npj | Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

A strategy to apply machine learning to small datasets in materials science

Ying Zhang[1] and Chen Ling[1]

There is growing interest in applying machine learning techniques in the research of materials science. However, although it is recognized that materials datasets are typically smaller and sometimes more diverse compared to other fields, the influence of availability of materials data on training machine learning models has not yet been studied, which prevents the possibility to establish accurate predictive rules using small materials datasets. Here we analyzed the fundamental interplay between the availability of materials data and the predictive capability of machine learning models. Instead of affecting the model precision directly, the effect of data size is mediated by the degree of freedom (DoF) of model, resulting in the phenomenon of association between precision and DoF. The appearance of precision–DoF association signals the issue of underfitting and is characterized by large bias of prediction, which consequently restricts the accurate prediction in unknown domains. We proposed to incorporate the crude estimation of property in the feature space to establish ML models using small sized materials data, which increases the accuracy of prediction without the cost of higher DoF. In three case studies of predicting the band gap of binary semiconductors, lattice thermal conductivity, and elastic properties of zeolites, the integration of crude estimation effectively boosted the predictive capability of machine learning models to state-of-art levels, demonstrating the generality of the proposed strategy to construct accurate machine learning models using small materials dataset.

npj Computational Materials (2018)4:25 ; doi:10.1038/s41524-018-0081-z

INTRODUCTION

In the past few decades the substantial advancement of machine learning (ML) has spanned the application of this data driven approach throughout science, commerce, and industry.[1] Recently, there has been an increasing interest in applying ML to solve problems in materials science.[2–7] In particular, ML techniques have been used to represent inorganic materials,[8–10] predict fundamental properties,[11–13] create atomic potential,[14–16] identify functional candidates,[17–21] analyze complex reaction networks,[22] and guide experimental design.[23–27] The key ingredient behind these successes is that the behavior in unknown domains can be accurately estimated by quantitatively learning the pattern from sufficient training examples. However, compared to others fields the materials data are typically much smaller and sometimes more diverse,[12] which undoubtable affects the construction of ML models. For example, Faber et al.[28] found the predictive accuracy of the ML model for the formation energy of Elpasolite compounds showed a systematic improvement with increasing training set size. Schmidt et al.[29] reported that the predicting error for the formation energy of perovskite compounds decreased monotonically with the size of training set following a power law, where doubling the training set decreased the error by around 20%. Lee et al. examined the ML models for band gaps of inorganic compounds and found the predicting accuracy converged for the ordinary least-square regression and LASSO models at certain sizes of training set, while for the support vector machine model the error still slowly decreased at the largest dataset in their study. While these studied unambiguously demonstrated that the less availability of training data not only renders the detection of patterns more difficult but also deteriorates the capability of making prediction in the unexplored domain, the role of materials dataset in constructing ML model has not been systematically investigated to the best of our knowledge. As a result, the possibility to establish accurate predictive rules using small available materials datasets remains unclear.

It is the focus of current work to comprehensively analyze the interplay between the availability of materials data and the predictive capability of ML models. Our study revealed an important phenomenon when the model is trained using limited available materials data: the association between the degree of freedom (DoF) of model and the precision of prediction, that is, the increase of precision is at the cost of higher DoF. Originated from the statistical bias-variance tradeoff, the appearance of precision–DoF association restricts the accuracy of prediction in unknown domains. We also propose a solution to improve the accuracy without causing higher DoF by incorporating the crude estimation of property (CEP) in the feature space. In three case studies, the integration of crude estimation effectively improved the predictive accuracy of ML models, demonstrating the generality of the proposed strategy to construct accurate ML models using small materials data.

RESULTS

Precision–DoF association

We started with a survey of reported ML models of materials properties,[12,13,18,28–33] focusing on the accuracy of prediction in

[1]Toyota Research Institute of North America, 1555 Woodridge Avenue, Ann Arbor, MI 48105, USA
Correspondence: Chen Ling (chen.ling@toyota.com)

Received: 19 December 2017 Revised: 10 April 2018 Accepted: 16 April 2018
Published online: 14 May 2018

unknown domains. Such a property is usually quantified by evaluating the predicting error by means of cross-validation (CV). To compare models trained to predict different materials properties, the CV errors were scaled by the spanning range of modeled properties. As seen in Fig. 1, the scaled error was in the range of 1–2% for models trained with 10^3–10^4 samples and increased to 10% and above for models established with ~100–200 examples, in full agreement with the intuition that the predictive accuracy increases with the availability of materials data. The seemly universality of the trend in Fig. 1 is, however, surprising, considering the diversity in the surveyed properties as well as the variety of ML techniques used to construct the models. Fitting the surveyed results led to the empirical power law of scaled error $= 0.67 \times \text{size}^{-0.372}$, agreed well with the decreasing of error to predict the formation energy of perovskite with the training size at a power of -0.297.[29] We acknowledge that the survey was conducted for a few recent studies and the observed universality is certainly subject to more examinations. Nonetheless, the challenge to improve the predicting capability with limited availability of materials data is clearly highlighted.

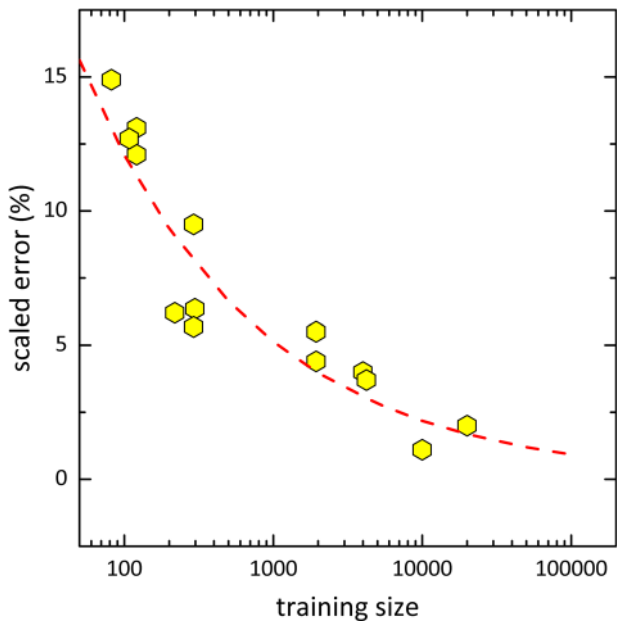

Fig. 1 Survey of scaled error and the size of training set in recent publications. The red dashed line shows the fitted curve of scaled error $= 0.67 \times \text{size}^{-0.372}$

A detailed study was then performed to model the band gaps of binary semiconductors (E_g) as the representative example to understand the effect of data size on the predictive precision. We used a series of manually crafted chemical parameters as possible descriptors, adapting the approach proposed by Ward et al.[34] The optimum set of features that best describes the patterns in the training data was selected using the stepwise forward search. We used the kernel ridge regression (KRR) to construct ML models because it gave the lowest predicting error as benchmarked with other ML algorithms (Supporting Information, Table S1). To manipulate the size of training data, subsets were randomly sampled from the full dataset. Figure 2a shows the average five-fold CV root mean-squared error (RMSE) of KRR models (unless otherwise mentioned, all the errors in the paper were evaluated using the CV method). As expected, the CV-RMSE continuously decreased with the expansion of dataset. Fitting the RMSE with data size gave a power law similar to that in Fig. 1. The smallest CV-RMSE was recorded at 0.51 eV when the full dataset (108 examples) was utilized. Although the prediction had a decent Pearson correlation of 0.94 with the training property, the relatively large error indicated even the "best" model cannot accurately predict E_g. The scaled error of 9.3% agreed well with the trend shown in Fig. 1, further indicating the model did not achieve any predictive capability beyond the observation in the survey.

To understand the origin of the large RMSE, we used the Bootstrap method to break down the contribution of bias and variance to the predicting error,[35,36]

$$\text{bias}^2 = \frac{1}{n_{\text{test}}} \sum_{l=1}^{n_{\text{test}}} \left(\bar{f}(\boldsymbol{x}_l) - y_l\right)^2, \quad (1)$$

$$\text{variance} = \frac{1}{n_{\text{test}}} \sum_{l=1}^{n_{\text{test}}} \frac{1}{B} \sum_{b=1}^{B} \left(f\left(\boldsymbol{x}_l; D^b\right) - \bar{f}(\boldsymbol{x}_l)\right)^2, \quad (2)$$

where n_{test} is the number of testing data, B is the number of training sets sampled from the original training data, y_l is the value of the target property, and $f(x_l;D^b)$ is the property predicted by the model using the training set of D_b. $\bar{f}(\boldsymbol{x}_l)$ is the average of predicted property for example l.

The squared bias was estimated at 0.26 eV^2, over four time to the variance of 0.06 eV^2. A large bias overwhelming small variance suggested that the selected features were not expressive enough to predict the property, or in other words signaled the statistical issue of underfitting. For the model falling in the region of underfitting, the inclusion of more features can effectively mitigate the statistical error. Figure 2b shows the RMSE as a function of the degree of freedom (DoF) of models, defined as the

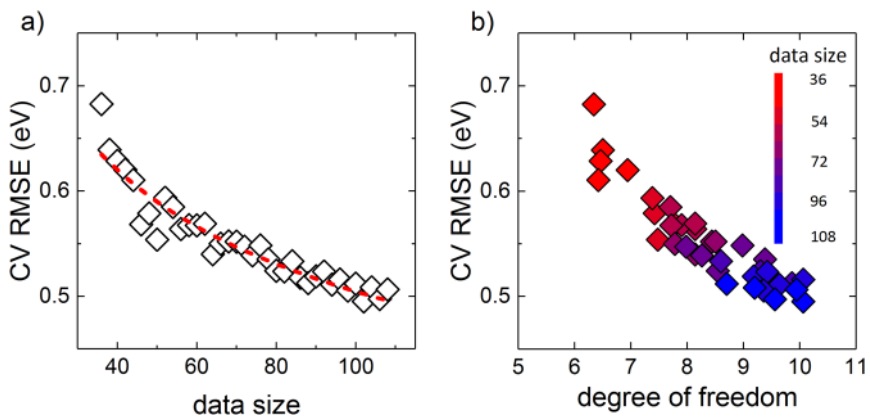

Fig. 2 Effect of data size (n) on the root mean squared error of KRR models for predicting experimental band gaps. **a** Averaged five-fold cross-validation RMSE as a function of data size. The red line shows the fitting curve of RMSE $= 1.42 \times \text{size}^{-0.23}$. **b** Averaged RMSE versus the averaged degree of freedom of KRR models

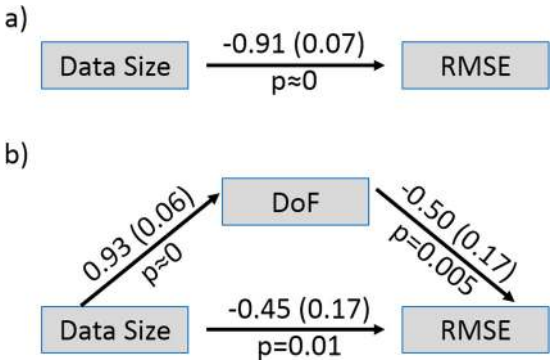

Fig. 3 Mediation analysis on the relationship of data size and precision (RMSE) mediated by model DoF. The top numbers are the standardized regression coefficients and standard errors in the bracket. The bottom numbers show the *p*-value in *t*-test

number of parameters with non-zero regression coefficients.[37] Clearly, both the model precision and DoF exhibited high correlation with training data and the improvement of precision was strongly associated with higher model DoF.

To unveil the underlying relationship between data size, DoF, and precision, a mediation analysis was conducted as illustrated in Fig. 3. In the mediation analysis, three variables are chosen as predictor, outcome and mediator and their relations are explored through statistical significant test.[38] In our study, the predictor variable was data size, the outcome variable was RMSE (precision) and the mediator variable was DoF. All binary associations of RMSE with data size, RMSE with DoF, data size, and DoF were statistically significant (p«0.01, t-test). Entering DoF in the regression greatly reduced the strength of the correlation between RMSE and data size and the p-value in t-test was increased by more than 10 orders of magnitude, confirming that the influence of data size on predictive precision was mediated by model complexity. Therefore, instead of affecting the precision directly, the variation of data size altered the DoF of optimized model, which then changed the accuracy of prediction. Naturally, this mediation effect resulted in the association between DoF and predictive precision as observed in Fig. 2b.

To examine whether the choice of ML method affects the conclusion that the effect of data size on the model precision is mediated by the DoF, we also analyzed the models established with the least absolute shrinkage and selection operator (LASSO) regression. The CV-RMSE to predict E_g using LASSO method was 0.71 eV. The less accuracy was probably attributed to the failure to capture the complicated physics with a linear regression algorithm. Figure 4a shows the selection of LASSO model using varied training sets. The decreasing of RMSE was associated with smaller tuning parameter of LASSO model, λ. In LASSO method, the tuning parameter determines the shrinkage of regression coefficient. A smaller λ applies less penalty for shrinkage, hence permitting the inclusion of more features in the optimized model. As shown in Supporting Information, Figure S1, the association between RMSE and DoF was clearly evidenced. Through the mediation analysis we confirmed that the influence of data size on predictive precision was mediated by the DoF of LASSO models (Supporting Information, Figure S2). These results demonstrated the precision–DoF association as a general statistical phenomenon when the model is trained with small sized materials data rather than a unique observation dependent on the choice of regression method.

These results reveal the influence of materials dataset on establishing ML model as following. Ideally, an ML model should be established to exhibit hidden relations between the property and features determined by the underlying physics. However, in

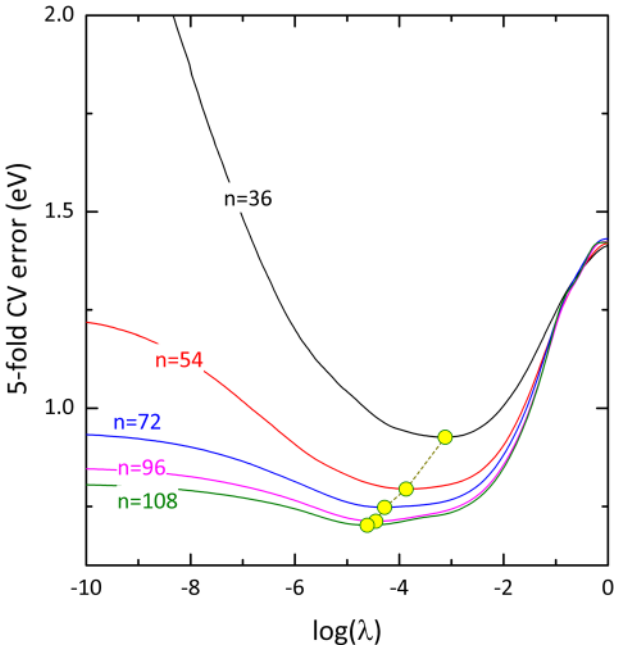

Fig. 4 Effect of data size (n) on the averaged cross-validation RMSE of LASSO models for predicting experimental band gaps. The orange circles is the position of the optimized models

practice, the ML model is constructed to best describe the structure in training data. Any change of internal pattern in training set will lead to the change of ML model. Especially, for models trained with small materials dataset, the DoF to select features is sensitive to the availability of training data. While inadequate selection of expressive features causes the underfitting of property, adding more training data allows the inclusion of more features to alleviate the issue of underfitting. Consequently, the predictive precision is improved with the cost of higher model complexity, resulting in the observed precision–DoF association.

Although it originates from the fundamental statistics, the precision–DoF association is more than just a statistical phenomenon. Because the association occurs as a result of the underfitting, the predicting error is largely dominated by characteristically large bias, which prevents to establish accurate predictive rules. In the above study, even the "best" model showed worse performance than modern density functional theory prediction of E_g. Therefore, the development of *effective strategy to improve model precision without the cost of higher DoF* becomes a crucial challenge to practice ML in modeling materials properties.

Strategy

In principle, the improvement of precision can be approached by appropriately manipulating the training data. For example, we can naturally consider adding more examples to the training set. However, simply expanding the dataset not only leads to highly complex model difficult to interpret the embedded physics but also is likely hindered by the expensive cost to conduct additional experiments. Using the empirical relation established from Fig. 1, doubling the data size roughly leads to the decrease of error by 23%. Hence the exponentially growing cost challenges the feasibility to improve the accuracy by adding new materials data. A model can be also constructed by restricting the configurational space of materials, such as predicting the band gaps of selected families of semiconductors with fixed composition or crystalline structure instead of modeling compounds spanning a wide

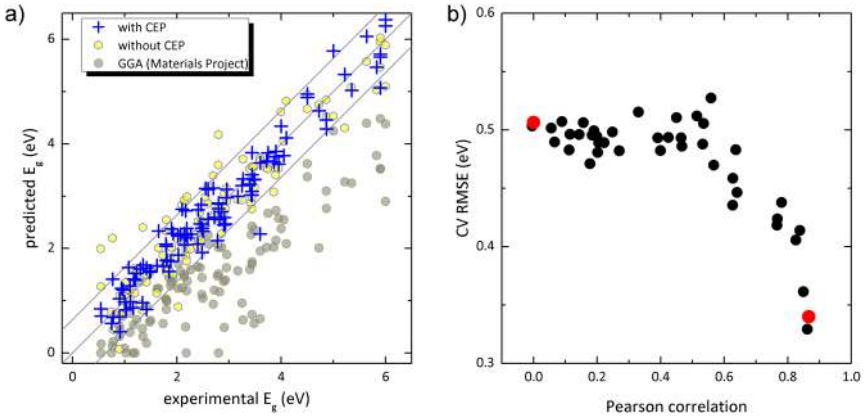

Fig. 5 KRR model to predict E_g using GGA calculated crude estimaiton. **a** Comparison of KRR models using GGA band gap and not using GGA band gap as a descriptor. The GGA-simulated values are also plotted for comparison. **b** Effect of using crude estimation having different correlation with the experimental measurements. The red circles show the positions of the model using and not using the GGA band gap as a descriptor

chemical space.[25,39] As a result, the constructed ML-estimator gained more precision, but sacrificed the generality when applying outside of the confined domain.

Considering the less flexibility of training data, a strategy should be developed from designing appropriate feature space for the modeling, which is widely recognized as a critical step in materials informatics.[30,34,40] Following the analysis in the previous section, let us imagine a simple toy model where the property is explicitly determined by a single feature. In this case, the precision–DoF association should disappear once this special feature is included in the modeling, even if the training data only contain, for example, two samples. Therefore, our intention is to design features to meet the consideration of (1) providing expressive information so that the property can be estimated (though the estimation may not be accurate) and (2) satisfying other requirements such as low dimensionality and cheap cost of acquiring.[30] Based on the consideration 1, we remind ourselves that the prediction of materials property has been carried out centuries before the era of ML. Although the empirical estimation may not be sufficiently accurate in terms of predicting absolute values, it may still provide at least qualitatively knowledge about targeted property. Because any prior knowledge of the targeted property should be considered when constructing the appropriate feature space, we are therefore motived to exercise the idea of using the CEP as a descriptor additional to the chemical descriptors in the ML models.

To give a more precise description of the proposed strategy, we define the CEP as the prediction of targeted property using methods at less accurate level, which includes zero or near zero computational demanding calculations, empirical models, and non-expensive experimental measurements. With this definition, the usage of CEP meets the consideration 2 for the cheap cost of acquiring. Our idea is to incorporate the CEP together with the previously used chemical descriptors to predict targeted property. In previous reports, the PBE-calculated band gaps, which is not accurate but also less computational demanding, were used to predict the band gaps calculated at more accurate and much more expensive level.[41,42] While the success of the earlier work may partially be related to the fact that both the descriptor and property were obtained using similar theory of foundation, we prove in current work that the usage of CEP as a descriptor improves the prediction of experimentally measured properties, where the model performance is strongly affected by the large noise contained in the training set. Furthermore, our study demonstrates that the means to obtain CEP is not restricted to density functional calculations, but can be extended to other non-expensive methods. In the following section, we constructed ML models in three exercises: the prediction of E_g using GGA-calculated values as the CEP, lattice thermal conductivity (κ_L) using empirical models to obtain the CEP, and elastic properties of zeolites using force field calculations to obtain the CEP. In all three studies the ML models achieved state-of-art predictive capability after integrating CEP in the feature space, demonstrating the generality of prosed strategy to construct accurate ML models with small available materials data.

Exercise 1: Band gap of binary semiconductors

We first examined the proposed method in the modeling of E_g. The band gaps simulated at the GGA-level were used as the crude estimation of E_g, the values of which were taken from Materials Project.[43] The performance of the new KRR model is evaluated in Fig. 5a using the leave-one-out cross-validation (LOOCV). In the LOOCV, the prediction of the property at one position is performed by removing that specific observation and using the rest as the training set. The relationship between ML-predicted value and E_g is clearly linear without large deviations even in the range of extremely low or high values. The CV-RMSE of the new ML model was 0.34 eV, decreased by 33% of the model using only chemical descriptors (0.51 eV). The scaled error was reduced to 6.2% after incorporating the GGA band gap in the feature space.

The construction of ML models to predict the band gap of semiconductors was attempted in several reports. Of particular relevance to current work, Lee et al.[41] used the PBE-band gap of inorganic compounds to predict the values from G_0W_0 calculation and achieved the CV error of 0.18 eV. Pilania et al.[42] predicted the band gap of elpasolite ($A_2BB'X_6$-type) compounds calculated at HSE06 level using a multi-fidelity co-kriging statistical learning framework and reported the accuracy of 0.1–0.2 eV on the validation set. At the first glance the error of current ML model seemed to be higher than these two reports. The larger predicting error of the current model can be attributed to two aspects. The first one is the different configuration spaces in the ML models. In the work of Pilania et al.,[42] the composition and crystalline structure was fixed to $A_2BB'X_6$ and elpasolite, respectively, leaving the only variant to be chemical constitutes. In general, the higher DoF in the configuration space, the more challenging to construct an accurate ML model. Another source of error came from the noise of measurement in the training data. While both Lee et al. and Pilania et al. employed values from first-principles calculations as training data with the noise of calculation determined by the accuracy of density functional theories, the current work modeled the experimental E_g measured by various techniques, in which the

uncertainty of measurement was expected to have larger contribution to the predictive error. We continue to discuss the contribution of experimental noise to predicting error in the next case study. To benchmark the performance of ML model, the predicted band gap was compared with that from GW calculation for 49 compounds.[44] While both GW simulation and ML exhibited reasonable predictive accuracy, the ML model showed smaller RMSE of 0.39 eV than that of GW simulation (0.52 eV; Supporting information, Figure S4), quantitatively demonstrating the predictive capability of the ML model.

Having established the predictive precision of the ML model, we now discuss the mechanism of the apparent improved performance after integrating GGA band gap in the feature space. Because of the well-known underestimation of band gap in GGA calculation, the improvement of ML-prediction cannot be attributed to the argument that GGA band gap is sufficiently accurate to predict the experimental values. In fact, the linear regression with only the feature of GGA band gap gave a CV-RMSE of 0.71 eV. Statistically, GGA band gap has a Pearson correlation coefficient of 0.86 with E_g. Therefore, although GGA simulation falls away from accurately predicting the experimental value, the value of GGA band gap provides a conditional range to estimate E_g.[45] Adding this expressive information reduced the squared bias from 0.26 to 0.09 eV^2 (Bootstrap estimation), demonstrating the significant alleviation of the issue of underfitting. More importantly, despite the greatly improved precision, the model DoF was in fact reduced from 12 to 9 after adding GGA band gap as a new descriptor, confirming that the improvement of precision was no longer associated with the increase of DoF. These results were in full agreement with the expectation that the usage of CEP descriptor enhances the accuracy of prediction without scarifying the complexity of ML models.

To further analyze the effect of integrating the statistically correlated CEP in the feature space, we added synthetic Gaussian noise to the GGA band gap, which acted as an irreducible error in a presumably badly controlled estimation. As we expect, integrating a feature weakly correlated with the target property barely affected the predictive performance, as shown in Fig. 5b. At the limit when the CEP was composed of random noise, the model behaved the same as that without knowing the crude estimation. The predicting error decreased rapidly once the Pearson correlation exceeded a threshold, about 0.5 in our study. Interestingly, this threshold was coincidently close to the highest Pearson correlation between E_g and chemical descriptors. We note that this result did not prove that an ML model should always select features exhibiting higher correlation with the property. Nonetheless, it clearly demonstrated that the integration of CEP improves the predicting capability if the estimation shows sufficient statistical correlation with property.

Exercise 2: Lattice thermal conductivity

In contrast to the study of E_g, the CEP of the lattice thermal conductivity (κ_L) cannot be directly obtained as output from ab initio calculation. Although first-principle-based methods can predict κ_L through accurately accounting for the anharmonic lattice dynamics, the high computational cost prevents the usage as descriptors in ML modeling. In our study, we used the empirical Slack model to obtain the CEP of κ_L,[46] assuming that the phonon scattering is dominated by the Umklapp process. In addition, we used a unified Grüneissen parameter for all compounds to avoid the expensive calculation of the anharmonicity of lattice dynamics. While these assumptions together with other sources of error such as ignoring the contribution from optical phonon evidently resulted in apparent inaccuracy to predict κ_L (Supporting Information, Figure S5), the estimation showed good statistical trend with the experimental values with the Pearson correlation of 0.87 and the Spearman ranking correlation of 0.85, suggesting the simplified Slack model was an inaccurate estimation but statistically correlated descriptor to establish the ML model.

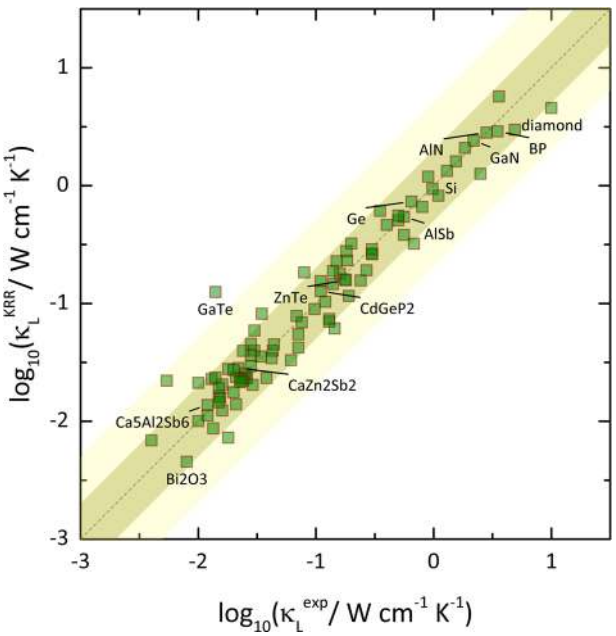

Fig. 6 Comparison of the lattice thermal conductivity from experimental measurement and from kernel ridge regression using leave-one-out calculation. The dark and light yellow shows the deviation within a factor of 2 and 5, respectively

Figure 6 compares the experimental κ_L with the prediction from ML model using the LOOCV method. Clearly, the ML model accurately predicted κ_L across a diverse range of compounds. The difference between the experimental κ_L and predicted value lied within a factor of 1.5 for 65% of the whole 93 compounds and within a factor of 2 for 90% of compounds. Only one outlier (GaTe) had the predicted value differing from experiment by more than half an order of magnitude. Quantitatively, we used the average factor difference (AFD) as proposed by Miller et al.[47] to evaluate the performance of the ML model. For the LOOCV prediction, the model has AFD of 1.38 (1.34 if not including GaTe in the modeling), exceeding the reported value of 1.48 using a modified Debye–Callaway model to predict κ_L.[47] The performance of the ML model was further verified by comparing the prediction with other reported models. For the compounds with simple crystalline structures of rocksalt, zincblende, and diamond, the current ML model achieved the AFD of 1.27 while for compounds with more complicated crystalline structures the AFD was 1.42. These values demonstrated the significant improvement of ML model compared to the calculation using Slack model with full Grüneissen parameter calculations or with the Grüneissen parameter estimated from the Mie–Grüneissen equation.[48,49]

It is interesting to note that the prediction for high κ_L seemed to exhibit higher accuracy than that for lower κ_L. The examination of AFD confirmed this observation by showing an uneven distribution of error: for compounds with κ_L higher than 0.1 W cm^{-1} K^{-1}, the AFD was 1.33 while for those with κ_L lower than 0.1 W cm^{-1} K^{-1} the AFD was 1.43. These results reflected the effect of the uncertainty of measurement on the predictive accuracy. In principle, a random noise affects the predicting error independently with the range of property. However, for models that learned the property in the logarithm format the predicting error is calculated on the relative scale, magnifying the noise for samples with small valued properties. Similar observations were noticed in the ML modeling of elastic properties of inorganic compounds, where the prediction seemed to be better for higher moduli materials.[39]

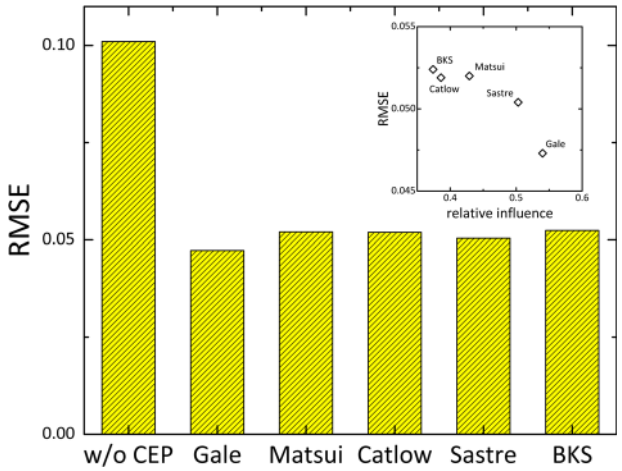

Fig. 7 Root mean square error of gradient boosting regressor to predict the bulk modulus (log(*K*)) of silica zeolite using crude estimation from different classical force field calculations. The RMSE of the model not using crude estimation is also shown. Insertion: the relative influence of different classical force field calculations in the machine learning model

Table 1. Summary of the results from case studies of modeling the experimental band gap (E_g), lattice thermal conductivity(κ_L), and elastics of zeolite (log(*K*))

Property	E_g	κ_L	log(*K*)
Data volume	108	93	102
Scaled error (%)	6.2	4.1	6.1
Scaled error before (%)	9.3	6.2	13
DoF	9	5	-[a]
DoF before	12	7	-[a]
Source of property	Experiment	Experiment	DFT
Source of CEP	DFT	Empirical model	Force field calculation
Regression method	Kernel ridge	Kernel ridge	Gradient boosting
Feature selection	Stepwise forward search	Stepwise forward search	—[a]

[a]Following the same approach used in ref. [31] no feature selection was performed for log(*K*)

Exercise 3: Elastic modulus of zeolites

The above two cases used the experimental values of E_g and κ_L as the training property and the precision of ML model was consequently affected by the uncertainty associated with different measurement techniques. In many studies the training set of property is created through high-throughput simulation where the uncertainty of measurements is better controlled. To examine the performance of the proposed strategy in the prediction of simulation derived properties, we modeled the Voigt–Reuss–Hill averaged bulk and shear moduli (*K* and *G*, respectively) of zeolites. Evans and Coudert[31] recently calculated *K* and *G* for over one hundred silica zeolites by means of the DFT method and used gradient boosting regressor to predict the DFT calculated values. Their ML model achieved significantly better predictive accuracy when compared to force fields for prediction of bulk modulus. For five classical force fields, the prediction of bulk modulus showed large deviation from DFT values with systematic errors exhibited in some calculations. Here we used the dataset of Evans and Coudert and established new ML models by incorporating the force field calculation as an additional descriptor in the feature space. Figure 6 compared the three-fold CV RMSE of different ML models for the prediction of log(*K*). In general, the integration of CEP from force field calculation improved the precision with RMSE reduced by around 50%. Similarly improved predictive precision was also observed for the prediction of log(*G*) (Supporting Information, Figure S6). Interestingly, although the Catlow potential gave better prediction of DFT-calculated bulk modulus compared to other classical potentials,[31] the corresponding ML models had the second worst predictive accuracy. As shown in the insertion of Fig. 7, the RMSE was strongly correlated to the relative influence of CEPs in the ML model. Therefore, the improved predictive precision was attributed to the statistical relation of CEP with the property but not to its absolute value.

DISCUSSION

We summarized the results from the cases studies in Table 1. All these studies utilized the available dataset of around 100 examples, which in our opinion represented a lower limit to apply ML in materials research. Although these studies varied in terms of data source, method to obtain CEP, the algorithm to select appropriate features, and regression method, in the vicinity of including CEP as a descriptor the predictive capability was effectively boosted with scaled error well below the trend observed in the aforementioned survey, demonstrating the capability of the proposed strategy in constructing accurate ML models with small available materials data. Of importance is that the success of proposed strategy relies on the statistical relation of CEP and property instead of requiring sufficiently accurate estimation of targeted property itself, which places the minimal hurdle to design appropriate descriptor. Considering the vast number of models and methods to empirically predict materials properties, we are optimistic that our proposed strategy permits a general solution to bridge machine learning techniques and the conventional wisdom of materials scientists to create better predictive models.

Developing the method to harvest the trend in a small materials data is not only of scientific significance but also of practical importance. Many materials properties are available in the quantity typically of the size of one to a few hundreds, necessitating the needs of special care when attempting to establish ML model. The current work studied the fundamental interplay between the data volume and predictive precision. We demonstrated that instead of affecting the precision directly the effect of data volume is mediated by the model DoF, resulting in the precision–DoF association when the model is trained with limited availability of materials data. The appearance of precision–DoF association is a signal of statistical underfitting and characterized by large bias of prediction, hence restricting the predictive capability in unknown domains. A solution to establish accurate ML models with small materials data is proposed by incorporating the CEP as a descriptor. In three case studies, the usage of crude estimation effectively boosted the predictive capability of ML models to state-of-art levels, demonstrating the generality of the proposed strategy to construct accurate ML models using small materials data.

METHODS

Data preparation

Property dataset. Except for indium nitride, the band gaps of A_xB_y binary compounds with experimental values in the range of 0.5–6 eV were compiled from two handbooks.[50,51] For indium nitride we used the value of 0.77 eV from the latest measurement by Wu et al.[52] Interestingly, the

model trained with the handbook value of 1.9 eV always predicted the band gap of indium nitride around 0.8 eV. The lattice thermal conductivity data were compiled from several resources.[47,49,51,53–57] The dataset was cleaned in the following procedure. For the materials studied in ref. [47], the same values of thermal conductivity were used. For other materials with duplicate measurements of the thermal conductivity, we used the values from the latest reports.

Crude estimation of properties. The modeling of E_g used the band gap calculated at the GGA level from Materials Project as the crude estimation. Note the dataset did not include compounds with transition metals. Therefore, we did not distinguish the GGA and GGA + U calculations in the modeling. The GGA band gap is well-known to underestimate the experimental band gap significantly. On average, the GGA band gap deviated from experimental values by 1.39 eV for the complied dataset.

The CEP of κ_L was obtained using the Slack model

$$\kappa_L = \frac{0.849 \times 3\sqrt[3]{4}}{20\pi^3(1-0.514\Upsilon^{-1}+0.228\gamma^{-2})} \times \frac{k_b^3\theta_D^3\rho V^{1/3}}{h^3 TN^{3/2}n\gamma^2}, \quad (3)$$

where the Debye temperature was calculated

$$\theta_D = \frac{h}{k_b}(6\pi^2 n)^{\frac{1}{3}} f(\sigma)\sqrt{K/\rho}. \quad (4)$$

Here γ is the Grüneissen parameter, T is the temperature, N is the number of atoms in the unit cell, V is the unit cell volume, ρ is the density, k_b is the Boltzmann constant, σ is the Poisson ratio, and K is the bulk modulus. The inputs of V, ρ, σ, and K were all obtained from first-principles calculations in Materials Project.[39] To avoid the expensive calculation of γ, we assume a value of 1 for all compounds. This simplification was made on the following considerations. First, the Grüneissen parameter can be roughly estimated as $0.5(\partial K/\partial P)-1$, where the value of the derivative for most materials is around 3–6.[58] Second, the value of 1 lies between that for tetrahedral compounds and for octahedral compounds,[47] which are two of the most common bonding environments in inorganic compounds. Third, we anticipated that the error caused by this simplification as other sources of error will be corrected in the ML modeling.

Chemical descriptors. Following the approach of Ward et al.,[34] the "fingerprint"-type chemical descriptors were categorized into the following: (1) stoichiometric attributes including the weight percentage and atomic percentage of the elements; (2) elemental properties including electronegativity, atomic radius, effective nuclear charge, Vander Waals radius, covalent radius, row number in the periodic table, block number, enthalpy of formation of gaseous atoms, ionization energy, and valence number; (3) compound descriptors including the molecular weight, density, volume, coordination number, and atomic number density; and (4) electronic structure contributes. The subset of optimal features was selected using the feature selection algorithm.

Machine learning

Regression. For the KRR the Scikit-learn package in Python was used.[59] The prediction value for the property is

$$\hat{f}_{KRR}(\boldsymbol{x}) = \sum_{i=1}^{n} a_i K(\boldsymbol{x}, \boldsymbol{x}_i), \quad (5)$$

where K is a kernel to measure the similarity between the training point $\boldsymbol{x}_i$ and the predicting point $\boldsymbol{x}$. A standard choice of the kernel is the radial basis function kernel $K(\boldsymbol{x}, \boldsymbol{x}_i) = \exp\left(-\frac{|\boldsymbol{x}-\boldsymbol{x}_i|^2}{2\sigma^2}\right)$ with the length scale σ used to tune the similarity. The weights $\alpha=(a_1,a_2,\ldots a_n)^\top$ are a set of variables that minimize the cost function

$$C(a_1, \ldots a_n) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \hat{f}_{KRR}(\boldsymbol{x}_i)\right)^2 + \eta \sum_{ij} a_i K(\boldsymbol{x}_i, \boldsymbol{x}_j) a_j, \quad (6)$$

where η is a tuning parameter to control the regularization term with the squared error. The hyperparameters σ and η are determined from CV.

For the LASSO method the glmnet package in R was used.[60] The LASSO prediction value for the target property at the point $\boldsymbol{x}$ is

$$\hat{f}_{LASSO}(\boldsymbol{x}) = \beta_0 + \sum_j^p \beta_j x_j, \quad (7)$$

where x_j is the jth feature of the predicting point $\boldsymbol{x}$. The coefficients β_j in LASSO model are a set of variables to minimize the objective function

$$L(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p}|\beta_j|, \quad (8)$$

where n is the number of data, x_{ij} is the jth feature of the $i$thdata $\boldsymbol{x}_i$, y_i is the target property of the ith data, and λ is a tuning parameter to control the impact of the shrinkage penalty, $\lambda \sum_{j=1}^{p}|\beta_j|$.

The prediction for Zeolite mechanic properties were utilizing the gradient boosting regression method as employed in the study of Evans and Coudert.[31] The calculation was performed using the gbm package in R.[61]

Feature selection. In the LASSO regression, the model was determined by varying the tuning parameter λ so that the shrinkage selects a subset of non-zero coefficients to minimize the CV error. In the KRR, the stepwise forward search procedure was used to select the features. The stepwise forward search started with zero feature and iteratively searched for the next feature with the largest reduction of CV error. The search stopped when the CV error cannot be reduced by adding new feature.

Mediation analysis. The mediation analysis was conducted following the procedure of Preacher and Kelly,[38] computed with MBESS package in R using an ordinary least squares regression-based analysis.[62] The standardized regression coefficient along with the standard error of the coefficients are reported. To determine the significance of the relation, the t-test was conducted at the statistical significance level of 0.01.

Data Availability

The datasets for the study are available from the corresponding author on reasonable request.

ACKNOWLEDGEMENTS

The authors want to thank their colleague Dr. Debasish Banerjee, Dr. Ryoji Asahi from Toyota Central R&D Labs, and Dr. Yoshiumi Kawamura from Toyota Motor Corporation for their instructive suggestions, Dr. Shihong Zhu and Dr. Hongshu Chen for the discussion of statistical methods. We are especially grateful to Dr. Jack D. Evans for providing us the raw data of zeolite calculations.

AUTHOR CONTRIBUTIONS

C.L. conceived the idea and designed the project. Y.Z. performed the experiment and prepared the data. Both authors analyzed the results, wrote, and revised the manuscript.

ADDITIONAL INFORMATION

Supplementary Information accompanies the paper on the *npj Computational Materials* website (https://doi.org/10.1038/s41524-018-0081-z).

Competing interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Jordan, M. I.., & Mitchell, T. M.. Machine learning: trends, perspectives, and prospects. *Science* **349**, 255–260 (2015).
2. Agrawal, A. & Choudhary, A. Perspective: materials informatics and big data: realization of the fourth paradigm of science in materials science. *APL Mater.* **4**, 053208 (2016).
3. Lookman, T., Alexander, F. J. & Rajan, K. Information Science for Materials Discovery and Design (Springer, Switzerland, 2016).
4. Hill, J. et al. Materials science with large-scale data and informatics: unlocking new opportunities. *MRS Bull.* **41**, 399–409 (2016).
5. Kalidindi, S. R. & Graef, M.D. Materials data science: current status and future outlook. *Ann. Rev. Mater. Res.* **45**, 171–193 (2015).
6. Rajan, K. Materials informatics: the materials "gene" and big data. *Ann. Rev. Mater. Res.* **45**, 153–169 (2015).
7. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater.* **3**, 54 (2017).
8. Seko, A., Hayashi, H., Nakayama, K., Takahashi, A. & Tanaka, I. Representation of compounds for machine-learning prediction of physical properties. *Phys. Rev. B* **95**, 144110 (2017).

9. Schütt, K. T. et al. How to represent crystal structures for machine learning: towards fast prediction of electronic properties. *Phys. Rev. B* **89**, 205118 (2014).
10. Isayev, O. et al. Materials cartography: representing and mining materials space using structural and electronic fingerprints. *Chem. Mater.* **27**, 735–743 (2015).
11. Medasani, B. et al. Predicting defect behavior in B2 intermetallics by merging ab initio modeling and machine learning. *npj Comput. Mater.* **2**, 1 (2016).
12. Jong, M.D. et al. A statistical learning framework for materials science: application to elastic moduli of k-nary inorganic polycrystalline compounds. *Sci. Rep.* **6**, 34256 (2016).
13. Legrain, F., Carrete, J., Roekeghem, A. V., Curtarolo, S. & Mingo, N. How chemical composition alone can predict vibrational free energies and entropies of solids. *Chem. Mater.* **29**, 6220–6227 (2017).
14. Chi, C. et al. Accurate force field for molybdenum by machine learning large materials data. *Phys. Rev. Mater* **1**, 043603 (2017).
15. Li, Z., Kermode, J. R. & Vita, A. D. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Phys. Rev. Lett.* **114**, 096405 (2015).
16. Takahashi, A., Seko, A. & Tanaka, I. Conceptual and practical bases for the high accuracy of machine learning interatomic potentials: application to elemental titanium. *Phys. Rev. Mater.* **1**, 063801 (2017).
17. Li, Z., Wang, S., Chin, W. S., Achenie, L. E. & Xin, H. High-throughput screening of bimetallic catalysts enabled by machine learning. *J. Mater. Chem. A* **5**, 24131–24138 (2017).
18. Ma, X., Li, Z., Achenie, L. E. K. & Xin, H. Machine-learning-augmented chemisorption model for CO2 electroreduction catalyst screening. *J. Phys. Chem. Lett.* **6**, 3528–3533 (2015).
19. Oliynyk, A. O. et al. High-throughput machine-learning-driven synthesis of full-Heusler compounds. *Chem. Mater.* **28**, 7324–7331 (2016).
20. Monnodi-Kanakkithodi, A., Huan, T. D. & Ramprasad, R. Mining materials design rules from data: the example of polymer dielectrics. *Chem. Mater.* **29**, 9901–9010 (2017).
21. Sendek, A. D. et al. Holistic computational structure screening of more than 12,000 candidates for solid lithium-ion conductor materials. *Energy Environ. Sci.* **10**, 306–320 (2017).
22. Ulissi, Z. W., Medford, A. J., Bligaard, T. & Nørskov, J. K. To address surface reaction network complexity using scaling relations machine learning and DFT calculations. *Nat. Commun.* **8**, 14621 (2016).
23. Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **553**, 73–77 (2016).
24. Xue, D. et al. Accelerated search for materials with targeted properties by adaptive design. *Nat. Commun.* **7**, 11241 (2015).
25. Dey, R. et al. Informatics-aided bandgap engineering for solar materials. *Comput. Mater. Sci.* **83**, 185–195 (2014).
26. Kim, E., Huang, K., Jegelka, S. & Olivetti, E. Virtual screening of inorganic materials synthesis parameters with deep learning. *npj Comput. Mater.* **3**, 53 (2017).
27. Kim, E. et al. Machine-learned and codified synthesis parameters of oxide materials. *Sci. Data* **4**, 170127 (2017).
28. Faber, F. A., Lindmaa, A., Lilienfeld, O. A. V. & Armiento, R. Machine learning energies of 2 million Elpasolite ($ABC2_D6$) crystals. *Phys. Rev. Lett.* **117**, 135502 (2016).
29. Schmidt, J. et al. Predicting the thermodynamic stability of solids combining density functional theory and machine learning. *Chem. Mater.* **29**, 5090–5103 (2017).
30. Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big data of materials science: critical role of the descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).
31. Evans, J. D. & Coudert, F.-X. Predicting the mechanical properties of zeolite frameworks by machine learning. *Chem. Mater.* **29**, 7833–7839 (2017).
32. Wu, H. et al. Robust FCC solute diffusion predictions from ab-initio machine learning methods. *Comput. Mater. Sci.* **134**, 160–165 (2017).
33. Meredig, B. et al. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **89**, 094104 (2014).
34. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**, 16028 (2016).
35. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman & Hall/CRC, New York, 1993).
36. Geman, S., Bienenstock, E. & Doursat, R. Neural networks and the bias/variance dilemma. *Neural Comput.* **4**, 1–58 (1992).
37. Zou, H., Hastie, T. & Tibshirani, R. On the "degrees of freedom" of the LASSO. *Ann. Stat.* **5**, 2173–2192 (2007).
38. Preacher, K. J. & Kelley, K. Effect size measures for mediation models: quantitative strategies for communicating indirect effects. *Psychol. Methods* **16**, 93–115 (2011).
39. Pilania, G. et al. Machine learning bandgaps of double perovskites. *Sci. Rep.* **6**, 19375 (2016).
40. Curtarolo, S. et al. The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
41. Lee, J., Seko, A., Shitara, K., Nakayama, K. & Tanaka, I. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Phys. Rev. B* **93**, 115104 (2016).
42. Pilania, G., Gubernatis, J. E. & Lookman, T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput. Mater. Sci.* **129**, 156–163 (2017).
43. Jain, A. et al. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
44. Lany, S. Band-structure calculations for the 3d transition metal oxides in GW. *Phys. Rev. B* **87**, 085112 (2013).
45. Setyawan, W., Gaume, R. M., Lam, S., Feigelson, R. S. & Curtarolo, S. High-throughput combinatorial database of electronic band structures for inorganic scintillator materials. *ACS Comb. Sci.* **13**, 382–390 (2011).
46. Slack, G. A. The thermal conductivity of nonmetallic crystals. *Solid State Phys.* **34**, 1–71 (1979).
47. Miller, S. A. et al. Capturing anharmonicity in a lattice thermal conductivity Model for high-throughput predictions. *Chem. Mater.* **29**, 2494–2501 (2017).
48. Madsen, G. K., Katre, A. & Bera, C. Calculating the thermal conductivity of the silicon clathrates using the quasi-harmonic approximation. *Phys. Status Solidi A* **213**, 802–807 (2015).
49. Toher, C. et al. High-throughput computational screening of thermal conductivity, Debye temperature, and Grüneisen parameter using a quasiharmonic Debye model. *Phys. Rev. B* **90**, 174107 (2014).
50. Weber, M. J. *Handbook of Optical Materials* (CRC Press, Boca Raton, FL, 2002).
51. Madelung, O. *Semiconductors: Data Handbook* 3rd edn (Springer-Verlag Berlin Heidelberg GmbH, New York, 2004).
52. Wu, J. et al. Unusual properties of the fundamental band gap of InN. *Appl. Phys. Lett.* **80**, 3967 (2002).
53. Yan, J. et al. Material descriptors for predicting thermoelectric performance. *Energy Environ. Sci.* **8**, 983–994 (2015).
54. Biswas, K., Zhao, L.-D. & Kanatzidis, M. G. Tellurium-free thermaoelectric: the anisotropic n-type semiconductor Bi2S3. *Adv. Energy Mater.* **2**, 634–638 (2012).
55. Plata, J. J. et al. An efficient and accurate framework for calculating lattice thermal conductivity of solids: AFLOW-AAPL automatic anharmonic phonon library. *Npj Comput. Mater.* **3**, 45 (2017).
56. Tan, Q. et al. Thermoelectrics with earth abundant elements: low thermal conductivity and high thermopower in doped SnS. *J. Mater. Chem. A* **2**, 17302 (2014).
57. Zhang, H. et al. Thermoelectric properties of polycrystalline SrZn2Sb2 prepared by spark plasma sintering. *J. Electron. Mater.* **39**, 1772–1776 (2010).
58. Vočadlo, N. L. & Price, G. D. The Grüneisen parameter—computer calculations via lattice dynamics. *Phys. Earth Planet. Inter.* **82**, 261–270 (1994).
59. Pedregosa, F et al. Scikit-learn: machine learning in Python. *J. Mach. Learn Res.* **12**, 2825–2830 (2011).
60. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
61. Ridgeway, G. gbm: Generalized boosted regression models, version 2.1. *The Comprehensive R Archive Network* 1–34 (2017).
62. Kelley, K. Confidence intervals for standardized effect sizes: theory, application, and implementation. *J. Stat. Softw.* **20**, 1–24 (2007).