Data-driven discovery of 3D and 2D thermoelectric materials

Kamal Choudhary, Kevin F Garrity and Francesca Tavazza

Materials Science and Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, United States of America

E-mail: kamal.choudhary@nist.gov

Received 23 March 2020, revised 5 June 2020
Accepted for publication 26 June 2020
Published 27 August 2020

Abstract
In this work, we first perform a systematic search for high-efficiency three-dimensional (3D) and two-dimensional (2D) thermoelectric materials by combining semiclassical transport techniques with density functional theory (DFT) calculations and then train machine-learning models on the thermoelectric data. Out of 36 000 three-dimensional and 900 two-dimensional materials currently in the publicly available JARVIS-DFT database, we identify 2932 3D and 148 2D promising thermoelectric materials using a multi-step screening procedure, where specific thresholds are chosen for key quantities like bandgaps, Seebeck coefficients and power factors. We compute the Seebeck coefficients for all the materials currently in the database and validate our calculations by comparing our results, for a subset of materials, to experimental and existing computational datasets. We also investigate the effect of chemical, structural, crystallographic and dimensionality trends on thermoelectric performance. We predict several classes of efficient 3D and 2D materials such as Ba(MgX)2 (X = P, As, Bi), X2YZ6 (X = K, Rb, Y = Pd, Pt, Z = Cl, Br), K2PtX2 (X = S, Se), NbCu3X4 (X = S, Se, Te), Sr2XYS6 (X = Ta, Zn, Y = Ga, Mo), TaCu3X4 (X = S, Se, Te), and XYN (X = Ti, Zr, Y = Cl, Br). Finally, as high-throughput DFT is computationally expensive, we train machine learning models using gradient boosting decision trees and classical force-field inspired descriptors for n- and p-type Seebeck coefficients and power factors, to quickly pre-screen materials for guiding the next set of DFT calculations. The dataset and tools are made publicly available at the websites: https://www.ctcms.nist.gov/~knc6/JVASP.html, https://www.ctcms.nist.gov/jarvisml/ and https://jarvis.nist.gov/.

Keywords: density functional theory, machine learning, thermoelectrics, two-dimensional materials, three-dimensional materials, materials genome, NIST-JARVIS

(Supplementary material for this article is available online)

1. Introduction
Thermoelectrics [1–4] are materials that can convert a temperature gradient into electric voltage, or vice versa. Thermoelectrics can be used to regenerate electricity from waste heat [5], refrigeration [6] and several other space-technology applications [7, 8]. The search for efficient thermoelectric materials is an area of intense research due the potential of converting waste heat into electrical power, and therefore improving energy efficiency and reducing fossil fuel usage. The figure of merit for thermoelectric materials is the dimensionless quantity $zT$:

$$zT = \frac{S^2 \sigma}{k_e + k_l} T$$

where $S$, $\sigma$, $k_e$, $k_l$, and $T$ are the Seebeck coefficient, electrical conductivity, electronic part of thermal conductivity,
lattice thermal conductivity, and temperature, respectively. The numerator, $S^2\sigma$, is referred to as the power-factor. To achieve a high $zT$, a material should have a high-power factor and low thermal conductivity. Experimental synthesis and characterization are ultimately the critical steps to prove the usefulness of a thermoelectric material; however, experiments are costly and time-consuming, and the list of potential thermoelectrics is very large. Computational methods based on first principles density functional theory (DFT) can be very useful in the initial screening process, as well as in interpreting experimental results. DFT [9–12] has successfully predicted the Seebeck coefficients and power factors for various classes of bulk materials. There has also been a series of high-throughput computational searches for bulk/three-dimensional (3D) thermoelectrics. Chen et al [9] and Ricci et al [13] compute thermoelectric properties of more than 48 000 materials and show a reasonably strong comparison between the maximum Seebeck-coefficient determined from DFT and experiment for a subset. Gorrati [14, 15] use high-throughput first-principles calculations to screen transition metal oxides, nitrides, and sulfides for candidate materials with high power factors and low thermal conductivity. Gorrati et al [16] develop TEdesignLab as a thermoelectrics-focused virtual laboratory that contains calculated thermoelectric properties using several thermoelectric metrics [15]. Carrete et al use high-throughput method to identify materials with low-thermal conductivity [17]. He et al [18] search for several transition metal oxides with high thermoelectric performance. In addition to the computational databases, there are several developments of experimental databases also such as UCSB-MRL thermoelectric database [19]. Similar to the 3D materials, there has been a huge upsurge in research on monolayer/two-dimensional (2D) materials due to their promising high Seebeck coefficients and low thermal conductivities [20–24]. Despite the above research, a systematic, combined database that allows for the comparison of bulk and monolayer thermoelectric properties is still lacking. In addition, such a systematic database of thermoelectric properties is necessary to develop machine learning models for predicting the thermoelectric properties of new materials, which would circumvent the high computational cost of additional DFT calculations and potentially guide materials discovery. There have been a few recent reports on the applications of machine learning for thermoelectric properties [25–27], but the field is still developing.

In this work, first, we present a high-throughput DFT database of bulk and monolayer thermoelectric properties. In this high-throughput work, we focus on finding high power factor materials, which is a necessary and less computationally expensive step in identifying thermoelectrics. We do not attempt to predict $zT$ values in this work. All of the data and tools are provided at the JARVIS-DFT website, which is a part of the materials genome initiative at the National Institute of Standards and Technology (NIST). The JARVIS-DFT database contains about 36 000 bulk and 900 low-dimensional materials with their DFT-computed structural, exfoliability [28], elastic [29], optoelectronic [30] solar-cell efficiency [31], and topologically non-trivial [32] properties. Using this database, we highlight a few novel 3D-bulk materials and 2D-monolayer materials that we predict have good thermoelectric properties. We also investigate correlations of thermoelectric properties with chemistry and structure of materials. Finally, we develop highly accurate machine learning models for quickly identifying efficient thermoelectric materials.

2. Methods

The computational workflow description adopted in this work can be divided into four sub-sections: (A) DFT calculations, (B) BoltzTrap calculations, (C) machine-learning model training, and (D) limitations.

(A) DFT calculations: the DFT calculations are carried out using Vienna Ab initio simulation package (VASP) [33, 34] software using the JARVIS-DFT workflow given on our github page (https://github.com/usnistgov/jarvis). Please note commercial software is identified to specify procedures. Such identification does not imply recommendation by NIST. We use OptB88vdW functional [35], which gives accurate lattice parameters for both vdW and non-vdW (3D-bulk) solids [28]. There have been several tests of vdW functionals [28, 62] and OptB88vdW predicts important physical quantities such as lattice constants, bulk modulus and atomization energies as well as or better than other available vdW functionals. We employ spin-polarized calculations, starting with a ferromagnetic spin-ordering, during the geometric optimization of each material. In this work, a material is termed as low-dimensional if it contains vdW-bonding in one (2D-bulk), two (1D-bulk), and three (0D-bulk) crystallographic directions. Details of the low-dimensional material database can be found in reference [28, 29, 36]. A monolayer/2D-material is simulated with broken periodicity in $z$-direction with a vacuum padding of at least 18 Å. Both the internal atomic positions and the lattice constants are allowed to relax in spin-unrestricted calculations until the maximal residual Hellmann–Feynman forces on atoms are smaller than 0.001 eV Å$^{-1}$ and energy-tolerance of $10^{-2}$ eV. We do not consider spin–orbit interactions or magnetic orderings besides ferromagnetic, because of a high computational cost. We note that nuclear spins are not explicitly considered during the DFT calculations. The list of pseudopotentials used in this work is given on the GitHub page.

(B) BoltzTrap calculations: the transport properties were calculated using the Boltzmann transport equation (BTE) implemented in the BoltzTrap code [37]. The BTE is used to investigate the non-equilibrium behavior of electrons and holes by statistically averaging all possible quantum states given by the equation below:

$$\frac{df(k, T, t)}{dt} = \left( \frac{\partial f(k, T, t)}{\partial t} \right) - \frac{dk}{dt} \nabla_k f(k, T, t) - v(k) \nabla_r f(k, T, t).$$  

(2)

$f$ is the electron distribution, which is a function of state $k$, temperature $T$ and time $t$, and $v(k)$ are the electron group velocities. The three terms on the right-hand side of equation (2) refer, respectively, to the temporal rate of change of $f$ due to
all scattering processes, the rate of change of $f$ due to external forces, and the diffusion from the carrier density gradient. If the external forces consist only of a low electric field, $E$, and no magnetic field, $B$, such that $\frac{\partial}{\partial t} = \frac{\partial}{\partial t}$ then the low-filed BTE is given by:

$$\frac{df(k,T,t)}{dt} + v(k) \nabla f(k,T) + \frac{eE}{\hbar} \nabla_k f(k,T) = \left( \frac{\partial f(k,T,t)}{\partial t} \right)_{\nu}.$$  \hspace{1cm} (3)

Now, $f$ can be described as a first-order (linear) perturbation from the (equilibrium) Fermi–Dirac distribution, $f_0$, due to scattering

$$\left( \frac{\partial f(k,T,t)}{\partial t} \right)_{\nu} = -\frac{f(k) - f_0(k)}{\tau}$$  \hspace{1cm} (4)

where

$$f_0[E(k)] = \frac{1}{e^{[E(k) - \mu]/k_B T} + 1}$$  \hspace{1cm} (5)

where the dependence of $E$ on $k$ is given by the electronic band structure, and the various scattering terms and time dependence are lumped into the electronic relaxation time, $\tau$. The computation of the relaxation time is very computationally expensive, especially in a high-throughput context [38]. This methodology has been used earlier for both bulk [20–24] and monolayer [40] materials. We converge the $k$-points and plane wave cut-off in DFT-calculations for all the materials in our database using energy convergence criteria [41] of 0.001 eV. These $k$-points and cut-offs values are generally higher than the usual 1000/atom $k$-points selected in previous high-throughput database studies. Note that for some metallic and low-bandgap materials it is difficult to converge the BoltzTrap code uses constant relaxation time approximation (CTRA) and the rigid band-approximation (RBA). In CTRA, the relaxation time cancels out for Seebeck coefficients, but for electrical conductivity, we choose a value of $10^{-14}$ s as the relaxation time [39]. The relaxation time can depend on both intrinsic factors like electron–phonon coupling and extrinsic factors like the presence of defects, which are not considered in this work. The RBA assumes that the shape of density of states (DOS) does not change by doping or increasing the operation temperature.

3. Results and discussion

We apply BoltzTrap to calculate the electronic transport properties for all the 36 000 bulk and 900 monolayer materials in our database, calculating the Seebeck coefficients ($S$), electrical-conductivities ($\sigma$), power factors, and electronic part of thermal conductivities, all as a function of temperature and doping. As mentioned above BoltzTrap is not suitable for calculating thermoelectric properties of various metallic systems, so we out of 36 000 bulk 3D and 900 2D monolayer materials we have calculated BoltzTrap data for 21 900 3D bulk 294 monolayer 2D materials data only. The resultant data is used for screening potential thermoelectric materials, analyzing trends and machine learning training processes. A flow-chart describing our computational search is shown in figure 1.

3.1. Experimental and theoretical benchmarking

To benchmark our computational methodology, we compare our Seebeck coefficients for a subset of 14 materials to experimental data [48–60], and we find a mean absolute deviation (MAD) of 54.7 $\mu$V K$^{-1}$ ($r^2 = 0.94$). The details of this comparison are shown in the supplementary information (figure S1(a) and table S1). Next, we compare our Seebeck coefficients (JV) to DFT results from another computational database, the materials-project (MP) [13], to ascertain how sensitive our results are to specific choices of DFT parameters (exchange–correlation functional, $k$-points density and energy cutoff). We look at 9434 compounds from the MP and JV databases, comparing the n-type Seebeck coefficient at 600 K and $10^{20}$/cm$^3$ doping. We obtained an MAD of 18.8 $\mu$V K$^{-1}$ ($r^2 = 0.87$), signifying that DFT data for Seebeck coefficients are closer to each other than to the experimental results. We
attribute the differences between the MP and JV datasets to the fact that MP uses the GGA–PBE [47] functional as well as fixed $k$-points and cutoffs for their DFT calculations, while JV uses the OptB88vdW functional and an automatic convergence procedure for $k$-points and cutoffs, which we expect to provide improved results especially for vdW-bonded materials. More details/data on the DFT comparisons are also provided in the supplementary information (figure S1(b)).

3.2. Screening and analysis of bulk thermoelectrics

After computing the Seebeck coefficients for all the materials in the database, we concentrate on developing a screening method that identifies high-efficiency thermoelectric materials. The thermoelectric performance of a material depends on several quantities, most notably, temperature, doping type, and doping concentration. We chose to concentrate on temperatures of 600 K and $10^{20}/\text{cm}^3$ doping, which represent a typical thermoelectric operating temperature and a doping level that is achievable for many semiconductors. However, we note that dopability of a semiconductor depends on several critical factors such as native defect energetics, which are too computationally expensive to predict from first principles in a systematic manner. Many experimentally relevant thermoelectrics are doped at similar concentration as shown in the supplementary information (table S1). As good thermoelectric materials are generally semiconducting or insulating, we first screen for materials with bandgap $>0.05$ and BoltzTrap data, which narrows the search to 8764 materials. We present an overview of the database in figure 2.

The n and p-type Seebeck coefficients are generally negative and positive values, respectively, with a maximum absolute value of 600 $\mu$V K$^{-1}$, as shown in figures 2(a) and (d). In figures 2(b) and (e), we show the distribution of power factors for n-type and p-type materials. Experimentally known high-efficiency thermoelectrics have power factors of more than $1000 \mu$W (mK)$^{-1}$ [2]. Although a high Seebeck-coefficient is necessary for a high power factor, there is typically a significant tradeoff between $S$ and $\sigma$, necessitating a more careful analysis [14]. This tradeoff can be seen in figure 2(c), which shows a scatter plot of $S$ versus $\sigma$ for n-type thermoelectrics. The sizes of dots are proportional to the bandgaps and color-coded based on their power factor values. A similar inverse-relationship applies to p-type materials as well. In figure 2(f), we look at the relationship between power factors and band gaps, finding that high power factors occur more often in low band gap materials. A similar relationship holds for the n-type materials as well.

For the next step in the screening procedure, we select only materials with n-type and p-type power factor at 600 K and $10^{20}/\text{cm}^3$ larger than $1000 \mu$W (mK)$^{-1}$ [2], which gives us 4330 and 4403 candidates, respectively. We find 2932 materials with both n-and p-type PF $>1000 \mu$W (mK)$^{-1}$ [2]. For this case, we analyzed the set of selected materials in terms of various chemical and physical attributes. To begin with, we classified their dimensionality, which is determined by lattice-constant and data-mining approaches [28]. As shown in figure 2(g), we find that 14.4% of the high efficiency thermoelectrics are low-dimensional, i.e. vdW-bonded, while the rest have three-dimensional bonding. Next, we find that these materials tend to be highly symmetric, as cubic and trigonal symmetry materials are over-represented.
Figure 2. A brief overview of the thermoelectric data for periodic bulk materials. Figures (a)–(f) have been computed on all 3D materials, while (g)–(i) display properties only computed on the set of bulk screened materials (bandgap $>0.05$ eV, n-type and p-type PF $>1000$ ($\mu$W/mK$^{-1}$) at 600 K and $10^{16}$/cm$^3$ doping concentration). (a) n-type Seebeck coefficient distribution, (b) n-type power factor of materials, (c) n-type electrical conductivity plotted against the absolute values of Seebeck-coefficient with color-coded power-factor and size of the dots proportional to bandgaps, (d) p-type Seebeck coefficient distribution, (e) p-type power factor of materials, (f) p-type power factor plotted against the bandgaps, (g) predicted dimensionality distribution of screened materials, (h) space-group distribution of the screened materials, (i) power-factor vs Seebeck distribution of the screened materials.

(as shown in figure 2(h)). As discussed earlier, we find a range of Seebeck coefficients among our set of high-power factor materials, due to the tradeoff between $S$ and $\sigma$, as shown in figure 2(i). Finally, we note that our screening process rediscovered several well-known thermoelectrics, such as Bi$_2$Te$_3$, SnSe, GeTe, Mg$_3$Si, PtSe$_2$, PbSe, PbTe and PtSe$_2$, as well as many potential new thermoelectrics. Some examples of families of thermoelectrics uncovered in our screening include Ba(MgX)$_2$ ($X = P, As, Bi$), X$_2$YZ$_6$ ($X = K, Rb, Y= Pd, Pt, Z = Cl, Br$), K$_2$PtX$_2$ ($X = S, Se$), NbCu$_3$X$_4$ ($X = S, Se, Te$), Sr$_2$XYO$_6$ ($X = Ta, Zn, Y= Ga, Mo$), TaCu$_3$X$_4$ ($X = S, Se, Te$). Some of these materials are 0D-bulk i.e. with vdW interactions along all three axes, as in AsI$_3$ (JVASP-3636), or along just two axes (1D-bulk materials such as BiSeI (JVASP-5200), TeBr$_2$ (JVASP-33839), SeI$_2$ (JVASP-33798)), or, as in most cases, along one axis only (2D-bulk materials such as BN (JVASP-17), YSnF$_3$ (JVASP-8344), HfS$_2$ (JVASP-210), MnBr$_3$ (JVASP-2041), GeTe (JVASP-1157), TiO$_2$ (JVASP-30586), NiO$_2$ (JVASP-8645), GaP (JVASP-28372), SbAsO$_4$ (JVASP-10177)). Many other newly identified materials for thermoelectrics applications are 3D-bulk materials, i.e. no vdW-bonding, such as LiMgN (JVASP-22546), CaO$_2$ (JVASP-22677), Li$_2$CuSb (JVASP-7820), SrZrO$_3$ (JVASP-8037), K$_2$PtSe$_2$ (JVASP-2838), Al$_2$ZnS$_4$ (JVASP-9688), ZnO$_2$ (JVASP-10252), Sr$_2$TaGaO$_6$ (JVASP-10974), TiSnO$_3$ (JVASP-35817), ZrSiPt (JVASP-40824), YAIO$_3$ (JVASP-50410), AlAs (JVASP-8183), RbAuC$_2$ (JVASP-7652). A full list is provided in the supplementary information. We will provide information on the stability of each compound using a convex-hull approach on the website soon. The JVASP-# denotes the JARVIS-IDs of the materials the details of which can be found at corresponding https://www.ctcms.nist.gov/~knc6/jsmol/JVASP-#.html. For example, the details of JVASP-8037 is available at https://www.ctcms.nist.gov/~knc6/jsmol/JVASP-8037.html. Note that all of the screenings and analysis in this work are based on CRTA and RBA approximations. As we develop more phonon related databases [63], it would be possible to go beyond the CRTA approach in future. But in this work, we mainly use CRTA approach because of its low computational cost.

In figure 3, we show the likelihood that a compound containing a given element has a high-power factor. More specifically, for every compound containing a given element, we calculate the percentage-probability that those materials have an $n$-power-factor greater than $1000$ ($\mu$W/mK$^{-1}$) [2]. We find...
Figure 3. Periodic table trends of high-power factor materials. The elements in a material were assigned 1 or 0 if the material has high or low n-type power-factor (>1000 ($\mu$W (m K)$^{-1}$)). Then the probability of finding the element in a high power-factor material was calculated.

that many of the alkaline earth metals, early transition metals, Ir, Pt, Cu, Ag and chalcogenides were found to contribute toward high efficiency materials, which is again consistent with previously known thermoelectric materials [1–4] such as Bi$_2$Te$_3$, SnSe, GeTe, Mg$_2$Si, PtSe$_2$. For example, 298 out of 587 Se-containing compounds in our database have power-factors greater than 1000 $\mu$W (m K)$^{-1}$ [2] so, the percentage is 50.77%. Such periodic table trends can help guide new materials searches or doping strategies that may result in improved thermoelectrics. There have been other material-class based studies (such as chalcogenides) for finding chemical trends in terms different types of chemically bonded systems [64]. However, as there are no global clear trends between high-PF materials and common chemical characteristics that would be obvious on a periodic table, there is a need for more sophisticated statistical machine learning techniques, which we carry out in a following section.

3.3. Screening and analysis of monolayer thermoelectrics

The screening process described and analyzed up to this point was applied exclusively to bulk materials. Next, we apply a similar procedure to monolayer 2D materials. The vdW-bonded materials (2D-bulk) can be exfoliated to produce 2D-monolayers, which have shown promise in thermoelectric applications [20–24]. The exfoliability of a 2D-bulk material is depends on its exfoliation energy, as described in detail in reference [28]. In order to find exfoliable high-efficiency 2D monolayers, we select materials in our database which meet the following criteria: (a) exfoliation energy < 200 meV/atom, (b) bandgap > 0.05 and (c) monolayer-n-type Seebeck-coefficient < $-100$ $\mu$V K$^{-1}$. In order to compare the thermoelectric properties of monolayer and bulk structures, it is necessary to express the doping and conductivity quantities in a way that does not depend on the arbitrary vacuum thickness of a monolayer simulation cell. To achieve this, we rescaled the volume of the monolayers using the thickness of a 2D-layer. Using this screening procedure, we identified 148 promising 2D monolayers among 900 materials. As seen in figure 4(a), comparing 2D and 3D Seebeck coefficients, we see that monolayers tend to have smaller absolute Seebeck coefficients than their bulk counterparts. This shows that interlayer coupling is important to thermoelectric behavior, and that 2D thermoelectric properties cannot be exactly obtained from their bulk counterpart. Nevertheless, the Spearman’s correlation between the bulk and monolayer Seebeck coefficients is 0.711 while the Pearson’s correlation is 0.721, as shown in tables 1 and S2, suggesting noteworthy correlation. We also investigated the correlation between 3D and monolayer DOS at the Fermi level, for 600 K and 10$^{20}$/cm$^3$ doping, as well as the correlation for the effective mass. Both results are given in tables 1 and S2. We find a strong correlation between bulk and monolayer effective masses (0.80), and an even stronger correlation (0.90) between bulk and monolayer DOS at the Fermi level, suggesting that the Seebeck-coefficient is more difficult to predict. A lower Spearman’s correlation of 0.43 was reported for bulk systems Seebeck-coefficient and DOS by Garrity et al [14] for transition metal oxides, sulfides and nitrides. Kumar et al [21] also showed similar correlations for monolayer and bulk WSe$_2$, using DFT calculations.

Some of the high-efficiency 2D-monolayer materials that we find are: AuBr (JVASP-27756), SnSe (JVASP-5929), SnS (JVASP-19989), PtSe$_2$ (JVASP-744), ZrS$_2$ (JVASP-792), GaSe (JVASP-687), WSe$_2$ (JVASP-652). The full database is available online. We note a special class of vdW-bonded materials, XYZ (X = Ti, Zr, Y = N, Z = Cl, Br, I) (such as
Figure 4. (a) 3D vs 2D Seebeck coefficients, (b) crystal structure of ZrBrN (JVASP-12027), Zr in gray, N in blue, Br in red (c) Phonon bandstructure for ZrBrN.

Table 1. Spearman correlation of monolayer (Mono) and bulk DOS (states/unitcell) at Fermi level, the effective mass of electrons ($m_e$) and Seebeck coefficients ($\mu V K^{-1}$).

| Spearman correlation | Mono-DOS | Bulk-DOS | Bulk-Seebeck | Mono-Seebeck | Bulk-$m_e$ | Mono-$m_e$ |
|---------------------|----------|----------|---------------|--------------|------------|------------|
| Mono-DOS            | —        | 0.897    | 0.592         | 0.553        | −0.474     | −0.432     |
| Bulk-DOS            | 0.897    | —        | 0.604         | 0.566        | −0.493     | −0.412     |
| Bulk-Seebeck        | 0.592    | 0.604    | —             | 0.711        | −0.859     | −0.723     |
| Mono-Seebeck        | 0.533    | 0.566    | 0.711         | —            | −0.596     | −0.668     |
| Bulk-$m_e$          | −0.474   | −0.493   | −0.859        | −0.596       | —          | 0.801      |
| Mono-$m_e$          | −0.432   | −0.412   | −0.723        | −0.668       | 0.801      | —          |

JVASP-6268, JVASP-6181, JVASP-6184), which show consistently high thermoelectric behavior, and which would be a promising target for future investigation. While many of these compounds have not yet been experimentally verified as 2D materials [28], they are almost all related to experimentally known 3D layered structures from the ICSD, and have exfoliation energies consistent with the small number of experimentally studied 2D materials. We hope that works like this one will in part encourage the study of a wider range of 2D materials.

As a first step in this direction, we carried out thermal conductivity calculations for a representative example, ZrBrN (JVASP-12027), in its 3D form, which resulted in a remarkable low lattice thermal conductivity of 0.3 W m$^{-1}$ K$^{-1}$ at 300 K. The unit cell and phonon dispersion curve of ZrBrN are shown in figures 4(b) and (c). We see that while the structure is dynamically stable (all positive frequency modes), there are low-frequency phonon modes which extend across the Brillouin zone, which are responsible for the strongly anharmonic behavior in this system. Such detailed analysis for candidate thermoelectric materials will be considered in the future.

3.4. Machine learning analysis

Table 2. Initial comparison of ML Classification Techniques Using DT, RF, KNN, MLP, GBDT implemented in scikit-Learn Package (SK-GB), GBDT in XGB, and GBDT in LGB.

| ROC-AUC | n-Seeb | p-Seeb | n-PF | p-PF |
|---------|--------|--------|------|------|
| DT      | 0.80   | 0.83   | 0.64 | 0.65 |
| RF      | 0.92   | 0.94   | 0.75 | 0.76 |
| KNN     | 0.91   | 0.92   | 0.74 | 0.74 |
| MLP     | 0.73   | 0.95   | 0.74 | 0.76 |
| SK-GB   | 0.92   | 0.94   | 0.75 | 0.76 |
| XGB     | 0.92   | 0.95   | 0.75 | 0.76 |
| LGB     | 0.94   | 0.96   | 0.75 | 0.80 |

Finally, to accelerate the DFT screening process, we train supervised classification machine learning models [43] for n and p-type Seebeck coefficients and power-factors for 3D bulk crystals only. In this scheme, we simply classify whether materials have a property value greater or lesser than certain thresholds using CFID descriptors [43]. The accuracies of the classification models are evaluated based on the AUC of the ROCs curves. The ROC curve illustrates the model’s ability to differentiate between high and low-performance materials, classifying a material to be high-performance if its Seebeck coefficient is less than −100 μV K$^{-1}$ for n-type, or more than 100 μV K$^{-1}$ for p-type, and if its n- and p-power factor is more than 1000 μW (m K$^{-1}$) [2] at 600 K and 10$^{20}$/cm$^3$ doping. The ROC curve plots the prediction rate for high-performance materials, correctly versus incorrectly predicted. A value of 0.5 implies random guessing, while a value of 1.0 implies a perfect model. We first train classification models with default parameters using DT, RF, KNNs, MLP, and gradient boosting models implemented in scikit-learn package and also GBDT implemented in XGB and LGB packages. As a standard practice, we use train-test split (90%:10%), five-fold cross-validation, and examining AUC for ROC curves on the 10% held set (as shown in table 2). Evidently, the LGB model already performs very well with the default parameters only.
We further tune LGB hyperparameters such as the number of estimators, the number of leaves, and the learning rate using a five-fold cross-validation grid search. Using the best model of grid search, we predict the ROC of the 10% held set (shown in figure 5). We achieve high accuracy for most of the models, with the model for p-type Seebeck being the best one as it corresponds to the highest value of 0.96 as shown in figure 5(b). We obtain at least 0.8 AUC for most of the ML models, signifying high prediction accuracies.

In addition to the thresholds specified above for classifying high/low performance materials ($100 \mu V K^{-1}$ for $|S|$, $1000 \mu W (mK)^{-1}$ [2] for PF), we also train models with more stringent thresholds as shown in table 3. We find similar ROC AUC for several different thresholds, allowing for more or less sensitive models to be chosen. We can apply these models to arbitrary materials to quickly pre-screen efficient thermoelectric materials, which would become the subject of the next set of DFT calculations. We followed a similar procedure to identify semiconducting 2D materials in our previous work [43] and successfully discovered several 2D materials. Clearly, the screening process can be much accelerated using machine learning models as a first step. All these models are provided on the JARVIS-ML website (https://www.ctcms.nist.gov/jarvisml/) for predicting thermoelectric performance of new compounds.

Furthermore, the GBDT algorithm allows us to get the feature importance information for each of the 1557 descriptors. We group the features in respective classes [43] such as chemical descriptors (Chem), radial distribution function, nearest neighbor, angular distribution up to first (ADF-1st) and second neighbor (ADF-2nd), dihedral distribution function up to first neighbor, charge descriptor (Chg), cell-size related descriptors (Cell) as shown in figure 6. We find that chemical, radial distribution, angle-distribution up to first neighbors, and dihedral angle distribution are almost equally important for achieving a high accuracy model. This is in contrast with the formation energy model [43] in our previous work, where we found that chemistry was the most dominant feature. This might explain why there is no clear trend visible in the periodic table discussed above (see figure 3). Out of all the 1557 descriptors, some of the most important descriptors for all the models are: cell-size related descriptors-density and logarithm of volume of the cell, radial distribution peak at 7.5 Å, 9.4 Å and 9.5 Å, first-neighbor based angular distribution peak at 178 degree, mean of product of polarizability and atomic mass, ratio of atomic radii and molar volume and refractive index of individual constituent elements [43]. Moreover, we observe

| Threshold          | Model | ROC AUC |
|--------------------|-------|---------|
| $-100 \mu V K^{-1}$ | p-Seeb | 0.96    |
| $-200 \mu V K^{-1}$ | p-Seeb | 0.96    |
| $-300 \mu V K^{-1}$ | p-Seeb | 0.95    |
| $1000 \mu W (mK)^{-1}$ [2] | p-PF | 0.82    |
| $2000 \mu W (mK)^{-1}$ [2] | p-PF | 0.83    |
| $3000 \mu W (mK)^{-1}$ [2] | p-PF | 0.84    |

Figure 5. Classification ROC curves for high Seebeck-coefficient and high power-factor materials. The dotted line shows the random guessing line with an AUC 0.5.
that the feature-importances are very similar for both the n-type and p-type ML models. This can be attributed to the fact that the majority of materials with high n-type Seebeck coefficients also have high p-type Seebeck coefficients, and vice versa, with a similar relationship for PFs. Note that ML is applied to the bulk 3D single crystals only in this work. Further studies would have to be done to understand the effects of nano structuring [65] such as nanowires. Application of ML to nanostructured materials requires a substantial increase in data-generation, which is beyond the scope of the current work.

4. Conclusions

In summary, we use semiclassical transport methods based on DFT calculations to evaluate the thermoelectric properties of both bulk (3D) and monolayer (2D) materials. In addition to identifying interesting candidate materials, we also show chemical, crystallographic, compositional and dimensionality trends for the whole dataset. We screen 2D materials and evaluate trends between the thermoelectric performance of bulk and monolayer geometries. We identify several compositional classes with high thermoelectric performance. We predict ultra-low lattice thermal conductivity in the ZrBrN class of materials. Although the constant-relaxation time approximation is a crude approximation, it allows the generation of large-scale database for initial screening of thermoelectric materials. Finally, we train machine learning models to accelerate the future screening processes. We believe that our data and tools for evaluating and predicting thermoelectric performance will greatly enhance the discovery and characterization of thermoelectric materials.

ORCID iDs

Kamal Choudhary https://orcid.org/0000-0001-9737-8074

References

[1] Snyder G J and Toberer E S 2011 Materials for Sustainable Energy: A Collection of Peer-Reviewed Research and Review Articles from Nature Publishing Group (Singapore: World Scientific) pp 101–10
[2] Chen G, Dresselhaus M, Dresselhaus G, Fleurial J-P and Caillat T 2003 Recent developments in thermoelectric materials Int. Mater. Rev. 48 45–66
[3] Wang H, Pei Y, LaLonde A D and Snyder G J 2013 Thermoelectric Nanomaterials (Berlin: Springer) pp 3–32
[4] Zebarjadi M, Esfarjani K, Dresselhaus M, Ren Z and Chen G 2012 Perspectives on thermoelectrics: from
fundamentals to device applications Energy Environ. Sci. 5 5147–62
[5] Bell L E 2008 Cooling, heating, generating power, and recovering waste heat with thermoelectric systems Science 321 1457–61
[6] Goldsmid H 2013 Thermoelectric Refrigeration (Berlin: Springer)
[7] Yang J and Caillat T 2006 Thermoelectric materials for space and automotive power generation MRS Bull. 31 224–9
[8] Rifats S B and Ma X 2003 Thermoelectrics: a review of present and potential applications Appl. Therm. Eng. 23 913–35
[9] Chen W et al 2016 Understanding thermoelectric properties from high-throughput calculations: trends, insights, and comparisons with experiment J. Mater. Chem. C 4 4414–26
[10] Zha H et al 2015 Computational and experimental investigation of TmAgTe3 and XYZ 2 compounds, a new group of thermoelectric materials identified by first-principles high-throughput screening J. Mater. Chem. C 3 10554–65
[11] Tan G et al 2014 High thermoelectric performance of p-type SnTe via a synergistic band engineering and nanostructuring approach J. Am. Chem. Soc. 136 7006–17
[12] Wang S, Wang Z, Setyawan W, Mingo N and Curtarolo S 2011 Assessing the thermoelectric properties of sintered compounds via high-throughput ab-initio calculations Phys. Rev. X 1 012012
[13] Ricci F et al 2017 An ab initio electronic transport database for inorganic materials Sci. Data 4 170085
[14] Garry K F 2016 First-principles search for n-type oxide, nitride, and sulfide thermoelectrics Phys. Rev. B 94 045122
[15] Gorai P et al 2016 TE Design Lab: a virtual laboratory for thermoelectric material design Comput. Mater. Sci. 112 368–76
[16] Gorai P, Stevanović V and Toberer E S 2017 Computationally guided discovery of thermoelectric materials Nat. Mater. 2 17053
[17] Carrete J, Li W, Mingo N, Wang S and Curtarolo S 2014 Finding unprecedented low-thermal-conductivity half-Heuslers semiconductors via high-throughput materials modeling Phys. Rev. X 4 011019
[18] He J et al 2017 Bi2PdO4: a promising thermoelectric oxide with high power factor and low lattice thermal conductivity Chem. Mater. 29 2529–34
[19] Gautlois M W et al 2013 Data-driven review of thermoelectric materials: performance and resource considerations Chem. Mater. 25 2911–20
[20] Huang W, Da H and Liang G 2013 Thermoelectric performance of MX2 (M = MO, W; X = S, Se) monolayers J. Appl. Phys. 113 104304
[21] Kumar S and Schwingenschlogl U 2015 Thermoelectric response of bulk and monolayer MoSe2 and WSe2 Chem. Mater. 27 1278–84
[22] Ding G, Gao G, Huang Z, Zhang W and Yao K 2016 Thermoelectric properties of monolayer MoSe2 (M = Zr, Hf): low lattice thermal conductivity and a promising figure of merit Nanotechnology 27 375703
[23] Qin D et al 2018 Monolayer PdSe2: a promising two-dimensional thermoelectric material Sci. Rep. 8 2764
[24] Dresselhaus M S et al 2007 New directions for low-dimensional thermoelectrics Adv. Mater. 19 1043–53
[25] Furmanchuk A o et al 2018 Prediction of seebeck coefficient for compounds without restriction to fixed stoichiometry: a machine learning approach J. Comput. Chem. 39 191–202
[26] Wang T, Zhang C, Snoussi H and Zhang G 2019 Machine learning approaches for thermoelectric materials research Adv. Funct. Mater. 30 1906041
[27] Gautlois M W et al 2016 Perspective: web-based machine learning models for real-time screening of thermoelectric materials properties APL Mater. 4 053213
[28] Choudhary K, Kalish I, Beams R and Tavazza F 2017 High-throughput identification and characterization of two-dimensional materials using density functional theory Sci. Rep. 7 5179
[29] Choudhary K, Cheon G, Reed E and Tavazza F 2018 Elastic properties of bulk and low-dimensional materials using van der Waals density functional Phys. Rev. B 98 014107
[30] Choudhary K et al 2018 Computational screening of high-performance optoelectronic materials using OptB86vdW and TB-mBJ formalisms Sci. Data 5 180082
[31] Choudhary K et al 2019 Accelerated discovery of efficient solar-cell materials using quantum and machine-learning methods Chem. Mater. 31 5900
[32] Choudhary K, Garrity K F and Tavazza F 2019 High-throughput discovery of topologically non-trivial materials using spin-orbit spillage Sci. Rep. 9 8534
[33] Kresse G and Furthmüller J 1996 Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set Phys. Rev. B 54 11169
[34] Kresse G and Furthmüller J 1996 Efficiency of ab initio total energy calculations for metals and semiconductors using a plane-wave basis set Comput. Mater. Sci. 6 15–50
[35] Klimeš J, Bowler D R and Michaelides A 2009 Chemical accuracy for the van der Waals density functional J. Phys.: Condens. Matter 22 022201
[36] Choudhary K et al 2020 (arXiv:2007.01831)
[37] Madsen G K and Singh D J 2006 BoltzTraP. A code for calculating band-structure dependent quantities Comput. Phys. Commun. 175 67–71
[38] Samsonidze G and Kozinsky B 2018 Accelerated screening of thermoelectric materials by first-principles computations of electron–phonon scattering Adv. Energy Mater. 8 1800246
[39] Pizzi G, Volja D, Kozinsky B, Fornari M and Marzari N 2014 BoltzTraP: a code for the evaluation of thermoelectric and electronic transport properties with a maximally-localized Wannier functions basis Comput. Phys. Commun. 185 422–9
[40] Jain A et al 2013 Commentary: the materials project: a materials genome approach to accelerating materials innovation APL Mater. 1 011002
[41] Choudhary K and Tavazza F 2019 Convergence and machine learning predictions of Monkhorst–Pack k-points and plane-wave cut-off in high-throughput DFT calculations Comput. Mater. Sci. 161 300–8
[42] Li W, Carrete J, Katcho N A and Mingo N S 2014 BTE: a solver of the Boltzmann transport equation for phonons Comput. Phys. Commun. 185 1747–58
[43] Choudhary K, DeCost B and Tavazza F 2018 Machine learning with force-field-inspired descriptors for materials: fast screening and mapping energy landscape Phys. Rev. Mater. 2 038301
[44] Pedregosa F et al 2011 Scikit-learn: machine learning in Python J. Mach. Learn. Res. 12 2825–30
[45] Chen T and Guerstein C 2016 Proceedings of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (ACM) pp 785–94
[46] Ge G et al 2017 Advances in Neural Information Processing Systems 30 (NIPS 2017) pp 3146–54
[47] Perdew J P, Burke K and Emzerhof M 1996 Generalized gradient approximation made simple Phys. Rev. Lett. 77 3865
[48] Zhang Y et al 2011 Surfactant-free synthesis of Bi2Te3–Te micro–nano heterostructure with enhanced thermoelectric figure of merit ACS Nano 5 3158–65
[49] Janácek P et al 2009 Thermoelectric properties of Ti-doped Bi2Se3 single crystals CrystEngComm. 11 4450–5
[50] Liu R et al 2012 Ternary compound CuInTe2: a promising thermoelectric material with diamond-like structure Chem. Commun. 48 3818–20
[51] Plirdpring T et al 2012 Chalcopyrite CuGaTe$_2$: a high-efficiency bulk thermoelectric material Adv. Mater. 24 3622–6
[52] Kurosaki K, Goto K, Muta H and Yamanaka S J 2007 Enhancement of thermoelectric figure of merit of AgTTe by tuning the carrier concentration J. App. Phys. 102 023707
[53] Kawano K, Kurosaki K, Sekimoto T, Muta H and Yamanaka S 2007 Effect of Sn doping on the thermoelectric properties of ErNiSb-based p-type half-Heusler compound Appl. Phys. Lett. 91 062115
[54] Matsushita H, Maeda T, Katsui A and Takizawa T J 2000 Thermal analysis and synthesis from the melts of Cu-based quaternary compounds Cu–III–IV–VI$_4$ and Cu$_2$–II–IV–VI$_4$ (II = Zn, Cd; III = Ga, In; IV = Ge, Sn; VI = Se) J. Cryst. Growth 208 416–22
[55] Ono Y, Inayama S, Adachi H and Kajitani T J 2006 Thermoelectric properties of doped half-Heuslers NbCoSn$_{1-x}$Sbx and Nb$_{0.97}$Ti$_{0.03}$CoSn$_{1-x}$Sbx Japan. J. Appl. Phys. 45 8740
[56] Kawaharada Y, Kurosaki K and Yamanaka S Thermophysical properties of Fe$_2$VA, 2003 J. Alloys Compd. 352 48–51
[57] Sekimoto T, Kurosaki K, Muta H and Yamasaka S 2005 ICT 2005. 24th International Conference on Thermoelectrics, 2005 pp 347–50
[58] Zhao L-D et al 2014 Ultralow thermal conductivity and high thermoelectric figure of merit in SnSe crystals Nature 508 373
[59] Ballikaya S, Chi H, Salvador J R and Uher C J 2013 Thermoelectric properties of Ag-doped Cu$_2$ Se and Cu$_2$ Te J. Mater. Chem. A 1 12478–84
[60] Kim S, Wiendlocha B, Jin H, Tobola J and Heremans J P 2014 Electronic structure and thermoelectric properties of p-type Ag-doped Mg$_2$Sn and Mg$_2$Sn$_{1-x}$Si$_x$ (x = 0.05, 0.1) J. App. Phys. 116 153706
[61] Klimeč J, Bowler D R and Michaelides A 2011 Van der Waals density functionals applied to solids Phys. Rev. B 83 195131
[62] Choudhary K et al 2020 High-throughput density functional perturbation theory and machine learning predictions of infrared, piezoelectric and dielectric responses npj Comput. Mater. 6 64
[63] Yu Y et al 2020 Chalcogenide thermoelectrics empowered by an unconventional bonding mechanism Adv. Funct. Mater. 30 1904862
[64] Bux S K et al 2010 Nanostructured materials for thermoelectric applications Chem. Commun. 46 8311