SPECIAL TOPIC

Prediction of lattice thermal conductivity with two-stage interpretable machine learning

SPECIAL TOPIC — Smart design of materials and design of smart materials

Prediction of lattice thermal conductivity with two-stage interpretable machine learning

Jinlong Hu(胡锦龙)[1,†], Yuting Zuo(左钰婷)[1,†], Yuzhou Hao(郝昱州)[1,†], Guoyu Shu(舒国钰)[1], Yang Wang(王洋)[1], Minxuan Feng(冯敏轩)[1], Xuejie Li(李雪洁)[1], Xiaoying Wang(王晓莹)[1], Jun Sun(孙军)[1], Xiangdong Ding(丁向东)[1], Zhibin Gao(高志斌)[1,‡], Guimei Zhu(朱桂妹)[2,§], and Baowen Li(李保文)[3,4,5]

[1] *State Key Laboratory for Mechanical Behavior of Materials, Xi'an Jiaotong University, Xi'an* 710049, *China*
[2] *School of Microelectronics, Southern University of Science and Technology, Shenzhen* 518055, *China*
[3] *Department of Materials Science and Engineering, Southern University of Science and Technology, Shenzhen* 518055, *China*
[4] *Department of Physics, Southern University of Science and Technology, Shenzhen* 518055, *China*
[5] *Paul M. Rady Department of Mechanical Engineering and Department of Physics, University of Colorado, Boulder, Colorado* 80305-0427, *USA*

(Received 31 December 2022; revised manuscript received 6 February 2023; accepted manuscript online 10 February 2023)

Thermoelectric and thermal materials are essential in achieving carbon neutrality. However, the high cost of lattice thermal conductivity calculations and the limited applicability of classical physical models have led to the inefficient development of thermoelectric materials. In this study, we proposed a two-stage machine learning framework with physical interpretability incorporating domain knowledge to calculate high/low thermal conductivity rapidly. Specifically, crystal graph convolutional neural network (CGCNN) is constructed to predict the fundamental physical parameters related to lattice thermal conductivity. Based on the above physical parameters, an interpretable machine learning model–sure independence screening and sparsifying operator (SISSO), is trained to predict the lattice thermal conductivity. We have predicted the lattice thermal conductivity of all available materials in the open quantum materials database (OQMD) (https://www.oqmd.org/). The proposed approach guides the next step of searching for materials with ultra-high or ultra-low lattice thermal conductivity and promotes the development of new thermal insulation materials and thermoelectric materials.

Keywords: low lattice thermal conductivity, interpretable machine learning, thermoelectric materials, physical domain knowledge

PACS: 63.20.–e, 63.20.Ry, 74.25.fg, 72.20.–i **DOI:** 10.1088/1674-1056/acbaf4

1. Introduction

Thermoelectric materials are a class of functional materials that directly convert thermal and electrical energy into each other through carrier (electron or hole) motion inside the material. They have great application prospects in waste heat conversion, and thermoelectric cooling, etc., both are very essential in our approach to realizing carbon neutrality. The relevant devices are made of simple structure, no noise, no waste emission, and are environmental friendly.[1] The dimensionless thermoelectric optimum ZT determines the conversion efficiency of a thermoelectric material: $ZT = S^2\sigma T/(\kappa_e + \kappa_L)$, where S, σ, T, κ_e, and κ_L are the Seebeck coefficient, electrical conductivity, absolute temperature, electronic thermal conductivity, and lattice thermal conductivity, respectively. The larger the ZT, the more efficient the conversion of thermal energy into electrical energy.[2]

The lattice thermal conductivity is a relatively independent term in the thermoelectric figure of merit. Thus, finding materials with lower lattice thermal conductivity is a critical way to improve the conversion efficiency of thermoelectric materials. Zhao *et al.*[3,4] have found SnSe and SnS materials with high thermoelectric conversion and ultralow thermal conductivity. However, the energy conversion efficiency of high-performance thermoelectric materials is still lower than that of conventional power generation and cooling technologies. Most of them are compounds of metallic lead elements with high preparation costs. Therefore, there is an urgent need to efficiently find new heat-conducting materials which are of great significance to reducing energy consumption and environmental pollution.

With the rapid development of artificial intelligence technology and big data science, material informatics has provided a new way to accelerate the design and development of new materials.[5–15] Material informatics is based on databases obtained from simulations or experimental measurements. Machine learning algorithms are used to find functional materials with specific properties or to predict unknown properties of target materials.[16] It enables direct mapping from material structure to thermal properties using machine learning algorithms, which allows the prediction of target materials based on predefined criteria and avoids repetitive human effort. As

[†] These authors contributed equally to this work.

[‡] Corresponding author. E-mail: zhibin.gao@xjtu.edu.cn

[§] Corresponding author. E-mail: zhugm@sustech.edu.cn one of the core technologies in materials informatics, data-driven machine learning methods have shown significant advantages in studying of thermal transport properties due to their efficiency and accuracy. Machine learning methods have become popular for accelerating the design of new materials with accuracy close to that of *ab initio* calculations. At the same time, the computational speed is several orders of magnitude faster.[17–19] Traditional machine learning methods such as decision trees, random forests, gradient boosting regression trees, extreme gradient boosting, and other algorithms have been widely used for thermal conductivity prediction.[20–26]

However, the above models only consider the fundamental physical quantities related to the elemental composition and neglect the influence of the crystal structure information of the material on the properties. To address the above problem, Xie *et al.*[27] developed a crystal graph convolutional neural networks framework to directly learn material properties from the connection of atoms in the crystal, providing a universal and interpretable representation of crystalline materials. The method provided a highly accurate prediction of density functional theory that calculated eight different properties of crystals with various structure types and compositions after being trained with 104 data points. Subsequently, for a small sample learning problem, Zhu *et al.*[28] combined graph neural networks and random forest approaches to predict the thermal conductivity of all known inorganic materials in the inorganic crystal structure database. Although graph neural network-based approaches can directly establish the mapping relationship between crystal structure and lattice thermal conductivity, they are black-box models[29] and have difficulty mining the physical mechanism behind the data.

In order to identify reliable material feature descriptors and capture the underlying physical mechanisms of target properties, Ouyang *et al.*[30] proposed a systematic approach to discover material property descriptors within a compressed sensing-based dimensional framework — sure independence screening and sparsifying operator (SISSO). The approach modeled the relationship between feature descriptors and targeted property concisely. Loftis *et al.*[31] introduced symbolic regression and obtained a concise expression to predict thermal conductivity. Liu *et al.*[32] introduced SISSO to obtain an explicit expression for lattice thermal conductivity. However, this method relies entirely on domain physical knowledge in selecting basic physical parameters, and the feature selection can only be performed in limited phase space.

In this paper, taking the advantages of the above two types of methods, we develop a two-stage machine-learning framework to predict lattice thermal conductivity efficiently. By combining graph convolution neural network and SISSO approach, we can predict the lattice thermal conductivity efficiently and accurately. For the first stage, CGCNN is introduced to predict the fundamental physical parameters related to lattice thermal conductivity. For the second stage, based on the above physical parameters, an interpretable machine learning model SISSO, is trained to predict the lattice thermal conductivity. We have predicted the lattice thermal conductivity of all materials in the open quantum materials database (OQMD). This work guides the next step of searching for materials with ultra-high or ultra-low lattice thermal conductivity and promotes the development of new functional thermal materials.

2. Predicting lattice thermal conductivity

Here, the overall schematic framework is shown in Fig. 1.

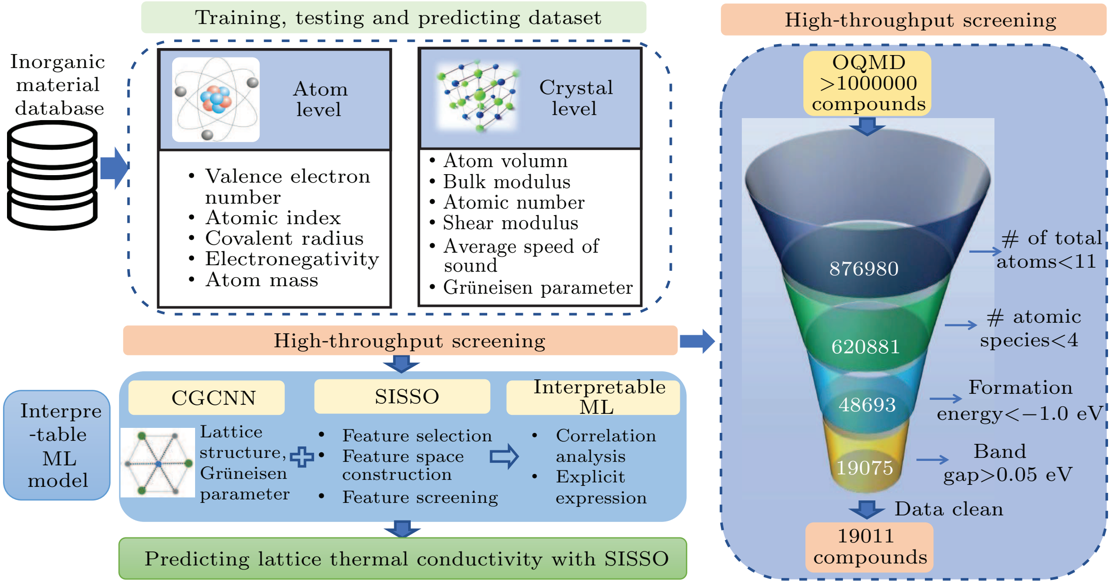

Fig. 1. The schematic framework of the proposed approach. First, the basic physical parameters of atom level and crystal level are selected from AFLOW database. After high-throughput screening, CGCNN is utilized to model the relationship between crystal structure information and Grüneisen parameter, shear modulus, bulk modulus and the average speed of sound which are related strongly to the lattice thermal conductivity of compound. Then, based on the above predicted parameters and primitive atomic parameters, SISSO is leveraged to construct an explicit expression for lattice thermal conductivity. Before model prediction for OQMD database, high-throughput screening is used to select non-metal compounds with four screening conditions.

We start model training and testing by learning from the AFLOW database, which is a globally available database of 3528653 material compounds with over 733959824 calculated properties and provides information on crystals' energy bands and thermal and mechanical properties. The basic physical parameters of atom and crystal levels are selected from this database.

First, CGCNN is trained to model the relationship between crystal structure information and the Grüneisen parameter, shear modulus, bulk modulus and the average sound speed, which are related strongly to the lattice thermal conductivity of the compound. Then, based on the above-predicted parameters and primitive atomic parameters, SISSO is leveraged to construct the explicit expression for lattice thermal conductivity. After the above two stages, we collected the dataset from the OQMD database for predicting lattice thermal conductivity. Specifically, high-throughput screening is used to select non-metal compounds with four screening conditions. After that, the lattice thermal conductivity of 19011 compounds is predicted efficiently with trained interpretable model.

2.1. Correlation analysis of physical parameters

Due to the large number of parameters affecting the lattice thermal conductivity, feature correlation analysis is used to select the critical parameters related to the lattice thermal conductivity. In this work, Pearson correlation analysis method is used to compute the correlation between each feature parameter and lattice thermal conductivity. Correlogram between the lattice thermal conductivity κ and basic physical parameters is shown in Fig. 2. The basic physical parameters include energy band gap, speed of sound v_{p}, Poisson's ratio P, bulk modulus B, shear modulus G, Young's modulus E, Debye temperature θ_{Da}, Grüneisen parameter γ, heat capacity C_{v} and lattice constant a. For binary compounds, basic physical parameters include atomic radius r_{a} and r_{b} , atomic mass m_{a} and m_{b}, electron affinity energy E_{ea} and E_{eb}, ground state energy of atoms E_{a0} and E_{b0}, valence electron number of atoms VE_{a} and VE_{b}, the electronegativity of atoms χ_{a} and χ_{b}, atomic number of atoms Z_{a} and Z_{b}. It is noted from the figure that speed of sound v_{p}, bulk modulus B, shear modulus G, Young's modulus E, and Debye temperature θ_{Da} have the strongest correlations with the lattice thermal conductivity, which has the most significant influence on the lattice thermal conductivity. As can be seen from the figure, Young's modulus is strongly correlated with bulk modulus B and shear modulus G, which is in agreement with physical domain knowledge. In this way, bulk modulus B and shear modulus G are selected as critical primary feature parameters with lattice thermal conductivity. Furthermore, according to the empirical Slack model,[33] which is described as follows:

$$\kappa_{\mathrm{L}} = \frac{2.43\times10^{-8}}{1-\frac{0.514}{\gamma}+\frac{0.228}{\gamma^{2}}}\cdot\frac{\bar{M}\theta_{\mathrm{Da}}^{3}V^{\frac{1}{3}}}{T\gamma^{2}}, \tag{1}$$

where γ denotes the Grüneisen parameter, θ_{Da}^{3} denotes Debye temperature, V denotes atomic volume, $\bar{M}$ denotes the average atomic mass, and T is the absolute temperature. Grüneisen parameter γ is an essential parameter with lattice thermal conductivity, which is not reflected in Fig. 2. In summary, based on the combination of data analysis and physical domain knowledge, Grüneisen parameter γ, bulk modulus B, shear modulus G, and average phonon velocity v_{p} are selected as the most critical primary feature descriptors for lattice thermal conductivity.

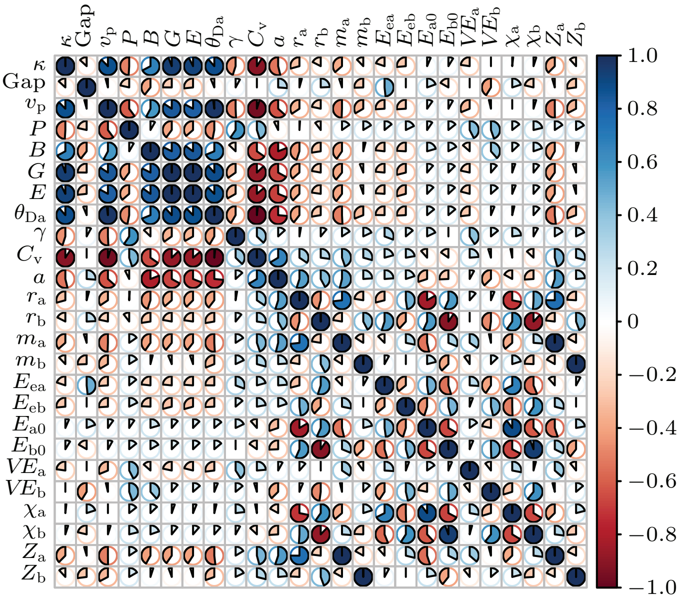

Fig. 2. Pearson correlation analysis diagram showing the correlation between the basic physical parameters and the lattice thermal conductivity κ. The larger the area of the pie chart and the closer the color to blue, the greater the correlation between the two parameters. The basic physical parameters include energy band gap, speed of sound v_{p}, Poisson's ratio P, bulk modulus B, shear modulus G, Young's modulus E, Debye temperature θ_{Da}, Grüneisen parameter γ, heat capacity C_{v}, lattice constant a. For binary compounds, basic physical parameters include atomic radius r_{a} and r_{b}, atomic mass m_{a} and m_{b}, electron affinity energy E_{ea} and E_{eb}, ground state energy of atoms E_{a0} and E_{b0}, valence electron number of atoms VE_{a} and VE_{b}, the electronegativity of atoms χ_{a} and χ_{b}, atomic number of atoms Z_{a} and Z_{b}. It is noted from the figure that speed of sound v_{p}, bulk modulus B, shear modulus G, Young's modulus E, and Debye temperature θ_{Da} have the strongest correlations with the lattice thermal conductivity, which has the most significant influence on the lattice thermal conductivity.

2.2. Predicting physical parameters with CGCNN

The Grüneisen parameter is directly related to the lattice thermal conductivity of the material and is an important parameter affecting the anharmonicity such as thermal expansion. However, due to its high computational complexity and unclear constitutive relationships, it is difficult to manually extract the fundamental physical parameters to describe the Grüneisen parameter. Furthermore, according to the Pearson correlation analysis in the above section, shear modulus, bulk modulus, and average sound speed are also strongly related to the lattice thermal conductivity.

In order to characterize the crystal structure information, it is necessary to transform the crystal structure into a specific feature vector. CGCNN[27] first transforms the crystal structure into a graph structure and then uses the graph convolutional neural network to extract and characterize the features of the graph structure to generate feature descriptors describing the crystal structure information. First, the crystal is formed into a graph structure by linking individual atoms. Atoms of nodes are convolved to update the information of each atom and the atomic environment and finally put together to form a feature representation of the entire crystal, which can be used to train a model for classification or regression. In this work, CGCNN is utilized to construct the relationship between crystal structure and critical parameters related to the lattice thermal conductivity, such as Grüneisen parameter, shear modulus, bulk modulus, and the average sound speed.

2.3. Predicting lattice thermal conductivity with SISSO

In order to predict lattice thermal conductivity efficiently, based on the predicted parameters in the last section, SISSO is utilized to construct the direct mapping relationship between primary physical parameters and lattice thermal conductivity.

2.3.1. Selection of fundamental physical parameters

In this work, binary functional materials are taken into consideration. Based on physical knowledge, primary feature descriptors consist of two categories: comprehensive physical parameters (Grüneisen parameter γ, bulk modulus B, average phonon velocity v_{p}, shear modulus G) and basic atomic information (valence electrons number VE, atomic number Z, covalent radius r, electronegativity χ, ground state energy E_0, atomic mass m, lattice constant a, density ρ). Due to the complexity of the thermal conductivity mechanism, the above two types of descriptors complement with each other to provide a rather comprehensive description of the factors affecting thermal conductivity.

2.3.2. Construction and screening of feature descriptor

For the SISSO approach, we assumed that the material properties can be expanded into the set of orthogonal perfect function spaces $\varPhi = \phi_1, \phi_2, \phi_3, \ldots$, and then the material properties $P(M)$ can be linearly represented by the set of perfect function spaces, i.e., $P(M) = \sum_{i=1}^{n} \beta_i \phi_i$, where β_i represents the coefficient of i-th function space, and ϕ_i denotes i-th function space. However, the orthogonal perfect function spaces are not easy to obtain. The high-dimensional feature space represents the material properties $P(M)$ instead of the orthogonal perfect space. Usually, the functions in the high-dimensional feature space ϕ constructed from the initial features have $n \sim 10^{10}$. Based on the theory of compressed perception, only the number of samples $m \sim \log(n)$ is required if the solution is sparse. Thus, the demand for samples by SISSO is not excessive. However, the feature space of $n \sim 10^{10}$ is still a big challenge for the algorithm and computational effort. To solve this problem, sure independence screening (SIS)[27] is used to reduce the feature space to a reasonable size.

The feature screening and descriptor construction process using SISSO is briefly described as follows. First, primary feature parameters correlated with the lattice thermal conductivity are selected as the input of SISSO based on physical knowledge. Then, a large feature space is constructed with algebraic operators $\hat{H}(m) \equiv\equiv I, +, -, \times, \div, /, \exp, \log, |-|, \surd, {}^{-1}, {}^{2}, \varphi$. Subsequently, the targeted attribute $P(M) = \sum_{i=1}^{n} \beta_i \phi_i = \varPhi \beta$ is expanded if β is sparse, where $\varPhi$ and β denote the function spaces set and the coefficients set respectively. Thus, based on the sparsity theory, SIS is utilized to screen the critical features from an ample feature space. Lastly, the explicit expression between feature descriptors and the lattice thermal conductivity can be described as $P(M) = \sum_{i=1}^{n} \beta_i \phi_i$.

3. Numerical experiments and result

The last section introduces a two-stage machine-learning framework to predict lattice thermal conductivity. It aims to uncover the hidden relationship between physical model parameters and the lattice thermal conductivity. Next, we will verify the feasibility and efficacy of the proposed framework. Firstly, we collect datasets from the public database AFLOW[29] and OQMD.[30] Taking crystal structure information as the input and primary physical parameters as output of the CGCNN, a graph convolutional neural network is utilized with high consistency between prediction and the primary physical parameters in AFLOW. For an interpretable machine learning model, the primary physical parameters are fed to SISSO, and the lattice thermal conductivity is taken as the output. Then, an interpretable model is achieved by SISSO. By validating the AFLOW database, we show that the proposed framework can predict the lattice thermal conductivity accurately and efficiently with the machine learning model.

3.1. Data preparation

In this work, the training and testing datasets are downloaded from the AFLOW database,[34] and the predicting dataset is downloaded from the OQMD database.[35] The AFLOW database provides abundant information on the crystal band gap, thermal and mechanical properties, etc. Then, the crystal structures and the corresponding thermal and mechanical properties such as Grüneisen parameter, shear modulus, bulk modulus, and the average speed of sound are collected from the AFLOW database and used as training and testing dataset to train the machine learning model and evaluate the model's performance.

After model training and testing, the predicting dataset is collected from the OQMD database, including the crystal structures of the compounds. The OQMD is a database of DFT-calculated thermodynamic and structural properties of 1022603 materials, created in Chris Wolverton's group at Northwestern University. There are two ways to download data. One is the entire OQMD as dumps of MySQL database, and the other is via API. The OQMD database provides two types of representational state transfer (REST) API to realize fully open data transfer. One is the OPTiMaDe API. This specification provides a detailed set of rules about keywords for materials data, specific rules about data lookup flexibility, and the format of the returned data. The other one is OQMD API with similar usage. However, there are a few notable differences between OQMD API and OPTiMaDe API because the former follows conventional qmpy data keywords, which have been used in the OQMD database since its inception. We choose OQMD API since its keywords fit better with the OQMD database.

The corresponding structure ID can be obtained by visiting the web page corresponding to the entry id in the retrieval result. Then the corresponding POSCAR file can be downloaded. As the crystal structure files in CIF format are fed to the machine learning model, the POSCAR files are converted to structure files in cif format in batches by modifying the code provided on https://github.com/dcccc/git_python/blob/master/poscar2cif.py. In this way, training, testing, and predicting datasets are collected for model training, testing, and predicting.

3.2. High-throughput screening

To screen for suitable compounds, four filters are used to obtain the data, including the total number of atoms less than 11, atomic species less than 4, formation energy less than -1.0 eV, and band gap greater than 0.05 eV. (1) Setting the filter condition as the total number of atoms less than 11 to get 876980 results. (2) Adding the filter that the atomic species should be less than 4 enables 620881 results to be found. (3) Screening the data that formation energy is less than -1.0 eV and obtaining 48693 data samples. (4) By selecting the filter condition as band gap > 0.05 eV, 19075 results meeting our requirements can be obtained. After removing errors and duplicate data, 19011 different structures and corresponding data can finally be downloaded. Each structural data contains different fields, such as name, entry id, composition, volume, the total number of atoms, band gap, formation energy, stability, etc., except for the density. Moreover, the density data can be calculated by volume and molar mass which can be obtained from the composition of the selected unit cell.

After data collection, we use the well-trained model trained with a dataset from the AFLOW database to predict the Grüneisen parameter, bulk modulus, shear modulus, and the average speed of sound corresponding to the crystal structures of the compounds.

3.3. Evaluation metrics

To evaluate the performance of the proposed approach, we employed four standard evaluation metrics for the regression task, mean absolute error (MAE), mean absolute percentage error (MAPE), root mean square error (RMSE) and goodness of fit (R^2), which are described as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} w_i |\hat{y}_i - y_i|, \tag{2}$$

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n} w_i \left|\frac{\hat{y}_i - y_i}{y_i}\right|, \tag{3}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} w_i (\hat{y}_i - y_i)^2}, \tag{4}$$

$$R^2 = \frac{\sum_{i=1}^{n} w_i (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y}_i)^2}. \tag{5}$$

Assuming a set of samples $y = \{y_1, y_2, \ldots, y_n\}$ and the predicted dataset $\hat{y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$, where n is the number of a sample dataset.

3.4. CGCNN results

First, the Grüneisen parameter, bulk modulus, shear modulus, and the average sound speed are predicted with the CGCNN model to predict the lattice thermal conductivity of the compounds. We obtain the crystal structures of all binary compounds in the AFLOW database and their corresponding Grüneisen parameters, bulk modulus, shear modulus, and the average speed of sound. Then, a nonlinear mapping relationship is established from the crystal structures to the Grüneisen parameters, bulk modulus, shear modulus, and the average speed of sound based on CGCNN via multi-task learning.

According to the above mapping relationship, the predicted performance of the model on the testing set is shown in Fig. 3. R^2_{test} indicates the goodness-of-fit of the trained model on the testing set, which describes the degree of fit between the predicted values of the Grüneisen parameters and the calculated values in the database, the closer the predicted values are to the calculated values, the better the fit is. The smaller the error value, the closer the predicted value is to the calculated value and the better the model performance. As can be seen from Fig. 3(a), the absolute error between the predicted Grüneisen parameter and the calculated Grüneisen parameter in the database lies between ± 0.1, and the average relative error is less than 0.05. It is noted from Figs. 3(b)–3(d) that for bulk modulus, shear modulus, and the average sound speed, the multi-task CGCNN model achieves superior prediction performance with a good fit above 0.9.

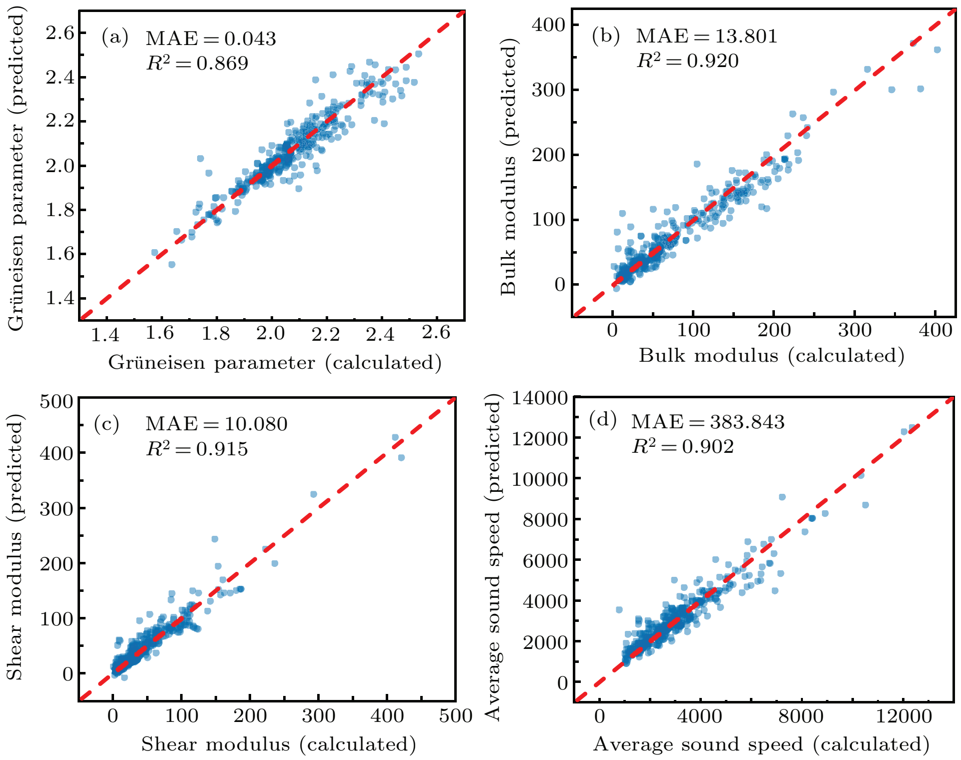

Fig. 3. Comparisons of the predicted and calculated values of Grüneisen parameter, shear modulus, bulk modulus and the average sound speed on testing dataset. (a) Comparison of the predicted and calculated values of Grüneisen parameter. (b) Comparison of the predicted and calculated values of Bulk modulus. (c) Comparison of the predicted and calculated values of shear modulus. (d) Comparison of the predicted and calculated values of the average sound speed on testing dataset.

3.5. SISSO results

Lattice thermal conductivity is a comprehensive parameter of materials with many influencing factors, and it is difficult to make a direct and exact prediction. Based on feature selection, four important primary feature parameters are selected via Pearson correlation coefficient analysis and domain knowledge in the above section. Then, SISSO is applied to construct the relationship between primary feature parameters and lattice thermal conductivity. Based on the predicted parameters with CGCNN, we obtain an explicit expression after the SISSO training. However, we find that the physical dimensions are misaligned, even though having good accuracy. The fact is that the dimension consistency of various descriptors is not considered in the feature construction process, resulting in misalignment of the dimensions for different descriptors.

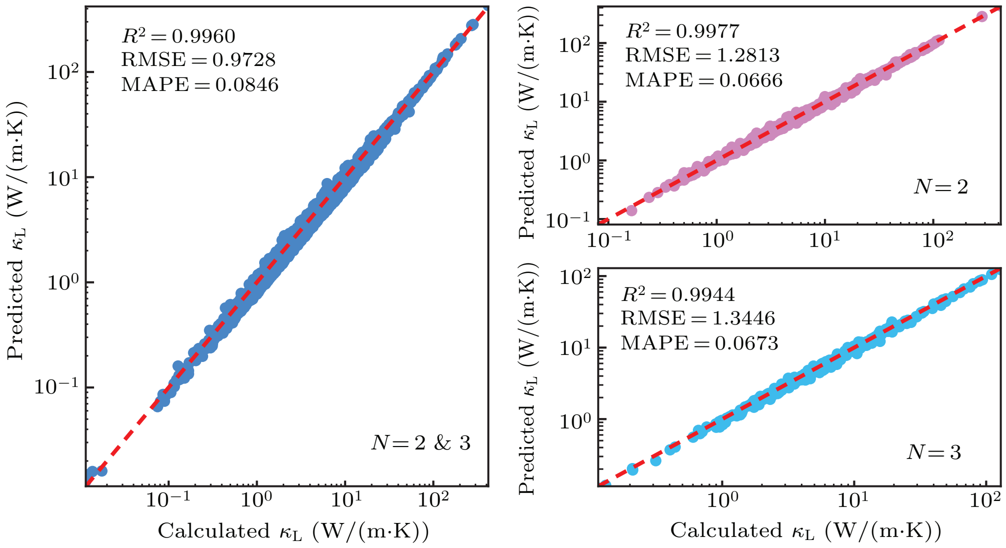

Fig. 4. The comparison of predicted lattice thermal conductivity and calculated values for binary compounds with different atomic numbers. It is noted from the figure that the model achieves superior performance for lattice thermal conductivity prediction, demonstrating the feasibility of our proposed framework.

To verify the feasibility and effectiveness of the model, the performance is validated on the testing dataset, and the results are shown in Fig. 4. The testing data in the AFLOW database has calculated the lattice thermal conductivity of binary compounds with different atomic numbers compared to the model's predicted values. The horizontal axis represents the calculated value of the lattice thermal conductivity in the database, and the vertical axis represents the predicted lattice thermal conductivity. The more concentrated the data points are on the red dotted line in the middle, the closer the predicted value is to the true value. RMSE and MAPE represent the root mean square error and the average relative error between the predicted and calculated values, respectively. The smaller the value, the better the prediction performance of the machine learning model. It is noted from the figure that the model achieves superior performance for lattice thermal conductivity prediction, demonstrating the feasibility of our proposed framework and providing a novel method for fast computation of lattice thermal conductivity.

4. Discussions

To analyze the composition distribution of thermoelectric materials,[36] we count the histogram of the elemental distribution of compounds in the OQMD database, shown in Fig. 5. The figure shows the distribution of elements in compounds with thermal conductivity less than 1. It is noted from the figure that the top three elements that appeared the most are chlorine, bromine, and cesium. In addition, F, O, Se, In, Te, and other elements with high-frequency occurrence are also common elements in low thermal conductivity materials. The electronegativity corresponding to the element is marked on the top of the Fig. 5 column. It can be seen from the comparison of electronegativity that the elements with more extreme electronegativity appear more often with stronger building ions. For example, it can be seen from the figure that the oxygen frequency is about half than that of fluorine. The strong electronegativity of fluorine can explain this phenomenon. Fluorine is usually accessible to form ionic solids with alkali metals (such as Cs, Rb, K, Na, etc.), alkaline earth metals (such as Ba, Sr, Ca, etc.), and elements from the 12th, 13th, and 14th groups (such as Tl, Sn, In, Cd, etc.).

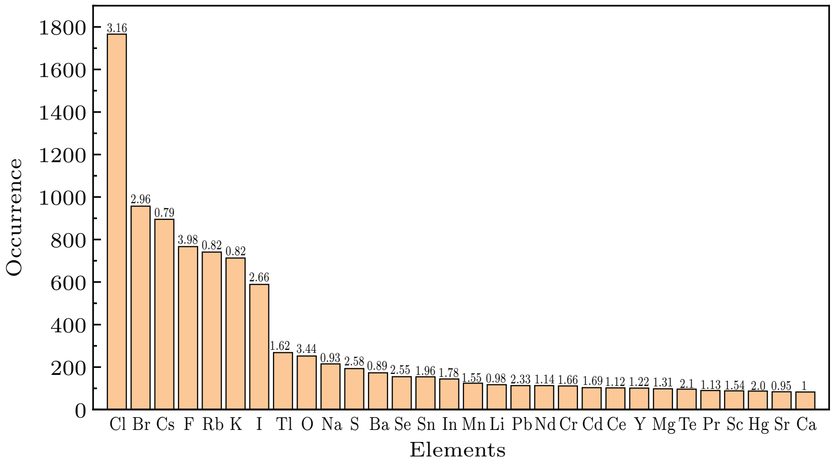

Fig. 5. The histogram of elemental distribution of compounds in OQMD database.

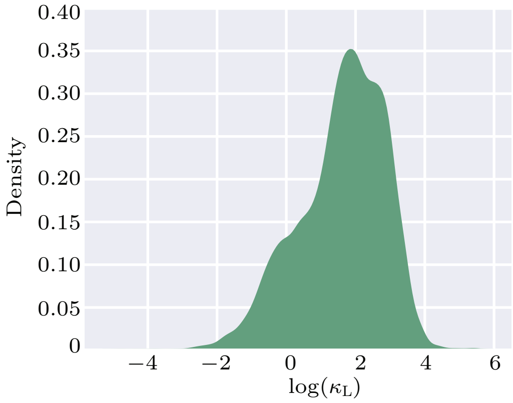

Fig. 6. The probability density function (PDF) of lattice thermal conductivity for compounds in OQMD database.

Furthermore, we also computed the PDF probability density function (PDF) of compounds in the OQMD database, which is shown in Fig. 6. As can be seen from the figure, the thermal lattice conductivity of the majority compounds in the OQMD database is mainly concentrated between $e^{1.0} = 2.72$ W/mK and $e^3 = 20.1$ W/mK, which agrees with the physical knowledge that the lattice thermal conductivity of the compounds conforms to a mixed Gaussian distribution, and can be fitted with a Gaussian mixture model. Materials with high and low thermal conductivity are only a very small percentage, less than 3%.

5. Conclusion and perspectives

We have proposed a two-stage interpretable machine learning framework to predict the lattice thermal conductivity with high accuracy and efficiency. To verify the feasibility of the proposed framework, a graph convolutional neural network CGCNN is utilized to predict complex physical parameters. An interpretable machine learning model SISSO is introduced to construct an explicit model between feature descriptors and the lattice thermal conductivity. The prediction results show that our proposed framework can accurately and efficiently predict the lattice thermal conductivity from crystal structures. It is worth stating that our approach applies only to semiconductors or insulators, where phonons contribute dominantly to the total thermal conductivity. This work provides a novel way for fast and accurate prediction of lattice thermal conductivity and guides the searching for materials with ultrahigh or ultralow lattice thermal conductivity.

Acknowledgments

We acknowledge the support of the National Natural Science Foundation of China (Grant Nos. 12104356 and 52250191), China Postdoctoral Science Foundation (Grant No. 2022M712552), the Opening Project of Shanghai Key Laboratory of Special Artificial Microstructure Materials and Technology (Grant No. Ammt2022B-1), and the Fundamental Research Funds for the Central Universities. We also acknowledge the support by HPC Platform, Xi'an Jiaotong University.

References

[1] He J and Tritt T M 2017 *Science* **357** eaak9997
[2] Bell L 2008 *Science* **321** 1457
[3] Chang C, Wu M, He D, Pei Y, Wu C F, Wu X, Hulei Y, Zhu F, Wang K, Chen Y, Huang L, Li J, He J and Zhao L 2018 *Science* **360** 778
[4] He W, Wang D, Wu H, Xiao Y, Yang Z, He D, Feng Y, Hao Y J, Dong J, Chetty R, Hao L, Chen D, Qin J, Yang Q, Li X, Song J M, Zhu Y, Xu W, Niu C and Zhao L 2019 *Science* **365** 1418
[5] Zhang Y and Chen L 2018 *NPJ Comput. Mater.* **4** 25
[6] Wen C, Zhang Y, Wang C, Xue D, Bai Y, Antonov S, Dai L, Lookman T and Su Y 2019 *Acta Mater.* **170** 109
[7] Rampi R, Rohit B, Ghanshyam P, Arun M K and Chiho K 2017 *NPJ Comput. Mater.* **3** 54
[8] Liu Y, Wu J, Wang Z, Lu X G, Avdeevd M, Shi S, Wang C and Yu T 2020 *Acta Mater.* **195** 454
[9] Mitchell J B 2014 *WIREs Comput. Mol. Sci.* **4** 468
[10] Bharat M, Anthony G, Hong D, Wei C, Kristin P, Mark A, Andrew C and Maciej H 2016 *NPJ Comput. Mater.* **2** 1
[11] Maarten J, Wei C, Randy N, Kristin P, Gerbrand C, Anubhav J, Mark A and Anthony G 2016 *Sci. Rep.* **6** 34256
[12] Paul R, Katherine E, Philip A, Casey F, Malia W, Aurelio M, Matthias Z, Sorelle F, Joshua S and Alexander N 2016 *Nature* **533** 73
[13] Xue D, Balachandran P, Hogden J, Theiler J, Xue D and Lookman T 2016 *Nat. Commun.* **7** 11241
[14] Edward K, Kevin H, Stefanie J and Elsa O 2017 *NPJ Comput. Mater.* **3** 53
[15] Edward K, Kevin H, Alex T, Sara M, Emma S, Adam S, Andrew M and Elsa O 2017 *Sci. Data* **4** 170127
[16] Wan X, Feng W, Wang Y, Wang H, Zhang X, Deng C and Yang N 2019 *Nano Lett.* **19** 3387
[17] Seko A, Togo A, Hayashi H, Tsuda K, Chaput L and Tanaka I 2015 *Phys. Rev. Lett.* **115** 205901
[18] Faber F A, Lindmaa A, von Lilienfeld O A and Armiento R 2016 *Phys. Rev. Lett.* **117** 135502
[19] Xue D, Balachandran P, Hogden J, Theiler J, Xue D and Lookman T 2016 *Nat. Commun.* **7** 11241
[20] Jaafreh R, Kang Y and Hamad K 2021 *ACS Appl. Mater. Interfaces* **13** 57204
[21] Juneja R, Yumnam G, Satsangi S and Singh A 2019 *Chem. Mater.* **31** 5145
[22] Miyazaki H, Tamura T, Mikami M, Watanabe K, Ide N, Ozkendir O M and Nishino Y 2021 *Sci. Rep.* **11** 13410
[23] Ju S, Yoshida R, Liu C, Wu S, Hongo K, Tadano T and Shiomi J 2021 *Phys. Rev. Mater.* **5** 053801
[24] Loftis C, Yuan K, Zhao Y, Hu M and Hu J 2021 *J. Phys. Chem. A* **4** 435
[25] Carrete J, Li W, Mingo N, Wang S and Curtarolo S 2014 *Phys. Rev. X* **4** 011019
[26] Roekeghem A, Carrete J, Oses C, Curtarolo S and Mingo N 2016 *Phys. Rev. X* **6** 041061
[27] Xie T and Grossman J C 2018 *Phys. Rev. Lett.* **120** 145301
[28] Zhu T, He R, Gong S, Xie T, Gorai P, Nielsch K and Grossman J 2021 *Energy Environ. Sci.* **14** 3559
[29] Rudin C 2019 *Nat. Mach. Intell.* **1** 206
[30] Ouyang R, Curtarolo S, Ahmetcik E, Scheffler M and Ghiringhelli L M 2018 *Phys. Rev. Mater.* **2** 083802
[31] Loftis C, Yuan K, Zhao Y, Hu M and Hu J 2020 *J. Phys. Chem. A* **125** 435
[32] Liu J, Han S, Cao G, Zhou Z, Sheng C and Liu H 2020 *J. Phys. D* **53** 315301
[33] Morelli D T and Slack G A 2006 *High Lattice Thermal Conductivity Solids* (New York: Springer) pp.37–68
[34] http://aflowlib.org/
[35] Kirklin S, Saal J, Meredig B, Thompson A, Doak J, Aykol M, Rhl S and Wolverton C 2015 *NPJ Comput. Mater.* **1** 15010
[36] Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, Persson K, Ceder G and Jain A 2019 *Nature* **571** 95