Applied Physics Reviews ARTICLE pubs.aip.org/aip/are

High-throughput thermoelectric materials screening by deep convolutional neural network with fused orbital field matrix and composition descriptors

Cite as: Appl. Phys. Rev. **11**, 021402 (2024); doi: 10.1063/5.0187855
Submitted: 16 November 2023 · Accepted: 14 March 2024 ·
Published Online: 1 April 2024

Mohammed Al-Fahdi,[1] Kunpeng Yuan,[2] Yagang Yao,[3] Riccardo Rurali,[4] and Ming Hu[1,a)]

AFFILIATIONS

[1]Department of Mechanical Engineering, University of South Carolina, Columbia, South Carolina 29208, USA
[2]College of New Energy, China University of Petroleum (East China), Qingdao 266580, China
[3]National Laboratory of Solid State Microstructures, College of Engineering and Applied Sciences, Jiangsu Key Laboratory of Artificial Functional Materials, and Collaborative Innovation Center of Advanced Microstructures, Nanjing University, Nanjing 210093, China
[4]Institut de Ciència de Materials de Barcelona, ICMAB–CSIC, Campus UAB, 08193 Bellaterra, Spain

[a)]**Author to whom all correspondence should be addressed:** hu@sc.edu

ABSTRACT

Thermoelectric materials harvest waste heat and convert it into reusable electricity. Thermoelectrics are also widely used in inverse ways such as refrigerators and cooling electronics. However, most popular and known thermoelectric materials to date were proposed and found by intuition, mostly through experiments. Unfortunately, it is extremely time and resource consuming to synthesize and measure the thermoelectric properties through trial-and-error experiments. Here, we develop a convolutional neural network (CNN) classification model that utilizes the fused orbital field matrix and composition descriptors to screen a large pool of materials to discover new thermoelectric candidates with power factor higher than 10 μW/cm K^2. The model used our own data generated by high-throughput density functional theory calculations coupled with *ab initio* scattering and transport package to obtain electronic transport properties without assuming constant relaxation time of electrons, which ensures more reliable electronic transport properties calculations than previous studies. The classification model was also compared to some traditional machine learning algorithms such as gradient boosting and random forest. We deployed the classification model on 3465 cubic dynamically stable structures with non-zero bandgap screened from Open Quantum Materials Database. We identified many high-performance thermoelectric materials with $ZT > 1$ or close to 1 across a wide temperature range from 300 to 700 K and for both *n*- and *p*-type doping with different doping concentrations. Moreover, our feature importance and maximal information coefficient analysis demonstrates two previously unreported material descriptors, namely, mean melting temperature and low average deviation of electronegativity, that are strongly correlated with power factor and thus provide a new route for quickly screening potential thermoelectrics with high success rate. Our deep CNN model with fused orbital field matrix and composition descriptors is very promising for screening high power factor thermoelectrics from large-scale hypothetical structures.

I. INTRODUCTION

Waste heat energy amounts to roughly 70% of the used energy in transportation and industrial operations.[1] Traditional approaches for converting waste heat into reusable energy are usually done by Rankine steam engines. However, such approaches can involve heavy machinery that constitute various complex moving parts that need constant maintenance.[2] Thermoelectric (TE) generators are solid-state devices that can directly convert waste heat into electricity, the highest form of reusable energy, and require no moving parts and additional complex machinery, making them a convenient and practical alternative approach for waste heat energy conversion applications.[2] A thermoelectric generator consists of two semiconducting materials with *n*- and *p*-type electronic conductivity, respectively. Those semiconducting materials are assembled and affect each other thermally and electrically through a heat source with high temperature (T_{hot}) and heat sink with low temperature (T_{cold}). The thermoelectric generator efficiency to a

large extent depends on the temperature difference between heat source and heat sink. More importantly, the efficiency also depends on the material properties of the semiconductors with *n*-type and *p*-type conductivities, which can be quantified by the dimensionless number, namely, the *figure of merit* (ZT) $ZT = \frac{S^2\sigma T}{\kappa_{tot}}$, where σ is the electrical conductivity, S is the Seebeck coefficient, T is the absolute temperature, and κ_{tot} is the total thermal conductivity, which is the sum of electronic thermal conductivity (κ_{el}) from heat diffusion carried by moving electrons and lattice thermal conductivity due to the lattice vibrations (phonons, κ_{ph}). The higher the dimensionless ZT in a material at a particular doping concentration, the more efficient the thermoelectric generators can become. The numerator ($S^2\sigma$) in ZT is referred to the power factor (PF). Therefore, a TE material with a high ZT demands high PF (i.e., large absolute S and σ) and low thermal conductivity, which is a difficult task due to coupling and competition between the properties in the ZT definition. Other aspects such as toxicity and abundance can further complicate the search for commercially efficient thermoelectric materials. Although the thermoelectricity phenomenon had been known since the early 1800s,[3,4] only a few materials with moderate ZT values have been discovered such as PbTe and Bi_2Te_3, but the toxicity of most popular thermoelectric materials still limits their widespread use.[5] A positive impact is expected to happen to the environment if thermoelectric generators become more ubiquitous, which also requires extensive research of new and highly efficient thermoelectric materials,[6] an indeed active and prominent field of research.[7,8] A few promising thermoelectric materials have been synthesized experimentally either by accident or chemical intuition. At room temperature or low temperature range in which thermoelectric materials are often used for low quality but large amount of heat recovery or cooling applications, top-performing thermoelectric materials are Bi_2Te_3-based alloys.[9] For example, $(Bi_{1-x}Sb_x)_2Te_3$ has $ZT \sim 1.2$ and $PF \sim 45\ \mu W/cm{\cdot}K^2$ at room temperature.[9] Materials similar to PbTe tend to have their best performance at higher temperatures between 500 and 900 K.[10] For example, $Pb_{1-x}Sr_xTe$ has a ZT equal to 2.5 and PF of $30\ \mu W/cm{\cdot}K^2$ at a temperature of approximately 800 K.[10] Other compound families that attracted substantial attention are Zintl phase compounds such as $YbZn_2Sb_2$,[11,12] Heusler and half-Heusler compounds such as ZrNiSn and TiNiSn,[13,14] metal chalcogenides such as Cu_2Se and SnSe,[15,16] clathrates such as $Sr_8Ga_{16}Ge_{30}$,[17] skutterudites such as $CoSb_3$ and $RhSb_3$,[18] or metal oxides such as $Ca_3Co_4O_9$ and $NaCo_2O_4$.[19,20] The highest ever experimentally proven value of ZT, to the best of our knowledge, is 3.1, which occurs in hole doped polycrystalline SnSe at 783 K.[21] The hope to obtain even higher ZT begs the need for faster and cheaper methods to search for candidate materials in the huge space of material compositions and structures. Obviously, traditional experimental means of trial and error are out of question since they are extremely time and resource consuming.

Machine learning (ML) methods have recently been used to discover novel materials with various desirable material properties such as mechanical,[22,23] magnetic,[24] thermal,[25,26] and optical properties.[27] Machine learning methods were in fact also utilized to calculate, screen, and/or predict thermoelectric properties using datasets such as by Ricci *et al.*,[28] and materials informatics platform (MIP) by Yao *et al.*[29] Choudhary *et al.* used his classification models to classify PF and S using Jarvis database materials at two doping concentrations of 10^{20} for both *n*- and *p*-type doping and at a temperature of 600 K.[30] Antunes *et al.* used an attention-based deep learning regression model to predict PF, S, and σ.[31] Sheng *et al.* used active learning to predict PF of diamond-like thermoelectric materials.[32] However, in lots of previous studies, transport properties were obtained through constant relaxation time approximation (CRTA) for electrons.[28,30,31] Despite some success of such rude approximation for predicting electronic transport properties and finding some promising thermoelectric materials, the low-resolution data produced by usage of CRTA naturally bring considerably large uncertainty to the quality of the trained ML models and the subsequent screening and prediction. To alleviate this problem, in this work we utilize high-resolution high-throughput density functional theory (DFT) calculations for electronic transport properties by the *ab initio* scattering and transport (AMSET) package, which computes electron lifetimes considering various scattering mechanisms, including polar and nonpolar electron–phonon coupling, scattering from ionized impurity, and piezoelectric scattering mechanisms. With such new DFT data, we trained a classification model, which fuses residual-like convolutional neural networks (CNNs) from both Magpie and hybrid orbital field matrix to screen materials at different temperatures and various doping concentrations. The trained classification model was then used for screening a large number of cubic structures taken from existing materials database and promising thermoelectric candidates were identified and further validated by DFT calculations.

II. DFT CALCULATION, WORKFLOW, AND ML MODEL TRAINING

All our cubic structures for training and screening were downloaded from the Open Quantum Material Database (OQMD).[33,34] We first screened the structures by the bandgap values provided by the OQMD itself and discarded the metallic structures with zero bandgap. We then re-optimized all structures by the density functional theory (DFT) calculations with the projector augmented wave (PAW) method as implemented in Vienna *Ab Initio* Simulation (VASP) software.[35–37] The exchange-correlation functional implemented in this work is Perdew–Burke–Ernzer (PBE) generalized gradient approximation (GGA).[37] A plane wave energy cutoff of 520 eV is used in the DFT calculations. For global structure optimization, a criterion of 10^{-8} eV and 10^{-6} eV/Å for energy and atomic force convergence, respectively, was applied. The Brillouin zone was sampled using the Monkhorst–Pack k-mesh depending on the lattice constants, with the product of number of k-points in a crystalline direction and the corresponding lattice constant larger than 60 Å to guarantee high-quality DFT calculations. After structure optimization, we continue to screen the structures by dynamic stability via our recently developed Elemental Spatial Density Neural Network Force Field (Elemental-SDNNFF).[38,39] The structures that go to our next step of thermoelectric property screening are predicted to be dynamically stable, i.e., no imaginary phonon frequencies are present in the full Brillouin zone. This way the prediction on thermoelectric performance of new structures will be physically meaningful. Finally, to prepare the training data for our classification model for thermoelectric performance prediction, the electronic transport properties were generated through AMSET package coupled with full DFT calculations.[40,41] AMSET implemented the acoustic deformation potential (ADP), ionized impurity scattering (IMP), and polar optical phonon scattering (POP) scattering types to obtain high-fidelity data without unreasonably assuming a constant relaxation time for electrons. These scattering processes in AMSET require computing

the following parameters: the electron charge wavefunction coefficients, high-frequency and static dielectric constants, elastic constants, polar optical phonon frequency, and deformation potentials, all of which were obtained by direct DFT calculations. The same k-points resolution was used for all DFT calculations to obtain these parameters. Lattice thermal conductivity (LTC) of selected materials was obtained by iteratively solving phonon Boltzmann transport equation (BTE) using ShengBTE package.[42] The second and third interatomic force constants (IFCs) required for phonon BTE run were computed by randomly displacing atoms in supercells by 0.03 Å (30 random configurations for each structure) and subsequently evaluating the atomic forces by high-precision self-consistent DFT calculations, followed by IFC fitting by compressive sensing lattice dynamics method.[43] Due to the high computational cost of LTC calculations by DFT, we only performed full DFT calculations for the LTC of a limited number of materials.

Figure 1 illustrates our entire workflow for accelerated search for thermoelectric materials. The first step is to screen OQMD database by extracting the structures that have non-zero bandgap (E_g). Due to the large number of hypothetical structures in the OQMD database, we only focus on cubic structures in this study. Expanding current workflow to full material space search is undergoing. After re-optimizing these structures by DFT with the aforementioned parameters, we screen dynamically stable structures by our recently developed Elemental-SDNNFF model.[38,39] We finally identified 5292 cubic structures that are predicted to have no negative frequencies in the Brillouin zone. This will be the pool for us to screen potential thermoelectric materials in this work. High-throughput AMSET + DFT calculations were performed to generate the classification dataset. The selected calculation cases are the temperatures of 300, 500, and 700 K, the *n*- and *p*-type doping concentrations of $1 \times 10^{19}\,\mathrm{cm}^{-3}$, $5 \times 10^{19}\,\mathrm{cm}^{-3}$, $1 \times 10^{20}\,\mathrm{cm}^{-3}$, and $5 \times 10^{20}\,\mathrm{cm}^{-3}$. The classification dataset was created based on the threshold value for PF of $10\,\mu\mathrm{W/cm\ K}^2$ (i.e., $\mathrm{PF} > 10\,\mu\mathrm{W/cm\ K}^2$ corresponds to target = 1; otherwise, target = 0) for all temperatures and *p*-type and *n*-type doping concentrations of $1 \times 10^{20}\,\mathrm{cm}^{-3}$. The classification dataset was then used to train the ML model with fused CNNs from orbital field matrix (OFM) and Magpie descriptors. Separate models were trained in parallel for each individual temperature and specific doping concentration. After training, the models were used to classify untested materials from the above pool to screen materials with high PF at different temperatures and doping concentrations. AMSET + DFT calculations were performed on selected potential thermoelectric materials to confirm whether such predicted materials truly have high PF. Finally, high precision but more expensive DFT calculations were performed to calculate LTC to further confirm possible high ZT in those materials.

III. RESULTS AND DISCUSSION

The Magpie descriptors[44] used for our composition descriptors are obtained using Matminer.[45] The 132-feature vector in Magpie is reshaped into a 12×11 matrix to be utilized as input for the 2D CNN. The descriptors for the OFM[46] are also obtained using Matminer.[45] The OFM descriptors have the 32×32 shape matrix (i.e., 2D), which allows it to be used as an input for a 2D CNN as well. Since the model uses two types of descriptors, each descriptor goes through a separate path of 2D CNN with various numbers of channels and pooling layers to extract their features. The number of channels of the 2D CNN for OFM features path is (1) 32, (2a) 32, (2b) 32, (3) 64, and (4) 64 (2D CNN or conv2D numbers are shown in Fig. 2). The number of channels in the 2D CNN layers for Magpie descriptors path is as follows:

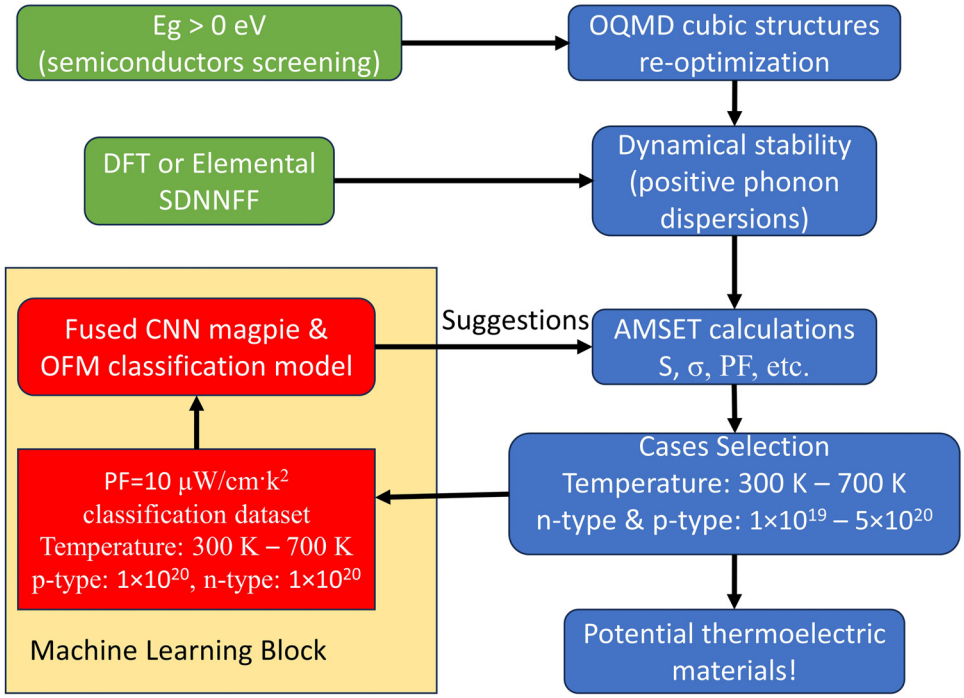

FIG. 1. Flow chart for the accelerated search for thermoelectric materials from OQMD database by combining machine learning classification model and high-throughput DFT calculations on thermoelectric properties with AMSET package.

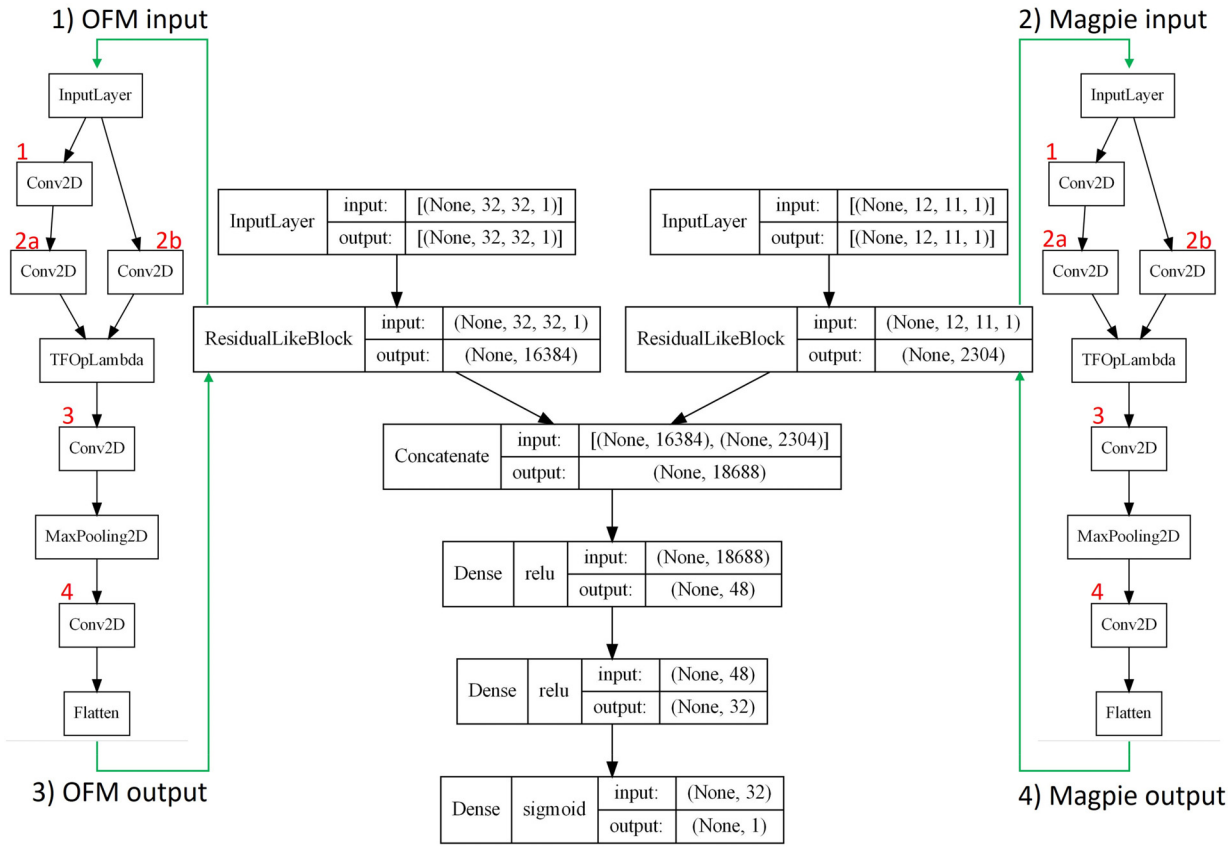

FIG. 2. Schematic of classification model architecture. The input matrix for orbital field matrix (OFM) is 32×32, and OFM output features are 16 384. The Magpie input matrix is 12×11, and Magpie output features are 2304. Both OFM and Magpie output features are then concatenated (i.e., 18 688 features in total) and input into two dense layers before classifying. The red numbers indicate the CNN layer number in each descriptor path (i.e., OFM and Magpie paths). More detailed OFM and Magpie layers are available in supplementary material.

(1) 32, (2a) 48, (2b) 48, (3) 64, and (4) 64 (2D CNN or conv2D numbers are also shown in Fig. 2). The residual-like 2D CNN or conv2D layer in (2b) is added to get the same shape of the 2D CNN or conv2D layer in 2a because implementing residual deep learning tends to improve training and feature extraction, which then yields better accuracy and lower error as demonstrated in Ref. 47. The maximum pooling layers in both OFM and Magpie descriptors have 2×2 kernels. Then, both features are concatenated and input into two dense layers with 48 and 32 neurons, respectively, and being used for predicting classes (i.e., prediction of "1" stands for high PF of the predicted structure, while "0" stands for low PF).

In total, 1438 *p*-type and 1499 *n*-type cubic structures were separately calculated by full DFT + AMSET, and this is the data used in our fused OFM and Magpie CNN model training. The training, validation, and test splits for all classification models are 80%, 10%, and 10%, respectively. Our developed fused OFM and Magpie CNN model was also compared to traditional ML algorithms such as random forest[48] and gradient boosting[49] adopted from scikit-learn or sklearn library[50] in python. The Magpie composition descriptors were utilized to train those models. Our model was developed using the Tensorflow package.[51] The metrics comparison was based on the accuracy score, which is defined as

$$\text{accuracy score} = \frac{TP + TN}{TP + TN + FP + FN},$$

where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative, respectively. Tables I and II show the accuracy scores of all trained models for all temperatures and doping concentrations in *p*- and *n*-type doping, respectively. Overall, all the models including traditional ML models have high accuracy, and the best accuracy is highlighted in bold. It seems that our developed model had a little higher accuracy compared to random forest and gradient boosting at some doping concentrations and temperatures. Therefore, we finally use our new fused OFM and Magpie CNN model to screen untested materials from OQMD database.

The loss function curves for training and validation are shown in Fig. 3. The binary cross-entropy loss function is used for our developed fused OFM and Magpie CNN model, which calculated the difference or dissimilarity between the predicted probability of the data points and actual labels of those data points. Binary cross-entropy is an ideal loss function for classifying two classes, which makes it an ideal loss function for our model since our model predicts two classes (i.e., 1 for high PF and 0 for low PF). The binary cross-entropy loss function is also selected because the sigmoid activation function in the last layer

TABLE I. Accuracy scores for *p*-type doping PF models. The best accuracy from each temperature and doping concentration is highlighted in bold.

Temperature	300 K				500 K				700 K			
Doping (cm^{-3})	1×10^{19}	5×10^{19}	1×10^{20}	5×10^{20}	1×10^{19}	5×10^{19}	1×10^{20}	5×10^{20}	1×10^{19}	5×10^{19}	1×10^{20}	5×10^{20}
Random forest	**0.93**	**0.92**	0.95	**0.91**	0.86	0.91	0.92	0.92	**0.92**	**0.86**	0.90	**0.93**
Gradient boosting	0.91	0.91	0.95	0.90	**0.87**	**0.92**	0.92	0.91	0.88	**0.86**	**0.93**	0.90
CNN OFM+Magpie	0.89	**0.92**	**0.96**	**0.91**	0.85	0.91	**0.93**	**0.94**	0.90	**0.86**	0.91	0.92

TABLE II. Accuracy scores for *n*-type doping PF models. The best accuracy from each temperature and doping concentration is highlighted in bold.

Temperature	300 K				500 K				700 K			
Doping (cm^{3})	-1×10^{19}	-5×10^{19}	-1×10^{20}	-5×10^{20}	-1×10^{19}	-5×10^{19}	-1×10^{20}	-5×10^{20}	-1×10^{19}	-5×10^{19}	-1×10^{20}	-5×10^{20}
Random forest	**0.90**	0.93	0.92	0.93	**0.86**	0.90	0.94	0.93	0.86	0.86	0.87	0.92
Gradient boosting	0.89	0.93	0.92	0.92	**0.86**	0.91	0.94	**0.95**	**0.87**	0.89	0.85	**0.93**
CNN OFM+Magpie	**0.90**	**0.95**	**0.95**	**0.94**	0.84	**0.93**	**0.95**	0.94	0.86	**0.90**	**0.90**	0.92

shown in Fig. 2 converts the outputs into probabilities between 0 and 1. Then, the sigmoid activation function probability output between 0 and 1 works perfectly with binary cross-entropy, which operates with probabilities. The batch size used in our model is 32, which means that the features from 32 materials are fed into the model to update its weights. After feeding all the batches from the training set into the model, the model makes predictions on the materials from the validation set. From the predictions of the training (validation) set, the training (validation) loss is calculated at that particular epoch. The number of epochs which the model is supposed to run is 70 (i.e., iterations). Our developed model seems to converge (i.e., reaches minimum validation loss) before 70 epochs (normally between 15 and 30 epochs) then the model starts to overfit in subsequent epochs. The model uses the callback functionality from TensorFlow[51] to find the epoch with the minimum validation loss (i.e., the best model) to save its weights and biases. Therefore, even if the model starts overfitting at some

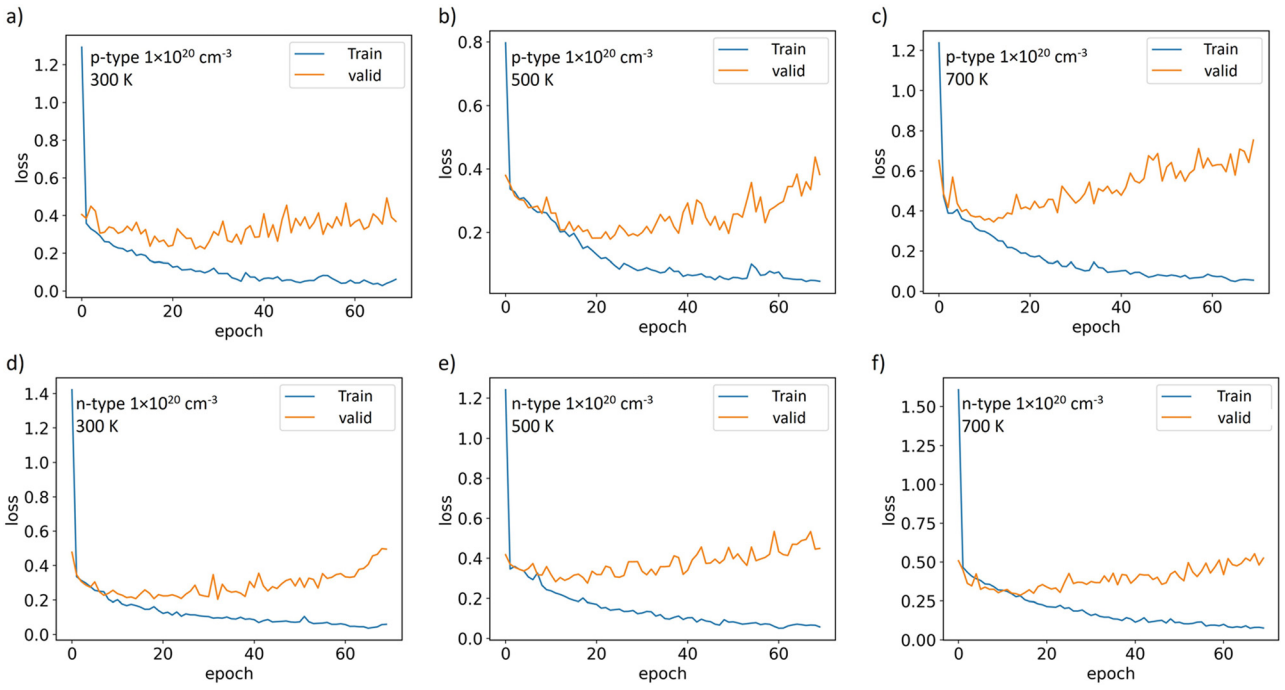

FIG. 3. Training and validation loss function curves for both *p*-type and *n*-type doping concentration of 1×10^{20} cm^{-3} at temperatures 300, 500, and 700 K. The rest of the training and validation curves for other doping concentrations and temperatures are available in the supplementary material.

epoch, the best model weights are still saved to make predictions on the testing set later. The cases shown in Fig. 3 are for all temperatures with *p*- and *n*-type doping concentration of 1×10^{20} cm^{-3}.

In Fig. 4, the receiver operating characteristic (ROC) curves are illustrated for *p*- and *n*-type doping levels of 1×10^{20} cm^{-3} and for all calculated temperatures. The ROC curve represents the model's ability of classification at various classification thresholds. The ROC curve uses two parameters: (1) the true-positive rate (TPR) and (2) false-positive rate (FPR), which can be defined as follows:

$$TPR = \frac{TP}{TP + FN},$$

$$FPR = \frac{FP}{FP + TN}.$$

Another metric that can be used from the ROC curve is the area under the curve (AUC). As the ROC AUC increases, more predictions are correct. A ROC AUC value of 0 means that no prediction is correct. In contrast, a ROC AUC value of 1 means 100% of the prediction are correct. In this work, the ROC AUC curve is used to inform us how much the model is capable of distinguishing between high and low PF materials classes, i.e., 1 means high PF and 0 means low PF. From Fig. 4, we can see that all the models exhibit high ROC AUC values > 0.9, implying that our trained ML models have outstanding performance of distinguishing whether a material possesses a high PF or not. The ROC curves for the rest of the doping levels and their temperatures can be found in supplementary material in Figs. S1 and S2 in *p*- and *n*-type, respectively.

In Fig. 5, the confusion matrices are shown for our models with *p*- and *n*-type doping concentrations of 1×10^{20} cm^{-3} for all calculated temperatures with "virdis" color map taken from the Matplotlib library.[52] The confusion matrix has a matrix size of number of classes $\times$ number of classes. In this work, two classes are trained on, so the matrix size is 2×2 for each model. The rows of the confusion matrix represent the true class labels, and the columns of the confusion matrix represent the predicted class labels. The order of the class labels between the rows and columns is the same. For example, if the true class label 1 has the zeroth index in rows, the predicted class label 1 has the zeroth index in columns. The percentage in the confusion matrices represent the number of materials with a predicted label divided by the total number of materials with that true label. Therefore, the diagonal elements and percentages in the confusion matrix should be as high as possible to have a high model accuracy, or alternatively the off-diagonal elements and percentages should be as low as possible. In Fig. 5, all subplots generally show high diagonal numbers with high corresponding percentages and low off-diagonal numbers with low corresponding percentages, which confirms the high accuracy of our classification models, i.e., most of the predictions are correct (either true positive or true negative). The illustrated confusion matrices reaffirm not only the high accuracy found from Tables I and II but also the high ROC AUC scores from Fig. 4, which indicates the models' high capability of distinguishing high PF materials from the low PF ones. It can also be noted that, at 700 K in the confusion matrices [Figs. 5(c) and 5(f)], the percentage is a bit lower for true label 1 and predicted label 1. That occurred because the training data are

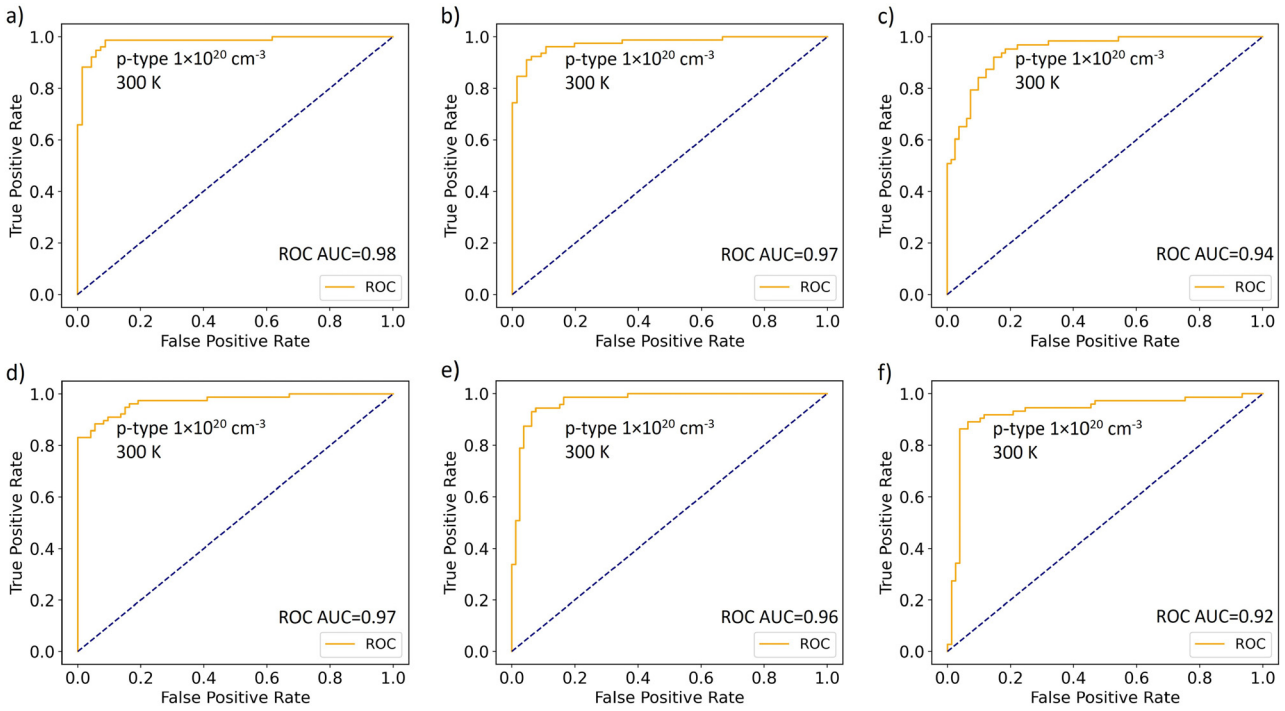

FIG. 4. The receiver operating characteristic (ROC) curves for both *p*-type and *n*-type doping concentration of 1×10^{20} cm^{-3} at temperatures 300, 500, and 700 K. The dashed line represents the 50% true-positive rate and 50% false-positive rate. The rest of ROC curves for other doping concentrations and temperatures are available in the supplementary material. The value of ROC AUC is indicated in each panel.

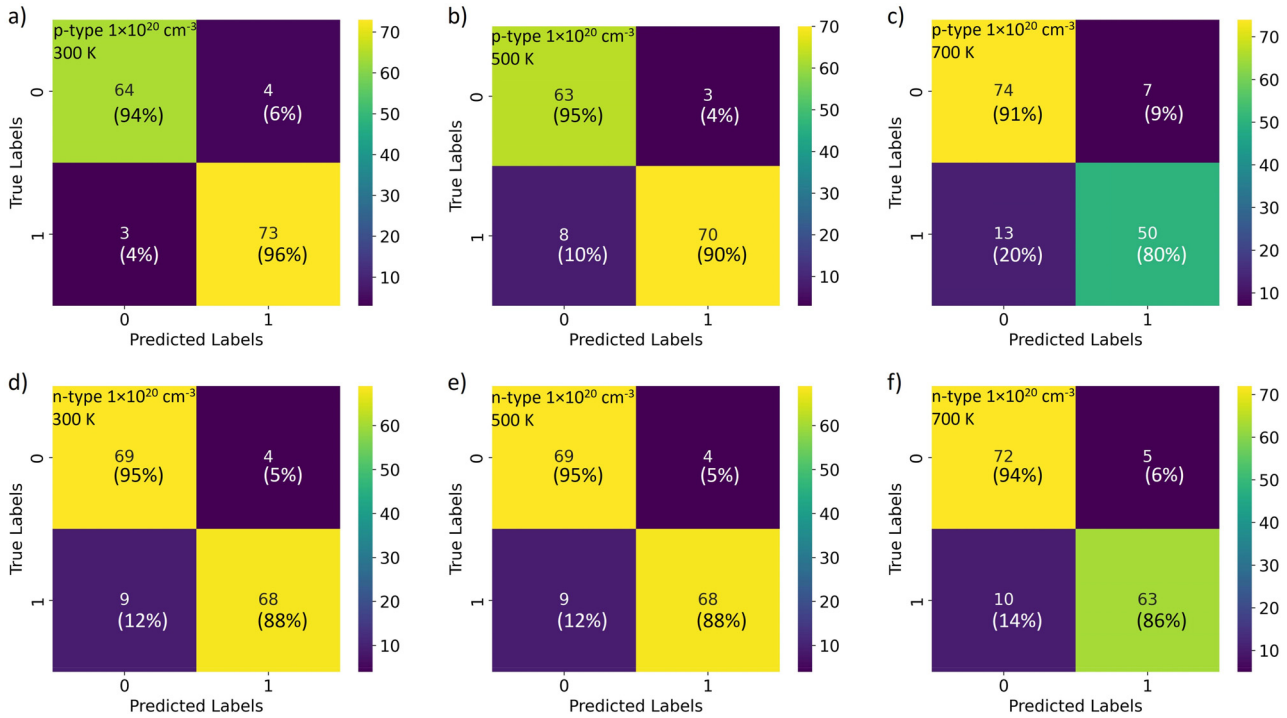

FIG. 5. The confusion matrix for both *p*- and *n*-type doping concentration of $1 \times 10^{20}\,\mathrm{cm}^{-3}$ at temperatures 300, 500, and 700 K. The numbers inside the box show the corresponding count by comparing with DFT data. The percentages shown in the parenthesis are the respective percentage by dividing the total count in the same row. The overall low off-diagonal percentages indicate the high-prediction performance by the trained classification model. The rest of confusion matrices for other doping concentrations and temperatures are available in the supplementary material.

imbalanced with more 0s than 1s, i.e., more materials with low PF ($\mathrm{PF} < 10\,\mu\mathrm{W/cm\,K^2}$) than high PF ($\mathrm{PF} > 10\,\mu\mathrm{W/cm\,K^2}$), which is an expected occurrence for imbalanced dataset. The rest of the confusion matrices for other doping levels and their temperatures can be found in supplementary material in Figs. S3 and S4 in *p*- and *n*-type, respectively. The results shown in Figs. 3–5 are for a single doping concentration of $1 \times 10^{20}\,\mathrm{cm}^{-3}$ because these models will be used later for screening new (untested) OQMD materials for the same doping concentration. We chose to focus on the doping concentration of $1 \times 10^{20}\,\mathrm{cm}^{-3}$ for all temperatures because it is usually an optimal doping concentration for broad operating temperatures and also achievable for many semiconductors[30,53–55] for which experimental dopability depends on native defect energetics, which is hard to compute using first-principles calculations in a systematic high-throughput manner. In fact, many relevant thermoelectric materials were doped at similar concentration level such as Bi_2Te_3 with a *p*-type doping of $7.79 \times 10^{19}\,\mathrm{cm}^{-3}$,[56] Bi_2Se_3 with an *n*-type doping of $2.2 \times 10^{19}\,\mathrm{cm}^{-3}$,[57] $AlFe_2V$ with an *n*-type doping concentration of $5 \times 10^{20}\,\mathrm{cm}^{-3}$,[58] CoSbZr with an *n*-type doping concentration of $2.72 \times 10^{20}\,\mathrm{cm}^{-3}$,[59] and SnTe with a *p*-type doping concentration of $1 \times 10^{21}\,\mathrm{cm}^{-3}$.[60]

After observing the high accuracy of our developed model from Tables I and II and Figs. 3–5 and its capacity in distinguishing the high PF structures from the low PF ones, the models, which were used to classify PF at 300, 500 and 700 K with *p*- and *n*-type doping levels of $1 \times 10^{20}\,\mathrm{cm}^{-3}$ (six models in total), were utilized to classify PF of 3465 untested OQMD materials to screen for new potential thermoelectric materials. The screened materials could overlap between the lists, but ZT is calculated for randomly selected materials after screening for dynamic stability. Calculating ZT requires computing LTC, and unfortunately, LTC information is not provided by the AMSET package. To this end, for selected promising thermoelectric candidates, the LTC was separately calculated by us at 300 K using the ShengBTE package after obtaining the second-order and third-order force constants by DFT. The LTC at other temperatures (500 and 700 K) was then computed using the physical law of temperature-dependent LTC ($\kappa_{\mathrm{ph}} \propto 1/T$), which is a reasonable assumption for most crystalline materials at higher temperatures, in particular for low LTC materials.[61–63] After calculating LTC with ShengBTE at 300 K and extrapolating to high temperatures by $1/T$ scaling law, all information to calculate ZT becomes available.

Moreover, we compare our AMSET calculation results with some experimental measurements and we found in the literature in Table III. The comparison shows that our DFT calculations using AMSET package but without CRTA are in good agreement with the experimental measurements and are better than our separate DFT results with CRTA, proving that the AMSET package has high accuracy to quantitatively predict thermoelectric properties. This gives us the confidence to use the PF data generated by the AMSET package as training data to train high-fidelity deep CNN with fused OFM and composition descriptors model. It is worth noting that the values in the table are interpolated from the figures in the respective references using the WebPlotDigitizer software.[74] Furthermore, the AMSET tool

TABLE III. Comparison of *ZT* of typical thermoelectric materials between experiments and our DFT calculations with both *n*- and *p*-type concentrations at various concentration levels and temperatures. The results with constant relaxation time approximation (CRTA) were obtained by AMSET with constant relaxation time of 10 fs as implemented in Refs. 28, 30, and 31.

Materials	Carrier concentration	*p*- or *n*-Type	Temperature (K)	*ZT* by experiments	*ZT* without CRTA	*ZT* by CRTA
PbS	5×10^{19}	n	500	0.4[64]	0.38	0.51
PbS	5×10^{19}	n	700	0.62[64]	0.7	1.18
PbS	5×10^{19}	n	850	0.7[64]	0.98	1.8
PbS	4×10^{18}	p	300	0.25[65]	0.09	0.11
PbS	4×10^{18}	p	600	0.31[65]	0.18	0.55
PbSe	3×10^{19}	n	500	0.45[64]	0.52	0.69
PbSe	3×10^{19}	n	700	0.9[64]	0.94	1.46
PbSe	3×10^{19}	n	850	1.18[64]	1.19	1.98
PbSe	5×10^{18}	p	300	0.1[66]	0.1	0.13
PbSe	5×10^{18}	p	400	0.15[66]	0.15	0.26
PbSe	5×10^{18}	p	500	0.21[66]	0.21	0.46
PbTe	4×10^{18}	n	300	0.24[67]	0.17	0.11
PbTe	4×10^{18}	n	400	0.27[67]	0.24	0.21
GeSe	5.4×10^{16}	p	600	0.006[68]	0.003	0.003
GeSe	5.4×10^{16}	p	700	0.02[68]	0.002	0.001
GeTe	7.84×10^{20}	p	337	0.04[69]	0.14	0.28
GeTe	7.84×10^{20}	p	372	0.06[69]	0.19	0.34
GeTe	7.84×10^{20}	p	471	0.16[69]	0.35	0.54
GeTe	7.84×10^{20}	p	521	0.3[69]	0.44	0.63
GeTe	7.84×10^{20}	p	570	0.44[69]	0.55	0.74
GeTe	7.84×10^{20}	p	622	0.76[69]	0.68	0.86
Mg_3Sb_2	2×10^{19}	n	322	0.45[70]	0.38	0.245
Mg_3Sb_2	2×10^{19}	n	400	0.56[70]	0.56	0.46
Mg_3Sb_2	2×10^{19}	n	500	0.66[70]	0.78	0.8
Mg_3Sb_2	2×10^{19}	n	600	0.76[70]	0.92	1.11
Mg_3Sb_2	5×10^{18}	p	300	0.04[71]	0.03	0.15
Mg_3Sb_2	5×10^{18}	p	400	0.06[71]	0.02	0.18
Mg_3Sb_2	5×10^{18}	p	500	0.08[71]	0.11	0.12
Mg_3Sb_2	5×10^{18}	p	600	0.12[71]	0.2	0.06
ZrNiSn	5.5×10^{19}	n	300	0.3[72]	0.39	0.08
ZrNiSn	0.8×10^{20}	p	400	0.15[72]	0.22	0.13
ZrNiSn	1.05×10^{20}	p	600	0.4[73]	0.5	0.49
ZrNiSn	1.09×10^{20}	p	700	0.47[73]	0.61	0.67
ZrNiSn	1.14×10^{20}	p	800	0.55[73]	0.6	0.73

without CRTA electron mobility calculations showed great consistency with experimental results.[41] Also note that experimental results may have some uncertainty or discrepancy when thermoelectric properties of single crystals or poly-crystals[70,75] are measured, or when the properties are measured by different methods such as Gasar process, solvothermal with sintering, NaCl flux method, solvothermal with spark plasma sintering (SPS), melting with annealing and SPS, solid-state reaction (SSR) with hot processing (HP), ball milling (BM) with SPS, solid-state explosion, and others.[75]

Tables IV–VII show the PF and ZT results for the selected promising materials for *p*-type and *n*-type doped with doping concentration $1 \times 10^{20}\,\mathrm{cm}^{-3}$, as well as the LTC at room temperature. Tables IV and V show some selected structures used for model training, validation, and testing, i.e., all data presented therein are calculated directly by full DFT + AMSET. Tables VI and VII are selected from the 3465 OQMD structures screened using the classification model. All materials have high PF at all temperatures except for *p*-type doped KPdRb with PF of 9.112 μW/cm K^2, which is still close to the threshold value of 10 μW/cm K^2. Note that the results in Tables VI and VII are separated from Tables IV and V to demonstrate that our fused OFM and Magpie CNN model can identify more materials with high PF. It can also be observed that all materials have $ZT > 1$ or close to 1 at higher temperatures, which makes them potential candidates for thermoelectric applications at elevated temperatures. There are also some materials, which are predicted to have ZT higher than or close to 1 even at room temperature, which are promising for recovering low quality but large

TABLE IV. Selected thermoelectric materials with PF and ZT results at *p*-type doping concentration of $1 \times 10^{20}\,cm^{-3}$ and at different temperatures from the training, validation, or testing datasets.

OQMD ID	Formula	LTC at 300 K (W/mK)	PF at 500 K (μW/cm K^2)	PF at 700 K (μW/cm K^2)	PF at 700 K (μW/cm K^2)	ZT at 300 K	ZT at 500 K	ZT at 700 K
942854	AlHgSb	1.567	244.9064	243.5590	178.4441	0.53	1.20	0.91
567963	NbInPt	12.173	242.9937	272.0877	248.5897	0.43	1.18	2.01
729608	MnBeVGe	6.401	106.8451	95.2929	39.6244	0.41	0.78	0.23
925506	LiAlSn	7.996	82.34713	131.9031	148.7458	0.20	0.65	1.01
997321	LiVInIr	5.613	81.4267	95.7431	68.0203	0.27	0.71	0.51
579059	NaAlSi	6.874	69.6700	86.0596	86.4577	0.26	0.78	1.38
995819	LiZrBiRu	8.551	48.8263	50.5872	45.9085	0.16	0.45	0.76
1250585	Y_2SiSe_4	4.825	48.1650	51.1759	46.3743	0.25	0.70	1.20
898822	ScSbPd	11.616	46.7572	73.3287	76.2484	0.11	0.42	0.68
1284087	Y_2SiSe_4	4.807	46.0826	48.4221	44.3779	0.24	0.67	1.15
1283934	Sc_2SnSe_4	3.319	36.1660	34.9674	30.9936	0.30	0.77	1.30
1250565	Y_2Te_4Pb	2.899	25.1695	23.9068	20.7267	0.24	0.62	1.05
1284073	Y_2GeSe_4	4.892	24.4851	23.2086	20.4811	0.14	0.37	0.62
1250557	Y_2SnSe_4	3.626	22.4878	21.1427	18.4563	0.18	0.45	0.76
1284089	Y_2SnS_4	1.838	14.7166	13.5523	11.7395	0.23	0.56	0.94
1575131	$Cs_2InRhBr_6$	0.218	14.6430	19.1570	19.8645	0.20	0.53	1.14
1041498	CsKRbBi	0.181	13.3645	8.66912	2.5114	1.82	2.36	0.34
1549329	$Cs_2HgPtCl_6$	0.295	9.3811	13.7371	16.0178	0.15	0.41	0.79
1575765	$Cs_2InRhCl_6$	0.750	7.7181	10.0938	10.5883	0.14	0.42	0.91

amount of waste heat near room temperature. The potential thermoelectric materials shown in Tables IV–VII include various family compounds such as quaternary Heuslers (BiLiMgZn and LiMgSbZn, space group no. 225), half-Heuslers (KNaPd, CdLiSb, and MgSiSr, space group no. 216), $ABCD_2$ prototype ($Cs_2HgPtCl$), cubic double perovskites ($Cs_2InRhBr_6$ and $Cs_2InRhBr_6$), AB_2C_6 type chalcogenides (Y_2GeSe_4, Sc_2SnSe_4, Y_2SnSe_4, and Y_2SnS_4), and binary cubic structure (AgBr, PbS, PbSe, PbTe, and $BiLi_3$). It is worth pointing out that our

TABLE V. Selected thermoelectric materials with PF and ZT results at *n*-type doping concentration of $1 \times 10^{20}\,cm^{-3}$ and at different temperatures from the training, validation, or testing datasets.

OQMD ID	Formula	LTC at 300 K (W/mK)	PF at 300 K (μW/cm K^2)	PF at 500 K (μW/cm K^2)	PF at 700 K (μW/cm K^2)	ZT at 300 K	ZT at 500 K	ZT at 700 K
729608	MnBeVGe	6.401	116.5457	104.9196	35.7103	0.45	0.85	0.19
1006467	MgVReSi	12.766	116.2575	153.9765	124.1762	0.22	0.62	0.47
942854	AlHgSb	1.567	108.7240	100.5835	15.7754	0.49	0.60	0.05
898822	ScSbPd	11.616	97.4349	99.6139	85.2037	0.23	0.61	0.73
1250548	KYSe	0.481	80.2304	81.9057	81.7615	0.51	1.16	1.85
898434	HfGePd	15.939	76.7029	94.0526	100.4008	0.14	0.43	0.85
579059	NaAlSi	6.874	65.5342	99.7444	113.0222	0.20	0.68	1.29
567963	NbInPt	12.173	57.7349	85.2277	95.1607	0.11	0.38	0.77
1250585	Y_2SiSe_4	4.825	49.0836	38.9358	31.8425	0.28	0.60	0.93
1575765	$Cs_2InRhCl_6$	0.7503	29.3976	21.6301	16.8329	0.94	1.81	2.64
432066	Rb_2BiO_4	0.0653	27.8250	35.9043	38.6706	0.12	0.40	0.89
1549329	$Cs_2HgPtCl_6$	0.295	23.2098	16.2718	12.2651	1.61	2.85	3.88
898037	RbNaPd	0.1702	21.9085	27.1496	26.7757	0.91	2.22	3.39
1546761	$Cs_2TlRhBr_6$	0.0808	13.0705	10.9407	9.2068	1.89	3.60	5.26
1721624	$Cs_2InIrCl_6$	0.5672	11.6465	9.4568	7.8977	0.53	1.14	1.79
1368340	Cs_2NaCdF_6	0.386	9.5626	14.8991	17.0851	0.09	0.31	0.63
974579	RbSrSb	0.346	8.5839	11.4278	12.8404	0.27	0.70	1.27

TABLE VI. Selected thermoelectric materials with PF and ZT results at *p*-type doping concentration of 1×10^{20} cm^{-3} and at different temperatures screened by our trained fused OFM and Magpie CNN model.

OQMD ID	Formula	LTC at 300 K (W/mK)	PF at 300 K (μW/cm K^2)	PF at 500 K (μW/cm K^2)	PF at 700 K (μW/cm K^2)	ZT at 300 K	ZT at 500 K	ZT at 700 K
899499	KNaPd	0.600 16	47.6144	36.049 8	25.7415	1.72	3.09	2.20
899241	KPdRb	0.681 83	14.3999	11.310 8	9.1116	0.58	1.20	1.85
734899	Li_2SiZn	4.669 5	94.0066	150.876 89	167.0331	0.22	0.72	1.17
713316	BiLiMgZn	1.191 1	94.9388	102.470 2	88.9047	0.42	1.06	1.10
713314	LiMgSbZn	1.065 6	142.0950	161.540 7	118.7023	0.66	1.53	0.83
1023013	AuCaLiSi	0.537 28	235.5487	225.579 8	168.0236	1.19	1.22	0.66
985659	CdLiSb	0.614 34	105.3302	85.941 0	46.3264	1.27	0.85	0.33
18757	BiLiMg	5.215 2	46.7374	54.220 5	53.4019	0.23	0.69	1.23
10938	$BiLi_3$	2.789 6	38.5131	45.596 2	44.0879	0.29	0.85	1.40
931975	BeNaSb	4.874 5	30.4577	48.457 8	55.8509	0.15	0.55	1.12
1104488	AgBr	0.222 48	11.9094	10.462 1	9.1866	1.24	2.71	4.24
1223796	PbS	1.6	11.8085	9.415 7	7.4826	0.20	0.44	0.68
1223799	PbSe	1.01	10.4161	10.283 0	8.9066	0.26	0.68	1.12
1106321	PbTe	1.622 1	36.6492	46.029 2	48.6060	0.37	1.13	2.18

fused OFM and Magpie model reproduce cubic PbX (X = S, Se, and Te) systems as high PF thermoelectric materials from the OQMD database, which have been well studied in the past.[5–8] This further demonstrates the accuracy of our model since these previously studied thermoelectric materials were not included in training, validation, or testing sets, and thus, our model is promising for discovering new thermoelectric materials.

In order to obtain deep insight into the structure–property relationship, feature importance values are computed and shown in Fig. 6 for the random forest model for *p*- and *n*-type doping concentration of 1×10^{20} cm^{-3} and the temperatures 300, 500, and 700 K. As mentioned earlier in this work, the random forest model utilized Magpie composition descriptors in training. The feature importance for a model can be defined as the relative importance of each feature when building that model. The sum of the feature importance values for all features used to train a model is equal to unity. The feature importance values can inform us which features are the most important in obtaining a model with such high accuracy which can give more insights into PF. It should be noted that feature importance values might change in each run due to stochastically random nature of machine learning models in changing their weights and biases. The top seven materials descriptors with highest feature importance values for each individual ML model we trained are presented in the respective panel in Fig. 6. It can be seen in all panels that the mean melting temperature and average deviation of electronegativity always occur in the top three highest importance values among the seven descriptors for all cases. This clearly demonstrates the crucial role of these descriptors in predicting PF of a structure as potential thermoelectric materials. In addition, it

TABLE VII. Selected thermoelectric materials with PF and ZT results at *n*-type doping concentration of 1×10^{20} cm^{-3} and at different temperatures screened by our trained fused OFM and Magpie CNN model.

OQMD ID	Formula	LTC at 300 K (W/mK)	PF at 300 K (μW/cm K^2)	PF at 500 K (μW/cm K^2)	PF at 700 K (μW/cm K^2)	ZT at 300 K	ZT at 500 K	ZT at 700 K
899241	KPdRb	0.681 83	10.102 6	12.489 6	12.687 8	0.32	0.95	1.69
899499	KNaPd	0.600 16	21.118 0	27.363 0	29.509 7	0.43	1.20	1.67
946541	AuGaTi	19.773	97.802 3	106.528 1	99.776 0	0.14	0.38	0.72
713316	BiLiMgZn	1.191 1	235.946 8	194.975 9	98.147 9	1.60	1.23	0.58
713314	LiMgSbZn	1.065 6	171.794 6	132.955 0	60.227 5	1.94	1.10	0.49
931975	BeNaSb	4.874 5	40.833 9	51.846 6	54.865 2	0.20	0.68	1.20
579163	MgSiSr	3.733 4	38.194 5	46.939 1	49.158 6	0.25	0.78	1.34
10938	$BiLi_3$	2.789 6	20.370 1	26.597 5	28.998 6	0.15	0.51	0.85
867124	GeMgSr	2.962	27.302 2	41.465 8	51.356 9	0.14	0.51	0.80
1223796	PbS	1.6	13.928 49	17.383 44	18.442 77	0.21	0.63	1.18
1223799	PbSe	1.01	16.028 85	19.600 51	20.694 77	0.35	1.00	1.75
1106321	PbTe	1.622 1	12.358 6	19.318 87	23.763 8	0.12	0.43	0.88

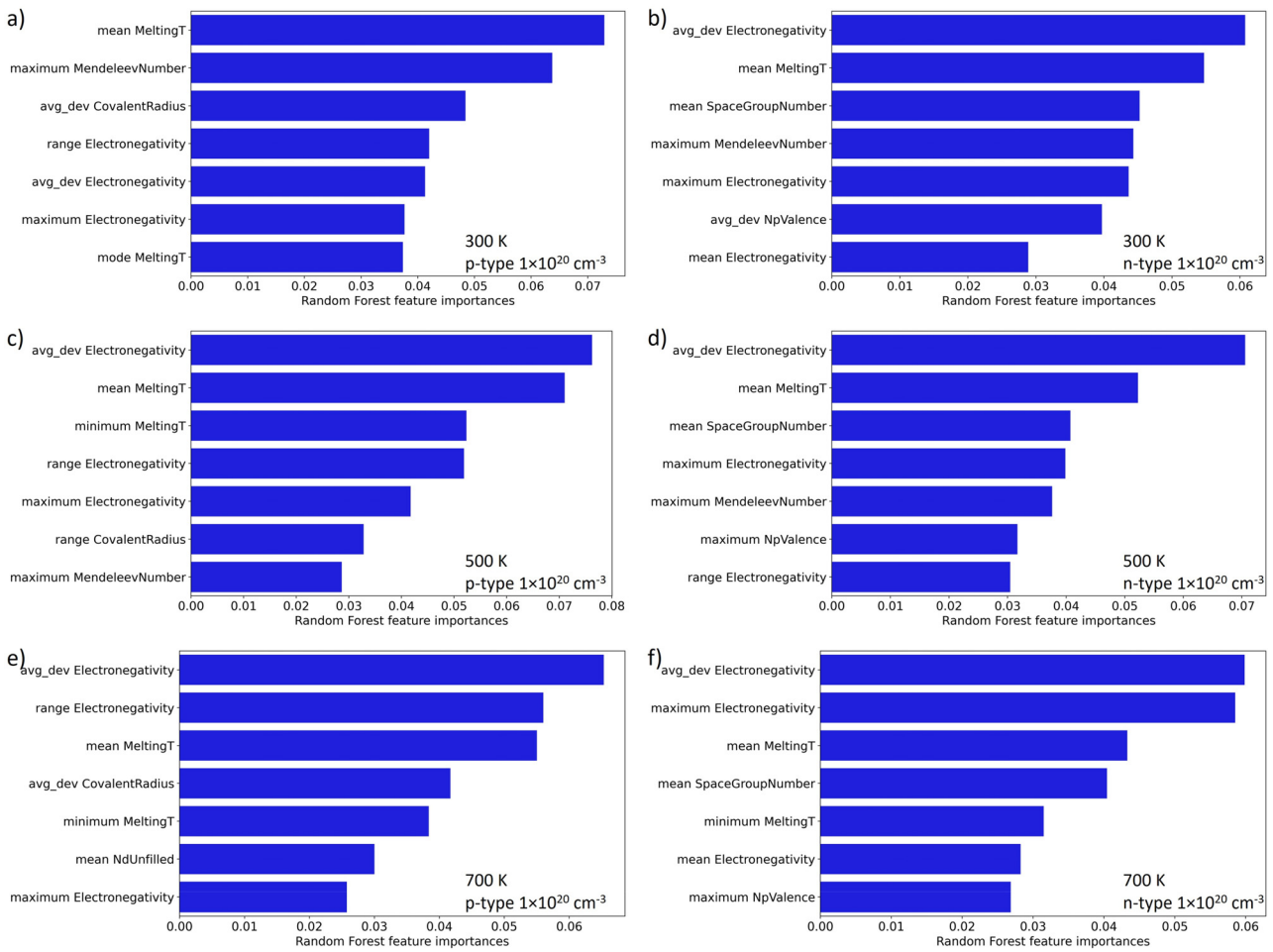

FIG. 6. Feature importance values from training a random forest model for PF of both *p*- and *n*-type doping concentration of $1 \times 10^{20}\,\mathrm{cm}^{-3}$ at temperatures 300, 500, and 700 K.

seems that the common features (excluding the statistical operation, e.g., mean, max, min, range, deviation) among all the panels are melting temperature, space group, electronegativity, number of unfilled d-orbital electrons, Mendeleev number, space group number, covalent radius, and number of valence electrons in *p*-orbitals. Similar machine learning analysis for feature importance results has been reported in various other works, such as chemical composition, radial distribution function, charge, angular distribution function up to the first and second nearest neighbors, and nearest neighbors from classical force field descriptors (CFID).[30,76] A similar study has reported that the maximum difference in melting temperature, mean number of *d*-orbitals, deviation of covalent radius, and mean ionic character are important features to predict thermoelectric properties using machine learning.[77] Finally, it is worth pointing out that these Magpie composition descriptors are easy to obtain, which provides a useful tool for quickly screening large-scale potential candidates.

Another common way to quantify the correlation among variables is to use maximal information coefficient (MIC).[78,79] The MIC score reflects both linearity and non-linearity strength of correlation between two variables. A MIC score can range between 0 and 1 values, with 0 corresponding to no correlation and 1 corresponding to strong correlation. The highest MIC values for PF with other Magpie composition features are shown in Fig. 7 for the same *n*- and *p*-type doping concentration of $1 \times 10^{20}\,\mathrm{cm}^{-3}$ and temperatures of 300, 500, and 700 K. Same as Fig. 6, the MIC scores shown in Fig. 7 are the top seven scores for each case. It can be observed that the MIC is again very high between PF and mean melting temperature and average deviation electronegativity. These results reaffirm the high correlation between PF and these two descriptors from feature importance values presented in Fig. 6. The MIC scores also confirm the importance of other material descriptors (excluding the statistical operation, e.g., mean, max, min, range, deviation) such as the number of unfilled electrons, Mendeleev number, column number in periodic table, number of valence electrons in *p*-orbitals, and number of unfilled electrons in d-orbitals. It can also be observed that cases at the same temperature, but different doping types (either *n*-type or *p*-type) have similar features (including the statistical operation, e.g., mean, max, min, range, deviation) with top MIC scores. For example, in Figs. 7(c) and 7(d), the mean melting

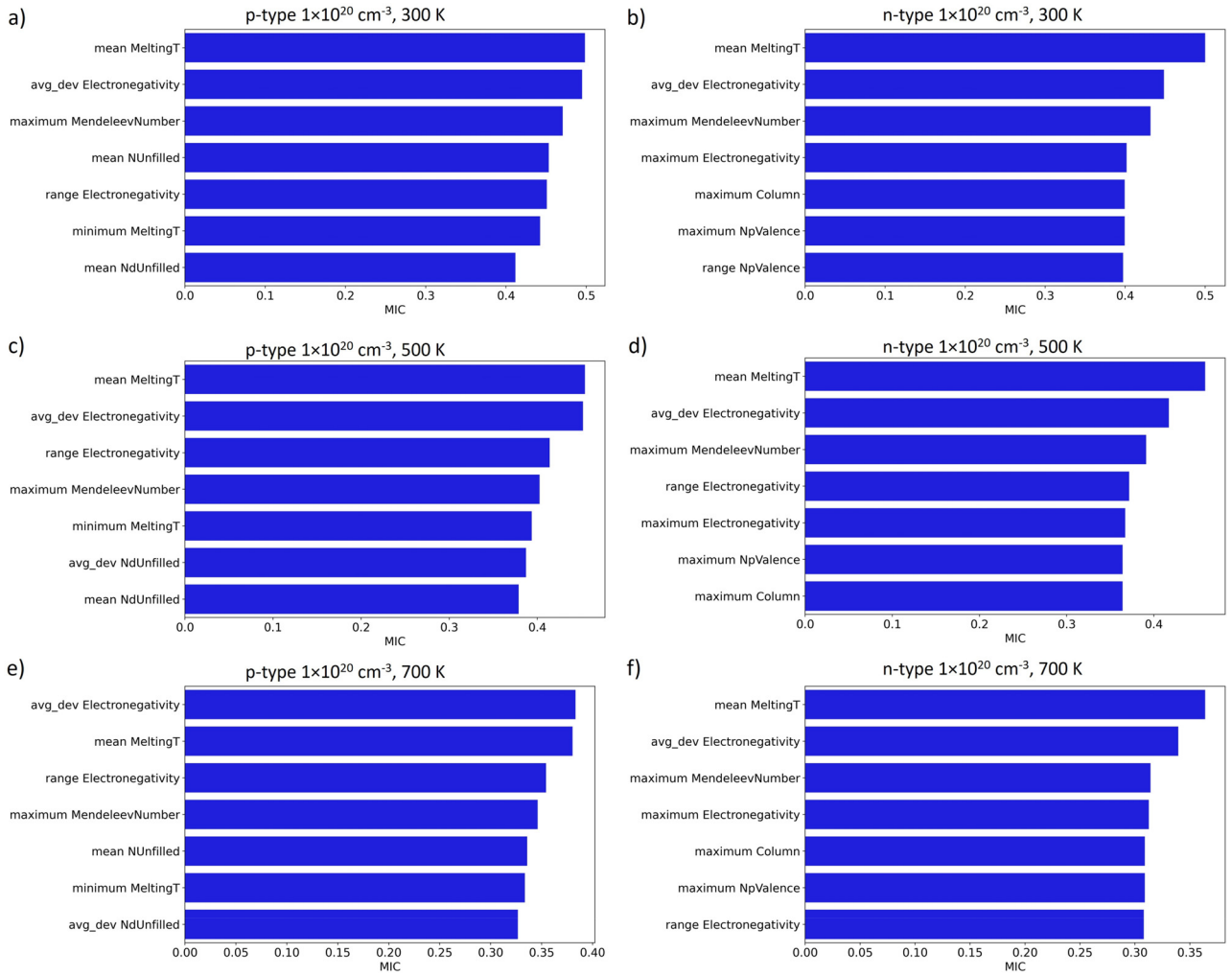

FIG. 7. Maximal information coefficient (MIC) value with PF for both *p*-type and *n*-type doping concentration of 1×10^{20} cm^{-3} at temperatures 300, 500, and 700 K.

temperature, average deviation electronegativity, maximum Mendeleev number, and range electronegativity are the top 4 features with high MIC scores.

Figure 8 shows the effect of average deviation of electronegativity and mean melting temperature on PF as a colormap having red representing high PF or 1 and blue representing low PF or 0 for both *p*- and *n*-type doping with doping concentration of 1×10^{20} cm^{-3} at temperatures of 300, 500, and 700 K. The average deviation of electronegativity and mean melting temperature are selected because both exist in all feature importance subplots in Fig. 6 and have the highest feature importance values. They also manifest the highest MIC scores in all cases as shown in Fig. 7. Figure 8 investigates the effect of these two descriptors with the highest MIC and feature importance values on PF of thermoelectric materials. It can be clearly observed from both *n*- and *p*-doping types, all doping concentrations, and all temperatures that, as mean melting temperature increases and average deviation of electronegativity decreases, the material more likely has a high PF. Although exceptions of both cases exist, i.e., materials with high mean melting temperature and low average deviation of electronegativity have low PF and vice versa as shown by a few red dots (high PF) in a blue majority region (low PF region) and vice versa, the trend seems to be obvious. This phenomenon can be explained by first examining how mean melting temperature and average deviation of electronegativity of the constituent elements are computed. The mean melting temperature is the average melting temperature of all atomic species in the primitive cell. The average deviation of electronegativity is calculated by computing the mean of electronegativity in the cell, and then, the absolute value is obtained after subtracting the mean value from each specie's electronegativity then the mean from those absolute values is calculated, or in simpler mathematical form, $\frac{\sum_{i=1}^{n} |X_i - X_{avg}|}{n}$, where X_i is the electronegativity of the site in the cell, and X_{avg} is the average electronegativity of the sites in the cell. The average deviation of electronegativity thus gives a measure of how much on average the electronegativity of the sites deviates from other sites, i.e., large average deviation of electronegativity means electronegativity of constituent

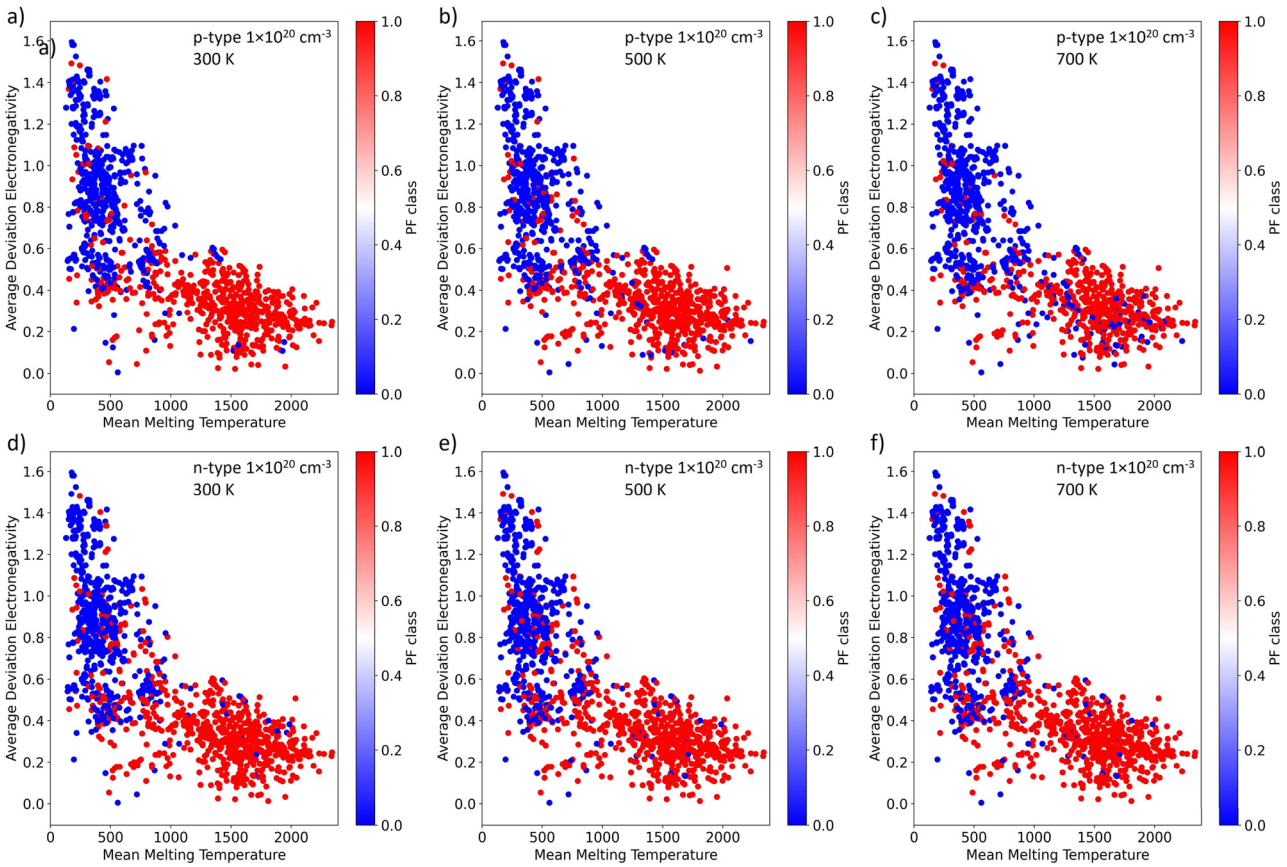

FIG. 8. Average deviation of electronegativity vs mean melting temperature with power factor (PF) classes as a color map (red represents "1" or high PF and blue represents "0" or low PF) for both *p*-type and *n*-type doping concentration of 1×10^{20} cm^{-3} at temperatures 300, 500, and 700 K.

elements differs significantly. Regarding the mean melting temperature, most thermoelectric materials have transition metals and rare earth elements.[80–82] Transition metals tend to have high melting temperature and high conductivity, which is a desirable property for high thermoelectric performance due to forming strong covalent bonds with other unfilled d-shell valence electrons.[45,82] Therefore, it is predicted that many materials with high PF, many of which have transition metals, have high mean melting temperature from the constituent elements, as can be seen from red dots concentrated on the high mean melting temperature region in all panels in Fig. 8. As for the average deviation of electronegativity, high average deviation of electronegativity indicates a higher tendency of forming ionic bonds between atoms in the primitive cell. Large ionic bonding tendency indicates more electron charges transferred from the anion to the cation, which reduces the electron carrier concentration, conductivity, and mobility and thus is not favorable for thermoelectrics, due to the less freedom that electrons experience compared to the shared electrons that can swim in an electron cloud more freely. This trend was observed in $Mg_3Sb_{2-x}Sn_x$ where the electronegativity difference is decreased, which enhances the covalent bonding character between the atoms and as a result increases electrical conductivity and the PF.[83] This pattern is also observed in other materials with ionic bonds and was confirmed that such materials have low electrical conductivity[84–86] due to less electron movement freedom, which reduces PF consequently. Therefore, high-performance thermoelectrics are clearly seen from the red dots concentrated in the low average deviation of electronegativity region in all panels in Fig. 8. From the above analysis, our MIC results of mean melting temperature and average deviation of electronegativity PF descriptors from Fig. 8 are supported by transition metals with high melting temperature and covalent chemical bonding, respectively. We notice that some exceptional materials of low (high) PF represented by the blue (red) dots located at the high (low) PF region exist in Fig. 8, which require a more comprehensive analysis to have better understanding of those materials and cannot be explained by merely two descriptors identified herein. Nevertheless, our analysis is still useful since it is consistent with a visually obvious pattern for the majority of the data in Fig. 8. Figure 8 provides a new route for quickly screening potential high PF materials by two simple material descriptors with low computational cost, which has not been reported in previous studies, to the best of our knowledge.

To further expand on the insights of PF from Fig. 8, we analyze electrical conductivity in base 10 logarithm value and absolute value of the Seebeck coefficient based on the two newly proposed descriptors, i.e., average deviation of electronegativity and mean melting

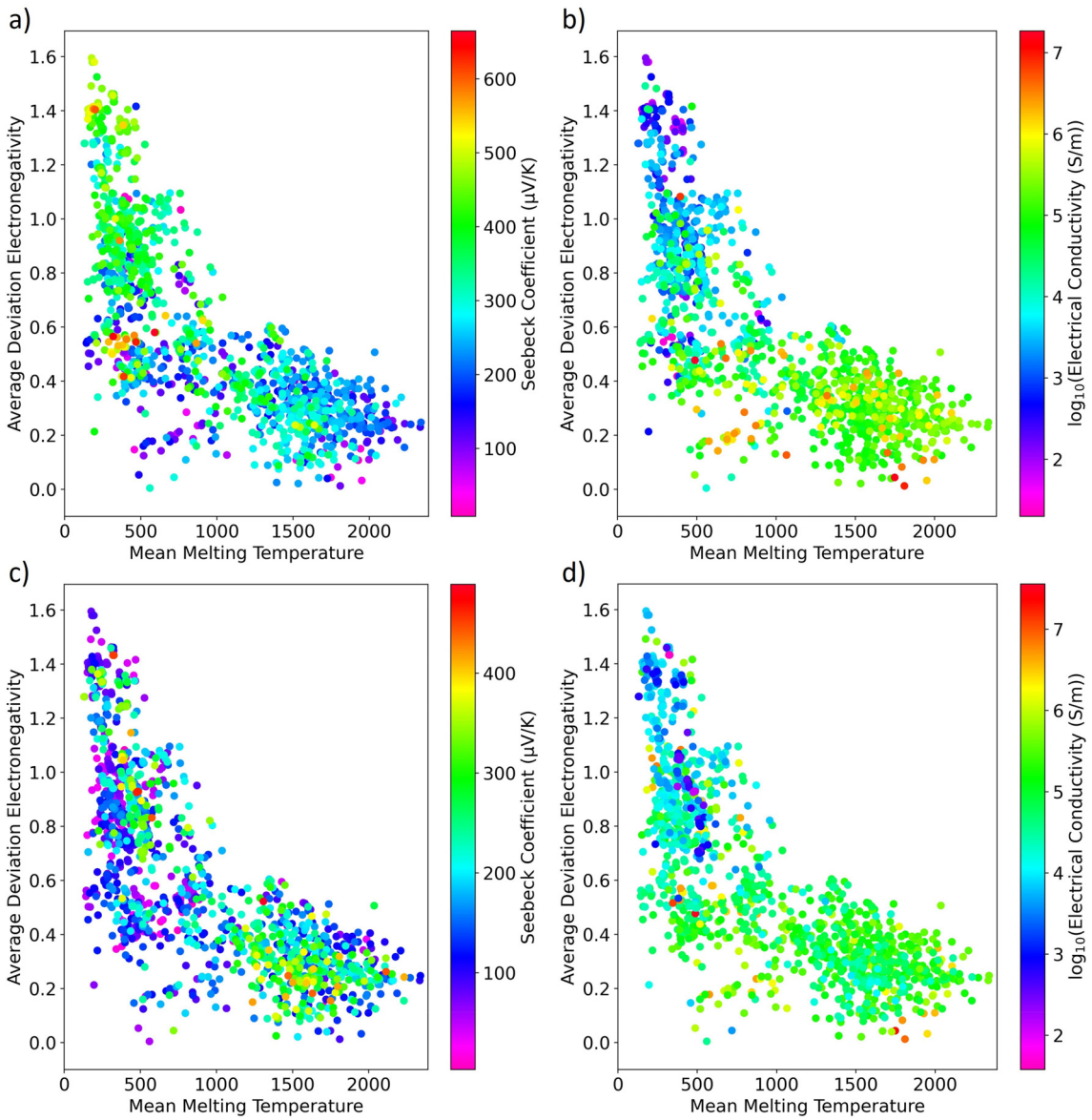

FIG. 9. Average deviation electronegativity vs mean melting temperature for *p*-type (top panels) and *n*-type (bottom panels) materials with carrier concentration of 1×10^{20} cm^{-3} color mapped with (a) and (c) Seebeck coefficient and (b) and (d) electrical conductivity.

temperature. The results are shown in Fig. 9. The logarithm value of electrical conductivity is used in the figure since it gives more variability of the color to distinguish the data. We analyzed both transport properties separately because they are the two components making up the PF, according to its definition (i.e., $S^2\sigma$). We select one temperature (300 K) for the plot in Fig. 9 because a similar analysis has been applied to the other two temperatures (500 and 700 K) in Fig. 8. We include *p*-type and *n*-type materials to see how the analysis differs between both carrier types. It can be seen from Fig. 9 that the *p*-type Seebeck coefficient has the opposite trend as PF, meaning that top left region of the data has most of the materials with a high Seebeck coefficient, whereas most of the materials with high PF occurs at the bottom right. The opposite trend is observed in the Seebeck coefficient with *n*-type carrier concentration where most of the *n*-type materials with a high Seebeck coefficient occur at bottom right, but there are still many *n*-type materials with a high Seebeck coefficient in the top left region. A clear trend, based on those two descriptors, does not emerge, because many exceptions occur in both carrier concentration types. Perhaps different descriptors are needed to explain or screen Seebeck coefficient only. On the other hand, *p*-type and *n*-type materials have high electrical conductivity in the bottom right region as can be seen from Figs. 9(b) and 9(d), which is also the same trend observed in PF.

Moreover, electrical conductivity is a more dominant factor in terms of achieving high PF, because the magnitude of electrical conductivity can differ by several orders among different materials. However, the Seebeck coefficient only differs by a few factors among different materials as can be seen from the color bar in Fig. 9, and thus, it has low contribution to high PF. The electrical conductivity trend can be attributed to the higher electrical conductivity observed for covalently and metavalently bonded materials with low to moderate transferred charge, which results from the low difference of electronegativity.[84–86] Yet, materials with high difference of electronegativity or high average deviation in electronegativity tend to have high transferred charge and form ionic bonds restricting electrons movement, which results in low electrical conductivity.[84–86] It can also be observed from electrical conductivity trend is that mean melting temperature does not play much role in determining which materials will have high or low electrical conductivity. The high mean melting temperature of the constituent atoms might be a byproduct of the existence of transition metals that have medium electronegativity and high melting temperatures due to their strong covalent bonds.[82] Note that we are not claiming that such compounds will necessarily have high melting temperatures. Instead, we are claiming that at least some of the constituent elements have high melting temperatures. It is known that transition metals have medium electronegativity between *s* and *p* block elements in the periodic table. Therefore, replacing a *p* block element with a transition metal to be paired with an *s* block element or vice versa should in principle reduce the electronegativity deviation and raise mean melting temperature of the constituent elements. For example, $MgBr_2$ has a 1.65 electronegativity difference between Mg and Br.[87] However, if Mg with 923 K melting temperature is replaced by Cu with 1358 K melting temperature, the electronegativity difference between Cu and Br reduces to 1.06, which decreases the ionic character and increase the covalent bonding character,[87] and the mean melting temperature of the constituent elements rises. This is demonstrated by doping $CuBr_2$ into $C_6H_4NH_2$, which enhanced the doped compound's electrical conductivity and thermoelectric properties.[88] From the previous analysis, it can be shown that transition metals with high melting temperature of the constituent atoms keeping in mind that such transition metals are used in lots of thermoelectric materials[80–82] might be an indirect descriptor to electrical conductivity and therefore potentially high PF.

Before closing, we would like to compare the computational efficiency of our proposed descriptors with those reported in the literature. The Fermi surface complexity factor ($N_V^* K^*$), which is essentially the number of Fermi surface pockets (N_V^*) and their anisotropy (K^*), has been proposed in previous study.[53] That descriptor shows a high correlation with PF computed by DFT with CRTA. Such calculations require self-consistent field (SCF) DFT calculations to compute the Fermi surface and extract such descriptor, and such calculations are not expensive for structures with a low number of atoms with low weights. However, calculations might be expensive if the cell size is big with a high number of heavy elements. The descriptor proposed by Graziosi *et al.*[89] is defined as $n_v \varepsilon_r / (D_0^2 m_{\text{cond}})$, where n_v is the number of valleys, ε_r is the dielectric constant, D_0 is the deformation potential, and m_{cond} is the effective mass conductivity. The number of valleys and conductivity effective mass can be obtained through SCF DFT calculations of the Fermi surface. Dielectric constant calculations require expensive DFPT calculations, which are moderately computationally expensive even if the cell size is small with light elements. The deformation potential calculations require SCF DFT calculations of multiple deformed cells to observe how the valence and conduction bands change in the band structure, with the number of SCF DFT runs depending on the symmetry and complexity of materials. This descriptor showed direct correlation with PF that was proven by the high Pearson correlation of 0.75 and 0.93 for *n*- and *p*-type doped materials. Although the results sound very promising, the descriptor was only analyzed on about 30 materials, which is not enough to be considered universal. A dimensionless parameter is the thermoelectric quality factor with the symbol "β,"[90] also known as the material factor named by Chasmar and Stratton in 1959.[91] The thermoelectric quality factor β is defined as $\beta = \left(\frac{k_B}{e}\right)^2 \frac{T}{k_l} \sigma_c$, where k_B is the Boltzmann constant, e is the electron charge, T is absolute temperature, k_l is lattice thermal conductivity, and σ_c is the electrical conductivity, which can be obtained from analyzing the electronic band structure of SCF calculations. The k_l requires several tens to even hundreds of SCF DFT calculations on displaced supercells instead of primitive cells, which is extremely computationally demanding. Therefore, the quality factor β is not computationally efficient for high-throughput screening. Overall, the previous descriptors are computationally expensive to obtain compared to our descriptors that can be easily calculated for hundreds of materials in mere seconds. The computational cost for our proposed descriptors is even cheaper than SCF DFT calculations for a single material.

IV. CONCLUSION

In summary, we have performed high-throughput DFT calculations coupled with AMSET package to generate a dataset of thermoelectric properties consisting of 1438 *p*-type and 1499 *n*-type cubic structures. Our thermoelectric property dataset considers various electron scattering mechanisms to acquire high-quality data without unreasonably assuming constant relaxation time of electrons as used in previous studies. We further developed a fused OFM and Magpie CNN model to classify high power factor materials screened from OQMD database. Traditional machine learning models only use global properties as end-property training without knowing dynamic stability of materials, which is critical for possible experimental synthesis. Our classification model is deployed on 3465 cubic structures whose dynamic stability is either verified by full DFT calculations or predicted by our recently developed high-accuracy Elemental-SDNNFF, which makes our screening of potential thermoelectrics more physically meaningful. Deep insight into the structure–thermoelectric property has been gained through feature importance analysis and maximal information coefficient method. Our results show that high mean melting temperature and low average deviation of electronegativity of the compound's constituent elements lead to high power factor of a material potential for thermoelectrics. These two new material descriptors with strong correlations with power factor provide a new route for quickly screening large-scale potential thermoelectric candidates, considering that both composition descriptors are simple, and their computational cost are minimal. Our high-throughput DFT calculations combined with screening by the fused OFM and Magpie CNN model have accelerated discovery of new thermoelectric materials with high ZT ($\geq$1) across broad operating temperatures from various family compounds, which hold the promise for future efficient screening of even larger-scale hypothetical structures in unexplored material space. While the current study focuses on recently screened stable cubic structures from OQMD database,[38,39] the techniques and workflow

established herein is anticipated to be straightforwardly extended to noncubic structures, and the relevant work is ongoing.

SUPPLEMENTARY MATERIAL

See the supplemental material for include the ROC curves and confusion matrices for the rest of the doping levels and their temperatures for both *p*-type and *n*-type materials.

ACKNOWLEDGMENTS

This work was supported by the NSF (Award Nos. 2030128, 2110033, 2311202, and 2320292), SC EPSCoR/IDeA Program under NSF OIA-1655740 (No. 23-GC01), and ASPIRE grant from the Office of the Vice President for Research at the University of South Carolina (Project No. 80005046). M.A.-F. acknowledges the financial support of the SPARC Graduate Research Grant (Project No. 80004800). R.R. acknowledges financial support by MCIN/AEI/10.13039/501100011033 under Grant No. PID2020-119777GB-I00, and the Severo Ochoa Centres of Excellence Program under Grant No. CEX2019-000917-S, and by the Generalitat de Catalunya under Grant No. 2021 SGR 01519. DFT calculations were performed at the Centro de Supercomputación de Galicia (CESGA) within action Phonon Database Generation and Analysis for Data Driven Materials Discovery (No. FI-2023-1-0003/FI-2023-2-0005) of the Red Española de Supercomputación (RES).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Mohammed Al-Fahdi: Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Validation (equal); Visualization (equal); Writing – original draft (equal). **Kunpeng Yuan:** Formal analysis (equal); Investigation (equal); Methodology (equal); Validation (equal); Writing – review & editing (equal). **Yagang Yao:** Formal analysis (equal); Investigation (equal); Visualization (equal); Writing – review & editing (equal). **Riccardo Rurali:** Data curation (equal); Investigation (equal); Methodology (equal); Resources (equal); Software (equal); Writing – review & editing (equal). **Ming Hu:** Conceptualization (equal); Data curation (equal); Funding acquisition (equal); Project administration (equal); Software (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are openly available in Github at https://github.com/Mofahdi, Ref. 92.

REFERENCES

[1]D. M. Rowe, *Thermoelectrics Handbook: Macro to Nano* (CRC/Taylor & Francis, 2006).
[2]G. J. Snyder, "Small thermoelectric generators," Electrochem. Soc. Interface **17**, 54–56 (2008).
[3]T. J. Seebeck, "Ueber Die magnetische polarisation der metalle und erze durch temperatur-differenz," Ann. Phys. **82**, 133–160 (1826).
[4]P. M. Roget, *Treatises on Electricity, Galvanism, Magnetism, and Electro-Magnetism* (Baldwin and Cradock, London, 1832).
[5]O. Caballero-Calero, J. R. Ares, and M. Martín-González, "Environmentally friendly thermoelectric materials: High performance from inorganic components with low toxicity and abundance in the Earth," Adv. Sustainable Syst. **5**(11), 2100095 (2021).
[6]R. Freer and A. V. Powell, "Realising the potential of thermoelectric technology: A roadmap," J. Mater. Chem. C **8**, 441–463 (2020).
[7]Z. Tie-Jun, "Recent advances in thermoelectric materials and devices," J. Inorg. Mater. **34**, 233 (2019).
[8]C. Gayner and K. K. Kar, "Recent advances in thermoelectric materials," Prog. Mater. Sci. **83**, 330–382 (2016).
[9]B. Poudel, Q. Hao, Y. Ma, Y. Lan, A. Minnich, B. Yu, X. Yan, D. Wang, A. Muto, D. Vashaee *et al.*, "High-thermoelectric performance of nanostructured bismuth antimony telluride bulk alloys," Science **320**, 634–638 (2008).
[10]G. Tan, F. Shi, S. Hao, L.-D. Zhao, H. Chi, X. Zhang, C. Uher, C. Wolverton, V. P. Dravid, and M. G. Kanatzidis, "Non-equilibrium processing leads to record high thermoelectric figure of merit in PBTE–SrTe," Nat. Commun. **7**, 12167 (2016).
[11]F. Gascoin, S. Ottensmann, D. Stark, S. M. Haïle, and G. J. Snyder, "Zintl phases as thermoelectric materials: tuned transport properties of the compounds $Ca_xYb_{1-X}Zn_2Sb_2$," Adv. Funct. Mater. **15**, 1860–1864 (2005).
[12]Z. Chang, J. Ma, K. Yuan, J. Zheng, B. Wei, M. Al-Fahdi, Y. Gao, X. Zhang, H. Shao, M. Hu *et al.*, "Zintl phase compounds $Mg_3Sb_{2-x}Bi_x$ (x = 0, 1, and 2) monolayers: Electronic, phonon and thermoelectric properties from ab initio calculations," Front. Mech. Eng. **8**, 876655 (2022).
[13]H. Hohl, A. P. Ramirez, W. Kaefer, K. Fess, C. Thurner, C. Kloc, and E. Bucher, "A new class of materials with promising thermoelectric properties: MNiSn (M = Ti, Zr, HF)," MRS Proc. **478**, 109–114 (1997).
[14]F. G. Aliev, V. V. Kozyrkov, V. V. Moshchalkov, R. V. Scolozdra, and K. Durczewski, "Narrow band in the intermetallic compounds MNiSn (M=Ti, Zr, HF)," Z. Phys. B: Condens. Matter **80**, 353–357 (1990).
[15]H. Liu, X. Shi, F. Xu, L. Zhang, W. Zhang, L. Chen, Q. Li, C. Uher, T. Day, and G. J. Snyder, "Copper ion liquid-like thermoelectrics," Nat. Mater. **11**, 422–425 (2012).
[16]M. Zhou, G. J. Snyder, L. Li, and L.-D. Zhao, "Lead-free tin chalcogenide thermoelectric materials," Inorg. Chem. Front. **3**, 1449–1463 (2016).
[17]G. S. Nolas, J. L. Cohn, G. A. Slack, and S. B. Schujman, "Semiconducting GE clathrates: Promising candidates for thermoelectric applications," Appl. Phys. Lett. **73**, 178–180 (1998).
[18]T. Caillat, J.-P. Fleurial, and A. Borshchevsky, "Bridgman-solution crystal growth and characterization of the skutterudite compounds $CoSb_3$ and $RhSb_3$," J. Cryst. Growth **166**, 722–726 (1996).
[19]K. Kurosaki, H. Muta, M. Uno, and S. Yamanaka, "Thermoelectric properties of $NaCo_2O_4$," J. Alloys Compd. **315**, 234–236 (2001).
[20]R. Tian, T. Zhang, D. Chu, R. Donelson, L. Tao, and S. Li, "Enhancement of high temperature thermoelectric performance in Bi, Fe Co-doped layered oxide-based material $Ca_3Co_4O_{9+\delta}$," J. Alloys Compd. **615**, 311–315 (2014).
[21]C. Zhou, Y. K. Lee, Y. Yu, S. Byun, Z.-Z. Luo, H. Lee, B. Ge, Y.-L. Lee, X. Chen, J. Y. Lee *et al.*, "Polycrystalline SnSe with a thermoelectric figure of merit greater than the single crystal," Nat. Mater. **20**, 1378–1384 (2021).
[22]M. Al-Fahdi, T. Ouyang, and M. Hu, "High-throughput computation of novel ternary B–C–N structures and carbon allotropes with electronic-level insights into superhard materials from machine learning," J. Mater. Chem. A **9**, 27596–27614 (2021).
[23]J. Ojih, M. Al-Fahdi, A. D. Rodriguez, K. Choudhary, and M. Hu, "Efficiently searching extreme mechanical properties via boundless objective-free exploration and minimal first-principles calculations," npj Comput. Mater. **8**, 143 (2022).
[24]W. Xia, M. Sakurai, B. Balasubramanian, T. Liao, R. Wang, C. Zhang, H. Sun, K.-M. Ho, J. R. Chelikowsky, D. J. Sellmyer *et al.*, "Accelerating the discovery of novel magnetic materials using machine learning–guided adaptive feedback," Proc. Nat. Acad. Sci. U. S. A. **119**, e2204485119 (2022).
[25]S. K. Kauwe, J. Graser, A. Vazquez, and T. D. Sparks, "Machine learning prediction of heat capacity for solid inorganics," Integr. Mater. Manuf. Innov. **7**, 43–51 (2018).

[26]G. Qin, Y. Wei, L. Yu, J. Xu, J. Ojih, A. Rodriguez, H. Wang, Z. Qin, and M. Hu, "Predicting lattice thermal conductivity from fundamental material properties using machine learning techniques," J. Mater. Chem. A **11**, 5801–5810 (2023).
[27]Y. Zhuo, A. Mansouri Tehrani, and J. Brgoch, "Predicting the band gaps of inorganic solids by machine learning," J. Phys. Chem. Lett. **9**, 1668–1673 (2018).
[28]F. Ricci, W. Chen, U. Aydemir, G. J. Snyder, G.-M. Rignanese, A. Jain, and G. Hautier, "An *ab initio* electronic transport database for inorganic materials," Sci. Data **4**, 170085 (2017).
[29]M. Yao, Y. Wang, X. Li, Y. Sheng, H. Huo, L. Xi, J. Yang, and W. Zhang, "Materials informatics platform with three dimensional structures, workflow and thermoelectric applications," Sci. Data **8**, 236 (2021).
[30]K. Choudhary, K. F. Garrity, and F. Tavazza, "Data-driven discovery of 3D and 2D thermoelectric materials," J. Phys.: Condens. Matter **32**, 475501 (2020).
[31]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).
[32]Y. Sheng, Y. Wu, J. Yang, W. Lu, P. Villars, and W. Zhang, "Active learning for the power factor prediction in diamond-like thermoelectric materials," npj Comput. Mater. **6**, 171 (2020).
[33]J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, "Materials design and discovery with high-throughput density functional theory: The Open Quantum Materials Database (OQMD)," JOM **65**, 1501–1509 (2013).
[34]S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, "The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies," npj Comput. Mater. **1**, 15010 (2015).
[35]G. Kresse and D. Joubert, "From ultrasoft pseudopotentials to the projector augmented-wave method," Phys. Rev. B **59**, 1758–1775 (1999).
[36]G. Kresse and J. Furthmüller, "Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set," Phys. Rev. B **54**, 11169–11186 (1996).
[37]J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," Phys. Rev. Lett. **77**, 3865–3868 (1996).
[38]A. Rodriguez, C. Lin, H. Yang, M. Al-Fahdi, C. Shen, K. Choudhary, Y. Zhao, J. Hu, B. Cao, H. Zhang *et al.*, "Million-scale data integrated deep neural network for phonon properties of Heuslers spanning the periodic table," npj Comput. Mater. **9**, 20 (2023).
[39]A. Rodriguez, C. Lin, C. Shen, K. Yuan, M. Al-Fahdi, X. Zhang, H. Zhang, and M. Hu, "Unlocking phonon properties of a large and diverse set of cubic crystals by indirect bottom-up machine learning approach," Commun. Mater. **4**, 61 (2023).
[40]Y. Zhang, J. Zhang, W. Gao, T. A. Abtew, Y. Wang, P. Zhang, and W. Zhang, "Near-edge band structures and band gaps of Cu-based semiconductors predicted by the modified Becke-Johnson potential plus an on-site Coulomb *U*," J. Chem. Phys. **139**(18), 184706 (2013).
[41]A. M. Ganose, J. Park, A. Faghaninia, R. Woods-Robinson, K. A. Persson, and A. Jain, "Efficient calculation of carrier scattering rates from first principles," Nat. Commun. **12**, 2222 (2021).
[42]W. Li, J. Carrete, N. A. Katcho, and N. Mingo, "Shengbte: A solver of the Boltzmann transport equation for phonons," Comput. Phys. Commun. **185**, 1747–1758 (2014).
[43]F. Zhou, W. Nielson, Y. Xia, and V. Ozoliņš, "Compressive sensing lattice dynamics. I. General formalism," Phys. Rev. B **100**, 184308 (2019).
[44]L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, "A general-purpose machine learning framework for predicting properties of inorganic materials," npj Comput. Mater. **2**, 16028 (2016).
[45]L. Ward, A. Dunn, A. Faghaninia, N. E. R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla *et al.*, "Matminer: An open source toolkit for materials data mining," Comput. Mater. Sci. **152**, 60–69 (2018).
[46]T. Lam Pham, H. Kino, K. Terakura, T. Miyake, K. Tsuda, I. Takigawa, and H. Chi Dam, "Machine learning reveals orbital interaction in materials," Sci. Technol. Adv. Mater. **18**, 756–765 (2017).
[47]K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2016).
[48]T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition* (IEEE, 1995), Vol. 1, pp. 278–282.
[49]J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Stat. **29**(5), 1189–1232 (2001).
[50]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, "Scikit-Learn: Machine learning in python," J. Mach. Learn. Res **12**(85), 2825–2830 (2011); available at https://www.jmlr.org/papers/v12/pedregosa11a.html.
[51]R. G. Babu, A. Nedumaran, G. Manikandan, and R. Selvameena, "Tensorflow: Machine learning using heterogeneous edge on distributed systems," in *Deep Learning in Visual Computing and Signal Processing* (Apple Academic Press, 2022), pp. 71–90.
[52]J. D. Hunter, "Matplotlib: A 2D graphics environment," Comput. Sci. Eng. **9**, 90–95 (2007).
[53]Z. M. Gibbs, F. Ricci, G. Li, H. Zhu, K. Persson, G. Ceder, G. Hautier, A. Jain, and G. J. Snyder, "Effective mass and Fermi surface complexity factor from ab initio band structure calculations," npj Comput. Mater. **3**, 8 (2017).
[54]L. Xi, S. Pan, X. Li, Y. Xu, J. Ni, X. Sun, J. Yang, J. Luo, J. Xi, W. Zhu *et al.*, "Discovery of high-performance thermoelectric chalcogenides through reliable high-throughput material screening," J. Am. Chem. Soc. **140**, 10785–10793 (2018).
[55]R. Li, X. Li, L. Xi, J. Yang, D. J. Singh, and W. Zhang, "High-throughput screening for advanced thermoelectric materials: Diamond-like ABX_2 compounds," ACS Appl. Mater. Interfaces **11**, 24859–24866 (2019).
[56]Y. Zhang, H. Wang, S. Kräemer, Y. Shi, F. Zhang, M. Snedaker, K. Ding, M. Moskovits, G. J. Snyder, and G. D. Stucky, "Surfactant-free synthesis of Bi_2Te_3–Te micro–nano heterostructure with enhanced thermoelectric figure of merit," ACS Nano **5**, 3158–3165 (2011).
[57]P. Janíček, C. Drašar, L. Beneš, and P. Lošťák, "Thermoelectric properties of TL-doped Bi_2Se_3 single crystals," Cryst. Res. Technol. **44**, 505–510 (2009).
[58]Y. Kawaharada, K. Kurosaki, and S. Yamanaka, "Thermophysical properties of Fe_2VAl," ChemInform **34**(40), 909–912 (2003).
[59]K. Kurosaki, A. Kosuga, K. Goto, H. Muta, and S. Yamanaka, "Thermoelectric properties of Ag-Tl-Te ternary system," MRS Proc. **886**, 907 (2005).
[60]G. Tan, L.-D. Zhao, F. Shi, J. W. Doak, S.-H. Lo, H. Sun, C. Wolverton, V. P. Dravid, C. Uher, and M. G. Kanatzidis, "High thermoelectric performance of p-type SnTe via a synergistic band engineering and nanostructuring approach," J. Am. Chem. Soc. **136**, 7006–7017 (2014).
[61]Y. Zhu, Y. Xia, Y. Wang, Y. Sheng, J. Yang, C. Fu, A. Li, T. Zhu, J. Luo, C. Wolverton *et al.*, "Violation of the T^{-1} relationship in the lattice thermal conductivity of Mg_3Sb_2 with locally asymmetric vibrations," Research **2020**, 4589786.
[62]L. Fu, J. Yang, J. Peng, Q. Jiang, Y. Xiao, Y. Luo, D. Zhang, Z. Zhou, M. Zhang, Y. Cheng *et al.*, "Enhancement of thermoelectric properties of YB-filled skutterudites by an Ni-induced "Core–Shell" structure," J. Mater. Chem. A **3**, 1010–1016 (2015).
[63]G. Yang, P. R. Romeo, A. Apostoluk, and B. Vilquin, "First principles study on the lattice thermal conductivity of α-phase Ga_2O_3," J. Vac. Sci. Technol. A **40**(5), 052801 (2022).
[64]H. Wang, J. Wang, X. Cao, and G. J. Snyder, "Thermoelectric alloys between PbSe and PBS with effective thermal conductivity reduction and high figure of merit," J. Mater. Chem. A **2**(9), 3169 (2014).
[65]Y. Zheng, S. Wang, W. Liu, Z. Yin, H. Li, X. Tang, and C. Uher, "Thermoelectric transport properties of p-type silver-doped PbS with *in situ* Ag_2S nanoprecipitates," J. Phys. D: Appl. Phys. **47**(11), 115303 (2014).
[66]S. Wang, G. Zheng, T. Luo, X. She, H. Li, and X. Tang, "Exploring the doping effects of AG in p-type PbSe compounds with enhanced thermoelectric performance," J. Phys. D: Appl. Phys. **44**(47), 475304 (2011).
[67]L. Yang, Z.-G. Chen, M. Hong, L. Wang, D. Kong, L. Huang, G. Han, Y. Zou, M. Dargusch, and J. Zou, "n-type Bi-doped PBTE nanocubes with enhanced thermoelectric performance," Nano Energy **31**, 105–112 (2017).
[68]D. Sarkar, S. Roychowdhury, R. Arora, T. Ghosh, A. Vasdev, B. Joseph, G. Sheet, U. V. Waghmare, and K. Biswas, "Metavalent bonding in GeSe leads to high thermoelectric performance," Angew. Chem. **133**(18), 10438–10446 (2021).

[69]Z. Zheng, X. Su, R. Deng, C. Stoumpos, H. Xie, W. Liu, Y. Yan, S. Hao, C. Uher, C. Wolverton, M. G. Kanatzidis, and X. Tang, "Rhombohedral to cubic conversion of GeTe via MnTe alloying leads to ultralow thermal conductivity, electronic band convergence, and high thermoelectric performance," J. Am. Chem. Soc. **140**(7), 2673–2686 (2018).
[70]K. Imasato, C. Fu, Y. Pan, M. Wood, J. J. Kuo, C. Felser, and G. J. Snyder, "Metallic N-type Mg_3Sb_2 single crystals demonstrate the absence of ionized impurity scattering and enhanced thermoelectric performance," Adv. Mater. **32**(16), 1908218 (2020).
[71]L. Song, J. Zhang, and B. B. Iversen, "simultaneous improvement of power factor and thermal conductivity *via* Ag Doping in p-type Mg_3Sb_2 thermoelectric materials," J. Mater. Chem. A **5**(10), 4932–4939 (2017).
[72]N. S. Chauhan, S. Bathula, A. Vishwakarma, R. Bhardwaj, K. K. Johari, B. Gahtori, M. Saravanan, and A. Dhar, "Compositional tuning of ZrNiSn half-Heusler alloys: Thermoelectric characteristics and performance analysis," J. Phys. Chem. Solids **123**, 105–112 (2018).
[73]H. Xie, H. Wang, C. Fu, Y. Liu, G. J. Snyder, X. Zhao, and T. Zhu, "The intrinsic disorder related alloy scattering in ZRNISN half-Heusler thermoelectric materials," Sci. Rep. **4**(1), 6888 (2014).
[74]A. Rohatgi, see https://automeris.io/WebPlotDigitizer for "WebPlotDigitizer" (accessed February 9, 2024).
[75]W.-D. Liu, L. Yang, and Z.-G. Chen, "Cu_2Se thermoelectrics: Property, methodology, and device," Nano Today **35**, 100938 (2020).
[76]K. Choudhary, B. DeCost, and F. Tavazza, "Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape," Phys. Rev. Mater. **2**(8), 083801 (2018).
[77]Y. Xu, L. Jiang, and X. Qi, "Machine learning in thermoelectric materials identification: Feature selection and analysis," Comput. Mater. Sci. **197**, 110625 (2021).
[78]D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," Science **334**, 1518–1524 (2011).
[79]F. Shao and H. Liu, "The theoretical and experimental analysis of the maximal information coefficient approximate algorithm," J. Syst. Sci. Inf. **9**, 95–104 (2021).
[80]P. Dwivedi, M. Miyata, K. Higashimine, M. Takahashi, M. Ohta, K. Kubota, H. Takida, T. Akatsuka, and S. Maenosono, "Nanobulk thermoelectric materials fabricated from chemically synthesized $Cu_3Zn_{1-x}Al_xSnS_{5-y}$ nanocrystals," ACS Omega **4**, 16402–16408 (2019).
[81]F. Ricci, A. Dunn, A. Jain, G.-M. Rignanese, and G. Hautier, "Gapped metals as thermoelectric materials revealed by high-throughput screening," J. Mater. Chem. A **8**, 17579–17594 (2020).
[82]M. V. Vedernikov, "The thermoelectric powers of transition metals at high temperature," Adv. Phys. **18**, 337–370 (1969).
[83]K. V. Prabu, V. Lourdhusamy, I. Paulraj, M. Sridharan, and C.-J. Liu, "Enhancing the thermoelectric power factor of $Mg_3Sb_2$with SN doping on electronegative sites of SB: Effects of reducing the electronegativity difference," Mater. Chem. Phys. **297**, 127379 (2023).
[84]J. Raty, M. Schumacher, P. Golub, V. L. Deringer, C. Gatti, and M. Wuttig, "A quantum-mechanical map for bonding and properties in solids," Adv. Mater. **31**(3), 1806280 (2018).
[85]Y. Yu, M. Cagnoni, O. Cojocaru-Mirédin, and M. Wuttig, "Chalcogenide thermoelectrics empowered by an unconventional bonding mechanism," Adv. Funct. Mater. **30**(8), 1904862 (2019).
[86]M. Wuttig, V. L. Deringer, X. Gonze, C. Bichara, and J. Raty, "Incipient metals: Functional materials with a unique bonding mechanism," Adv. Mater. **30**(51), 1803777 (2018).
[87]A. Helmenstine, see https://sciencenotes.org/electronegativity-definition-and-trend/ for "Electronegativity definition and trend" (accessed February 14, 2024).
[88]Y. Liu, X. Li, J. Wang, L. Xu, and B. Hu, "An extremely high power factor in Seebeck effects based on a new n-type copper-based organic/inorganic hybrid $C_6H_4NH_2CuBr_2I$ film with metal-like conductivity," J. Mater. Chem. A **5**(26), 13834–13841 (2017).
[89]P. Graziosi, C. Kumarasinghe, and N. Neophytou, "Material descriptors for the discovery of efficient thermoelectrics," ACS Appl. Energy Mater. **3**(6), 5913–5926 (2020).
[90]A. Suwardi, J. Cao, Y. Zhao, J. Wu, S. W. Chien, X. Y. Tan, L. Hu, X. Wang, W. Wang, D. Li, Y. Yin, W.-X. Zhou, D. V. M. Repaka, J. Chen, Y. Zheng, Q. Yan, G. Zhang, and J. Xu, "Achieving high thermoelectric quality factor toward high figure of merit in GeTe," Mater. Today Phys. **14**, 100239 (2020).
[91]R. P. Chasmar and R. Stratton, "The thermoelectric figure of merit and its relation to thermoelectric generators," J. Electron. Control **7**(1), 52–72 (1959).
[92]See https://github.com/Mofahdi for "Github."