*Supporting information for*

# 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>*

## Note S1: optimized Slack model

To verify the prediction accuracy of the trained ML models, the optimized Slack model[1] with an updated coefficient ($A$) is used to predict the $\kappa$ for the 350 materials based on the basic properties, which are taken as descriptors in the ML models [Eq. (8) in the maintext], and the results are shown in Fig. S1. The $\kappa_{\text{Slack}}$ predicted by the optimized Slack model reasonably agrees with the experimentally measured $\kappa_{\text{Exp.}}$ with the discrepancy in about one order of magnitude, which has better performance than the previous prediction of the 350 types of materials[2] using the original Slack model. By comparing the performance of ML models (Fig. 2 in the main text) and the optimized Slack model, it is distinctly shown that the ML models have a great advantage over the Slack model for the accurate prediction of $\kappa$.

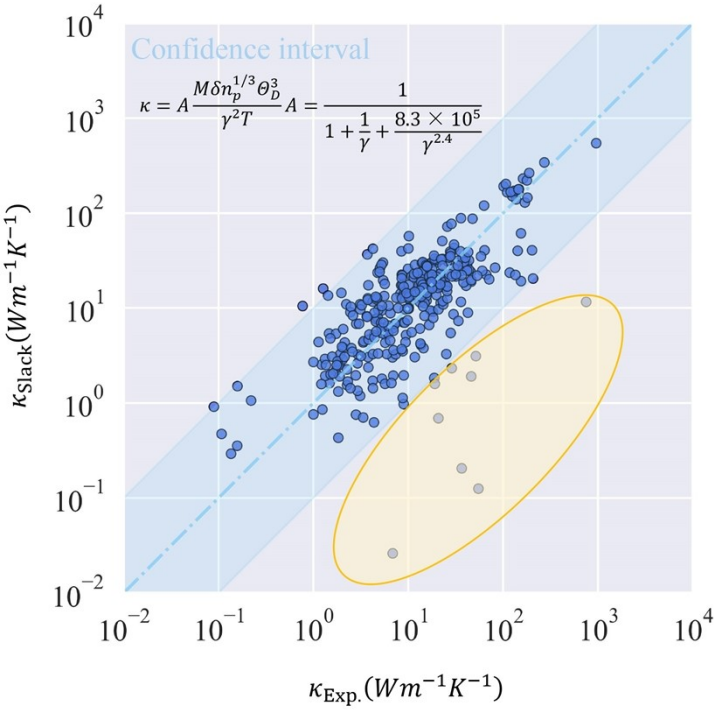


Fig. S1 The prediction accuracy of the optimized Slack model [Eq. (8)] against the experimental $\kappa$. The formula of the optimized Slack model is displayed as an illustration. The orange shaded ellipse marks the prediction with large discrepancy (generally more than one order of magnitude) from the $\kappa_{\text{Exp.}}$ The blue shade marks the boundary of the discrepancies by one order of magnitude higher and lower.

## Note S2: data visualization and feature engineering

A detailed analysis of the experimental datasets is conducted and the more in-depth exploration of the internal relationship between the variables has been carried out. Firstly, we use the t-SNE technique to perform dimensionality reduction analysis on some feature variables. T-SNE technology is a data structure visualization technology that can embed high-dimensional datasets into two-dimensional or three-dimensional spaces[3,4].

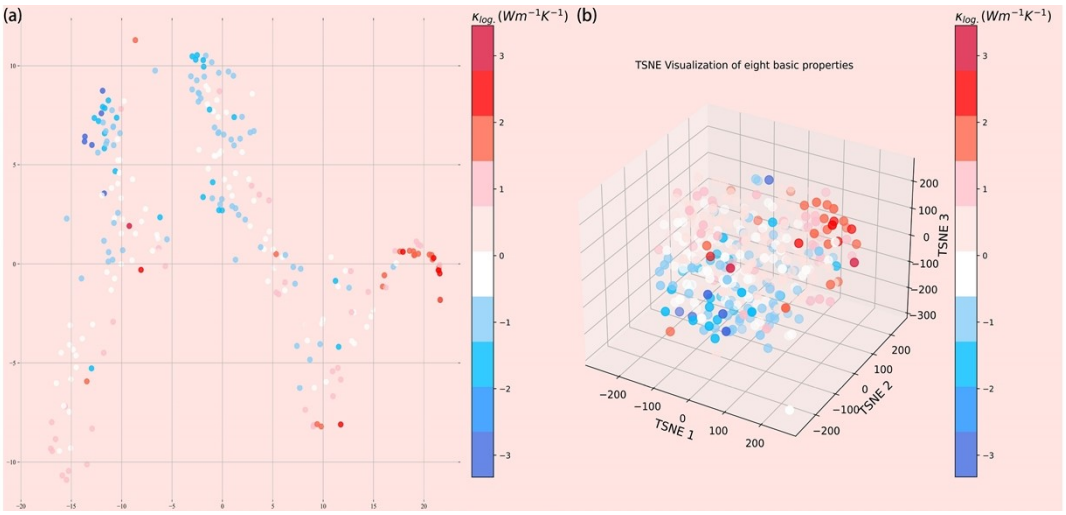


Fig. S2 T-SNE visualization of 8 basic properties containing $V$, $M$, $n$, $n_p$, $B$, $G$, $B'$, and $G'$. (a)2D t-SNE visualization. (b)3D t-SNE visualization. Each dot represents one type of material, and the coordinate axis is the measure of the two-dimensional or three-dimensional hidden variable obtained after dimensionality reduction of the 8-dimensional datasets.

At the same time, the feature embedding method of the materials is explored using the popular embedding network as well as the t-SNE technology. There are a large number of textual and categorical features, *i.e.* the names of the materials, space groups, ***etc***. The information can not only help to identify different materials when performing the visualization operations, but also will play an equally important role for thermal conductivity prediction and material representation. However, most of the information has high dimensionality, which is not suitable for traditional one-hot encoding method. Therefore, we use the training set containing 350 kinds of materials, and take the material names (350 categories) and space group types (25 categories) as the input of the embedding network, in order to map high-dimensional categorical variables to a low-dimensional learned representation, and then reduce the dimension to a 3-dimensional vector output. Furthermore,

we use t-SNE to convert 3D vectors to 2D vectors and visualize them, revealing the correlation between material characteristics and thermal conductivity as shown in Fig. S3.

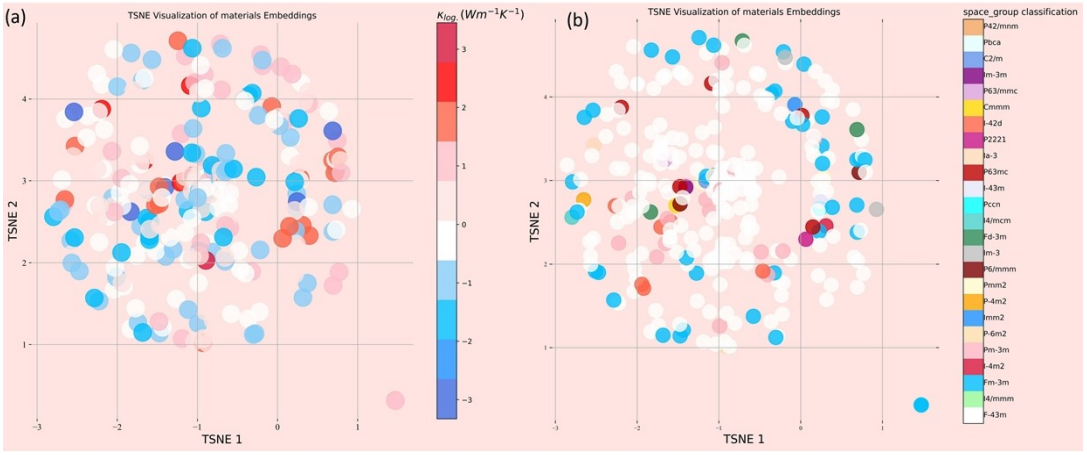


Fig. S3 T-SNE Visualization of materials Embeddings. (a) Colored by the magnitude of thermal conductivity (b) Colored by the space group categories. Axes provide a good description of material names and space group categories.

We have plotted a clustered heatmap with Seaborn Clustermap to explore the internal correlation and similarity between variables. The clustering between feature descriptors has been visualized using hierarchical clustering analysis dendrogram, and the results are shown in Fig. S4. Some variables have shown strong correlation with each other *i.e.* the similarity between $\Theta_D$ and the variables $\nu_L$, $\nu_S$, $\nu_a$, while the variables $V$, $M$, $n$ also showed a certain correlation.

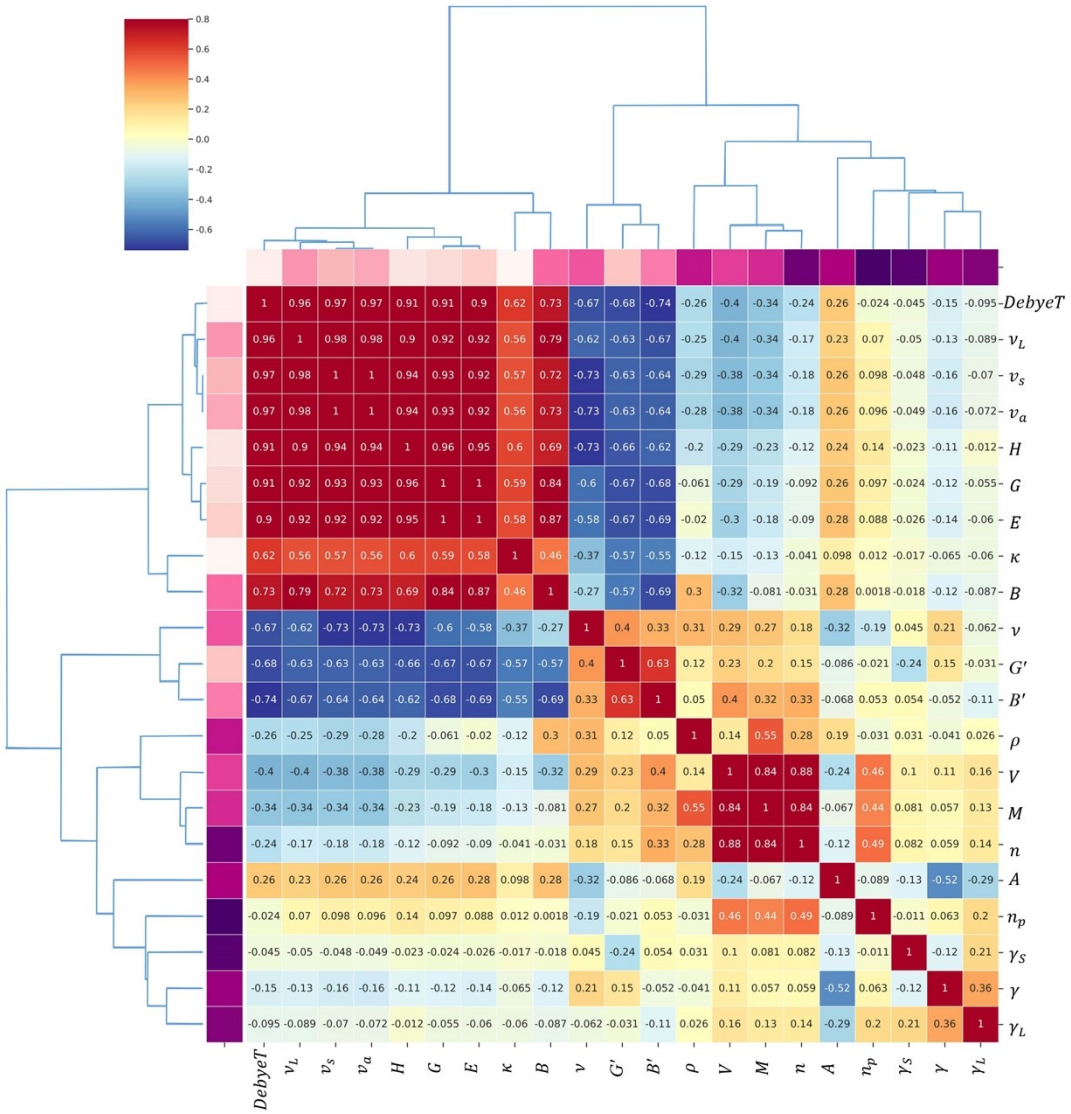


Fig. S4 Clustermap shows Pearson correlation among 21 feature descriptors. Rows and columns are grouped by cluster analysis using Euclidean distance as a measure. We use the clusters in the dendrogram to distinguish more similar materials, and the color labels represent the feature similarity between the variables.

Among the 21 properties in Table 1 of the manuscript, all the other properties can be derived from the 8 basic properties of $V$, $M$, $n$, $n_p$, $B$, $G$, $B'$, and $G'$. We use the 8 basic material properties to represent all the other properties from the perspective of formula, and show the relationships between these material properties in Table S1[5].

**Table S1:** The symbols and the relationship between material properties

| Symbols | expression |
|---|---|
| $V$(Å$^3$) | -- |
| $M$ | -- |
| $n$ | -- |
| $n_p$ | -- |
| $\rho$(g/cm$^3$) | $\frac{nM}{VN_A}$ |
| $B$(GPa) | --[a] |
| $G$(GPa) | --[a] |
| $E$(GPa) | $9 * B * G/(3 * B + G)$ |
| $\nu$ | $(3 * B - 2 * G)/2/(3 * B + G)$ |
| $H$(GPa) | $2 * (G^3/B^2)^{0.585}$ |
| $B'$ | $dB/dV$ |
| $G'$ | $dG/dV$ |
| $\nu_L$(10$^3$ m/s) | $\sqrt{\frac{(B + 4/3G)}{\rho}}$ |
| $\nu_S$(10$^3$ m/s) | $\sqrt{\frac{G}{\rho}}$ |
| $\nu_a$(10$^3$ m/s) | $\left[\frac{1}{3}\left(\frac{1}{v_L^3} + \frac{1}{v_S^3}\right)\right]^{-1/3}$ |
| $\Theta_D$ | $\frac{h}{k_B}\left[\frac{3V}{4\pi}\right]^{1/3}\nu_a n_p^{-1/3}$ |
| $\gamma_L$ | $-\frac{1}{2B + 4G/3}\frac{V}{\partial V}\frac{\partial(B + 4G/3)}{\partial V} - \frac{1}{6}$ |
| $\gamma_S$ | $-\frac{1V\partial G}{2G\partial V} - \frac{1}{6}$ |
| $\gamma$ | $\sqrt{\left[(\gamma_L)^2 + 2(\gamma_S)^2\right]/3}$ |
| $A$ | $\frac{1}{1 + 1/\gamma + 8.3 \times 10^5/\gamma^{2.4}}$ |
| $\kappa$ | --[b] |

[a] Bulk modulus ($B$) and shear modulus ($G$) describe the material's response to different kinds of stress in material science, which arise in the generalized Hooke's law:

The bulk modulus $B$ describes the material's response to uniform hydrostatic pressure, while the shear modulus $G$ describes the material's response to shear stress. As for isotropic materials, bulk modulus ($B$) and shear modulus ($G$) are not independent, which are connected via the equations[6]

$$E = 2G(1 + v) = 3B(1 - 2v) \quad \text{(S1)}$$

where $E$ is the Young's modulus describing the material's strain response to uniaxial stress in the direction of this stress, and $v$ is Poisson's ratio. All the properties can be found in Table 1 of the manuscript. However, complex anisotropic materials exhibit differing material response to stress or strain when tested in different directions. In those cases, Eq.(S1) does not hold, and the full generalized Hooke's law must be used to evaluate the stress on material[7].

[b] The $\kappa$ can be calculated theoretically by the Slack model, where the accuracy is limited by the lacking of phonon transport details.

## Note S3: Semi-supervised learning

The training idea of semi-supervised learning model using pseudo labelling technique[8,9] is briefly shown in Fig. S5. By training the labeled samples, the unlabeled samples are predicted and the prediction results are taken as new labels. Then, the unlabeled and labeled data are combined for further training, and the second time prediction is performed. The MSE value of the unlabeled data between the two prediction results should be evaluated, and the new labels can be defined by the part of the samples with the minimum MSE. Finally, it is necessary to iterate repeatedly according to the above steps until the error converges[10,11].

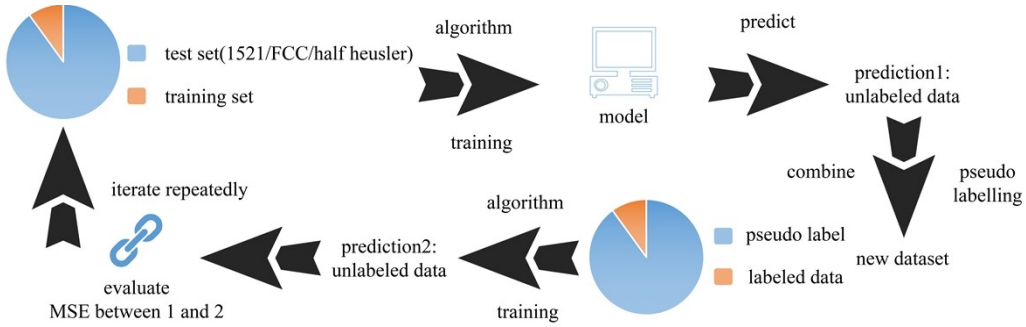


Fig. S5 working flow of semi-supervised learning.

## Note S4: Active learning

The general idea of active learning is to query the most useful unlabeled samples through some specific methods, and hand them over to experts for marking. Then, use the queried samples to train the model to improve model performance[12] (Fig. S6). A model with better performance can be obtained with fewer labeled samples, and thus it has been widely used in terms of machine learning.

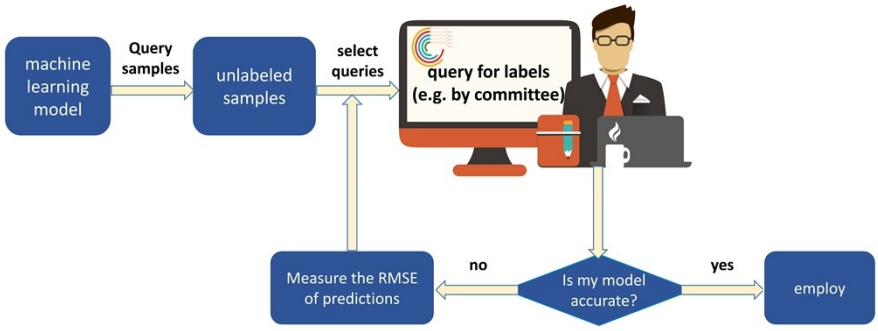


Fig. S6 working flow of active learning.

Aiming at achieving the active learning strategies, we integrate the encapsulated four Keras models into modAL workflow. Among the 350 materials, 10 of them have been randomly selected as initial data without repetition, and then we shall generate the pool by removing the initial data from the training dataset. After defining and initializing the active learner, the data are queried by Committee Regressor and put into y_new cyclically.

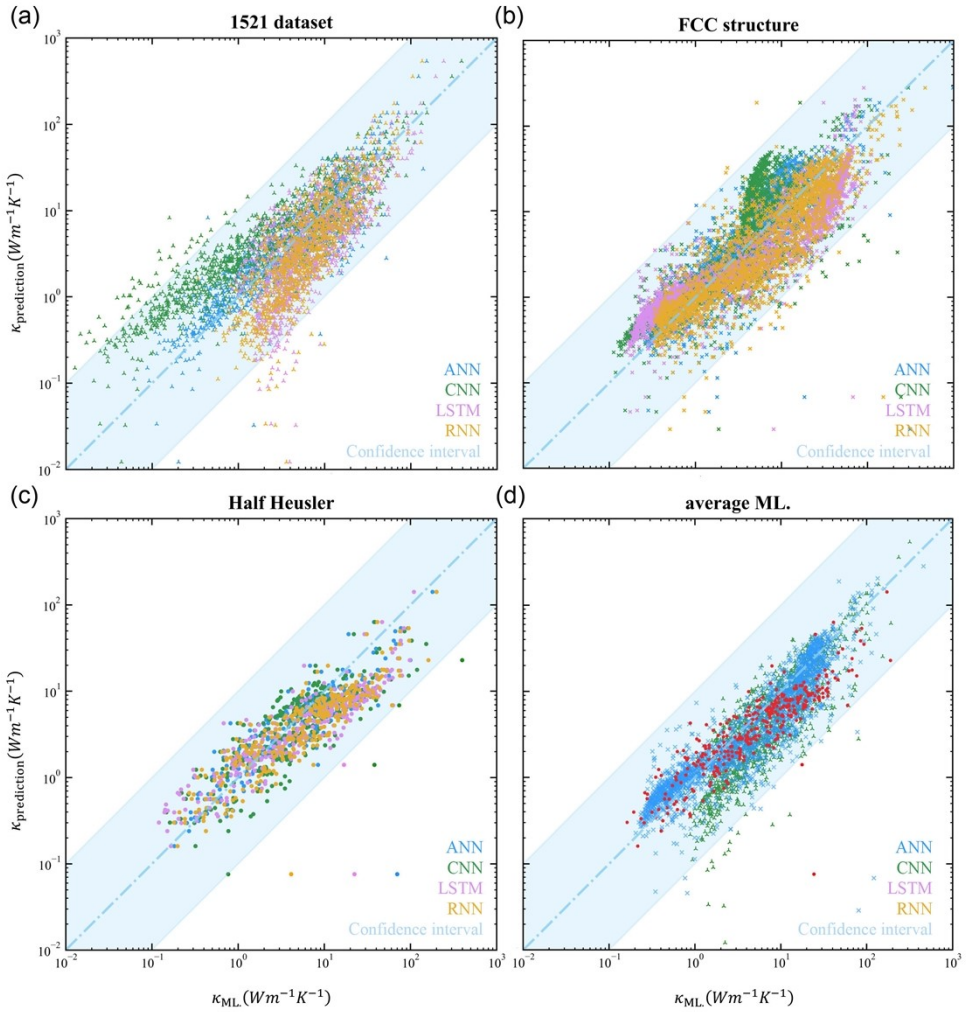


Fig. S7 Comparison between the $\kappa_{prediction}$ calculated by the optimized Slack model [Eq.(8)] and the $\kappa_{ML.}$ predicted by the four deep active learning models for a large set of materials: (a) the 1521 dataset, (b)FCC structures, (c)half Heusler, (d)average prediction of the four deep learning models. The blue shade marks the boundary of the discrepancies by one order of magnitude higher and lower.

## Note S5: K-fold cross-validation

As an important means to effectively prevent overfitting and reduce errors, K-fold cross-validation has been widely used in previous literatures. The method is used in this study as a means of model training and prediction[13,14]. Fig. S8 shows the principle of K-fold cross-validation, which can be described as: 1) divide the data set into *n* folds, while training the model on (*n*-1) folds, and the remaining 1-fold is used as the validation set; 2) the data should be iterated until each fold already participates in training, and the model score is calculated as the average of the *n*-folds validation score.

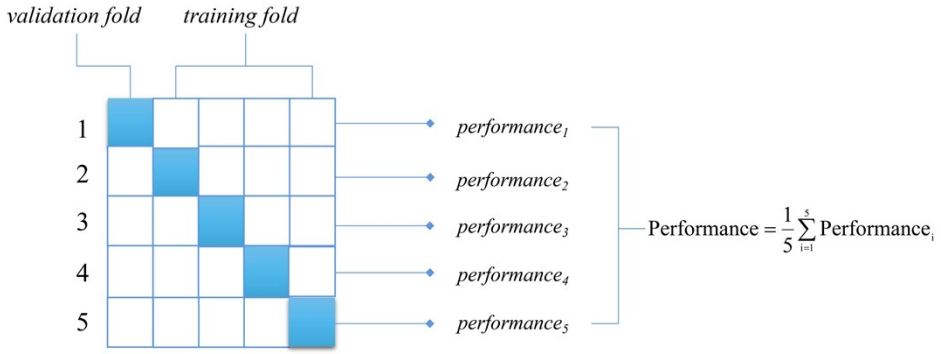


Fig. S8 Working principle of K-fold cross-validation taking the most widely used 5-fold cross-validation as an example.

The results of prediction using K-fold cross-validation alone are shown in Fig. S9. From the comparative analysis, it can be concluded that the use of incompletely supervised learning seems to have a better performance in obtaining results close to the true thermal conductivity values.

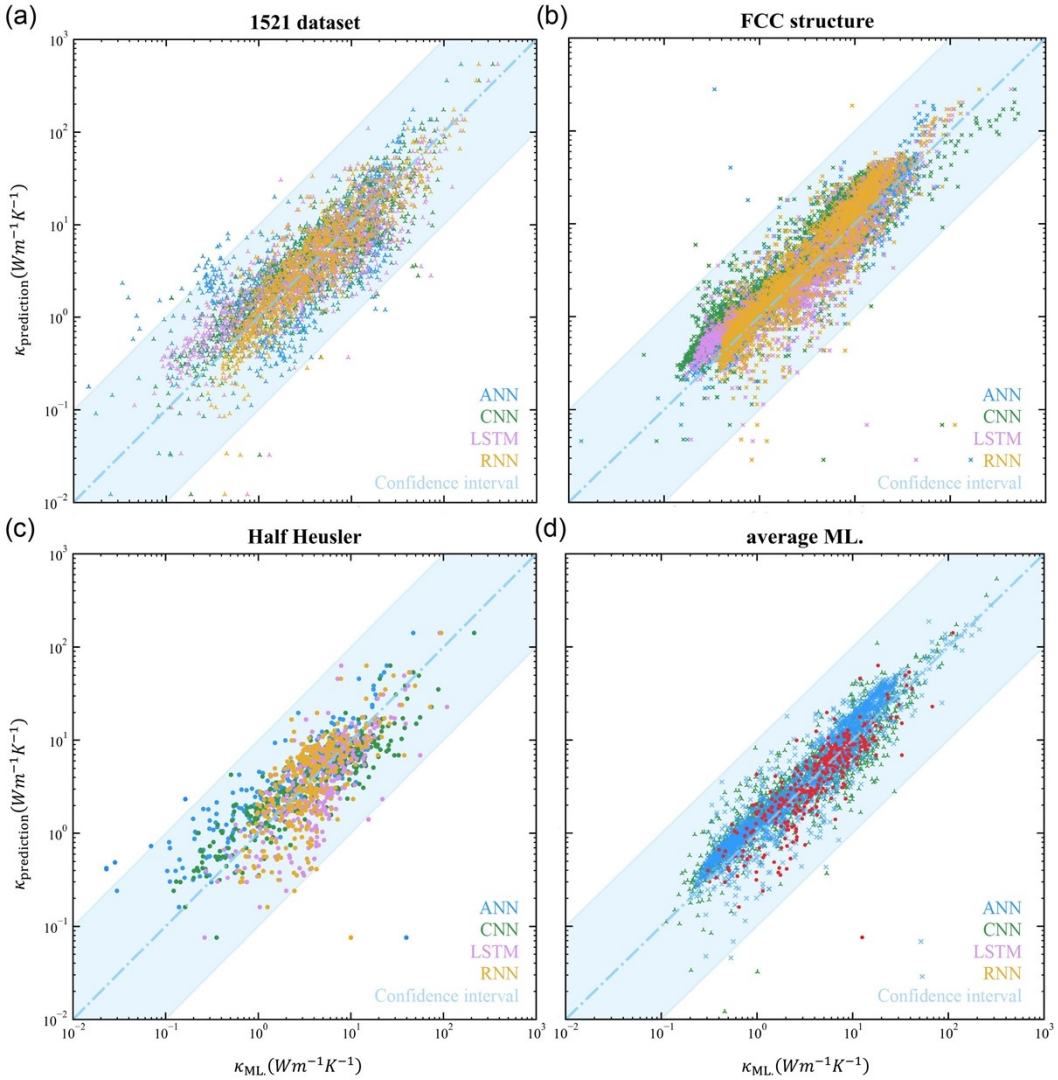


Fig. S9 Comparison between the $\kappa_{prediction}$ calculated by the optimized Slack model [Eq.(8)] and the $\kappa_{ML.}$ predicted by the four deep learning models using K-fold cross-validation strategy for a large set of materials: (a)the 1521 dataset, (b)FCC structures, (c)half Heusler, (d)average prediction of the four deep learning models. The blue shade marks the boundary of the discrepancies by one order of magnitude higher and lower.

## Note S6: Feature importance of eight basic material properties

Most of the models optimized for performance, especially deep learning models, are black-box models. Consequently, it is difficult to have a deep insight into feature importance of these models. To further explore interpretability of the models with good performance as revealed, we further try to explain our deep learning models with SHapley Additive exPlanations (SHAP) values.

The SHAP[15] provides a convenient way to explain the output of any machine learning model. We calculate feature importance of eight basic material properties based on validation set by SHAP, and use the shap.summary_plot() function to plot the analysis results in Fig. S10 and Fig. S11. The advantage of this tool is that it can clearly show SHAP values for all the features and all samples in validation set. In Fig. S10 and Fig. S11, the SHAP values have been sorted by the importance, and the first one is the most important feature. In addition, we provide useful information of how each feature affects the model output.

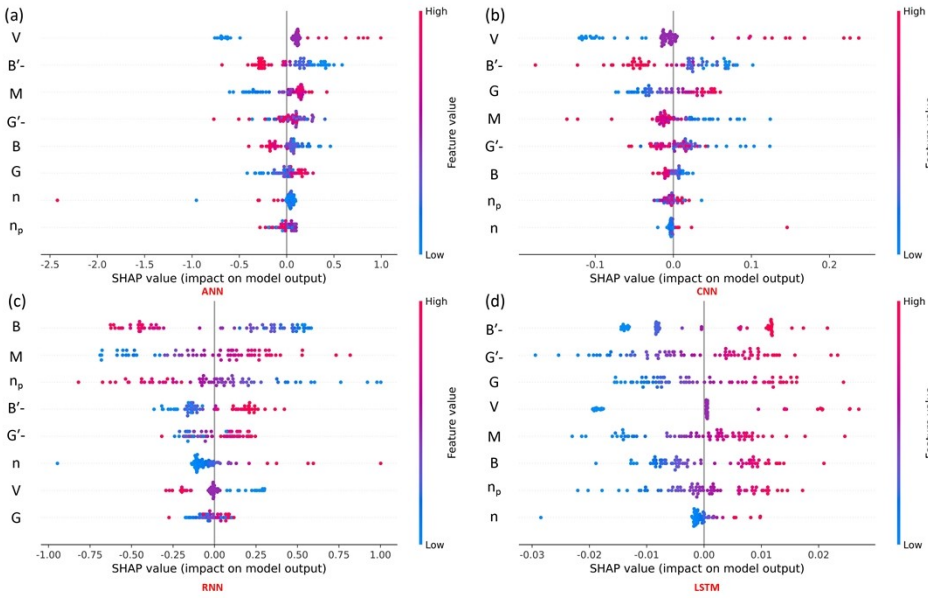


Fig. S10 Sorted SHAP values for all the features and all samples in models of (a) ANN, (b) CNN, (c) RNN, (d) LSTM validation set. The color represents the relationship between the size of the feature value and the predicted impact, where the characteristic value distribution is also displayed.

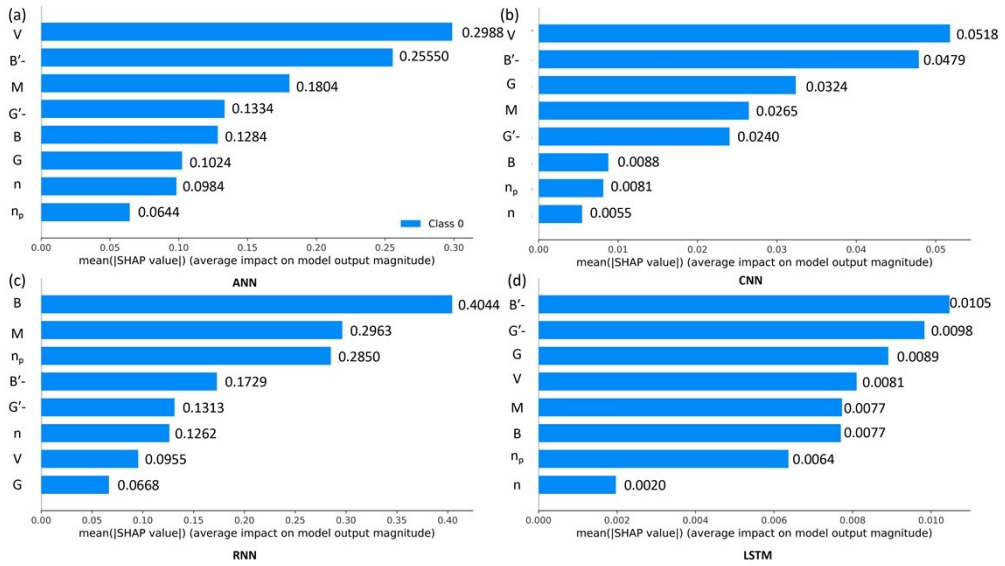


Fig. S11 Global feature importance of eight basic material properties in models of (a) ANN, (b)CNN, (c)RNN, (d)LSTM validation set. The standard bar plot is obtained by calculating the mean absolute value of the SHAP values for each feature.

## Note S7: Prediction performance of deep learning models for low-$\kappa$ materials

By analyzing the results of model training and testing, it can be easily concluded that four deep learning models mentioned in the manuscript have better ability to predict $\kappa$ values spanning over four orders of magnitude. Therefore, to show the prediction performance of these models for low-$\kappa$ materials more objectively and accurately, some low-$\kappa$ materials with thermal conductivities lower than 3 W/mK are selected from training set(280) and test set(70), and the comparison between their predicted values and first principle calculations of four deep learning models for selected materials with low thermal conductivity are listed in the following table.

**Table S2:** Prediction performance of deep learning models for low-$\kappa$ materials. Some low thermal conductivity materials containing Na, Ag and Cu elements are listed below, in which $\kappa_{exp.}$ represents first principle calculations, while $\kappa_{LSTM}$, $\kappa_{ANN}$, $\kappa_{CNN}$ and $\kappa_{RNN}$ stand for the predicted values of LSTM, ANN, CNN and RNN.

| Name of Materials | $\kappa_{exp.}$ | $\kappa_{LSTM}$ | $\kappa_{ANN}$ | $\kappa_{CNN}$ | $\kappa_{RNN}$ | Dataset |
|---|---|---|---|---|---|---|
| 22913-CuBr | 2.94 | 2.72 | 3.66 | 2.76 | 2.40 | training set |
| 20118-$Ag_2Cr_4Te_8$ | 0.107 | 0.143 | 0.140 | 0.0856 | 0.683 | training set |
| 22922-AgCl | 1.22 | 1.81 | 4.42 | 1.67 | 2.31 | training set |
| 22925-AgI | 1.51 | 1.92 | 2.66 | 1.10 | 3.040 | test set |
| 5342-$AgGaS_2$ | 1.79 | 1.66 | 1.32 | 0.62 | 1.28 | test set |
| 1100443-SeKNa | 2.13 | 2.53 | 3.31 | 1.56 | 3.67 | test set |
| 22916-NaBr | 1.32 | 1.48 | 2.21 | 0.912 | 2.55 | test set |
| 542680-$Au_4In_8Na_{12}$ | 0.158 | 0.264 | 0.188 | 0.0947 | 0.840 | training set |
| NaAsBa_1766119 | 2.91 | 3.05 | 4.09 | 4.44 | 2.92 | training set |
| NaBaSn_2750457 | 1 | 1.29 | 1.89 | 2.15 | 1.26 | training set |
| NaBiSr_2805096 | 1.83 | 1.76 | 2.66 | 2.72 | 2.04 | test set |
| NaGeBa_2694166 | 2.44 | 1.64 | 2.31 | 2.50 | 1.37 | training set |
| NaKTe_2672659 | 1.82 | 1.38 | 2.18 | 1.68 | 1.38 | training set |

## References


1 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.

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

3 L. van der Maaten and G. Hinton, *Journal of Machine Learning Research*, 2008, **9**, 2579–2605.

4 L. van der Maaten, in *Artificial Intelligence and Statistics*, PMLR, 2009, pp. 384–391.

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

6 L. D. Landau, E. M. Lifshitz, J. B. Sykes, W. H. Reid and E. H. Dill, *Physics Today*, 1960, **13**, 44–46.

7 *Wikipedia*, 2022.

8 S. Jain, *URL= https://www. analyticsvidhya. com/blog/2017/09/pseudo-labellingsemi-supervised-learning-technique*.

9 X. Zhu and A. B. Goldberg, *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2009, **3**, 1–130.

10 Z. Wu, J. Liang, Z. Zhang and J. Lei, *Journal of Biomedical Informatics*, 2021, **115**, 103683.

11 G. Kostopoulos, S. Karlos, S. Kotsiantis and O. Ragos, *Journal of Intelligent & Fuzzy Systems*, 2018, **35**, 1483–1500.

12 T. Danka and P. Horvath, , DOI:10.48550/arXiv.1805.00979.

13 P. Refaeilzadeh, L. Tang and H. Liu, in *Encyclopedia of Database Systems*, eds. L. LIU and M. T. ÖZSU, Springer US, Boston, MA, 2009, pp. 532–538.

14 T. Fushiki, *Stat Comput*, 2011, **21**, 137–146.

15 S. M. Lundberg and S.-I. Lee, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017, vol. 30.