Journal of Materials Chemistry A

Predicting Lattice Thermal Conductivity from Fundamental Material Properties Using Machine Learning Techniques

Journal:	*Journal of Materials Chemistry A*
Manuscript ID	TA-ART-11-2022-008721.R1
Article Type:	Paper
Date Submitted by the Author:	26-Jan-2023
Complete List of Authors:	Qin, Guangzhao; Hunan University Wei, yi; Hunan University, Yu, Linfeng; Hunan University Xu, jinyuan; hu nan da xue Ojih, Joshua; University of South Carolina Rodriguez, Alejandro; University of South Carolina Wang, Huiming; Xiangtan University Qin, Zhenzhen; Zhengzhou University School of Physics and Engineering Hu, Ming; University of South Carolina, Department of Mechanical Engineering

Predicting Lattice Thermal Conductivity from Fundamental Material Properties Using Machine Learning Techniques

Guangzhao Qin[1,2,*], Yi Wei[1], Linfeng Yu[1], Jinyuan Xu[1], Joshua Ojih[2], Alejandro David Rodriguez[2], Huimin Wang[3,1,2], Zhenzhen Qin[4], and Ming Hu[2,*]

[1]*State Key Laboratory of Advanced Design and Manufacturing for Vehicle Body, College of Mechanical and Vehicle Engineering, Hunan University, Changsha 410082, P. R. China*

[2]*Department of Mechanical Engineering, University of South Carolina, Columbia, SC 29208, USA*

[3]*Hunan Key Laboratory for Micro-Nano Energy Materials & Device and School of Physics and Optoelectronics, Xiangtan University, Xiangtan 411105, Hunan, China*

[4]*International Laboratory for Quantum Functional Materials of Henan, School of Physics and Engineering, Zhengzhou University, Zhengzhou 450001, China*

Correspondence E-mail: G.Q. <gzqin@hnu.edu.cn>, M. H. <hu@sc.edu>

Abstract

High-throughput screening and material informatics have shown a great power in novel materials discovery including batteries, high entropy alloys, photocatalysts, *etc*. However, the lattice thermal conductivity (κ) oriented high-throughput screening of advanced thermal materials is still limited to the intensive use of first principle calculations, which is inapplicable to fast, robust, and large-scale material screening due to the unbearable computational cost demanding. In this study, 15 machine learning algorithms are utilized for fast and accurate κ prediction from basic physical and chemical properties of materials. The well-trained models successfully capture the inherent correlation between these fundamental materials properties and κ for different types of materials. Moreover, deep learning combined with semi-supervised technique show the capability of accurately predicting diverse κ values spanning 4 orders of magnitude, especially the power of extrapolative prediction on 3,716 new materials. The developed models provide a powerful tool for large-scale thermal materials screening with target thermal transport property.

1. Introduction

In many fields of modern science and engineering, knowledge of the thermophysical properties, in particular the lattice thermal conductivity (κ) of materials, is becoming more and more important.[1,2] The interest to precisely predict κ of crystalline materials has been triggered by the collective realization of the dramatic consequences of climate change[3]. For instance, a very typical application is the recovery of waste heat, which could be realized through a very simple and clean method, *i.e.* the thermoelectric (TE) effect. The energy conversion efficiency of a TE device is characterized by the figure of merit, *ZT*,[4] which is inversely proportional to the κ. In the past decades, searching for high-efficient TEs has been guided by the concept of 'phonon glass-electron crystal',[5] *i.e.* an ideal TE material should have high electrical carrier mobility and low κ simultaneously. Therefore, there is a strong quest for designing complex crystalline structures with unprecedentedly low κ. Besides, the κ of semiconductor materials is a key parameter for designing high performance electronic devices. Due to the significant amount of excess heat during operation, thermal management for high performance heat dissipation must be taken to prolong the durability and to increase the operating reliability. Thus, searching for materials with ultrahigh κ is extremely important for the disruptive development of micro-/nano-electronics. For non-metallic solids such as semiconductors, the heattransfer is viewed as being transferred via lattice vibrations and the quanta of such lattice vibration in a solid is called phonons[6]. Historically, the classical kinetic theory provides a rough estimation of κ based on the phonon gas model[7]. Since κ is one of the intrinsic physical properties of materials, which relates to its ability of conducting heat energy, demands of understanding and characterizing thermophysical properties of materials are ever increasing for a wide range of modern science, advanced engineering, and materials-based energy technologies.

Despite the significance of understanding and controlling thermal transport ability of materials, accurately predicting κ of a crystalline material from its atomic structure is not an easy task. Historically, κ can only be calculated theoretically by some empirical models, such as the Debye-Callaway model,[8–10] the Slack model,[11] *etc.* The empirical models could be very fast but with less accuracy because of the limitations of capturing phonon transport details.

Besides, classical equilibrium and nonequilibrium molecular dynamics (EMD/NEMD) simulations based on empirical potentials and the Newton's second law have been widely used to characterize thermal transport properties of various materials in the past decades[12]. The difficulty and limitation of classical MD simulations lie in the description of interatomic interactions by the empirical potentials[12]. Beyond that, direct numerical calculation of κ of a single crystalline material from its atomic structure by accurate first-principles coupled with phonon Boltzmann transport equation (BTE) without any other inputs has just been made available for a few years[13]. However, such computations are usually tedious and very computationally demanding even for primitive cells that are not too complicated[14–20]. Because of the huge computational loads, current density functional theory (DFT) based on first-principles method for κ calculation is out of the question for high-throughput screening thermal materials. Despite the very few successes of high-throughput computational screening low κ materials,[2,21] fast and robust κ oriented material design is still limited, not only because of the complex relationship between the intrinsic κ and the atomic structures, but also due to the unbearable computational costs. Thus, it is necessary to develop efficient and accurate κ prediction models for high-throughput screening thermal materials.

The recent success of AlphaGo in 2016-2017 fully let people appreciate the tremendous development potential of artificial intelligence (AI) technology. At present, machine learning (ML) technique has been widely used in lots of fields such as computer vision, natural language processing, data mining, robot application, *etc*[22]. With the powerful capacity, ML has been widely used by researchers to conduct research and design functional materials. In particular, ML has been increasingly used in material properties prediction and computational screening[23–31]. Most of these studies are defined as a regression problem, which is usually composed of three parts: property dataset acquisition, feature engineering, and selection of the ML algorithms. Commonly used ML algorithms include linear regression(LR), support vector regression(SVR), ridge regression(RR), neural networks, Random Forests, gradient boosting decision trees(GBDT), *etc*[24,28]. So far, only general materials properties have been used as prediction targets such as different kinds of energies, band gap, bulk modulus, shear modulus, Poisson's ratio, hardness, e*tc*,[24,26,30] while the applications of ML in predicting thermal transport properties are still limited and need to be explored.

In this paper, by using typical ML algorithms for fast and accurate prediction of κ from basic properties of materials, it is found that the well-trained ML models successfully capture the inherent correlation between basic materials properties and κ for different types of materials. Compared with the optimized Slack model, a few selected ML models show the capability of accurately predicting κ spanning 4 orders of magnitude. Moreover, the Pearson correlation coefficient map for 21 thermal-related properties of materials is generated to achieve insight into the performance of the ML models. The development of ML models for fast and accurately predicting κ provides a powerful tool for the large-scale thermal functional materials screening with targeted thermal transport property.

2. Computational methodology

ML. **—** All the ML models are built based on the ML library of TensorFlow[32]. A total of 15 different ML models have been constructed and trained for thermal transport property prediction. According to the different types of ML algorithms, these models can be classified into the following four categories[24,28]:

(1) Generalized linear regression models[33]: multiple linear regression (MLR) model, optimization of multiple linear regression models using Stochastic Gradient Descent (SGD), and ridge regression model (RR).

(2) Support vector regression (SVR) models with four different kernel functions,[34] including Linear kernel function, Gaussian kernel function, Sigmoid kernel function, and Polynomial kernel function.

(3) Tree-based models[35,36]: The classification and Regression Tree (CART), as well as some ensemble learning models[37,38] which contain Random Forests, gradient boosting decision trees (GBDT), and light gradient boosting machine (LGB).

(4) Neural network models[22,39]: artificial neural network (ANN), convolutional neural network (CNN), recurrent neural network (RNN), and long short-term memory network (LSTM).

To build and train the ML models, we firstly need to obtain the experimental data for the dataset. The large amount of κ for the training and testing procedures are collected from previously published papers and databases,[21,40–45] which consist of experimentally measured κ for 350 different materials.

Pre-process data. **—** To optimize the performance of the ML models, it is vital to select appropriate basic material properties as descriptors. The descriptors of materials are chosen based on three principles: 1) the descriptors should be basic properties of materials and should be representative; and 2) the descriptors can be easily collected from literature or calculations with limited effort. In this work, the descriptors are chosen as a combination of V, M, n, n_p, B, G, B', and G' (detailed explanation of the descriptors can be found in Table 1). Additionally, normalization processing is necessary, and the experimental data will conform to the normal distribution, which would help improve the prediction accuracy of the models. Then, the data is transformed into standard normal distribution through standardized processing. When the training process is finished, the inverse transformation is performed on the results, for the purpose of facilitating the comparison with the original data.

The training process of the these ML models are based on the well-known n-fold procedure, with the typical n = 5,[46] which means that 280 (80%) types of materials are used to train these ML models and 70 (20%) types of materials are used to test the trained model. In the process of training and testing, the material types in the training and test set of each model should be exactly the same, which is important for effectively comparing the performance of these models by controlling variables.

Model Performance Evaluation Metrics. **—** By training ML models, the goal is to select the model with optimal performance, which can be evaluated by a series of statistical indicators[47]. Some of the most important evaluation metrics as listed below have been applied to evaluate the performance of different models.

(1) root mean square error (RMSE)

$$\mathrm{RMSE} = \left[\frac{1}{\mathrm{N}}\sum_{\mathrm{i}}^{\mathrm{N}}\left(\kappa_{\mathrm{ML}}^{\mathrm{i}} - \kappa_{\mathrm{Exp.}}^{\mathrm{i}}\right)^2\right]^{\frac{1}{2}} \qquad (1)$$

(2) coefficient of determination (R^2)

$$R^2 = 1 - \frac{\kappa^i_{Exp.} - \kappa^i_{ML}}{\kappa^i_{Exp.} - \kappa^i_{avg}} \tag{2}$$

(3) mean absolute error (MAE)

$$MAE = \frac{1}{N}\left|\kappa^i_{ML} - \kappa^i_{Exp.}\right| \tag{3}$$

, where *i* specifies *i*th material sample and *N* is the total number of samples in the dataset. In addition to the above metrics, we also take into account the running time of models to measure the prediction efficiency of different models.

We also introduced K-fold cross-validation algorithm in both training and testing process, not only to better reflect the average performance of the model and obtain relatively accurate evaluation metrics, but also to serve as a comparison for incompletely supervised learning models in the subsequent testing process. More details can be found sin Note S4.

LSTM. — The LSTM is introduced to overcome the exploding/vanishing gradient problems when training very deep neural networks[48]. The principle of LSTM is shown in Fig. 1. The concept of three thresholds is introduced as input gate i_t, output gate o_t and forgetting gate f_t, which can be written as

$$i_t = \sigma(\omega_i[h_{t-1}, x_t] + b_i) \tag{4}$$

$$o_t = \sigma(\omega_o[h_{t-1}, x_t] + b_o) \tag{5}$$

$$f_t = \sigma(\omega_f[h_{t-1}, x_t] + b_f) \tag{6}$$

,where ω_i, $\omega_o$and ω_f are the parameter matrices to be trained, and b_i, b_o and b_f are offset terms. The input gate stores the information in the cellular C_t, the forgetting gate discards it according to the specified proportion, and the output gate selectively exports the information. At this moment, the long-term memory in the cellular state can be defined as

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{7}$$

,where C_{t-1} represents the long-term memory stored in the cell state at the previous time and $\tilde{C}_t$ means the candidate state memory at the current time. Therefore, the long-term memory at the current time is the sum of the long-term memory at the previous time through the forgetting gate and the candidate memory through the input gate. The model used in this work is constructed with an input layer, two LSTM layers, two dropout layers, a dense fully connected layer, and an output layer.

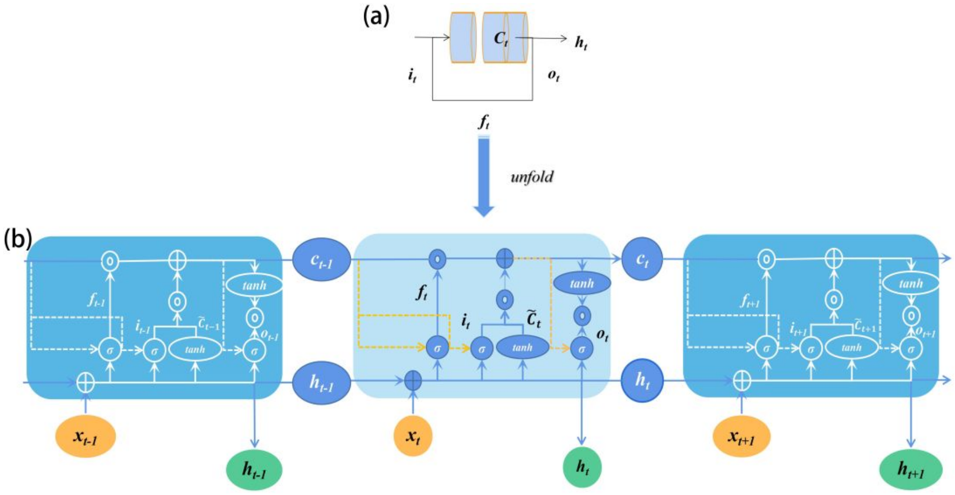


Fig. 1 The LSTM network architecture. (a) The cyclic kernel structure with a threshold. (b) The detailed structure within a single LSTM neuron and its relationship with the two neurons before and after in the forward direction.

Before feeding into the LSTM network, the data needs to be converted into a three-dimensional form. The 8 neurons in the input layer correspond to the 8 fundamental properties of materials. With the aim of updating network parameters iteratively, we use the Adam optimizer as the activation function. The first hidden layer is an LSTM layer with a total of 50 LSTM neurons (Fig. 1), which is used to extract features from the input layers. After the dropout layer (dropout rate is set as 20%), the LSTM_1 layer with 20 LSTM neurons converts the three-dimensional data into two-dimensional form and transfers the data to the dropout_1 layer (dropout rate is set as 20%). Finally, the result is passed to the fully connected layer and the predicted value of thermal conductivity can be obtained.

Incomplete supervision. — Only 350 samples are labeled in the whole dataset, while the test set contains a great number of unlabeled data samples. Generally, it is quite costly to label the samples one by one. Therefore, incomplete supervision has been developed to make full and effective use of unlabeled samples, so as to further improve the generalization ability of the model. Both active learning and semi-supervised learning belong to the incomplete supervision, which are utilized to predict the thermal conductivity of materials. Their predictive performance will also be compared in the following contents. More details about the principles of active learning and semi-supervised learning can be found in the Supplementary Note S2 and S3.

Slack model. — To compare with the ML models, an optimized Slack model[49] is further used to predict κ based on the basic properties that are used as descriptors in ML models, and the prediction results are shown in the Supplementary Note S1. The κ is expressed as[49]

$$\kappa = A\frac{M\delta n_p^{1/3}\Theta_D^3}{\gamma^2 T}, \tag{8}$$

where $\delta = V/n$ is the cubic root of the average volume per atom, T is the absolute temperature, and the explanations of other symbols are available in Table 1. The coefficient A is calculated as[49] $A = \frac{1}{1 + {}^{1}/_{\gamma} + 8.3 \times 10^{5}/_{\gamma^{2.4}}}$. All the properties in the Slack equation, such as Debye temperature and Grüneisen parameter, can be calculated from the elastic properties of bulk modulus (B) and shear modulus (G), and their derivations (B' and G') can be obtained with respect to the change of volume[50]. According to the Voigt-Reuss-Hill (VRH) theory,[51–53] the elastic properties can be evaluated from the elastic constants, which can be obtained based on accurate first-principles calculations. The above formula has been applied for the evaluation of κ for 353 materials,[49] which has been verified to have better performance than the widely used Slack model.

Table 1: The symbols and the corresponding properties of materials.

Symbols	Properties
V	The volume of conventional cell (Å^3)
M	The total mass of conventional cell
n	The number of atoms in conventional cell
n_p	The number of atoms in primitive cell
ρ	Mass density (g/cm^3)
B	Bulk Modulus (GPa)
G	Shear Modulus (GPa)
E	Young's Modulus (GPa)
ν	Poisson's ratio
H	Hardness (GPa)
B'	The derivative of B with respect to volume

G'	The derivative of G with respect to volume
ν_L	Sound velocity of longitude waves (10^3 m/s)
ν_S	Sound velocity of shear waves (10^3 m/s)
ν_a	The averaged sound velocity (10^3 m/s)
Θ_D	The Debye temperature
γ_L	The Grüneisen parameter of longitude waves
γ_S	The Grüneisen parameter of shear waves
γ	The overall Grüneisen parameter
A	The parameter in the Slack model
κ	The Lattice thermal conductivity

First-principles. — All the basic properties mentioned above can be obtained from first-principles calculations on the basis of the density functional theory (DFT), and the calculation uses the projector augmented wave (PAW) technique[54], which is implemented in the Vienna *ab initio* simulation package (VASP)[55]. The generalized gradient approximation (GGA) Perdew-Burke-Ernzerhof (PBE)[56] is taken as the exchange-correlation functional, while the kinetic energy cutoff of wave functions for each material is set as the default maximum energy cutoff. For sampling the Brillouin Zone (BZ), A Monkhorst-Pack[57] *k*-mesh with the grid density of 0.42 $\pi/\text{Å}$ is used to sample the Brillouin Zone (BZ). The self-consistent field (SCF) calculations are converged with energy difference smaller than 10^{-5} eV. Before any further calculations, all the geometries are fully optimized with the maximal Hellmann-Feynman force smaller than 0.01 $\text{eV}/\text{Å}$. The elastic constants are calculated using the density functional perturbation theory (DFPT). The derivative of elastic properties (bulk and shear modulus) is evaluated by changing the volume from -1.5% to 1.5% (5 points in total).

3. Results and discussion

By performing elaborate testing, it is found that the neural network models, especially the LSTM, have better performance in predicting thermal conductivity of materials compared to other ML models. The best performance of LSTM is confirmed by both the lowest RMSE of 8.3593 and the lowest R^2 of 0.8866 when testing on the test dataset of 70 materials, while the lowest MAE value of 0.8799 is achieved by the CNN model.

The corresponding numerical data of material descriptors were obtained based on the *state-of-the-art* first-principles calculations (Table 1), which were used as input of the ML models. Fifteen different ML models were constructed, and each model was trained separately using 280 experimental data. To test the performance of the trained model, the trained model is used to predict the thermal conductivity of 70 separate materials in the test set. Different root mean square error (RMSE) values between the predictive values and the true values in the test set have been collected in Table 2.

Table 2: Comparison of evaluation metrics for predicting thermal conductivity among the 15 machine learning models.

ML model	RMSE of test set	R^2 of test set	MAE of test set	time cost
Linear	26.4930	0.8096	13.5803	6.49s
Ridge	26.3697	0.8103	13.5384	3.46s
SGD	17.4929	0.8241	9.5713	3.58s
linearSVR	11.9823	0.8479	6.8762	3.68s
sigmoidSVR	20.6119	0.3432	9.7734	1.94s
rbfSVR	14.6274	0.7547	6.9582	1.80s
ploySVR	12.0221	0.7496	7.2365	3.61s
Decisiontree	19.3780	0.5348	8.6358	4.56s
DGBT	10.3623	0.8158	6.9570	5.69s
Random Forests	9.6385	0.8767	6.0574	3.76s
lightGBM	12.9994	0.7398	7.7365	4.56s
ANN	8.7211	0.8593	5.7933	18.19s
CNN	8.4061	0.8799	5.1674	19.57s
RNN	8.3726	0.8748	5.3209	61.03s
LSTM	8.3593	0.8866	5.4011	125.46s

The performance of the 15 ML models is comparably visualized in Fig. 2, and the following observations can be concluded from Fig. 2 combined with Table 2 and Table S2: 1) The performance of testing even exceeds that of the training process, indicating the excellent fitting of the ML models. 2) The trained ML models show the capability of accurately predicting *κ* over 4 orders of magnitude. In particular, the high *κ* ~1000 W/mK is successfully predicted by the trained ML models, showing the ability of extrapolation prediction. 3) Nonlinear models, including tree-based models, ensemble learning models and neural network models, have better prediction performance. Compared with generalized linear models, they have great advantages in predicting thermal conductivity, which also reflects the highly complex nonlinear relationship between thermal conductivity and basic properties of materials.

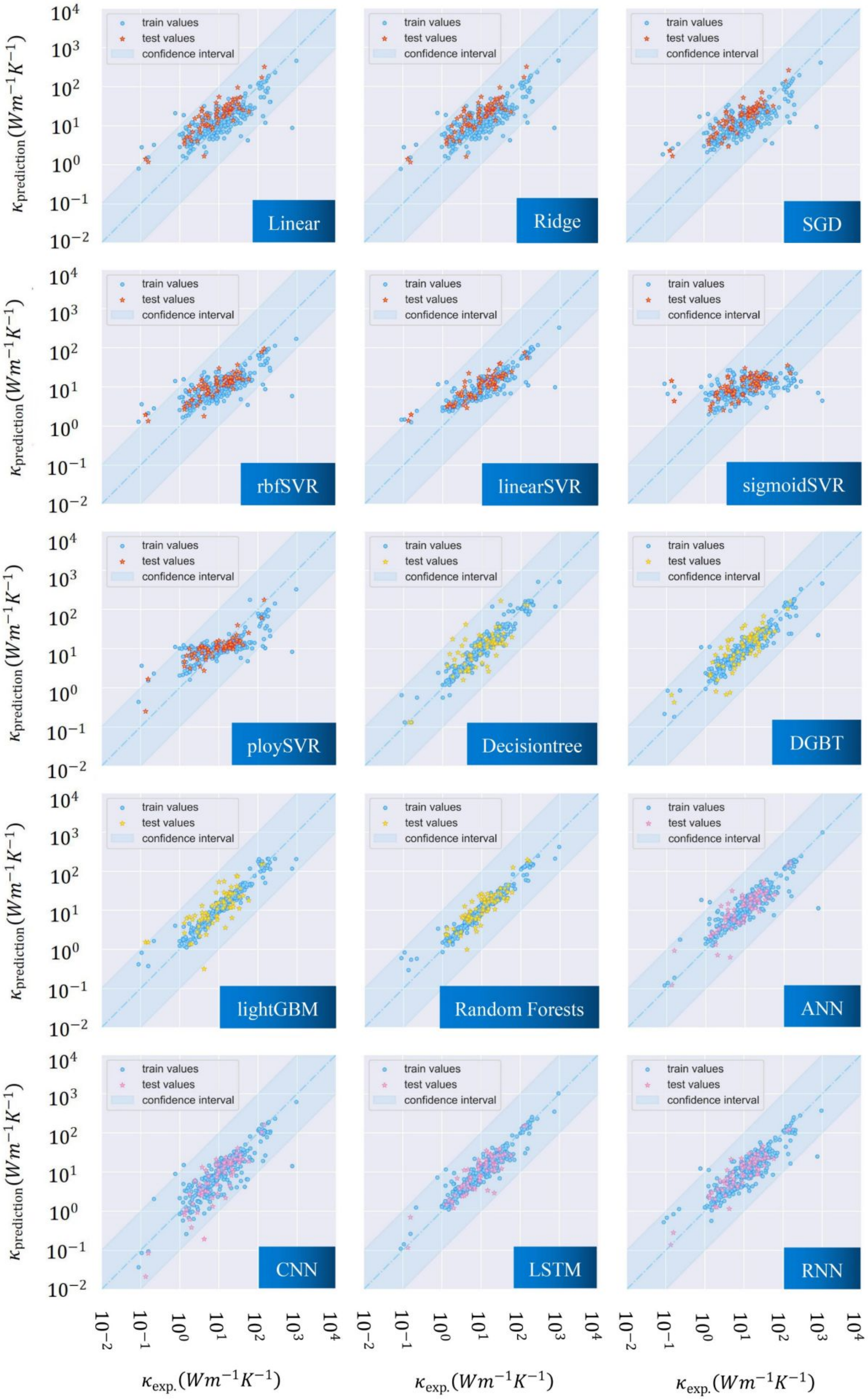

train values
test values
confidence interval
Linear
Ridge
SGD
rbfSVR
linearSVR
sigmoidSVR
ploySVR
Decisiontree
DGBT
lightGBM
Random Forests
ANN
CNN
LSTM
RNN
$\kappa_{\text{prediction}}(Wm^{-1}K^{-1})$
$\kappa_{\text{exp.}}(Wm^{-1}K^{-1})$
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}

Fig. 2 Comparative analysis of thermal conductivity (κ) predicted by 15 different machine learning models with comparison to experimental measurements. [21,40–45]. Both the training (80%, 280 materials) and testing (20%, 80 materials) data are plotted.

With the trained ML models, we further explore the thermal transport properties of more materials in three types: half Heusler (328),[21] materials with fcc structure (2,249),[58] and the 1521 dataset (1,139)[41]. The numbers in the parenthesis denote the numbers of materials in the specific structure type, where the non-stable structures are excluded. In combination with the performance of the above-mentioned 15 ML models in the thermal conductivity predictions (Fig. 2), four deep learning models with relatively better performance have been selected to predict the thermal conductivities of 3,716 materials. The selected models are ANN, CNN, RNN, and LSTM. The thermal conductivities predicted by the ML models are collected to those predicted by the optimized Slack model[49] to evaluate the performance. More details on the optimized Slack model[49] can be found in Supplementary Note S1.

To improve the predicting performance of these models, training set (80%, 280 materials) and test set (20%, 70 materials) are synthesized into a big dataset involving 350 materials. The combined dataset is used as a labeled training set. In addition, two incomplete supervised algorithms of active learning and semi-supervised learning algorithms, as well as K-fold cross-validation are used to predict the 1521, FCC, and Half Heusler datasets separately. The semi-supervised learning algorithm has shown relatively superior performance in this process as shown in Fig. 3, which may be due to the fact that the semi-supervised learning algorithm can efficiently utilize a large amount of unlabeled data to improve the generalization ability of the model. If the detailed prediction results of K-fold cross-validation and active learning are needed, please refer to the Supplementary Note S4 and S5. Moreover, the efficiency of the developed models is successfully demonstrated by the fact that each model only takes less than 30 seconds to predict the thermal conductivity of all materials in a single test set.

Moreover, the averaged predicted values of the four models are taken to reflect the average performance of these ML models. At the same time, this method is also taken to avoid the impact of prediction error from a single model on the overall prediction results. The comparison of the developed four deep semi-supervised learning models and the Slack model is shown in Fig. 3.

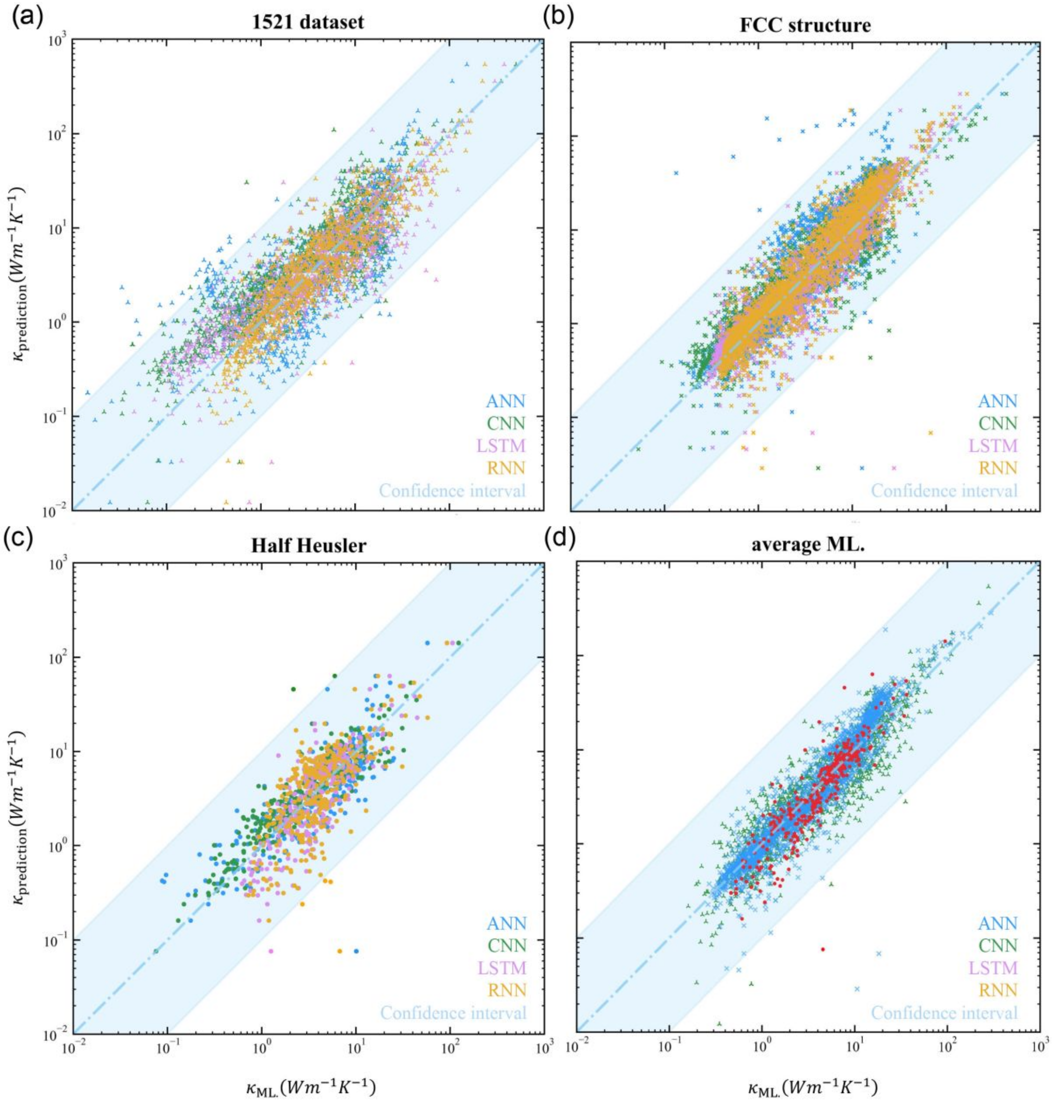


Fig. 3 Comparison between the $\kappa_{\text{prediction}}$ calculated by the optimized Slack model [Eq.(8)] and the $\kappa_{\text{ML.}}$ predicted by the four deep semi-supervised learning models for a large set of materials: (a)the 1521 dataset, (b)FCC structures, (c)half Heusler, (d)average prediction values of the four deep learning models. The blue shade marks the boundary of the discrepancies by one order of magnitude higher and lower.

Overall, the $\kappa_{\text{prediction}}$ agrees well with the $\kappa_{\text{ML.}}$, particularly in terms of high thermal conductivity predictions. Such an excellent agreement verifies the outstanding performance of the well-trained ML models in predicting thermal conductivity. The discrepancy mainly lies in about one order of magnitude, which might stem from the uncertainty in the κ prediction of the optimized Slack model as revealed in previous studies[11,49,50].

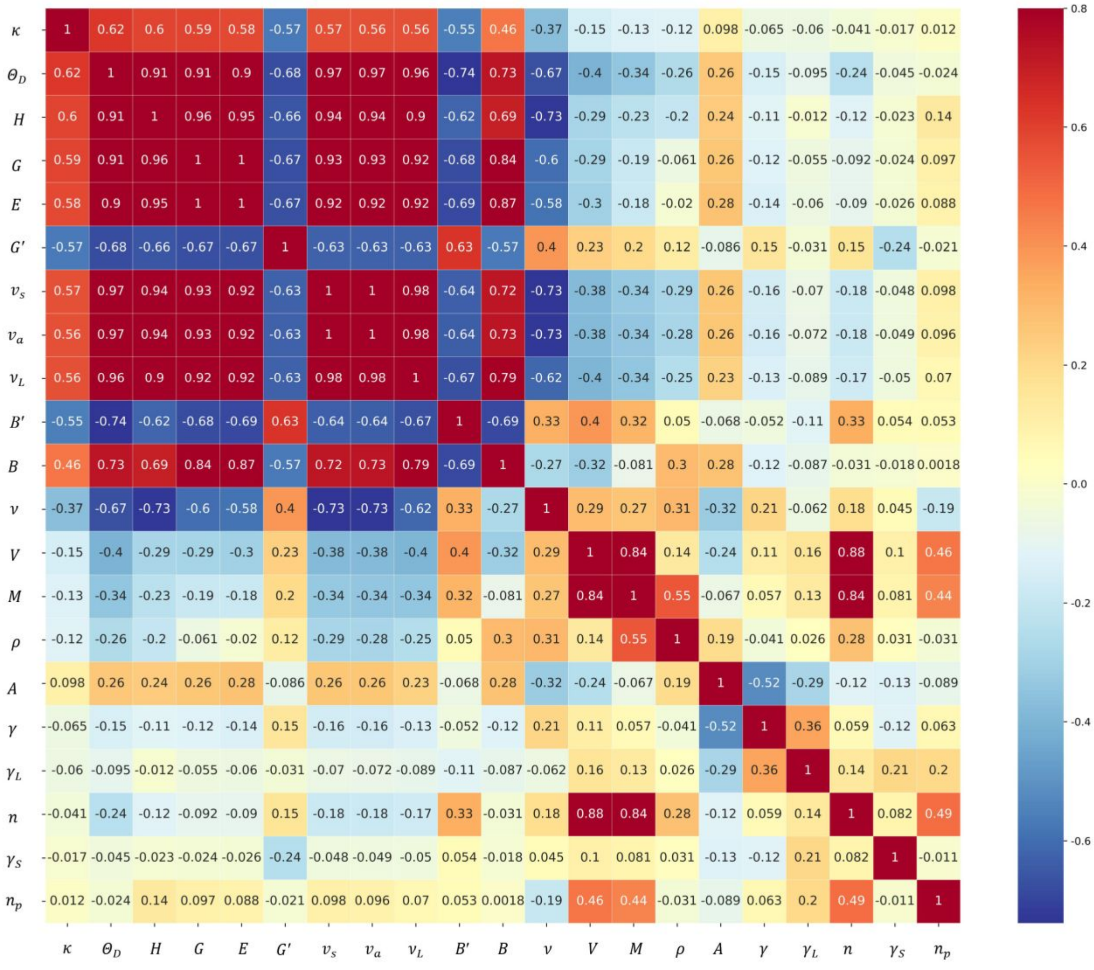
It is interesting to note that some of the $\kappa_{\mathrm{prediction}}$ predicted by the optimized Slack model for the 3716 materials deviate largely from those $\kappa_{\mathrm{ML.}}$ predicted by the ML models by more than one order of magnitude, as shown in Fig. 3. Similar performance of the optimized Slack model is also observed when compared with the experimentally measured $\kappa_{\mathrm{Exp.}}$ for the 350 materials as marked in the Supplementary Fig. S1. Besides, the κ predicted by the Slack model agree very well with those from the ML models for the high-κ materials with κ >~300 W/mK.

To have a deep insight into the performance of the ML models, the Pearson correlation coefficient map for the 21 properties (Table 1) of materials is generated. Among the 21 properties, all the other properties can be derived from the 8 basic properties of V, M, n, n_p, B, G, B', and G' (Table 1). The relationship between these properties can be found in Supplementary Note S6. The Pearson correlation coefficient is calculated based on the 350 materials with κ available from experiments using the formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{9}$$

,where $\bar{x}$ and $\bar{y}$ denote the average value of x and y respectively, n is 350, and r is the Pearson correlation coefficient between the properties of x and y, which indicates the correlation strength. The values of 1 and -1 represent totally positive and negative linear correlations, respectively. As shown in Fig. 4, the κ positively correlates with B, G, E, H, ν_L, ν_S, ν_a and Θ_D, where E, H, ν_L, ν_S, ν_a and Θ_D can be derived from the basic properties of V, M, n, n_p, B and G[50]. Among these properties, B and G can be used to represent the harmonicity of materials. On the contrary, the κ negatively correlates with v, γ_L, γ_S, γ, B', and G', where the γ_L, γ_S, γ can be derived from the basic properties of B' and G' [50]. These properties can be used to represent the anharmonicity of materials. Thus, the combination of the basic properties including V, M, n, n_p, B, G, B', and G' as descriptors is effective for the prediction of κ using the ML models.

Note that the overall Grüneisen parameter (γ) quantifying the phonon anharmonicity can be also derived from the basic properties of B' and G',[50] which shows almost no correlation with the κ. The weak correlation is abnormal from first look because they are supposed to be strongly negatively correlated. The reason might be that, 1) the γ is not correctly calculated in the Slack model, which also explains the generally more than one order of magnitude discrepancy of the $\kappa_{\mathrm{Slack\ model}}$ from the κ_{Exp}; 2) a single value of γ is not sufficient to fully describe the complex phonon anharmonicity of crystalline materials, which, rigorously speaking, should be phonon mode dependent $\gamma(\omega, q)$. The results suggest that further improvement to the Slack model and more accurate formula for γ is needed to better describe the phonon thermal transport.



	κ	Θ_D	H	G	E	G'	v_s	v_a	ν_L	B'	B	ν	V	M	ρ	A	γ	γ_L	n	γ_S	n_p
κ	1	0.62	0.6	0.59	0.58	-0.57	0.57	0.56	0.56	-0.55	0.46	-0.37	-0.15	-0.13	-0.12	0.098	-0.065	-0.06	-0.041	-0.017	0.012
Θ_D	0.62	1	0.91	0.91	0.9	-0.68	0.97	0.97	0.96	-0.74	0.73	-0.67	-0.4	-0.34	-0.26	0.26	-0.15	-0.095	-0.24	-0.045	-0.024
H	0.6	0.91	1	0.96	0.95	-0.66	0.94	0.94	0.9	-0.62	0.69	-0.73	-0.29	-0.23	-0.2	0.24	-0.11	-0.012	-0.12	-0.023	0.14
G	0.59	0.91	0.96	1	1	-0.67	0.93	0.93	0.92	-0.68	0.84	-0.6	-0.29	-0.19	-0.061	0.26	-0.12	-0.055	-0.092	-0.024	0.097
E	0.58	0.9	0.95	1	1	-0.67	0.92	0.92	0.92	-0.69	0.87	-0.58	-0.3	-0.18	-0.02	0.28	-0.14	-0.06	-0.09	-0.026	0.088
G'	-0.57	-0.68	-0.66	-0.67	-0.67	1	-0.63	-0.63	-0.63	0.63	-0.57	0.4	0.23	0.2	0.12	-0.086	0.15	-0.031	0.15	-0.24	-0.021
v_s	0.57	0.97	0.94	0.93	0.92	-0.63	1	1	0.98	-0.64	0.72	-0.73	-0.38	-0.34	-0.29	0.26	-0.16	-0.07	-0.18	-0.048	0.098
v_a	0.56	0.97	0.94	0.93	0.92	-0.63	1	1	0.98	-0.64	0.73	-0.73	-0.38	-0.34	-0.28	0.26	-0.16	-0.072	-0.18	-0.049	0.096
ν_L	0.56	0.96	0.9	0.92	0.92	-0.63	0.98	0.98	1	-0.67	0.79	-0.62	-0.4	-0.34	-0.25	0.23	-0.13	-0.089	-0.17	-0.05	0.07
B'	-0.55	-0.74	-0.62	-0.68	-0.69	0.63	-0.64	-0.64	-0.67	1	-0.69	0.33	0.4	0.32	0.05	-0.068	-0.052	-0.11	0.33	0.054	0.053
B	0.46	0.73	0.69	0.84	0.87	-0.57	0.72	0.73	0.79	-0.69	1	-0.27	-0.32	-0.081	0.3	0.28	-0.12	-0.087	-0.031	-0.018	0.0018
ν	-0.37	-0.67	-0.73	-0.6	-0.58	0.4	-0.73	-0.73	-0.62	0.33	-0.27	1	0.29	0.27	0.31	-0.32	0.21	-0.062	0.18	0.045	-0.19
V	-0.15	-0.4	-0.29	-0.29	-0.3	0.23	-0.38	-0.38	-0.4	0.4	-0.32	0.29	1	0.84	0.14	-0.24	0.11	0.16	0.88	0.1	0.46
M	-0.13	-0.34	-0.23	-0.19	-0.18	0.2	-0.34	-0.34	-0.34	0.32	-0.081	0.27	0.84	1	0.55	-0.067	0.057	0.13	0.84	0.081	0.44
ρ	-0.12	-0.26	-0.2	-0.061	-0.02	0.12	-0.29	-0.28	-0.25	0.05	0.3	0.31	0.14	0.55	1	0.19	-0.041	0.026	0.28	0.031	-0.031
A	0.098	0.26	0.24	0.26	0.28	-0.086	0.26	0.26	0.23	-0.068	0.28	-0.32	-0.24	-0.067	0.19	1	-0.52	-0.29	-0.12	-0.13	-0.089
γ	-0.065	-0.15	-0.11	-0.12	-0.14	0.15	-0.16	-0.16	-0.13	-0.052	-0.12	0.21	0.11	0.057	-0.041	-0.52	1	0.36	0.059	-0.12	0.063
γ_L	-0.06	-0.095	-0.012	-0.055	-0.06	-0.031	-0.07	-0.072	-0.089	-0.11	-0.087	-0.062	0.16	0.13	0.026	-0.29	0.36	1	0.14	0.21	0.2
n	-0.041	-0.24	-0.12	-0.092	-0.09	0.15	-0.18	-0.18	-0.17	0.33	-0.031	0.18	0.88	0.84	0.28	-0.12	0.059	0.14	1	0.082	0.49
γ_S	-0.017	-0.045	-0.023	-0.024	-0.026	-0.24	-0.048	-0.049	-0.05	0.054	-0.018	0.045	0.1	0.081	0.031	-0.13	-0.12	0.21	0.082	1	-0.011
n_p	0.012	-0.024	0.14	0.097	0.088	-0.021	0.098	0.096	0.07	0.053	0.0018	-0.19	0.46	0.44	-0.031	-0.089	0.063	0.2	0.49	-0.011	1

Fig. 4 The Pearson correlation coefficient map for the 21 properties of materials as listed in Tab. 1, which is calculated based on the 350 materials with κ available from experiments. The onsite values indicate the correlation strength, with 1 and -1 representing totally positive and negative linear correlations, respectively.

4. Conclusion

In summary, fifteen ML models are constructed and trained for accurate κ prediction. During the training process, 8 basic properties of materials are used as descriptors (inputs) and the experimentally measured κ are used as targets (output). The trained ML models, especially the deep learning models show the capability of accurately predicting thermal conductivity spanning 4 orders of magnitude, which have a great advantage over the widely used empirical Slack model. With the trained 4 deep learning models, combined with semi-supervised learning strategy, the thermal transport properties of 3,716 materials are further predicted, and the results are also verified by the optimized Slack model. Furthermore, the Pearson correlation coefficient map for 21 thermal-related properties of materials is generated to gain a deep insight into the performance of the ML models. It is confirmed that the combination of the basic properties of B (+), G (+), B′ (−), and G′ (−) as descriptors is effective for the prediction of κ using the ML models, where (+) and (−) denote the positive and negative correlation with κ, respectively. The developed ML models in our work for fast and accurately predicting thermal conductivity provide a powerful tool for the large-scale thermal material screening with targeted thermal transport property.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 52006057), the Fundamental Research Funds for the Central Universities (Grant Nos. 531119200237 and 541109010001), and the State Key Laboratory of Advanced Design and Manufacturing for Vehicle Body at Hunan University (Grant No. 52175013). This work was also supported in part by the NSF (award number: 2030128, 2110033) and ASPIRE grant from the Office of the Vice President for Research at the University of South Carolina (project 80005046). The numerical calculations in this paper have been done on the E.T. Cluster and the supercomputing system of the National Supercomputing Center in Changsha.

References

1 D. G. Cahill, P. V. Braun, G. Chen, D. R. Clarke, S. Fan, K. E. Goodson, P. Keblinski, W. P. King, G. D. Mahan, A. Majumdar, H. J. Maris, S. R. Phillpot, E. Pop and L. Shi, *Applied Physics Reviews*, 2014, **1**, 011305.

2 S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito and O. Levy, *Nature Mater*, 2013, **12**, 191–201.

3 E. R. Burns, Y. Zhu, H. Zhan, M. Manga, C. F. Williams, S. E. Ingebritsen and J. B. Dunham, *Water Resources Research*, 2017, **53**, 3341–3351.

4 G. Qin, Q.-B. Yan, Z. Qin, S.-Y. Yue, H.-J. Cui, Q.-R. Zheng and G. Su, *Sci Rep*, 2014, **4**, 1–11.

5 G. Qin, K.-R. Hao, Q.-B. Yan, M. Hu and G. Su, *Nanoscale*, 2019, **11**, 5798–5806.

6 G. Qin, Z. Qin, S.-Y. Yue, Q.-B. Yan and M. Hu, *Nanoscale*, 2017, **9**, 7227–7234.

7 G. Qin, Z. Qin, H. Wang and M. Hu, *Phys. Rev. B*, 2017, **95**, 195416.

8 Y. Zhang, E. Skoug, J. Cain, V. Ozoliņš, D. Morelli and C. Wolverton, *Phys. Rev. B*, 2012, **85**, 054306.

9 J. Callaway, *Phys. Rev.*, 1959, **113**, 1046–1051.

10 D. T. Morelli, J. P. Heremans and G. A. Slack, *Phys. Rev. B*, 2002, **66**, 195304.

11 *Journal of Physics and Chemistry of Solids*, 1973, **34**, 321–335.

12 X. Zhang, H. Xie, M. Hu, H. Bao, S. Yue, G. Qin and G. Su, *Phys. Rev. B*, 2014, **89**, 054310.

13 W. Li, J. Carrete, N. A. Katcho and N. Mingo, *Computer Physics Communications*, 2014, **185**, 1747–1758.

14 S. Li, L. Yu, C. Qi, K. Du, G. Qin and Z. Xiong, *Front. Mater.*, , DOI:10.3389/fmats.2021.725219.

15 H. Wang, G. Qin, G. Li, Q. Wang and M. Hu, *Phys. Chem. Chem. Phys.*, 2017, **19**, 12882–12889.

16 Z. Qin, G. Qin, X. Zuo, Z. Xiong and M. Hu, *Nanoscale*, 2017, **9**, 4295–4309.

17 E. Zhou, J. Wu, C. Shen, H. Zhang and G. Qin, *Journal of Applied Physics*, 2022, **131**, 185702.

18 C. Shen, N. Hadaeghi, H. K. Singh, T. Long, L. Fan, G. Qin and H. Zhang, *J. Mater. Chem. C*, 2022, **10**, 1436–1444.

19 Y. Han, Y. Zhou, G. Qin, J. Dong, D. S. Galvao and M. Hu, *Carbon*, 2017, **122**, 374–380.

20 H. Wang, G. Qin, G. Li, Q. Wang and M. Hu, *2D Mater.*, 2017, **5**, 015022.

21 J. Carrete, W. Li, N. Mingo, S. Wang and S. Curtarolo, *Phys. Rev. X*, 2014, **4**, 011019.

22 J. Schmidhuber, *Neural Networks*, 2015, **61**, 85–117.

23 G. Pilania, C. Wang, X. Jiang, S. Rajasekaran and R. Ramprasad, *Sci Rep*, 2013, **3**, 1–6.

24 R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi and C. Kim, *npj Comput Mater*, 2017, **3**, 1–13.

25 J. L. McDonagh, A. F. Silva, M. A. Vincent and P. L. A. Popelier, *J. Chem. Theory Comput.*, 2018, **14**, 216–224.

26 G. Pilania, J. E. Gubernatis and T. Lookman, *Computational Materials Science*, 2017, **129**, 156–163.

27 G. Pilania and X.-Y. Liu, *J Mater Sci*, 2018, **53**, 6652–6664.

28 A. Seko, H. Hayashi, K. Nakayama, A. Takahashi and I. Tanaka, *Phys. Rev. B*, 2017, **95**, 144110.

29 L. Ward, R. Liu, A. Krishna, V. I. Hegde, A. Agrawal, A. Choudhary and C. Wolverton, *Phys. Rev. B*, 2017, **96**, 024104.

30 Y. Zhuo, A. Mansouri Tehrani and J. Brgoch, *J. Phys. Chem. Lett.*, 2018, **9**, 1668–1673.

31 L. Ward, A. Agrawal, A. Choudhary and C. Wolverton, *npj Comput Mater*, 2016, **2**, 1–7.

32 M. Abadi, in *Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming*, Association for Computing Machinery, New York, NY, USA, 2016, p. 1.

33 J. A. Nelder and R. W. M. Wedderburn, *Journal of the Royal Statistical Society. Series A (General)*, 1972, **135**, 370.

34 J. Kivinen, A. J. Smola and R. C. Williamson, *IEEE Transactions on Signal Processing*, 2004, **52**, 2165–2176.

35 E. Gatnar, *Classification, Clustering, and Data Analysis*, 2002, 399–407.

36 W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S-PLUS*, 1999, 303–327.

37 L. Breiman, *Mach Learn*, 1996, **24**, 123–140.

38 Y. Freund and R. E. Schapire, in *In: Thirteenth International Conference on ML*, 1996, pp. 148–156.

39 X. Wan, W. Feng, Y. Wang, H. Wang, X. Zhang, C. Deng and N. Yang, *Nano Lett.*, 2019, **19**, 3387–3395.

40 S. A. Miller, P. Gorai, B. R. Ortiz, A. Goyal, D. Gao, S. A. Barnett, T. O. Mason, G. J. Snyder, Q. Lv, V. Stevanović and E. S. Toberer, *Chem. Mater.*, 2017, **29**, 2494–2501.

41 G. Petretto, S. Dwaraknath, H. P. C. Miranda, D. Winston, M. Giantomassi, M. J. van Setten, X. Gonze, K. A. Persson, G. Hautier and G.-M. Rignanese, *Sci Data*, 2018, **5**, 1–12.

42 J. J. Plata, P. Nath, D. Usanmaz, J. Carrete, C. Toher, M. de Jong, M. Asta, M. Fornari, M. B. Nardelli and S. Curtarolo, *npj Comput Mater*, 2017, **3**, 1–10.

43 A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput and I. Tanaka, *Phys. Rev. Lett.*, 2015, **115**, 205901.

44 C. Toher, J. J. Plata, O. Levy, M. de Jong, M. Asta, M. B. Nardelli and S. Curtarolo, *Phys. Rev. B*, 2014, **90**, 174107.

45 A. van Roekeghem, J. Carrete, C. Oses, S. Curtarolo and N. Mingo, *Phys. Rev. X*, 2016, **6**, 041061.

46 J. Schmidt, M. R. G. Marques, S. Botti and M. A. L. Marques, *npj Comput Mater*, 2019, **5**, 1–36.

47 F. S. Guthery, K. P. Burnham and D. R. Anderson, *The Journal of Wildlife Management*, 2003, **67**, 655.

48 G. Van Houdt, C. Mosquera and G. Nápoles, *Artif Intell Rev*, 2020, **53**, 5929–5955.

49 G. Qin, A. Huang, Y. Liu, H. Wang, Z. Qin, X. Jiang, J. Zhao, J. Hu and M. Hu, *Mater. Adv.*, 2022, **3**, 6826–6830.

50 T. Jia, G. Chen and Y. Zhang, *Phys. Rev. B*, 2017, **95**, 155206.

51 D. H. Chung and W. R. Buessem, *Journal of Applied Physics*, 1967, **38**, 2010–2012.

52 J. M. J. den Toonder, J. A. W. van Dommelen and F. P. T. Baaijens, *Modelling Simul. Mater. Sci. Eng.*, 1999, **7**, 909–928.

53 O. L. Anderson, *Journal of Physics and Chemistry of Solids*, 1963, **24**, 909–917.

54 G. Kresse and D. Joubert, *Phys. Rev. B*, 1999, **59**, 1758–1775.

55 G. Kresse and J. Furthmüller, *Phys. Rev. B*, 1996, **54**, 11169–11186.

56 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.

57 H. J. Monkhorst and J. D. Pack, *Phys. Rev. B*, 1976, **13**, 5188–5192.

58 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Materials*, 2013, **1**, 011002.