mater.scichina.com link.springer.com

Published online 27 July 2022 | https://doi.org/10.1007/s40843-022-2134-3
Sci China Mater 2023, 66(1): 343–351

A universal model for accurately predicting the formation energy of inorganic compounds

Yingzong Liang[1,2], Mingwei Chen[1], Yanan Wang[1,2], Huaxian Jia[2,3], Tenglong Lu[2], Fankai Xie[2], Guanghui Cai[2], Zongguo Wang[4], Sheng Meng[1,2*] and Miao Liu[1,2,5*]

ABSTRACT Harnessing recent advances in data science and materials engineering, it is feasible today to build reliable models for predicting materials properties. Here we employ a comprehensive dataset of 170,714 inorganic crystalline compounds obtained from high-throughput accurate quantum mechanics calculations, to train a machine learning model for the precise prediction of the formation energy of inorganic compounds. Distinct from previous studies, our model can be universally applied to a large phase space of inorganic materials as all the data is utilized for the training, and the model reaches a fairly good predictive ability (R^2 = 0.982 and mean absolute error = 0.072 eV atom^{-1}, DenseNet model). The improvement comes from several effective structure-dependent descriptors, which are carefully designed to take into account the information of the electronegativity difference between neighboring atoms and local atomic structure. This model provides a useful tool to predict the energy landscape of the compound systems in a fast and cost-effective manner.

Keywords: machine learning, formation energy, electronegativity

INTRODUCTION

The formation energy (E_{form}) of crystals, i.e., the energy to bind atoms together to form condensed phase of matters, is one of the most important physical properties of a material. The E_{form} represents the strength of the adhesion of atoms to each other within a material system, and is responsible for many thermodynamic- and kinetic-related properties, such as stability [1] and synthesizability [2] of a compound. The value of the E_{form} can be obtained from the first-principles calculations, which are relatively computation-intensive and expensive. Harnessing the advances of computational materials science databases, as well as the knowledge of data science, the E_{form} prediction can now be achieved in an easier and more effective way.

Previously, attempts have been made to quickly gauge the E_{form} of inorganic materials. For example, Cao *et al.* [3] utilized a convolution neural network (CNN) model and mixed Magpie descriptors [4] and orbital-field matrix (OFM) descriptors [5] as input for predicting the E_{form} using data of more than 4000 crystalline materials including transition metal binary alloys, lanthanide metal and transition metal binary alloys. The prediction of E_{form} achieved a mean absolute error (MAE) of 70 meV atom^{-1}. Ye *et al.* [6] used the Pauling electronegativity [7,8] and ionic radii [9] of the constituent species as the input descriptors with artificial neural networks (ANNs), and obtained a model with extremely low MAEs of 7–10 and 20–34 meV atom^{-1} in predicting the E_{form} of garnets and perovskites, respectively. Ward *et al.* [10] built a machine learning (ML) decision tree model by training with a dataset of inorganic compounds taken from the Open Quantum Materials Database (OQMD) [11], and achieved an MAE of 80 meV atom^{-1} in cross validation for predicting E_{form} by employing structural descriptors from the Voronoi tessellation of the structure of crystalline. Xie and Grossman [12] proposed a generalized graph convolutional neural networks (CGCNN) model for material property predictions, and their model for the E_{form} prediction reached an MAE of 39 meV atom^{-1} with the dataset of 28,046 samples taken from the Materials Project [13]. Li *et al.* [14] proposed to use deep neural network-based transfer learning and a set of hybrid descriptors for perovskite E_{form} prediction [15].

If we look at the existing models closely, nearly all the published papers trim their training dataset in some way, and those treatments would for sure improve the predicting ability. For example, Ye *et al.* [6] limit their model to the $C_3A_2D_3O_{12}$ garnets and ABO_3 perovskites; the CGCNN paper [12] states that only 28,000 of compounds are used out of ~130,000 data points from Materials Project, making their model not generally applicable, especially for those removed structures; the Roost model [16] uses the "subset contains only the lowest energy polymorph for each stoichiometry" from OQMD database, meaning their model works for the stable compounds only.

Fig. 1 shows the actual predictive power of the artificial intelligence (AI) models by plugging in the entire dataset without data cleaning. All these models are reproduced using the codes, parameters, and data treatment as stated in the original papers, and all the models work as well as they are demonstrated in the paper for the cleaned dataset. But if we incorporate the entire dataset (139,368 structures from Materials Project) as the input to validate the model, the predicting power of the models becomes awful. For example, The MAE of the CGCNN model increases from 39 meV atom^{-1} to 0.137 eV atom^{-1}.

[1] Songshan Lake Materials Laboratory, Dongguan 523808, China
[2] Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China
[3] Tencent AI Lab, Tencent, Shenzhen 518075, China
[4] Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China
[5] Center of Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
* Corresponding authors (emails: mliu@iphy.ac.cn (Liu M); smeng@iphy.ac.cn (Meng S))

 © Science China Press and Springer-Verlag GmbH Germany, part of Springer Nature 2022

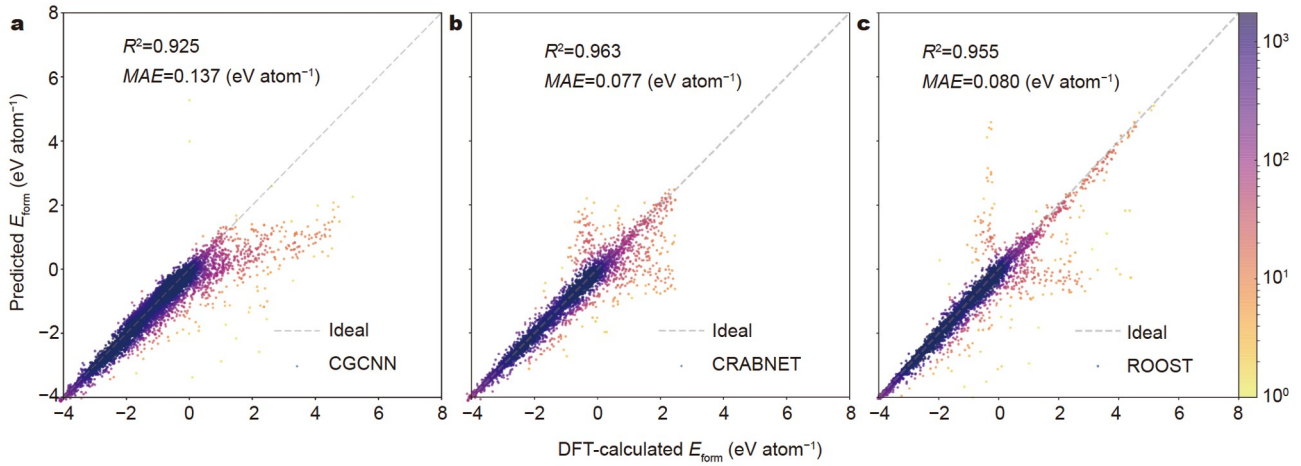

Figure 1 The validation results of the AI models for (a) CGCNN, (b) Crabnet, and (d) Roost when plugging in all the data of 139,368 compounds from Materials Project. The color represents the density of data points, and the darker the denser the points are.

It is indeed a common issue for AI research that the ML model can be tuned with biased data to make it looking good, but it hurts the ability of extrapolation of the model. It suggests that, when we judge the AI models, both the predicting accuracy and the ability of extrapolation matter, and hence both should be validated to make sure the model is feasible and generally applicable. In this work, we build an AI model for predicting the E_{form} of inorganic compounds, and the model has an improved capability of extrapolation compared with several existing models.

Evidence has been demonstrated that AI models can advance materials science significantly [17], e.g., accelerating the materials discovery of energy materials [18,19] and incorporating expert knowledge into the materials design [20]. There are three important things for building a good and reasonable ML model: (1) good data and more data (shown in Fig. S1); (2) good ML algorithm that can react to or pass over the information effectively between neurons; (3) improved descriptors those can extract the intrinsic connections between the properties. In this work, (1) we do not perform any data "cleaning" trick to purposely boost the model predictive power, rather the entire dataset is included for the training process to make sure the model is as universal as possible to cover a large configuration space of inorganic materials; (2) a DenseNet (DN) algorithm is employed to make sure the neurons are directly and effectively connected; (3) to boost the model accuracy, it is found that the electronegativity difference and structural information are indeed the most important descriptors for determining the E_{form} of compounds, and the prediction performance of E_{form} is greatly improved after the addition of these descriptors. Hence, our ML model can reach a fairly good predictive ability up to R^2 = 0.982 and MAE = 0.072 eV atom^{-1} for all the compounds.

METHODS

Models

Eight conventional ML models as well as a neural network model are adopted for predicting the density functional theory (DFT)-calculated formation energy per atom (E_{form}), which are AdaBoost Regression (ABR), Linear Support Vector Regression (LSVR), SVR, Ridge Regression (RR), GradientBoosting Regression (GBR), K-Nearest Neighbors Regression (KNNR), Random Forest Regression (RFR), ExtraTrees Regression (ETR), and DN [21,22]. In this study, the commonly used ML models are implemented by *scikit-learn* [23] and the DN is implemented by *pytorch* [24]. DN is a neural network model, which has three hidden dimensions with node counts of [256, 128, 64], in addition to an input layer with 119 nodes and an output layer with 1 node. The DN model is trained using the MAE loss criterion, or called $L1$ loss criterion, shown in Formula (1) and an Adam optimizer with a learning rate of 10^{-3}.

$$L1 = |f(x) - Y|, \quad (1)$$

where $f(x)$ represents the predicted value and Y represents the actual DFT-calculated value retrieved from the Atomly.net materials-data infrastructure. The predictive performances of models are described as variance (R^2), MAE, root mean squared error (RMSE).

The training of the AI models in this work usually takes ~20 h on a single NVIDIA Tesla V100 card and the prediction of the formation energy for single structures takes less than 1 s.

Descriptors

In order to obtain a useful AI model, several descriptors are constructed for describing a crystal material. In general, there are two distinct types of descriptors: composition-dependent (CD) descriptors and structure-dependent (SD) descriptors. The later incorporate the atomistic structure of a material, hence making the model geometry-sensitive. The CD descriptors are semantically extracted from the chemical formula of a crystal material based on the elemental properties, like the descriptors proposed by Magpie [4] and Oliynyk [25]. In this study, 44 different elemental properties (listed on Table S1) have been chosen for generating CD descriptors. To each elemental property, six kinds of statistical quantities (average, variance, range, skewness, kurtosis, and sum) are derived as CD descriptors, ending up with 264 CD descriptors.

For the SD descriptors, beside density and band gap, we also propose some new/in-house SD descriptors related to the chemical bond (*CB*), coordination number (*CN*), and electronegativity difference ($\Delta\chi$), which are all extracted from the local environment of each atom in a structure. The density is calculated from the crystal structure *via* the Pymatgen [19], and the band gap was directly retrieved from the "atomly.net" database.

© Science China Press and Springer-Verlag GmbH Germany, part of Springer Nature 2022

The local environment of an atom represents the interactions between the atom and its nearest neighbors, and directly affects the E_{form} of compounds. The nearest neighbors of each atom in a structure are extracted by using the Voronoi tessellation method [10], which has also been used for generating many other descriptors related to the structure, e.g., the aforementioned OFM descriptors [5]. By using this method, the real space can be divided into a number of convex polyhedrons by constructing vertical bisectors between neighboring atoms, similar to the Wigner-Seitz cell, as shown in Fig. 2a. One can see that there is only one atom in each polyhedron, whose nearest neighbors are represented by the surface of the polyhedron.

Similar to the CD descriptors, we also use six statistical quantities to generate the new/in-house SD descriptors. That is to say, there are six *CB* SD descriptors (CB^a, CB^v, CB^r, CB^s, CB^k, CB^u) which can be calculated as follows:

$$CB^{a} = \frac{1}{N}\sum_{i=1}^{N} CB_i, \tag{2}$$

$$CB^{v} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(CB_i - \frac{1}{N}\sum_{i=1}^{N} CB_i\right)^2}, \tag{3}$$

$$CB^{r} = \max(CB_i) - \min(CB_i), \tag{4}$$

$$CB^{s} = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(CB_i - \frac{1}{N}\sum_{i=1}^{N} CB_i\right)^3}{\left(\frac{1}{N}\sum_{i=1}^{N}\left(CB_i - \frac{1}{N}\sum_{i=1}^{N} CB_i\right)^2\right)^{\frac{3}{2}}}, \tag{5}$$

$$CB^{k} = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(CB_i - \frac{1}{N}\sum_{i=1}^{N} CB_i\right)^4}{\left(\frac{1}{N}\sum_{i=1}^{N}\left(CB_i - \frac{1}{N}\sum_{i=1}^{N} CB_i\right)^2\right)^{2}}, \tag{6}$$

$$CB^{u} = \sum_{i=1}^{N} CB_i, \tag{7}$$

where N is the number of atoms in the structure and CB_i is the average of the chemical bonds of the i^{th} atom with its nearest neighbors. As shown in Fig. 2b, the *CB* between the i^{th} atom and its j^{th} nearest neighbor is expressed in the spherical coordinates as $(R, \theta, \varphi)_{ij}$, which is represented by the yellow vector $CB_{i,j}$. And CB_i is defined by

$$CB_i = \frac{1}{n_i}\sum_{j=1}^{n_i} CB_{i,j} = \frac{1}{n_i}\sum_{j=1}^{n_i}(R, \theta, \varphi)_{ij}, \tag{8}$$

where n_i represents the number of neighbors of the i^{th} atom. Fig. 2c shows the coordination number CN_i of the i^{th} atom counted *via* the Voronoi tessellation. The six *CN* SD descriptors (CN^a, CN^v, CN^r, CN^s, CN^k, CN^u) related to statistical quantities are calculated with the formulas similar to those of the *CB* SD descriptors.

The strength of a chemical bond should depend on the electronegativity difference ($\Delta\chi$) between the two bonding atoms. Usually, for ionic compounds, the bigger the electronegativity difference, the stronger the bond. The $\Delta\chi$ between atoms bonded together will lead to the charge transfer from one to another, greatly affecting the charge density distribution of the local

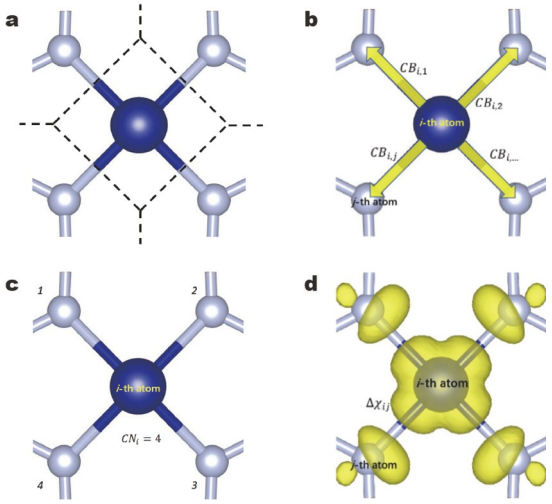

Figure 2 (a) The nearest neighbors of the i^{th} atom found by the Voronoi tessellation method. The construction of the SD descriptors: (b) *CB*; (c) *CN*; (d) $\Delta\chi$.

structure, and thereby the E_{form} of a system will be affected because the total energy would be a functional of charge density [26]. Therefore, the $\Delta\chi$ is essential to the prediction of E_{form}, and thus we propose several SD descriptors of $\Delta\chi$. Fig. 2d describes the electron distribution caused by the $\Delta\chi$ between bonding atoms. Here, we have chosen five different electronegativity definitions, Allred and Rochow electronegativity (AR χ), Pauling electronegativity (Pa χ), Martynov and Batsanov electronegativity (MB χ), Gordy electronegativity (Go χ) and Mullikan electronegativity (Mu χ), to generate descriptors related to the $\Delta\chi$. To each electronegativity definition, the SD descriptors are calculated by the following steps: firstly, the absolute value of the $\Delta\chi$ between the i^{th} atom and its j^{th} neighbor ($\Delta\chi_{i,j}$) is given by

$$\Delta\chi_{i,j} = \left|\chi_i - \chi_j\right|, \tag{9}$$

where χ_i and χ_j are the electronegativity of the i^{th} atom and its j^{th} neighbor, respectively. Then, the $\Delta\chi_i$ of the i^{th} atom with its neighbors are calculated as the calculation as follows:

$$\Delta\chi_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\Delta\chi_{i,j}, \tag{10}$$

where n_i denotes the number of neighbors of the i^{th} atom. Finally, to each electronegativity definition, the six electronegativity difference SD descriptors ($\Delta\chi^a$, $\Delta\chi^v$, $\Delta\chi^r$, $\Delta\chi^s$, $\Delta\chi^k$, $\Delta\chi^u$) for the whole structure are calculated similarly to the aforementioned calculation method of the *CB* SD descriptors.

In the end, we got another set of descriptors which consists of the 234 CD descriptors and the 44 SD descriptors based on 9 different structure-related properties as detailed in Table S2. The 30 CD descriptors with respect to 5 kinds of electronegativity definitions are removed and replaced by the corresponding SD descriptors related to the electronegativity definitions. Although the treatments to build the descriptors, such as the Voronoi tessellation, are well-known, those descriptors of electronegativity difference in this work are the newly invented, and the following paper will show that they are fairly important.

Data and data availability

The dataset used in this work is obtained from the "atomly.net"

 © Science China Press and Springer-Verlag GmbH Germany, part of Springer Nature 2022

[27], which is an-access DFT database. The details of the data can be found in Supplementary information. The codes that are used to generate and validate the models can be found at https://github.com/atomly-materials-research-lab/Descriptor.

RESULTS AND DISCUSSION

Nine ML models are built to predict E_{form}, and their performance are benchmarked by comparing their cross-validated coefficients of determination (R^2), RMSE, and MAE values of the test dataset, which is the entire database without data cleaning (the optimized parameters of the eight conventional ML models are listed in Table S3). The top 20 important descriptors are screened out by using the RFR method, and the correlations between each descriptor pair are studied by the Pearson correlation coefficient.

It is found that the newly proposed SD descriptor along with the entire dataset (without data cleaning) can significantly optimize the models, making it much better than the models in Fig. 1. According to the validation results illustrated in Fig. 3, the DN model (Fig. 3i) gives the best performance with R^2 = 0.982, while the RR method (Fig. 3a) has the lowest prediction accuracy with R^2 = 0.858. Other four classical ML models, GBR (Fig. 3e), ABR (Fig. 3f), ETR (Fig. 3g), and SVR (Fig. 3h), can also achieve a very high prediction accuracy of E_{form} with R^2 > 0.95. For the two tree-based models, the ETR with R^2 = 0.961 is better than the RFR (Fig. 3d) with R^2 = 0.945. As most classical boosting algorithms, both ABR (with extremely randomized trees as the weak learners) and GBR (with decision trees as the weak learners) can implemented very high R^2 = 0.957 and 0.952, respectively. The R^2 value of the KNNR model is close to 0.9 as illustrated in Fig. 3c. Comparing Fig. 3b and h, it is found that using the SVR model with the nonlinear RBF (radial basis function) kernel function can get a better prediction performance than that with the linear kernel function.

With the objective to estimate the effect of the new SD descriptors proposed in this study on predicting the E_{form} of

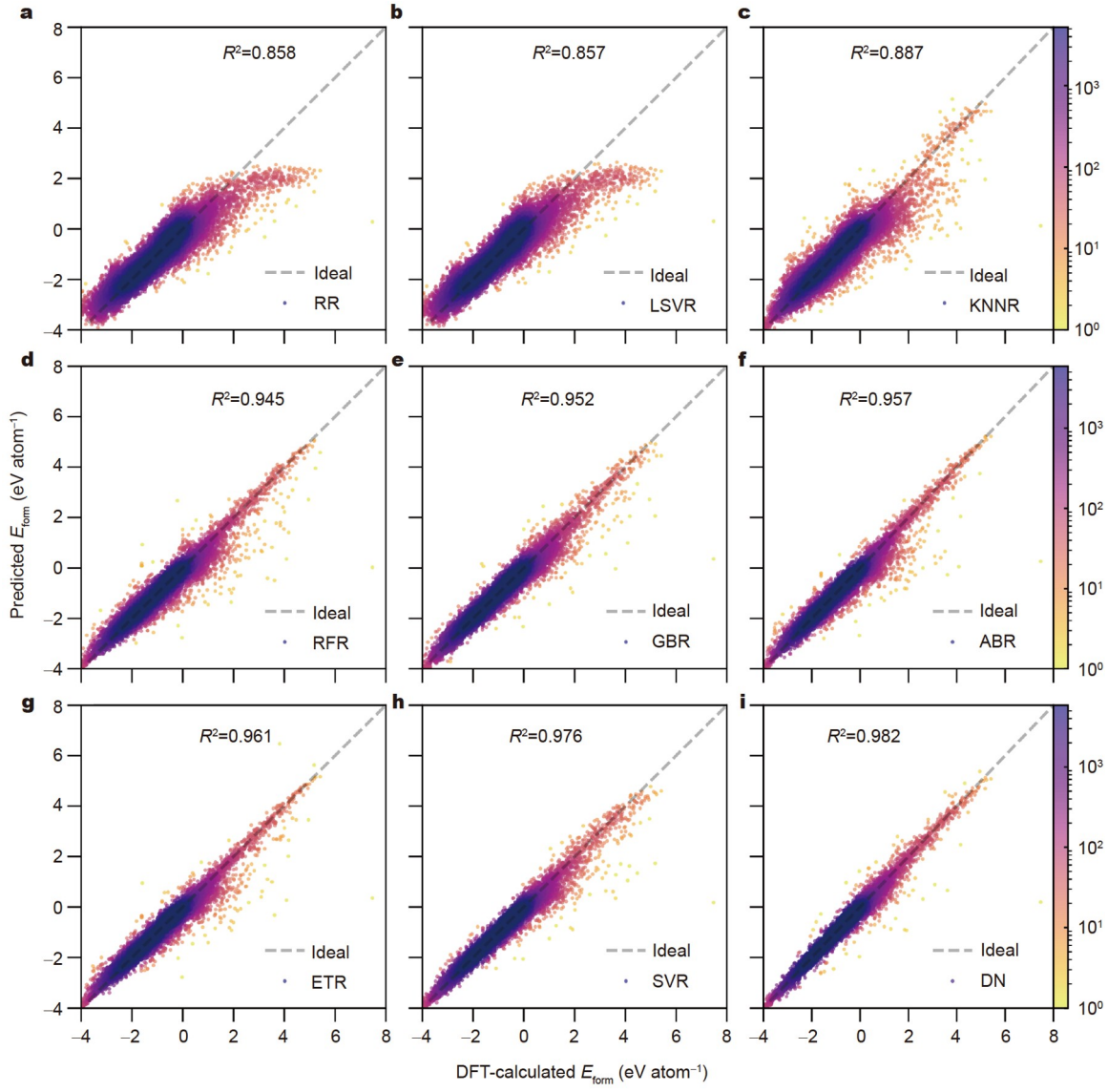

Figure 3 The DFT-calculated *vs.* predicted scatter plots of E_{form} for the nine different ML methods with the new set of descriptors including both the CD and SD descriptors: (a) RR; (b) LSVR; (c) KNNR; (d) RFR; (e) GBR; (f) ABR; (g) ETR; (h) SVR; (i) DN. The gray dash line in each figure represents the ideal curve $y = x$. The color represents the density of data points.

© Science China Press and Springer-Verlag GmbH Germany, part of Springer Nature 2022

crystals, we investigated the changes in evaluation metrics (R^2, RMSE, and MAE) of the ML models with and without the SD descriptors. As shown in Fig. 4 and Fig. S2, the prediction performances of E_{form} for all the nine ML models have been improved by increasing the SD descriptors, which indicate that the local structure information describes E_{form} very well for the inorganic materials. Fig. 4a depicts that the increase rate of R^2 values varies from 6.3% to 17.0% in combination of the SD descriptors with the CD ones. The SD descriptors dramatically improve the R^2 value of the ABR model for E_{form} from 0.829 to 0.957. For DN model, which has the highest predictive power, the R^2 value increase from 0.885 to 0.982, which is 11.0% increase. The changes of RMSE and MAE values are shown in Fig. 4b, c. One can see that the decreasing rates of RMSE and MAE values fluctuate in the range of [20.5%, 60.4%] and [10.0%, 50%], respectively. These results prove that the newly proposed SD descriptors, in particular the electronegativity difference of neighboring atoms, can significantly improve the prediction performances of E_{form} in the inorganic materials.

It is also observed that the electronegativity difference, unlike the average electronegativity employed in previous studies, indeed improves the model significantly. To analyze the control factors of E_{form}, the RFR method is chosen for ranking the importance of different descriptors, since the RFR method is a tree-based learning algorithm and has advantages on both accuracy and interpretability [28]. Fig. 5 shows the top 20 descriptors ranked for E_{form} prediction, and the correlation between descriptor pair in the way of Pearson correlation coefficient matrix. It is found that the Pauling electronegativity is the most important factor for the prediction model of E_{form} without and with SD descriptors (47.1% of Pa χ^r in Fig. 5a and 46.4% of ΔPa χ^a in Fig. 5b, respectively). Although the ratio of ΔPa χ^a slightly drops when using the SD descriptors, the overall ratio of the top 20 important descriptors increases from 74.6% to 82.6%. In Fig. 5a, b, there are five descriptors (Pa χ^r, Pa χ^k, AR χ^k, AR χ^s, Mu χ^k) related to electronegativity and two descriptors (ΔPa χ^a, ΔAR χ^a) related to "electronegativity difference" appearing in the top 20 important descriptors, respectively. Metallic valence represents the oxidation state of the metal in alloy similarly to that of covalent compound. In the previous studies for predicting the thermodynamic stability of perovskites, the most common oxidation state is one of the most important descriptors. The three descriptors (MV^s, MV^v, MV^r) related to the metallic valence [29,30] appear in the 20 important descriptors in both Fig. 5a, b, respectively. Fig. 5b illustrates that the density (ρ), coordination number (CN^a), band gap (E_g) are also important descriptors for improving the prediction performance of E_{form}, which is consistent with the previous results [14,31,32]. In the prior study of Ward *et al.* [4], it is pointed out that the E_{form} of intermetallic compounds is best described by the variances in the melting point (MP^v) and number of d electrons between constituent elements (V_d^v). The similar result is also observed in this study, because there exist linear relationships between the melting point and boiling point. As shown in Fig. 5a, b, the variance of the boiling point (BP^v) as well as the variance of d electrons and unfilled d electrons (V_d^v, UV_d^v) appear in the top 20 important descriptors. The variance of cohesive energy, the skewness and kurtosis of the 1st ionization

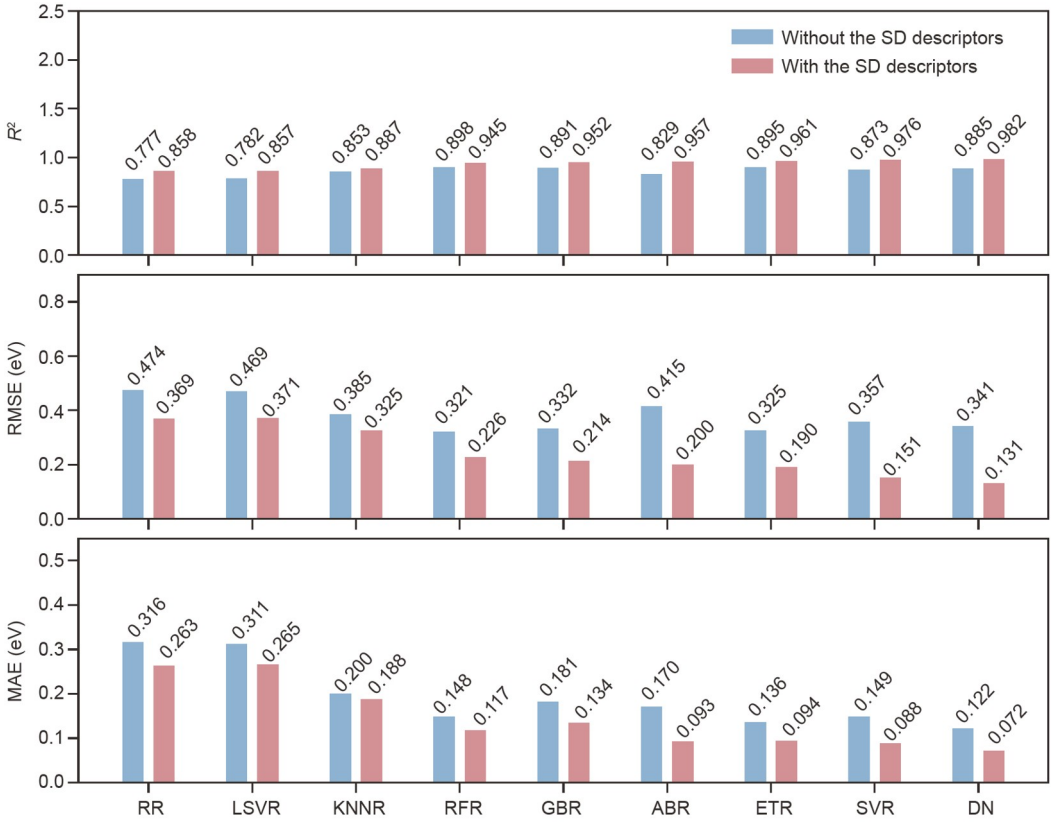

Figure 4 (a) R^2, (b) RMSE, and (c) MAE values of the nine different ML methods for E_{form} prediction with and without the new proposed SD descriptors. The red and blue bars represent the corresponding values before and after adding these descriptors, respectively.

 © Science China Press and Springer-Verlag GmbH Germany, part of Springer Nature 2022

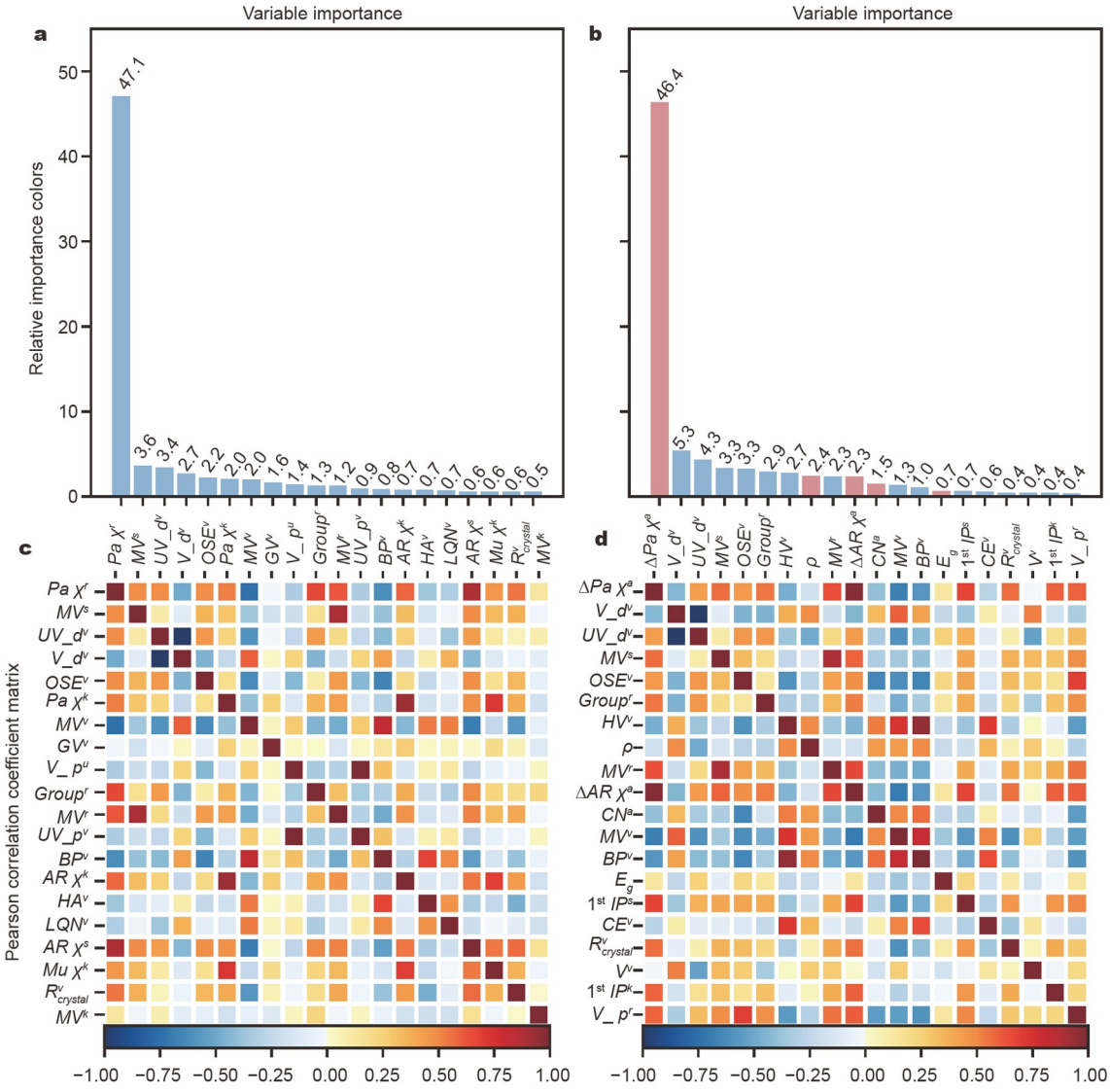

Figure 5 The importance ranking of all descriptors (a) without and (b) with the new SD descriptors in predicting E_{form}. The bars marked by the red color in (b) represent the relative importance of the SD descriptors proposed in this study. Pearson correlation coefficient matrix of (c) without and (d) with the SD descriptors. The lower-case letters a, v, r, s, k, and u labeled in the upper-right corner of *x*- and *y*-axis labels are the abbreviation of "average, variance, range, skewness, kurtosis, and sum".

potential (CE^v, 1^{st} $1P^s$, 1^{st} $1P^k$), also appear in the top 20 important descriptors.

Fig. 5c, d show the Pearson correlation matrix for the top 20 important descriptors without and with the SD descriptor set. For each pair of descriptors, *A* and *B*, the correlation coefficient R_{AB} is defined as

$$R_{AB} = \frac{\sum_{i=1}^{n}(A_i-\overline{A})(B_i-\overline{B})}{\sqrt{\sum_{i=1}^{n}(A_i-\overline{A})}\sqrt{\sum_{i=1}^{n}(B_i-\overline{B})}}, \quad (11)$$

where $\overline{A}$ and $\overline{B}$ are the sample means of *A* and *B* descriptors over the total *n* crystal materials, while A_i and B_i are the descriptors of the i^{th} material. As shown in Fig. 5c, strong correlations are found in several pairs of descriptors: (1) two descriptors related to the d valence electrons orbitals (UV_d^v, V_d^v); (2) three descriptors related to the electronegativity (Pa χ^k, AR χ^k, Mu χ^k); (3) two descriptors related to the thermodynamic property (BP^v, HA^v). Comparing Fig. 5c, d, the changes in the color distributions at the lower right corner (marked by the black rectangles) show that the correlations of the top 20 descriptors decrease obviously by adding the SD descriptors. Our results show that there is a correlation between different definitions of electronegativity. As shown in Figs S3 and S4, when the most important descriptor ΔPa χ^a in Fig. 5b is removed from the collection of the input descriptors, the ΔAR χ^a descriptor becomes the most important descriptor after ranking again by the RFR method. Repeatedly with the above steps, when removing ΔAR χ^a, the ΔAR χ^u becomes the most important descriptor, which is consistent with the observed result in Fig. 5d that there exist strong correlations among the descriptors (ΔPa χ^a, ΔAR χ^a) and the relation on prediction performance of E_{form} follows the trend ΔPa χ^a > ΔAR χ^a > ΔAR χ^u.

Then, the E_{form} of compounds can be predicted with the constructed models. Fig. 6 shows the thermodynamic phase diagram for compounds in Ti-O, V-O, Mn-O and Li-P chemical systems obtained from the DN model and DFT. These four

© Science China Press and Springer-Verlag GmbH Germany, part of Springer Nature 2022

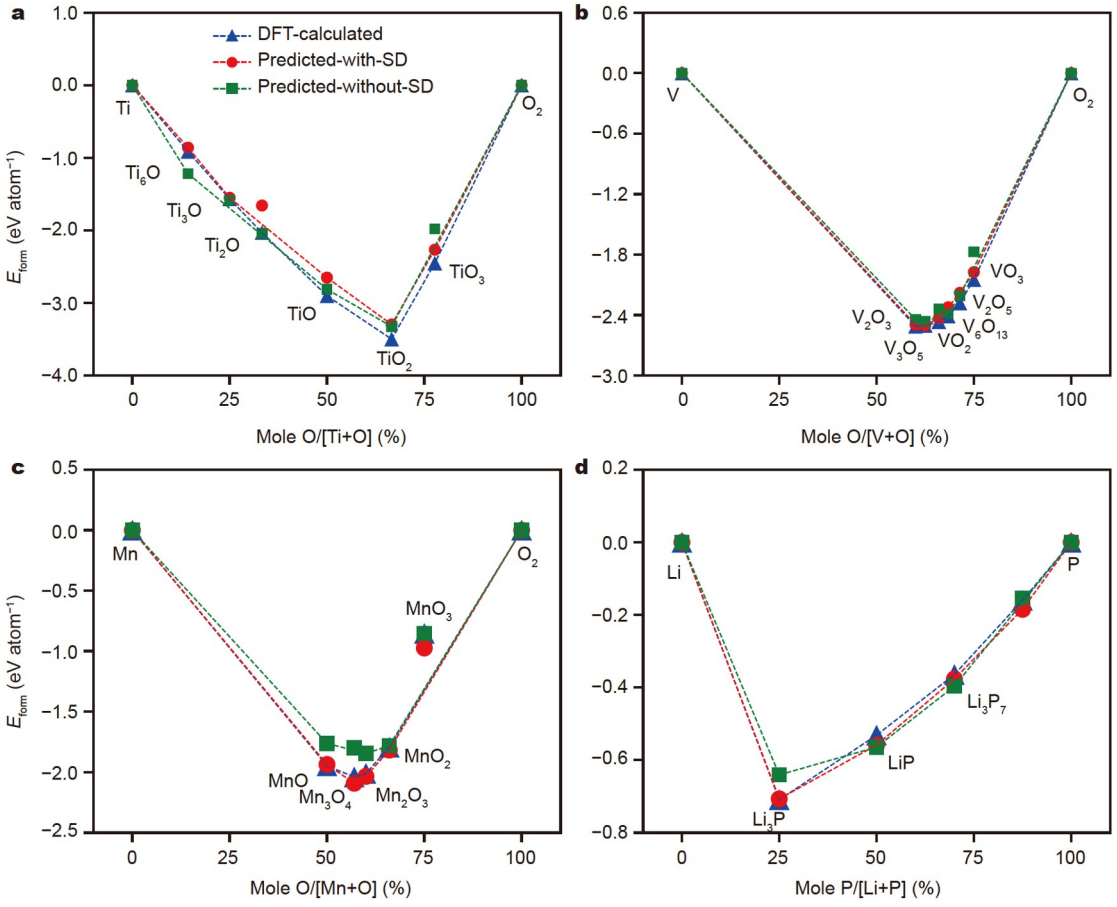

Figure 6 The thermodynamic phase diagram for compounds consisted in (a) Ti-O, (b) V-O, (c) Mn-O, and (d) Li-P chemical systems. The blue triangles, red circles and green rectangles indicate the DFT-calculated E_{form}s, predicted ones with and without the SD descriptors proposed in this study, respectively. The blue envelopes are made up of the DFT-calculated E_{form}s of the stable structures in each chemical system.

chemical systems are selected to demonstrate the accuracy of the model for their importance, since TiO_x materials are commonly used in photovoltaic devices to improve device performance [33,34], VO_x is a promising cathode material and likely to be commercialized applicable in the future [35], MnO_x is widely studied as catalysts [36–38], and LiP_x is the intermedia product in certain anode reactions of Li-ion batteries [39]. The thermodynamic phase diagrams are generated using the same energy correction mechanism as the Materials Project [1,40], to make sure the energy hulls in this work are comparable to either Materials Project [1,40] or Atomly [27] (also adopts the same energy correction for generating the phase diagram). As shown in Fig. 6a–d, The DFT data tell us that there are 6 stable crystalline compounds (Ti_6O, Ti_3O, Ti_2O, TiO, TiO_2 and TiO_3) in the Ti-O chemical system, 4 stable crystalline compounds (V_2O_3, V_3O_5, VO and VO_2) and 2 unstable ones (V_6O_{13} and VO_3) in the V-O chemical system, 4 stable crystalline compounds (MnO, Mn_3O_4, Mn_2O_3 and MnO_2) and 1 unstable crystalline compound (MnO_3) in the Mn-O chemical system, and 4 stable crystalline compounds (Li_3P, LiP, Li_3P_7 and LiP_7) in the Li-P chemical system, respectively. Note that the DN model can assess the E_{form} of those compounds fairly well. It is seen from Fig. 6c that in the Mn-O chemical system, there is a distinct advantage of the model with the SD descriptors on predictive performance over that without the SD descriptors. The other panels in Fig. 6 show that both models, with or without the SD descriptors, can capture the shape of the energy hull to some extent. When the SD descriptors are taken into account in the model, the predicted E_{form} is off only by ~135 meV atom^{-1} on average (ranging from 0 to 373 meV atom^{-1}) for the Ti-O chemical system, ~40 meV atom^{-1} (ranging from 0 to 102 meV atom^{-1}) for the V-O chemical system, ~33 meV atom^{-1} (ranging from 0 to 120 meV atom^{-1}) for the Mn-O chemical system, and ~11 meV atom^{-1} (ranging from 0 to 27 meV atom^{-1}) for the Li-P chemical system, compared with the DFT values. Hence, our model makes the prediction of E_{form} of inorganic materials easy and accurate.

CONCLUSIONS

In summary, we develop a universal model for predicting the formation energy of compounds as all the data points are employed for training without the "data cleaning". The "data cleaning", as performed by many other literature, makes their model looking good, but worsens the capability of extrapolation of the model. We have also identified several new SD descriptors related to the electronegativity difference and coordination numbers for predicting the formation energy of a crystal material, derived from the local environment of the material *via* the Voronoi tessellation method. We demonstrate that these new descriptors can significantly improve the prediction accuracy of formation energy not only for eight classical ML methods but also for another neural network method. The neural network

 © Science China Press and Springer-Verlag GmbH Germany, part of Springer Nature 2022

model achieves the high prediction accuracy of R^2 = 0.982 and MAE = 0.072 eV atom^{-1}. The tree-based RFR method is chosen for screening the key descriptors for best describing the formation energy. It is found that the Pauling electronegativity difference between bonding atoms is the most important factor with a ratio of 46.4% for the prediction of formation energy. Our work shows that by adopting the physics-based descriptor as well as a good dataset, the predictive power of ML models can be significantly improved. There might be still a large room out there to keep enhancing the predictive power, if we can find better descriptors with better data. We expect that the model presented here can provide a useful tool for materials science community immediately towards extremely fast, large-scale, and accurate materials modelling and prediction without costly computations.

Received 2 February 2022; accepted 24 May 2022; published online 27 July 2022

1 Ong SP, Wang L, Kang B, *et al.* Li-Fe-P-O_2 phase diagram from first principles calculations. Chem Mater, 2008, 20: 1798–1807
2 Miura A, Bartel CJ, Goto Y, *et al.* Observing and modeling the sequential pairwise reactions that drive solid-state ceramic synthesis. Adv Mater, 2021, 33: 2100312
3 Cao Z, Dan Y, Xiong Z, *et al.* Convolutional neural networks for crystal material property prediction using hybrid orbital-field matrix and Magpie descriptors. Crystals, 2019, 9: 191
4 Ward L, Agrawal A, Choudhary A, *et al.* A general-purpose machine learning framework for predicting properties of inorganic materials. npj Comput Mater, 2016, 2: 16028
5 Lam Pham T, Kino H, Terakura K, *et al.* Machine learning reveals orbital interaction in materials. Sci Tech Adv Mater, 2017, 18: 756–765
6 Ye W, Chen C, Wang Z, *et al.* Deep neural networks for accurate predictions of crystal stability. Nat Commun, 2018, 9: 3800
7 Pauling L. The nature of the chemical bond. IV. The energy of single bonds and the relative electronegativity of atoms. J Am Chem Soc, 1932, 54: 3570–3582
8 Allred AL. Electronegativity values from thermochemical data. J InOrg Nucl Chem, 1961, 17: 215–221
9 Shannon RD. Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. Acta Cryst A, 1976, 32: 751–767
10 Ward L, Liu R, Krishna A, *et al.* Including crystal structure attributes in machine learning models of formation energies *via* Voronoi tessellations. Phys Rev B, 2017, 96: 024104
11 Kirklin S, Saal JE, Meredig B, *et al.* The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies. npj Comput Mater, 2015, 1: 15010
12 Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys Rev Lett, 2018, 120: 145301
13 Jain A, Ong SP, Hautier G, *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. APL Mater, 2013, 1: 011002
14 Li X, Dan Y, Dong R, *et al.* Computational screening of new perovskite materials using transfer learning and deep learning. Appl Sci, 2019, 9: 5510
15 Ong SP, Richards WD, Jain A, *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. Comput Mater Sci, 2013, 68: 314–319
16 Goodall REA, Lee AA. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. Nat Commun, 2020, 11: 6280
17 Xu Y, Liu X, Cao X, *et al.* Artificial intelligence: A powerful paradigm for scientific research. Innovation, 2021, 2: 100179
18 Liu Y, Zhao T, Ju W, *et al.* Materials discovery and design using machine learning. J Materiomics, 2017, 3: 159–177
19 Liu Y, Guo B, Zou X, *et al.* Machine learning assisted materials design and discovery for rechargeable batteries. Energy Storage Mater, 2020, 31: 434–450
20 Liu Y, Wu JM, Avdeev M, *et al.* Multi-layer feature selection incorporating weighted score-based expert knowledge toward modeling materials with targeted properties. Adv Theor Simul, 2020, 3: 1900215
21 Wang AYT, Murdock RJ, Kauwe SK, *et al.* Machine learning for materials scientists: an introductory guide toward best practices. Chem Mater, 2020, 32: 4954–4965
22 Wang AYT, Kauwe SK, Murdock RJ, *et al.* Compositionally restricted attention-based network for materials property predictions. npj Comput Mater, 2021, 7: 77
23 Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: Machine learning in Python. J Mach Learn Res, 2011, 12: 2825–2830
24 Paszke A, Gross S, Chintala S, *et al.* Automatic differentiation in PyTorch. NIPS 2017 Workshop Autodiff Decision Program Chairs, 2017
25 Oliynyk AO, Antono E, Sparks TD, *et al.* High-throughput machine-learning-driven synthesis of full-Heusler compounds. Chem Mater, 2016, 28: 7324–7331
26 Parr RG, Donnelly RA, Levy M, *et al.* Electronegativity: The density functional viewpoint. J Chem Phys, 1978, 68: 3801–3807
27 Atomly. Available from: https://atomly.net
28 Xu YF, Rao HS, Wang XD, *et al. In situ* formation of zinc ferrite modified Al-doped ZnO nanowire arrays for solar water splitting. J Mater Chem A, 2016, 4: 5124–5129
29 Snow AI. Metallic valences. J Chem Phys, 1950, 18: 233
30 Pauling L. Atomic radii and interatomic distances in metals. J Am Chem Soc, 1947, 69: 542–553
31 Im J, Lee S, Ko TW, *et al.* Identifying Pb-free perovskites for solar cells by machine learning. npj Comput Mater, 2019, 5: 37
32 Emery AA, Wolverton C. High-throughput DFT calculations of formation energy, stability and oxygen vacancy formation energy of ABO_3 perovskites. Sci Data, 2017, 4: 170153
33 Park JS, Jeong JK, Mo YG, *et al.* Impact of high-k TiO_x dielectric on device performance of indium-gallium-zinc oxide transistors. Appl Phys Lett, 2009, 94: 042105
34 Wang DH, Im SH, Lee HK, *et al.* Enhanced high-temperature long-term stability of polymer solar cells with a thermally stable TiO_x interlayer. J Phys Chem C, 2015, 113: 17268–17273
35 Liu P, Zhu K, Gao Y, *et al.* Recent progress in the applications of vanadium-based oxides on energy storage: From low-dimensional nanomaterials synthesis to 3D micro/nano-structures and free-standing electrodes fabrication. Adv Energy Mater, 2017, 7: 1700547
36 Kijlstra WS, Brands DS, Smit HI, *et al.* Mechanism of the selective catalytic reduction of NO with NH_3 over MnO_x/Al_2O_3. J Catal, 1997, 171: 219–230
37 Qi G, Yang RT, Chang R. MnO_x-CeO_2 mixed oxides prepared by co-precipitation for selective catalytic reduction of NO with NH_3 at low temperatures. Appl Catal B-Environ, 2004, 51: 93–106
38 Wu Z, Jiang B, Liu Y, *et al.* Experimental study on a low-temperature SCR catalyst based on MnO/TiO_2 prepared by sol-gel method. J Hazard Mater, 2007, 145: 488–494
39 Marino C, Boulet L, Gaveau P, *et al.* Nanoconfined phosphorus in mesoporous carbon as an electrode for Li-ion batteries: Performance and mechanism. J Mater Chem, 2012, 22: 22713–22720
40 Jain A, Hautier G, Ong SP, *et al.* Formation enthalpies by mixing GGA and GGA + U calculations. Phys Rev B, 2011, 84: 045115

Acknowledgements The computational resource is provided by the Platform for Data-Driven Computational Materials Discovery of Songshan Lake laboratory. We especially thank Atomly database for data sharing. We would also acknowledge the financial support from the Chinese Academy of Sciences (CAS-WX2021PY-0102, ZDBS-LY-SLH007, and XDB33020000).

Author contributions Liang Y and Liu M conceived the idea, designed the experiments, and wrote the paper with support from Meng S and Wang Z; Liang Y, Chen M, Wang Y, Jia H, Lu T, and Xie F prepared the data, built the database, wrote the machine learning code, and analyzed the training as well

 © Science China Press and Springer-Verlag GmbH Germany, part of Springer Nature 2022

as testing results. All authors contributed to the general discussion.

Conflict of interest The authors declare that they have no conflict of interest.

Supplementary information Supporting data are available in the online version of the paper.

Yingzong Liang is a senior engineer at Songshan Lake Materials Laboratory. He received his PhD degree in systems innovation from the University of Tokyo in 2017 and worked as a postdoctoral fellow at the Institute of Physics, Chinese Academy of Sciences (CAS) from 2019 to 2021. His current occupancy is a senior IC design engineer at Beijing Smartchip Microelectronics Technology Co., Ltd. and he is an expert in data science, material science, device physics as well as IC design.

Sheng Meng is a professor of physics at the Institute of Physics, CAS since 2009, and the director of Platform for Data-Driven Computational Materials Discovery at Songshan Lake Materials Laboratory. His research interests focus on excited state quantum dynamics in condensed matters, energy conversion mechanism for sustainable society, as well as new algorithms and tools in materials computation.

Miao Liu is an associate professor at the Institute of Physics, CAS, and Songshan Lake Materials Laboratory. His research revolves around data-driven materials science for energy materials, alloys, quantum materials, etc. He is the founder of the atomly.net materials science database.

一种预测无机晶体形成能的高精度泛化模型

梁英宗[1,2], 陈明威[1], 王亚南[1,2], 贾华昱[2,3], 芦腾龙[2], 谢帆恺[2], 蔡光辉[2], 王宗国[4], 孟胜[1,2*], 刘淼[1,2,5*]

摘要 随着数据科学和材料科学的进步，人们如今可构建出较为准确的人工智能模型，用于材料性质预测. 本文中，我们以170,714个无机晶体化合物的高通量第一性原理计算数据集为基础，训练得到了可精确预测无机化合物形成能的机器学习模型. 相比于同类工作，本项研究以超大数据集为出发点，构建出无机晶体形成能的高精度泛化模型，可外推至广阔相空间，其中的DenseNet神经网络模型精度可以达到R^2 = 0.982和平均绝对误差(MAE) = 0.072 eV atom^{-1}. 上述模型精度的提升源自一系列新型特征描述符，这些描述符可有效提取出原子与领域原子间的电负性和局域结构等信息，从而精确捕捉到原子间的相互作用. 本文为新材料搜索提供了一种高效、低成本的结合能预测手段.

 © Science China Press and Springer-Verlag GmbH Germany, part of Springer Nature 2022