machine-learning approach to obtain physical formulas based on simple and easily-accessible ingredients, such as atomic properties. The latter are used to build materials features that are finally employed, through Linear Regression, to predict the energetic stability of semiconducting binary compounds with respect to zincblende and rocksalt crystal structures. The adopted models are trained using dataset built from first-principles calculations. Our results show that already one-dimensional (1D) formulas well describe the energetics; a simple grid-search optimization of the automatically-obtained 1D-formulas enhances the prediction performances at a very small computational cost. In addition, our approach allows to highlight the role of the different atomic properties involved in the formulas. The computed formulas clearly indicate that "spatial" atomic properties (*i.e.* radii indicating maximum probability densities for $s, p, d$ electronic shells) drive the stabilization of one crystal structure with respect to the other, suggesting the major relevance of the radius associated to the $p$-shell of the cation species.

## I. INTRODUCTION

Modeling material properties with high accuracy and low computational cost is one of the grand-challenges in materials science and engineering. The development of ab-initio methods have provided accurate tools for material properties prediction and their further optimization; nevertheless, one disadvantage of approaches relying only on first-principles simulations is the high cost required in terms of computational resources and simulation time. In recent years, the continuous growth of available computational power[1] has stimulated scientists to move in the direction of high-throughput simulations[2–10]. Along this line, open access databases, such as OQMD[11][12], NOMAD[13,14], Aflowlib[15], C2DB[16,17], QPOD[18], Materials Project[19], Materials Cloud[20] and related AiiDa[21,22], provide researchers with a huge collection of basic first-principles results. A large amount of ab-initio data is thus available, which can be used for deeper analyses and studies, provided one can count on proper tools to extract relevant information out of them. Therefore, in the last years, materials scientists have developed different Machine Learning (ML) methods to rationalize the data analysis[23–32]. Each method has its own specific advantages and limitations. Methods like Neural Network (NN)[33] or Random Forest [34] are very efficient[35] but not always transparent, blurring the comprehension of the role played by the input variables in the final results; ML methods, based for instance on linear regression (LR)[36,37], appear to be more suitable to obtain predictive and comprehensible models[38,39]. Nevertheless, finding a linear dependence between input and output properties is not always an easy task.

In this work, we thus propose a ML-based approach to build sets of features (or descriptors) starting from a given set of basic variables (e.g., atomic properties), which are subsequently used to construct LR models (or formulas). To test our method, we target a prototypical case in material science: the classification of the most stable crystal structure between rock-salt (RS) and zinc-blende (ZB) for semiconductor AB binary compounds[40]. In our approach, we adopt both simple one-dimensional and multi-dimension LR. To identify useful features, we generate combinations of basic atomic properties (*i.e.* the independent variables in our approach) of the material constituents through a combinatorial approach[41]. We then carry out an analysis of the emerging best-performing formulas, identifying the role of specific atomic features in determining the final stabilization of the crystal structure. Finally, we test the predictive capability of the obtained formulas by applying them to "new" compounds (*i.e.* outside the dataset used for training the model), finding an overall satisfactory agreement with first-principles results. We remark that our approach is similar to what originally proposed by Ghiringhelli *et al.*[40], though with some differences and further extensions, which will be carefully discussed in what follows.

## II. METHODOLOGY

The approach we present here can be regarded as a combinatorial machine-learning: a set of basic atomic properties (APs, listed in Table S.2 in Supplementary Information) are randomly combined (though under certain initial constraints detailed below), to build a set of material features (MFs). The generated features are then used to train a LR model, where the energy difference between rocksalt and zincblende structures is the dependent variable (i.e., the label). Then, we select the best performing model according to standard performances metrics, such as the Root Mean Squared Error (RMSE). The final result of this procedure is a "formula",

which is a concise and clear representation of the relationship between the used atomic properties and the energy difference between RS and ZB phases. In the following, we describe in detail the different steps of our approach.

## A.    Dataset preparation and materials

As mentioned, we aim at predicting the total energy difference ($\Delta E = E^{RS} - E^{ZB}$) between RS and ZB phases of cubic crystal structures for 82 semiconductor binary AB compounds (the dataset is reported in table S.2 in SI). We employed total energies reported in Ref.[40], which were calculated through density functional theory (DFT)[42][43] within the local density approximation (LDA[44]).

The construction of the material features (MFs), is based on primary atomic properties of the constituents, also taken from Ref.[40]. To facilitate the physical interpretation of each MF, the APs are subdivided into two different kinds: (*i*) "energy" properties, including highest occupied Kohn-Sham level (HOMO), lowest unoccupied Kohn-Sham level (LUMO), Ionization Potential (IP), Electron affinity (EA); (*ii*) "spatial" properties, including $r_s$, $r_p$, and $r_d$, *i.e.* the radii where the radial probability density of the valence $s$, $p$, and $d$ orbitals, respectively, reaches its maximum.

### 1.    Formulas construction

We rely on the LR[36][37] approach to obtain a direct interpretation of the dependent and independent variables. The construction of a useful LR model can become troublesome, requiring a linear dependence between features. In Ref.[40], the authors implemented an automated feature selection method employing the LASSO regression analysis method[40][46]. In our work, we use a combinatorial approach to generate the dependent variable (material features) to be used within the linear equations, and thus to finally obtain the formulas.

In Fig. 1, we illustrate the workflow of the formula generation and selection using LR. The process starts with the selection of the APs to be combined. Afterwards, we choose prototype functions that are simple analytical operations applied to the APs. In our case, we selected 5 prototype functions, $f(x)$, namely $x, x^2, x^3, \sqrt{x}, e^x$. where $x$ is an AP. Then, we obtain the final set of MFs by combining different prototype functions via the combinatorial approach (see for instance[41]), and applying the following additional set of rules:

- *GEN*1: combine two prototype functions in the numerator, forcing them to belong to the same kind of APs, that is both "spatial"-like or both "energy"-like; one prototype function is at the denominator with the only constraint to be non-zero, such as

$$MF = \frac{f_1(AP_1) \pm f_2(AP_2)}{f_3(AP_3)} \tag{1}$$

- *GEN*2: combine two prototype functions with same kind of APs at the numerator, and a single prototype

function at the denominator with argument of a different kind with respect to the numerator ones. For instance, if $AP_1$ in $f_1(AP_1)$ and $AP_2$ in $f_2(AP_2)$ is an "energy" term (*i.e. EA* or *HOMO*), then $AP_3$ must be a "spatial" term, (*i.e.* $r_p$)

$$MF = \frac{f_1(AP_1) \pm f_2(AP_2)}{f_3(AP_3)} \tag{2}$$

- *GEN*3: combine two prototype functions at both the numerator and denominator without any constraints

$$MF = \frac{f_1(AP_1) \pm f_2(AP_2)}{f_3(AP_3) \pm f_4(AP_4)} \tag{3}$$

- *GEN*4: combine two prototype functions with the same physical dimensions at both the numerator and denominator

$$MF = \frac{f_1(AP_1) \star f_2(AP_2)}{f_3(AP_3) \star f_4(AP_4)} \tag{4}$$

where $\star = + - \times \div$

Each one of these set of rules corresponds to a different MFs generator.

From the implementation point of view, each generator is a Python[47] function that produces a set of strings. Therefore, we can easily exploit the Python capability to parse a source code and run Python expression (code) within a program[48] to compute all the MFs' values starting from the generated sets of strings. This allows for an easy implementation and plugin of other generators, leaving the workflow unchanged: a new generator can be introduced implementing a Python function returning a list of strings, each one being a valid MF.

Finally, in order to choose the optimal formula, we build a LR model for each of the generated MF. To practically select the best model, *i.e.* the "best formula", we randomly split the full dataset into : 90% as training set to train/initialize the model; 10% as a test set to check model's performance. We perform this random splitting $N$ times (with $N = 150$) for each model, and we calculate the RMSE from the test set for each run. Afterwards, we again verify the top 10 resulting best formulas with a higher value of training set and test set splitting, with $N = 1000$. We average it over all $N$ splitting, and we obtain $avg(RMSE)$, as reported in our Tables.

We mention that different metrics for evaluating regression model can lead to different formulas ranking. In this work, we rank the obtained models based on the lowest $avg(RMSE)$ for direct comparison with a previous work[40].

## B.    Formula optimization

In order to further improve the performance of our models, we introduce an additional step, which we refer to as "formula optimization". In detail, we focus on the top 10 formulas obtained using each generator and the subsequent LR, as described in the previous section. After that, we use a grid
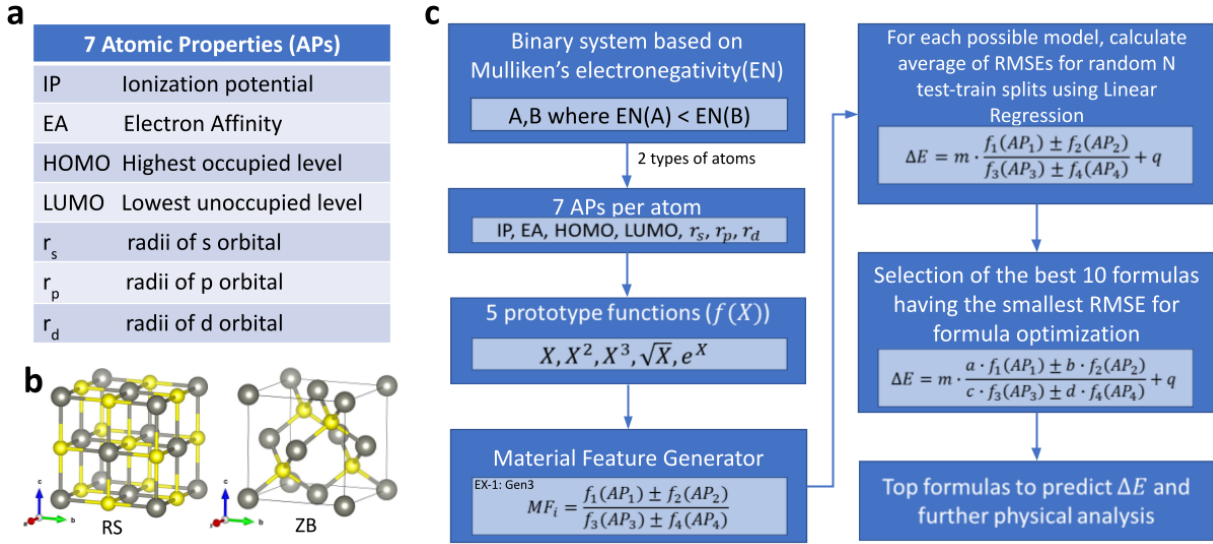
FIG. 1. a) Basic atomic properties (APs) used to construct the material features. b) Crystal structures of RS and ZB (plot made using the VESTA tool)[45]. Grey (yellow) spheres represent A (B) atoms. c) Workflow for formulas construction, machine learning methodology, validation, and MF selection procedures. In the AB compounds, A is the atom with the lowest electronegativity.

search to find the relative weights of each prototype function of the atomic properties (i.e., each $f_i(AP_i)$) within the formula. A first grid search ranging between -1 to 1 with the increasing step of 0.1 is used. We multiply each $f_i(AP_i)$ of the formula by the weight coefficient and we optimize the final RMSE value. Once the procedure finds a set of optimal weight coefficients, two subsequent grid searches, with reduced incremental step values (0.01 and 0.001 respectively) and range of search are performed to obtain the final set of refined weight coefficients. Noteworthy, for each set of weight coefficients generated during the grid search, we also run the linear regression. Thus, we are performing a proper formula optimization, as at each step of the grid search we are updating both the weight coefficients as well as the slope and intercept coming from the LR.

To further clarify the procedure, we show here an exemplary equation:

$$\Delta E = m \cdot \frac{a \cdot f_1(AP_1) \star b \cdot f_2(AP_2)}{c \cdot f_3(AP_3) \star d \cdot f_4(AP_4)} + q \qquad (5)$$

where $\Delta E$ is the targeted material feature (MF), $a, b, c, d$ denote the weight coefficients scanned during the grid search, $f_1(AP_1), f_2(AP_2), f_3(AP_3), f_4(AP_4)$ are the prototype functions build on the primary atomic properties $AP_i$, and $m$ and $q$ are the the the slope or angular coefficient and intercept, respectively, recursively determined upon LR.

In Table II, we report the optimized, best performing formula from the different generators; the top 10 formulas are reported in Table S.1 of the Supplementary Material.

To benchmark our grid search, we also used automated coefficient-optimizing methods: Nelder-Mead[49], Conjugate Gradient (CG)[50], Broyden–Fletcher–Goldfarb–Shanno (BFGS)[51] and Truncated Newton method (TNC)[52]. Although the resulting sets of coefficients are different in terms of single values with respect to those obtained via the grid search,

the ratios between them is almost preserved as well as the associated RMSE. In particular, for the case of $GEN1$ and $GEN2$, the ratio between the numerator coefficients $a$ and $b$ is preserved; for $GEN3$ and $GEN4$ also the denominator coefficients ratio, between $c$ and $d$, is preserved. In Fig. S.3 of the Supplementary Material, we show the evolution of the RMSE and different ratios for different methods using 1D feature generated by $GEN3$.

### C.   Higher-dimensional features

For the construction of higher dimensional 2D-formulas, we combined in all possible ways pairs of MFs extracted from the best 1000 ones and checked the $avg(RMSE)$ using multiple LR for $N$ test-train set splits. We followed the same process to construct the 3D formulas, where three different 1D MFs are combined. The comparison between performances is discussed in the following Section.

### D.   Test of predictive power of $\Delta E$ formula for novel AB compounds

After obtaining the optimised 1D formulas for $\Delta E$ in the case of AB compounds, we aimed at further verifying their validity and predictive power, by considering additional AB systems (*i.e.* which were not originally included in the ML training set) and by comparing values obtained from ML-predicted $\Delta E$ formula with corresponding *ab-initio* calculated values. In closer detail, we focused on different alloys, obtained by changing respectively the concentration of A-site atoms, such as $[A_x A'_{1-x}]B$, and of B-site atoms, such as $A[B_x B'_{1-x}]$. Accordingly, one can test the efficiency of the formulas by check-

ing the energy difference for intermediate concentrations as obtained from optimised 1D formulas and compare their trend with respect to first-principles results. To this end, ab-initio electronic-structure simulations were carried out within DFT and LDA functional. Calculations were performed using the VASP[53–55] code, employing a $8 \times 8 \times 8$ k-mesh for the Brillouin zone sampling. We verified that the results obtained with the pseudopotential VASP for the parent binary compounds were consistent with those reported by Ghiringhelli *et al.*, calculated with the all-electron FHI-aims code[56]. For simulations at different concentrations, we adopted the so called "Virtual Crystal Approximation" (VCA), based on virtual atoms interpolating between the real constituent atoms[57,58]. However, as well known from the literature, the VCA approach neglects some effects, such as local distortions around atoms and, as such, should not be expected to reproduce fine details of disordered alloys properties[59]. Accordingly, in some cases (*i.e.* for $Mg_xCa_{1-x}Se$ alloys), in order to mimic disordered structures with an improved accuracy, we calculated total energies using supercell structures, rather than using the VCA method on primitive unit cells. Specifically, the considered supercell is the cubic unit cell composed by four AB formula units with planes of cations alternating along the **c** direction (see Figure S.4). The k-mesh was modified accordingly, to maintain the same density of points employed in the simulations of primitive cells.

## III.   RESULTS AND DISCUSSION

In this section, we will analyse the final formulas as obtained from different generators. The results are shown in Tables I,II,III,IV; in the first row we report the results obtained by Ghiringhelli *et al.*[40] for comparison.

First, by comparing the $avg(RMSE)$ values, we note that all 1D formulas obtained from our different generators better perform with respect to the 1D ones reported in[40], where the authors used the automated feature selection method LASSO[46]. Noteworthy, some atomic primary features appearing in 1D formulas of Ref.[40] also appear in our obtained list of 1D formulas using $GEN1$ and $GEN2$; nevertheless, those are characterized by a higher $avg(RMSE)$ than other formulas we obtained via our combinatorial approaches. Additionally, formulas from $GEN3$ show the lowest $avg(RMSE)$ among all the others. We also note, from Table I, that $GEN1$ and $GEN3$ provide lower $avg(RMSE)$ compared to $GEN2$ and $GEN4$ respectively; however, $GEN2$ and $GEN4$ have a higher success rate in terms of classification prediction. This testifies the fact that the choice of the performances metric to rank the material features can be different according to the target problem to be studied; different models' performances metric are, in fact, not always correlated.

In order to gather hints on the relative contribution of the individual primary atomic properties to the stabilization of either the rocksalt or the zincblende structure, we extracted the best ten formulas with the lowest $avg(RMSE)$ from each generator (so called "original" formulas) and then apply the formula optimization, as detailed in the previous section. This

procedure attributes relative weights to each $f(AP)$, allowing to measure the importance of the individual atomic properties in driving the energy stabilization. In principle, the $avg(RMSE)$ value depends on random test-train splits that we perform to our dataset. Therefore, to reduce the effect of randomization, as a target model performances metric, we rank our optimized formula based on the RMSE of the whole dataset, rather than based on $avg(RMSE)$. By comparing Table I and Table II, it is evident that the optimization procedure can further change the formulas ranking, providing a different final "best formula" with respect to the non-optimized formulas. In particular, we notice an improvement in RMSE around 5-10% after the formula optimization.

Interestingly, our results reveal the size of the A-ion to play a leading role in the phase stabilization; in fact, the $r_p(A)$ radius appears in the best performing formulas more frequently than the other basic atomic properties. Therefore, we further analysed the dependence of $\Delta E$ on $r_p(A)$. In Fig. 2, we show $\Delta E$ as a function of $r_p(A)$, including fitting curves proportional to $r_p(A)^{-2}$ and $r_p(A)^{-3}$. What can be observed is a clear dependence of $\Delta E$ on $r_p(A)$: larger (smaller) $r_p(A)$ favors RS (ZB). Moreover, there is an overall good agreement with the fit, particularly using the $r_p(A)^{-3}$ function. The latter is, in fact, the most recurrent prototype function detected by the ML models. Such a strong dependence for the energy is not observed with respect to the other atomic properties; other comparative plots of $\Delta E$ as a function of other $f(p)$ are reported in Fig. S.2 of the Supplementary Material.

From the obtained results, we remark that formulas based on "spatial" atomic properties achieve higher ranking, thus better performance, with respect to those including atomic energy terms, both in the original models and in the optimized ones. Accordingly, this behaviour further confirms the primary role played by the atomic size (or, equivalently, steric effects), in determining the energetics of the AB compounds, *i.e.* in selecting the preferred crystal structure[40].
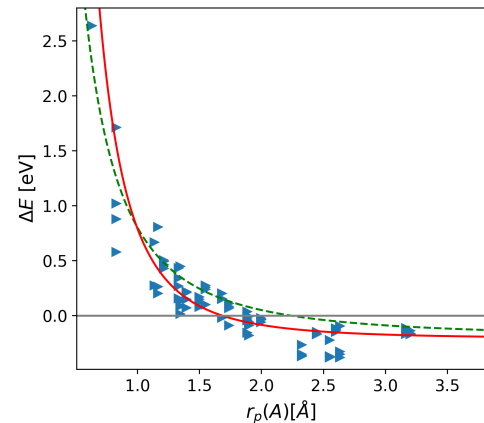


FIG. 2.    Energy difference between rocksalt and zincblende, $\Delta E$ (in eV), as a function of $r_p(A)$ for different binary compounds (blue triangles). Data fit functions are also shown, using proportionality to $r_p(A)^{-2}$ and $r_p(A)^{-3}$ via green dashed line and red straight line, respectively.

In the aim of further proving such trends and validate the implemented combinatorial ML method, we study the energetics in alloys of the type $[A_xA'_{1-x}]B$ and $A[B_xB'_{1-x}]$, where $x$ is the relative concentration of the mixing ions, monotonically tuning thus the average size of one ion with respect the other. All the alloy input properties were linearly interpolated between corresponding values for end binaries (*i.e.* $AB$ and $A'B$ in the $[A_xA'_{1-x}]B$ case), according to the Vegard's law[60]. For the A-ion mixing case, we considered SrSe, CaSe, MgSe, BeSe as parent AB compounds, already included in the original dataset. We then predicted the energy differences between RS and ZB phases for varying concentrations using the original and optimized 1D formulas constructed via $GEN3$ and $GEN4$ generators (Table I and Table II, respectively). To confirm the obtained predictions, we thus calculated the energy difference via DFT simulations, for a few intermediate concentrations. The results, shown in Fig. 3, demonstrate an overall agreement between first-principles calculated and machine-learning predicted energetics. In particular, we notice a change of sign in $\Delta E$, reflecting the change in the stability of the RS with respect to the ZB phase, when moving from the larger Strontium to the smaller Beryllium at the A-site, in line with the previously discussed relation between atomic radii of the A-ions and phase stabilization. At variance, no such change of phase is observed when mixing ions at the B-site, keeping fixed the A-type one. This is confirmed, by looking at the energetics in $B[Sb_{1-x}P_x]$ and $Sr[Se_{1-x}S_x]$ alloys, shown in Fig.4(a) and Fig. 4(b), respectively. Despite the changing size of the average B-site, the two systems preserve the crystal structure adopted by the the parent compounds, *i.e.* rocksalt for the Sr-based compounds and zincblende for the B-based compounds. Such a behavior is still in line with preferred atomic structure fixed by the ion at the A-site, consistently with Strontium being larger than Boron. Qualitative agreement between ML-predicted and DFT-calculated energetics is observed again.

After discussing the results related to 1D models, we now comment about the higher dimensional formulas. Our best 2D and 3D formulas from different generators are reported in Tables III and IV, respectively.

To visualize the performance of the obtained formulas, we reproduce in Fig. 5 the scatter plots of DFT-calculated energies as a function of model-predicted energy differences for the best formulas obtained by $GEN3$ - in terms of $avg(RMSE)$ - for 1D, 1D after formula optimization, 2D, and 3D models. From these, one can infer the quality of the prediction for the different approaches: the narrower the area between red lines (representing $2 \times avg(RMSE)$), the smaller the error or, equivalently, the more reliable the prediction. Notably, this is the case when building higher dimension formulas.

In addition, a careful comparison between our results and those reported in the reference paper, Ref.[40], is reported in Table S.1 of the Supplementary Material. In particular, in Fig. S.1 we compared the scatter plot of the 1D formula from $GEN3$ and Ref.[40], with bar graphs of errors for individual compounds. To check the improvement with respect to 1D

formulas, we considered the $avg(RMSE)$ value, as also chosen in Ref.[40]. One can observe the improvement in $avg(RMSE)$ if we examine 1D and 2D formulas in Tables I and III. We notice around 10-20% improvement from the original 1D to 2D, but less than 10% of optimized 1D to original 2D formulas. Furthermore, we also notice that original and optimized 1D formulas from $GEN3$ and $GEN4$ better perform with respect the corresponding 2D ones reported in Ref[40].

We remark that the process of formula optimization is less computationally expensive than the construction of higher-dimensional formulas. In addition, from the formula optimization one can gain a better physical insights about the contribution of individual primary atomic properties. These comments overall suggest that lower-dimensional formulas constitute a better choice in terms of physical interpretation and computational efficiency.

## IV.   CONCLUSIONS

The knowledge of a material stable crystal structure constitutes the starting point for any ab–initio modelling, since materials properties crucially depend on the periodic atomic arrangement in the crystal. Within this general framework, our aim here has been to exploit ML methods to correlate the energetic stability of different crystal structures (zincblende vs rocksalt) for popular binary semiconducting compounds with primary properties of their atomic constituents, the latter representing simple and easily-accessible ingredients. Based on atomic properties, we therefore built the material features using a combinatorial approach, we trained the machine learning model using the created features over a density–functional–theory dataset and we obtained simple mathematical expressions to quantitatively predict the energetic stability of one crystal structure over the other (i.e., a formula). In addition, we have also introduced an extra step following the linear regression to explore the relative contributions of individual basic atomic properties.

To investigate the performance of the combinatorial approach, we compared our results with a reference paper[40], where the authors predicted the stability of the crystal structure using an automated feature selection method. We found that our 1D formulas constructed using the combinatorial approach achieved a higher accuracy with respect to the reference ones. Furthermore, we also learned more about the underlying mechanism from the formula optimization, where we found that the stability of RS and ZB heavily depends on the $r_p$ radius of A-sites. This kind of understanding is, in general, much more difficult to achieve in heavily-automated artificial–intelligence methods, such as neural networks, where it is not possible to interpret directly the model results. In this respect, our approach based on linear regression allows the construction of physical models supported by machine-driven suggestions of relevant ingredients; as such, it should be regarded as a methodology offering a huge range of applications in addressing microscopic mechanisms underlying different phenomena, calling for extensive investigations in the nearby future.

## V.   AUTHOR DECLARATIONS

### A.   Conflict of Interest

The authors have no conflicts to disclose.

### B.   Data Availability

The data that supports the findings of this study are available within the article and its supplementary material. The code for machine learning is available at https://github.com/lstorchi/ matinformatics .

## VI.   SUPPLEMENTARY MATERIAL

See Supplementary Material for technical details related to LR, DFT calculations of the alloy supercell, dataset and for additional results related to 1D, 2D, 3D formulas.

## VII.   ACKNOWLEDGMENTS

| Formulas | avg(RMSE) | RMSE | $R^2$ | Success rate | Generator type |
|---|---|---|---|---|---|
| $0.117 \cdot \frac{EA(B)-IP(B)}{r_p(A)^2} - 0.342$ | 0.1455 | 0.1423 | 0.89 | 89% | 1D descriptor[40] |
| $-0.751 \cdot \frac{r_p(B)^3 - exp[r_s(B)]}{r_p(A)^2} - 0.317$ | 0.1296 | 0.1193 | 0.92 | 90% | GEN1 |
| $0.285 \cdot \frac{\sqrt{|IP(B)|}+\sqrt{|EA(A)|}}{r_p(A)^2} - 0.387$ | 0.1367 | 0.1309 | 0.91 | 91% | GEN2 |
| $0.774 \cdot \frac{r_p(B)+\sqrt{|r_d(A)|}}{r_p(A)^3+r_p(B)^3} - 0.303$ | 0.0995 | 0.0963 | 0.95 | 94% | GEN3 |
| $1.155 \cdot \frac{r_s(B)+r_s(A)}{r_p(B)^3+r_p(A)^3} - 0.368$ | 0.1103 | 0.1058 | 0.94 | 96% | GEN4 |

TABLE I. 1D formulas, along with related statistics: $avg(RMSE)$ denotes the root mean squared error for average over 1000 random train-test splits of dataset. Instead, the RMSE is the root mean squared error for the entire dataset as training and test. Similarly, the $R^2$ values are calculated considering the entire dataset and they show the quality of fit between predicted and actual values. The success rate (in percent) shows how many RS or ZB phases out of 82 have been correctly identified by the descriptor. The "Generator type" column indicates the different generators used to produce the corresponding descriptor. RMSEs are in eV.

| Formula | avg(RMSE) | RMSE | $R^2$ | Success Rate | Generator type |
|---|---|---|---|---|---|
| $0.127 \cdot \frac{0.800 \cdot EA(B) - 1.000 \cdot IP(B)}{1.110 \cdot r_p(A)^2} - 0.352$ | 0.1457 | 0.1419 | 0.89 | 89% | 1D descriptor[40] |
| $-1.870 \frac{0.801 \cdot \sqrt{r_p(B)} - 0.606 \cdot exp[r_p(A)]}{1.010 \cdot r_p(A)^3} - 0.968$ | 0.1191 | 0.1143 | 0.93 | 91% | GEN1 |
| $0.477 \cdot \frac{0.876 \cdot \sqrt{|HOMO(B)|}+0.468 \cdot \sqrt{|LUMO(B)|}}{1.110 \cdot r_p(A)^2} - 0.372$ | 0.1340 | 0.1296 | 0.91 | 91% | GEN2 |
| $1.609 \cdot \frac{0.642 \cdot r_p(B)+0.502 \cdot \sqrt{|r_d(A)|}}{1.170 \cdot r_p(A)^3+1.170 \cdot r_p(B)^3} - 0.309$ | 0.0991 | 0.0961 | 0.95 | 94% | GEN3 |
| $1.207 \cdot \frac{0.878 \cdot r_s(B)+0.200 \cdot r_p(A)}{0.512 \cdot r_p(B)^3+0.610 \cdot r_p(A)^3} - 0.359$ | 0.1045 | 0.1016 | 0.94 | 99% | GEN4 |

TABLE II. 1D formulas after the optimization step, along with related statistics. Notation as in table-I.

| Descriptor | avg(RMSE) | RMSE | $R^2$ | Success Rate | Generator type |
|---|---|---|---|---|---|
| $0.113 \cdot \frac{EA(B)-IP(B)}{r_p(A)^2} - 1.558 \cdot \frac{|r_s(A)-r_p(B)|}{exp[r_s(A)]} - 0.133$ | 0.1041 | 0.0988 | 0.95 | 96% | 2D descriptor[40] |
| $-0.342 \cdot \frac{r_p(B)^3 - exp[r_p(A)]}{r_p(A)^3} - 1.042 \cdot \frac{r_p(A)^2 - \sqrt{|r_d(A)|}}{exp[r_p(A)]} - 0.062$ | 0.0989 | 0.0944 | 0.95 | 89% | GEN1 |
| $-0.081 \cdot \frac{IP(B)+\sqrt{|IP(A)|}}{r_p(A)^3} - 0.001 \cdot \frac{r_s(A)^3 - \sqrt{r_d(A)}}{exp(HOMOKS(A))} - 0.062$ | 0.1163 | 0.1100 | 0.93 | 86% | GEN2 |
| $-1.175 \cdot \frac{r_p(A)-\sqrt{|r_d(A)|}}{r_s(B)^3+r_p(A)^3} + 0.513 \cdot \frac{r_s(B)+\sqrt{|r_p(B)|}}{r_p(B)^3+r_s(A)^3} - 0.250$ | 0.0911 | 0.0878 | 0.96 | 87% | GEN3 |
| $0.618 \cdot \frac{r_d(A)/r_p(B)}{r_p(A)^3 * \sqrt{r_d(A)}} + 1.097 \cdot \frac{r_p(A)*\sqrt{|r_p(B)|}}{r_p(B)^3+r_p(A)^3} - 0.384$ | 0.0995 | 0.0955 | 0.95 | 92% | GEN4 |

TABLE III. 2D descriptors, along with related statistics. Notation as in table-I.

| Descriptor | avg(RMSE) | RMSE | $R^2$ | Success Rate | Generator type |
|---|---|---|---|---|---|
| $0.108 \cdot \frac{EA(B)-IP(B)}{r_p(A)^2} - 1.806 \cdot \frac{|r_s(A)-r_p(B)|}{exp[r_s(A)]} - 3.782 \cdot \frac{|r_p(B)-r_s(B)|}{exp[r_d(A)]} - 0.023$ | 0.0818 | 0.0756 | 0.97 | 93% | 3D descriptor[40] |
| $0556 \cdot \frac{r_p(B)^3 - exp[r_p(A)]}{r_p(A)^3} + 0.364 \cdot \frac{r_p(A)^2 - \sqrt{|r_d(A)|}}{exp[r_p(A)]}, -0.124 \cdot \frac{r_p(B)^2 - \sqrt{|r_d(A)|}}{r_p(A)^3} - 1.87$ | 0.1003 | 0.0933 | 0.95 | 90% | GEN1 |
| $-0.056 \cdot \frac{(LUMOKS(A)+HOMOKS(B))}{r_p(A)^3} + 0.266 \cdot \frac{\sqrt{|EA(B)|}+exp(EA(B))}{r_s(A)^3} - 0.016 \cdot \frac{HOMOKS(A)-exp(LUMOKS(B))}{(r_p(A)^3)} - 0.310$ | 0.1300 | 0.1205 | 0.92 | 91% | GEN2 |
| $-0.885 \cdot \frac{r_p(B)-exp[r_p(A)]}{r_p(A)^2+r_p(A)^3} - 0.417 \cdot \frac{r_s(A)-exp[r_s(B)]}{r_s(A)^3+r_p(B)^3} - 0.579 \cdot \frac{r_p(A)-\sqrt{|r_d(A)|}}{r_p(B)^2+r_s(A)^3} - 0.616$ | 0.0875 | 0.0834 | 0.96 | 98% | GEN3 |
| $0.635 \cdot \frac{\sqrt{IP(B)/\sqrt{IP(A)}}}{r_p(A)^3+r_p(B)^3} + 0.730 \cdot \frac{r_p(B)*\sqrt{|r_d(A)|}}{r_p(A)^3+r_p(B)^3} + 0.038 \cdot \frac{IP(A)^2 - EA(A)^2}{exp(r_p(A))*exp(r_d(A))} - 0.358$ | 0.0989 | 0.0919 | 0.96 | 93% | GEN4 |

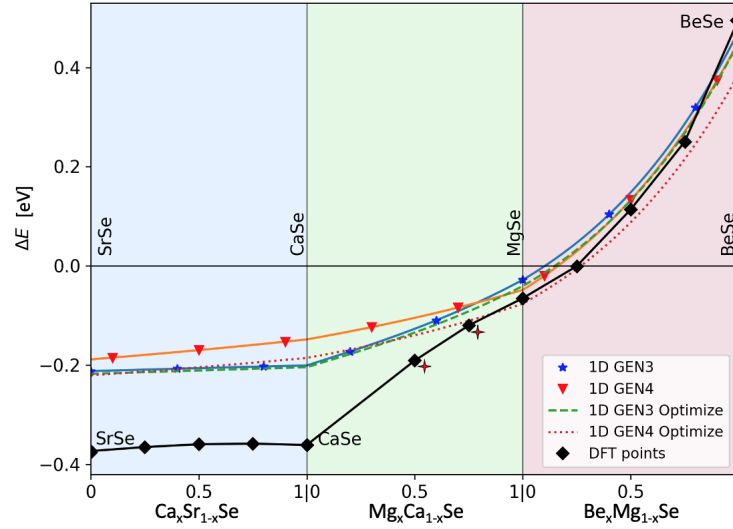TABLE IV. 3D descriptors, along with related statistics. Notation as in table-I.

FIG. 3. Total energy difference $\Delta E$ as a function of concentration($x$) for $[Ca_xSr_{1-x}]Se$, $[Mg_xCa_{1-x}]Se$ and $[Be_xMg_{1-x}]Se$ alloys, highlighted in blue, green, and pink regions respectively. Energy differences are predicted using original and optimized 1D descriptors constructed using $GEN3$ and $GEN4$ and verified using DFT (black line with diamond points) within VCA. For an improved accuracy, the two asterisk-highlighted intermediate points in the $[Mg_xCa_{1-x}]Se$ region are calculated using the supercell approach rather than VCA.
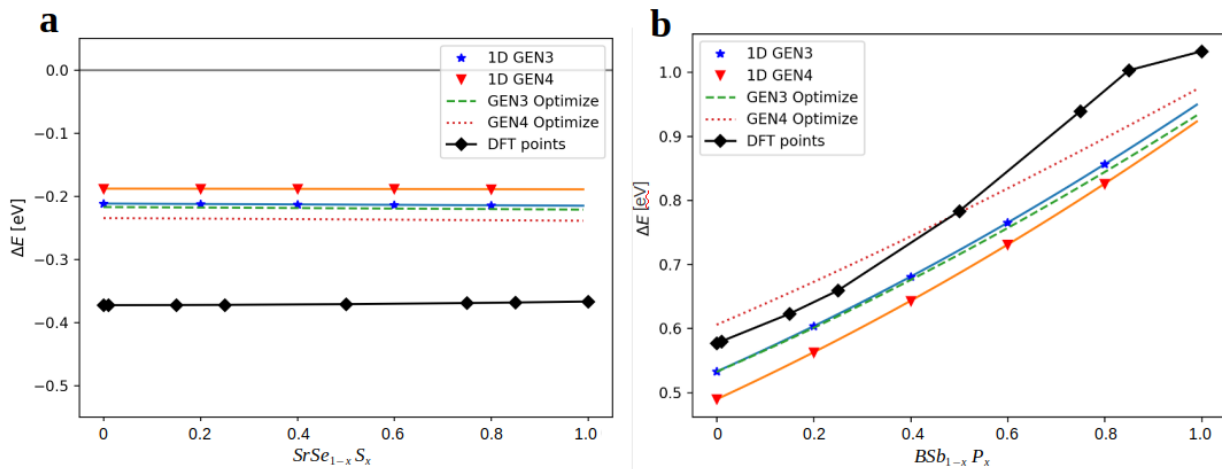


FIG. 4. Total energy difference $\Delta E$ as a function of concentration ($x$) for a) $Sr[S_xSe_{1-x}]$ and b) $B[P_xSb_{1-x}]$ alloys, predicted from original and optimized 1D descriptors constructed using $GEN3$ and $GEN4$. Model predictions are verified using energy differences calculated via DFT[42,43] (black-line with diamond points).
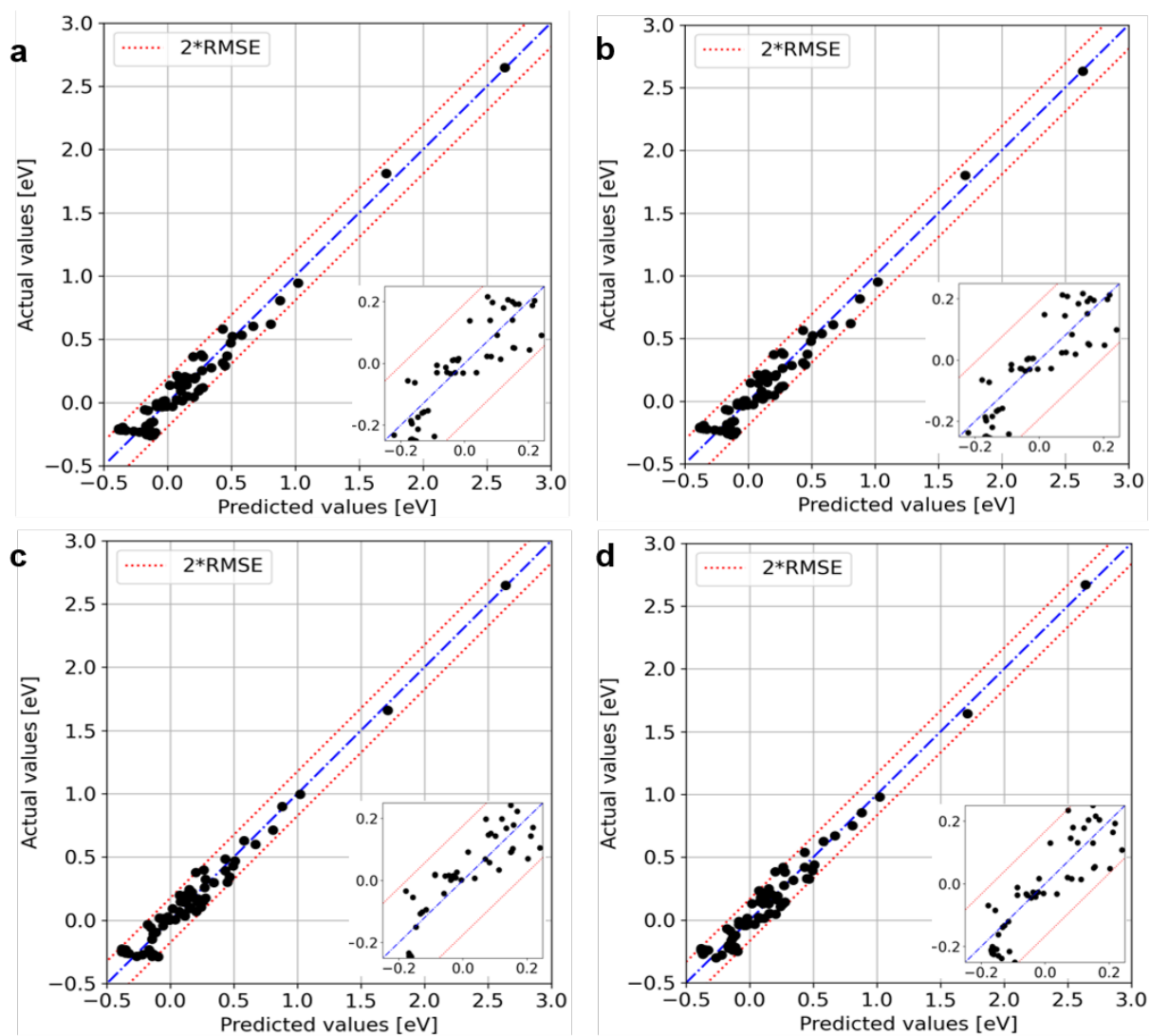
FIG. 5. Comparison of actual (*i.e.* DFT) vs predicted total energy difference $\Delta E$ for a) 1D, c) 2D and d) 3D descriptors constructed using *GEN*3. Panel b) shows the best 1D descriptors after formula optimization. Lower-right insets show a zoom in the relevant region where many compounds are concentrated. Red dotted lines correspond to $2\times avg(RMSE)$ value. The respective descriptors can be inferred from tables-I, II, III, IV

[1] G. E. Moore, "Cramming more components onto integrated circuits," Electronics **38** (1965).

[2] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, "Machine learning in materials informatics: recent applications and prospects," npj Comput. Mater. **3**, 54 (2017).

[3] M. Fukuda, J. Zhang, Y.-T. Lee, and T. Ozaki, "A structure map for AB$_2$ type 2D materials using high-throughput DFT calculations," Mater. Adv. **2**, 4392–4413 (2021).

[4] D. Schwalbe-Koda, S. Kwon, C. Paris, E. Bello-Jurado, Z. Jensen, E. Olivetti, T. Willhammar, A. Corma, Y. Román-Leshkov, M. Moliner, and R. Gómez-Bombarelli, "A priori control of zeolite phase competition and intergrowth with high-throughput simulations," Science **374**, 308–315 (2021).

[5] E. R. Homer, "High-throughput simulations for insight into grain boundary structure-property relationships and other complex microstructural phenomena," Comput. Mater. Sci. **161**, 244–254 (2019).

[6] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, "The high-throughput highway to computational materials design," Nat. Mate. **12**, 191–201 (2013).

[7] M. L. Green, C. L. Choi, J. R. Hattrick-Simpers, A. M. Joshi, I. Takeuchi, S. C. Barron, E. Campo, T. Chiang, S. Empedocles, J. M. Gregoire, A. G. Kusne, J. Martin, A. Mehta, K. Persson, Z. Trautt, J. V. Duren, and A. Zakutayev, "Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies," Appl. Phys. Rev. **4**, 011105 (2017).

[8] A. Walsh, "The quest for new functionality," Nat. Chem **7**, 274–275 (2015).

[9] J. Shen, V. I. Hegde, J. He, Y. Xia, and C. Wolverton, "High-Throughput Computational Discovery of Ternary Mixed-Anion Oxypnictides," Chem. Mater. **33**, 9486–9500 (2021).

[10] S. D. Griesemer, L. Ward, and C. Wolverton, "High-throughput crystal structure solution using prototypes," Phys. Rev. Mater. **5**, 105003 (2021).

[11] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, "Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD)," JOM, **65**, 1501–1509 (2013).

[12] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, "The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies," NPJ Comput. Mater. **1**, 1–15 (2015).

[13] C. Draxl and M. Scheffler, "NOMAD: The FAIR concept for big data-driven materials science," MRS Bull. **43**, 676–682 (2018).

[14] C. Draxl and M. Scheffler, "The NOMAD laboratory: from data sharing to artificial intelligence," J. Phys. Mater. **2**, 036001 (2019).

[15] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, and O. Levy, "AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations," Comput. Mater. Sci. **58**, 227–235 (2012).

[16] M. N. Gjerding, A. Taghizadeh, A. Rasmussen, S. Ali, F. Bertoldo, T. Deilmann, N. R. Knøsgaard, M. Kruse, A. H. Larsen, S. Manti, T. G. Pedersen, U. Petralanda, T. Skovhus, M. K. Svendsen, J. J. Mortensen, T. Olsen, and K. S. Thygesen, "Recent progress of the Computational 2D Materials Database (C2DB)," 2D Mater. **8**, 044002 (2021).

[17] S. Haastrup, M. Strange, M. Pandey, T. Deilmann, P. S. Schmidt, N. F. Hinsche, M. N. Gjerding, D. Torelli, P. M. Larsen, A. C. Riis-Jensen, J. Gath, K. W. Jacobsen, J. Jørgen Mortensen, T. Olsen, and K. S. Thygesen, "The Computational 2D Materials Database: high-throughput modeling and discovery of atomically thin crystals," 2D Mater. **5**, 042002 (2018).

[18] F. Bertoldo, S. Ali, S. Manti, and K. S. Thygesen, "Quantum point defects in 2D materials: The QPOD database," arXiv:2110.01961 [cond-mat, physics] (2021).

[19] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation," APL Mater. **1**, 011002 (2013).

[20] L. Talirz, S. Kumbhar, E. Passaro, A. V. Yakutovich, V. Granata, F. Gargiulo, M. Borelli, M. Uhrin, S. P. Huber, S. Zoupanos, C. S. Adorf, C. W. Andersen, O. Schütt, C. A. Pignedoli, D. Passerone, J. VandeVondele, T. C. Schulthess, B. Smit, G. Pizzi, and N. Marzari, "Materials Cloud, a platform for open computational science," Sci. Data **7**, 299 (2020).

[21] G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari, and B. Kozinsky, "AiiDA: automated interactive infrastructure and database for computational science," Comput. Mater. Sci. **111**, 218–230 (2016).

[22] S. P. Huber, E. Bosoni, M. Bercx, J. Bröder, A. Degomme, V. Dikan, K. Eimre, E. Flage-Larsen, A. Garcia, L. Genovese, C. Johnston, G. Petretto, S. Poncé, G.-M. Rignanese, C. J. Sewell, B. Smit, V. Tseplyaev, M. Uhrin, D. Wortmann, A. V. Yakutovich, A. Zadoks, P. Zarabadi-Poor, B. Zhu, N. Marzari, and G. Pizzi, "Common workflows for computing material properties using different quantum engines," NPJ Comput. Mater. **7**, 136 (2021).

[23] H. Park, A. Ali, R. Mall, H. Bensmail, S. Sanvito, and F. El-Mellouhi, "Data-driven enhancement of cubic phase stability in mixed-cation perovskites," Mach. Learn.: Sci. Technol. **2**, 025030 (2021).

[24] C. Kim, G. Pilania, and R. Ramprasad, "From Organized High-Throughput Data to Phenomenological Theory using Machine Learning: The Example of Dielectric Breakdown," Chem. Mater. **28**, 1304–1311 (2016).

[25] E. Tsymbalov, Z. Shi, M. Dao, S. Suresh, J. Li, and A. Shapeev, "Machine learning for deep elastic strain engineering of semiconductor electronic band structure and effective mass," NPJ Comput. Mater. **7**, 1–10 (2021).

[26] C. J. Bartel, C. Sutton, B. R. Goldsmith, R. Ouyang, C. B. Musgrave, L. M. Ghiringhelli, and M. Scheffler, "New tolerance factor to predict the stability of perovskite oxides and halides," Sci. Adv. **5**, eaav0693 (2019).

[27] A. G. Kusne, T. Gao, A. Mehta, L. Ke, M. C. Nguyen, K.-M. Ho, V. Antropov, C.-Z. Wang, M. J. Kramer, C. Long, and I. Takeuchi, "On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets," Sci. Rep **4**, 6367 (2014).

[28] H. Koinuma and I. Takeuchi, "Combinatorial solid-state chemistry of inorganic materials," Nat. Mate. **3**, 429–438 (2004).

[29] S. Manti, M. K. Svendsen, N. R. Knøsgaard, P. M. Lyngby, and K. S. Thygesen, "Predicting and machine learning structural instabilities in 2D materials," arXiv:2201.08091 [cond-mat] (2022).

[30] K. Kim, L. Ward, J. He, A. Krishna, A. Agrawal, and C. Wolverton, "Machine-learning-accelerated high-throughput materials screening: Discovery of novel quaternary Heusler compounds," Phys. Rev. Mater. **2**, 123801 (2018).

[31] J. Pal, C. W. Park, Y. Xia, J. Shen, and C. Wolverton, "Scale-invariant Machine-learning Model Accelerates the Discovery of Quaternary Chalcogenides with Ultralow Lattice Thermal Conductivity," arXiv:2109.03751 [cond-mat] (2021).

[32] M. Kuban, S. Rigamonti, M. Scheidgen, and C. Draxl, "Density-of-states similarity descriptor for unsupervised learning from materials data," arXiv:2201.02187 [cond-mat] (2022).

[33] K. Gurney, *Introduction to Neural Networks* (UCL Press Limited, London, 1997).

[34] L. Breiman, "Random forests," Mach. Learn. **45**, 5–32 (2001).

[35] T. Xie and J. C. Grossman, "Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties," Phys. Rev. Lett. **120**, 145301 (2018).

[36] S. Chatterjee and J. S. Simonoff, *Handbook of regression analysis* (Wiley, Hoboken, New Jersey, 2013).

[37] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of Massive Datasets* (Stanford University, Stanford, 2010).

[38] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," Artificial intelligence **267**, 1–38 (2019).

[39] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," Adv. Neural Inf. Process. Syst. **29** (2016).

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

[41] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, "Combinatorial screening for new materials in unconstrained composition space with machine learning," Phys. Rev. B **89**, 094104 (2014).

[42] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," Phys. Rev. **136**, B864–B871 (1964).

[43] W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects," Phys. Rev. **140**, A1133–A1138 (1965).

[44] J. P. Perdew and Y. Wang, "Accurate and simple analytic representation of the electron-gas correlation energy," Phys. Rev. B **45**, 13244–13249 (1992).

[45]K. Momma and F. Izumi, "*VESTA*: a three-dimensional visualization system for electronic and structural analysis," J. Appl. Crystallogr. **41**, 653–658 (2008).

[46]S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms* (Cambridge University Press, , New York, NY 10013-2473, USA, Cambridge, 2014).

[47]G. Van Rossum and F. L. Drake Jr, *Python tutorial* (Centrum voor Wiskunde en Informatica Amsterdam, 1995).

[48]L. Storchi, "Open source code," `https://github.com/lstorchi/matinformatics` (2022).

[49]F. Gao and L. Han, "Implementing the Nelder-Mead simplex algorithm with adaptive parameters," Comput Optim Appl **51**, 259–277 (2012).

[50]G. H. Golub and C. F. Van Loan, *Matrix computations*, fourth edition ed., Johns Hopkins studies in the mathematical sciences (The Johns Hopkins University Press, Baltimore, 2013).

[51]C. G. Broyden, "The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations," IMA Journal of Applied Mathematics **6**, 76–90 (1970), https://academic.oup.com/imamat/article-pdf/6/1/76/2233756/6-1-76.pdf.

[52]R. dembo, S. Eisenstat, and T. Steihaug, "Inexact Newton Methods," SIAM J. Numer. Anal. **19**, 400–408 (1982).

[53]G. Kresse and J. Hafner, "Ab initio molecular dynamics for liquid metals," Phys. Rev. B **47**, 558–561 (1993).

[54]G. Kresse and J. Furthmüller, "Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set," Comput. Mater. Sci. **6**, 15 – 50 (1996).

[55]G. Kresse and J. Furthmüller, "Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set," Phys. Rev. B **54**, 11169–11186 (1996).

[56]V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, "Ab initio molecular simulations with numeric atom-centered orbitals," Computer Physics Communications **180**, 2175–2196 (2009).

[57]C. Eckhardt, K. Hummer, and G. Kresse, "Indirect-to-direct gap transition in strained and unstrained $sn_x ge_{1-x}$ alloys," Phys. Rev. B **89**, 165201 (2014).

[58]L. Bellaiche and D. Vanderbilt, "Virtual crystal approximation revisited: Application to dielectric and piezoelectric properties of perovskites," Phys. Rev. B **61**, 7877–7882 (2000).

[59]D. Amoroso, A. Cano, and P. Ghosez, "First-principles study of $(Ba,Ca)tio_3$ and $Ba(Ti,Zr)o_3$ solid solutions," Phys. Rev. B **97**, 174108 (2018).

[60]L. Vegard, "Die konstitution der mischkristalle und die raumfüllung der atome," Zeitschrift für Physik **5**, 17–26 (1920).