Journal of Applied Physics

RESEARCH ARTICLE | OCTOBER 15 2024

Compositionally restricted atomistic line graph neural network for improved thermoelectric transport property predictions

Zeyu Wang ; Run Hu ; Xiaobing Luo ; Jinlong Ma

Check for updates

J. Appl. Phys. 136, 155103 (2024)
https://doi.org/10.1063/5.0226327

View Online
Export Citation

Articles You May Be Interested In

KoopmanLab: Machine learning for solving complex physics equations

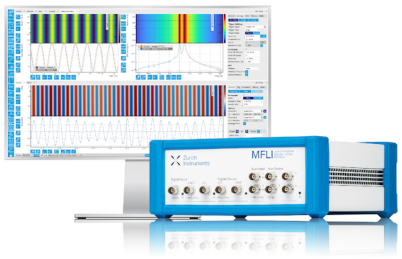
APL Mach. Learn. (September 2023)

Experimental realization of a quantum classification: Bell state measurement via machine learning

APL Mach. Learn. (September 2023)

16 October 2024 02:04:58

Challenge us.
What are your needs for periodic signal detection?
Zurich Instruments
Find out more



AIP Publishing

Journal of Applied Physics ARTICLE pubs.aip.org/aip/jap

Compositionally restricted atomistic line graph neural network for improved thermoelectric transport property predictions

Cite as: J. Appl. Phys. **136**, 155103 (2024); doi: 10.1063/5.0226327
Submitted: 1 July 2024 · Accepted: 2 October 2024 ·
Published Online: 15 October 2024

View Online Export Citation CrossMark

Zeyu Wang, Run Hu, Xiaobing Luo, and Jinlong Ma[a)]

AFFILIATIONS

School of Energy and Power Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

[a)]**Author to whom correspondence should be addressed:** majinlong@hust.edu.cn

ABSTRACT

Graph neural networks (GNNs) have evolved many variants for predicting the properties of crystal materials. While most networks within this family focus on improving model structures, the significance of atomistic features has not received adequate attention. In this study, we constructed an atomistic line GNN model using compositionally restricted atomistic representations which are more elaborate set of descriptors compared to previous GNN models, and employing unit graph representations that account for all symmetries. The developed model, named as CraLiGNN, outperforms previous representative GNN models in predicting the Seebeck coefficient, electrical conductivity, and electronic thermal conductivity that are recorded in a widely used thermoelectric properties database, confirming the importance of atomistic representations. The CraLiGNN model allows optional inclusion of additional features. The supplement of bandgap significantly enhances the model performance, for example, more than 35% reduction of mean absolute error in the case of 600 K and 10^{19} cm^{-3} concentration. We applied CraLiGNN to predict the unrecorded thermoelectric transport properties of 14 half-Heusler and 52 perovskite compounds, and compared the results with first-principles calculations, showing that the model has extrapolation ability to identify the thermoelectric potential of materials.

16 October 2024 02:04:58

© 2024 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/). https://doi.org/10.1063/5.0226327

I. INTRODUCTION

Thermoelectric (TE) materials can be used for devices that realize direct conversion between heat and electricity. This provides a solid-state solution of environmentally friendly recovery for low-grate waste heat, and enable TE cooling with many advantages such as compact in size and precise temperature control.[1,2] The conversion efficiency of TE materials is gauged by the dimensionless figure of merit, $ZT = S^2\sigma T/\kappa$, where S is the Seebeck coefficient, σ is the electrical conductivity, κ is the thermal conductivity composed of electronic thermal conductivity (κ_e) and lattice thermal conductivity (κ_l). Improving ZT necessitates the collaborative modulation of these multiple parameters, which is not an easy task due to their strongly adverse interdependence. Over the past two decades, although the ZT has been explosively improved thought multidimensional strategies,[3–5] the discovery of new high-potential TE materials remains one of the fundamental requirements for pushing TE applications toward greater efficiency and maturity.

In general, a critical step to establish the feasibility of TE materials depends on experimental synthesis and characterization. However, these procedures are economically expensive and time-consuming due to the vast scale of the materials search space.[6] Recent advances in computational materials science provide innovative avenues for material discovery, which have uncover thousands of previously unsuspected compounds and metastable structures.[7,8] Compared with experimental trial-and-error methods, density functional theory (DFT) calculations have been proven to be an efficient tool in pinpointing good TE candidates. One scheme is to calculate the electrical transport and lattice thermal transport properties as accurately as possible, by solving the Boltzmann transport equations (BTEs) with the mode-dependent relaxation times of charge and heat carriers computed from DFT calculations.[9,10] This approach gives considerably good agreement with experimental measurements,[9–12] but the heavy computational workload makes it difficult to undertake the task of selecting promising TE candidates from numerous materials. The other scheme is

© Author(s) 2024

the high-throughput approach, which aims at computing as many compounds as possible using simplified relaxation time. The high-throughput approach is a valuable tool in the initial screening of new classes of TE materials with favorable efficiency.[13–15]

The obtained results from high-throughput DFT calculations can be recorded as material databases[7,8,16,17] and, on one hand, for further processing such as more accurate first-principles examinations. On the other hand, with the availability of these databases, data-driven materials design and discovery using machine learning (ML) have gained great attention due to its potential to predict new materials with desired properties under much less time and computational cost than DFT calculations.[18,19] Among the various ML frameworks, the graph neural network (GNN) architecture applied to crystal materials has been rapidly evolving due to its capture of structure–property paradigm. The most typical model is the crystal graph convolutional neural networks (CGCNNs),[20] which constructs a crystal graph to represent the material with nodes for each constituent atom and edges for interatomic bonds. The elemental properties are used as node features and interatomic distances are used as edge features. Subsequently, different variants based on CGCNN framework have been proposed by developing representations that enable more unique and accurate identification of materials. One representative improvement is the geometric-information-enhanced crystal graph neural network (GeoCGNN) which encodes full topological and spatial geometric structures by considering the distance vector between each node and corresponding neighbors.[21] The other is the atomistic line graph neural network (ALIGNN) which explicitly incorporates the angle-based information by deriving a line graph in addition to the basis crystal graph.[22] Benefiting from these improved descriptions for structural features, GeoCGNN and ALIGNN achieve better performance than CGCNN in predicting many properties.

In addition to structural features, the enhanced atomistic features have also been proven to play an important role in improving CGCNN accuracy,[23] and leveraging additional features with transfer learning method can improve the predictions of various properties such as heat capacity and bandgap.[24,25] Recently, a more elaborate set of descriptors based on each elements and their chemical composition has been constructed and used in compositionally restricted attention-based network (CrabNet) model.[26] However, when applying the compositionally restricted representations to GNN architectures, there is a problem that the number of element representations may be smaller than the number of nodes in crystal graph for materials where an element occupies multiple atomic sites. To solve this problem, we performed translation and reflection symmetry operations for a crystal unit cell,[27] as a result, all symmetry-equivalent atoms are merged into one node in the graph. Eventually, we can build the compositionally restricted atomistic representations upon the ALIGNN framework and present a compositionally restricted atomistic line graph neural network (CraLiGNN) model. In the prediction of TE involved properties such as Seebeck coefficient, electrical conductivity and electronic thermal conductivity which are recorded in the widely used Ricci database,[17] the CraLiGNN model has the better performance than previously reported CGCNN, GeoCGNN, ALIGNN, and CrabNet models. Furthermore, additional features can be optionally incorporated into the CraLiGNN architecture. When the bandgap is utilized as supplement input, the predictive accuracy of CraLiGNN model is significantly improved.

II. METHODOLOGY

The CraLiGNN model passes and updates messages on both the atomistic bond graph and its line graph. Figure 1 shows the schematic of model architectures.

A. Compositionally restricted atomistic representation

Different from previous GNN models, the atomistic representation utilized in CraLiGNN is based on chemical compositions, including not only the atomic numbers but also the fractions of each chemical element in the compositions. The atomic number is used to match the element representations from mat2vec embedding,[28] which is unsupervised word embedding that captures valuable knowledge of the connections and relationships between property items as retrieved from scientific literature. It is different from the traditional elemental property descriptors, but there exist directions in the mat2vec embedding space that well correlate with elemental properties such as atomic weight, melting temperature, covalent radius, and so on. The element representations in mat2vec embedding have the advantages of being pre-scaled and normalized, as well as the absence of missing elements or element features. The retrieved element representations are used to generate an elemental embedding matrix via a fully connected (FC) layers. Inspired by the positional encoder as performed in CrabNet,[26] we also use sine and cosine functions of various periods to project the fractional amounts into a high-dimensional space, which preserves smooth interpolation between fractional values. The fractional amounts are represented by two fractional encodings. The first part utilizes normalized fractional amounts on a linear scale with a resolution of 0.01, which allows for a straightforward representation of stoichiometric ratios. The second part maps these stoichiometric values to a logarithmic scale, which is beneficial in preserving small fractional amounts. The two parts of the fractional embedding are concatenated to form a fractional embedding matrix that represents the stoichiometric information and has the same size as the element embedding matrix.[26] Then, the elemental and fractional matrices are added together to create the element-derived matrix in a size of $(n_{\text{element}}, d_{\text{model}})$, where d_{model} is the size of the element embedding and n_{element} is the number species of elements. These compositionally restricted representations are translated into atomic features, which depict the node features in the crystal graph.

B. Crystal graph construction

In the traditional representation of GNNs, the structure is modeled as a unit cell graph under periodic boundary conditions. A single node denotes all the translation-equivalent atoms of the crystal, and the multiple edges corresponding to the bonds between atoms include distance information matching their translated lattice position. This periodic graph only considers translation symmetry, as a result, different nodes in a unit cell graph can be the same element, which conflicts with the n_{element}-dimensional element-derived matrix description. To solve this issue, a unit graph considering all symmetries of crystal structures inspired by Ruff *et al.*[27] is adopted. In this representation, all symmetry-equivalent atoms are represented by one element and then merged into a single node in the graph. Therefore, the

16 October 2024 02:04:58

© Author(s) 2024

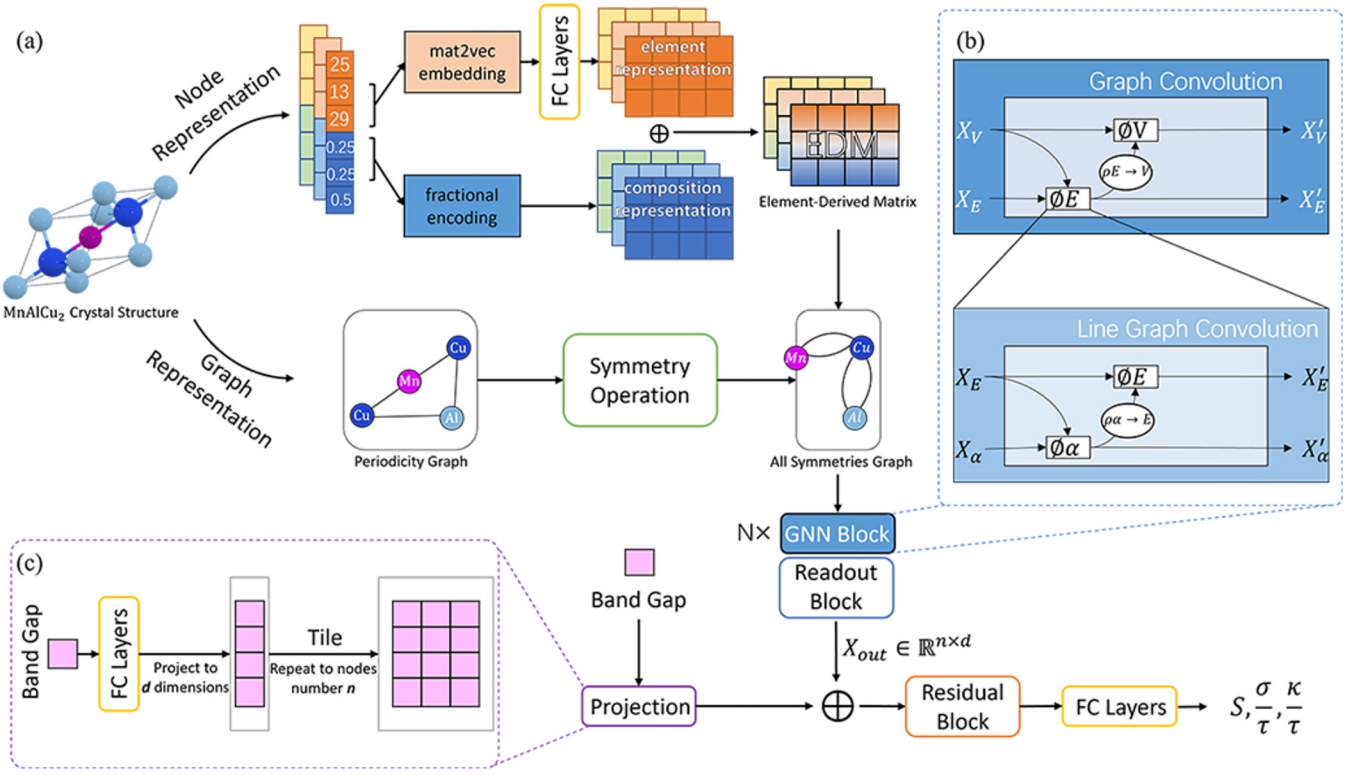


FIG. 1. The schematic architecture of the CraLiGNN model with the supplement of bandgaps. (a) is the overview of property prediction, (b) is the architecture of GNN blocks, and (c) is the projection process of bandgaps.

number of nodes means the number of elements in a crystal. Since the crystal undergoes more symmetric transformations, the reduced number of nodes can substantially decrease the time and memory cost required to train the GNN without any detriment of prediction accuracy.

After representing the crystal into a graph, as shown in Fig. 1(b), the GNN block is used to extract information from the graph and perform edge gated graph convolution message passing updates on atomistic bond graph and corresponding line graph.[22] For the atomistic graph, the node features h_i are converted from element-derived matrix embeddings and the initial edge features e_{ij} are interatomic bond distances operated in a radial basis function expansion with support from 0 to 8 Å for crystals. Therefore, the node representations h^l are updated from the layer l to layer $l+1$ by a edge gated graph convolution according to

$$h_i^{l+1} = h_i^l + \mathrm{SiLU}\left(\mathrm{Norm}\left(W_{src}^l h_i^l + \sum_{j\in N_i} \hat{e}_{ij}^l W_{dst}^l h_j^l\right)\right), \quad (1)$$

where W_{src}^l and W_{dst}^l are the weight matrix of source and destination node features, respectively. SiLU is the sigmoid linear unit activation and Norm is the layer normalization function. $\hat{e}_{ij}^l$ is defined as

$$\hat{e}_{ij}^l = \frac{\sigma(e_{ij}^l)}{\sum_{k\in N_i}\sigma(e_{ik}^l)+\varepsilon}, \quad (2)$$

where σ is the sigmoid function, ε is the noise to prevent existence of zero. The edge messages is updated according to the formula of

$$e_{ij}^l = e_{ij}^{l-1} + \mathrm{SiLU}\left(\mathrm{Norm}\left(W_{edge}^l Z_{ij}^{l-1}\right)\right), \quad (3)$$

where W_{edge}^l is the weight matrix of bond features in conjunction with nodes i and j, and Z_{ij} is the augmented edge features with a function of $Z_{ij} = h_i \oplus h_j \oplus e_{ij}$. In line graph, the node corresponds to an bond in atomistic graph and the edge corresponds to bond pairs. Analogously, the edge-gated graph convolution is implemented on the angle features of line graph, which generates bond messages that are passed to atomistic graph and then updated with the atomic features.

C. Bandgap supplement

The material properties may also be associated to additional information beyond atomic and structure characteristics. Specifically,

16 October 2024 02:04:58

© Author(s) 2024

the significant role of electronic bandgap in electrical transport properties has been confirmed.[29] Therefore, we incorporated the bandgap as an additional feature to augment our CraLiGNN model. For the updated graph representation of GNN block output, an aggregation function is applied to read out the features of global graph by summing the features of all nodes, yielding a output matrix $X_{out} \in \mathbb{R}^{n\times d}$. Here, n represents the count of nodes in a crystal graph, and d represents the embedding dimensions. Since the bandgap needs to be incorporated into the output matrix X_{out}, we performed a projection operation on the bandgap features. As shown in Fig. 1(c), the bandgap is projected to d dimensions using FC layers and then tiled. Thus, this projection converts the bandgap features into a matrix P with dimensions $n \times d$, aligning with X_{out}. Subsequently, we implemented a concatenation operation to combine X_{out} and P, generating a new matrix and passed it through a residual block. Following the residual block, a FC layer is employed to make the final predictions for the material properties.

D. Model training

The Pytorch and Deep Graph Library (DGL) packages[30] are employed to construct the models and build the DataLoaders classes, then train the models relied on Pytorch-ignite package.[31] The foundational hyperparameters remain consistent across different models, for instance, the embedding feature is 64, and the hidden feature is 256. To avoid overfitting issues, the layer normalization and dropout are operated in each GNN block, and robust L1 criterion[32] is adopted for the loss functions, which are robust as they allow the model to learn to attenuate the importance of potentially anomalous training points. For a transport property p, the loss is given by

$$L_p = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{m}\left(\sqrt{2}\left|\hat{y}_{ij}^{p} - y_{ij}^{p}\right|\exp\left(-\ln\hat{s}_{ij}^{p}\right) + \ln\hat{s}_{ij}^{p}\right), \quad (4)$$

where k is the number of samples in the dataset, m is the number of components of the output vector y_i^p. Thus, $\hat{y}_{ij}^{p}$ is the jth component in the i-th predictions, corresponding to the target value y_{ij}^{p}. $\hat{s}_{ij}^{p}$ is the mean of the element-uncertainties giving the estimated aleatoric uncertainty. We choose a batch size of 64 and undertook the training of all models for 120 epochs, which strikes a balance between hardware performance and training expenses. For the hyperparameter optimization, the AdamW optimizer with 10^{-5} weight decay was used. In addition, we conducted an experiment to ascertain the optimal initial learning rate. Upon configuring the learning rate to 10^{-3}, the model training loss manifested as NaN and triggered a gradient explosion. When the learning rate was set to 10^{-5}, the prediction accuracy fell short compared to that of 10^{-4}. Thus, an initial learning rate of 10^{-4} was used and subsequently adjusted based on the one-cycle policy for effective training.

III. RESULTS AND DISCUSSION

The models are trained on the dataset provided by Ricci *et al.*,[17] which is freely available and provides a comprehensive collection of computationally derived electrical transport properties for about 48 000 inorganic compounds with stoichiometric compositions. The crystal and band structures of these materials are calculated using Vienna *ab initio* Simulation Package (VASP)[33] with Perdew–Burke–Ernzerhof (PBE) approximated projected augmented wave (PAW) pseudopotentials. The ZT involved S, σ/τ, and κ_e/τ are calculated using BoltzTraP under constant relaxation time approximation.[34] These aforementioned properties are retrievable at different temperatures, for n- and p-type doping with concentrations from 10^{16} to 10^{20} cm^{-3} at an order of magnitude increment. Although the information of bond lengths and angles are included in the graph representations, which can capture the anisotropy of geometry to some extent, we used the mean values of the diagonal elements of S, σ and κ_e tensors and thus the

16 October 2024 02:04:58

TABLE I. Regression performances on the dataset of electrical transport properties under 600 K, 900 K, and 10^{19} cm^{-3} carrier concentration using CrabNet, CGCNN, GeoCGNN, ALIGNN, and CraLiGNN models, in terms of MAE. Boldface denotes the minimum value among the results of these models.

	Temperature	Doping	CrabNet	CGCNN	GeoCGNN	ALIGNN	CraLiGNN
	600 K	n	86.831	104.920	96.606	82.383	**74.645**
S		p	87.846	108.300	102.38	82.975	**79.790**
	900 K	n	84.818	95.088	101.675	83.042	**74.685**
		p	93.878	98.014	109.782	86.891	**81.209**
	Average time (s)		3402	15 100	21 894	24 126	15 772
	600 K	n	0.581	0.564	0.544	0.484	**0.454**
$\log(\sigma/\tau)$		p	0.598	0.559	0.545	0.483	**0.460**
	900 K	n	0.597	0.524	0.532	0.477	**0.429**
		p	0.617	0.534	0.552	0.491	**0.438**
	Average time (s)		3188	12 733	24 661	23 997	15 867
	600 K	n	0.621	0.617	0.582	0.534	**0.476**
$\log(\kappa_e/\tau)$		p	0.654	0.630	0.601	0.548	**0.504**
	900 K	n	0.604	0.554	0.570	0.502	**0.449**
		p	0.613	0.551	0.571	0.507	**0.449**
	Average time (s)		3241	14 677	22 866	23 870	16 184

© Author(s) 2024

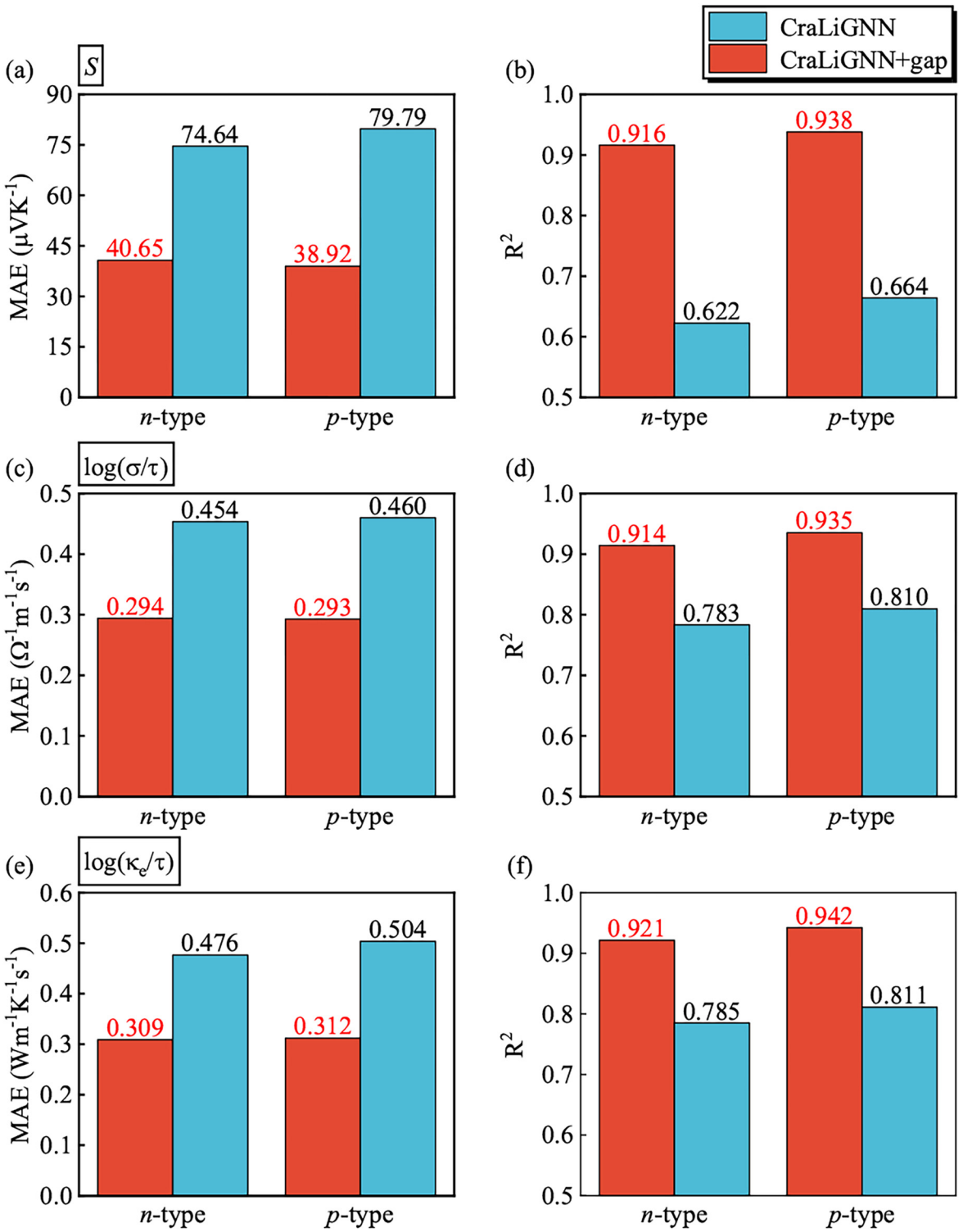


FIG. 2. Regression performance for *n*- and *p*-type (a) and (b) *S*, (c) and (d) $\log(\sigma/\tau)$, and (e) and (f) $\log(\kappa_e/\tau)$ using CraLiGNN with and without the supplement of bandgaps for datasets at 600 K and 10^{19} cm^{-3} concentration, in terms of MAE and R^2.

16 October 2024 02:04:58

© Author(s) 2024

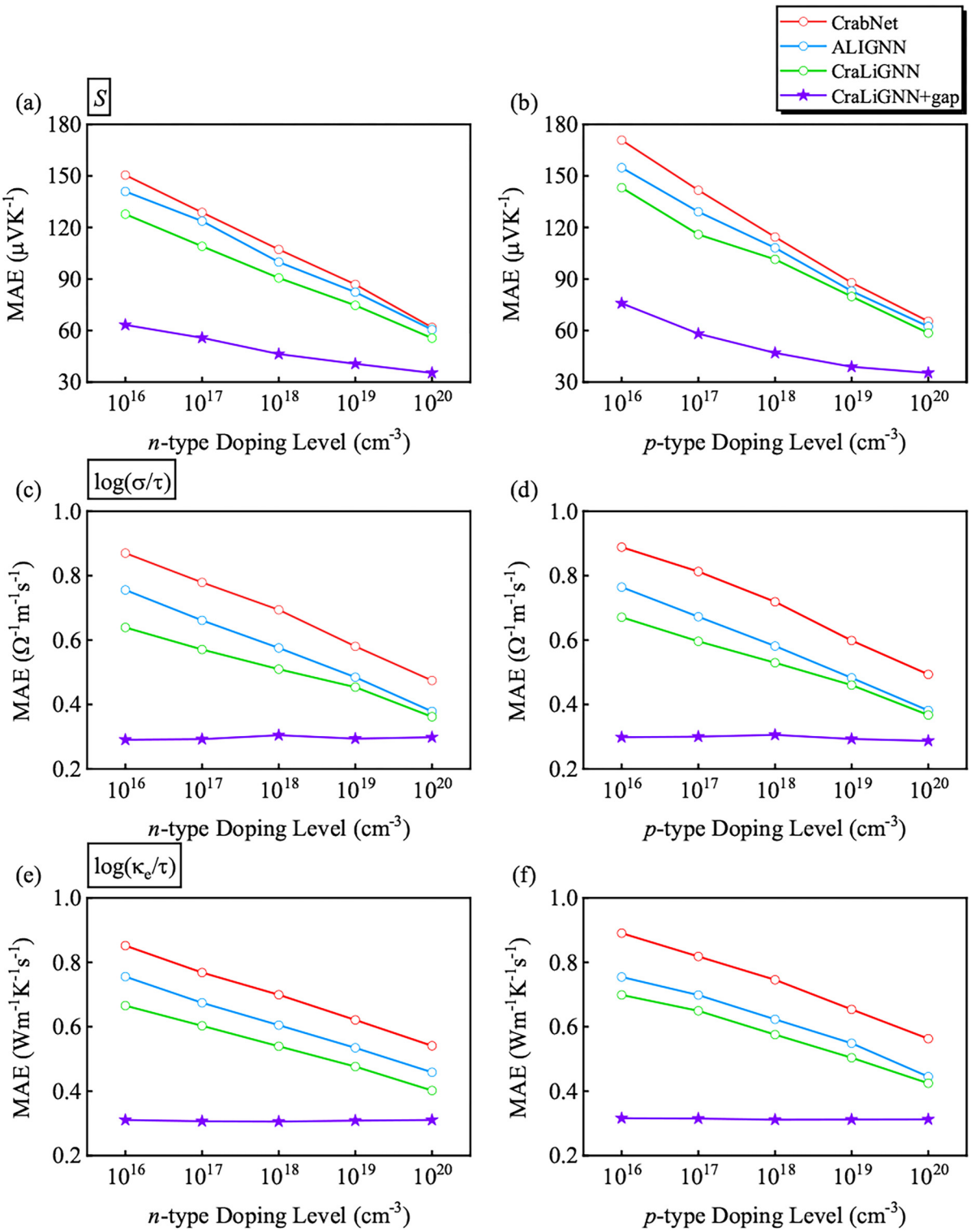


FIG. 3. The MAE of CrabNet, ALIGNN, CraLiGNN, and CraLiGNN+gap models for 600 K and different doping levels.

16 October 2024 02:04:58

© Author(s) 2024

anisotropy was not considered. The data at representative temperatures of 600 and 900 K are chosen for calculation. Since the values of σ/τ and κ_e/τ differ by orders of magnitude, as a usual course, their logarithms are used as the targets of the models. The dataset is randomly splitted by 80% for training, 10% for validation, and 10% for testing. The mean absolute error (MAE) metric evaluates the discrepancy between the model predictions and the true values. The baseline for the better performance of models is determined to have a smaller MAE value.

The regression performances of CraLiGNN are compared with previous GNN models of CrabNet, CGCNN, GeoCGNN, and ALIGNN. Table I comprehensively lists the MAE results of various models for *n*- and *p*-type *S*, $\log(\sigma/\tau)$ and $\log(\kappa_e/\tau)$ under temperatures of 600, 900 K, with a carrier concentration of 10^{19} cm^{-3}. It can be seen that the GeoCGNN and ALIGNN have better predictive accuracy than CGCNN. This is in line with what is expected for them on prediction due to their improved descriptions in the crystal structure. More noteworthy is that the CraLiGNN outperforms previous models. At 600 K, the MAE of CraLiGNN for *n*-type *S*, $\log(\sigma/\tau)$ and $\log(\kappa_e/\tau)$ are 29%, 20%, and 23% smaller than that of CGCNN, while the reductions for *p*-type properties are 26%, 18%, and 20%, respectively. Even compared to ALIGNN, the CraLiGNN still exhibits significant improvements, achieving 9%, 6%, 11% enhancements for *n*-type properties and 4%, 5%, 8% for *p*-type properties, respectively. These reveal the important role of compositionally restricted atomic representations. Further comparison with CrabNet shows that the CraLiGNN model has performance enhancements by 14%, 22%, 23% for *n*-type properties, and 9%, 23%, 23% for *p*-type properties, respectively, indicating the importance of geometric information. Notably, these trends persist at 900 K, underscoring the consistent performance gains offered by the CraLiGNN model across varying temperature conditions.

In terms of the training time on NVIDIA RTX A5000, as shown in Table I, the CrabNet model has significantly less computational cost than these GNN models, which is attributed to the avoidance of graph convolution for structure information. As compared with the CGCNN model, the GeoCGNN and ALIGNN models incorporate more geometry-topological information and angle-based information, respectively, as a result, their average computational time increase by 45%–94% and 60%–88% in the cases listed in Table I. By contrast, the CraLiGNN model, considering more detailed atomistic and crystal representations, has the best performance but with only 4%–25% increase in computational time as compared with the CGCNN model. More importantly, compared with ALIGNN model, the computational time reduces about 30%. It can be seen that the CraLiGNN model gives improved prediction but without the computational cost increased significantly.

To further improve the predictive capability of the CraLiGNN model, the bandgap is incorporated as a supplementary descriptor.[29] Figure 2 shows the MAE and R^2 results for *S*, $\log(\sigma/\tau)$ and $\log(\kappa_e/\tau)$ using CraLiGNN with and without band gap supplement. The MAE of *S* decreases from 74.64 (79.79) to 40.65 (38.92) for *n*-type (*p*-type) doping, by about 46% (51%) reduction, respectively. The corresponding R^2 of *S* raises from 0.622 (0.664) to 0.916 (0.938), by an increase of 47% (41%), respectively. The enhancement caused by bandgap also manifests in $\log(\sigma/\tau)$ and $\log(\kappa_e/\tau)$, with the MAE reduced by more than 35% (35%) and R^2 increased by more than 15% (16%) for *n*-type (*p*-type) doping, respectively. The results confirm that our model incorporating the bandgap gains significant enhancement in the predictive accuracy of TE properties.

The electrical transport properties are highly dependent on carrier concentrations. Therefore, we examine the performance of

16 October 2024 02:04:58

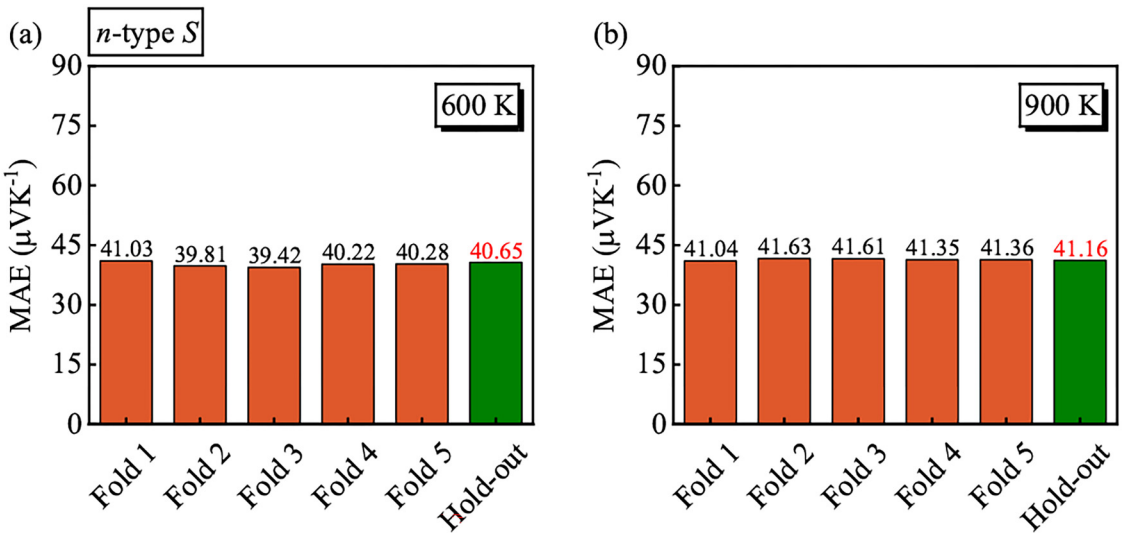


FIG. 4. The MAE analysis of five-fold cross-validation and hold-out methods for *n*-type *S* using CraLiGNN+gap model under representative temperatures of (a) 600 K and (b) 900 K with 10^{19} cm^{-3} carrier concentration.

© Author(s) 2024

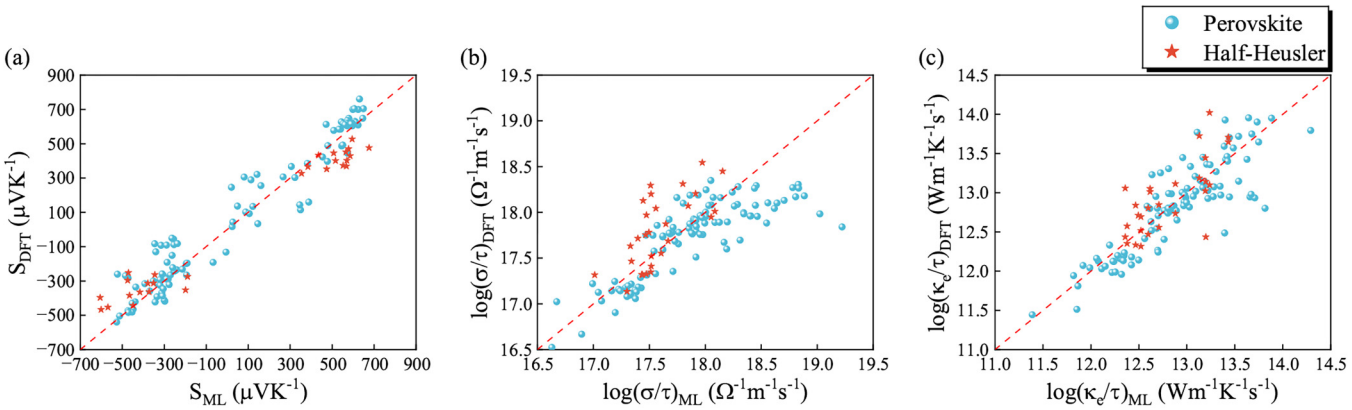


FIG. 5. Prediction of CraLiGNN supplemented with bandgap for 14 half-Heusler and 52 perovskite compounds for (a) S, (b) $\log(\sigma/\tau)$ and (c) $\log(\kappa_e/\tau)$ vs the corresponding DFT calculated values.

the CraLiGNN model on different doping levels, as shown in Fig. 3. It is evident that the CraLiGNN model remains superior performance compared with CrabNet and ALIGNN models for S, $\log(\sigma/\tau)$, and $\log(\kappa_e/\tau)$. This finding underscores the efficacy of the CraLiGNN model in accurately predicting and surpassing the performance of other models in these key TE properties. It is demonstrated that the CraLiGNN model showcases its ability to extract a wealth of information from materials, enabling a more comprehensive understanding of their TE behaviors. Moreover, when the CraLiGNN model is supplemented by the bandgap, the MAE is almost halved at a concentration of 10^{16} cm^{-3}. As concentration increases, the enhancement of prediction performance reduces, but still more than 36%, 18%, and 23% reduction in MAE at 10^{20} cm^{-3} for S, $\log(\sigma/\tau)$, and $\log(\kappa_e/\tau)$, respectively. The CraLiGNN with bandgaps expresses the best predictive performance in TE properties and can be considered for the application to screen potential TE materials. These results are not suffered from overfitting issue. As an illustrative example, Fig. 4 presents a fivefold cross-validation for the n-type S under conditions of 600 and 900 K with a carrier concentration of 10^{19} cm^{-3}. It can be observed that the results for each fold are excellently consistent with the findings derived from our selected hold-out split method, indicating the reliability and robustness of our analyses and conclusions.

One important purpose is to use the model for quick identification of the thermoelectric potential of materials. Here, the thermoelectric transport properties of 14 half-Heusler compounds (XYZ type in $F\bar{4}3m$ space group) and 52 perovskites (ABX3 type in $Pm\bar{3}m$ space group), which are documented in the Materials Project but not in the thermometric training database, were predicted using CraLiGNN and calculated using BoltzTraP.[34] The predictions vs DFT calculations of S, $\log(\sigma/\tau)$ and $\log(\kappa_e/\tau)$ are shown in Fig. 5. The R^2 reaches as high as 0.936 for S, while they are 0.444 and 0.628 for $\log(\sigma/\tau)$ and $\log(\kappa_e/\tau)$, respectively. The better prediction for S is not surprising because S is dominated by band structures while σ and κ_e are also significantly related to the complex electron and phonon properties as well as their coupling. Although R^2 is not high, the relatively high and low properties can still be fairly identified for both Half-Heusler and perovskite structures. This implies that our model has ability to generalize well across diverse structural datasets. Trained on a wide range of structures, our model demonstrates commendable predictive capabilities for a specified structure. This achievement highlights the robustness and transferability of our model's learned representations.

IV. CONCLUSIONS

In conclusion, a CraLiGNN model was developed for the prediction of thermoelectric transport properties of crystal materials. As compared to previous GNN models, CraLiGNN enhanced the description of atomistic representations, and the cell graph was streamlined with all symmetries. The two improvements increase the prediction performance of models but do not increase the computational cost. The supplement with bandgaps could significantly enhance the prediction accuracy, by a reduction of more than 36%, 18%, and 23% in the MAE of S, $\log(\sigma/\tau)$, and $\log(\kappa_e/\tau)$ for the datasets at 600 K within the concentration from 10^{16} to 10^{20} cm^{-3}. In addition, the thermoelectric transport properties of 14 half-Heusler and 52 perovskite compounds whose properties are not recorded were calculated using first principles. The comparison between the calculated values and the predicted values of the CraLiGNN model suggests the powerful prediction ability of the model in the thermoelectric materials screening scenario.

ACKNOWLEDGMENTS

The authors acknowledge support from Hubei Provincial Natural Science Foundation of China (No. 2022CFB104) and National Key Research and Development Program of China (No. 2022YFC2204400).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

16 October 2024 02:04:58

© Author(s) 2024

Author Contributions

Zeyu Wang: Investigation (equal); Methodology (equal); Writing – original draft (equal). **Run Hu:** Conceptualization (equal). **Xiaobing Luo:** Resources (equal); Supervision (equal). **Jinlong Ma:** Conceptualization (equal); Investigation (equal); Supervision (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The training data that support this study are openly available. The thermoelectric transport property data from Ricci database[35] are publicly available online at https://datadryad.org/stash/dataset/doi:10.5061/dryad.gn001. The structure and bandgap data from Materials Project[36] are publicly available online at https://next-gen.materialsproject.org/. The data are available directly in Zenodo[37]: https://doi.org/10.5281/zenodo.11057422.

REFERENCES

[1]Q. Yan and M. G. Kanatzidis, Nat. Mater. **21**, 503 (2022).
[2]J. Mao, G. Chen, and Z. Ren, Nat. Mater. **20**, 454 (2021).
[3]J. He and T. M. Tritt, Science **357**, eaak9997 (2017).
[4]T. Zhu, Y. Liu, C. Fu, J. P. Heremans, J. G. Snyder, and X. Zhao, Adv. Mater. **29**, 1605884 (2017).
[5]D. Liu, D. Wang, T. Hong, Z. Wang, Y. Wang, Y. Qin, L. Su, T. Yang, X. Gao, Z. Ge, B. Qin, and L.-D. Zhao, Science **380**, 841 (2023).
[6]D. Beretta, N. Neophytou, J. M. Hodges, M. G. Kanatzidis, D. Narducci, M. Martin- Gonzalez, M. Beekman, B. Balke, G. Cerretti, W. Tremel, A. Zevalkink, A. I. Hofmann, C. Müller, B. Dörling, M. Campoy-Quiles, and M. Caironi, Mater. Sci. Eng. R. Rep. **138**, 100501 (2019).
[7]S. Curtarolo, W. Setyawan, G. L. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko, and D. Morgan, Comput. Mater. Sci. **58**, 218 (2012).
[8]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 Mater. **1**, 011002 (2013).
[9]S. Poncé, W. Li, S. Reichardt, and F. Giustino, "First-principles calculations of charge carrier mobility and conductivity in bulk semiconductors and two-dimensional materials," Rep. Prog. Phys. **83**, 036501 (2020).
[10]W. Li, J. Carrete, N. A. Katcho, and N. Mingo, Comput. Phys. Commun. **185**, 1747 (2014).
[11]J. Ma, A. S. Nissimagoudar, and W. Li, Phys. Rev. B **97**, 045201 (2018).
[12]Z. Han, X. Yang, W. Li, T. Feng, and X. Ruan, Comput. Phys. Commun. **270**, 108179 (2022).
[13]S. Wang, Z. Wang, W. Setyawan, N. Mingo, and S. Curtarolo, "Assessing the thermoelectric properties of sintered compounds via high-throughput *ab-initio* calculations," Phys. Rev. X **1**, 021012 (2011).
[14]P. Gorai, V. Stevanović, and E. S. Toberer, Nat. Rev. Mater. **2**, 17053 (2017).
[15]J. Recatala-Gomez, A. Suwardi, I. Nandhakumar, A. Abutaha, and K. Hippalgaonkar, ACS Appl. Energy Mater. **3**, 2240 (2020).
[16]K. Choudhary, K. F. Garrity, A. C. E. Reid, B. DeCost, A. J. Biacchi, A. R. Hight Walker, Z. Trautt, J. Hattrick-Simpers, A. G. Kusne, A. Centrone, A. Davydov, J. Jiang, R. Pachter, G. Cheon, E. Reed, A. Agrawal, X. Qian, V. Sharma, H. Zhuang, S. V. Kalinin, B. G. Sumpter, G. Pilania, P. Acar, S. Mandal, K. Haule, D. Vanderbilt, K. Rabe, and F. Tavazza, npj Comput. Mater. **6**, 173 (2020).
[17]F. Ricci, W. Chen, U. Aydemir, G. J. Snyder, G.-M. Rignanese, A. Jain, and G. Hautier, Sci. Data **4**, 170085 (2017).
[18]T. Wang, C. Zhang, H. Snoussi, and G. Zhang, Adv. Funct. Mater. **30**, 1906041 (2020).
[19]X. Qian and R. Yang, Mater. Sci. Eng. R. Rep. **146**, 100642 (2021).
[20]T. Xie and J. C. Grossman, Phys. Rev. Lett. **120**, 145301 (2018).
[21]J. Cheng, C. Zhang, and L. Dong, Commun. Mater. **2**, 92 (2021).
[22]K. Choudhary and B. DeCost, npj Comput. Mater. **7**, 185 (2021).
[23]Z. Wang, J. Ma, R. Hu, and X. Luo, Appl. Phys. Lett. **122**, 152106 (2023).
[24]X.-G. Li, B. Blaiszik, M. E. Schwarting, R. Jacobs, A. Scourtas, K. J. Schmidt, P. M. Voyles, and D. Morgan, J. Chem. Phys. **155**, 154702 (2021).
[25]H. Huang and A. Barati Farimani, J. Appl. Phys. **135**, 165104 (2024).
[26]A. Y.-T. Wang, S. K. Kauwe, R. J. Murdock, and T. D. Sparks, npj Comput. Mater. **7**, 77 (2021).
[27]R. Ruff, P. Reiser, J. Stühmer, and P. Friederich, Digital Discovery **3**, 594 (2023).
[28]V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, and A. Jain, Nature **571**, 95 (2019).
[29]L. M. Antunes, K. T. Butler, and R. Grau-Crespo, "Predicting thermoelectric transport properties from composition with attention-based deep learning," Mach. Learn.: Sci. Technol. **4**, 015037 (2023).
[30]C. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, and K. F. Jensen, Chem. Sci. **10**, 370 (2019).
[31]V. Fomin, J. Anmol, S. Desroziers, J. Kriss, and A. Tejani, See https://github.com/pytorch/ignite for "High-level library to help with training neural networks in pytorch" (2020).
[32]R. E. A. Goodall and A. A. Lee, Nat. Commun. **11**, 6280 (2020).
[33]G. Kresse and J. Furthmuller, Comput. Mater. Sci. **6**, 15 (1996).
[34]G. K. Madsen and D. J. Singh, Comput. Phys. Commun. **175**, 67 (2006).
[35]F. Ricci *et al.* (2018), "Data from: An ab initio electronic transport database for inorganic materials," Dryad, Dataset. https://doi.org/10.5061/dryad.gn001
[36]See https://next-gen.materialsproject.org/ for the structure and bandgap data from the Materials Project, version v2023.11.1.
[37]Z. Wang, "The dataset of CraLiGNN," Zenodo, https://doi.org/10.5281/zenodo.11057422

16 October 2024 02:04:58

© Author(s) 2024