Journal of Materials Chemistry C

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d4tc03403a

Combining machine-learning models with first-principles high-throughput calculations to accelerate the search for promising thermoelectric materials†

Tao Fan * and Artem R. Oganov

Received 9th August 2024,
Accepted 3rd November 2024

DOI: 10.1039/d4tc03403a

rsc.li/materials-c

Thermoelectric materials can achieve direct energy conversion between electricity and heat, and thus can be applied to waste-heat harvesting and solid-state cooling. The discovery of new thermoelectric materials is mainly based on experiments and first-principles calculations. However, these methods are usually expensive and time-consuming. Recently, the prediction of properties *via* machine learning has emerged as a popular method in materials science. Herein, we firstly did first-principles high-throughput calculations for a large number of chalcogenides and built a thermoelectric database containing 796 compounds. Many novel and promising thermoelectric materials were discovered. Then, we trained four ensemble learning models and two deep learning models to distinguish the promising thermoelectric materials from the others for n-type and p-type doping, respectively. All the presented models achieve a classification accuracy higher than 85% and area under the curve (AUC) higher than 0.9. In particular, the M3GNet model for n-type data achieves accuracy, precision and recall all higher than 90%. Our work demonstrates a very efficient way of combining machine-learning prediction and first-principles high-throughput calculations to accelerate the discovery of advanced thermoelectric materials.

1 Introduction

Thermoelectric (TE) materials could play an important role in building clean and alternative energy sources due to their ability to realize the direct conversion between heat and electricity.[1–3] Thermoelectric devices are of small size, and produce no noise or pollution; thus they have wide application potential in space power, industrial waste-heat harvesting, small and mobile refrigerators, and other fields.[4–6] The energy conversion efficiency of thermoelectric materials depends on the dimensionless figure of merit (ZT). ZT is defined as $ZT = \alpha^2\sigma T/(\kappa_e + \kappa_L)$, where α is the Seebeck coefficient, σ is the electrical conductivity, T is the absolute temperature, κ_e is the electronic thermal conductivity, and κ_L is the lattice thermal conductivity. In particular, $\alpha^2\sigma$ is called the power factor (PF). In order to obtain a high ZT, both α and σ must be maximized, while κ_e and κ_L need to be minimized. However, the interdependence of these parameters makes improving the ZT of a material a great challenge.[7,8]

Traditional thermoelectric materials discovery has been led by experiments, while computations are becoming more and more important with the advances in theory and the increase of computing power.[9–11] First-principles methods, such as DFT, have been widely used in calculating thermoelectric properties.[12–15] However, full first-principles calculation of transport properties is usually computationally expensive. Thus, there are many simplified models being proposed to calculate electronic and phonon transport properties, leading to many interesting and important discoveries.[16–20] However, they all face the problem of accuracy–computational-cost trade-off. Recently, machine learning (ML) has achieved much progress in both its theory and available models.[21] Data science and machine learning have become an integral part of natural sciences, as the fourth pillar in science, next to experiment, theory, and simulation.[22–25] ML algorithms find patterns in high-dimensional training data and build a mathematical model to make predictions or decisions without explicit human knowledge. This approach has been applied successfully to various materials science studies, such as structure prediction,[26,27] construction of force fields,[28–30] and predictions of the properties of materials.[31–33] A variety of ML algorithms provide an alternative to costly and complex DFT calculations, providing similarly accurate results in a fraction of the time. For TE materials, ML-assisted research has also been conducted, focusing on

Skolkovo Institute of Science and Technology, Bolshoy Boulevard 30, bld. 1, 121205 Moscow, Russia. E-mail: Tao.Fan@skoltech.ru

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4tc03403a

predicting the band structure, Seebeck coefficient, electrical conductivity, power factor, lattice thermal conductivity and figure of merit directly with carefully designed features or features learned by the model itself.[34–39] These works are based on supervised learning. Besides these, there are some works using unsupervised ML, which does not require well-labeled training data and can discover hidden patterns in the unlabeled datasets based on the input feature values. For example, Jia *et al.* proposed a strategy to discover a series of promising half-Heusler thermoelectric materials through the iterative combination of unsupervised ML with labeled known half-Heusler thermoelectric materials.[40]

The training data for ML-assisted thermoelectric studies are usually from experiments. However, such data are not only distributed over a huge number of publications, but also have varied synthesis conditions and measurement conditions. Even worse is that the data from different sources may conflict with each other. While first-principles calculations can guarantee the consistency of the data, the huge computational cost for high-throughput study limits the amount of the obtainable data. In this work, we explore the possibility of using ML methods to speed up first-principles high-throughput screening work. Using the AICON code developed in our group,[41] we first built a database comprising 796 chalcogenides with their n-type and p-type TE properties, including Seebeck coefficient, electrical conductivity, power factor, *etc.* We found many novel and promising TE materials, and some of them, such as Ge_5Te_4Se, $KBiSe_2$, GeTe (*Pnma*) and $BaCu_2Te_2$, are predicted to be much better than state-of-the-art TE materials. Then, we applied ML methods to this dataset and trained four ensemble learning models – random forest (RF), gradient boosting decision tree (GBDT), adaptive boosting (AdaB) and extreme gradient boosting (XGB)[42] – and two deep learning models – MatErials graph network (MEGNet)[43] and materials three-body graph network (M3GNet)[44] – to identify the promising TE materials, those having high power-factor values, from the others. The predictions on test sets show that all the trained models can achieve classification accuracy higher than 85%. Furthermore, the trained models were analyzed using the SHAP method, revealing that these models truly capture the underlying physics.

2 Methods

2.1 First-principles calculations

All first-principles calculations were performed using the Vienna *Ab initio* Simulation Package (VASP) with the Perdew–Burke–Ernzerhof generalized gradient approximation (PBE-GGA) and projector augmented wave (PAW) pseudo-potentials.[45–47] For structure relaxation, the plane-wave kinetic energy cut-off was set to 600 eV and the Brillouin zone was sampled using Γ-centered meshes with a reciprocal-space resolution of $2\pi \times 0.03$ Å^{-1}. Kohn–Sham equations were solved self-consistently with the total energy tolerance of 10^{-7} eV per cell and structures were relaxed until the maximum force became smaller than 10^{-3} eV Å^{-1}. The dielectric constants were calculated using density functional perturbation theory (DFPT),[48] and the elastic constants were calculated using the finite difference method as implemented in VASP. To obtain the deformation potential constants of a compound, three band-structure calculations were run: one at the equilibrium volume, and the other two at volumes −0.1% and +0.1% with respect to the equilibrium one.

After all necessary first-principles calculations were finished, the resulting files were collected and key parameters, including the conductivity effective mass m_c^*, density of states effective mass m_d^*, deformation potential constant Ξ, band degeneracy N, band gap E_g, elastic constant tensor C and dielectric constant ε, were extracted from these files as input to the AICON code to calculate the TE transport properties. To enable high-throughput screening, the automated workflow control, from structural relaxations, over band-structure calculations *etc.*, to transport property calculation, was also developed and implemented in AICON based on the Materials Project high-throughput infrastructure.[49–51] The detailed workflow and settings for each step can be found in our previous work.[41,52]

2.2 ML implementation

2.2.1 Feature engineering. Feature engineering is a crucial step in ML, as it directly impacts the performance of predictive models. The input features for a compound should be quick to compute and should capture all relevant features of that compound in a compact list of attributes. In this work, we designed a feature vector consisting of three different groups of descriptors:

Composition: the group of composition descriptors are similar to those used by Ward *et al.*[53] They include stoichiometric attributes, elemental property attributes, valence shell attributes and ionicity attributes. Different from the previous work, elemental property attributes are formed by the mean, maximum, minimum, range, and mean absolute deviation of 23 different elemental properties of all atoms in a compound. The list of these 23 elemental properties can be found in the ESI,† Table S1.

Structure: the group of structure descriptors have three different types, the Voronoi tessellation of crystal structure (VORONOI) as used by Ward *et al.*,[54] the partial radial distribution function (PRDF) as used by Schütt *et al.*[55] and similar to the fingerprint function of Oganov & Valle,[56] and the generalized radial distribution function plus bond order parameter (GRDF + BOP) as used by Seko *et al.*[57] These three types of structure descriptors are commonly adopted general-purpose structure descriptors. Specifically, for the PRDF descriptors we generated four feature vectors with different lengths by fixing the cutoff parameter as 20 Å and changing the bin_size parameter; they are PRDF_10 (bin_size = 0.1 Å), PRDF_16 (bin_size = 0.16 Å), PRDF_20 (bin_size = 0.2 Å), and PRDF_25 (bin_size = 0.25 Å).

Band structure: the group of band-structure descriptors include five easily calculated band-structure parameters for the conduction band minimum (CBM) or valence band maximum

(VBM), including the band degeneracy N, conductivity effective mass m_c^*, density of states effective mass m_d^*, deformation potential constant Ξ, and band gap E_g. These parameters are also used by AICON to calculate the thermoelectric transport properties.

The composition and structure descriptors were generated by Matminer.[58] According to the involved structure descriptors, 6 different feature vectors were used in this work, namely PRDF_10, PRDF_16, PRDF_20, PRDF_25, VORONOI and GRDF_BOP.

2.2.2 Model selection. In this work, four common ensemble machine-learning algorithms, namely random forest (RF), gradient-boosted decision tree (GBDT), adaptive boosting (AdaB) as implemented in the scikit-learn (sklearn) package,[59] and the extreme gradient-boosting model as implemented in the XGBoost package with the sklearn interface,[42] were used to train the classification models. Before entering the ML algorithm, the input feature vectors were standardized to eliminate the influence of variance of each descriptor. A grid search method with 5-fold cross validation on the training set as implemented in sklearn was used to optimize the hyperparameters in these models.

For deep learning, MEGNet and M3GNet models, as implemented in Materials Graph Library (MGL), were used.[43,44] The binary cross entropy loss was used as a loss function. We trained all models for 550 epochs using the Adam optimizer and a batch size of 64. The initial learning rate is set to 0.001 and the LinearLR scheduler is used to adjust the learning rate per epoch. The final learning rate decayed to 10% of the original value after 500 epochs, then remained constant for another 50 epochs. During the optimization, the loss function values of the validation set were used to monitor the model's performance.

2.2.3 Model evaluation. The performance of ML models was evaluated with metrics such as accuracy, precision, recall, F1-score, ROC curve and AUC. For a binary classification problem, all samples can be divided into four categories according to the combination of their real label and predicted label, as shown in Table 1.

Then, the definitions of those metrics are as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \quad (1a)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (1b)$$

$$\text{Recall} = \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (1c)$$

Table 1 Confusion matrix for binary classification

	Prediction	
Reality	Positive	Negative
Positive	TP	FN
Negative	FP	TN

$$\text{F1} = \frac{2 \times \text{Pre} \times \text{Rec}}{\text{Pre} + \text{Rec}} \quad (1d)$$

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}} \quad (1e)$$

The receiver operating characteristic (ROC) curve is the plot of the true positive rate (TPR) against the false positive rate (FPR) at each threshold setting. This means that the top left corner of the plot is the "ideal" point – a FPR of zero, and a TPR of one. AUC is the area under the ROC curve. For a perfect classifier, AUC equals 1. All of the above metrics are implemented in sklearn.metrics.

3 Results and discussion

3.1 Thermoelectric dataset

All structures were extracted from the Materials Project database[60] with five searching criteria: (1) S, Se and Te as anions; (2) the band gap should be larger than 0 eV but smaller than 1.2 eV, since good TE materials are usually narrow-gap semiconductors; (3) the energy above the convex hull line should be less than 0.1 eV per atom to ensure the structure is thermodynamically stable or at least potentially synthesizable under experimental conditions; (4) the material should belong to the cubic, tetragonal or orthorhombic crystal system; and (5) it should have a nonferromagnetic phase. These criteria resulted in over 1000 entries in the database. Then, calculations of the electronic transport properties for these compounds were carried out. Finally, 796 compounds finished the complete process of such calculations. Among them, 752 items were of n-type and 757 items of p-type. Other structures could fail because of various reasons. For example, since the GGA is known to underestimate the band gap, the calculated band gaps of some structures with very small gap values could be zero. In addition, the eigenvalues of the elastic constant matrix of some compounds have negative values. Such structures were discarded.

This work is an extension of our previous work.[52] In that work, we only calculated 94 cubic compounds' thermoelectric transport properties. Here we mainly calculated tetragonal and orthorhombic chalcogenide compounds' properties. Then both datasets were merged to build a whole thermoelectric database. Fig. 1 shows the maximum power factor with respect to the carrier concentration at which this maximum value is reached, for n- and p-type compounds, respectively, in the temperature range from 300 K to 1000 K. The maximum power factors for both n-type and p-type compounds are distributed across three orders of magnitude and most of them are within the range of 1 $\mu W\ cm^{-1}\ K^{-2}$–10 $\mu W\ cm^{-1}\ K^{-2}$. Some representative compounds are marked in the plots. Among these compounds, PbTe, PbS (cubic), PbSe (cubic), GeTe (cubic), SnTe, and SnSe are already well-known TE materials and have high power-factor values according to our calculations, which validates our methods. In the ESI,† Tables S2 and S3, we list the top 50 non-cubic TE materials found in this work for n- and p-type

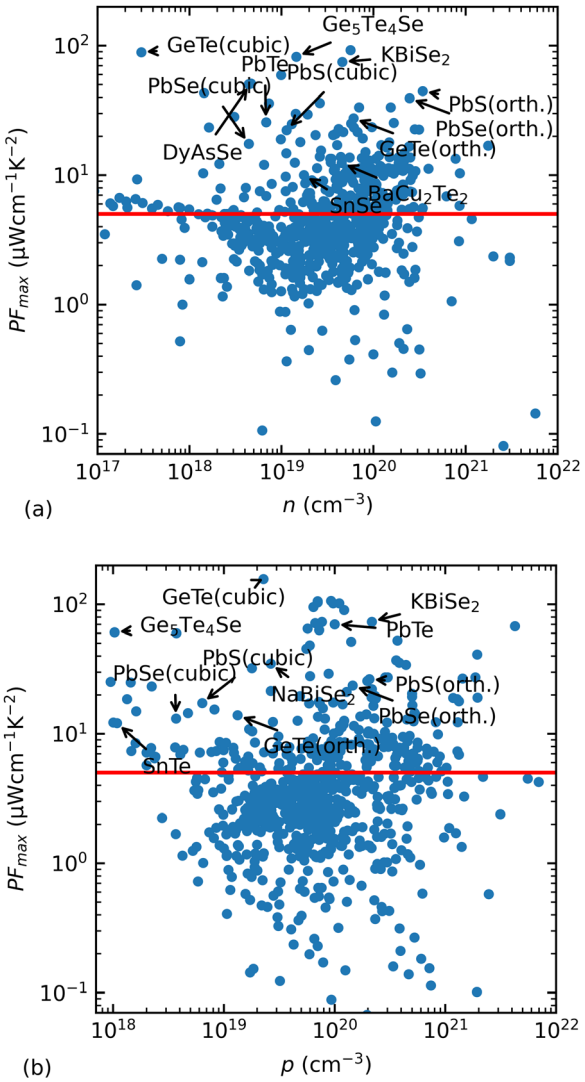

Fig. 1 Maximum power factor as a function of the corresponding carrier concentration for the studied compounds in the temperature range from 300 K to 1000 K for (a) n-type and (b) p-type data. Some compounds with high power factor are marked.

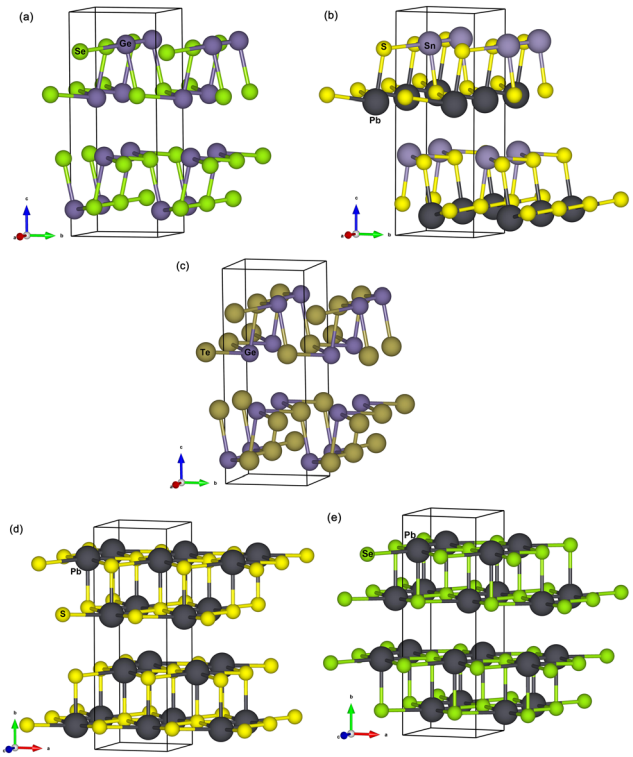

Fig. 2 Crystal structures of (a) GeSe, (b) $PbSnS_2$, (c) GeTe, (d) PbS, and (e) PbSe.

doping, respectively. All of them have their PF_{max} larger than 10 μW cm^{-1} K^{-2}, similar to those well-known TE materials. The tables also list the band-structure parameters for each compound. Generally, they all have large band degeneracy N and small conductivity effective mass m_c^*. A small m_c^* is beneficial for carrier mobility, while a large N, which means there are many carrier pockets involved in the transport, can make the DOS effective mass m_d^* large, and thus is beneficial for the Seebeck coefficient. Some of the listed compounds, such as Ge_5Te_4Se, $KBiSe_2$, TbAsSe, DyAsSe, YAsSe, PbS (*Cmcm*), PbSe (*Cmcm*), GeTe (*Pnma*), *etc.*, have high PF_{max} for both n- and p-type doping, which is good for building TE devices.

We picked up several interesting compounds and further calculated their lattice thermal conductivity κ_L and figure of merit ZT using AICON.[41] For example, SnSe is a famous TE material and its single crystal has a record high ZT = 2.6 at around 900 K.[61] SnSe has two phases: the low-temperature α phase with space group *Pnma* and high-temperature β phase with space group *Cmcm*. In Tables S2 and S3 (ESI†), there are several compounds that are similar to SnSe. Among them, the structures of GeTe, $PbSnS_2$ and GeSe are the same as α-SnSe, while the structures of PbS and PbSe are the same as β-SnSe. The crystal structures of the above compounds are shown in Fig. 2. Fig. 3 shows the PF of GeTe, $PbSnS_2$ and GeSe as a function of temperature and carrier concentration. Our calculations suggest that $PbSnS_2$ and GeSe are promising n-type TE materials, while GeTe is good for both n- and p-type. Fig. 4 shows the κ_L of these three compounds. Similar to that of SnSe, the κ_L values of these compounds are very low within the calculated temperature range. Therefore, their ZT values can exceed 1 in a wide range of temperatures and carrier concentrations, as shown in Fig. 5. Here we need to mention that AICON's lattice thermal conductivity model tends to underestimate the κ_L value of strongly anharmonic compounds. Thus, the ZT values of these compounds are overestimated. Still, the values can be a sign of great potential of these compounds. Fig. S1–S3 in the ESI† show the detailed PF, κ_L and ZT values of *Cmcm* PbS and PbSe. These two compounds are excellent n- and p-type TE materials; their ZT values could be higher than 1 in a wide range of temperatures and carrier concentrations. Another interesting compound we want to introduce is $BaCu_2Te_2$. This compound has a high PF for n-type doping. The crystal structure of $BaCu_2Te_2$ is shown in Fig. S4 (ESI†). The Cu atom is tetrahedrally coordinated by

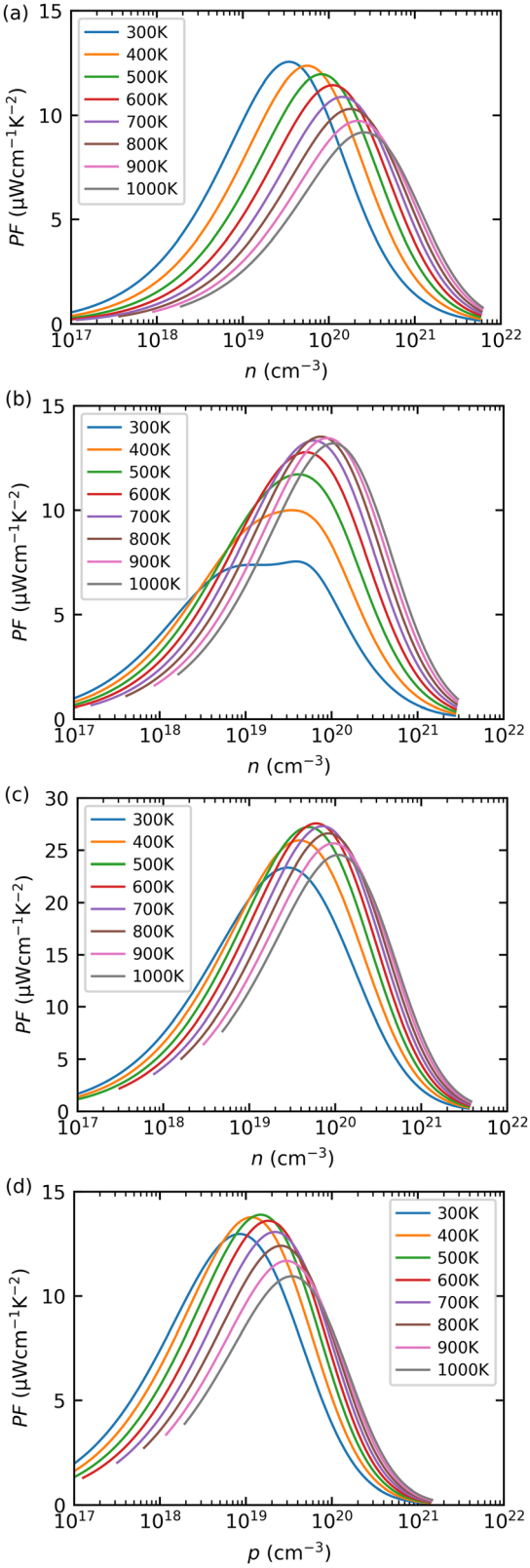

Fig. 3 Power factor at varying temperatures and carrier concentrations for (a) GeSe, (b) $PbSnS_2$, and (c) and (d) GeTe.

four Te atoms. Meanwhile, the Cu and Te atoms form one-dimensional frameworks with channels extended along the

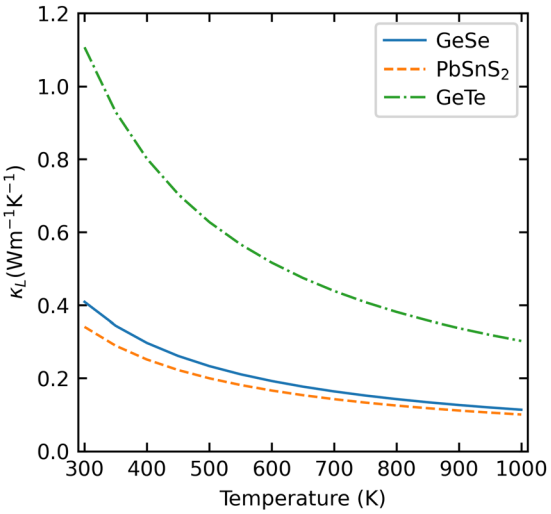

Fig. 4 Lattice thermal conductivity of GeSe, $PbSnS_2$, and GeTe.

a axis, while the Ba atoms locate inside these channels. Such a structure is beneficial for impeding the transport of phonons, and thus is expected to have low lattice thermal conductivity. According to our calculations, it indeed has very low κ_L values, together with high PF values, and the *ZT* of this compound is very large in a wide range of temperatures and carrier concentrations (see Fig. S5, ESI†). We plan to release the complete database in the near future.

Although our high-throughput framework AICON is quite fast for transport property calculation, it still has space to improve. The overall calculation process involves several different types of first-principles calculations. Some of these calculations, such as the elastic constant and dielectric constant, are quite time-consuming. Since good thermoelectric materials are always a minority, most of the computing resources are consumed on unpromising materials. If one could evaluate whether a compound is promising, especially if it has a high power factor before doing computationally demanding first-principles calculations, it would save a lot of time and computing resources. The simplest idea is to train a classification model to distinguish good and poor thermoelectric materials. At the first step, we need to label each sample in the dataset as positive (label 1) or negative (label 0). The maximum power factor value PF_{max}, as shown in Fig. 1, is a good index to be used to split the dataset. Here we used PF_{max} = 5 μW cm^{-1} K^{-2} as the dividing line (the red line in Fig. 1). The compounds above the line were labeled as 1, while those under the line were labeled as 0. The explanation for using 5 μW cm^{-1} K^{-2} as the boundary is in the ESI.† According to such a split, for n-type data, #positive : #negative = 308 : 444, while for p-type data, #positive : #negative = 244 : 513.

3.2 Ensemble learning models

Ensemble learning combines predictions of several base estimators built with a given learning algorithm in order to improve generalizability and robustness over a single estimator.[62] Multiple individual learners predict their own results, which are combined

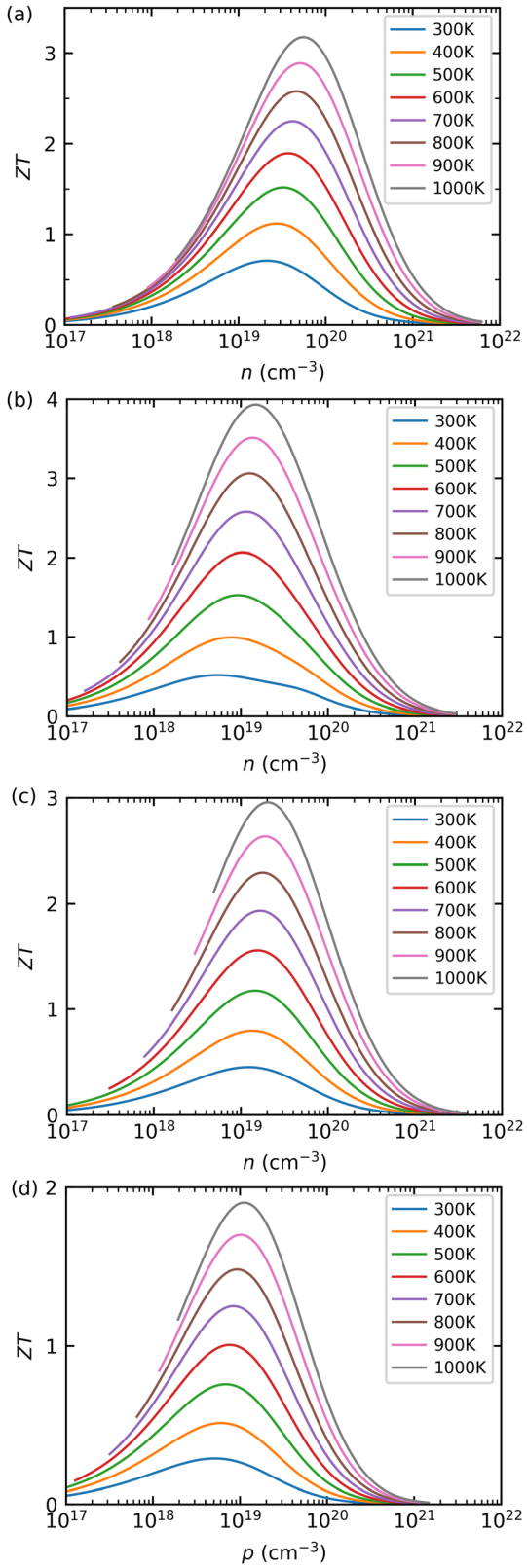

Fig. 5 Figure of merit at varying temperatures and carrier concentrations for (a) GeSe, (b) $PbSnS_2$, and (c) and (d) GeTe.

with a strategy such as weighted averages or voting. The performance of the ensemble learning model is usually obviously better than a single best model. In this work, four common ensemble machine-learning algorithms, including random forest (RF), gradient-boosting decision tree (GBDT), adaptive boosting (AdaB) and extreme gradient boosting (XGB), combined with six kinds of input feature vectors (see details in the Methods), were used to train the classification models. Both the n-type and the p-type datasets were divided into 90% training and 10% test sets.

Table 2 Performance measures on test sets of the best model of each ensemble learning algorithm with corresponding input feature vectors

	Model	Features	Acc.	Prec.	Recall	F1	AUC
N	GBDT	GRDF_BOP	0.86	0.78	0.90	0.84	0.97
	XGB	VORONOI	0.88	0.81	0.94	0.87	0.95
	AdaB	PRDF_16	0.89	0.85	0.90	0.88	0.95
	RF	GRDF_BOP	0.88	0.81	0.94	0.87	0.96
P	GBDT	PRDF16	0.92	0.95	0.79	0.86	0.96
	XGB	GRDF_BOP	0.92	0.95	0.79	0.86	0.96
	AdaB	PRDF_25	0.89	0.90	0.75	0.82	0.93
	RF	GRDF_BOP	0.89	0.90	0.75	0.82	0.95

Table 2 shows the performance measure on test sets of the best model of each ensemble learning algorithm. The optimal hyperparameters for each model are listed in Table S4 in the ESI.† The complete table for each choice of input feature vector of each learning algorithm can also be found in the ESI† (Tables S5–S8). The accuracies of models trained on n-type data are higher than 85%, while those trained on p-type data are higher than 90%. Since our objective is to reduce the first-principles computational cost as much as possible, we prefer other metrics, especially precision and recall. These two metrics directly reflect the efficiency of a model for picking up truly good thermoelectric materials. The models trained on n-type data have precision higher than 80% (except the GBDT model), and recall higher than 90%. In contrast, the models trained on p-type data have precision higher than 90%, with recall higher than 75%. Therefore, the F1 values, as the harmonic average of precision and recall, of these models are similar. From the column of AUC, all these models achieve AUC values larger than 0.9. Thus, all of them are very good classifiers. Generally, the performance metrics of different algorithms are similar for n-type data and p-type data, respectively. Although the best input feature vectors are different for different algorithms, GRDF_BOP appears four times in this table. It seems this kind of feature vector is better than other used feature vectors.

In the above trained models, the input feature vectors include band-structure descriptors, such as the band effective mass and band gap. Therefore, it is meaningful to compare the performance of the models with and without these descriptors in order to see if we can further reduce the computational cost. In the ESI,† Tables S9 and S10 show the performance measures using GBDT and XGB algorithms together with input feature vectors with and without band-structure descriptors. Generally, compared with the performance of those models with input features having band-structure descriptors, the models without band-structure descriptors perform worse for almost all metrics we evaluated. Therefore, these band-structure parameters are important for accurate predictions.

ML is criticized for being a black box. Interpretability is always a hot topic in the ML field. Here we used SHapley Additive exPlanations (SHAP) to explain our trained models.[63] SHAP connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions. The most attractive point of SHAP is that it can decompose the output of a model into the contribution of each input feature for each sample. Fig. 6 shows the use of SHAP to analyze the GBDT models presented in Table 2 (Fig. S6 in the ESI† shows the analysis of the XGB models in Table 2). Fig. 6(a) and (b) show summary plots of the top ten important features – they take the mean absolute SHAP value of each feature over all the samples of the dataset. All five band-structure parameters are among the top ten features for both n-type and p-type models, which explains why there is an obvious difference between the performance of models with and without band-structure parameters as input. Fig. 6(c) and (d) show beeswarm plots to summarize the entire distribution of SHAP values for each feature, where the color of each point represents the feature value of that individual. These two pictures can reveal the direction of the feature's effect. For example, the higher the band degeneracy N_c or N_v is, the higher the SHAP value is, which means a higher probability to be a good thermoelectric material. For the conductivity effective mass, in contrast, the lower m_c^* is, the higher the SHAP value is. This is consistent with our analysis in the last section. Another important feature revealed by SHAP analysis is the deformation potential constant Ξ – larger Ξ leads to a negative impact on the SHAP value. This also matches with theory, since large Ξ causes a large electronic scattering rate, and thus a small carrier mobility. Therefore, our models truly capture the underlying physics by just learning from a small amount of data.

3.3 Deep learning models

Deep learning models are mostly used in computer vision and natural language processing fields, since these fields have accumulated a large amount of data. Recently, with the development of several large materials databases, such as the Materials Project, OQMD and AFLOW, deep learning models appeared in materials science in large numbers. Different from ensemble learning, one of the advantages of a deep learning model is that it learns the representation of materials by itself. The graph neural network is one of the most used deep learning models in materials science. Graphs are a powerful non-Euclidean data structure method for establishing relationships between nodes and their edges. Thus, it is natural to represent a crystal structure as a graph. In this work, we used both MEGNet and M3GNet to predict the label of the candidate materials. Compared with MEGNet, M3GNet also incorporates three-body interactions by building an additional line graph for bonds, which makes it one of the state-of-the-art models.[64]

Instead of training the model from scratch, it is better to make use of those pretrained models and do transfer learning. Both MEGNet and M3GNet have pretrained models that were

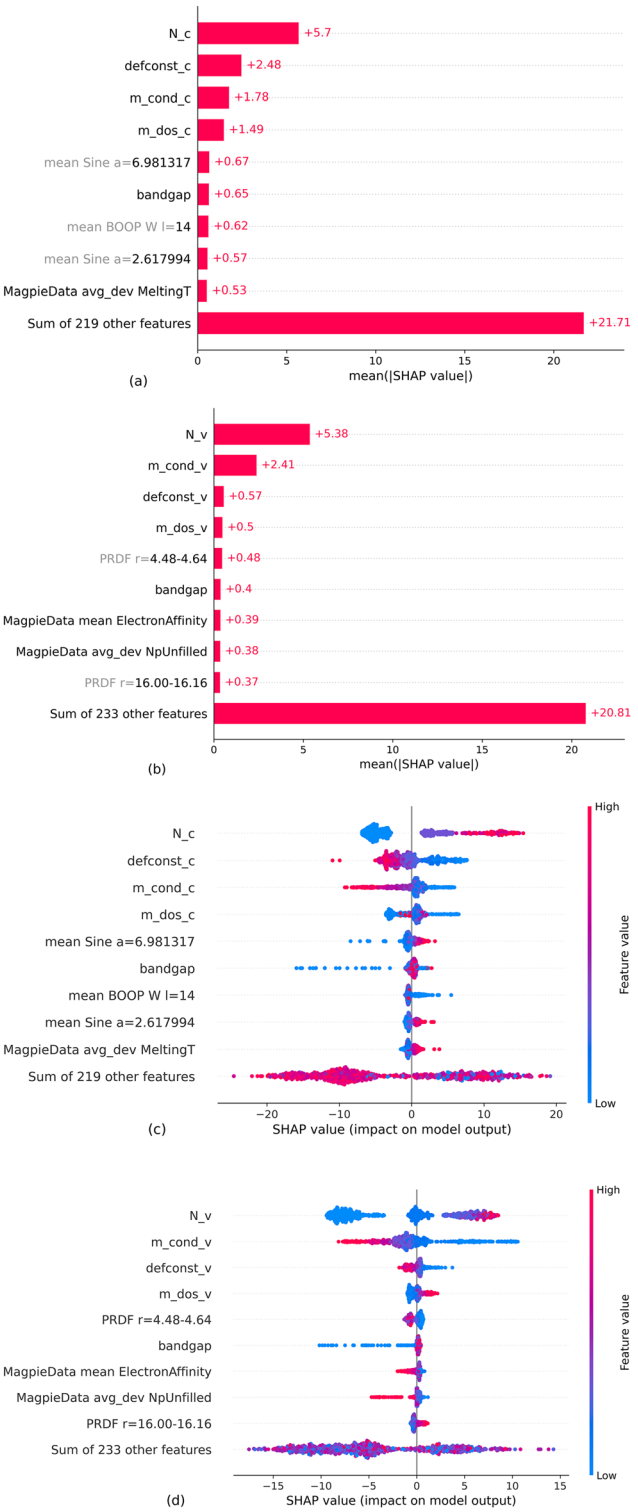

Fig. 6 SHAP analysis based on the GBDT models in Table 2, (a) and (c) for the n-type model, and (b) and (d) for the p-type model. (a) and (b) Bar charts of the average SHAP value magnitude for each input feature. (c) and (d) A set of beeswarm plots, where each dot corresponds to a sample. The dot's position on the x axis shows the impact that feature has on the model's prediction for that sample. When multiple dots are located at the same x position, they pile up to show density.

trained on larger datasets from the Materials Project to predict the energy of formation. Here we reused their element embedding layers since such embedding should be universal for all kinds of tasks. After the readout stage, the node features for each atom were combined into the crystal feature vector. At this time, we added the band-structure descriptors, and the concatenated vector was passed to a multi-layer perceptron (MLP) to predict the target value. Both the n-type and the p-type datasets were divided into 80% training, 10% validation and 10% test sets. The hyperparameters, including the learning rate, number of units in each layer, weight decay, *etc.*, were optimized based on the loss of the validation set. Then, the final models were trained on the datasets merged from the training sets and validation sets. Tables S11 and S12 (ESI†) list the settings for the hyperparameters. Fig. S7 (ESI†) shows the training loss and test loss during the final training process. The test loss converged at the end of training for all selected models.

Besides training deep models directly, we also used the pre-trained M3GNet model as a way to generate structure feature vectors, then combined them with composition features and band-structure features as inputs to GBDT and XGB models. By doing this, we wanted to see if features learned from one task can be applicable to another.

The performance measures on test sets for the two deep models and for two models combining an ensemble learning algorithm and M3GNet structure feature vectors are shown in Table 3. Although the datasets are small, the performance of the two deep models is very good. In particular, the M3GNet model for n-type data has accuracy, precision and recall all higher than 90%, making it the best model that we have obtained. For p-type data, although the M3GNet model is not much better than those in Table 2, its performance is more balanced – both precision and recall are higher than 80%. The performance of MEGNet models for both n-type and p-type data is a bit worse than that of the M3GNet models. In addition, models trained on n-type data are generally better than those trained on p-type data, mainly because the p-type dataset is less balanced than the n-type dataset. For the two composite models, their performance is not better than that of deep models. Moreover, compared with GBDT and XGB models with general purpose structure features (see Tables S5 and S6, ESI†), these two models do not show clear advantages, suggesting the trained structure features for one task may not be the optimal choice for another. Finally, we combined four ensemble learning models in Table 2 with the M3GNet model in Table 3 to form a voting classifier. In this classifier, the predicted class label for a particular sample is the class label that represents the majority of the class labels predicted by each individual classifier. Although the voting classifiers are not better than the best individual classifier for both n-type and p-type data, they should be more robust for the unseen data.

Fig. 7 shows the ROC curves of the four ensemble learning models in Table 2 and two deep learning models in Table 3 on test sets. The AUC values of all these models are higher than 0.9, close to the perfect classifier. Then, these trained models were integrated into AICON's electronic transport property workflow. Before doing elastic constant and dielectric constant

Table 3 Performance measures on test sets of the deep learning models and composite models with deep learned input structure features

	Model	Acc.	Prec.	Recall	F1	AUC
N	MEGNet	0.88	0.82	0.90	0.86	0.96
	M3GNet	0.93	0.91	0.94	0.92	0.96
	GBDT-M3G	0.80	0.71	0.87	0.78	0.95
	XGB-M3G	0.84	0.74	0.94	0.83	0.94
	Voting	0.87	0.80	0.90	0.85	
P	MEGNet	0.87	0.77	0.83	0.80	0.91
	M3GNet	0.88	0.80	0.83	0.82	0.92
	GBDT-M3G	0.83	0.72	0.75	0.73	0.93
	XGB-M3G	0.86	0.76	0.79	0.78	0.90
	Voting	0.92	0.95	0.79	0.86	

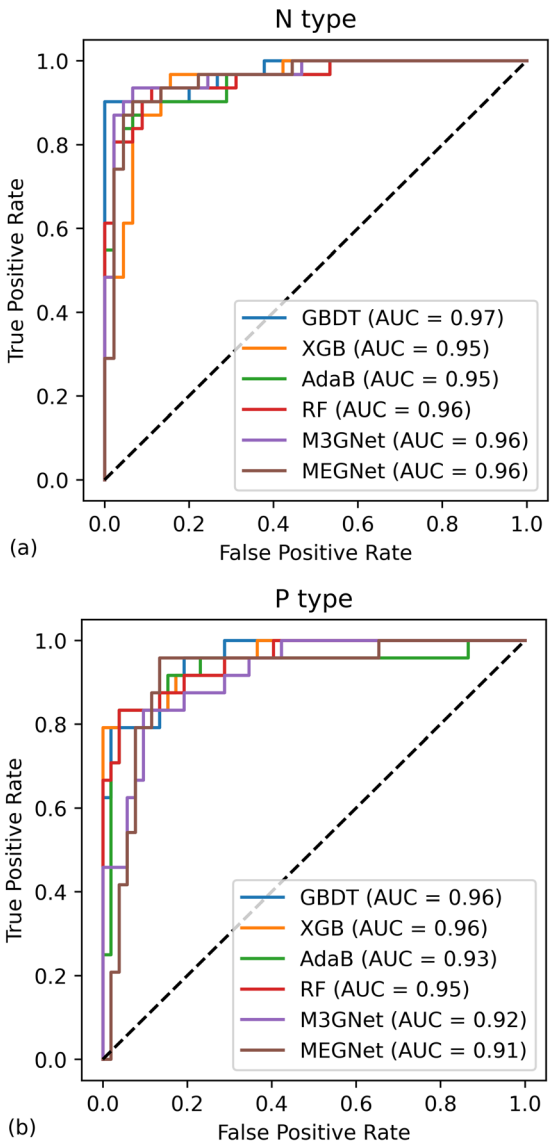

Fig. 7 Receiver operating characteristic (ROC) curves of models listed in Tables 2 and 3, for (a) n-type and (b) p-type data. The black dashed line is the chance level line (AUC = 0.5).

calculations, the ML models should be used to make a prediction. Any compound that was not predicted as a good one would be discarded and its following calculation would be cancelled. In this way, a lot of time and computational resources can be saved.

4 Conclusions

In this work, we did first-principles high-throughput transport property calculations and built a database containing 796 compounds' electronic transport properties, including the Seebeck coefficient, electrical conductivity, power factor, *etc.* Then, we picked up several compounds with high power factors and calculated their lattice thermal conductivity and figure of merit further. We have found many novel and promising TE materials. Some of them, such as Ge_5Te_4Se, $KBiSe_2$, GeTe (*Pnma*) and $BaCu_2Te_2$, may have performance better than that of state-of-the-art TE materials. Then, in order to reduce the amount of costly computations, we trained several types of ML models to identify the TE compounds with high power factor from the others, with only crystal structures and parameters extracted from band structures as input. Specifically, four ensemble learning models and two deep learning models based on a graph neural network were trained and compared. Among them, the M3GNet model for n-type data achieved accuracy, precision and recall all higher than 90%, making it the best among the models we obtained. Moreover, all of the trained models achieve AUC values higher than 0.9, and their ROC curves are close to that of a perfect classifier. Integrating these models into the calculation workflow of electronic transport properties in AICON can speed up the process of screening of TE materials greatly. In the future, we plan to include more chalcogenides (hexagonal, monoclinic, *etc.*) into the high-throughput calculations and expand our database. Another important thing is to train a ML model only using structure as an input, which would even improve the efficiency of screening work an order of magnitude. We believe our work will greatly reduce the workload to find good thermoelectric materials and, in combination with experimental works, accelerate the discovery of superior thermoelectric materials.

Author contributions

T. F. initiated the project, performed the calculations, analyzed the data, and wrote the manuscript. A. R. O. participated in discussions of the methods and results and revised the manuscript.

Data availability

The data for this article, including composition and structure information, band-structure descriptors and power factors, and the code for training and testing ML models, are available online at **https://github.com/Baijianlu/ML4TE**.

Conflicts of interest

There are no conflicts of interest to declare.

Acknowledgements

We acknowledge the usage of the Skoltech HPC cluster ARKUDA for obtaining the results presented in this paper. This work is funded by the Russian Science Foundation (grant 24-43-00162).

Notes and references

1 D. M. Rowe, *Thermoelectrics handbook: macro to nano*, CRC Press, 2018.
2 G. Tan, L. Zhao and M. G. Kanatzidis, *Chem. Rev.*, 2016, **116**, 12123–12149.
3 X. Shi, L. Chen and C. Uher, *Int. Mater. Rev.*, 2016, **61**, 379–415.
4 L. E. Bell, *Science*, 2008, **321**, 1457–1461.
5 N. Jaziri, A. Boughamoura, J. Müller, B. Mezghani, F. Tounsi and M. Ismail, *Energy Rep.*, 2020, **6**, 264–287.
6 A. Nozariasbmarz, H. Collins, K. Dsouza, M. H. Polash, M. Hosseini, M. Hyland, J. Liu, A. Malhotra, F. M. Ortiz and F. Mohaddes, *et al.*, *Appl. Energy*, 2020, **258**, 114069.
7 L. Zhao, S. Hao, S.-H. Lo, C.-I. Wu, X. Zhou, Y. Lee, H. Li, K. Biswas, T. P. Hogan and C. Uher, *et al.*, *J. Am. Chem. Soc.*, 2013, **135**, 7364–7370.
8 T. Zhu, Y. Liu, C. Fu, J. P. Heremans, J. G. Snyder and X. Zhao, *Adv. Mater.*, 2017, **29**, 1605884.
9 P. Gorai, V. Stevanović and E. S. Toberer, *Nat. Rev. Mater.*, 2017, **2**, 17053.
10 S. Hao, V. P. Dravid, M. G. Kanatzidis and C. Wolverton, *npj Comput. Mater.*, 2019, **5**, 58.
11 J. J. G. Moreno, J. Cao, M. Fronzi and M. H. N. Assadi, *Mater. Renewable Sustainable Energy*, 2020, **9**, 1.
12 J.-J. Zhou, J. Park, I.-T. Lu, I. Maliyov, X. Tong and M. Bernardi, *Comput. Phys. Commun.*, 2021, **264**, 107970.
13 V. Askarpour and J. Maassen, *Phys. Rev. B*, 2023, **107**, 045203.
14 W. Li, J. Carrete, N. A. Katcho and N. Mingo, *Comput. Phys. Commun.*, 2014, **185**, 1747–1758.
15 K. Pal, Y. Xia, J. Shen, J. He, Y. Luo, M. G. Kanatzidis and C. Wolverton, *npj Comput. Mater.*, 2021, **7**, 82.
16 H. Zhu, G. Hautier, U. Aydemir, Z. M. Gibbs, G. Li, S. Bajaj, J.-H. Pöhls, D. Broberg, W. Chen and A. Jain, *et al.*, *J. Mater. Chem. C*, 2015, **3**, 10554–10565.
17 J. Carrete, N. Mingo, S. Wang and S. Curtarolo, *Adv. Funct. Mater.*, 2014, **24**, 7427–7432.
18 L. Xi, S. Pan, X. Li, Y. Xu, J. Ni, X. Sun, J. Yang, J. Luo, J. Xi and W. Zhu, *et al.*, *J. Am. Chem. Soc.*, 2018, **140**, 10785–10793.
19 P. Gorai, P. Parilla, E. S. Toberer and V. Stevanovic, *Chem. Mater.*, 2015, **27**, 6213–6221.
20 T. Jia, Z. Feng, S. Guo, X. Zhang and Y. Zhang, *ACS Appl. Mater. Interfaces*, 2020, **12**, 11852–11864.

21 Y. LeCun, Y. Bengio and G. Hinton, *Nature*, 2015, **521**, 436–444.
22 G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang and L. Yang, *Nat. Rev. Phys.*, 2021, **3**, 422–440.
23 K. Choudhary, B. DeCost, C. Chen, A. Jain, F. Tavazza, R. Cohn, C. W. Park, A. Choudhary, A. Agrawal and S. J. Billinge, *et al.*, *npj Comput. Mater.*, 2022, **8**, 59.
24 P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. van Hoesel, H. Schopmans and T. Sommer, *et al.*, *Commun. Mater.*, 2022, **3**, 93.
25 H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk and A. Deac, *et al.*, *Nature*, 2023, **620**, 47–60.
26 K. Ryan, J. Lengyel and M. Shatruk, *J. Am. Chem. Soc.*, 2018, **140**, 10158–10168.
27 E. V. Podryabinkin, E. V. Tikhonov, A. V. Shapeev and A. R. Oganov, *Phys. Rev. B*, 2019, **99**, 064114.
28 J. Behler, *Chem. Rev.*, 2021, **121**, 10037–10072.
29 O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko and K.-R. Müller, *Chem. Rev.*, 2021, **121**, 10142–10186.
30 T. Zubatiuk and O. Isayev, *Acc. Chem. Res.*, 2021, **54**, 1575–1585.
31 J. Lee, A. Seko, K. Shitara, K. Nakayama and I. Tanaka, *Phys. Rev. B*, 2016, **93**, 115104.
32 A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi and M. Ceriotti, *Sci. Adv.*, 2017, **3**, e1701816.
33 M. Li, L. Dai and Y. Hu, *ACS Energy Lett.*, 2022, **7**, 3204–3226.
34 Y. Gan, G. Wang, J. Zhou and Z. Sun, *npj Comput. Mater.*, 2021, **7**, 176.
35 Y.-L. Lee, H. Lee, T. Kim, S. Byun, Y. K. Lee, S. Jang, I. Chung, H. Chang and J. Im, *J. Am. Chem. Soc.*, 2022, **144**, 13748–13763.
36 Y. Li, J. Zhang, K. Zhang, M. Zhao, K. Hu and X. Lin, *ACS Appl. Mater. Interfaces*, 2022, **14**, 55517–55527.
37 X. Jia, H. Yao, Z. Yang, J. Shi, J. Yu, R. Shi, H. Zhang, F. Cao, X. Lin and J. Mao, *et al.*, *Appl. Phys. Lett.*, 2023, **123**, 203902.
38 Q. Ren, D. Chen, L. Rao, Y. Lun, G. Tang and J. Hong, *J. Mater. Chem. A*, 2024, **12**, 1157–1165.
39 Y. Luo, M. Li, H. Yuan, H. Liu and Y. Fang, *npj Comput. Mater.*, 2023, **9**, 4.
40 X. Jia, Y. Deng, X. Bao, H. Yao, S. Li, Z. Li, C. Chen, X. Wang, J. Mao and F. Cao, *et al.*, *npj Comput. Mater.*, 2022, **8**, 34.
41 T. Fan and A. R. Oganov, *Comput. Phys. Commun.*, 2021, **266**, 108027.
42 T. Chen and C. Guestrin, Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.
43 C. Chen, W. Ye, Y. Zuo, C. Zheng and S. P. Ong, *Chem. Mater.*, 2019, **31**, 3564–3572.
44 C. Chen and S. P. Ong, *Nat. Comput. Sci.*, 2022, **2**, 718–728.
45 G. Kresse and J. Furthmüller, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1996, **54**, 11169.
46 G. Kresse and J. Furthmüller, *Comput. Mater. Sci.*, 1996, **6**, 15–50.
47 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865.
48 X. Gonze and C. Lee, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1997, **55**, 10355.
49 S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson and G. Ceder, *Comput. Mater. Sci.*, 2013, **68**, 314–319.
50 K. Mathew, J. H. Montoya, A. Faghaninia, S. Dwarakanath, M. Aykol, H. Tang, I.-H. Chu, T. Smidt, B. Bocklund and M. Horton, *et al.*, *Comput. Mater. Sci.*, 2017, **139**, 140–152.
51 A. Jain, S. P. Ong, W. Chen, B. Medasani, X. Qu, M. Kocher, M. Brafman, G. Petretto, G.-M. Rignanese, G. Hautier, *et al.*, *Concurrency and Computation: Practice and Experience*, 2015, vol. 27, pp. 5037–5059.
52 T. Fan and A. R. Oganov, *J. Mater. Chem. C*, 2021, **9**, 13226–13235.
53 L. Ward, A. Agrawal, A. Choudhary and C. Wolverton, *npj Comput. Mater.*, 2016, **2**, 1–7.
54 L. Ward, R. Liu, A. Krishna, V. I. Hegde, A. Agrawal, A. Choudhary and C. Wolverton, *Phys. Rev. B*, 2017, **96**, 024104.
55 K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K.-R. Müller and E. K. Gross, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2014, **89**, 205118.
56 A. R. Oganov and M. Valle, *J. Chem. Phys.*, 2009, **130**, 104504.
57 A. Seko, H. Hayashi, K. Nakayama, A. Takahashi and I. Tanaka, *Phys. Rev. B*, 2017, **95**, 144110.
58 L. Ward, A. Dunn, A. Faghaninia, N. E. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom and M. Dylla, *et al.*, *Comput. Mater. Sci.*, 2018, **152**, 60–69.
59 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg, *et al.*, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
60 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner and G. Ceder, *et al.*, *APL Mater.*, 2013, **1**, 011002.
61 L.-D. Zhao, S.-H. Lo, Y. Zhang, H. Sun, G. Tan, C. Uher, C. Wolverton, V. P. Dravid and M. G. Kanatzidis, *Nature*, 2014, **508**, 373–377.
62 Z.-H. Zhou, *Machine learning*, Springer Nature, 2021.
63 S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal and S.-I. Lee, *Nat. Mach. Intell.*, 2020, **2**, 56–67.
64 J. Riebesell, R. E. Goodall, A. Jain, P. Benner, K. A. Persson and A. A. Lee, *arXiv*, 2023, preprint, arXiv:2308.14920, DOI: **10.48550/arXiv.2308.14920**.

