npj Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

Active learning for the power factor prediction in diamond-like thermoelectric materials

Ye Sheng[1], Yasong Wu[1,2], Jiong Yang[1]✉, Wencong Lu[3,1], Pierre Villars[4] and Wenqing Zhang[5,6]✉

The Materials Genome Initiative requires the crossing of material calculations, machine learning, and experiments to accelerate the material development process. In recent years, data-based methods have been applied to the thermoelectric field, mostly on the transport properties. In this work, we combined data-driven machine learning and first-principles automated calculations into an active learning loop, in order to predict the p-type power factors (PFs) of diamond-like pnictides and chalcogenides. Our active learning loop contains two procedures (1) based on a high-throughput theoretical database, machine learning methods are employed to select potential candidates and (2) computational verification is applied to these candidates about their transport properties. The verification data will be added into the database to improve the extrapolation abilities of the machine learning models. Different strategies of selecting candidates have been tested, finally the Gradient Boosting Regression model of Query by Committee strategy has the highest extrapolation accuracy (the Pearson $R = 0.95$ on untrained systems). Based on the prediction from the machine learning models, binary pnictides, vacancy, and small atom-containing chalcogenides are predicted to have large PFs. The bonding analysis reveals that the alterations of anionic bonding networks due to small atoms are beneficial to the PFs in these compounds.

npj Computational Materials (2020)6:171 ; https://doi.org/10.1038/s41524-020-00439-8

INTRODUCTION

Thermoelectric (TE) materials have aroused widespread interest owing to their potential applications in waste heat harvesting and refrigeration[1–3]. The conversion efficiency of TE materials is evaluated by the dimensionless TE figure-of-merit ZT, defined as $ZT = \frac{S^2\sigma T}{\kappa_L + \kappa_e}$, where S, σ, κ_L, κ_e, and T, respectively, stand for the Seebeck coefficient, electrical conductivity, lattice thermal conductivity, electronic thermal conductivity, and the absolute temperature. Because of the intercorrelation between the transport parameters, the improvement of ZT values is challenging[4–6].

As computational materials science is emerging, the high-throughput (HTP) calculation methods have been introduced to the TE material field. In 2014, Carrete et al. scanned ~79,000 half-Heusler structures and finally recommended 3 semiconductors with low lattice thermal conductivities[7]. Chen et al. screened 25,000 semiconductors out of 48,000 inorganic compounds and performed the calculations of their electrical transport properties[8,9]. In 2018, Xi et al. applied HTP ab initio calculation to 161 p-type chalcogenides and experimentally verified the recommended TE compound $Cd_2Cu_3In_3Te_8$ with ZT >1.0[10]. Li et al. studied both p-type pnictides and chalcogenides in the atomic ratio 1:1:2, and pnictides showed exceptionally high power factors (PFs)[11].

Although HTP theoretical and experimental means bring a revolutionary leap in predicting properties of energy materials, their scales are limited by the high cost. Meanwhile, data-driven machine learning (ML) methods have attracted a lot of attention because it can efficiently search the huge space with extremely low cost. Recently, ML has been widely used in the development and design of TE materials. In 2017, Zhan et al. trained the ML model based on the collected experimental thermal boundary resistance data and achieved better prediction accuracy than the commonly used acoustic mismatch model[12]. In 2018, Miller et al. viewed diamond-like semiconductors from the perspective of carrier concentration range with ML method and quantified their dopabilities by linear regression[13]. In 2019, an ML model for predicting the κ_L was proposed based on the experimentally measured κ_Ls of ~100 inorganic materials[14]. In the same year, Tshitoyan et al. employed the text mining method on the material literature and sought potential TE materials by their similarities with the word "thermoelectric"[15].

In most of the ML works, the train–test splitting scores or cross-validation results are usually adopted to evaluate the accuracy of the ML models[16]. However, the high scores on the testing set do not necessarily represent superior extrapolation ability. On the other hand, the model extrapolation plays a decisive role in seeking potential materials. Although some algorithms can improve the model extrapolation ability in some degree[17], the poor extrapolation performance is fundamentally inevitable due to the lack of information outside the data set. Thus iterative data verification that provides external information to ML models is a promising method to improve the model extrapolation. To build reliable models with as less validation samples as possible, active learning, a verification-by-learning framework, is suitable[18].

In this work, active learning is used in the TE field to accurately predict the p-type PFs. Our active learning loop contains both the ML module and density functional theory (DFT) verification. As long as the extrapolation accuracy of the model is not high enough, the DFT verification will continue to provide reliable data to the ML module. We adopt three strategies of selecting validation candidates, including Top, Random, and Query by

[1]Materials Genome Institute, Shanghai University, 200444 Shanghai, China. [2]Qianweichang College, Shanghai University, 200444 Shanghai, China. [3]Department of Chemistry, College of Science, Shanghai University, 200444 Shanghai, China. [4]Material Phases Data System, CH-6354 Vitznau, Switzerland. [5]Department of Physics and Shenzhen Institute for Quantum Science & Technology, Southern University of Science and Technology, 518055 Shenzhen, Guangdong, China. [6]Guangdong Provincial Key Lab for Computational Science and Material Design, and Shenzhen Municipal Key Lab for Advanced Quantum Material and Device, Southern University of Science and Technology, 518055 Shenzhen, Guangdong, China. ✉email: jiongy@t.shu.edu.cn; zhangwq@sustech.edu.cn

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

npj nature partner journals

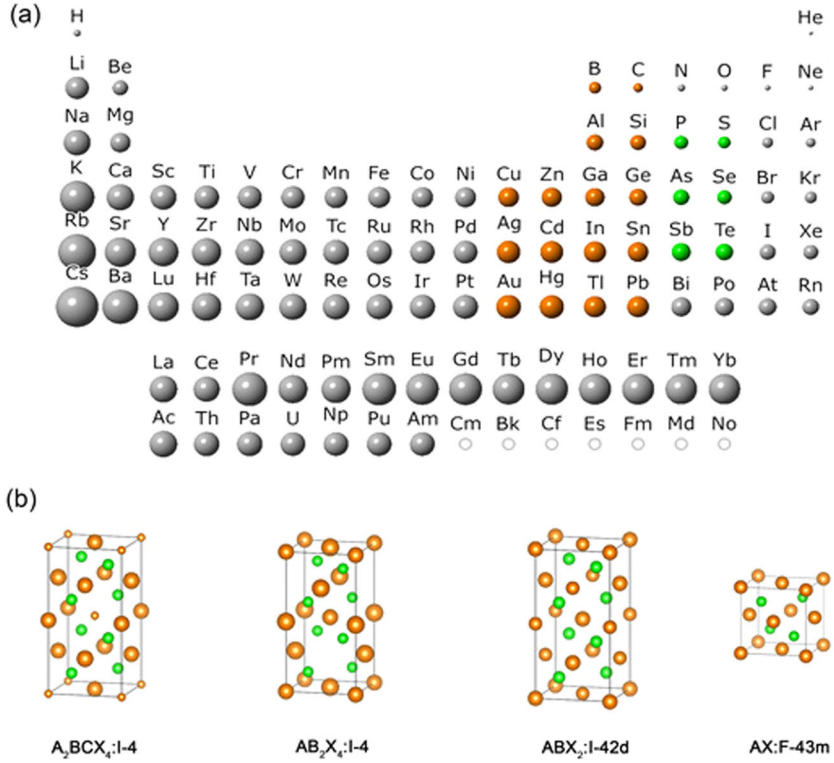

Fig. 1 The search space and schematic diagram of the crystal structure. a Element replacement range; orange indicates the cation range, green is the anion range. **b** Schematic diagram of four types of diamond-like materials.

Committee (QBC)[19]. The model with the highest extrapolation accuracy comes from the QBC strategy. Finally, the bonding analysis on the screened high PF compounds is conducted to reveal the physical reasons for the good TE performance.

RESULTS

Data source

The diamond-like materials investigated include four types of compositions (Fig. 1b), ABX_2, AX, A_2BCX_4, and A_2BX_4, where the X site is a chalcogen or phosphorus group element. The atoms on A, B, and C sites are ordered by the valence of the elements among the IB, IIB, IIIA, and IVA groups (orange-marked in Fig. 1a). The original 158 entries of chalcogenides and pnictides are quoted from our previous HTP works and referred in the DFT database later[10,11]. By exhausting all the possible combinations among the aforementioned cations and anions, we construct a search space of diamond-like compounds with 482 entries (158 DFT calculated and 324 uncalculated)[10]. The target properties, maximum p-type PFs, are calculated with the constant electron–phonon coupling approximation (with the uniform deformation potential 4 eV and Young's modulus 100 GPa) at 700 K under optimized carrier concentrations in theory, similar to our previous works[10,11]. PF obtained by this method purely reflects the influence of electronic structure on group velocities and electronic relaxation times.

Active learning workflow

A classic active learning strategy, Bayesian optimization, has been used many times to find materials with breakthrough properties. These works prove the effectiveness of active learning[20,21]. However, models in Bayesian optimization are limited to the probabilistic regression ones, excluding many other ML methods that also have outstanding performance, such as Support Vector Regression (SVR)[22]. In this work, the active learning strategies with unlimited model types are adopted to integrate active learning with more effective ML algorithms.

Figure 2 shows our active learning loop with the key ingredients, i.e., the search space including DFT database, the ML module (including the models and strategies for candidate selections), and the DFT verification module. In order to be available for both calculated and uncalculated materials with diamond-like structures, the descriptors are all element-related, such as valence electron number, atomic weight, electronegativity, Mendeleev number, etc., with ~60 descriptors per atom. The reason for not taking structure-related descriptors into account is that their generation for the uncalculated materials require the DFT structural relaxation, which is costly if applied to the whole search space.

Based on the DFT database, the models to predict the unexplored materials are built by ML algorithms. Then the candidate selection strategies are carried out according to the model results. There are three strategies, including one several-model strategy and two single-model strategies. The several-model strategy means that the selection of candidates requires the prediction results from multiple different models. In this work, QBC is a several-model strategy in which 15 candidates with large ambiguity are selected. The ambiguity is measured by the variance of five ML models, respectively, using different algorithms, SVR[22], Gradient Boosting Regression (GBR)[23], Random Forest Regression (RFR)[24], Adaptive Boosting Regression[25], and Kernel Ridge Regression (KRR)[26]. The other two single-model strategies are, respectively, Top and Random. Top strategy chooses the 15 candidates with high predicted PFs, and Random strategy just randomly recommends the candidates.

In each round, the recommended 15 candidates are verified by the DFT calculations. Based on the package TransOpt of the electrical transport calculation method and the HTP workflow, the entire verification process can be automatically proceeded[10,27]. Since the validation set is independent of model-learned data, the

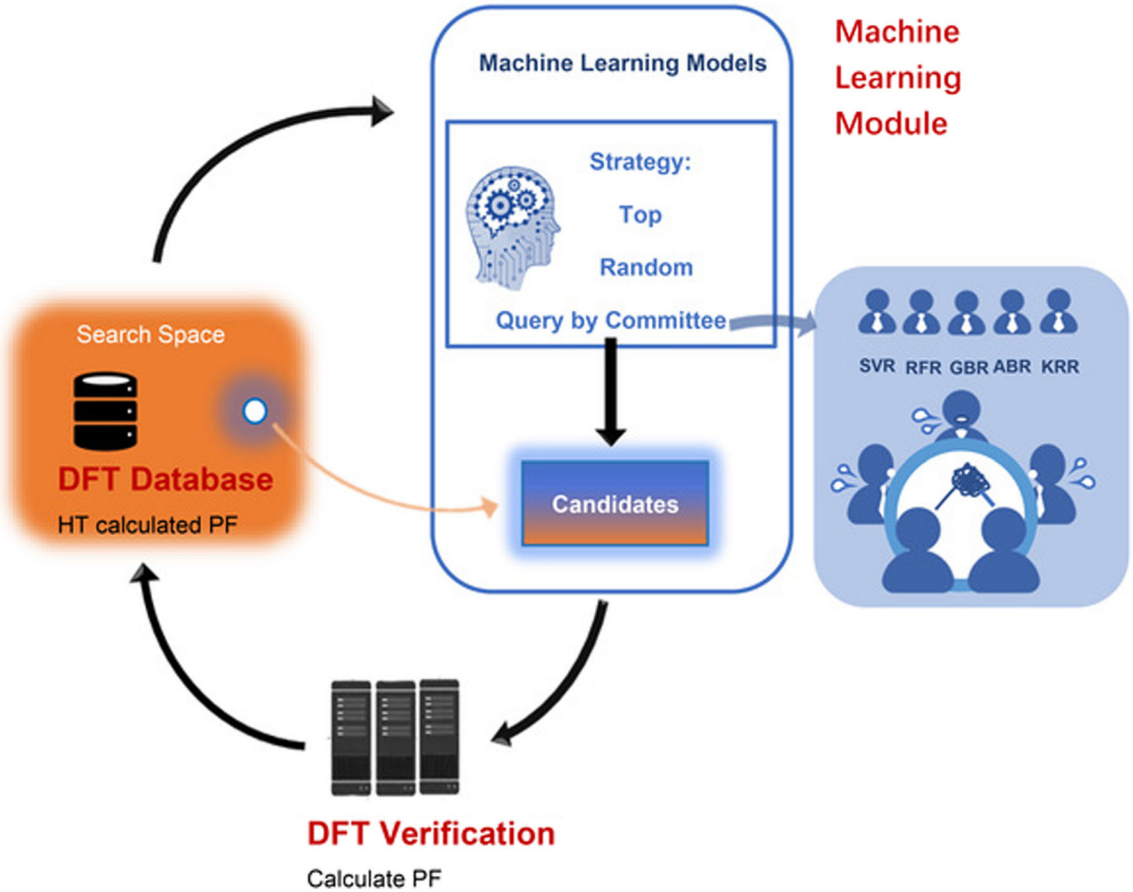

Fig. 2 The workflow of active learning loop. The loop contains three ingredients, Search Space including DFT database, machine learning module and DFT verification. Machine learning model is built based on the DFT database to recommend candidates, and then the candidates are verified with DFT calculations. If the extrapolation accuracy does not meet the convergence criterion, the verification data will be merged into the DFT database, and the whole loop will proceed.

score of the validation set can be regarded as the measure of the model extrapolation accuracy. If the extrapolation is not satisfactory, the already verified samples will be added to the DFT database and the whole loop updated. The active learning loop is terminated when the extrapolated Pearson R is >0.9 or the number of iterations reaches the set maximum 10.

Active learning results

From Fig. 3a, the root mean square error (RMSE) curves of all strategies with RFR algorithm show a generally decreasing trend with the number of iterations. Notably, the results of the first round in the active learning loop are equivalent to the performance of supervised learning for the untrained data (RMSE ~20 μW cm^{-1} K^{-2}, Pearson R −0.11). The poor performance of the first round implies that the introduction of active learning is essential for ML methods to improve the prediction power on unknown data. The performance of Random and QBC strategies is similar; the falling RMSE curve and the rising Pearson R curve show that the accuracies are gradually improved with the number of iterations. Although there is a small range of RMSE fluctuations (~ 3 μW cm^{-1} K^{-2}) in both Random and QBC strategies, it is reasonable because of the sample difference in each iteration. However, the Pearson R curve of the Top strategy does not maintain an upward trend after the first round, indicating that the extrapolation ability of the Top strategy does not improve with iterations. Nevertheless, the RMSE curve has a slight downward trend. It is possibly caused by lowered absolute values of PFs due to the nature that the Top strategy selects candidates with PFs from high to low.

Because the PFs cover a large range of absolute values (10–100 μW cm^{-1} K^{-2}), RMSE cannot fully describe the accuracy of models. Therefore, we introduce a measure for the relative error, i.e., the mean absolute percentage error (MAPE). The formula is expressed as $\mathrm{MAPE} = \sum_{i=0}^{n} \left|\frac{\mathrm{PF_{DFT}} - \mathrm{PF_{pre}}}{\mathrm{PF_{DFT}}}\right| \times \frac{100}{n}\%$, where n represents the number of samples in each iteration. As shown in Supplementary Fig. 3 with both RMSE and MAPE, there is no downward trend in the MAPE curve after the first round of the Top strategy. The overall trends of RMSE and MAPE curves are similar. After the sixth generation, the values of MAPE for QBC and Random strategies basically fluctuate between 10 and 15%, while the MAPE values for Top strategy float between 20 and 30%.

As shown in Fig. 3b, all the ML models in the QBC strategy eventually converged to high accuracies indicated by low RMSE (~4 μW cm^{-1} K^{-2}) and high Pearson R (>0.9, Supplementary Fig. 1). These models have been improved tremendously after ten round iterations, especially for the KRR model. From the results of the first round, the RMSE of KRR model reaches the maximum 40 μW cm^{-1} K^{-2}. Figure 3c, d show the data deviations of predicted and DFT PFs in the first and tenth round, respectively. From Fig. 3c, the sample points of KRR are the farthest from the line with a slope of 1, implying that KRR model performs the worst. Some PF values of the predictions of KRR model are even unreasonably negative. On the other hand, after ten rounds of iterations, the points of all algorithms, including KRR, are obviously close to the line with a

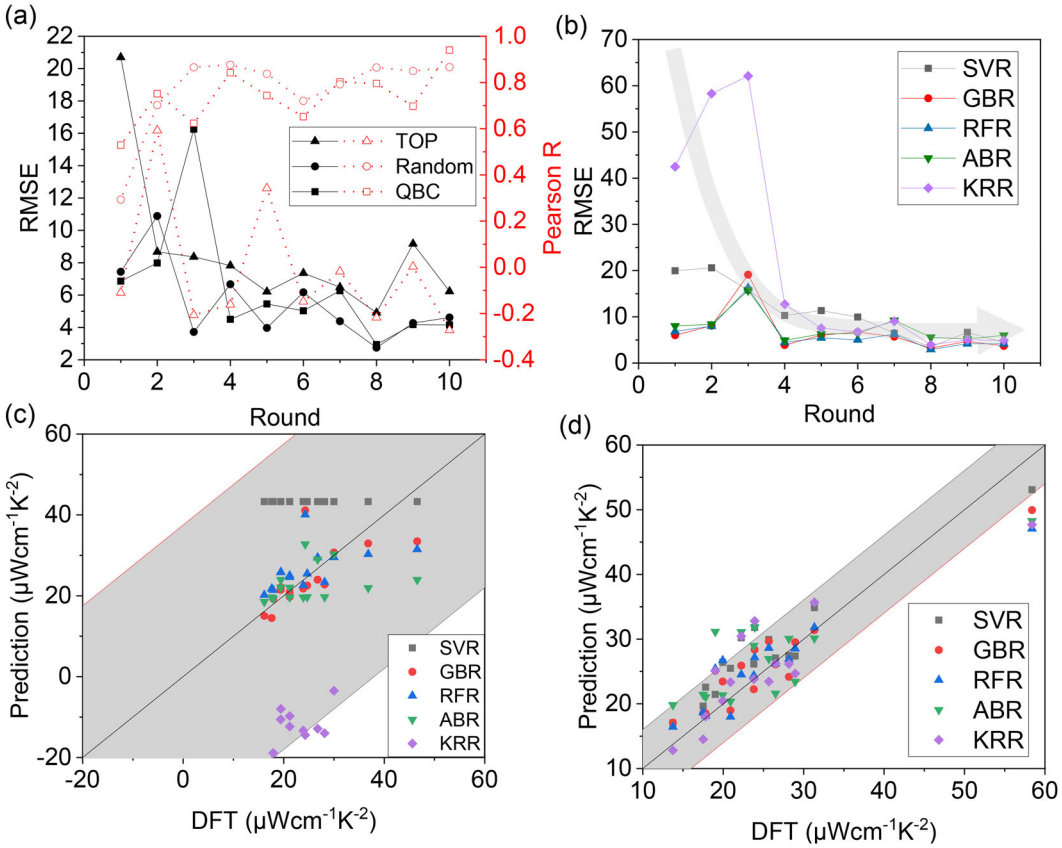

Fig. 3 Extrapolation accuracy results of different strategies. a Comparison of Pearson *R* and RMSE of RFR algorithms in, respectively, Top, Random, and QBC strategies. **b** Changes in RMSE of five algorithms in QBC strategy with the number of iterations. **c** The comparison of the first-generation validation set between the ML prediction and the DFT calculation with QBC strategy. The shaded part indicates the largest root mean square error (RMSE) range of the five algorithms. **d** The comparison of the last-generation (tenth) validation set between the ML prediction and the DFT calculation with QBC strategy.

slope of 1 (Fig. 3d). The dramatic improvement demonstrates that the DFT verification provides sufficient external data to enhance their extrapolation capabilities.

The efficiency of the selection strategy can be considered from two aspects, the divergence and information. Compared with the other two strategies, the candidates of the Top strategy are localized in high PF area in each iteration (low divergence). The accuracy of the model still increases in the first round because the data in the high PF area is sparse (high information). After the first round, the provided PFs contain less information due to the decreasing of the data sparsity. The low divergence of the Top strategy sometimes reduces the extrapolation ability. On the other hand, the divergence of the QBC strategy is comparable to the Random strategy in the PF prediction. Based on the fact that the Random and QBC perform comparably (Fig. 3a), thus in the case of PF prediction, the data divergence plays a vital role.

Material analysis

The last-round GBR algorithm in the QBC strategy, which performs the best (Pearson *R* 0.95), was used to predict the p-type PFs of the whole search space. The compounds in top 20% PFs are shown in Fig. 4a. The overall TE performance depends not only on PFs but also on many other factors. Here we choose two other parameters for further screening potential high-performance TE materials, including "band gap," relating to electrical properties, and "average atomic weight," relating to lattice thermal conductivity. The band gap criterion is 0.7 ± 0.4 eV, considering the uncertainties of the band gap in DFT calculations and the optimal band gaps for TE applications ($10k_BT_{op}$, where T_{op} is the operating temperature)[28]. In addition, the compounds with average atomic weight >80 might have low lattice thermal conductivities and therefore be screened out. Figure 4a shows the results under the two criteria with the highlighted box. The compounds with a relatively large PF are marked with triangles and their chemical formulas are labeled, and they are HgB_2Te_4, $ZnSiSb_2$, $AuBSe_2$, Zn_2GeTe_4, and Zn_2SnTe_4. Combining with the calculations of electronic and lattice thermal conductivities[11], the DFT predicted ZT_{max}s at 700 K of HgB_2Te_4, $ZnSiSb_2$, $AuBSe_2$, Zn_2GeTe_4, and Zn_2SnTe_4 are, respectively, 1.19, 0.97, 1.26, 1.30, and 1.41. ZT_{max} represents the maximum value of DFT-calculated ZT when the carrier concentration is fully optimized within the range of 5×10^{19}–1×10^{21} cm^{-3}.

In order to explore the underlying mechanisms for high PFs, all the compounds with top 20% PFs and the corresponding optimal carrier concentrations are plotted in Fig. 4b (pnictides) and Fig. 4c (chalcogenides). Three major phenomena relating to the PFs can be concluded: (1) among all the studied diamond-like materials, the PFs of pnictides are generally larger; (2) among chalcogenides, the PFs of the compounds in IIB_1:$IIIA_2$:VIA_4 atomic ratio are relatively large; (3) the PFs of IIB_1:$IIIA_2$:VIA_4 chalcogenides with smaller atomic radius elements such as Si or B are relatively large.

Pnictides own extremely high PFs, mainly due to the low valence band effective masses, and therefore high group velocities and low scattering phase space in relaxation times[11]. For quantitative comparison, we calculated the effective masses and group velocities of the pnictide GaAs and chalcogenide ZnSe. The effective mass of the valence band maximum (VBM) in GaAs (2.12 m_e, m_e is the mass of a free electron) is smaller than that of

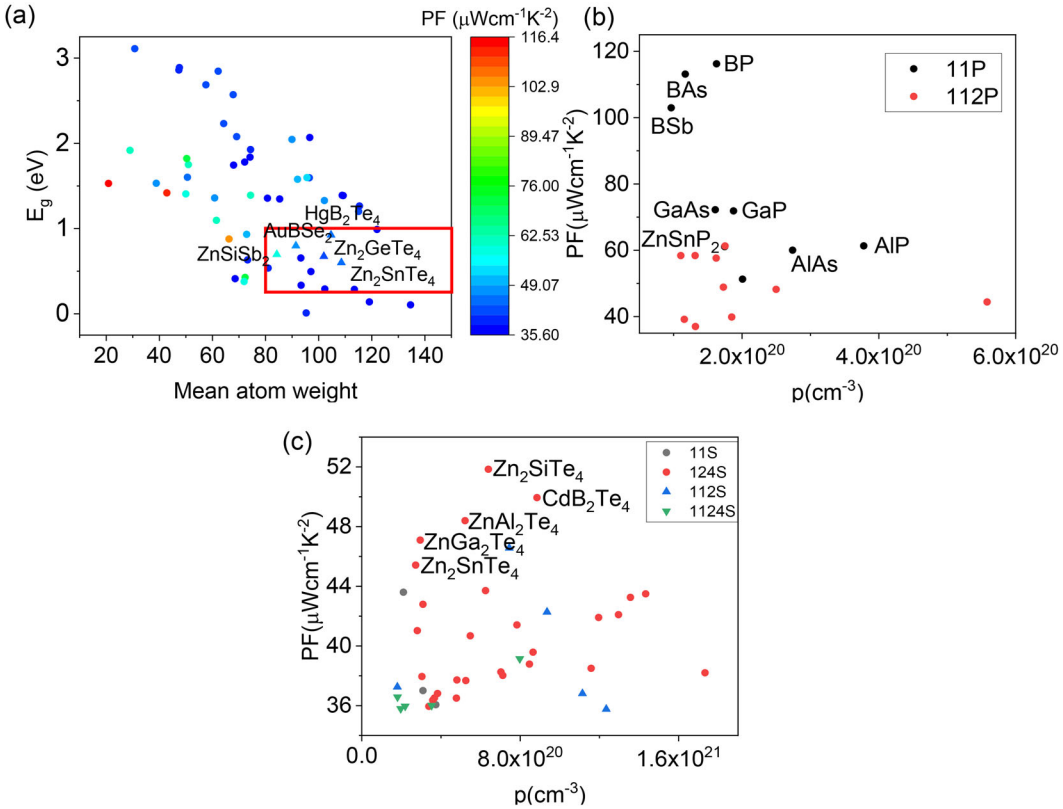

Fig. 4 The distribution of the top 20% ML-predicted PFs. a Band gaps and average atomic weights for compounds with the top 20% ML-predicted PFs. **b** Pnictides with top 20% of PF and PF > 60 μW cm^{-1} K^{-2} is marked. 11P and 112P means the pnictides in the form of AX and ABX_2. **c** Chalcogenides with top 20% of PF and PF > 45 μW cm^{-1} K^{-2} is marked. 11S, 124S, 112S, and 1124S, respectively, means the pnictides in the form of AX, AB_2X_4, ABX_2, and A_2BCX_4.

ZnSe (2.62 m_e), and the electron group velocity of GaAs (2.93×10^5 m s^{-1}) is higher than that of ZnSe (2.13×10^5 m s^{-1}). Meanwhile, the relaxation time of GaAs (4.83×10^{-14} s) is increased compared to that of ZnSe (1.50×10^{-14} s).

Observing chalcogenide in compounds with the top 20% p-type PFs (Fig. 4b), we found that a large percentage of compounds are in IIB_1:$IIIA_2$:VIA_4 atomic ratio. This conclusion is consistent with our previous work[10]. From the crystal structures, this series of compounds can be seen as vacancy-containing chalcogenides (VCCs). In order to further explain why the PFs of VCCs are relatively high, two compounds with similar atomic masses but in different chemical formulas, $ZnGa_2Te_4$ and $CuGaTe_2$, were investigated (Supplementary Fig. 2). We introduce the energy integral of the negative density of energy (-DOE) at the VBM to quantify the degree of the destabilizing contribution, which is written as $E_{band} = \int_{E_f-2}^{E_f} -\mathrm{DOE}(E)\mathrm{d}E$[29]. The E_{band} of $ZnGa_2Te_4$ and $CuGaTe_2$ are, respectively, −19.67 and −31.89 eV. A smaller E_{band} means that the anti-bonding interaction at VBM is weaker, resulting in a flat band structure and high density of states (DOS) at the Fermi levels (Supplementary Table 1). Although the relaxation time and group velocity are slightly decreased, the electrical conductivity increased significantly due to the large enhancement in DOS and carrier concentration.

In addition to vacancies, the lattice distortion caused by small atoms might further increase the PFs. For example, both Zn_2SnTe_4 and Zn_2SiTe_4 are in the vacancy-containing structure, but the PF of Zn_2SiTe_4 is ~6 μW cm^{-1} K^{-2} higher than Zn_2SnTe_4. From the view of the structure, small silicon atoms cause the short Si-Te bonds, thereby shortening the distance between neighboring Te atoms (Fig. 5a). The anti-bonding interactions are raised between the originally non-interacting Te-Te in Zn_2SiTe_4 (Fig. 5c). Comparing with the band structure of Zn_2SnTe_4 (Fig. 5b), the anti-bonding interaction of Te-Te leads to the increase of band energy at X point, causing a better band convergence with the VBM at Γ point.

DISCUSSION

The scores of the train-test splitting in supervised learning models are generally good; however, the accuracy of extrapolation could be very poor. In most material problems, the reason for the inaccurate extrapolation results from ML models lies in the lack of samples. Therefore, a method of guiding material exploration is needed, which aims at providing reasonable estimate of the material property in the whole search space by supplying a small scale of samples. Hence, active learning, a framework for updating ML models through external verification, is implemented to improve the extrapolation accuracy, exemplified by the TE PFs for chalcogenides and pnictides with diamond-like structures. Several candidate selection strategies in active learning are tested. Finally, the extrapolation accuracy of the GBR model in QBC strategy is the highest (Pearson R 0.95), ensuring the reliability of extrapolation. Hence, this model is applied to predict the full search space to seek high PF materials. Materials with the top 20% PFs are analyzed by band structures and bonding conditions. It is found that the diamond-like materials with three special structures are more likely to have higher PFs: (1) binary pnictides, (2) IIB_1:$IIIA_2$:VIA_4 compounds with VCC structure, and (3) materials containing elements with small atomic radii. This work demonstrates the ability of active learning on accurately proposing potential materials based on small sample set.

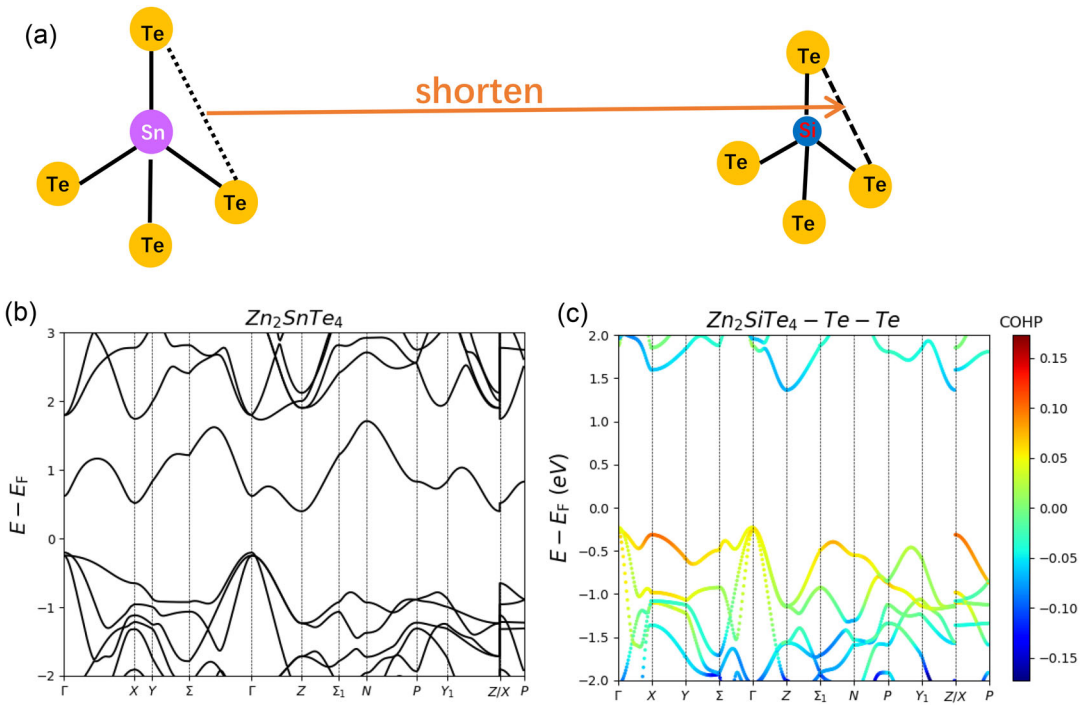

Fig. 5 Schematic diagram of the bond length and band structure of Zn_2SnTe_4 and Zn_2SiTe_4. a Schematic diagram of lattice distortion. **b** Band structure of compound Zn_2SnTe_4. **c** The band-resolved projected crystal orbital Hamilton populations (COHPs) for Te-Te offsite of Zn_2SiTe_4.

METHODS

DFT computational methods

DFT calculations are carried out using projector augmented wave method as implemented in the Vienna ab initio Simulation Package[30,31]. Perdew–Burke–Ernzerhof-type generalized gradient approximation (GGA) is applied as exchange–correlation functional[32]. Self-consistent calculation is performed with an energy convergence criterion of 10^{-4} and 520 eV plane-wave energy cutoff. The strongly constrained and appropriately normed meta-GGA potential is adopted[33]. The Monkhorst–Pack uniform *k*-point sampling was used with $k = 180/L$ (L represents the lattice parameter) for electrical transport properties[34]. Chemical-bonding information was obtained using the band-resolved projected crystal orbital Hamilton populations as implemented in the Local Orbital Basis Suite Towards Electronic-Structure Reconstruction package[35–39].

In order to get the ZT value, the electrical properties, including the Seebeck coefficient, electrical conductivity, and the electronic thermal conductivity are calculated by Boltzmann transport theory. The lattice thermal conductivity is obtained by the Slack model, which has proved to be suitable for diamond-like compounds[11,40,41].

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Received: 2 August 2020; Accepted: 29 September 2020;
Published online: 10 November 2020

REFERENCES

1. Goldsmid, H. *Thermoelectric Refrigeration* (Springer, 2013).
2. Sales, B. C. Smaller is cooler. *Science* **295**, 1248–1249 (2002).
3. Tritt, T. & Rowe, D. *Thermoelectrics Handbook: Macro to Nano* (CRC Press, Boca Raton, FL, 2005).
4. Liu, W., Yan, X., Chen, G. & Ren, Z. Recent advances in thermoelectric nanocomposites. *Nano Energy* **1**, 42–56 (2012).
5. Zhu, T. et al. Compromise and synergy in high-efficiency thermoelectric materials. *Adv. Mater.* **29**, 1605884 (2017).
6. Yang, J. et al. On the tuning of electrical and thermal transport in thermoelectrics: an integrated theory–experiment perspective. *npj Comput. Mater.* **2**, 15015 (2016).
7. Carrete, J., Li, W., Mingo, N., Wang, S. & Curtarolo, S. Finding unprecedentedly low-thermal-conductivity half-heusler semiconductors via high-throughput materials modeling. *Phys. Rev. X* **4**, 011019 (2014).
8. Chen, W. et al. Understanding thermoelectric properties from high-throughput calculations: trends, insights, and comparisons with experiment. *J. Mater. Chem. C* **4**, 4414–4426 (2016).
9. Ricci, F. et al. An ab initio electronic transport database for inorganic materials. *Sci. Data* **4**, 170085 (2017).
10. Xi, L. et al. Discovery of high-performance thermoelectric chalcogenides through reliable high-throughput material screening. *J. Am. Chem. Soc.* **140**, 10785–10793 (2018).
11. Li, R. et al. High-throughput screening for advanced thermoelectric materials: diamond-like ABX_2 compounds. *ACS Appl. Mater. Interfaces* **11**, 24859–24866 (2019).
12. Zhan, T., Fang, L. & Xu, Y. Prediction of thermal boundary resistance by the machine learning method. *Sci. Rep.* **7**, 7109 (2017).
13. Miller, S. A. et al. Empirical modeling of dopability in diamond-like semiconductors. *npj Comput. Mater.* **4**, 71 (2018).
14. Chen, L., Tran, H., Batra, R., Kim, C. & Ramprasad, R. Machine learning models for the lattice thermal conductivity prediction of inorganic materials. *Comp. Mater. Sci.* **170**, 109155 (2019).
15. Tshitoyan, V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
16. Mueller, T., Kusne, A. G. & Ramprasad, R. in *Reviews in Computational Chemistry* (eds Parrill, A. L. & Lipkowitz, K. B.) 186–273 (Wiley-Blackwell, 2016).
17. Chin, T. J. & Suter, D. Out-of-sample extrapolation of learned manifolds. *IEEE T Pattern Anal.* **30**, 1547–1556 (2008).
18. Settles, B. *Active Learning Literature Survey* (University of Wisconsin-Madison Department of Computer Sciences, 2009).
19. Burbidge, R., Rowland, J. J. & King, R. D. Active learning for regression based on query by committee. In *International Conference on Intelligent Data Engineering and Automated Learning* (eds Yin, H., Tino, P., Corchado, E., Byrne, W. & Yao, X.) 209–218 (Springer, 2007).
20. Ju, S. et al. Designing nanostructures for phonon transport via bayesian optimization. *Phys. Rev. X* **7**, 021024 (2017).
21. Hou, Z., Takagiwa, Y., Shinohara, Y., Xu, Y. & Tsuda, K. Machine-learning-assisted development and theoretical consideration for the $Al_2Fe_3Si_3$ thermoelectric material. *ACS Appl. Mater. Interfaces* **11**, 11545–11554 (2019).

22. Smola, A. J. & Scholkopf, B. A tutorial on support vector regression. *Stat. Comput.* **14**, 199–222 (2004).
23. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
24. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
25. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**, 119–139 (1997).
26. Robert, C. Machine learning, a probabilistic perspective. *CHANCE* **27**, 62–63 (2014).
27. Li, X. et al. TransOpt. A code to solve electrical transport properties of semiconductors in constant electron-phonon coupling approximation. *Comp. Mater. Sci.* **186**, 110074 (2021).
28. Ioffe, A. Semiconductor thermoelements and thermoelectric cooling. *Phys. Today* **12**, 42 (1959).
29. Küpers, M. et al. Unexpected Ge–Ge contacts in the two-dimensional Ge_4Se_3Te Phase and analysis of their chemical cause with the density of energy (DOE) function. *Angew. Chem. Int. Ed.* **56**, 10204–10208 (2017).
30. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758 (1999).
31. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169 (1996).
32. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
33. Sun, J., Ruzsinszky, A. & Perdew, J. P. Strongly constrained and appropriately normed semilocal density functional. *Phys. Rev. Lett.* **115**, 036402 (2015).
34. Monkhorst, H. J. & Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **13**, 5188 (1976).
35. Maintz, S., Deringer, V. L., Tchougréeff, A. L. & Dronskowski, R. Analytic projection from plane-wave and PAW wavefunctions and application to chemical-bonding analysis in solids. *J. Comput. Chem.* **34**, 2557–2567 (2013).
36. Dronskowski, R. & Blöchl, P. E. Crystal orbital Hamilton populations (COHP): energy-resolved visualization of chemical bonding in solids based on density-functional calculations. *J. Phys. Chem.* **97**, 8617–8624 (1993).
37. Deringer, V. L., Tchougréeff, A. L. & Dronskowski, R. Crystal orbital Hamilton population (COHP) analysis as projected from plane-wave basis sets. *J. Phys. Chem. A* **115**, 5461–5466 (2011).
38. Maintz, S., Deringer, V. L., Tchougréeff, A. L. & Dronskowski, R. LOBSTER: a tool to extract chemical bonding from plane-wave based DFT. *J. Comput. Chem.* **37**, 1030–1035 (2016).
39. Sun, X. et al. Achieving band convergence by tuning the bonding ionicity in n-type Mg_3Sb_2. *J. Comput. Chem.* **40**, 1693–1700 (2019).
40. Slack, G. A. Nonmetallic crystals with high thermal conductivity. *J. Phys. Chem. Solids* **34**, 321–335 (1973).
41. Jia, T., Chen, G. & Zhang, Y. Lattice thermal conductivity evaluated using elastic properties. *Phys. Rev. B* **95**, 155206 (2017).

ACKNOWLEDGEMENTS

This work was supported by the National Key Research and Development Program of China (Nos. 2018YFB0703600 and 2017YFB0701600), Natural Science Foundation of China (Grant Nos. 11674211, 51632005, and 51761135127), and the 111 Project D16002. W.Z. also acknowledges the support from the Guangdong Innovation Research Team Project (No. 2017ZT07C062), Guangdong Provincial Key-Lab program (No. 2019B030301001), Shenzhen Municipal Key-Lab program (ZDSYS20190902092905285), and Shenzhen Pengcheng-Scholarship Program. Part of the calculations were supported by Center for Computational Science and Engineering at Southern University of Science and Technology.

AUTHOR CONTRIBUTIONS

The initial idea was developed by Y.S. and J.Y., and its implementation was discussed with W.Z. The descriptors are provided by P.V. and Y.W. All authors participated in the data analysis and writing and reading of the paper. J.Y. managed the project.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41524-020-00439-8.

Correspondence and requests for materials should be addressed to J.Y. or W.Z.

Reprints and permission information is available at http://www.nature.com/reprints

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