A Public Database of Thermoelectric Materials and System-Identified Material Representation for Data-Driven Discovery of High-Performance Thermoelectric Materials

Gyoung S. Na and Hyunju Chang
Korea Research Institute of Chemical Technology
Republic of Korea
ngs0@krict.re.kr, hjchang@krict.re.kr

July 7, 2022

Abstract

Thermoelectric materials have received much attention for energy harvesting devices and power generators. However, discovering novel high-performance thermoelectric materials is a challenging task due to the diversity and the structural complexities of the thermoelectric materials containing alloys and dopants. For efficient data-driven discovery of novel thermoelectric materials, we constructed a public dataset that contains experimentally synthesized thermoelectric materials and their experimental thermoelectric properties. In our dataset, we achieved R^2-scores greater than 0.9 in the regression problems for predicting experimentally measured thermoelectric properties of the materials from their chemical compositions. Furthermore, we devised a new material descriptor for the chemical compositions of the materials to improve extrapolation capabilities of machine learning methods. Based on transfer learning with the proposed material descriptor, we greatly improved R^2-score from 0.13 to 0.71 in predicting experimental ZTs of the materials from completely unseen material groups.

1 Introduction

Thermoelectric material is a class of the materials that convert heat energy to electrical energy based on the Seebeck and Peltier effect [1]. These thermoelectric materials have been widely applied to scientific applications, such as energy harvesting [2], thermoelectric cooling [3], and thermopower generators [4]. Recently, thermoelectric materials have received much attention for renewable energy [5]. To discover high-performance thermoelectric materials, various thermoelectric materials were explored around the promising host materials, such as tine selenide [6], silicon-germanium [7], and lead telluride [8]. In particular, various doped materials have been widely studied in thermoelectric materials to improve the thermoelectric performances of the promising host materials [6, 9, 10].

Density functional theory (DFT) [11] have been widely applied to estimate and interpret the relationships between electronic structures and their materials properties in many applications of materials science, such as solar cell materials [12], 2d materials [13], and electrocatalysis [14]. However, although DFT achieved numerous successes as a generally applicable method for the crystalline systems, the applicability of DFT is still limited to the materials of small unit cells due to exponential computational costs of huge unit cells [15, 16]. For this reason, calculating the thermoelectric properties of the doped materials remains a challenging problem in physical science.

In another direction, machine learning has been widely studied to efficiently approximate the relationships between the materials and their physical properties [17, 18]. Several machine learning methods outperformed the conventional calculation- and simulation-based methods in predicting materials properties of the inorganic materials [19, 20]. In particular, graph neural networks (GNNs) [21] have shown remarkable prediction capabilities in various regression tasks on crystal structures [19, 22]. The great successes of GNNs in materials science came from using a crystal graph representation that preserves structural information of the materials as well as their atomic attributes. However, the general applicability of GNNs in real-world applications is still limited because the crystal structures of the doped materials are not available in most cases.

In materials science, several machine learning methods that predict target properties from chemical compositions of the materials have been proposed to extend the general applicability of machine learning to the applications where the crystal structures are not available. Representation learning from stoichiometry (Roost) tried to learn latent embeddings from the chemical compositions of the materials by representing the chemical compositions as an elemental graph [23]. In the experimental evaluations, Roost achieved state-of-the-art accuracies in predicting band gaps from the chemical compositions of the materials. However, its applicability is still limited to the pristine materials because the elemental graph is not defined only for the pristine materials. A lot of technical modifications should be performed to apply the elemental graph to the alloy and doped materials.

Based on a material space embedding approach, DopNet was proposed to predict the physical properties of the alloy and doped materials from their chemical compositions [18]. By separately representing the host materials and the dopants, DopNet was able to learn more informative and latent features of the doped materials and consequently achieved state-of-the-art accuracies in predicting thermoelectric properties of the doped materials. However, the thermoelectric materials have different and complex relationships with their thermoelectric properties for each group of the materials [24], i.e., a simple splitting of the host materials and the dopants cannot fully handle the complexities of the regression problems on the thermoelectric materials. In addition to DopNet, various machine learning approaches have been studied for thermoelectric materials. Neural networks were applied to determine the thermophysical properties of amino acid based ionic liquids [25]. Several ensemble methods were used to predict thermoelectric properties of the materials [26]. However, to the best of our knowledge, there is no public database and material descriptor for data-driven extrapolation to discover novel thermoelectric materials.

To accelerate the data-driven discovery of the thermoelectric materials, we constructed a public materials dataset containing 5,205 experimental observations of the chemical compositions and their thermoelectric properties. We named the collected dataset as experimentally synthesized thermoelectric materials (ESTM) dataset. ESTM dataset covers 880 unique thermoelectric materials and provides five experimentally measured thermoelectric properties: Seebeck coefficient, electrical conductivity, thermal conductivity, power factor, and figure of merit (ZT). In addition, we predicted the thermoelectric properties of the collected materials from their chemical compositions to validate the usefulness of ESTM dataset in machine learning. In machine learning on ESTM dataset, we achieved R^2-scores [27] greater than 0.9 in predicting five thermoelectric properties of the materials, and our prediction model showed a mean

Table 1: Collected features of the data in ESTM dataset. Each data row of the thermoelectric materials is represented as a vector containing string and numerical values. N/A means that this value is not available.

Name	Type	Unit	Range	Mean	Description
Chemical composition	String	N/A	N/A	N/A	Chemical composition of the thermoelectric material
Temperature	Numeric	K	[10, 1275]	539.28±192.42	Temperature in measuring the thermoelectric properties
Seebeck coefficient	Numeric	μV/K	[-1174, 1052.4]	73.18±208.92	Experimentally measured Seebeck coefficient
Electrical conductivity	Numeric	S/m	(0, 9.47E+07)	1.10E+05±1.47E+06	Experimentally measured electrical conductivity
Thermal conductivity	Numeric	W/mK	[0.07, 77.16]	2.25±3.29	Experimentally measured thermal conductivity
Power factor	Numeric	W/mK^2	(0, 7.61E-03)	9.92E-04±1.12E-03	Experimentally measured or theoretically calculated power factor
ZT	Numeric	N/A	(0, 2.28)	0.35±0.35	Experimentally measured or theoretically calculated ZT
DOI	String	N/A	N/A	N/A	Source literature of the collected data row

absolute error (MAE) less than 0.06 in predicting ZTs of the materials.

In addition to dataset construction, we devised a new representation method for the materials, called system-identified material descriptor (SIMD), to accurately predict the target physical properties of the thermoelectric materials from their chemical compositions. SIMD makes a material cluster by collecting similar materials based on the chemical and physical attributes of the materials. Then, SIMD characterizes the relationships between the clustered materials and their target properties for each material cluster based on the least-square method to solve the system of equations. By SIMD, each material cluster is summarized as a vector that can be used to the input data of the machine learning algorithms.

Based on transfer learning with SIMD, we were able to improve R^2-score from 0.13 to 0.71 in an extrapolation problem for ZT prediction (Section 2.4.1), which is a key problem for the data-driven discovery of high-performance thermoelectric materials. We conducted a data-driven search of the materials space to evaluate the usefulness of ESTM and SIMD in real-world applications for material discovery. In the experiments of materials discovery, a machine learning model based on SIMD showed a screening accuracy of 0.61 measured by F1-score [28] as shown in Section 2.4.1, even though the high-throughput screening was conducted on the thermoelectric materials from completely-unknow material systems. Furthermore, we conducted a data-driven search based on SIMD to discover high-ZT materials under the temperature constraints, and the machine learning model with SIMD reduced the false positive by 50% for all search tasks. ESTM dataset and all resources of SIMD with the search results are publicly available at ———.

2 Results

2.1 A Public Dataset of Experimentally Synthesized Thermoelectric Materials

For data-driven discovery of the thermoelectric materials, we performed the literature search to collected the chemical compositions and the experimentally measured thermoelectric properties of the materials. We collected 5,205 experimental observations that are uniquely defined by a pair of the chemical composition and the measuring temperature. Each observation contains five target thermoelectric properties: Seebeck coefficient, electrical conductivity, thermal conductivity, power factor, and ZT. Table 1 describes the data row of ESTM dataset. The first column of chemical composition presents the chemical composition of the thermolectric materials, and the second column is the measuring temperature of the thermoelectric properties. The first column presents the chemical composition of the collected thermoelectric materials. For machine learning, the chemical composition should be converted to the numerical data. The second column is the measuring temperature of

the thermoelectric properties. The last column is DOI of the source literature of the collected experimental observation. The remaining columns are the thermoelectric properties that were experimentally collected or theoretically calculated from the experimental observations.

ESTM covers 880 unique thermoelectric materials of 65 elements from Li to Bi. The visualized distributions of the elements in ESTM dataset are given in Supplementary Information (SI). The most common elements in ESTM dataset were Se, Sb, and Te, which have been widely studied for high-performance thermoelectric materials [29, 30]. In addition, ESTM contains popular and promising thermoelectric materials and their variants, such as lead tellurides (PbTe), bismuth chalcogenides (Bi_2Te_3), and tin selenide (SnSe). The maximum ZTs of the collected thermoelectric materials at high temperature ($\geq$ 700 K) and near room temperature ($\approx$ 300 K) were 2.16 and 1.17, respectively.

2.2 Machine Learning Interpolation for Predicting Thermoelectric Properties

For the collected ETML dataset, we trained machine learning models to predict the thermoelectric properties from the chemical compositions of the materials. We predicted four experimentally measured thermoelectric properties: Seebeck coefficient, electrical conductivity, thermal conductivity, and ZT. In the experiment, we evaluated six different machine learning methods as follows.

- Ridge regression (RidgeReg) [31]: It is a baseline linear regression model with weight regularization by the L2-norm.

- K-nearest neighbor regression (KNNR) [32]: KNNR predicts the target value of the input data by interpolating the target values associated with K nearest neighbor data in the training dataset.

- Support vector regression (SVR) [33]: SVR is a variant of support vector machine for the regression problems. It employs kernel methods to capture the nonlinear relationships between the input and the target data.

- Gaussian process regression (GPR) [34]: GPR is a regression model of a Gaussian process. GPR assumes that the input variables are multivariate Gaussian random variables, and they are drawn from a multivariate normal distribution.

- Fully-connected neural network (FCNN) [35]: FCNN is a feedforward neural network to approximate the relationships between the vector-shaped inputs and the scalar targets. We stacked two hidden layers between the input and the output layers of FCNN to extract latent and nonlinear information from the input data.

- XGB [36]: XGB is an ensemble method that integrates multiple week prediction models to improve the prediction and generalization capabilities. XGB employs

Table 2: MAEs and R^2-scores of the six machine learning methods in the interpolation problems to predict the thermoelectric properties on ESTM dataset. The mean and the standard deviation of the evaluation metrics are presented for each machine learning method. The bold font highlights the best prediction performance for each target property. Note that N/A means the negative R^2-score that indicates the failure of the machine learning method in predicting the thermoelectric property of the materials.

| Target Property (Unit)	Evaluation Metric	RidgeReg	KNNR	SVR	GPR	FCNN	XGB
Seebeck coefficient (μV/K)	MAE	128.57±0.31	101.55±3.31	N/A	N/A	62.41±2.40	**21.10±0.48**
	R^2-score	0.31±0.01	0.47±0.04	N/A	N/A	0.74±0.02	**0.96±0.01**
Electrical conductivity (S/m)	MAE	1.38±0.01	1.16±0.03	N/A	N/A	0.81±0.03	**0.28±0.02**
	R^2-score	0.33±0.02	0.42±0.04	N/A	N/A	0.70±0.03	**0.92±0.01**
Thermal conductivity (W/mK)	MAE	0.50±0.01	0.43±0.01	0.70±0.01	N/A	0.24±0.01	**0.09±0.01**
	R^2-score	0.46±0.01	0.56±0.02	0.06±0.02	N/A	0.84±0.02	**0.97±0.01**
ZT	MAE	0.21±0.01	0.16±0.01	0.23±0.01	0.19±0.01	0.12±0.01	**0.06±0.01**
	R^2-score	0.35±0.01	0.52±0.02	0.17±0.02	0.41±0.03	0.68±0.03	**0.91±0.01**
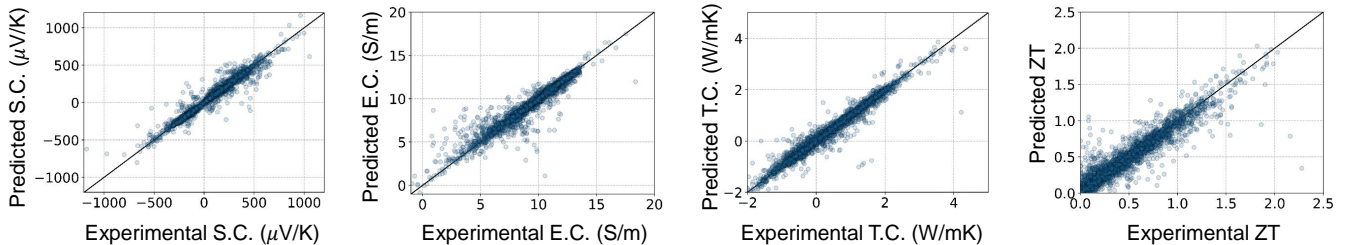
Figure 1: Interpolation results of XGB that was the best prediction model in the interpolation problems to predict the thermoelectric properties of the 5,205 observations in ESTM dataset. Each abbreviation in the axis label means: S.C. (Seebeck coefficient), E.C. (electrical conductivity), and T.C. (thermal conductivity).

a gradient boosting method based on decision trees. XGB has shown state-of-the-art prediction accuracies in many scientific fields, such as organic chemistry and material science.

In the experiments, we used k-fold cross-validation method to train and evaluate the machine learning prediction models. We divided the entire ESTM dataset into three non-duplicated subsets (3-folds). In the training process, two folds of the dataset containing 3,435 observations were used for training of the prediction models, and the remaining subset containing 1,770 observations was used for evaluating the generalization capabilities of the trained prediction models. For each machine learning method, we repeated the training and evaluation processes until were subsets are used for the evaluation. We used the sparse encoding to convert the chemical compositions into machine-readable numerical vectors. The encoding process is described in Section 3 of Supplementary Information. For electrical and thermal conductivity, we applied the logarithm to adjust their high variances.

Table 2 presents the means and the standard deviations of the measured MAEs for the six machine learning methods in the interpolation problems that predict the thermoelectric properties on ESTM dataset. For the comparison in a normalized metric, the mean and the standard deviations of the measured R^2-scores are reported together in the table. In the evaluation, RidgeReg, SVR, and GPR failed to predict the thermoelectric properties for the input chemical compositions and measuring temperatures, and their R^2-scores were less than 0.5 for all prediction tasks. Although KNNR and FCNN showed relatively high R^2-scores over RidgeReg, SVR, and GPR, their prediction capabilities were still limited. By contrast, XGB achieved R^2-scores greater than 0.9 for all prediction tasks. In addition, MAEs of XGB in predicting Seebeck coefficient, electrical conductivity, thermal conductivity, and ZT were 21.10±0.48, 0.28±0.02, 0.09±0.01, and 0.06±0.01, respectively. Fig. 1 shows the prediction results of XGB, and the thermoelectric properties of the materials in ESTM dataset were accurately predicted from the chemical compositions and the measuring temperature. The prediction results of XGB in predicting ZT show the availability of our ESTM dataset for rapid estimation of the experimentally measured thermoelectric performance of the materials in real-world applications.

2.3 System-Identified Material Descriptor

Machine learning methods showed remarkable prediction capabilities in the interpolation problems [17, 19, 22], and we were also able to observe the outstanding interpolation capabilities of machine learning methods on ESTM dataset as shown in Fig. 1. However, it is not sufficient for the data-driven discovery of novel thermoelectric materials because our interest is in the extrapolation that predicts the thermoelectric properties of unexplored material systems. In other words, we should evaluate the prediction capabilities of machine learning methods before using them for the data-driven material discovery. For this reason, we evaluated the prediction capabilities of FCNN and XGB in the extrapolation problems on ESTM dataset by randomly splitting the dataset based on the material groups rather than each material, i.e., the materials in the same material group were entirely removed in the training dataset. In the evaluation of the extrapolation, FCNN and XGB failed to predict ZTs of the materials from unknown material groups, and their R^2-scores were just -0.15 and 0.13, respectively. This failure is natural because conventional machine learning methods are not effective in the extrapolation problems [37, 38].

To improve the extrapolation capabilities of the machine learning methods, we devised a new material representation called system-identified material descriptor (SIMD) that extracts system-level input features for each material group. The overall process of SIMD consists of three steps: (1) material cluster generation, (2) material cluster characterization, and (3) system-identified feature generation. Fig. 2 illustrates the overall process of SIMD to generate the system-conditioned material representations for input tabular data of the materials. In the following subsections 2.3.1-2.3.3, we will formally describe each step of SIMD to generate the system-conditioned material representations.

2.3.1 Material Cluster Generation

The purpose of material cluster generation is to construct the material clusters that cover alloy and doped materials derived from the same pristine materials. To this end, we define a cluster identifier that uniquely represent the material cluster. Formally, the cluster identifier for an input chemical composition s is defined as:

$$id(s) = \{(e, \texttt{round}(r_e)) \mid e \in s \text{ and } r_e > 0.5\}, \quad (1)$$

where e is the symbol of the element in s, r_e is the ratio of the element represented by e, and $\texttt{round}()$ is the
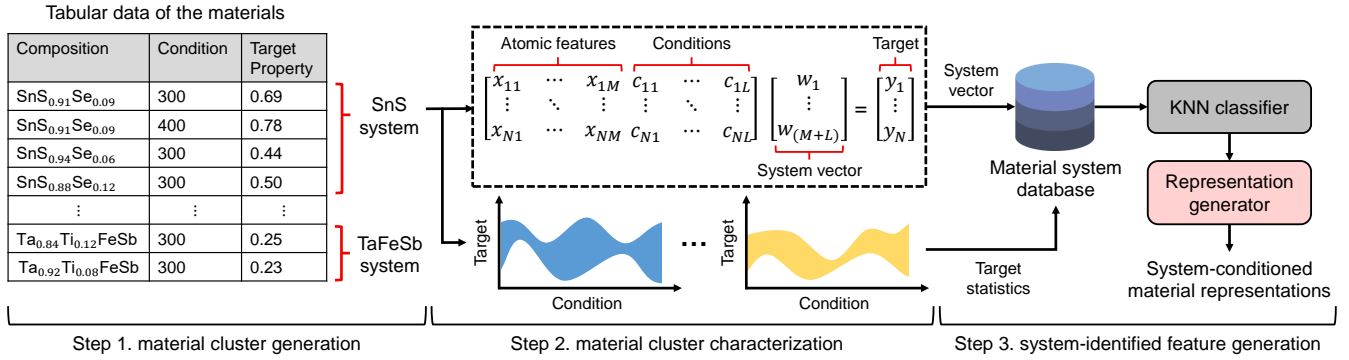
Figure 2: The overall process of SIMD to generate the material representations for an input tabular data of the materials. SIMD calculates the materials representations through the three steps. Step 1: the input materials are clustered based on their compositions. Step 2: the system vectors and target statistics are calculated for each material cluster. Step 3: the KNN classifier allocates the material cluster for each input material, and the representation generator concatenates the system vectors and the target statistics to the input features of the input material.

mathematical rounding operator to convert floating point values to integer values. For example, the input compositions of $SnS_{1-x}Sn_x$ ($x = 0.09, 0.06, 0.12$) and $Ta_{1-x}Ti_xFeSb$ ($x = 0.08, 0.12$) generate the material clusters identified by SnS and TaFeSb, respectively. Note that the input engineering and measuring conditions of the material data are ignored in the process of the material clusters construction in order to cluster the input materials based on the chemical and physical attributes of the materials. After generating the material clusters, the input materials are clustered into the generated material clusters based on the compositions of the input materials. For example, the input materials of the compositions $SnS_{0.91}Se_{0.09}$, $SnS_{0.94}Se_{0.06}$, and $SnS_{0.88}Se_{0.12}$ are clustered into the material cluster of SnS regardless of their alloy and doping elements, as shown in Fig. 2. In the next step of the material cluster generation, the latent vector for each constructed material cluster is calculated to describe the material cluster as a machine-readable vector-shaped representation.

2.3.2 Material Cluster Characterization

In this step, we generate a vector representation of the constructed material clusters based on their cluster identifiers. Through this process, we can convert the material clusters described by chemical compositions into the vector-shaped representations that can be used for the input of the machine learning methods. Specifically, we extract two latent information called system vector and target statistics vector to generate the vector-shaped representations of the material clusters. The system vector for a material cluster represents a projection function from the input material space to the target space for the set of the materials in the material cluster, i.e., the system vector describes the relationships between the materials and the target properties in the local material space defined by the material cluster. The target statistics vector is defined by the mean, standard deviation, minimum, and maximum of the target properties of the materials in the material cluster. It briefly presents the distributions of the target properties in the material clusters, as shown in Fig. 2.

To calculate the system vector of the material cluster u, we represent the relationship between the input materials and the target properties in the material cluster as a mathematical system of material-wise equations as:

$$\begin{bmatrix} x_{11} & \cdots & x_{1M} & c_{11} & \cdots & c_{1L} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ b_{|u|1} & \cdots & b_{|u|M} & c_{|u|1} & \cdots & c_{|u|L} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_{|u|} \end{bmatrix}, \quad (2)$$

where $\mathbf{x}_n = [x_{n1}x_{n2}, ..., x_{nM}]$ is a M-dimensional atomic feature vector calculated from the chemical composition of the n-th material in the material cluster, $\mathbf{c}_n = [c_{n1}, c_{n2}, ..., c_{nL}]$ is a L-dimensional condition vector of the n-th material, y_n is the target property of the n-th material, and $d = M + L$ is the dimensionality of the system vector $\mathbf{w} = [w_1, w_2, ..., w_d]$. In this equation system, the system vector $\mathbf{w}$ is the solution of the system, and it can be

efficiently calculated by least-square method [39] in linear algebra. In the implementation of SIMD, the atomic feature vector $\mathbf{x}_n$ can be defined in various ways to transform the chemical composition of the material into the feature vector. We used a sparse encoding for the atomic feature vector $\mathbf{x}_n$, and the formal description of the atomic feature vector is given in the method section.

The target statistics vector $\mathbf{v}$ of the material cluster u is defined as a 4-dimensional vector of the mean, standard deviation, minimum, and maximum of the target properties of the materials in the material cluster. The target statistics vector is formally defined as:

$$\mathbf{v} = [\bar{y}, y_\sigma, y_{min}, y_{max}], \quad (3)$$

where $\bar{y} = \frac{1}{|u|} \sum_{i=1}^{|u|} y_i$ is the mean of target properties, and $y_\sigma = \sqrt{\sum_{i=1}^{|u|} (y_i - \bar{y})^2 / |u|}$ is the standard deviation of the target properties. Note that y_{min} and y_{max} mean the minimum and maximum values among the target values $\{y_1, y_2, ..., y_{|u|}\}$, respectively. As a result, the material clusters are represented as the $(d+4)$-dimensional concatenated vector of the system and target statistics vectors.

2.3.3 System-Identified Feature Generation

The purpose of system-identified feature generation is to convert the chemical compositions of the materials based on their physical attributes and the constructed material clusters. First, we determine the material clusters of the input chemical compositions. To this end, we define an anchor space where the material clusters are defined as anchor vectors corresponding to their cluster identifiers. The anchor vector is defined as an attribute vector based on the chemical attributes of the cluster identifiers and the chemical compositions. The implementation details of the anchor vector are provided in the method section. Then, the KNN classifier selects K nearest material clusters for the input chemical compositions in the anchor space. For the selected K nearest material clusters, the system and target statistics vectors are combined by a distance-weighted sum as:

$$\mathbf{w}_s^{(K)} = \sum_{u \in \mathcal{N}_s} \phi_{u,s} \mathbf{w}^{(u)}, \quad (4)$$

$$\mathbf{v}_s^{(K)} = \sum_{u \in \mathcal{N}_s} \phi_{u,s} \mathbf{v}^{(u)}, \quad (5)$$

where $\phi_{u,s} = r(a_u, a_s) / \sum_{h \in \mathcal{N}_s} r(a_h, a_s)$ is a distance-based weight for the input chemical composition s and a set of its nearest neighbor material cluster $\mathcal{N}_s$, and r is a distance function for the two input vectors. Finally, the system-identified material representation $\mathbf{z}_s$ of the input chemical composition s is calculated as:

$$\mathbf{z}_s = \mathbf{x}_s \oplus \mathbf{c}_s \oplus \mathbf{w}_s^{(K)} \oplus \mathbf{v}_s^{(K)}, \quad (6)$$

where $\oplus$ is the vector concatenation operator. Note that $\mathbf{x}_s$ and $\mathbf{c}_s$ are the atomic feature vector and the input conditions of the input chemical composition s. In machine learning with SIMD, we use the system-identified material representation $\mathbf{z}_s$ rather than the original input $\mathbf{x}_s$ and $\mathbf{c}_s$.
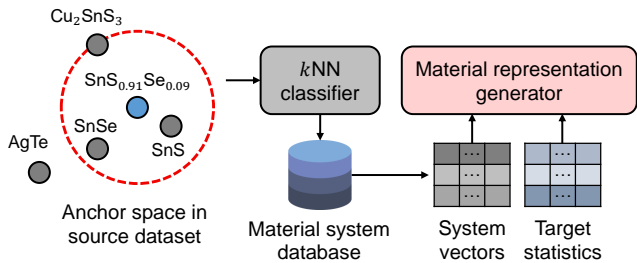
Figure 3: The overall process of SIMD to generate the system-identified features of the input chemical composition in the transfer learning environments.

2.4 Transfer Learning Based on SIMD for Machine Learning Extrapolation

The main feature of SIMD is to generate the vector-shaped representations of the material groups, and it can be used to summarize the large materials datasets. From this perspective, we applied SIMD to transfer learning that aims to transfer knowledge gained from source datasets in solving different but related problems. For the transfer learning on the thermoelectric materials, we used a large source dataset called Starry dataset[1] that contains 215,683 observations of the thermoelectric materials and their ZT. However, although Starry dataset covers extensive thermoelectric materials and their thermoelectric properties, it is not suitable for machine learning due to the following two reasons. (1) The experimentally collected and theoretically calculated thermoelectric materials and their properties are mixed without source labels in Starry dataset, which makes the prediction models unreliable. (2) The parsing errors in the collected thermoelectric properties are inevitable in Starry dataset because the data was automatically collected by a parsing algorithm. For this reason, we used Starray dataset as a source dataset for the transfer learning rather than the training dataset.

Fig. 3 illustrates the overall process of SIMD to generate the system-identified features in the transfer learning environments. The transfer learning based on SIMD is performed through the following four steps.

1. SIMD constructs the material clusters and the system-identified features on the merged dataset of the source and training datasets.

2. The KNN classifier of SIMD determines K nearest material clusters of the input chemical compositions among the material clusters constructed on the source and training datasets.

3. SIMD transforms the original training dataset $\mathcal{D}_{train} = \{(s_1, \mathbf{c}_1, y_1), (s_2, \mathbf{c}_2, y_2), ..., (s_N, \mathbf{c}_N, y_N)\}$ into $\mathcal{Z}_{train} = \{(\mathbf{z}_1, y_1), (\mathbf{z}_2, y_2), ..., (\mathbf{z}_N, y_N)\}$ based on Eq. (6), where $N = |\mathcal{D}_{train}|$ is the number of observations in the training dataset.

4. The prediction model to predict the target property y_n is trained on the transformed training dataset $\mathcal{Z}_{train}$.

As a result, our transfer learning problem based on SIMD is formally defined as an optimization problem as:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \sum_{n=1}^{N} L(y_n, f(\texttt{SIMD}(s_n, c_n, \mathcal{D}_{train}, \mathcal{D}_s); \boldsymbol{\theta})), \quad (7)$$

where L is a loss function to measure the prediction errors, $f(\cdot; \boldsymbol{\theta})$ is a prediction model parameterized by $\boldsymbol{\theta}$, $\texttt{SIMD}()$ is a function to generate $\mathbf{z}_n$ for the input $(s_n, \mathbf{c}_n)$, and $\mathcal{D}_s$ is a source dataset of transfer learning. In the following subsections, we will evaluate the effectiveness of SIMD in transfer learning based extrapolation to predict thermoelectric efficiency of unknown materials.

2.4.1 SIMD-Based Transfer Learning to Extrapolate ZTs of Unknown Material Groups

In this experiment, we conducted machine learning extrapolation to predict the target properties of the materials from

[1]https://www.starrydata2.org/

Table 3: R^2-scores of machine learning methods trained by transfer learning approaches in an extrapolation problem that predicts ZTs for the thermoelectric materials in unknown material groups.

Without Transfer Learning		With Transfer Learning	
Method	R^2-score	Method	R^2-score
FCNN	N/A	FCNN$_f$	0.22±0.07
		FCNN$_d$	0.48±0.10
XGB	0.13±0.07	XGB$_d$	0.52±0.09
		SXGB$_d$	**0.71±0.05**

Figure 4: Confusion matrices of XGB$_d$ and SXGB$_d$ in the high-throughput screening to discover high-ZT (≥ 1.5) thermoelectric materials from unknown material groups. The true label indicates whether the experimentally measured ZT of the material is actually 1.5 or more. On the other hand, the predicted label indicates whether ZT of the material is predicted to be 1.5 or more. Two abbreviations T and F mean the true and false labels, respectively.

unknown material groups. It is essential for material discovery based on machine learning, as we should explore unknown material groups to discover novel materials. To make an extrapolation problem on ESTM dataset, we divided the entire dataset into the training and test datasets based on the material groups, i.e., none of the materials in the test material groups have never been included in the training dataset. For example, if a pristine material SnS is selected as a test material group, all alloy and doped materials derived from SnS (e.g., $SnS_{0.91}Se_{0.09}$ and $SnS_{0.94}Se_{0.06}$) are entirely removed in the training dataset. That is, the prediction models should predict the target properties of the materials that has never been seen in the training dataset, which is called extrapolation problem in machine learning.

To validate the effectiveness of SIMD, we generated four prediction models based on transfer learning approaches as:

- FCNN$_f$: FCNN is pretrained on the source Starray dataset. Then, FCNN is re-trained on the training dataset of ESTM dataset.

- FCNN$_d$: FCNN is trained on the merged training dataset of the source Starry dataset and the training dataset of ESTM dataset.

- XGB$_d$: XGB is trained on the merged training dataset.

- SXGB$_d$: XGB is trained on the merged training dataset transformed by SIMD.

After the training, the extrapolation capabilities of these four transfer learning methods were evaluated on the test dataset $\mathcal{D}_{test}$ that contains completely unseen materials.

Table 3 summarizes the measured R^2-scores of the four transfer learning methods in an extrapolation problem that predicts ZTs of the materials in unknown material groups. As shown in the results, the extrapolation capabilities of all machine learning methods were improved by employing the transfer learning approaches based on Starry dataset. Specifically, FCNN$_d$ and XGB$_d$ showed R^2-scores close to 0.5. However, SXGB$_d$ showed further improvement over the conventional XGB$_d$ and achieved a R^2-score of 0.7. The improvement of SXGB$_d$ in R^2-score was 0.58 and 0.19 compared to R^2-scores of the baseline XGB model and the transfer learning based XGB$_d$ model, respectively. The R^2-score of SXGB$_d$ around 0.70 in the extrapolation environment means that SXGB$_d$ roughly predicted the relationships between the test materials and their experimental ZTs, even

5

a Screening results for the thermoelectric materials with ZT ≥ 0.5 at 290 ≤ T ≤ 310

b Screening results for the thermoelectric materials with ZT ≥ 0.8 at 300 ≤ T ≤ 600

c Screening results for the thermoelectric materials with ZT ≥ 1.5 at 600 ≤ T

Confusion matrix of XGB_d (F1-score: 0.36)

Confusion matrix of $SXGB_d$ (F1-score: 0.65)

Confusion matrix of XGB_d (F1-score: 0.50)

Confusion matrix of $SXGB_d$ (F1-socre: 0.63)

Confusion matrix of XGB_d (F1-score: 0.55)

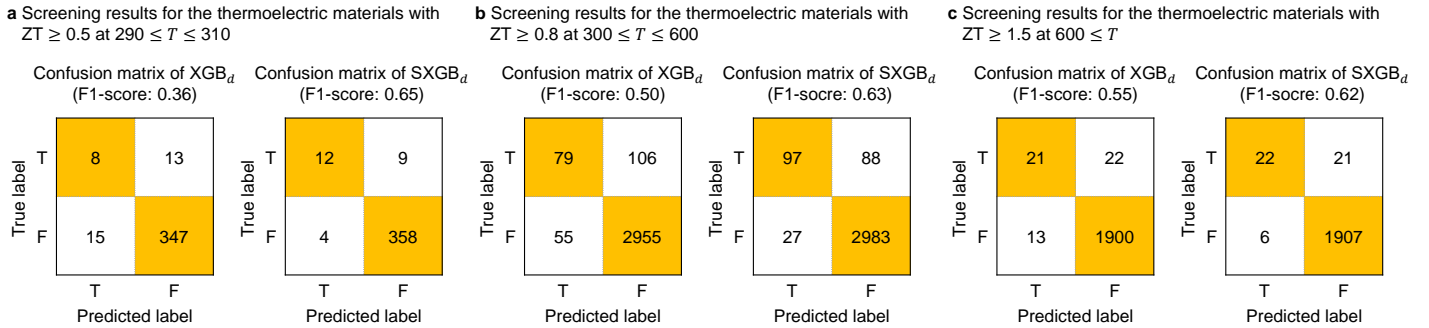Confusion matrix of $SXGB_d$ (F1-socre: 0.62)

Figure 5: Confusion matrices of XGB_d and $SXGB_d$ in the high-throughput screening to discover high-ZT thermoelectric materials for given temperature ranges.

for the test materials that have never been seen in the training dataset. In the following experiments, we will perform high-throughput screening based on $SXGB_d$ to demonstrate the effectiveness of SIMD in real-world applications for discovering novel thermoelectric materials.

2.4.2 High-Throughput Screening for Discovering High-Performance Thermoelectric Materials

The thermoelectric performance defined by ZT essentially determines the efficiency of power generation and energy harvesting in various real-world applications of the thermoelectric materials [2, 3, 5]. To discover novel materials of high ZTs, many experimental analyses and demonstrations have been conducted for various candidate material groups [9, 10, 24]. To validate the effectiveness of SIMD in material discovery, we conducted the high-throughput screening based on $SXGB_d$ for the thermoelectric materials that have never been provided in the training dataset of $SXGB_d$.

We used $SXGB_d$ to predict the experimental ZTs of the materials from their chemical compositions and the given measuring temperatures. The high-throughput screening for discovering high-ZT materials can be defined as a binary classification problem determining whether ZTs of the given materials will actually be greater than the threshold ZT. In this classification problem, the true and false labels indicate whether ZTs of the materials are greater than the threshold ZT. For the high-throughput screening results, we calculated the screening accuracy using F1-score [28] that can comprehensively evaluate the binary classification accuracy based on true positive, false positive, and false negative. Fig. 4 shows the confusion matrices of the binary classification results of XGB_d and $SXGB_d$ in the high-throughput screening for discovering the materials with ZTs of 1.5 or more. $SXGB_d$ achieved an F1-score of 0.61, and the improvement by $SXGB_d$ in F1-score was 0.12 compared to the F1-score of XGB_d. In particular, $SXGB_d$ significantly reduced the number of false positives from 21 to 6. In the high-throughput screening for material discovery, a low false positive is crucial because it guarantees a high probability that the suggested materials will actually have desired properties. In other words, the low false positive can prevent the waste of time and labor to synthesize the materials incorrectly suggested by the prediction models. Quantitatively, the materials suggested by XGB_d were actually the high-ZT materials with a probability of 50.00% ($=100\times21/42$), whereas the materials suggested by $SXGB_d$ were actually the high-ZT materials with a probability of 78.57% ($=100\times22/28$). These high-throughput screening results of $SXGB_d$ show the potential usability of SIMD in the data-driven discovery of high-performance novel thermoelectric materials.

2.4.3 High-Throughput Screening under Temperature Constraints

Since the applications of the thermoelectric materials are mainly categorized by the target temperatures, it is crucial to discover high-performance thermoelectric materials for a given temperature [9, 10, 24]. In this experiment, we evaluated the screening accuracies of XGB_d and $SXGB_d$ in a high-throughput screening for discovering high-ZT materials under the given temperature ranges. We performed the high-throughput screening for three target ranges of the

a Predicted ZTs of Ag-doped $Bi_{0.5}Sb_{1.5}Te_3$

b Predicted ZTs of Ti-doped $Bi_{0.5}Sb_{1.5}Te_3$

— $Bi_{0.5}Sb_{1.5}Te_3$ — $Ag_{0.003}Bi_{0.5}Sb_{1.5}Te_3$
— $Ag_{0.035}Bi_{0.5}Sb_{1.5}Te_3$

— $Bi_{0.5}Sb_{1.5}Te_3$ — $Ti_{0.04}Bi_{0.5}Sb_{1.5}Te_3$
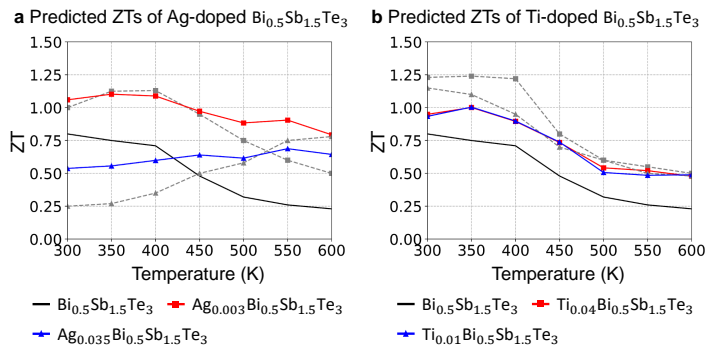— $Ti_{0.01}Bi_{0.5}Sb_{1.5}Te_3$

Figure 6: Experimentally measured and predicted ZTs of Ag- and Ti-doped $Bi_{0.5}Sb_{1.5}Te_3$ materials. The grey dotted lines indicate the experimentally measured ZTs of the materials presented with the same marker.

temperatures in kelvin: (1) near room temperature ($290 \leq T \leq 310$), (2) common thermoelectric temperature ($300 \leq T \leq 600$), and (3) high temperature ($T \geq 600$). For the three target temperature ranges, we searched the materials of ZTs greater than or equal to 0.5, 0.8, and 1.5, respectively.

Fig. 5 shows the confusion matrices of the classification results based on the predicted ZTs of XGB_d and $SXGB_d$ in the high-throughput screening for discovering high ZT materials under the temperature constraints. As shown in the confusion matrices, $SXGB_d$ showed higher F1-scores than XGB_d for all high-throughput screening tasks. The false positive of XGB_d was 65.22% in the high-throughput screening task in Fig. 5-a. By contrast, the false positive of $SXGB_d$ was 25%. In addition, the false positive of XGB_d was 41.04% in the tasks of Fig. 5-b, but the false positive of $SXGB_d$ was 21.77%. Also, the false positives of XGB_d and SGB_d were 38.24% and 21.43% in the task of Fig. 5-c, respectively. That is, we was able to reduce the number of false positive samples by more than 50% for all high-throughput screening tasks by applying the proposed SIMD. As we emphasized before, the low false positive is crucial in machine learning based high-throughput screening because it can prevent the waste of time and labor to synthesize the materials incorrectly suggested by the prediction models.

2.5 Exploration of Virtual Dopant Spaces for Discovering High-ZT Materials

As shown in the experimental results, $SXGB_d$ achieved R^2-score of 0.71 in the extrapolation problem of ZT prediction and showed reliable results in high-throughput screening of high-ZT materials. One of the most beneficial advantages of the extrapolation models is that we can efficiently explore unknown material spaces to discover novel materials without time-consuming experiments and simulations. In this section, we explored virtual dopant spaces using $SXGB_d$ to discover promising dopants for target given host materials. To this end, we generated the virtual dopant spaces for given host materials by concatenating the chemical compositions of the host materials and the candidate dopant elements. For example, we generated candidate materials $Cu_{0.001}SnSe$, $Cu_{0.002}SnSe$, ..., $Cu_{0.1}SnSe$ for a given host material SnSe and a target dopant Cu and. Then, we predicted ZTs of the materials for the target measuring temperatures.

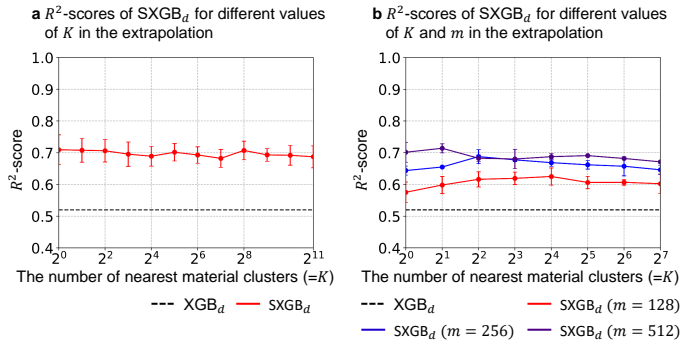We conducted the virtual screening of the dopant space

a R^2-scores of $SXGB_d$ for different values of K in the extrapolation

b R^2-scores of $SXGB_d$ for different values of K and m in the extrapolation

The number of nearest material clusters (=K)

--- XGB_d — $SXGB_d$

--- XGB_d — $SXGB_d$ ($m = 128$)

— $SXGB_d$ ($m = 256$)   — $SXGB_d$ ($m = 512$)

Figure 7: Measured R^2-scores of $SXGB_d$ in the extrapolation problem to predict ZTs for different values of the hyper-parameter and the data size. K is the number of nearest material clusters in Eq. (4) and (5). m is the number of material clusters in the training dataset. For the quantitative comparisons, we present the R^2-score of the baseline XGB_d using the black dotted line.

Figure 8: Distribution of the thermoelectric materials around the $Mg_{1-x}Li_xGe_{0.9}Si_{0.1}$ system in the chemical space based on the sparse encoding of the chemical compositions.

to discover novel dopants for a host material $Bi_{0.5}Sb_{1.5}Te_3$ that was showed promising thermoelectric properties at low temperature. We generated virtual materials by concatenating $Bi_{0.5}Sb_{1.5}Te_3$ and the elements from H to Fm with the doping concentrations in $\{0.001, 0.002, ..., 0.1\}$. In other words, we predicted ZTs of 10^4 candidate virtual materials for the host material $Bi_{0.5}Sb_{1.5}Te_3$. Then, we predicted ZTs of the generated materials at the temperatures in $\{300\ K, 350\ K, ..., 800\ K\}$. After that, we selected top 10% materials based on their predicted ZTs at 300 K.

In the exploration results, most selected materials contained the dopants of Ti, Fe, Ga, Se, and Ag, and we were able to crosscheck the improved thermoelectric performances by Ag and Ti in experiments [40, 41, 42]. We plotted ZTs of Ag- and Ti-doped $Bi_{0.5}Sb_{1.5}Te_3$ for different measuring temperatures, as shown in Fig. 6. The red square and blue triangle lines present the predicted ZTs of the materials that were experimentally reported to have the highest and lowest ZTs in the Ag- and Ti-doped $Bi_{0.5}Sb_{1.5}Te_3$, respectively. The grey dotted lines present the experimentally measured ZTs of the Ag- and Ti-doped $Bi_{0.5}Sb_{1.5}Te_3$. As shown in the results, $SXGB_d$ accurately predicted the promising dopants Ag and Ti that were experimentally demonstrated to improve the thermoelectric performance of $Bi_{0.5}Sb_{1.5}Te_3$. Moreover, $SXGB_d$ captured the tendency of ZTs of the doped materials for different measuring temperatures.

3 Discussion

3.1 Extrapolation Accuracy for Different Number of Nearest Material Clusters

The number of nearest material clusters (defined as K) is an important hyper-parameter of SIMD, as shown in Eqs. (4) and (5). In this section, we evaluated the extrapolation capabilities of SIMD for the changes in the values of K. We measured R^2-scores of $SXGB_d$ in predicting ZTs in the extrapolation problems for different values of K in $\{1, 2, ..., 2048\}$, as shown in Fig. 7-a. Note that the number of material clusters in the training dataset was 2,201 ($\sim$ 2,048). As shown in the results, $SXGB_d$ showed the consistent R^2 scores for the changes in the values of K. We also measured the effectiveness of K in the R^2-scores of $SXGB_d$ for different data sizes. Fig. 7-b shows the measured R^2-scores, and m means the number of material clusters in the training dataset. The small value of m indicates the small training dataset. In the experiment, $SXGB_d$ showed the highest R^2-scores for K of 4 and 8 in the small training datasets containing 128 and 256 material systems (red and blue lines, respectively), which these small training dataset may not contain a material cluster that covers the input material. For these small training datasets, the values of K larger than 1 were helpful because SIMD was able to generate material descriptors by collecting the information from similar material clusters. By contrast, $SXGB_d$ achieved the
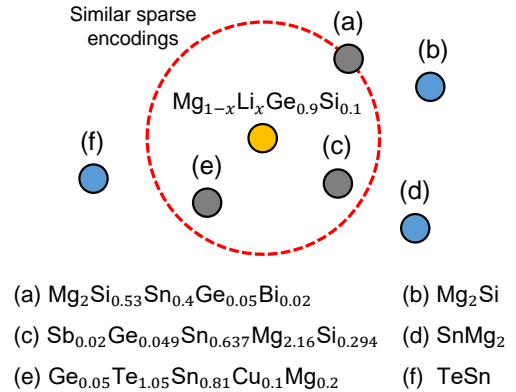
best R^2-score for K of 1 and 2 for more large dataset containing 512 material systems (indigo line). We can interpret this results that $K = 1$ was sufficient because various material systems covering the input materials were contained in the relatively large training dataset. Therefore, hyper-parameter selection of K is not a big problem in the implementation of SIMD if we have a large training dataset. However, when the training dataset is small, a sufficiently large K will help to improve the extrapolation ability of SIMD.

3.2 SIMD for System-Conditioned Prediction

In the high-throughput screening to discover novel thermoelectric materials, SIMD reduced the false positive of the prediction models by about 50% compared to the conventional methods. To reveal physical or chemical insights from the high-throughput screening results, we conducted a case study for the $Mg_{1-x}Li_xGe_{0.9}Si_{0.1}$ system that generated most false positive samples in the baseline XGB_d. Fig. 8 illustrates the distribution of the thermoelectric materials around the $Mg_{1-x}Li_xGe_{0.9}Si_{0.1}$ system in the chemical space based on the sparse encoding of the chemical compositions. In this chemical space, three high-ZT materials denoted by (a), (c), and (e) in Fig. 8 are distributed near the materials from the $Mg_{1-x}Li_xGe_{0.9}Si_{0.1}$ system because they commonly contain two elements Ge and Mg. Basically, the similar feature representations yields the similar target values in machine learning. Hence, although the thermoelectric materials from the $Mg_{1-x}Li_xGe_{0.9}Si_{0.1}$ system actually have the low-ZT (< 0.4), XGB_d incorrectly predicted them as the high-ZT (≥ 1.5) materials because the similarly-encoded materials (a), (c), and (e) in the training dataset have the high-ZT (> 1.4).

By contrast, $SXGB_d$ with SIMD greatly reduced the false positive from the $Mg_{1-x}Li_xGe_{0.9}Si_{0.1}$ system in the high-throughput screening. As shown in Fig. 8, the three materials (a), (c), and (e) belong to the material clusters (b), (d), and (f) in SIMD. The materials from the $Mg_{1-x}Li_xGe_{0.9}Si_{0.1}$ system have different sparse encodings to the materials (b), (d), and (f) because they contain only one element Mg in common or no element in common. Therefore, SIMD can generate the representations of the $Mg_{1-x}Li_xGe_{0.9}Si_{0.1}$ system that are different to high-ZT materials (b), (d), and (f). As a result, the prediction models based on SIMD were able to correctly learn that the thermoelectric materials from the $Mg_{1-x}Li_xGe_{0.9}Si_{0.1}$ system are low-ZT materials.

3.3 Material Space Exploration Based on Global Search Method with SIMD

In the result section, SIMD was successfully applied to discover high-ZT thermoelectric materials based on high-throughput screening. However, we can further extend SIMD to an automated global search method for discovering novel materials in unexplored material space by inte-

grating SIMD with randomized iterative search algorithms called metaheuristic algorithms [43]. In chemical science, metaheuristics have been successfully applied to discover novel molecules and materials [44, 45, 46]. Various metaheuristic optimization algorithms have been proposed based on evolutionary method [47], swarm intelligence strategy [48], and physics-inspired approach [49]. Although existing metaheuristic algorithms employ different optimization schemes, we can integrate SIMD without algorithmic modifications because we can identically define a problem to discover novel materials based on SIMD as a constrained optimization problem as:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x}; \boldsymbol{\theta}^*) + g(\mathbf{x}), \qquad (8)$$

where $\mathbf{x}$ is the sparse encoding of the elements in the chemical composition, $f(\cdot; \boldsymbol{\theta}^*)$ is a trained extrapolation model to predict target material properties from the input $\mathbf{x}$, and $g(\mathbf{x})$ is a penalty term for checking violation of user-defined constraints on discovered materials. In this formulation, $f(\cdot; \boldsymbol{\theta}^*)$ can be implemented by SXGB_d to extrapolate ZTs of the unexplored materials. Furthermore, $g(\mathbf{x})$ can be defined as domain-specific constraints regardless of whether it is differentiable or not. For example, if we focus on the materials containing maximum three elements, we can define the penalty term as:

$$g(\mathbf{x}) = \begin{cases} 1, & \text{if } |\{\mathbf{x}_i \mid \mathbf{x}_i > 0, \forall i\}| \geq 4 \\ 0, & \text{otherwise.} \end{cases} \qquad (9)$$

In addition to the example constraint, we can impose various domain-specific constraints on the chemical characteristics of discovered materials, such as target elements, ranges of target properties, and the number of dopants. Note that if we want to discover the materials maximizing the target properties, the optimization problem in Eq. (8) can be defined as a maximization problem with a negative penalty term. In this study, we conducted the global search using equilibrium optimizer [49] based on SXGB_d to discover novel high-ZT materials, and the search results are provided in SI.

4   Methods

4.1   Sparse Encoding of Chemical Compositions

In the implementation of SIMD in Eqs. (6) and (8), we used the sparse encoding $\mathbf{x}$ to represent the input chemical compositions as the vector-shaped data. Formally, the sparse encoding of the input chemical composition s is defined as:

$$\mathbf{x}_i = \begin{cases} r_e, & \text{if } i = n_e \text{ and } e \in s \\ 0, & \text{otherwise,} \end{cases} \qquad (10)$$

where e is an element in the input chemical composition s, r_e is the ratio of e, and n_e is the atomic number of e. In the implementation, we considered the elements from H to Fm, i.e., the dimensionality of the sparse encoding was 100.

4.2   Anchor Space for Material Clusters Allocation

To determine the material clusters for the input chemical compositions based on the KNN classifier, the cluster identifiers defined by the chemical compositions and the input chemical composition should be converted into compact low-dimensional vectors. To this end, we defined an anchor space where the chemical compositions are represented as compact 12-dimensional vectors. Formally, the chemical compositions of the material clusters and the input materials are defined as an anchor vector as:

$$\mathbf{a} = \mathbf{q}_{mean} \oplus \mathbf{q}_{std} \oplus \mathbf{q}_{min} \oplus \mathbf{q}_{max}, \qquad (11)$$

where $\mathbf{q}_{mean}$ is a 3-dimensional vector of the average atomic numbers, atomic volumes, and atomic weights of the elements in the input chemical composition. Similarly, $\mathbf{q}_{std}$,

$\mathbf{q}_{min}$, and $\mathbf{q}_{max}$ are calculated as the standard deviations, minimum values, and the maximum values of the three atomic attributes of the elements in the input chemical composition, respectively. Based on the anchor vectors of the material clusters and the input materials, we determined the material clusters of the input materials by calculating the Euclidean distances between the anchor vectors of the material clusters and the input materials.

Data Availability

The collected dataset and all resources of the proposed method are publicly available at ———.

Code Availability

All experimental scripts and source codes of SGL are publicly available at ———.

Acknowledgements

References

[1] Rademann, K., Raghuwanshi, V. & Hoell, A. Chapter 3 - crystallization and growth mechanisms of nanostructures in silicate glass: From complete characterization toward applications. In *Glass Nanocomposites*, 89–114 (William Andrew Publishing, 2016).

[2] Nozariasbmarz, A. *et al.* Review of wearable thermoelectric energy harvesting: From body temperature to electronic systems. *Appl. Energy* **258**, 114069 (2020).

[3] Zhao, D. & Tan, G. A review of thermoelectric cooling: materials, modeling and applications. *Appl. Therm. Eng.* **66**, 15–24 (2014).

[4] Tan, G., Ohta, M. & Kanatzidis, M. G. Thermoelectric power generation: from new materials to devices. *Philos. Trans. R. Soc. A* **377**, 20180450 (2019).

[5] Elsheikh, M. H. *et al.* A review on thermoelectric renewable energy: Principle parameters that affect their performance. *Renew. Sustain. Energy Rev.* **30**, 337–355 (2014).

[6] Zhao, L.-D. *et al.* Ultralow thermal conductivity and high thermoelectric figure of merit in snse crystals. *Nature* **508**, 373–377 (2014).

[7] Joshi, G. *et al.* Enhanced thermoelectric figure-of-merit in nanostructured p-type silicon germanium bulk alloys. *Nano Lett.* **8**, 4670–4674 (2008).

[8] Sharma, P. K., Senguttuvan, T., Sharma, V. K. & Chaudhary, S. Revisiting the thermoelectric properties of lead telluride. *Mater. Today Energy* **21**, 100713 (2021).

[9] Yang, L., Chen, Z.-G., Dargusch, M. S. & Zou, J. High performance thermoelectric materials: progress and their applications. *Adv. Energy Mater.* **8**, 1701797 (2018).

[10] Tan, G., Zhao, L.-D. & Kanatzidis, M. G. Rationally designing high-performance bulk thermoelectric materials. *Chem. Rev.* **116**, 12123–12149 (2016).

[11] Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133–A1138 (1965). URL https://link.aps.org/doi/10.1103/PhysRev.140.A1133.

[12] Jacobs, R., Booske, J. & Morgan, D. Understanding and controlling the work function of perovskite oxides using density functional theory. *Adv. Funct. Mater.* **26**, 5471–5482 (2016).

[13] Patra, A. *et al.* Efficient band structure calculation of two-dimensional materials from semilocal density functionals. *J. Phys. Chem. C* **125**, 11206–11215 (2021).

[14] Liao, X. *et al.* Density functional theory for electrocatalysis. *Energy Environ. Mater.* **5**, 157–185 (2022).

[15] Schuch, N. & Verstraete, F. Computational complexity of interacting electrons and fundamental limitations of density functional theory. *Nat. Phys.* **5**, 732–735 (2009).

[16] Cohen, A. J., Mori-Sánchez, P. & Yang, W. Insights into current limitations of density functional theory. *Science* **321**, 792–794 (2008).

[17] 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).

[18] Na, G. S., Jang, S. & Chang, H. Predicting thermoelectric properties from chemical formula with explicitly identifying dopant effects. *Npj Comput. Mater.* **7** (2021).

[19] Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* (2018).

[20] Schleder, G. R., Padilha, A. C., Acosta, C. M., Costa, M. & Fazzio, A. From dft to machine learning: recent approaches to materials science–a review. *J. Phys. Mater.* **2**, 032001 (2019).

[21] Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. *ICLR* (2017).

[22] Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater* (2019).

[23] Goodall, R. E. & Lee, A. A. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *Nat. Commun.* **11**, 1–9 (2020).

[24] Alam, H. & Ramakrishna, S. A review on the enhancement of figure of merit from bulk to nano-thermoelectric materials. *Nano energy* **2**, 190–212 (2013).

[25] Cancilla, J. C., Perez, A., Wierzchoś, K. & Torrecilla, J. S. Neural networks applied to determine the thermophysical properties of amino acid based ionic liquids. *Physical Chemistry Chemical Physics* **18**, 7435–7441 (2016).

[26] Wang, T., Zhang, C., Snoussi, H. & Zhang, G. Machine learning approaches for thermoelectric materials research. *Advanced Functional Materials* **30**, 1906041 (2020).

[27] Draper, N. R. & Smith, H. *Applied Regression Analysis, 3rd Edition* (Wiley-Interscience, 1998).

[28] Powers, D. M. W. Evaluation: From precision, recall and f-measure to roc., informedness, markedness and correlation. *J. Mach. Learn. Technol.* **2**, 37–63 (2011).

[29] Chen, Z.-G., Shi, X., Zhao, L.-D. & Zou, J. High-performance snse thermoelectric materials: Progress and future challenge. *Prog. Mater. Sci.* **97**, 283–346 (2018).

[30] Wei, J. *et al.* Review of current high-zt thermoelectric materials. *J. Mater. Sci.* **55**, 12642–12704 (2020).

[31] McDonald, G. C. Ridge regression. *Wiley Interdiscip. Rev. Comput. Stat.* **1**, 93–100 (2009).

[32] Taunk, K., De, S., Verma, S. & Swetapadma, A. A brief review of nearest neighbor algorithm for learning and classification. In *ICCS*, 1255–1260 (IEEE, 2019).

[33] Smola, A. J. & Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **14**, 199–222 (2004).

[34] Schulz, E., Speekenbrink, M. & Krause, A. A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions. *J. Math. Psychol.* **85**, 1–16 (2018).

[35] Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **65**, 386 (1958).

[36] Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *KDD*, 785–794 (2016).

[37] Xu, K. *et al.* How neural networks extrapolate: From feedforward to graph neural networks. In *ICLR* (2021).

[38] Na, G. S., Jang, S. & Chang, H. Nonlinearity encoding to improve extrapolation capabilities for unobserved physical states. *Phys. Chem. Chem. Phys.* **24**, 1300–1304 (2022).

[39] Axelsson, O. A generalized conjugate gradient, least square method. *Numer. Math.* **51**, 209–227 (1987).

[40] Lee, J. *et al.* Control of thermoelectric properties through the addition of ag in the bi0. 5sb1. 5te3alloy. *Electron. Mater. Lett.* **6**, 201–207 (2010).

[41] Cao, S. *et al.* Enhanced thermoelectric properties of ag-modified bi0. 5sb1. 5te3 composites by a facile electroless plating method. *ACS Appl. Mater. Interfaces* **9**, 36478–36482 (2017).

[42] Feng, B., Tang, Y. & Lei, J. The influential mechanism of ti doping on thermoelectric properties of bi0. 5sb1. 5te3 alloy. *J. Mater. Sci.: Mater. Electron.* **32**, 28534–28541 (2021).

[43] Gogna, A. & Tayal, A. Metaheuristics: review and application. *J. Exp. Theor. Artif. Intell.* **25**, 503–526 (2013).

[44] Leguy, J., Cauchy, T., Glavatskikh, M., Duval, B. & Da Mota, B. Evomol: a flexible and interpretable evolutionary algorithm for unbiased de novo molecular generation. *J. Cheminformatics* **12**, 1–19 (2020).

[45] Pirgazi, J., Alimoradi, M., Esmaeili Abharian, T. & Olyaee, M. H. An efficient hybrid filter-wrapper metaheuristic-based gene selection method for high dimensional datasets. *Sci. Rep.* **9**, 1–15 (2019).

[46] Mlinar, V. Utilization of inverse approach in the design of materials over nano-to macro-scale. *Annalen der Physik* **527**, 187–204 (2015).

[47] Katoch, S., Chauhan, S. S. & Kumar, V. A review on genetic algorithm: past, present, and future. *Multimed. Tools. Appl.* **80**, 8091–8126 (2021).

[48] Poli, R., Kennedy, J. & Blackwell, T. Particle swarm optimization. *Swarm Intell.* **1**, 33–57 (2007).

[49] Faramarzi, A., Heidarinejad, M., Stephens, B. & Mirjalili, S. Equilibrium optimizer: A novel optimization algorithm. *Knowl. Based Syst.* **191**, 105190 (2020).

Author Contributions

Competing Interests

The authors declare no competing interests