npj | Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

Predicting thermoelectric properties from chemical formula with explicitly identifying dopant effects

Gyoung S. Na[1]✉, Seunghun Jang[1] and Hyunju Chang[1]

Dopants play an important role in synthesizing materials to improve target materials properties or stabilize the materials. In particular, the dopants are essential to improve thermoelectic performances of the materials. However, existing machine learning methods cannot accurately predict the materials properties of doped materials due to severely nonlinear relations with their materials properties. Here, we propose a unified architecture of neural networks, called DopNet, to accurately predict the materials properties of the doped materials. DopNet identifies the effects of the dopants by explicitly and independently embedding the host materials and the dopants. In our evaluations, DopNet outperformed existing machine learning methods in predicting experimentally measured thermoelectric properties, and the error of DopNet in predicting a figure of merit (ZT) was 0.06 in mean absolute error. In particular, DopNet was significantly effective in an extrapolation problem that predicts ZTs of unknown materials, which is a key task to discover novel thermoelectric materials.

npj Computational Materials (2021)7:106 ; https://doi.org/10.1038/s41524-021-00564-y

INTRODUCTION

In physical science, various calculation methods to predict materials properties have been developed because the materials properties determine applications of materials[1–3]. However, extensive computation costs of the calculation methods frequently limit the applicability of them in practical applications[4,5]. In particular, the conventional calculation methods are not applicable to the doped materials due to impractical computation costs caused by large cells of the doped materials[6]. For this reason, most experiments to discover novel materials of desired thermoelectric properties have been conducted relying on the intuition of domain experts.

With the rapidly growing public materials databases, machine learning began to be studied widely in physical science to efficiently predict the materials properties[7–9]. In the early stage of materials machine learning, the materials were described as vector-shaped representations based on global characteristics of the materials or statistical information from atomic attributes. Then, conventional machine learning methods (e.g., Gaussian process regression[10]) were applied to predict materials properties based on these vector-shaped representations[7,11]. Recently, advanced machine learning methods that explore structural information of input data, as well as input features, have been studied in physical science to fully utilize structural information from the crystal structures. In particular, graph neural networks (GNNs)[12] have been successfully applied to various scientific applications of physical science because the crystal structure is natively represented as a mathematical graph. In various chemical and physical applications, GNNs have achieved state-of-the-art performances beyond the conventional machine learning methods[8,13,14].

Predicting the materials properties of the doped materials is the next challenge of machine learning in physical science. The doped materials are popular in real-world applications due to their superior performances and stability[15–17]. In particular, the doped materials are dominant in the thermoelectric materials due to their superior thermoelectric performances[18,19]. However, although machine learning has been successfully applied to various scientific applications in physical science, existing machine learning methods are not effective to predict the materials properties of the doped materials. There are three problems in predicting materials properties of the doped materials based on machine learning:

- Lack of information: the crystal structures of the doped materials are not available in most cases because impractical computation costs are required to calculate the crystal structures of the doped materials.
- Dopant effect vanishing: the chemical formula-based materials representations cannot precisely describe the effects of the dopants in the doped materials because numerical changes by the dopants are tiny in the materials representations due to small proportions of the dopants.
- Severely nonlinear relations: relations between the doped materials and their materials properties are severely nonlinear because the dopants sometimes dramatically change the materials properties of the host materials.

Therefore, to accurately predict the materials properties of the doped materials, we need a machine learning method to effectively approximate severely nonlinear functions from the chemical formulas with identifying the dopant effects.

In this paper, we propose a unified architecture of neural networks, called DopNet, to accurately predict the materials properties of the doped materials. DopNet utilizes the chemical formula of the materials to predict materials properties without the crystal structures. To overcome the problems from the dopant effect vanishing and severely nonlinear relations, DopNet explicitly describes the host materials and the dopants. In Discussion Section, we will show that the doped materials can be clearly identified according to their materials properties by explicitly embedding the host materials and the dopants. Another benefit of DopNet is that it does not require additional information about the materials other than the chemical formulas of them. Hence, DopNet can be universally applied to both experimental and calculation materials databases.

[1]Korea Research Institute of Chemical Technology, Daejeon, South Korea. ✉email: ngs0@krict.re.kr

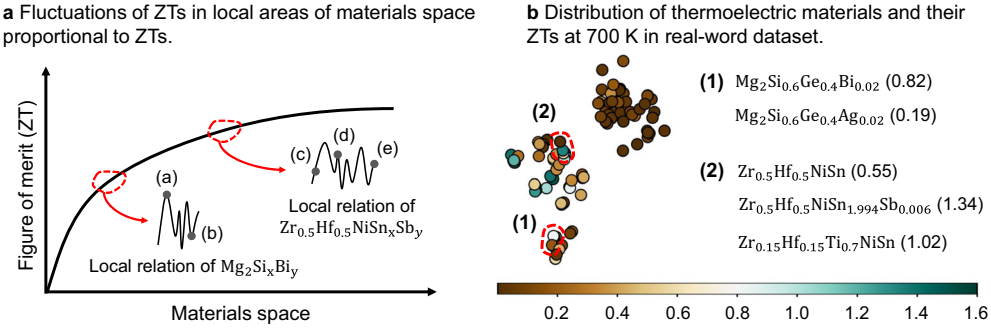

Fig. 1 Mixed distribution of thermoelectirc materials and their ZTs. a Conceptual materials space proportional to ZT at 700 K. Five example doped materials denoted by (a)–(e) are presented to describe severely nonlinear relations between the doped materials and ZT. (a) $Mg_2Si_{0.9985}Bi_{0.0015}$ (ZT = 0.64); (b) $Mg_2Si_{0.999}Bi_{0.001}$ (ZT = 0.32); (c) $Zr_{0.5}Hf_{0.5}NiSn$ (ZT = 0.54). (d) $Zr_{0.5}Hf_{0.5}NiSn_{1.9998}Sb_{0.002}$ (ZT = 1.45); (e) $Zr_{0.5}Hf_{0.5}NiSn_{1.9994}Sb_{0.006}$ (ZT = 1.34). **b** Distribution of doped materials and their ZTs at 700 K in real-world dataset[26]. Each point indicate a material, and they were visualized in 2-dimensional space by t-SNE[27]. For the five example material, ZTs at 700 K are presented in parentheses.

In our evaluations on a real-world materials dataset, DopNet achieved state-of-the-art accuracies in predicting experimentally measured thermoelectric properties of 573 different materials and conditions. Specifically, the prediction error of DopNet in predicting a figure of merit (ZT) was 0.06 in mean absolute error (MAE). In particular, DopNet was significantly more effective than the most popular gradient boosting tree regression (GBTR)[20] in predicting ZTs of completely unseen materials. The errors of GBTR and DopNet in predicting ZTs of these unseen materials were 0.41 and 0.13, respectively, and the performance improvement by DopNet is 68.29%. The significant improvement of DopNet in this extrapolation problem is noteworthy because an accurate prediction for unseen materials is a key task of inverse design. Although we focused on the prediction of the thermoelectric properties in this paper, DopNet can be generally applied to predict any materials properties of the doped materials. For the future works of machine learning in materials science, we publicly open the source code of DopNet at https://github.com/ngs00/DopNet.

RESULTS

Doped materials in regression of materials property

Usually, only a small amount of these dopants are added to the host materials, but the materials properties can be changed drastically[21,22]. Hence, the doped material have severely nonlinear relations with their materials properties in local areas of materials space as shown in Fig. 1a. For instance, the changes from (b) $Mg_2Si_{0.999}Bi_{0.001}$ to (a) $Mg_2Si_{0.9985}Bi_{0.0015}$ is tiny in the entire materials space, but ZT was significantly improved from 0.32 to 0.64[21]. Also, ZTs of (c)-(e) fluctuated from 0.54 to 1.45[22]. However, existing machine learning methods are not suitable to approximate these severely nonlinear relations[23]. Although some machine learning methods to approximate the severely nonlinear functions were proposed[24,25], they require large-scale training datasets, which are impractical in experimental materials databases.

This severely nonlinear relations between the thermoelectric materials and their ZTs are observed in real-world datasets. Figure 1b shows distribution of thermoelectric materials collected from a real-world database[26]. Each point is a material, and the colors of the points indicate the values of ZTs at 700 K. For the five example material, their ZTs at 700 K are presented in parentheses. The materials were visualized in 2-dimensional space by t-SNE[27]. As shown in the figure, the doped materials are highly mixed in terms of their ZTs. As a result, this mixed distribution forms a severely nonlinear relation between the doped materials and their ZTs. It is consistent with our common sense in the conceptual materials space of Fig. 1a. Specifically, $Mg_2Si_{0.6}Ge_{0.4}Bi_{0.02}$ and $Mg_2Si_{0.6}Ge_{0.4}Ag_{0.02}$ are distributed almost the same region despite their completely different ZTs. That is, the effect of the dopants $Bi_{0.02}$ and $Ag_{0.02}$ are not identified. The similar problem was observed in the example materials $Zr_{0.5}Hf_{0.5}NiSn$, $Zr_{0.5}Hf_{0.5}NiSn_{1.994}Sb_{0.006}$, and $Zr_{0.15}Hf_{0.15}Ti_{0.7}NiSn$. In the next section, we propose a neural network for the accurate prediction of thermoelectric properties by explicitly identifying the dopant effects in the doped materials.

Architecture of DopNet

In the existing machine learning methods, to predict materials properties from the chemical formulas, the materials are represented based on the statistical information from the elemental attributes of the atoms in the materials regardless of identifying the dopants[7,11]. However, the atoms in the host material and the dopants are independently embedded in DopNet. This explicit embedding mechanism of DopNet improves the prediction performance for the doped materials by capturing the dopant effects, which are numerically tiny in the materials representations.

DopNet consists of three parts: (1) host embedding networks to extract latent embeddings representing the host materials, (2) dopant embedding networks to generate latent embeddings of the dopants, and (3) dense network to predict target materials property from the embeddings of the host materials and the dopants. Figure 2 illustrates the architecture and forward step of DopNet to predict target materials property **y** from the input chemical formula through four steps:

- The input chemical formula is decomposed into the host material and the dopant(s). Each atom in the material is classified as a dopant when their proportion is less than or equal to γ, where $\gamma \geq 0$ is a pre-defined hyperparameter of DopNet. For instance, $Zr_{0.5}Hf_{0.5}Sn_{1.998}Sb_{0.002}$ is decomposed into a host material $Zr_{0.5}Hf_{0.5}Sn_{1.998}$ and a dopant $Sb_{0.002}$ for a given $\gamma = 0.1$.
- The host material is described as a vector-shaped representation $\mathbf{x}_h$ based on statistical information from the elemental attributes of the constituent atoms. For the host feature vector $\mathbf{x}_h$, an autoencoder $g_\psi(h_\phi(\mathbf{x}_h))$[28] is applied to generate a compact latent embedding of the host material. Then, the host embedding $\mathbf{z}_h$ is calculated by feeding the latent feature vector of the host material into the host embedding network.
- The feature vectors of the dopants are stored in a set S_d that can contain maximum K dopants, where the maximum number of dopants K is a hyperparameter of DopNet. In doped materials including $M < K$ dopants, the $K - M$ undefined dopant feature vectors are set to zero vectors. Then, the dopants are embedded independently of the host material through the dopant embedding networks that share model

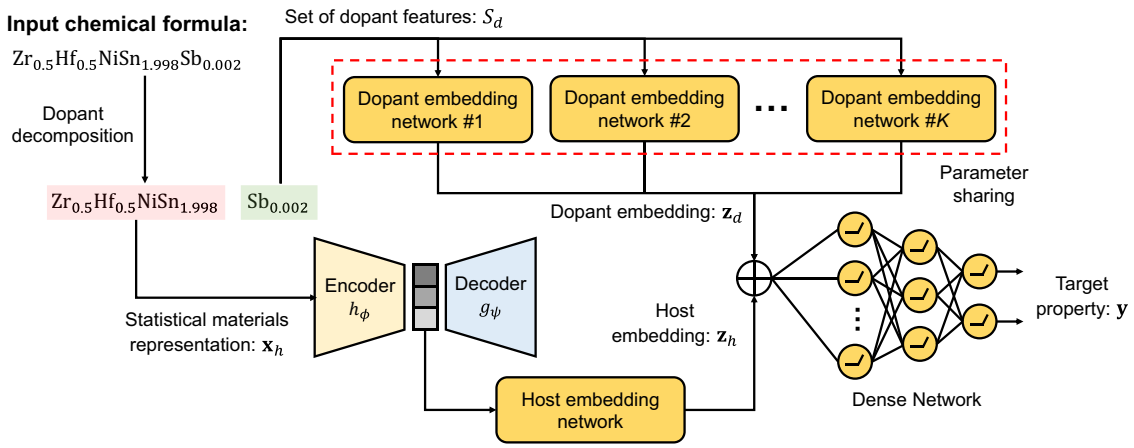

Fig. 2 Architecture of DopNet and its forward process to predict target materials property y **from the chemical formula $Zr_{0.5}Hf_{0.5}Sn_{1.998}Sb_{0.002}$.** The yellow circle and squircle in the network indicates artificial neuron with ReLU activation[52] and ReLU feedforward network, respectively.

parameters with each other. After the embedding process, the generated dopant embeddings are concatenated as a single vector representation $\mathbf{z}_d$.

- The target property **y** is predicted through the dense network by feeding the final materials representation $\mathbf{z}_h \oplus \mathbf{z}_d$.

The hyperparameter settings and network configurations of DopNet are provided in the method section, and the implementation details including selected elemental attributes are presented in Supplementary Table 1.

Prediction of thermoelectric properties

About 70% of the primary energy is lost in the form of heat during the energy conversion process[29]. To utilize the wasted energy, thermoelectric materials that convert heat into electricity by Seebeck effect[30] have been widely studied in physical science[31,32]. The efficiency of the energy conversion process originated from the thermoelectric materials are typically given by a figure of merit (ZT) as:

$$ZT = \frac{S^2 \sigma T}{\kappa}, \quad (1)$$

where S is Seebeck coefficient, σ is electrical conductivity, T is temperature, and κ is thermal conductivity.

Despite the importance of the thermoelectric materials, machine learning to predict thermoelectric performances of the materials was hardly studied in physical science. The main obstacle is the lack of training datasets because simulation and calculation methods are not applicable to estimate the thermoelectric properties of the materials in most cases. However, although there is no dataset for the thermoelectric materials, Materials Research Laboratory (MRL) opened about 500 materials with their thermoelectric properties in their website[26], and we manually collected the chemical formulas of the thermoelectric materials and their thermoelectric properties at the website of MRL to predict the thermoelectric properties of the materials. In this paper, we refer this collected thermoelectric dataset to MRL dataset. This MRL dataset contains 573 thermoelectric materials from various combinations of the host materials and the dopants with several thermoelectric properties measured experimentally at 300 K, 400 K, and 700 K. The collected materials systems in the MRL dataset are summarized in Supplementary Table 2.

In the evaluations, we predicted five thermoelectric properties: Seebeck coefficient, electrical conductivity, thermal conductivity, power factor, and ZT. For prediction, the chemical formulas of the materials were converted into the vector-shaped materials representations based on elemental attributes of the constituent atoms. The selected elemental attributes and the representation method of the materials are provided in Supplementary Note 1. We compared the prediction performance of DopNet with support vector regression (SVR)[33], Gaussian process regression (GPR)[10], gradient boosting tree regression (GBTR)[20], and deep neural network (DNN)[34]. SVR is effective to prevent overfitting due to its margin-based loss formulation. GPR is widely used in scientific applications due to its extrapolation capabilities[35]. GBTR is the most popular method in scientific applications and achieved state-of-the-art performance in various applications[36,37]. The prediction performances were measured by mean absolute error (MAE) and coefficient of determination (R^2 score)[38]. All machine learning methods were evaluated with 3-fold cross-validation, and the evaluation was repeated 10 times. We reported the average of the prediction performances measured by the 10 times repetitions of the evaluations.

Table 1 summarizes the evaluation results of SVR, GPR, GBTR, DNN, and DopNet on the MRL dataset. For all thermoelectric properties, DopNet showed the best prediction performances, as highlighted by the bold fonts. In particular, DopNet outperformed GBTR that showed state-of-the-art performances in various scientific applications[36,37,39,40]. Furthermore, DopNet achieved R^2 score of 0.86 ± 0.02 in predicting ZT that determines the thermoelectric capability of the materials, and its prediction error for ZT was $0.06 \pm 3.00e{-}3$. Figure 3 shows the prediction results of GPR, GBTR, and DopNet for power factor and ZT. Note that the prediction results of SVR are not presented due to its low R^2 scores -4.10 ± 0.34 and 0.17 ± 0.01 in predicting power factor and ZT, respectively. Although GPR and GBTR well predicted the power factors of the materials, there are severe outliers in their prediction results as shown in Fig. 3a. By contrast, the severe outliers were removed in the prediction results of DopNet. In predicting ZT, GPR and GBTR also well predicted the ZTs of the materials. However, they significantly underestimated the ZTs of the high-ZT materials, as highlighted in Fig. 3b. By contrast, DopNet roughly predicted the ZTs of the high-ZT materials.

In addition to power factor and ZT, we also present the prediction results for the transport properties of the materials, as shown in Fig. 4. GPR completely failed to predict the transport properties of the test materials, even though it roughly predicted ZTs. By contrast, GBTR and DopNet accurately predicted Seebeck coefficients of the materials. In particular, many outliers in the prediction results of GBTR were removed as shown in the prediction results of DopNet. However, GBTR and DopNet showed large prediction errors for the materials of the low electrical

Table 1. Prediction errors of SVR, GPR, GBTR, DNN, and DopNet.

Prediction Method	Seebeck coefficient	Electrical conductivity	Thermal conductivity	Power factor	ZT
SVR	148.62 ± 1.75	1464.15 ± 193.08	2.56 ± 0.14	2.44e−3 ± 6.05e−5	0.16 ± 5.00e−3
	(−0.03 ± 0.04)	(−0.02 ± 0.00)	(−0.08 ± 0.01)	(−4.10 ± 0.34)	(0.17 ± 0.01)
GPR	148.62 ± 1.75	2160.61 ± 107.03	2.15 ± 0.06	8.63e−4 ± 1.94e−4	0.15 ± 6.00e−3
	(−0.03 ± 0.04)	(−0.01 ± 0.01)	(0.43 ± 0.04)	(−5.98 ± 7.23)	(0.47 ± 0.04)
GBTR	45.40 ± 1.42	795.96 ± 233.47	1.21 ± 0.09	3.05e−4 ± 1.55e−5	0.07 ± 3.00e−3
	(0.80 ± 0.02)	(0.57 ± 0.27)	(0.55 ± 0.09)	(0.74 ± 0.05)	(0.78 ± 0.03)
DNN	56.53 ± 2.51	1325.92 ± 197.95	1.27 ± 0.06	3.69e−4 ± 1.40e−5	0.09 ± 0.01
	(0.74 ± 0.03)	(0.03 ± 0.10)	(0.53 ± 0.06)	(0.69 ± 0.02)	(0.77 ± 0.02)
DopNet	**39.46 ± 1.34**	**763.66 ± 208.02**	**1.12 ± 0.09**	**2.75e−4 ± 1.15e−5**	**0.06 ± 3.00e−3**
	(**0.86 ± 0.04**)	(**0.64 ± 0.13**)	(**0.61 ± 0.08**)	(**0.79 ± 0.03**)	(**0.86 ± 0.02**)

The prediction errors measured by MAE are reported with their standard deviations. For each machine learning method, R^2 scores are presented in the parenthesis below the prediction errors. The measured Seebeck coefficient, electrical conductivity, thermal conductivity, power factor, and ZT are distributed within [−752.00, 1235.00], [6.90e−5, 1.32e+5], [0.20, 48.70], [1.77e−10, 6.73e−3], and [6.76e−8, 1.60], respectively. The best prediction performance was highlighted by the bold font.

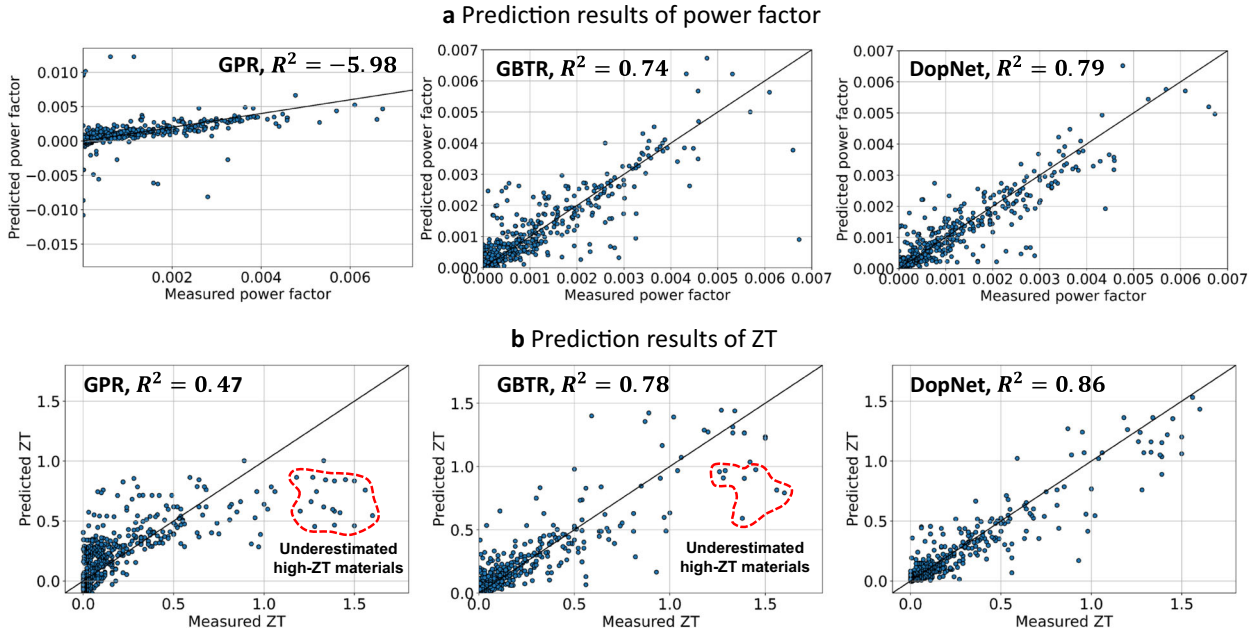

Fig. 3 Prediction results of the machine learning methods on the MRL dataset. a Scatter plots of the prediction results for the test materials in predicting power factor. **b** Scatter plots of the prediction results in predicting ZT. X and Y axes are measured and predicted materials properties, respectively.

conductivities as shown in the yellow areas of Fig. 4b. This happened because the data with small target values are sometimes omitted in the training of ML algorithms. As future work, we can employ a weighted surrogate loss function to reduce the prediction errors for the materials of the low electrical conductivities. In the prediction results of thermal conductivity, GBTR and DopNet showed severe outliers marked by red circles in Fig. 4c. These data are $Zn_{0.9975}Al_{0.0025}O$ at 300 K and 400 K. Experimentally, they have low electrical conductivities less than 3. However, similar materials, such as $Zn_{0.995}Al_{0.005}O$ and $Zn_{0.95}Al_{0.05}O$, have very high electrical conductivities larger than 30. That is, GBTR and DopNet were failed to predict the electrical conductivities of $Zn_{0.9975}Al_{0.0025}O$ at 300 K and 400 K because GBTR and DopNet were overfitted to $Zn_{0.995}Al_{0.005}O$ and $Zn_{0.95}Al_{0.05}O$. This overfitting problem is common in ML and can be solved by collecting larger training datasets.

Prediction of high-ZT materials

The ultimate goal of machine learning in materials science is to discover a novel material, which is called inverse design. For this purpose, an accurate prediction of the high-ZT materials is important because the goal of the inverse design for the thermoelectric materials is to discover a novel material with high ZT. To evaluate the effectiveness of the machine learning methods in the inverse design, we measured the prediction errors of the machine learning methods in predicting the high-ZT materials. Table 2 shows the predicted ZTs of GPR, GBTR, and DopNet for the top 10 high-ZT materials in the MRL dataset. As shown in the table, GPR and GBTR significantly underestimated ZT of the high-ZT materials. The prediction errors of GBTR are 0.12–0.81, and its MAE for the high-ZT materials was 0.45. By contrast, DopNet showed prediction errors lower than 0.5 for all materials, and its MAE for the high-ZT materials was 0.26. Hence, DopNet improved the prediction

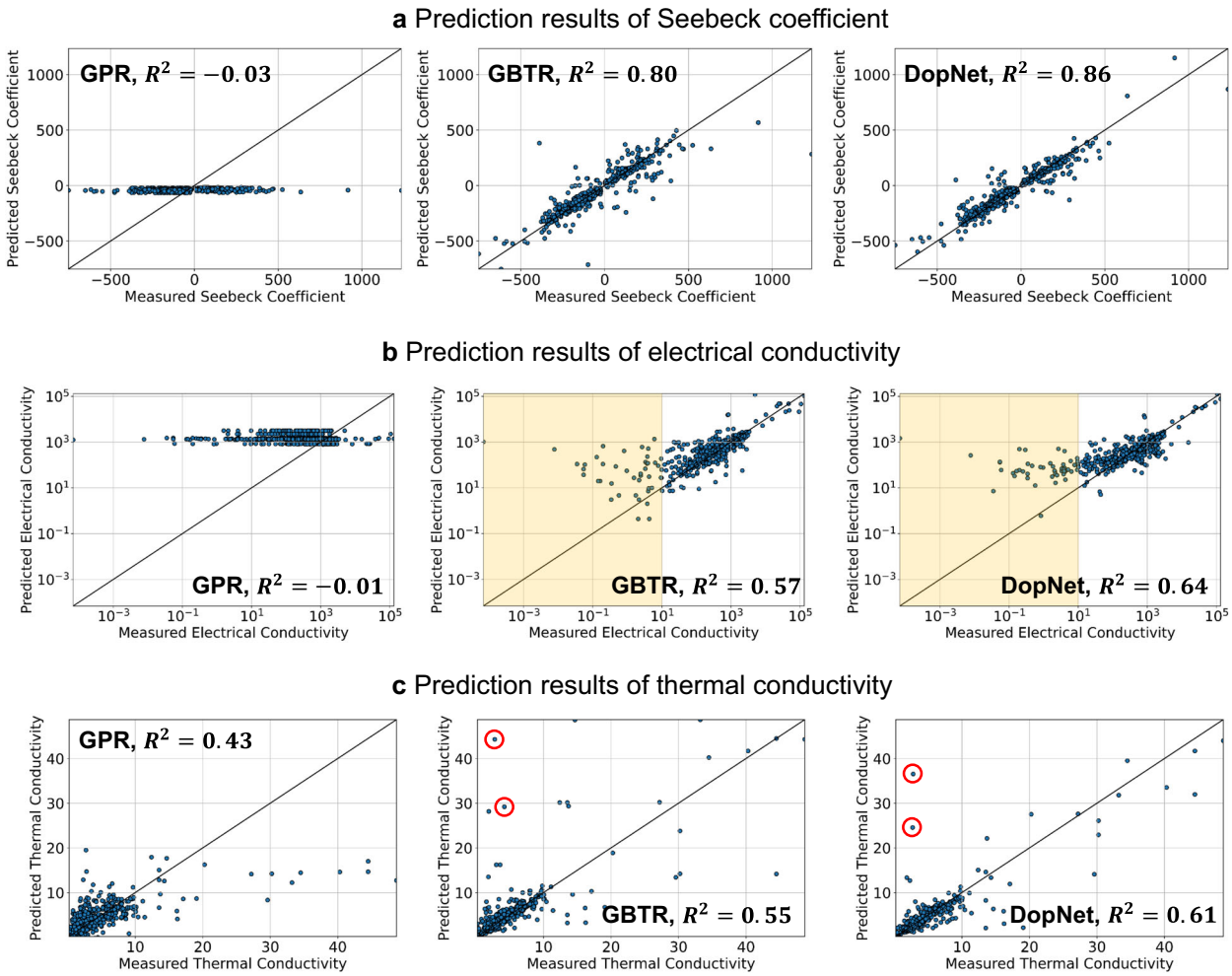

Fig. 4 Regression results of the ML algorithms in predicting Seebeck coefficient, electrical conductivity, and thermal conductivity. **a** Scatter plots of the prediction results for the test materials in predicting Seebeck coefficient. **b** Scatter plots of the prediction results with large error regions for electrical conductivity. **c** Scatter plots of the prediction results with large error data for thermal conductivity. The prediction results of electrical conductivity were presented by log scale. Two marked materials (**a**) and (**b**) in the prediction results of thermal conductivity are $Zn_{0.9975}Al_{0.0025}O$ at 300 K and 400 K, respectively.

Table 2. Predicted ZTs of the machine learning algorithms for top 10 high-ZT materials in the MRL dataset.

Chemical formula	Ground truth (= y)	f_{GP}	f_{GB}	f_{DN}	$\|y - f_{GP}\|$	$\|y - f_{GB}\|$	$\|y - f_{DN}\|$
$Pb_{0.96}Sr_{0.4}TeNa_{0.2}$	1.60[53]	0.55	0.79	1.43	1.05	0.81	0.17
$Pb_{0.98}Sr_{0.2}TeNa_{0.1}$	1.56[53]	0.76	0.81	1.53	0.80	0.75	0.03
$Na_{0.02}PbTe_{0.85}Se_{0.15}$	1.50[54]	0.84	1.24	1.21	0.66	0.26	0.29
$In_{0.25}Co_4Sb_{12}$	1.50[55]	0.46	1.22	1.06	1.04	0.28	0.44
$Zr_{0.5}Hf_{0.5}NiSn_{1.998}Sb_{0.002}$	1.45[22]	0.85	0.98	1.35	0.60	0.47	0.10
$Zr_{0.25}Hf_{0.25}Ti_{0.5}NiSn$	1.42[22]	0.62	1.04	1.22	0.80	0.38	0.20
$Tl_{0.02}Pb_{0.98}Te$	1.39[56]	0.90	0.91	0.91	0.49	0.48	0.48
$Na_{0.02}PbTe_{0.75}Se_{0.25}$	1.39[54]	0.84	1.27	1.15	0.55	0.12	0.24
$In_{0.2}Co_4Sb_{12}$	1.39[55]	0.47	1.26	1.04	0.92	0.13	0.35
$Ag_{0.15}Sb_{0.15}Te_{1.15}Ge_{0.85}$	1.38[57]	0.60	0.59	1.05	0.78	0.79	0.33
Average error (MAE)					0.77	0.45	0.26

ZTs of all materials were measured at 700 K. The predicted values of GPR, GBTR, and DopNet are denoted by f_{GP}, f_{GB}, and f_{DN}, respectively.

accuracy for the high-ZT materials by 42.22% compared to GBTR. This significant improvement by DopNet in predicting high ZT is important and has wide impacts because the accurate prediction of the high-ZT materials is a key task in the machine-based inverse design of the materials.

ZT prediction of unseen materials from external databases

Since the MRL dataset contains ZTs of the same materials for each temperature, some known materials with ZTs measured at different temperatures can be included in the training dataset. To evaluate the machine learning methods in the extrapolation problem, we measured the prediction accuracies of GBTR and DopNet for completely unseen thermoelectric materials. The test thermoelectric materials for this evaluation were collected in previous literature[22,41,42]. ZTs of the collected test materials were measured at 700 K. The test thermoelectric materials for the evaluation can be categorized as:

- Known combination: the combinations of the host atoms and dopants were already provided in the MRL dataset, but the same doping concentrations were not given in the training.
- Unknown combination: both the combinations of the atoms and the doping concentrations are completely unseen in the MRL dataset.

Table 3 summarizes the evaluation results. For the two test cases, GBTR showed relatively large errors in predicting ZTs of the test thermoelectric materials because the tree-based methods are not suitable for the extrapolation problems. In particular, GBTR couldn't capture the changes by the different doping concentrations of Sb in $Zr_{0.25}Hf_{0.25}Te_{0.5}NiSn_{1-x}Sb_x$, and ZTs of these materials are predicted as the same value of 0.32. MAE of GBTR for the materials from the external databases was 0.41. By contrast, DopNet predicted ZT of the test materials more accurately, and the MAE of DopNet was 0.13. Furthermore, while GBTR failed to identify the dopant effects in $Zr_{0.25}Hf_{0.25}Te_{0.5}NiSn_{1-x}Sb_x$ system, DopNet roughly predicted the order of ZTs in the $Zr_{0.25}Hf_{0.25}Te_{0.5}NiSn_{1-x}Sb_x$ system. As shown in the prediction results in Table 3 and the MAE of DopNet, it can be used to roughly investigate the thermoelectric performances of new material before synthesizing it (Table 2).

Hyperparameter analysis

Compared to conventional artificial neural networks, DopNet has two additional hyperparameters denoted by γ and K. The dopant threshold γ determines whether an atom in a given material is an atom in the host material or a dopant. If the proportion of the atom is less than or equal to γ, it is identified as a dopant. Another hyperparameter K defines the maximum number of dopants allowed in DopNet. However, K is automatically determined to cover all doped materials in the dataset for a given γ. Hence, we measured the prediction errors of DopNet as γ changes on the MRL dataset. Figure 5a shows the evaluation results of DopNet for different values of the dopant threshold γ. The prediction errors were measured by MAE. As shown in Fig. 5, DopNet achieved lower errors than GBTR for all values of γ. This evaluation result shows the robustness of DopNet for the hyperparameter γ.

In addition to γ, there are two important hyperparameters in the training of deep neural networks, called initial learning rate and batch size. We also evaluated the prediction errors of DopNet for different initial learning rates and batch sizes in predicting ZT. As shown in Fig. 5b, DopNet showed smaller prediction errors with reasonable choices of the initial learning rates in { 5e−3, 1e−2, 5e−2, 1e−1}. However, the prediction error of DopNet was larger than the errors of GBTR for the initial learning rate of 5e−1 because the gradient descent method to train GBTR was not converged. For the different batch sizes, DopNet always outperformed GBTR as shown in Fig. 5c. As a result, DopNet was robust to the changes in the dopant threshold (γ) and the batch size, and it will enhance the general applicability of DopNet to real-world applications.

In a practical implementation, the dopant threshold γ is an important hyperparameter of DopNet because it determines the host materials and the additives in a given material. However, the dopant threshold should be selected in appropriate ranges.

Table 3. Predicted ZTs of the machine learning algorithms for external test materials of two test cases.

Test case	Chemical formula	Temperature (K)	Ground truth (= y)	f_{GB}	f_{DN}
Known combination	$Na_{0.01}Pb_{0.99}Te$	700	1.22[41]	1.31	1.26
	$Na_{0.02}Pb_{0.98}Te$	700	1.37[41]	1.32	1.29
	$Na_{0.03}Pb_{0.97}Te$	700	1.49[41]	1.32	1.37
Unknown combination	$Pb_{0.95}Ce_{0.05}Te$	300	0.24[42]	0.06	0.09
	$Zn_{0.02}PbTe$	300	0.41[42]	0.22	0.09
	$Pb_{0.95}Ce_{0.05}Te$	673	0.88[42]	1.16	0.98
	$Na_{0.025}Mg_{0.03}Pb_{0.95}Te$	700	1.07[42]	1.29	1.22
	$Na_2TeSr_{0.01}PbTe$	700	1.24[42]	0.43	1.32
	$Mg_{3.05}Nb_{0.15}Sb_{1.5}Bi_{0.49}Te_{0.01}$	673	1.57[42]	0.40	1.66
	$Zr_{0.25}Hf_{0.25}Te_{0.5}NiSn_{0.998}Sb_{0.002}$	700	1.50[22]	0.32	1.27
	$Zr_{0.25}Hf_{0.25}Te_{0.5}NiSn_{0.996}Sb_{0.004}$	700	1.38[22]	0.32	1.25
	$Zr_{0.25}Hf_{0.25}Te_{0.5}NiSn_{0.99}Sb_{0.01}$	700	1.21[22]	0.41	1.24
	$Zr_{0.25}Hf_{0.25}Te_{0.5}NiSn$	700	1.30[22]	0.32	1.27
	$Mg_{3.2}Sb_{1.5}Bi_{0.49}Te_{0.01}$	716	1.50[42]	0.33	1.43
	$PbTeCd_{0.02}$	773	1.50[42]	1.19	1.31
	$Tl_{0.02}Pb_{0.98}Te$	800	1.50[42]	0.93	1.36
	$Ce_{0.1}In_{0.1}Tb_{0.2}Co_4Sb_{12}$	800	1.34[42]	0.64	0.95
	$Ba_{0.06}La_{0.05}Tb_{0.02}Co_4Sb_{12}$	850	1.28[42]	0.63	1.25
Average error (MAE)				0.41	0.13

ZTs of the test materials were measured at 700 K. The predicted ZT of GBTR and DopNet are denoted by f_{GB} and f_{DN}, respectively.

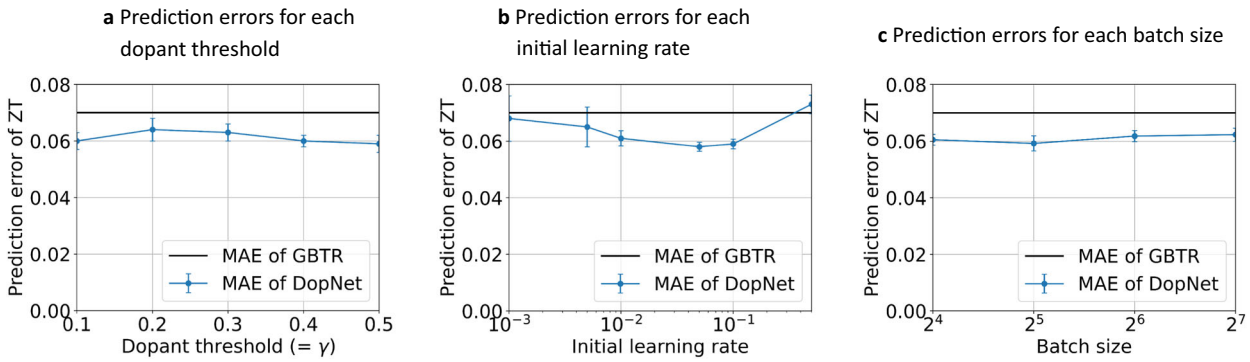

Fig. 5 Prediction errors of DopNet for different values of the hyperparameters. a Prediction errors for different dopant thresholds. **b** Prediction errors for different initial learning rates of the gradient method. **c** Prediction errors for different batch sizes.

For a large dopant threshold, too many elements can be identified as the additives, and the host materials were not defined. To prevent this implementation issue, we propose a rule to select the dopant threshold. Let β is defined as the maximum value of the proportions of the elements in a chemical composition. In the implementation of DopNet, the dopant threshold should satisfy the following inequality for the chemical compositions $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ in a given dataset, where $\mathbf{x}_i$ is a chemical composition.

$$\gamma < \min\{\beta_1, \beta_2, \ldots, \beta_N\} \tag{2}$$

If the dopant threshold does not satisfy the inequality in Eq. (2), some materials are recognized as having no host materials. Thus, the dopant threshold should be selected in $(0, \min\{\beta_1, \beta_2, \ldots, \beta_N\})$ to properly separate the materials into the hosts and the additives.

DISCUSSION

The doped materials are common in thermoelectric materials. Since the materials can have completely different materials properties by a small amount of the dopants, the doped materials usually have mixed distribution for the materials properties, as shown in Fig. 1. In particular, the thermoelectric properties of the materials can be dramatically changed by the dopants. For this reason, predicting the materials properties of the doped materials using machine learning algorithms is challenging because the relations between the doped materials and their materials properties are severely nonlinear. In this paper, we proposed a unified architecture of the neural networks, called DopNet, to accurately predict the thermoelectric properties of the doped materials. DopNet is designed to explicitly and independently representing the host materials and the dopants to identify the effects of the dopants in the entire materials, as illustrated by the architecture of DopNet in Fig. 2. To the best of our knowledge, DopNet is the first machine learning algorithm to predict materials properties from the chemical formulas of the materials by identifying the dopant effects.

We evaluated DopNet in predicting the five thermoelectric properties of various doped materials. For the evaluations, we manually collected the chemical formulas of 573 materials and their thermoelectric properties from the MRL database[26]. As shown in Table 1, DopNet outperformed state-of-the-art machine learning methods in predicting Seebeck coefficient, electrical conductivity, thermal conductivity, power factor, and ZT. In particular, DopNet achieved R^2 scores of 0.79 and 0.86 in predicting power factor and ZT, respectively. Furthermore, DopNet was significantly more effective than GPR and GBTR in predicting the high-ZT materials. Specifically, the MAEs of GPR and GBTR for the top 10 high-ZT materials were 0.77 and 0.45, respectively. However, the MAE of DopNet was 0.26, and the improvement of DopNet is 42.22% compared to GBTR. This improvement of DopNet in predicting high-ZT materials is noteworthy because our ultimate goal is to discover a novel high-ZT material.

Since the MRL dataset contains ZTs of the same material measured at different temperatures, some materials in the test dataset can be shown in the training dataset with different temperatures. To evaluate DopNet in predicting ZTs of completely unseen materials, we predicted ZTs of the materials collected from external databases as shown in Table 3. In this experiment, GBTR was not effective to predict ZTs of the unseen materials, and the MAE of GBTR increased explosively from 0.07 for the test dataset of the MRL dataset to 0.41 for the completely unseen materials. The inaccurate prediction performance of GBTR in the extrapolation limits the applicability of GBTR to real-world applications despite its superior interpolation capabilities. This problem of GBTR in the extrapolation is unsolvable because the functions approximated the tree-based methods are defined only in the ranges of the training datasets. By contrast, the MAE of DopNet increased from 0.06 for the MRL dataset to 0.13 for the unseen materials. That is, DopNet was significantly more effective than the popular GBTR in the extrapolation problem. We will investigate the reason for the performance improvement of DopNet in the extrapolation problem in the next section.

To clarify the reason for the performance improvement of DopNet, we investigated the embedding results of DopNet and compared the embedding results of DopNet with the embeddings of DNN. Note that we did not compare the embedding results with SVR, GPR, and GBTR because they do not generate the latent embeddings of the input data. To visualize the embedding results, t-SNE[27] was applied to the outputs of the last hidden layers of DNN and DopNet. Figure 6 shows the visualization results of the embeddings generated by DNN and DopNet. In the figure, each point is the data (pair of material and temperature) in the MRL dataset, and the colors of the points indicate the measured ZTs. As shown in Fig. 6a, the data is disorderly distributed in the initial stage. After 100 epochs of the training of DopNet, the data was roughly clustered into the low-ZT data and the high-ZT data. Finally, the data was arranged in a direction proportional to the measured ZTs. That is, DopNet generated a latent data representation that makes the regression problem easier. We also compared the embedding results of DopNet with the embeddings of DNN. As shown in Fig. 6b, DNN did not generate a proper data embedding that separates the input data according to their target values. For instance, despite the completely different ZTs of $In_{0.2}Co_4Sb_{12}$ and $In_{0.05}Co_4Sb_{12}$ at 700 K, they were embedded into the same area as shown in the embedding results of DNN. By contrast, they were separately embedded by DopNet. In addition to this case, DNN did not properly represent $Zr_{0.25}Hf_{0.25}Ti_{0.5}NiSn$ and $Zr_{0.5}Hf_{0.25}NiSn$. This embedding result of DNN shows that DNN cannot effectively identify the subtle changes by the dopants

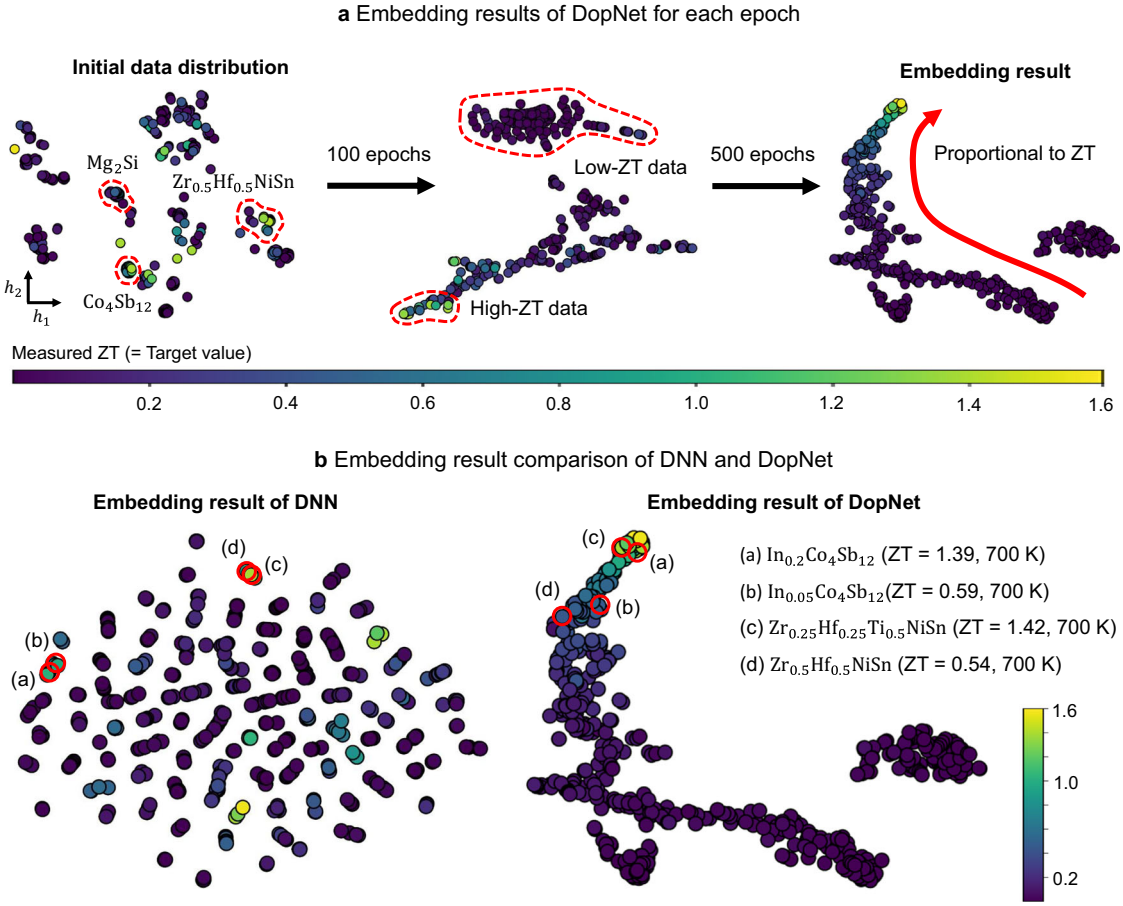

Fig. 6 2-dimensional visualization of embedding results of DNN and DopNet. a embedding results of DopNet for each epoch of the training. **b** embedding results comparison between FNN and DopNet. The embedding results were visualized by t-SNE[27] applied to the outputs of the last hidden layers of FNN and DopNet. Each point is the data (pair of material and temperature) in the MRL dataset, and the colors of the points indicate the values of ZTs. Two axes h_1 and h_2 indicate the first and second latent features calculated by t-SNE, respectively.

in the materials. However, DopNet properly represented the doped materials according to their ZTs because it was specifically designed to identify the dopant effects in the materials, as illustrated in Fig. 2.

From the embedding results, we can also rationalize the performance improvement of DopNet in predicting ZTs of the unseen materials, which is called the extrapolation problem. We can observe that the data is monotonically arranged according to ZTs in the embedding results of DopNet, as shown in Fig. 6b. That is, the linearity of the relation between the materials and their ZTs increased in the embedding space generated by DopNet. Hence, the relation to be approximated the prediction model was linearized by encoding the nonlinearity into the input data of the prediction model. The positive effects of improving the prediction accuracy in the extrapolation problems by the nonlinearity encoding were well investigated theoretically and experimentally in[43].

Recently, Fan et al.[44] calculated the electronic structure, electron relaxation time, and thermoelectric properties for Pb doped Mg_2Si structure, and reported the calculated thermoelectric properties at different temperature. Also, Põhls et al.[45] tried to calculate the thermoelectric properties of $RECuZnP_2$ (RE = Pr, Nd, Er) using sophisticated calculation methods based on ab initio scattering and transport (AMSET) and compressive sensing lattice dynamics. According to their calculation results at 300 K, 400 K, and 700 K, MAE between the experimental and calculation ZTs are 0.12[44,45]. Similarly, DopNet showed the error of 0.13 in predicting ZTs of the completely unseen materials as shown in Table 3. As a result, DopNet achieved the comparable extrapolation capabilities with the traditional calculation methods even though DopNet does not require human labor and extensive computing resources.

We also predicted ZTs of $RECuZnP_2$ (RE = Pr, Nd, Er) at 300 K, 400 K, 700 K. We omitted the Mg_2Si structure because it was contained the training dataset of DopNet. By contrast, the $RECuZnP_2$ (RE = Pr, Nd, Er) system have never been shown in the training dataset. Moreover, Pr and Er have never been shown in any data in the training dataset. For the calculation results of[45], MAE was 0.12 for the $RECuZnP_2$ (RE = Pr, Nd, Er) system. However, DopNet achieved MAE of 0.06 ± 0.01 in 10 times repetitions of the training and prediction processes for the same materials systems. Furthermore, although the calculation methods should be manually modified to improve the prediction accuracies, the prediction accuracy of DopNet can be improved just by collecting more training data. Thus, DopNet can be used as a computation tool to discover novel thermoelectric materials.

DopNet can provide a rapid prediction with reasonable prediction accuracies in discovering new thermoelectric materials. One of the most practical benefits of DopNet is that additional information generated by experimental analyses and density functional theory (DFT)[46] calculations are not required. Thus, DopNet can be used for fast screening in large materials databases or combinatorially generated candidates. This screening method based on DopNet will significantly accelerate the process of discovering novel materials because we can sum up several candidate materials in thousands of materials from combinations of host materials, doped elements, and doping concentrations.

For a trained DopNet $f_\theta(\mathbf{z};\boldsymbol{\theta}^*)$, the computational screening process based on DopNet can be conducted as the following three steps.

- Step 1: The chemical compositions of the candidate materials are generated combinatorially. For instance, the compositions of our target materials system $Tl_aPb_bTe_c$ are generated combinatorially for all possible values of the proportions *a*, *b*, and *c*. Then, the combinatorially generated compositions are validated based on chemical rules, such as valency checking.
- Step 2: The trained DopNet predicts the target materials properties for the generated compositions. After the prediction, the compositions are sorted according to user-defined criteria.
- Step 3: For top *k* materials in the prediction results, domain experts synthesize the selected materials to validate their properties experimentally.

Usually, the first and second steps are finished within an hour, i.e., promising materials can be identified from thousands of candidate materials within an hour. Thus, experimenters can significantly reduce the time required to synthesize thousands of materials to the time required to synthesize only a few *k* materials.

METHODS

Forward process of DopNet

Forward process of DopNet consists of host embedding, dopant embedding, and prediction. For a host feature vector $\mathbf{x}_h$, a latent embedding of the host material is calculated via an the autoencoder and the host embedding network as:

$$\mathbf{z}_h = f_\omega(u_\omega(h_\phi(\mathbf{x}_h))), \tag{3}$$

where h_ϕ is an encoder network of the autoencoder, and f_ω is the host embedding network. Simultaneously, the dopant embedding $\mathbf{z}_d$ is calculated from the set of dopant features S_d via dopant embedding networks as:

$$\mathbf{z}_d = f_\mu(\mathbf{x}_{d_1}) \oplus f_\mu(\mathbf{x}_{d_2}) \oplus \cdots \oplus f_\mu(\mathbf{x}_{d_K}), \tag{4}$$

where $\mathbf{x}_{d_i}$ is a feature vector of *i*th dopant in the input material, f_μ is the dopant embedding network, and $\oplus$ indicates vector concatenation. After generating the latent embeddings of the host material and the dopants, the target materials property $\mathbf{y}$ is predicted via f_θ as:

$$\mathbf{y} = f_\theta(\mathbf{z}_h \oplus \mathbf{z}_d). \tag{5}$$

By the independent embedding processes of the host material and the dopants in Eqs. (3) and (4), DopNet can easily capture the dopant effects in the entire materials. The forward step of DopNet is formally described in Algorithm1.

Algorithm 1: Forward process of DopNet

Input : Input chemical formula c (e.g., $Zr_{0.5}Hf_{0.5}NiSn_{1.998}$)
Output: Predicted target property $\mathbf{y}$

// Decompose input chemical formula into host material and dopants.
$\mathbf{x}_h, S_d = \text{Decompose}(c, \gamma, K)$
// Calculate latent embedding of the host material.
$\mathbf{z}_h = f_\omega(u_\omega(h_\phi(\mathbf{x}_h)))$
// Calculate latent embedding of the dopants.
$\mathbf{z}_d = f_\mu(\mathbf{x}_{d_1}) \oplus f_\mu(\mathbf{x}_{d_2}) \oplus \cdots \oplus f_\mu(\mathbf{x}_{d_K})$
// Predict target property.
$\mathbf{y} = f_\theta(\mathbf{z}_h \oplus \mathbf{z}_d)$
// Return the predicted target materials property.
return y

Model parameter optimization

The materials representations of the host materials are converted into the latent and compact embeddings via autoencoder in DopNet. In the training of DopNet, the autoencoder for the host material and the dopant embedding network are independently trained on the basis of the decomposed materials representations $\mathbf{x}_h$ and S_d. Autoencoders are designed to extract latent features of given data and trained in unsupervised manner by minimizing reconstruction loss. For a given training dataset $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), ..., (\mathbf{x}_N, \mathbf{y}_N)\}$, the training problem of autoencoders are defined by:

$$\phi^*, \psi^* = \arg\min_{\phi,\psi} \frac{1}{N} \sum_{i=1}^{N} ||\mathbf{x}_{h,i} - g_\psi(h_\phi(\mathbf{x}_{h,i}))||_2^2, \tag{6}$$

where *N* is the number of materials in the training dataset. Note that the label data $\mathbf{y}_i$ is not used in the training of the autoencoder. In DopNet, there is no restriction in choosing autoencoder to embed the representations of the host materials. For instance, a probabilistic model of autoencoder[47] can be used in DopNet rather than the traditional autoencoder defined by Eq. (6).

After the training of the autoencoder, the dopant embedding network f_μ and the dense network f_θ is simultaneously trained in supervised manner. For a trained autoencoder $g_\psi(h_\phi(\mathbf{x}_h))$, the host embedding network, the *K* dopant embedding networks, and the dense network are trained by directly minimizing the surrogate loss, such as MAE and root mean square error (RMSE). For instance, the training problem of the networks can be defined based on MAE as:

$$\theta^*, \omega^*, \mu^* = \arg\min_{\theta,\mu,\omega} \frac{1}{N} \sum_{i=1}^{N} |\mathbf{y}_i - f_\theta(\mathbf{z}_{h,i} \oplus \mathbf{z}_{d,i})|, \tag{7}$$

where θ, ω, and μ are the model parameters of the dense, host embedding, and dopant embedding networks, respectively. In the training of the neural networks in DopNet, the dropout technique[48] was applied to improve the generalization capability. Adam optimizer[49] and stochastic gradient descent method with learning rate decay were used to optimize the model parameters of the autoencoder and the other model parameters in DopNet, respectively. Algorithm2 formally describes the training process of DopNet based on the gradient descent methods.

Algorithm 2: Training process of DopNet based on gradient descent method

Input : Training dataset $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), ..., (\mathbf{x}_N, \mathbf{y}_N)\}$,
Dopant threshold $\gamma > 0$,
Maximum number of dopants $K > 0$,
Initial learning rate $\alpha > 0$
Output: Optimized model parameters $\psi^*, \phi^*, \theta^*, \omega^*, \mu^*$

// Train autoencoder.
repeat
  // Decompose input chemical formula into host material and dopants.
  $\mathbf{x}_h, S_d = \text{Decompose}(c, \gamma, K)$
  // Calculate reconstruction loss.
  $L_r = \frac{1}{N} \sum_{i=1}^{N} ||\mathbf{x}_{h,i} - g_\psi(h_\phi(\mathbf{x}_{h,i}))||_2^2$
  // Optimize model parameters of autoencoder.
  $\psi = \psi - \alpha \frac{\partial L_r}{\partial \psi}$
  $\phi = \phi - \alpha \frac{\partial L_r}{\partial \phi}$
until ψ and ϕ converged;
// Train host embedding, dopant embedding, and dense networks.
repeat
  // Calculate latent embedding of the host material.
  $\mathbf{z}_h = f_\omega(u_\omega(h_\phi(\mathbf{x}_h)))$
  // Calculate latent embedding of the dopants.
  $\mathbf{z}_d = f_\mu(\mathbf{x}_{d_1}) \oplus f_\mu(\mathbf{x}_{d_2}) \oplus \cdots \oplus f_\mu(\mathbf{x}_{d_K})$
  // Predict target property.
  $\mathbf{y} = f_\theta(\mathbf{z}_h \oplus \mathbf{z}_d)$
  // Calculate surrogate loss.
  $L_s = \frac{1}{N} \sum_{i=1}^{N} |\mathbf{y}_i - f_\theta(\mathbf{z}_{h,i} \oplus \mathbf{z}_{d,i})|$
  // Optimize model parameters of host embedding, dopant embedding, and dense networks.
  $\omega = \omega - \alpha \frac{\partial L_s}{\partial \omega}$
  $\mu = \mu - \alpha \frac{\partial L_s}{\partial \mu}$
  $\theta = \theta - \alpha \frac{\partial L_s}{\partial \theta}$
until θ, ω, and μ converged;
// Return the optimized model parameters.
return $\psi^*, \phi^*, \theta^*, \omega^*, \mu^*$

Hyperparameter settings

We applied greedy search with validation dataset to select hyperparameters of the machine learning methods. All hyperparameters were set as optimal values that minimize the prediction errors of the validation datasets. Two hyperparameters of SVR, called margin ϵ and regularization coefficient *C*, were selected within {0.01, 0.1, 0.5, 1.0} and {0.1, 0.2, 0.4}, respectively. For GBTR, maximum depth of tree and number of estimators

were searched in {3, 4, 5, 6, 7, 8} and {100, 200, 300, 400}, respectively. The hyperparameters of DopNet were also selected manually. The selected value of the dopant threshold γ and the maximum number of dopants K were set to 5e−1 and 3, respectively. DopNet was trained by stochastic gradient descent (SGD)[50]. The autoencoder of DopNet to embedded the host materials were trained by SGD with the initial learning rate 1e−3, the L_2 regularization coefficient 1e−5, and the batch size 32. The other parts of DopNet (dopant embedding network and dense network) were also trained by SGD with the initial learning rate 1e−1, the L_2 regularization 1e−7, and the batch size 32. For all experiments, the autoencoder of DopNet was defined as fc(256)-fc(64)-fc(256)-fc(n_1), where fc indicates fully-connected layer, the numbers in the fc are the number of output neurons, and n_1 is dimensionality of the host feature $\mathbf{x}_h$. The host and dopant embedding networks of DopNet were defined as simple dense networks with one fc(256). The dense network of DopNet to predict the target materials properties was implemented by three fully-connected layers as fc(512)-fc(16)-fc(1), and the dropout technique was applied to each fc layer.

To convert the chemical formulas into the numerical feature vectors, intrinsic elemental features are assigned for each atom in the materials. These feature vectors of the atoms is converted into a feature vector of a material by calculating statistics of the atomic feature vectors. We calculated average, standard deviation, minimum value, and maximum value of the atomic feature vectors. Total 31 intrinsic elemental features were assigned for each atom, such as atomic number, atomic weight, and electronegativity. Hence, the materials were represented by 124 features from the mean, standard deviation, minimum, and maximum of the 31 atomic attributed of the constituent atoms. Finally, the feature vector of the materials were concatenated with the temperatures at which the thermoelectric properties of the materials were measured. The elemental features used in this paper are available in Python Mendeleev Package[51].

DATA AVAILABILITY

The MRL dataset used in the evaluations was manually collected from http://www.mrl.ucsb.edu:8080/datamine/thermoelectric.jsp. The collected MRL dataset is available at https://github.com/ngs00/DopNet.

CODE AVAILABILITY

The source code of DopNet and the experiment scripts are publicly available at https://github.com/ngs00/DopNet.

Received: 18 March 2021; Accepted: 28 May 2021;
Published online: 14 July 2021

REFERENCES

1. Wang, X.-P. et al. Time-dependent density-functional theory molecular-dynamics study on amorphization of sc-sb-te alloy under optical excitation. *npj Comput. Mater.* **6**, 31 (2020).
2. Tsai, Y.-C. & Bayram, C. Band alignments of ternary wurtzite and zincblende iii-nitrides investigated by hybrid density functional theory. *ACS Omega* **5**, 3917–3923 (2020).
3. Jang, S. et al. First-principles calculation of metal-doped caalsin3: material design for new phosphors. *RSC Adv.* **5**, 39319–39323 (2015).
4. Umari, P., Mosconi, E. & Angelis, F. D. Relativistic GW calculations on $CH_3NH_3PbI_3$ and $CH_3NH_3SnI_3$ perovskites for solar cell applications. *Sci. Rep.* **4**, 4467 (2014).
5. Govoni, M. & Galli, G. Large scale gw calculations. *J. Chem. Theory Comput.* **11**, 2680–2696 (2015).
6. Shim, J., Lee, E.-K., Lee, Y. J. & Nieminen, R. M. Density-functional calculations of defect formation energies using supercell methods: defects in diamond. *Phys. Rev. B* **71**, 035206 (2005).
7. Zhuo, Y., Mansouri Tehrani, A. & Brgoch, J. Predicting the band gaps of inorganic solids by machine learning. *J. Phys. Chem. Lett* **9**, 1668–1673 (2018).
8. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
9. Zhao, Z.-W., del Cueto, M., Geng, Y. & Troisi, A. Effect of increasing the descriptor set on machine learning prediction of small molecule-based organic solar cells. *Chem. Mater.* **32**, 7777–7787 (2020).
10. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)* (The MIT Press, 2005).
11. Lee, J., Seko, A., Shitara, K., Nakayama, K. & Tanaka, I. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Phys. Rev. B* **93**, 115104 (2016).
12. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)* (2017).
13. Wu, Z. et al. Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
14. Morawietz, T. & Artrith, N. Machine learning-accelerated quantum mechanics-based atomistic simulations for industrial applications. *J. Comput. Aided Mol. Des.* **35**, 557–586 (2020).
15. Zitolo, A. et al. Identification of catalytic sites for oxygen reduction in iron- and nitrogen-doped graphene materials. *Nat. Mater.* **14**, 937–942 (2015).
16. Shui, J., Wang, M., Du, F. & Dai, L. N-doped carbon nanomaterials are durable catalysts for oxygen reduction reaction in acidic fuel cells. *Sci. Adv.* **1**, 1–7 (2015).
17. Das Adhikari, S., Guria, A. K. & Pradhan, N. Insights of doping and the photoluminescence properties of mn-doped perovskite nanocrystals. *J. Phys. Chem. Lett.* **10**, 2250–2257 (2019).
18. Pei, Y., Wang, H. & Snyder, G. J. Band engineering of thermoelectric materials. *Adv. Mater.* **24**, 6125–6135 (2012).
19. Wei, J. et al. Review of current high-zt thermoelectric materials. *J. Mater. Sci.* **55**, 12642–12704 (2020).
20. Chen, T. & Guestrin, C. Xgboost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery (2016).
21. Bux, S. K. et al. Mechanochemical synthesis and thermoelectric properties of high quality magnesium silicide. *J. Mater. Chem.* **21**, 12259–12266 (2011).
22. Sakurada, S. & Shutoh, N. Effect of ti substitution on the thermoelectric properties of (zr,hf)nisn half-heusler compounds. *Appl. Phys. Lett.* **86**, 082105 (2005).
23. Tavanaei, A., Ghodrati, M., Kheradpisheh, S. R., Masquelier, T. & Maida, A. Deep learning in spiking neural networks. *Neural Netw.* **111**, 47–63 (2019).
24. Bian, W. & Chen, X. Neural network for nonsmooth, nonconvex constrained minimization via smooth approximation. *IEEE Trans. Neural Netw. Learn. Syst.***25**, 545–556 (2014).
25. Weinberger, K. Q., Blitzer, J. & Saul, L. K. Distance metric learning for large margin nearest neighbor classification. In *Conference on Neural Information Processing Systems (NIPS)* (MIT Press, 2009).
26. Gaultois, M. W. et al. Data-driven review of thermoelectric materials: performance and ressource considerations. *Chem. Mater.* **25**, 2911–2920 (2013).
27. van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
28. Baldi, P. Autoencoders, unsupervised learning and deep architectures. In *Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning Workshop - Volume 27*, UTLW'11, 37-50 (JMLR.org, 2011).
29. Forman, C., Muritala, I., Pardemann, R. & Meyer, B. Estimating the global waste heat potential. *Renew. Sust. Energy Rev.* **57**, 1568–1579 (2016).
30. Seebeck, T. Ueber die magnetische polarisation der metalle und erze durch temperatur-diferenz. *Ann. Phys.* **82**, 133–160 (1826).
31. Snyder, G. J. & Toberer, E. S. Complex thermoelectric materials. *Nat. Mater.* **7**, 105–114 (2008).
32. Julio Gutiérrez Moreno, J., Cao, J., Fronzi, M. & Assadi, M.H.N. A review of recent progress in thermoelectric materials through computational methods. *Mater. Renew. Sustain. Energy* **9**, 16 (2020).
33. Awad, M. & Khanna, R. Support vector regression. *Efficient Learning Machines.* (Springer, 2015).
34. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
35. Wilson, A. G. & Adams, R. P. Gaussian process kernels for pattern discovery and extrapolation. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13 (JMLR.org, 2013).
36. Wang, Z., Zhang, H. & Li, J. Accelerated discovery of stable spinels in energy systems via machine learning. *Nano Energy* **81**, 105665 (2021).
37. Sheridan, R. P., Wang, W. M., Liaw, A., Ma, J. & Gifford, E. M. Extreme gradient boosting as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model.* **56**, 2353–2360 (2016).
38. Draper, N. R. & Smith, H. *Applied Regression Analysis, 3rd ed.* (Wiley-Interscience, 1998).
39. Rothschild, D. et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
40. Zhang, D. et al. A data-driven design for fault detection of wind turbines using random forests and xgboost. *IEEE Acess* **6**, 21020–21031 (2018).
41. Jood, P. et al. Na doping in pbte: solubility, band convergence, phase boundary mapping, and thermoelectric properties. *J. Am. Chem. Soc.* **142**, 15464–15475 (2020).
42. Hasan, M. N., Wahid, H., Nayan, N. & Mohamed Ali, M. S. Inorganic thermoelectric materials: a review. *Int. J. Energy Res.* **44**, 6170–6222 (2020).

43. Xu, K. et al. How neural networks extrapolate: From feedforward to graph neural networks. In *International Conference on Learning Representations* (2021).
44. Fan, T., Xie, C., Wang, S., Oganov, A. R. & Cheng, L. First-principles study of thermoelectric properties of Mg_2Si-Mg_2Pb semiconductor materials. *RSC Adv.* **8**, 17168–17175 (2018).
45. Pőhls, J.-H. et al. Experimental validation of high thermoelectric performance in RECuZnP_2 predicted by high-throughput dft calculations. *Mater. Horiz.* **8**, 209–215 (2021).
46. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133–A1138 (1965).
47. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. In International Conference on Learning Representations (ICLR) (2014).
48. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
49. Kingma, D. P. & Ba, J. L. Adam: A method for stochastic optimization.In *International Conference on Learning Representations (ICLR)* (2015).
50. Bottou, L. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*. (Physica-Verlag HD, 2010).
51. Python mendeleev package. https://github.com/lmmentel/mendeleev (2020). Accessed 12 March 2021.
52. Agarap, A. F. Deep learning using rectified linear units (ReLU). Preprint at https://arxiv.org/abs/1803.08375 (2018).
53. Biswas, K. et al. High-performance bulk thermoelectrics with all-scale hierarchical architectures. *Nature* **489**, 414–418 (2012).
54. Pei, Y. et al. Convergence of electronic bands for high performance bulk thermoelectrics. *Nature* **473**, 66–69 (2011).
55. He, T., Chen, J., Rosenfeld, H. D. & Subramanian, M. A. Thermoelectric properties of indium-filled skutterudites. *Chem. Mater.* **18**, 759–762 (2006).
56. Heremans, J. P. et al. Enhancement of thermoelectric efficiency in pbte by distortion of the electronic density of states. *Science* **321**, 554–557 (2008).
57. Skrabek, E. Properties of the general tags system. In *CRC Handbook of Thermoelectrics*, 267–275 (CRC Press,1995).

ACKNOWLEDGEMENTS

This study was supported by a project from the Korea Research Institute of Chemical Technology (KRICT) [grant number: SI2151-10].

AUTHOR CONTRIBUTIONS

G.S.N. and H.J. supervised the research. G.S.N. and S.J. contributed to design of experiments and G.S.N. conducted experiments. G.S.N. and S.J. wrote the original manuscript and analyzed the results. G.S.N. and S.J. equally contributed this work. All the authors were involved in writing the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-021-00564-y.

Correspondence and requests for materials should be addressed to G.S.N.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.