Prediction of thermoelectric performance for layered IV-V-VI semiconductors by high-throughput ab initio calculations and machine learning

Yu Gan, Guanjie Wang, Jian Zhou and Zhimei Sun

Layered IV-V-VI semiconductors have immense potential for thermoelectric (TE) applications due to their intrinsically ultralow lattice thermal conductivity. However, it is extremely difficult to assess their TE performance via experimental trial-and-error methods. Here, we present a machine-learning-based approach to accelerate the discovery of promising thermoelectric candidates in this chalcogenide family. Based on a dataset generated from high-throughput ab initio calculations, we develop two highly accurate-and-efficient neural network models to predict the maximum $ZT$ ($ZT_{\text{max}}$) and corresponding doping type, respectively. The top candidate, n-type Pb$_2$Sb$_2$S$_5$, is successfully identified, with the $ZT_{\text{max}}$ over 1.0 at 650 K, owing to its ultralow thermal conductivity and decent power factor. Besides, we find that n-type Te-based compounds exhibit a combination of high Seebeck coefficient and electrical conductivity, thereby leading to better TE performance under electron doping than hole doping. Whereas p-type TE performance of Se-based semiconductors is superior to n-type, resulting from large Seebeck coefficient induced by high density-of-states near valence band edges.

**INTRODUCTION**

Due to the depletion of fossil fuels and ever-increasing environmental concerns, it is urgent to explore sustainable and clean energies, which has become a global consensus. The thermoelectric (TE) conversion technique creates a very promising opportunity to directly convert waste heat into electricity. The energy conversion efficiency is quantified by the TE materials’ dimensionless figure of merit $ZT$: $ZT = S^2\sigma/(\kappa_\text{el} + \kappa_\text{th})$, where $S$, $\sigma$, $\kappa_\text{el}$ and $\kappa_\text{th}$ are the Seebeck coefficient, electrical conductivity, lattice thermal conductivity, and electronic thermal conductivity, respectively. Undoubtedly, for a material to achieve high TE performance, it should maintain the high power factor (i.e., $S^2\sigma$) and low thermal conductivity (i.e., $\kappa_\text{el} + \kappa_\text{th}$) simultaneously. To date, several methods have been employed to improve the $ZT$ value, including band structure engineering, nanostructure engineering, carrier modulation, and enhancement of phonon scattering through extrinsic strategies, such as the introduction of all-scale hierarchical architectures. Besides, searching crystalline solids with intrinsically low $\kappa_\text{th}$ is a significant avenue, and it is also beneficial to optimize the electrical transport properties independently.

Very recently, we identified a large family of stable layered IV-VI (IV (A) = Si, Ge, Sn, Pb; V (B) = As, Sb, Bi; VI (C) = S, Se, Te) semiconductors (70 compounds in total). It is noted that the chemical composition of the current advanced TE materials is mainly composed of main group IV to VI elements, such as PbTe, Bi/Sb–Te-based alloys, and recently reported Sb$_2$S$_3$Te$_6$, most of which also adopt layered crystal structures. More importantly, all these IV-VI compounds exhibit ultralow lattice thermal conductivity (0.28–2.02 W m$^{-1}$ K$^{-1}$ at 300 K) and suitable electronic bandgaps (>0.1 eV) within the range of good TE materials. The above inherently favorable characteristics indicate the immense potential of this ternary chalcogenide family as TE materials. Thus it is of great importance to explore their TE performance.

It is well known that the traditional trial-and-error approach is a cost-inefficient way to develop unknown materials with target properties, which may even end up with fruitlessness. In contrast, the high-throughput (HT) density functional theory (DFT) calculations have emerged as a powerful and useful tool for accelerating the discovery of various types of functional materials in recent years. However, the full HT calculations will be computationally extremely expensive when identifying suitable materials from numerous candidates. Nowadays, driven by the advances of artificial intelligence technology, the machine learning (ML) approach has become a much attractive technique in materials science, which has been successfully utilized to predict diverse materials properties. By combing HT calculations and ML methodology, therefore, one can not only efficiently search the huge space with low computational cost but also gain quantitative insights into the structure-property relationship.

In this work, using our developed HT intelligent computational platform (ALKEMIE), we first generate a database comprising of a series of descriptors (features) and TE labels (characteristics) of 40 compounds to compose an input data matrix for the ML models training and validating. Then, based on extensive examinations of hyperparameters of deep neural networks, two highly accurate ML models are successfully developed and used to predict the maximum $ZT$ ($ZT_{\text{max}}$) and corresponding doping type at different temperatures for the remaining 30 semiconductors, respectively. Among this IV-VI family, we identify several compounds that exhibit good TE performance with $ZT_{\text{max}}$ values above 0.8 at 650 K, especially the n-type Pb$_2$Sb$_2$S$_5$ achieving a high $ZT_{\text{max}}$ of 1.2. Meanwhile, it is also observed that Se-based compounds have higher $ZT_{\text{max}}$ under p-type doping compared to n-type doping while Te-based compounds show better n-type TE performance.

---

1School of Materials Science and Engineering, Beihang University, Beijing 100191, China. 2Center for Integrated Computational Materials Engineering, International Research Institute for Multidisciplinary Science, Beihang University, Beijing 100191, China. 3These authors contributed equally: Yu Gan, Guanjie Wang. *email: zmsun@buaa.edu.cn

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences
than p-type. Finally, with the intention of understanding this behavior and unraveling the origin of the excellent TE performance of Pb$_2$Sb$_2$S$_5$, we comprehensively investigate the electronic structures, electrical and phonon thermal transport properties.

RESULTS

Dataset for machine learning

The magnitude of ZT directly describes the performance of a TE material: the larger the ZT value, the higher the TE energy conversion efficiency. Meanwhile, the optimal doping type (i.e., n-type or p-type) at which the maximum ZT value is achieved for a given temperature provides significant insights for experimental research. Therefore, the ZT$_{\text{max}}$ and corresponding optimal doping type are used as the two TE labels for ML prediction. It is known that the input dataset is one of the key components in developing efficient ML models. Here, we created a database of altogether 480 instances comprising of 40 compounds, as each compound is investigated at 12 different temperature levels (from 100 to 650 K with an interval of 50 K). These 40 semiconductors are selected randomly from the 70 compounds to ensure the full coverage of all data (see "Methods" section for details). Each instance is described by 31 initial descriptors: the number of atoms, valence electronic configurations, features related to the periodic table (row numbers and atomic numbers), atomic and covalent radii, electronegativities, average atomic mass, lattice constants, interatomic bond lengths, and temperature (see Supplementary Table 1 of Supplementary Information for details). Figure 1a shows the correlation coefficients between all 31 features. Clearly, many different features exhibit extremely high correlation coefficients closely approached to 1, indicating that they are strongly coupled. In order to avoid the extremely complex ML models, it is very important to select a suitable number of descriptors that could perfectly reflect the TE performance. Hence, we have performed a feature reduction by removing those redundant descriptors with correlation coefficients above the absolute 0.90 (Supplementary Fig. 1), leaving 11 optimal features (Fig. 1b), which are very easy to obtain without any complicated and expensive calculations. Finally, an input data matrix used to build ML models is obtained, in which the total number of data points is 5280 (i.e., 480 × 11). We divided this dataset into 75% training (3960 data points comprising of 30 compounds) and 25% validation sets (1320 data points containing 10 compounds). In addition, for preventing overfitting, 70% of the training data is adopted for training ML models, while the remaining 30% is used for real-time testing.

Machine learning models training

Owing to the rapid development of computing power and the acceleration of graphics processing unit (GPU), here we directly employed the deep neural networks as the machine learning model. To avoid overfitting or underfitting models and ensure high accuracy simultaneously, we adopted three methods of adding batch norm layer, dropout layer, and cross-validation. In addition, a set of appropriate hyperparameters is critical for obtaining effective neural network models, which generally includes the activation function, the optimization algorithm, the number of hidden layers, the number of nodes in each hidden layer, and the initialization of weights and bias. We have extensively tested the effects of different hyperparameters on the convergence and accuracy of the ML models, as shown in Fig. 2.

Firstly, we examined three popular types of activation functions, namely, Sigmod, Tanh, and rectified linear unit (Relu)37. As seen in Fig. 2a, the mean square error (MSE) of the training set using the Sigmod activation function is 0.00531, which is highly comparable to that of Tanh (0.00576) (Fig. 2b) and is relatively lower than the Relu (0.01037) (Fig. 2c), certifying the higher training accuracy of the Sigmod and Tanh functions. However, the root mean square error (RMSE) of Sigmod (0.16156) and Tanh (0.27545) is significantly larger than the value of Relu (0.06520), indicating the much lower prediction accuracy of the former two functions, as clearly evidenced in Fig. 2a–c (Sigmod: 83.844%, Tanh: 72.455%, and Relu: 93.480%). Also, Sigmod and Tanh functions appear serious underfitting problems, as the MSE of their testing dataset is substantially larger than that of the training dataset. In contrast, the Relu function exhibits a better fitting behavior because both training and testing datasets have similar learning curves and comparable MSE values, in spite of minor deviations. Meanwhile, among these three activation functions, Relu also possesses the largest coefficient of determination value38 ($R^2 = 0.95014$), substantiating its superior globally fitting performance compared to the other two functions, and thus the Relu function is selected. Then, based on the Relu activation function, we tested different backpropagation algorithms39. By comparing with the Adaptive Moment Estimation (Adam) (Fig. 2c), the convergence processes during training and testing of stochastic gradient descent (SGD)41 are almost the same, and the corresponding MSE difference could be ignored (Fig. 2d), although SGD has a slower convergence speed and moderately smaller prediction accuracy (91.765%). Hence, the algorithm of SGD is adopted as the optimizer to further eliminate the underfitting problem that existed in Adam. In addition, using the Relu activation function and the SGD optimizer, we also demonstrated that the enhancement of hidden layers (Fig. 2e) and nodes (Fig. 2f) has little impact on the fitting performance but increases the training cost. Therefore, the combination of Relu activation function, SGD optimization algorithm, and three hidden layers with corresponding nodes of 100, 50, and 20 were selected as the favorable ML model (defined as mode-1 here) for predicting ZT$_{\text{max}}$.

Notably, the prediction of ZT$_{\text{max}}$ is a kind of continuous problem, while the estimation of optimal doping type is the
classification problem, which has only two outputs (i.e., n-type and p-type) and thus is more prone to overfitting or underfitting. Hence, we trained a different ML model through transfer learning from model-I to predict optimal doping type at different temperatures. Supplementary Fig. 2 shows the testing process for various hyperparameters. In order to avoid overfitting or underfitting behaviors and to reduce the training error simultaneously, the combination of Tanh activation function, SGD algorithm and four hidden layers with corresponding nodes of 100, 100, 50, and 20 have been adopted (model-II) to estimate the optimal doping type (Supplementary Fig. 2e). This is because the cross-entropy difference between the training (0.20965) and testing datasets (0.20990) in model-II is negligible, suggesting the non-existence of overfitting and underfitting situations. Besides, it also shows the lowest training error compared to the other two non-overfitting and non-underfitting ML models.

**Fig. 2** Machine learning model. ML-predicted $ZT_{max}$ based on the deep neural networks with different activation functions, optimizers, and hidden layers, in comparison with the DFT-calculated values. a-c ML models using Adam optimization algorithm, three hidden layers of [100, 50, 20] and a Sigmoid, b Tanh, or c Relu activation functions. d-f ML models using Relu activation function, SGD optimization algorithm and d three hidden layers of [100, 50, 20], e four hidden layers of [100, 100, 50, 20], or f four hidden layers of [500, 100, 50, 20]. The insets indicate the convergence process and mean square error (MSE) during training (blue line) and testing (red line).
Pb$_2$Sb$_2$S$_5$, Pb$_2$Sb$_2$Se$_5$, and Pb$_2$Sb$_2$Te$_5$ using the generalized properties, using three Pb$_2$Sb$_2$VI$_5$ (VI systematically investigated the electronic structures and transport doping type difference in Se- and Te-based compounds, we where the MSE is ~0.008, semiconductors show p-type semiconductors have been identified, which exhibit the $ZT_{\text{max}}$ above 0.8 at 650 K, especially the top compound Pb$_2$Sb$_2$S$_5$ having $ZT_{\text{max}}$ of 1.2 under n-type doping. Moreover, for a given temperature, it is interesting to observe that Te-based compounds achieve the $ZT_{\text{max}}$ under n-type doping while the Se-based semiconductors show p-type $ZT_{\text{max}}$. Such behavior also appears in almost all input Se/Te-based compounds used to train and validate ML models (Supplementary Fig. 4). In order to explore the superior n-type TE performance of Pb$_2$Sb$_2$S$_5$ and understand the doping type difference in Se- and Te-based compounds, we systematically investigated the electronic structures and transport properties, using three Pb$_2$Sb$_2$VI$_5$ (VI = S, Se, Te) compounds as examples in the following.

**Electronic structures**

Figure 4 shows the calculated electronic band structures of Pb$_2$Sb$_2$S$_5$, Pb$_2$Sb$_2$Se$_5$, and Pb$_2$Sb$_2$Te$_5$ using the generalized gradient approximations (GGA) of Perdew–Burke–Ernzerhof (PBE) exchange-correlation functional. Note that we have demonstrated the band dispersion features determined by the Heyd–Scuseria–Ernzerhof screened hybrid functional (HSE06) are quite similar to those obtained from the GGA-PBE method (Supplementary Fig. 5), indicating that the HSE06 approach has little effect on the shape of the band structures. Therefore, the efficient GGA-PBE band structures instead of the computational expensive HSE band structures were used in this work. It is clearly seen that all the three compounds are direct-gap semiconductors, as their conduction band minimum (CBM) and valence band maximum (VBM) are both located at the $\Gamma$ point (Fig. 4a–c). The resulting bandgaps are 0.53, 0.12 and 0.34 eV for Pb$_2$Sb$_2$S$_5$, Pb$_2$Sb$_2$Se$_5$, and Pb$_2$Sb$_2$Te$_5$, respectively, which are appreciably lower than those determined from the HSE06 functional (0.91 eV for Pb$_2$Sb$_2$S$_5$, 0.47 eV for Pb$_2$Sb$_2$Se$_5$, and 0.60 eV for Pb$_2$Sb$_2$Te$_5$). To eliminate the bipolar conduction effect, therefore, the more accurate HSE bandgaps were used in our calculations of electrical transport properties (see below). In addition, these three compounds exhibit similar projected band structures (Fig. 4d–f). Specifically, the valence bands are basically dominated by the p electrons of VI atoms, and the conduction bands are mainly contributed from Pb 6p and Sb 5p states. Since the Pb/Sb p orbitals have slight contributions to the valence band edge, it is possible to modulate the hole carrier concentration via Pb/Sb defect engineering without considerably influencing the valence band structures, thereby enhancing the TE performance. For example, the isosctructural compound Ge$_2$Sb$_2$Te$_5$ gives the maximal $ZT$ value of 0.78 at 700 K by substituting the Ge sites with In, ~2-fold improvement compared with the undoped system because the introduction of indium as a potentially donor-like dopant lowers the hole carriers density. From band structures, it is also found that the highest valence band (HVB) of Pb$_2$Sb$_2$S$_5$ is much flatter than that of Pb$_2$Sb$_2$Se$_5$ and Pb$_2$Sb$_2$Te$_5$, indicating the larger density of states (DOS) near the VBM of the former. Meanwhile, Pb$_2$Sb$_2$Te$_5$ is expected to possess the highest DOS near the CBM due to its flattest lowest conduction band (LCB). We will further discuss this later.

To improve the understanding of electronic structures, the projected density of states (PDOS) of Pb$_2$Sb$_2$S$_5$, Pb$_2$Sb$_2$Se$_5$, and Pb$_2$Sb$_2$Te$_5$ are plotted in Fig. 5a–c. It clearly reveals that the top of valence bands is predominated by the VI p electrons, and the electronic states near the CBM primarily come from the p states of Pb and Sb and partly from the VI p orbital. Pb 6p and Sb 5p electrons are widely distributed between $-6$ and 0 eV, which strongly overlap with VI p valence electrons in the same range, suggesting the covalent bonding characters of Pb-VI and Sb-VI bonds. Noticeably, there exists small Pb 6s and Sb 5s peaks close to the VBM, although they are mainly localized in the deep energy.
region below the Fermi level \( E_F \) (−10 to −6 eV). In order to gain further knowledge of the electronic interactions between different atoms, we calculated the projected crystal orbital Hamilton populations (pCOHP)\textsuperscript{46} as shown in Fig. 5d–f. In principle, a negative value of −pCOHP indicates the antibonding interaction while the bonding state exhibits a positive one. Obviously, these three compounds are considered to be electronic stable because no antibonding states are located at the \( E_F \). In the energy range slightly below the \( E_F \), the interactions between \( \text{Pb}/\text{Sb} \) and \( \text{VI} \) p electrons have negative −pCOHP values, indicating the antibonding states of \( \text{Pb} 6s-\text{IV} 5p \) and \( \text{Sb} 5s-\text{IV} 5p \), which would play an important role in the hole carrier transport. It is also found that the \( \text{Pb}/\text{Sb} \) p-\( \text{VI} \) p exhibit positive −pCOHP values in the range from −6 to 0 eV, forming bonding states and hence stabilizing the electronic structure.

Figure 6 illustrates the total DOS of valence and conduction band edges of \( \text{Pb}_2\text{Sb}_2\text{S}_5 \), \( \text{Pb}_2\text{Sb}_2\text{Se}_5 \) and \( \text{Pb}_2\text{Sb}_2\text{Te}_5 \). In general, a rapid change in the DOS with energy is favorable for obtaining a...
large Seebeck coefficient according to the relations given by $^6$

$$S = \frac{8n^2k_B^2}{3e\hbar^2} m_d T \left( \frac{m}{3n} \right)^{2/3},$$  \hfill (1)

$$\text{DOS} = g(E) = \frac{8n (m_d^{2/3} / 2E)^{1/2}}{\hbar^3},$$  \hfill (2)

where $n$ is the carrier concentration and $m_d$ represents the DOS effective mass. It turns out that the valence DOS near the VBM of Pb$_2$Sb$_2$S$_5$ is indeed much larger than that of Pb$_2$Sb$_2$Se$_5$ and Pb$_2$Sb$_2$Te$_5$ (Fig. 6a) while the increase of DOS close to the CBM is considerably faster in Pb$_2$Sb$_2$Te$_5$ (Fig. 6b), which agrees well with our above analysis of band structures. Hence, among these three compounds, we can expect that Pb$_2$Sb$_2$S$_5$ gives the highest Seebeck coefficient under p-type doping and Pb$_2$Sb$_2$Te$_5$ should have the largest value in n-type doping. At the same p-type doping level, the $S$ value of Pb$_2$Sb$_2$S$_5$ is much larger than that of Pb$_2$Sb$_2$Se$_5$ and Pb$_2$Sb$_2$Te$_5$; meanwhile, Pb$_2$Sb$_2$Te$_5$ has the largest n-type $|S|$ among these three compounds. By comparison, it is also observed that the p-type performance of Pb$_2$Sb$_2$S$_5$ and Pb$_2$Sb$_2$Se$_5$ is superior to the n-type, whereas Pb$_2$Sb$_2$Te$_5$ is much better under n-type doping. These results are well consistent with our above analysis of the total DOS around the Fermi energy.

Notably, the electronic relaxation time ($\tau$) is a critical parameter in the calculation of electrical conductivity. Here, by comparing the $\sigma/\tau$ with available experimental data of $\sigma$, the $\tau$ values of four GST compounds were obtained, namely, 10 fs for Ge$_2$Sb$_2$Te$_5$ and Ge$_2$Sb$_2$Te$_6$, 12 fs for Ge$_2$Sb$_2$Te$_5$ and Ge$_2$Sb$_2$Te$_6$, which were adopted for compounds with the same stoichiometry as GST compounds. Figure 7b, f shows the electrical conductivity of
Pb\textsubscript{2}Sb\textsubscript{2}S\textsubscript{5} compounds. Normally, a system with larger $S$ would give lower $\sigma^{49}$. Indeed, this happens for all three compounds under p-type doping. Specifically, at the same hole carrier density, the $\sigma$ of Pb\textsubscript{2}Sb\textsubscript{2}Te\textsubscript{5} is slightly larger than that of Pb\textsubscript{2}Sb\textsubscript{2}Se\textsubscript{5}, while the Pb\textsubscript{2}Sb\textsubscript{2}S\textsubscript{5} has the smallest value; the Seebeck coefficient decrease in the order of Pb\textsubscript{2}Sb\textsubscript{2}S\textsubscript{5}, Pb\textsubscript{2}Sb\textsubscript{2}Se\textsubscript{5}, and Pb\textsubscript{2}Sb\textsubscript{2}Te\textsubscript{5}. For n-type doping, however, the fact is that Pb\textsubscript{2}Sb\textsubscript{2}Te\textsubscript{5}, which has the largest $S$, also possesses the highest $\sigma$ among Pb\textsubscript{2}Sb\textsubscript{2}VI\textsubscript{5} compounds. Such a behavior can be attributed to the existence of multiple conduction valleys in Pb\textsubscript{2}Sb\textsubscript{2}Te\textsubscript{5}, which contribute significantly to electrical conduction when the Fermi level across the conduction band edge (Supplementary Fig. 11). By substituting the constant relaxation time into the ratio of $\kappa / \sigma$ given by BoltzTraP, the electronic thermal conductivity as a function of carrier concentration was determined, as shown in Fig. 7c, g. Clearly, for both p-type and n-type dopings, the $\kappa$ of Pb\textsubscript{2}Sb\textsubscript{2}S\textsubscript{5}, Pb\textsubscript{2}Sb\textsubscript{2}Se\textsubscript{5}, and Pb\textsubscript{2}Sb\textsubscript{2}Te\textsubscript{5} all increases with increasing the carrier density, exhibiting similar trends to the $\sigma$–$n$ relation (Fig. 7b, f). It is important to note that Pb\textsubscript{2}Sb\textsubscript{2}Se\textsubscript{5} has very small n-type $\kappa$, even at a relatively high carrier concentration of $1 \times 10^{20}$ cm$^{-3}$, which plays a significant role in achieving the high n-type TE performance. The power factor (PF) was further extracted from the Seebeck coefficient and electrical conductivity (Fig. 7d, h). The results demonstrate that Pb\textsubscript{2}Sb\textsubscript{2}S\textsubscript{5} is moderately larger than that of Pb\textsubscript{2}Sb\textsubscript{2}Se\textsubscript{5} and Pb\textsubscript{2}Sb\textsubscript{2}Te\textsubscript{5} in the whole studied concentration range, owing to the higher $S$ of the former. For n-type doping, the combination of high $S$ and $\sigma$ results in the much larger PF of Pb\textsubscript{2}Sb\textsubscript{2}Te\textsubscript{5} compared with Pb\textsubscript{2}Sb\textsubscript{2}S\textsubscript{5} and Pb\textsubscript{2}Sb\textsubscript{2}Se\textsubscript{5}. In addition, the p-type PF of Pb\textsubscript{2}Sb\textsubscript{2}Se\textsubscript{5} is larger than that of n-type, which is opposite in Pb\textsubscript{2}Sb\textsubscript{2}S\textsubscript{5}. verifying the fact that the PE performance of Pb\textsubscript{2}Sb\textsubscript{2}Te\textsubscript{5} in n-type doping is superior to p-type doping, while Pb\textsubscript{2}Sb\textsubscript{2}Se\textsubscript{5} is better when p-type doped. Supplementary Figs. 12–15 illustrates the electrical transport properties of some Se-based and Te-based compounds, which exhibit similar electrical transport behaviors to Pb\textsubscript{2}Sb\textsubscript{2}Se\textsubscript{5} and Pb\textsubscript{2}Sb\textsubscript{2}Te\textsubscript{5}, respectively.

**Phonon thermal transport properties**

The lattice thermal conductivity is one of the key quantities that determine the TE performance of materials. Based on the linearized phonon Boltzmann transport equation (BTE), the temperature dependence of $\kappa_l$ for the Pb\textsubscript{2}Sb\textsubscript{2}VI\textsubscript{5} system was estimated, as shown in Fig. 8a. Evidently, the $\kappa_l$ values of all compounds gradually decrease with the increase of temperature and exhibit a typical $T^{-1}$ behavior in the temperature range of 100–650 K, manifesting that the phonon transport is dominated by the phonon–phonon Umklapp scattering. Among them, for a given temperature, Pb\textsubscript{2}Sb\textsubscript{2}S\textsubscript{5} possesses the lowest lattice thermal conductivity while Pb\textsubscript{2}Sb\textsubscript{2}Se\textsubscript{5} gives the highest value. At 300 K, the $\kappa_l$ value is 0.55 W m$^{-1}$K$^{-1}$ for Pb\textsubscript{2}Sb\textsubscript{2}S\textsubscript{5}, 0.71 W m$^{-1}$K$^{-1}$ for Pb\textsubscript{2}Sb\textsubscript{2}Se\textsubscript{5}, and 1.83 W m$^{-1}$K$^{-1}$ for Pb\textsubscript{2}Sb\textsubscript{2}Te\textsubscript{5}, which decreases to 0.25, 0.33, and 0.85 W m$^{-1}$K$^{-1}$ at 650 K, respectively. To unravel the low-$\kappa_l$ mechanism in Pb\textsubscript{2}Sb\textsubscript{2}S\textsubscript{5}, we calculated the phonon group velocity of each mode derived from the phonon dispersion. In general, the slow phonon transport in the crystal lattice suppresses the phonon thermal conduction, thereby leading to small lattice thermal conductivity$^{36,51}$. As seen from Fig. 8b, the group velocity basically decreases in the order of Pb\textsubscript{2}Sb\textsubscript{2}S\textsubscript{5}, Pb\textsubscript{2}Sb\textsubscript{2}Se\textsubscript{5}, and Pb\textsubscript{2}Sb\textsubscript{2}Te\textsubscript{5}, which is good agreement with the trend of mean sound speed estimated from the elastic moduli (Pb\textsubscript{2}Sb\textsubscript{2}S\textsubscript{5} (2392 m s$^{-1}$) > Pb\textsubscript{2}Sb\textsubscript{2}Se\textsubscript{5} (2195 m s$^{-1}$) > Pb\textsubscript{2}Sb\textsubscript{2}Te\textsubscript{5} (1919 m s$^{-1}$)). However, the fact is that the lattice thermal conductivity of Pb\textsubscript{2}Sb\textsubscript{2}Se\textsubscript{5} is smaller than that of Pb\textsubscript{2}Sb\textsubscript{2}Se\textsubscript{5} and Pb\textsubscript{2}Sb\textsubscript{2}Te\textsubscript{5}, indicating that the factor of sound speed fails to explain the lower $\kappa_l$ of Pb\textsubscript{2}Sb\textsubscript{2}S\textsubscript{5}.

Further, we estimated the Grüneisen parameter ($\gamma$) to measure the strength of phonon anharmonicity. A larger $|\gamma|$ usually indicates the higher anharmonicity of lattice vibrations and stronger anharmonic phonon–phonon scattering, which would result in a shorter phonon lifetime and the lower thermal conductivity. Figure 8c indicates that Pb\textsubscript{2}Sb\textsubscript{2}S\textsubscript{5} has relatively larger $|\gamma|$ values compared to those of Pb\textsubscript{2}Sb\textsubscript{2}Se\textsubscript{5} and Pb\textsubscript{2}Sb\textsubscript{2}Te\textsubscript{5}, especially in the low-frequency range (<75 cm$^{-1}$), indicating the stronger anharmonic phonon interactions in Pb\textsubscript{2}Sb\textsubscript{2}S\textsubscript{5}. For better
the concentration of $1.68 \times 10^{20} \text{ cm}^{-3}$ given in Fig. 9. Their which can be attributed to the larger p-type PF and lower $10^{19} \text{ cm}^{-3}$ suggesting its great promise of being a good mid-temperature TE former. Importantly, the Seebeck coef

type at different temperatures for the IV-V-VI family of layered compounds. As predicted, the maximum $ZT$ value of Pb$_2$Sb$_2$S$_5$ at their respective optimal carrier doping levels $(1.7 \times 10^{20} \text{ cm}^{-3}$ for Pb$_2$Sb$_2$S$_5$ and $4.3 \times 10^{19} \text{ cm}^{-3}$ for Pb$_2$Sb$_2$Te$_5$), which can be attributed to the larger p-type PF and lower $\kappa_l$ of the former. Importantly, the $ZT_{\text{max}}$ of n-type Pb$_2$Sb$_2$S$_5$ increases to 1.21 ($n = 4.9 \times 10^{19} \text{ cm}^{-3}$), comparable to and even slightly larger than that of the n-type Bismuth–Telluride-based system (1.1–1.2)$^{20,32}$, suggesting its great promise of being a good mid-temperature TE material. The optimal $ZT$ value of Pb$_2$Sb$_2$Te$_5$ in n-type doping ($ZT_{\text{max}} = 0.73$ at the concentration of $3.6 \times 10^{19} \text{ cm}^{-3}$) is relatively larger than that of p-type due to the combination of high n-type Seebeck coefficient and electrical conductivity; whereas the n-type TE performance of Pb$_2$Sb$_2$Se$_5$ is inferior to p-type ($ZT_{\text{max}} = 0.38$ at the concentration of $1.68 \times 10^{20} \text{ cm}^{-3}$) arising from the lower n-type S. This substantiates the fact that Pb$_2$Sb$_2$Te$_5$ is a better n-type TE material while Pb$_2$Sb$_2$Se$_5$ does better under p-type doping, which can also be applied to other Se/Te-based compounds.

**DISCUSSION**

In summary, by combing high-throughput DFT calculations and deep neural networks, we have trained two different ML models to predict the maximum $ZT$ value and corresponding optimal doping type at different temperatures for the IV-VI family of layered semiconductors. These two models exhibit outstanding global fitting performance and have no overfitting and underfitting problems. Moreover, using the validation dataset, the comparison between DFT-calculated and ML-predicted TE properties further substantiates their robust reliability due to the exceptionally low MSE ($\approx 0.008$), large $R^2$ ($\approx 0.95$) and high prediction accuracy ($\approx 90\%$). Among this IV-VI crystal system, several compounds with high $ZT_{\text{max}}$ ($\approx 0.8$ at 650 K), in particular Pb$_2$Sb$_2$S$_5$ under n-type doping ($ZT_{\text{max}} = 1.21$), were successfully identified. It is demonstrated that the ultralow thermal conductivity is mainly responsible for the excellent TE performance of n-type Pb$_2$Sb$_2$S$_5$. The significant lattice anharmonicity and short phonon lifetime suppress the phonon thermal conduction in Pb$_2$Sb$_2$S$_5$, thereby leading to the exceptionally low-$\kappa_l$ (0.55 W m$^{-1}$ K$^{-1}$ at 300 K, and 0.25 W m$^{-1}$ K$^{-1}$ at 650 K).

In addition, we also found that the Te-based semiconducting systems exhibit better n-type TE performance than that of p-type, whereas the Se-based compounds show a completely opposite situation. This is due to that n-type Te-based compounds possess a combination of high S and $\sigma$, while Se-based semiconductors have large Seebeck coefficients for p-type doping arising from the steep DOS near the VBM. This work not only provides several promising TE semiconductors for further studies and applications but also highlights the method to overcome the obstacles in the traditional trial-and-error ways, which is applicable to the discovery of various types of functional materials.

### METHODS

#### Density functional theory

Within the framework of the DFT, all calculations were carried out in the Vienna ab initio simulation package (VASP)$^{19}$. The GGA-PBE functional was employed and the valence electronic configurations are $s^p$, $s^p^2$, and $s^p^3$ for group IV, V, and VI elements, respectively. We used the semi-empirical DFT-D2 method$^{19}$ to better describe the van der Waals forces between IV/VI atomic layers. The force and energy convergence criteria were set to $1 \times 10^{-2}$ eV Å$^{-1}$ and $1 \times 10^{-6}$ eV, respectively, and a kinetic energy cutoff of 400 eV was adopted for the plane-wave basis set. We employed the LOBSTER code to calculate pCOHP.

#### Machine learning

Among all the 70 compounds, the number of A$_1$B$_2$C$_4$, A$_1$B$_4$C$_7$, A$_2$B$_2$C$_5$ A$_3$B$_2$C$_6$ (A = Si, Ge, Sn, Pb; B = As, Sb, Bi; C = S, Se, Te) are 19, 18, 20, 13, respectively. By using the random module in the Numpy package of Python, we randomly selected 40 compounds as the input dataset for machine learning. Meanwhile, a fixed random factor of 31 was set to ensure the repeatability of the random process. Finally, we obtained 13 A$_1$B$_2$C$_4$, 10 A$_1$B$_4$C$_7$, 9 A$_2$B$_2$C$_5$, and 8 A$_3$B$_2$C$_6$ input compounds. The code used to randomly generate the 40 input compounds is available online at https://github.com/AlienMarkWong/ztml/blob/main/ztml/tools.py.py. In this work, the machine learning was carried out using the package of ztml, which was developed by Python program language and PyTorch machine learning framework. All dataset and origin codes are available online at https://github.com/AlienMarkWong/ztml/.

#### Transport properties

On the basis of the constant relaxation time and approximation and rigid band approach, the electrical transport properties were calculated by solving the semi-classic BTE as implemented in the BoltzTraP software$^{35}$. The HSE06 electronic bandgaps obtained from our previous work$^{19}$ were used to obtain the more realistic TE properties. The temperature-
dependent lattice thermal conductivity was determined from the single-mode relaxation time approximation using the ShengBTE package, and the calculation details can be found in our recent research. The detailed setting parameters in this study have been summarized in Supplementary Table 2. Note that the hexagonal conventional unit cell was employed to calculate the electrical and phonon transport properties for all compounds. For better reflecting the electronic structures, the primitive rhombohedral cell was adopted to calculate the density of states and band structures for $A\text{I}_2\text{B}_2\text{C}_6$ and $A\text{II}_2\text{B}_2\text{C}_6$ compounds.

**DATA AVAILABILITY**

The data that support the findings of this study are available from the corresponding author, and the dataset used to generate the ML models is available online at https://github.com/AlienMarkWong/ztml.

**CODE AVAILABILITY**

The codes employed to develop the ML models are available online at https://github.com/AlienMarkWong/ztml.

Received: 12 July 2021; Accepted: 1 October 2021; Published online: 29 October 2021

**REFERENCES**

1. Chu, S. & Majumdar, A. Opportunities and challenges for a sustainable energy future. *Nature* **488**, 294–303 (2012).
2. DiSalvo, F. J. Thermoelectric cooling and power generation. *Science* **285**, 703–706 (1999).
3. Bell, L. E. Cooling, heating, generating power, and recovering waste heat with thermoelectric systems. *Science* **321**, 1457–1461 (2008).
4. Pei, Y. et al. Convergence of electronic bands for high performance bulk thermoelectrics. *Nature* **473**, 66–69 (2011).
5. Pei, Y., Wang, H. & Snyder, G. J. Band engineering of thermoelectric materials. *Adv. Mater.* **24**, 6125–6135 (2012).
6. Heremans, J. P. et al. Enhancement of thermoelectric efficiency in PbTe by distortion of the electronic density of states. *Science* **321**, 554–557 (2008).
7. Yang, Y. et al. Pyroelectric nanogenerators for harvesting thermoelectric energy. *Nano Lett.* **12**, 2833–2838 (2012).
8. Li, Z. et al. High thermoelectric performance of few-quintuple Sb$_2$Te$_3$ nanofilms. *Nano Energy* **43**, 285–290 (2018).
9. Poudel, B. et al. High-temperature performance of nanostructured bismuth antimony telluride bulk alloys. *Science* **320**, 634–638 (2008).
10. Zebbarjadi, M. et al. Power factor enhancement by modulation doping in bulk nanocomposites. *Nano Lett.* **11**, 2225–2230 (2011).
11. Yu, B. et al. Enhancement of thermoelectric properties by modulation-doping in silicon germanium alloy nanocomposites. *Nano Lett.* **12**, 2077–2082 (2012).
12. Berry, T., Fu, C., Auffermann, G., Fecher, G. H. & Felser, C. Enhancing thermoelectric performance of Ti$_{11}$Sn$_5$ half-Heusler compound via modulation doping. *Chem. Mater.* **29**, 7042–7048 (2017).
13. Biswas, K. et al. High-performance bulk thermoelectrics with all-scale hierarchical architectures. *Nature* **489**, 414–418 (2012).
14. Zhang, Q. et al. Improved thermoelectric performance of silver nanoparticles-dispersed Bi$_2$Te$_3$ composites deriving from hierarchical two-phased heterostructure. *Adv. Funct. Mater.* **25**, 966–976 (2015).
15. Qiu, P., Zhang, T., Qiu, Y., Shi, X. & Chen, L. Sulfide bismuth telluride material: a natural mineral with ultralow thermal conductivity. *Energy Environ. Sci.* **7**, 4000–4006 (2014).
16. Raychowdhury, S. et al. Soft phonon modes leading to ultralow thermal conductivity and high thermoelectric performance in AgCuTe. *Angew. Chem. Int. Ed.* **130**, 4107–4111 (2018).
17. Sootsman, J., Chung, D. & Kanatzidis, M. New and old concepts in thermoelectric. *Mater. Angew. Chem. Int. Ed.* **48**, 8616–8639 (2010).
18. Jana, M. K. & Biswas, K. Crystalline solids with intrinsically low lattice thermal conductivity for thermoelectric energy conversion. *ACS Energy Lett.* **3**, 1315–1324 (2018).
19. Gan, Y., Huang, Y. D., Miao, N. H., Zhou, J. & Sun, Z. M. Novel IV-VI semi-conductors with ultralow lattice thermal conductivity. *J. Mater. Chem. C* **9**, 4189–4199 (2021).
20. Hu, L., Zhu, T., Liu, X. & Zhao, X. Point defect engineering of high-performance bismuth-telluride-based thermoelectric materials. *Adv. Funct. Mater.* **24**, 5211–5218 (2014).

**CODE AVAILABILITY**

The codes employed to develop the ML models are available online at https://github.com/AlienMarkWong/ztml.
ACKNOWLEDGEMENTS
This work is supported by the National Natural Science Foundation of China (51872017) and the high-performance computing (HPC) resources at Beihang University.

AUTHOR CONTRIBUTIONS
Y.G. performed the DFT calculations and prepared training data from simulations. G.W. developed the ML models. Y.G. and G.W. contributed to discussing the results. J.Z. and Z.S. initialized and led the project, and provided professional guidance to this research. All authors contributed to writing the manuscript.

COMPETING INTERESTS
The authors declare no competing interests.

ADDITIONAL INFORMATION
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-021-00645-y.

Correspondence and requests for materials should be addressed to Zhimei Sun.

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.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021