Data analytics accelerates the experimental discovery of new thermoelectric materials with extremely high figure of merit

Sergey Levchenko (S.Levchenko@skoltech.ru)
Skolkovo Institute of Science and Technology https://orcid.org/0000-0001-5813-8473
Yaqiong Zhong
Ningbo University of Technology
Xiaojuan Hu
Fritz-Haber-Institute of the Max Planck Society
Debalaya Sarker
Skolkovo Institute of Science and Technology
Qingrui Xia
Ningbo University of Technology
Liangliang Xu
Hanyang University
Chao Yang
Ningbo University of Technology
Zhongkang Han
Skolkovo Institute of Science and Technology
Jiaolin Cui
Ningbo University of Technology https://orcid.org/0000-0002-4110-8347

Article

Keywords: thermoelectric materials, energy efficiency, active-learning framework

Posted Date: October 4th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-926972/v1

Abstract

Thermoelectric (TE) materials are among very few sustainable yet feasible energy solutions of present time. This huge promise of energy harvesting is contingent on identifying/designing materials having higher efficiency than presently available ones. However, due to the vastness of the chemical space of materials, only its small fraction was scanned experimentally and/or computationally so far. Employing a compressed-sensing based symbolic regression in an active-learning framework, we have not only identified a trend in materials' compositions for superior TE performance, but have also predicted and experimentally synthesized several extremely high performing novel TE materials. Among these, we found polycrystalline p-type $Cu_{0.45}Ag_{0.55}GaTe_2$ to possess an experimental figure of merit as high as ~2.8 at 827 K. This is a breakthrough in the field, because all previously known thermoelectric materials with a comparable figure of merit are either unstable or much more difficult to synthesize, rendering them unusable in large-scale applications. The presented methodology demonstrates the importance and tremendous potential of physically informed descriptors in material science, in particular for relatively small data sets typically available from experiments at well-controlled conditions.

Background

The ever-increasing energy demand of present era and the resulting limitless combustion of fossil fuels have already lifted global pollutant levels to an alarming threshold. Scavenging waste heat with thermoelectric generators is one of the few viable ways towards sustainable energy revolution. In thermoelectric materials a temperature gradient produces an electric potential, giving us a handle to convert waste heat to electricity.[1-4] However, the efficiency of such a conversion remained too low for its economically viable large-scale utilization. The progress in the field was limited by the need of optimizing a variety of conflicting parameters. A large absolute Seebeck coefficient (α), a low thermal conductivity (κ), and a high electrical conductivity (σ) are required to maximize the thermoelectric figure of merit $zT = \alpha^2\sigma T/\kappa$. Here κ is the sum of electronic (κ_e) and lattice (κ_L) components of thermal conductivity. The interdependence of the TE transport properties implies the crucial importance of synergistic optimization of the electronic and thermal transport properties to achieve high zT values.

Excellent thermoelectric performances have been achieved in different types of thermoelectric materials due to the devotement of a substantial amount of research efforts to identify material compositions and configurations with optimized transport characteristics providing best possible TE performance. Among them, half-Heusler alloys show great potential for high-temperature power generation applications owing to their high stability at high temperature (a high zT of ~1.5 at 1200 K was achieved in p-type NbFeSb).[5] Inorganic $Ba_8Ga_{16}Sn_{30}$ clathrates have achieved high peak zT of 1.45 at around 500 K which indicates their great potential for low- and medium-temperature applications.[6] Skutterudites achieved a high zT of ~1.8 at 823 K for $(Sr, Ba, Yb)_yCo_4Sb_{12}$ + 9.1 wt% $In_{0.4}Co_4Sb_{12}$.[7] In recent years, chalcogenides were

identified as very promising thermoelectric materials due to the discovery of a few members of this class with extremely high thermoelectric performance (zT > 2).[8-17] In particular, the binary chalcogenides, including PbTe, Cu_2Se, GeTe, and SnSe, were “superstars” in the thermoelectric family.[9-14] The ternary chalcogenide compounds are also extensively investigated due to their large variety and high TE performance (zT ~1.9 at 585 K was achieved in $AgSb_{0.96}Zn_{0.04}Te_2$).[15, 16] The mechanisms governing the performance of TE materials were also proposed. While substitution of Te by Se in PbTe has reportedly enhanced the zT by increasing phonon scattering,[18] doping with thallium has improved the TE performance of PbTe by introducing new states in the band gap and thereby improving the Seebeck coefficient.[19] The “liquid-like” behavior of copper ions around the Se sublattice in $Cu_{2-x}Se$ results in low lattice thermal conductivity, which enables high zT.[20] The n-type SnSe single crystals have an extremely high zT of ~2.8 at 773 K by Br-doping via well-controlled synthesis techniques due to a unique 3D charge and 2D phonon transport behavior.[13] Despite all these substantial efforts over the years, only a handful of materials and their optimized working conditions (i.e. temperatures) for best performances have been identified hitherto. Thus, a quick yet reliable means to scan the huge compositional space of TE materials is essential for the accelerated discovery of new high-performance TE materials.

With the advance of computational resources, first-principles calculations have been proven very useful not only for the initial screening process but also for understanding and interpretation of experimental results.[21, 22] Electronic transport properties of more than 48,000 inorganic materials have been calculated and reported as a database to be used in many fields from thermoelectricity to electronics and photovoltaics.[23] While a lot of high-throughput screening with density-functional theory (DFT) inputs have been performed to identify materials with high power factors or low thermal conductivities, including transition-metal oxides, nitrides, sulphides, etc., the actual performance and stability of the predicted materials have remained unconfirmed.[24-26] This is unfortunately a typical situation in many fields of materials science: powerful theoretical and data analytics methodologies are intensively developed and employed to screen materials space, but experimental confirmation is very scarce. The reason is not only the relatively high cost of experiments, but also a lack of effort to produce consistent experimental datasets with well-documented metadata (including experimental conditions, materials synthesis protocol, etc.). Recently an attempt to address this problem systematically has been made in the field of heterogeneous catalysis.[27] There have been several attempts in developing experimental databases in the field of thermoelectricity,[28] but the sparsity of the available datasets make the screening either biased or wrong.

Herein, we generate ~600 data points for in-house experimentally synthesized ternary p-type chalcogenide materials $A_{1-x}A_x^*$-$B_{1-y}B_y^*$-$C_{2-z}C_z^*$ (where A = Cu, Ag; A^*=Cu, Ag, Zn, Na; B = Bi, In, Sb, Ga; B^*=Bi, In, Sb, Ga, Zn, Sn; C = Se, Te; C^*=Te, Cl) with proper doping to ensure consistency of environmental and instrumental effects. Using the recently developed compressed-sensing based data analytics approach

SISSO (sure independence screening and sparsifying operator),[29] we have identified descriptors that can be used to predict the TE performance of ternary A-B-C_2 type chalcogenide materials in a fast yet reliable manner. SISSO enables us to identify the best low-dimensional descriptors (sets of descriptive parameters of materials that are either known or can be easily obtained) in an immensity of offered candidates. The search for materials with optimized zT values is performed in an active learning framework. Starting with an initial pool of experimental data, we first identify a descriptor and predict new candidates. Followed by that, we synthesize few of the predicted high-performance TE materials and include them in the data pool. This procedure (Figure 1) has been repeated several times until the SISSO model has converged.

Thus, our study goes beyond the traditional expensive and time-consuming intuition-driven or trial-and-error experimental/theoretical approaches. It reveals a relationship between TE figure of merit (zT), elemental composition, and physical features of atoms of involved species. Using data driven active learning, we have successfully predicted and experimentally verified several new TE materials of the ternary A-B-C_2 type chalcogenide family, which are not only high-performing but also are stable over a broad temperature range.

Results

Around 600 data points (for zT, Seebeck coefficient α, electronic conductivity σ, total thermal conductivity κ, and lattice thermal conductivity κ_L) for experimentally synthesized ternary A-B-C_2 type chalcogenide compounds within the temperature window of 300-850K are used as the training dataset. The in-house synthesis provides our dataset the consistency necessary for machine learning (ML).

Before discussing predictive analytic SISSO model for zT, we analyze qualitatively all measured properties contributing to thermoelectric performance using a data-mining approach subgroup discovery (SGD).[30-33] SGD finds statistically exceptional subgroups in a dataset described by statements (selectors) of the kind (feature 1 <) AND (feature 2 >) AND ... . The features include only temperature and experimentally known materials properties listed in Table 1 and Table S1. The subgroups are characterized by a quality value calculated according to a quality function. In our case, the quality is maximized when a subgroup, while being not very small, contains materials with optimal target properties (see Methods section for details).[34] We apply SGD to identify which combinations of the considered primary features maximize zT, α^2, σ, 1/κ, and 1/κ_L. In fact, we explore a range of the quality function parameter constraining the minimum value of a target property within the subgroups. The obtained results are summarized in Table S2.

We find a subgroup of 34 out of 602 data points with zT > 1.5, characterized by doping with Ga or Zn at the B site, doping levels at the A site $C_{A*} > 0.05$, and higher temperatures T > 696 K. The top subgroup for the squared Seebeck coefficient $\alpha^2 >$ ($10^{-12}V^2K^{-2}$) contains materials with Ag but not Cu and no doping ($C_{A*} = 0$) at the A site, and constrains temperature to a moderate range (344 K ≤ T ≤ 618 K). Clearly, the conditions on temperature and C_{A*} for the Seebeck coefficient are in contradiction with the conditions on these features for maximizing zT. The top subgroup for $\alpha^2 >$ ($10^{-12}V^2K^{-2}$) is characterized by a narrower temperature range (489 K ≤ T ≤ 618 K) and exclusion of Zn as B-site dopant. The identified subgroups show that higher but not too high temperatures, Ag at the A site and no Zn at the B-site are beneficial for increasing Seebeck coefficient α^2 within the considered materials class. However, contrary to the zT subgroups, the subgroups with maximum α^2 do not include all the materials with large α^2, indicating that there may be alternative ways of increasing α^2, but the available data are insufficient to identify them as a statistically significant subgroup.

The materials properties leading to a small total thermal conductivity are similar to the properties resulting in increased Seebeck coefficient: Ag at A site, low ($C_{A*} \leq 0.24$) or no doping at A site, and no Zn as a dopant at B site). However, the conditions on temperature are opposite: for lower thermal conductivity, higher temperatures are preferred. For lower thermal lattice conductivity, higher temperature, Ag as a majority species at the A-site, and no Zn dopant at B-site are also preferred, but in addition Bi and In should be excluded as majority species at B-site.

Finally, electrical conductivity is maximized at not too high temperatures and when dopant concentration at A site is small. The conductivity is further increased when also dopant concentration at B-site is small. In addition, the dopants at the A site should not be Cu or Ag (i.e., Zn or Na are better).

We now focus on the analysis of zT using SISSO. In the SISSO method, a huge pool of candidate features of increasing complexity is first constructed iteratively by applying a set of mathematical operators to a set of primary features. The same features (Table 1) are used as primary features for SISSO. The target

$$\mathrm{zT} = \sum_{i=1}^{N} c_i d_i + A,$$

property (zT) is expressed as a linear combination of the complex features ：

The set of is called a descriptor with dimension . *A* is the intercept. Compressed sensing is used to identify best model and the corresponding descriptor for each dimension up to a maximum value. The initial choice of primary features is crucial for the predictive performance of descriptors identified by SISSO. Materials properties such as atomic species and their relative concentrations,[35, 36] atomic

radii,[37, 38] atomic weights,[38] electronegativities,[39, 40] ionization energies,[41] and heats of fusion/vaporization[42-44] are reported to have strong impact on TE performance of different thermoelectric materials. Therefore, we have constructed the primary feature space with these properties.

Table 1. Primary features used for the descriptor construction for the in-house experimentally synthesized ternary based chalcogenide materials A-B-C_2 (where A = Cu or Ag; B = Bi, In, Sb, or Ga; C = Se or Te) with proper doping.

Primary Feature	Element	Symbol
Temperature (K)	–	T
Dopant concentration	A, A*, B, B*,C,C*	C_{A*},C_{B*},C_{C*}
Electronegativity (eV)		EN_A,EN_{A*},EN_B,EN_{B*},EN_C,EN_{C*}
Ionization Energy (eV)		IE_A,IE_{A*},IE_B,IE_{B*},IE_C,IE_{C*}
Heat of fusion (eV)		HF_A,HF_{A*},HF_B,HF_{B*},HF_C,HF_{C*}
Heat of vaporization (eV)		HV_A,HV_{A*},HV_B,HV_{B*},HV_C,HV_{C*}
Atomic Radius (Å)		AR_A,AR_{A*},AR_B,AR_{B*},AR_C,AR_{C*}
Atomic Weight (a.u.)		AW_A,AW_{A*},AW_B,AW_{B*},AW_C,AW_{C*}

The dopants at A, B, and C sites are represented by A* (A* = Cu, Ag, Zn, or Na), B* (B* = Ga, In, Bi, Sb, Zn, or Sn), and C* (C* = Te or Cl), respectively. C_{M*} (M = A, B, C) are concentrations of dopants defined as the fraction of corresponding sites occupied by M*. Minority species ($C_{M*} \leq 0.5$) are always considered as dopant, so that C_{M*} varies between 0 and 0.5. When the concentration of a dopant is zero ($C_{M*} = 0$), the values of EN_{M*}, IE_{M*}, HF_{M*}, HV_{M*}, AR_{M*}, and AW_{M*} are equal to the values of EN_M, IE_M, HF_M, HV_M, AR_M, and AW_M.

In SISSO overfitting may occur with increasing dimensionality of the descriptor. To avoid over-fitting, 10-fold cross-validation (CV10)[45] is employed to identify the optimal dimension of the model. For each cycle displayed in Figure 1, the materials pool is first split into 10 subsets (the materials datasets for each cycle are collected in the Supplementary Data), and the descriptor identification along with the model training is performed using 9 subsets. Then the error in predicting properties of the systems in the remaining subset is evaluated with the obtained model. The CV10 error is calculated as the average value of the test errors obtained for these ten subsets. The results for each cycle are collected in Figure S1. As can be seen in Figure S1, for each cycle the CV10 error reduces gradually from to (1D to 7D) descriptors, which suggests that overfitting does not occur for dimensions up to 7D. This is due to the relatively large

number of training data points for SISSO (although it is still small compared to typical dataset sizes needed for traditional ML approaches such as neural networks). The root-mean-square fitting error (RMSE) for the 7D descriptor for each cycle is already small and is only slightly smaller than for the 6D descriptor. Since higher dimensions mean higher model complexity and computational cost of training, we have carried out all our analysis with up to 7D descriptors. To further confirm the predictive power of our best model (7D) at the final iteration, we have validated the model based on the zT values from previous reports on chalcogenide materials from other groups in Table S3 alongside with the SISSO-predicted values. Despite experimental uncertainties, the predictions remain overall consistent with experiment.

The overall good performance of our best model at the final iteration is demonstrated by low RMSE in zT (0.14). The distribution of errors in predicted zT for different temperature ranges, different zT ranges, and the overall distribution are shown in Figures 2Sa-c. Figure 2 shows the distribution of zT for different temperature ranges in the final dataset. The top five largest deviations between SISSO model predicted zT values and the experimentally measured zT values are collected in Table S4. The maximum absolute error reaches 0.76 for $Ag_{0.55}Cu_{0.45}GaTe_2$ at 730.3 K. In general, the absolute errors larger than 0.6 are all from $Cu_{1-x}Ag_xGaTe_2$ systems with experimentally measured zT values larger than two at temperature larger than 730 K. Fewer data points measured at high temperatures (>750K) and high-zT materials (> 1.2) explain why the prediction errors are larger for these cases. However, for all the materials the SISSO model underestimates zT, i.e., all materials predicted to have good thermoelectric properties can be expected to have even better properties in reality.

The descriptor components , the coefficients , importance score, and the occurrence number of each component during the CV10 processes for the best SISSO model at the final iteration are given in Table 2 (the results for other iterations are collected in Table S5-6). One of the advantages of SISSO over other data analytics approaches is the physical interpretability of the models. By inspecting the descriptor components, one can see how SISSO selects physically meaningful correlations between the target property (zT) and the combinations of primary features. In particular, occurrence of T^2 in reflects the fact that $zT/T = \alpha^2\sigma/\kappa$ consistently increases with temperature for the materials in the considered class. The occurrence of $1\text{-}C_{A*}$ in the denominator of indicates that TE efficiency increases with increasing disorder in the A-site, which reduces thermal conductivity.[4, 46, 47] This is consistent with the results of SGD analysis. It should be noted that the dependence of zT on the dopant concentration is strongly nonlinear. A small dopant concentration can have a strong effect on the TE performance due to the formation of impurity band within the gap.[19] The dependence of and on atomic features indicates that zT is increased for smaller and lighter atoms at the B-site, with at least the minority species (B*) being as light as possible relative to the majority species at the C-site. This is also in agreement with SGD analysis. According to , additional gain in zT can be achieved by choosing B atoms lighter than majority of atoms

at A sites. The disparity of size and weight between B and C/A atoms maximizes the rattling effect, which is responsible for a decrease of thermal conductivity in clathrates.[48]

At the same time, smaller atomic radius of majority species at site B and disorder at site A are detrimental for electrical conductivity, consistent with the dependence of on these features. This reflects the well-known dilemma in solid solutions that impurity scattering will reduce both the thermal conductivity and carrier mobility.[38, 49] The appearance of the heat of vaporization HV_A in can be related to the ability of element A with higher heat of vaporization to form stronger bonds with surrounding atoms and thus also reduce disorder and maintain higher electrical conductivity at elevated temperatures.[42-44] For these reasons, concentration of A* will only improve zT up to a limit, namely while the phonon scattering rises due to the lattice disorder but the carrier mobility remains unaffected.[18] In $Cu_{1-x-\delta}Ag_xInTe_2$, zT reportedly increases only until $x = 0.2$,[50] while in $Cu_{1-x}Ag_xGa_{0.6}In_{0.4}Te_2$ zT reaches its highest value of 1.64 (at 873K) for $x = 0.3$.[15] Thus, our data analysis shows that selecting impurity atoms with small radius is indeed an effective strategy to suppress lattice thermal conductivity and at the same time maintain the carrier mobility.

Table 2. The descriptor components , the coefficients , importance score[$], and the occurrence number of each component during the CV10 processes for the best SISSO model at the final iteration.

descriptor component	coefficient	importance score	occurrence number
$T^2/((1-C_{A*})\bullet AW_B)$	0.78685E-04	0.65 (0.61)	10
$(T/AR_{B*})\bullet\lvert AW_C-AW_{B*}\rvert$	0.13073E-04	0.38 (0.17)	7
$((1-C_{A*})\bullet AR_B\bullet HV_A)/T$	0.12837E+03	0.21 (0.14)	1
$\lvert AW_B-AW_C\rvert/\lvert AW_A-AW_{B*}\rvert$	-0.47799E-01	0.18 (0.16)	1
$((1-C_{A*})\bullet AR_{A*})/(EN_B-EN_C)$	-0.74883E-01	0.04 (0.06)	0
$(AW_A/T)/\lvert HV_B-HV_{A*}\rvert$	0.30740E-01	0.06 (0.03)	0
$(T\bullet C_{B*})/(AR_A-AR_C)$	-0.85704E-04	0.02 (0.01)	2

[$]The importance score based on RMSE (MaxAE) for each descriptor component is calculated as follows. The component is removed from the descriptor, the model is refit with the remaining components, and RMSE and MaxAE are calculated. The score is calculated as 1 – RMSE (MaxAE) (all components) / RMSE (MaxAE) (all – 1 component). Score 0 means that removing the component does not change RMSE (MaxAE) of the model.

Since the SISSO model clearly captures the physics governing the performance of the thermoelectric materials, we proceed with the high-throughput (HT) screening of the huge compositional space of ternary based chalcogenide materials A-B-C_2 (see Supplementary Methods for details). Results of HT screening of more than 10,000 data points (zT values) from ternary A-B-C_2 type chalcogenide compounds are presented in Figure 3a. We have found many new promising TE materials that have not been realized experimentally before. Moreover, based on these results, we have synthesized the compound $Ag_{0.55}Cu_{0.45}GaTe_2$, showing an experimental zT value as high as ~2.8 at 827K, which have been repeated for several times (Figure S3). The results of thermal diffusivity are also reproduced from a third-party laboratory (see Figure S4 and Supplementary Appendix for details). The maximum zT values for other types of thermoelectric materials known to date are displayed in Figures 3b and c for comparison.

The only material with similar or slightly higher zT is SnSe.[17] However, compared to our materials, SnSe with high zT is much harder to synthesize, which inhibits its large-scale use.

The phase stability of $Ag_{0.55}Cu_{0.45}GaTe_2$ can be seen from the XRD pattern in Figure S5. One can also see in Figure 4 that, although increasing the Ag concentration beyond 0.55 increases Seebeck coefficient, it does not reduce the lattice thermal conductivity any further. Rather, the additional Ag lowers the electrical conductivity so that zT decreases. It is exactly this interplay of various parameters that makes the search for high-performance thermoelectric materials so non-trivial.

In summary, we have not only overcome one of the major hurdles of traditional TE material synthesis, namely exploring the very vast configurational space, but have also demonstrated the power of data analytics and machine learning in accelerating thermoelectric materials design. By learning from our own experimental dataset, obtained at consistent well-controlled conditions, we have developed a ML model that allows us to predict many new materials with exceptional TE performance effortlessly and reliably. Due to physical interpretability of our model, we were able to clearly disentangle interconnected, often conflicting effects of basic materials parameters on TE performance, and find a way to overcome the optimization difficulties by fully utilizing the large variability of the parameters within the explored materials class. The complex relationship between the primary features and the target property zT emphasizes the absolute necessity of advanced data analytics tools for the advancement of new functional materials discovery. Moreover, the successful synthesis and characterization of our predicted materials have paved out new lanes in TE research. In addition to the fact that these materials have exceptionally high zT values and stabilities and are therefore very promising for several TE applications, the success of our active-learning ML strategy clearly shows that such concerted approaches have a great potential for future materials design.

Methods

Data-analytics:

The descriptors are obtained with SISSO, using ~600 experimental zT inputs as training data. SISSO is meant to single out a simple yet physically intuitive descriptor from an immensely large set of candidates. Initially, a huge pool consisting of more than ten billion candidate descriptors, is constructed. Then an iterative approach is employed to search the descriptors by combining pre-defined primary features and a set of mathematical operators (+, −, ·, /, log, exp, exp−, $^{-1}$, ^{2}, ^{3}, $\sqrt{}$, $\sqrt[3]{}$, |−|). The complexity of the obtained descriptors depends on how many times the operators are employed. In the present study, we have considered feature space of complexity level up to two, Φ_1 and Φ_2.[29] Any given feature space of complexity *n* (Φ_n) also contains all of the lower (i.e. *n*-1) feature spaces. The details of the SISSO model identification procedures and the high-throughput screening of new materials/compositions at each iteration are given in the Supplementary Methods.

Experimental:

Sample preparation

Bulk samples of polycrystalline $Cu_{1-x}Ag_xGaTe_2$ (x=0-0.6) were prepared by vacuum melting–annealing combined with spark plasma sintering (SPS) using elemental Cu, Ag, Ga and Te (99.999%, Emei Semicon. Mater. Co., Ltd. Sichuan, CN). All the raw materials were weighed and mixed according to the above formulae and sealed in quartz tubes under vacuum, which was gradually heated up to 1273 K at a heating rate of 100 Kh^{-1}, and incubated for 28 h. Afterwards, the ampoules were slowly cooled to 873 K at a rate of 15 Kh^{-1} followed by quenching in water and then dwelt at 813 K for 72 h. The obtained chunks were ball-milled into fine powders for 10 h, and then sintered by the spark plasma sintering apparatus (SPS-1030) at 673 K with a pressure of 55 MPa.

Transport property measurements

The densified bulk samples of size ~2.5×3×12 mm^3 and ϕ10×1.5 mm were prepared for electrical property and thermal diffusivity measurement. The Seebeck coefficients and electrical conductivities were performed with a ZEM-3 device (ULVAC-RIKO, Japan) under a helium atmosphere from room temperature to ~870 K with an uncertainty of < 5.0%. Thermal conductivity (κ) was calculated via $\kappa=DC_p\rho$, where the thermal diffusivity (D) was measured by the laser flash method (TC-1200RH, ULVAC-RIKO, Japan) with a precision of ~10.0% and confirmed by NETZSCH LFA457, Germany. The heat capacities (C_p) were estimated following the Dulong-Petit rule, $C_p = 3nR$ (here n is the number of atoms per formula unit and R gas constant). The sample density (ρ) was measured by the Archimedes method. When calculating the electronic thermal conductivities (κ_e) according to the equation $\kappa_e = L\sigma T$, the Lorenz numbers L were estimated by using the formula L=1.5+exp(-|α|/116) (where L is in 10^{-8} $W\Omega K^{-2}$ and α in

μVK^{-1}).[51] The three physical parameters (α, σ, and κ) were finalized after three measurements. The total uncertainty for zT is ~20%.

Declarations

Acknowledgments

The data-analytics methodology development is supported by RSF grant 21-13-00419. J.C. is supported by the National Natural Science Foundation of China (51671109).

Author contributions

J.C. and Z.-K.H. created the idea and conceived the work. J.C., Z.-K.H., and S.V.L. designed and supervised the project. J.C. supervised the experimental synthesis and analysis. Y.Z. and Q.X. synthesized the materials and measured the properties of the thermoelectrics. X.H. and D.S. performed the data analytics. Y.Z., X.H., D.S., Z.-K.H., J.C., and S.V.L. co-wrote the manuscript. All authors contributed to the analysis and interpretation of the results. All the authors commented on the manuscript and have given approval to the final version of the manuscript.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary Information is available for this paper at http://www.nature.com/nature.

Correspondence and requests for materials should be addressed to J.C. or Z.-K.H. or S.V.L.

References

1. Snyder, G.J. & Toberer, E.S. in Materials for sustainable energy: a collection of peer-reviewed research and review articles from Nature Publishing Group 101-110 (World Scientific, 2011).

2. Sootsman, J.R., Chung, D.Y. & Kanatzidis, M.G. New and old concepts in thermoelectric materials. *Angewandte Chemie International Edition* **48**, 8616-8639 (2009).

3. Dresselhaus, M.S. et al. New directions for low-dimensional thermoelectric materials. *Advanced Materials* **19**, 1043-1053 (2007).

4. Hu, L., Zhu, T., Liu, X. & Zhao, X. Point defect engineering of high-performance Bismuth-Telluride-based thermoelectric materials. *Advanced Functional Materials* **24**, 5211-5218 (2014).

5. Fu, C. et al. Realizing high figure of merit in heavy-band p-type half-Heusler thermoelectric materials. *Nature Communications* **6**, 1-7 (2015).

6. Saiga, Y., Du, B., Deng, S., Kajisa, K. & Takabatake, T. Thermoelectric properties of type-VIII clathrate Ba8Ga16Sn30 doped with Cu. *Journal of Alloys and Compounds* **537**, 303-307 (2012).

7. Rogl, G. et al. In-doped multifilled n-type skutterudites with ZT= 1.8. *Acta Materialia* **95**, 201-211 (2015).

8. Hong, M. et al. Achieving zT> 2 in p-Type AgSbTe2− xSex Alloys via Exploring the Extra Light Valence Band and Introducing Dense Stacking Faults. *Advanced Energy Materials* **8**, 1702333 (2018).

9. Wu, Y. et al. Lattice strain advances thermoelectrics. *Joule* **3**, 1276-1288 (2019).

10. Zhong, B. et al. High superionic conduction arising from aligned large lamellae and large figure of merit in bulk Cu1. 94Al0. 02Se. *Applied Physics Letters* **105**, 123902 (2014).

11. He, Y. et al. Ultrahigh thermoelectric performance in mosaic crystals. *Advanced Materials* **27**, 3639-3644 (2015).

12. Zhao, K. et al. Are Cu2Te-Based Compounds Excellent Thermoelectric Materials? *Advanced Materials* **31**, 1903480 (2019).

13. Chang, C. et al. 3D charge and 2D phonon transports leading to high out-of-plane ZT in n-type SnSe crystals. *Science* **360**, 778-783 (2018).

14. Hong, M. et al. Rashba effect maximizes thermoelectric performance of GeTe derivatives. *Joule* **4**, 2030-2043 (2020).

15. Zhang, J. et al. Design of domain structure and realization of ultralow thermal conductivity for record-high thermoelectric performance in chalcopyrite. *Advanced Materials* **31**, 1905210 (2019).

16. Roychowdhury, S., Panigrahi, R., Perumal, S. & Biswas, K. Ultrahigh Thermoelectric Figure of Merit and Enhanced Mechanical Stability of p-type AgSb1−x Zn x Te2. *ACS Energy Letters* **2**, 349-356 (2017).

17. Zhou, C. et al. Polycrystalline SnSe with a thermoelectric figure of merit greater than the single crystal. *Nature Materials*, 1-7 (2021).

18. Biswas, K. et al. High-performance bulk thermoelectrics with all-scale hierarchical architectures. *Nature* **489**, 414-418 (2012).

19. Heremans, J.P. et al. Enhancement of thermoelectric efficiency in PbTe by distortion of the electronic density of states. *Science* **321**, 554-557 (2008).

20. Liu, H. et al. Copper ion liquid-like thermoelectrics. *Nature Materials* **11**, 422-425 (2012).

21. Tan, G. et al. High thermoelectric performance of p-type SnTe via a synergistic band engineering and nanostructuring approach. *Journal of the American Chemical Society* **136**, 7006-7017 (2014).

22. Wang, S., Wang, Z., Setyawan, W., Mingo, N. & Curtarolo, S. Assessing the thermoelectric properties of sintered compounds via high-throughput ab-initio calculations. *Physical Review X* **1**, 021012 (2011).

23. Ricci, F. et al. An ab initio electronic transport database for inorganic materials. *Scientific Data* **4**, 170085 (2017).

24. Garrity, K.F. First-principles search for n-type oxide, nitride, and sulfide thermoelectrics. *Physical Review B* **94**, 045122 (2016).

25. Gorai, P., Stevanović, V. & Toberer, E.S. Computationally guided discovery of thermoelectric materials. *Nature Reviews Materials* **2**, 1-16 (2017).

26. Carrete, J., Li, W., Mingo, N., Wang, S. & Curtarolo, S. Finding unprecedentedly low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling. *Physical Review X* **4**, 011019 (2014).

27. Trunschke, A. et al. Towards Experimental Handbooks in Catalysis. *Topics in Catalysis*, 1-17 (2020).

28. Gaultois, M.W. et al. Data-driven review of thermoelectric materials: performance and resource considerations. *Chemistry of Materials* **25**, 2911-2920 (2013).

29. Ouyang, R., Curtarolo, S., Ahmetcik, E., Scheffler, M. & Ghiringhelli, L.M. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Physical Review Materials* **2**, 083802 (2018).

30. Wrobel, S. in European symposium on principles of data mining and knowledge discovery 78-87 (Springer, 1997).

31. Friedman, J.H. & Fisher, N.I. Bump hunting in high-dimensional data. *Statistics and Computing* **9**, 123-143 (1999).

32. Meeng, M. & Knobbe, A. For real: a thorough look at numeric attributes in subgroup discovery. *Data Mining and Knowledge Discovery* **35**, 158-212 (2021).

33. Goldsmith, B.R., Boley, M., Vreeken, J., Scheffler, M. & Ghiringhelli, L.M. Uncovering structure-property relationships of materials by subgroup discovery. *New Journal of Physics* **19**, 013031 (2017).

34. Mazheika, A. et al. Ab initio data-analytics study of carbon-dioxide activation on semiconductor oxide surfaces. *arXiv preprint arXiv:1912.06515* (2019).

35. Yusufu, A. et al. Thermoelectric properties of Ag1− x GaTe2 with chalcopyrite structure. *Applied Physics Letters* **99**, 061902 (2011).

36. Shen, J. et al. Vacancy scattering for enhancing the thermoelectric performance of CuGaTe 2 solid solutions. *Journal of Materials Chemistry A* **4**, 15464-15470 (2016).

37. Yang, J., Meisner, G. & Chen, L. Strain field fluctuation effects on lattice thermal conductivity of ZrNiSn-based thermoelectric compounds. *Applied Physics Letters* **85**, 1140-1142 (2004).

38. Liu, Y. et al. Lanthanide contraction as a design factor for high-performance half-Heusler thermoelectric materials. *Advanced Materials* **30**, 1800881 (2018).

39. Lee, J.-H., Wu, J. & Grossman, J.C. Enhancing the thermoelectric power factor with highly mismatched isoelectronic doping. *Physical Review Letters* **104**, 016602 (2010).

40. Zhang, J., Song, L., Borup, K.A., Jørgensen, M.R.V. & Iversen, B.B. New insight on tuning electrical transport properties via Chalcogen doping in n-type Mg3Sb2-based thermoelectric materials. *Advanced Energy Materials* **8**, 1702776 (2018).

41. Awadalla, S.A. et al. Isoelectronic oxygen-related defect in CdTe crystals investigated using thermoelectric effect spectroscopy. *Physical Review B* **69**, 075210 (2004).

42. Ando, Y., Miyamoto, N., Segawa, K., Kawata, T. & Terasaki, I. Specific-heat evidence for strong electron correlations in the thermoelectric material (N a, C a) Co 2 O 4. *Physical Review B* **60**, 10580 (1999).

43. Cui, T., Xuan, Y., Yin, E., Li, Q. & Li, D. Experimental investigation on potential of a concentrated photovoltaic-thermoelectric system with phase change materials. *Energy* **122**, 94-102 (2017).

44. Cui, T., Xuan, Y. & Li, Q. Design of a novel concentrating photovoltaic–thermoelectric system incorporated with phase change materials. *Energy Conversion and Management* **112**, 49-60 (2016).

45. Shao, J. Linear model selection by cross-validation. *Journal of the American Statistical Association* **88**, 486-494 (1993).

46. Liu, Z. et al. Understanding and manipulating the intrinsic point defect in α-MgAgSb for higher thermoelectric performance. *Journal of Materials Chemistry A* **4**, 16834-16840 (2016).

47. Mao, J. et al. Defect engineering for realizing high thermoelectric performance in n-type Mg3Sb2-based materials. *ACS Energy Letters* **2**, 2245-2250 (2017).

48. Tadano, T. & Tsuneyuki, S. Quartic anharmonicity of rattlers and its effect on lattice thermal conductivity of clathrates from first principles. *Physical Review Letters* **120**, 105901 (2018).

49. Deng, R. et al. High thermoelectric performance in Bi 0.46 Sb 1.54 Te 3 nanostructured with ZnTe. *Energy & Environmental Science* **11**, 1520-1535 (2018).

50. Liu, R. et al. Thermoelectric performance of Cu 1− x− δ Ag x InTe 2 diamond-like materials with a pseudocubic crystal structure. *Inorganic Chemistry Frontiers* **3**, 1167-1177 (2016).

51. Kim, H.-S., Gibbs, Z.M., Tang, Y., Wang, H. & Snyder, G.J. Characterization of Lorenz number with Seebeck coefficient measurement. *APL materials* **3**, 041506 (2015).

Figures

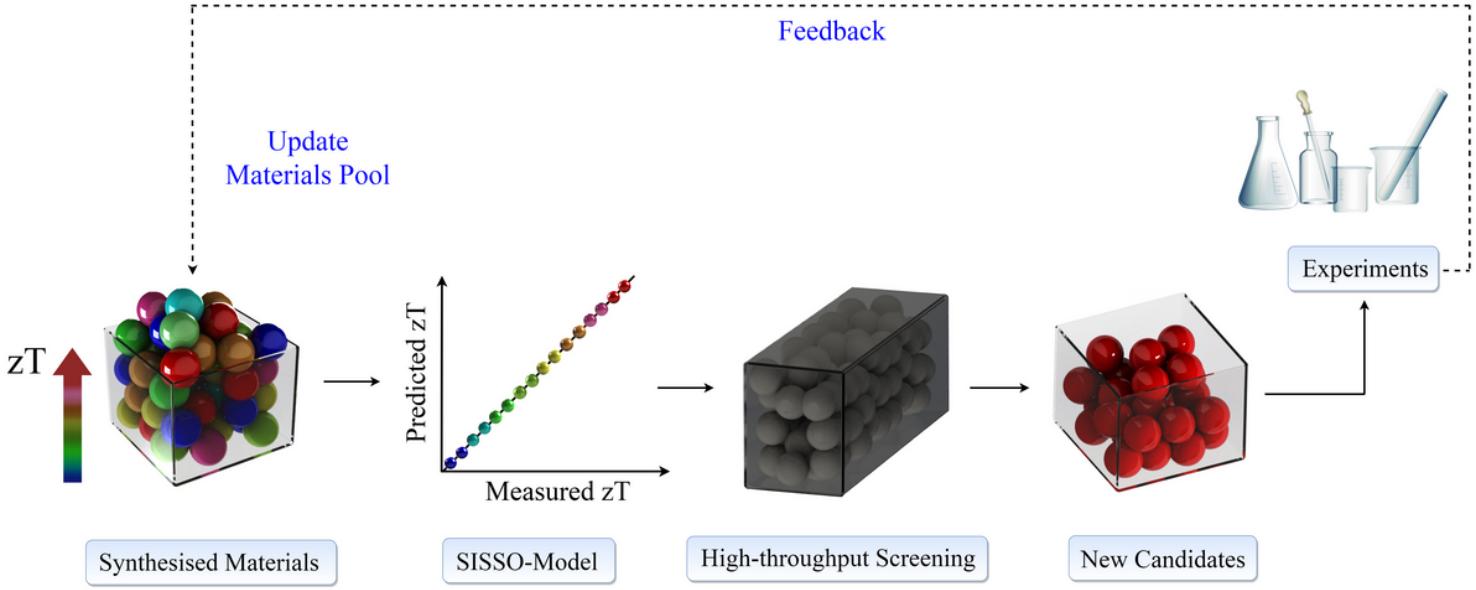

Figure 1

Schematic representation of materials design with active learning.

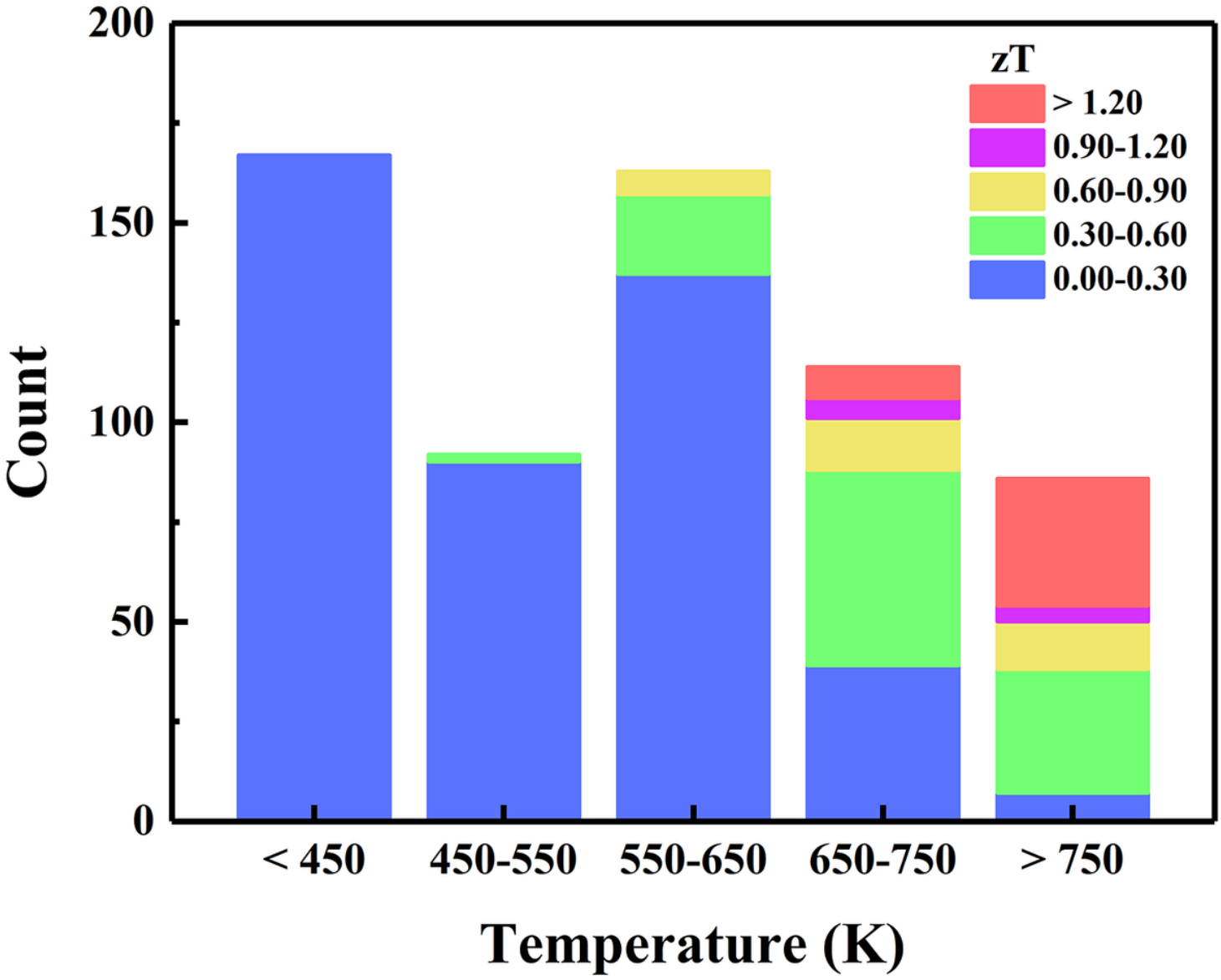

Figure 2

Distribution of the data within the final dataset for different temperature ranges and zT ranges.

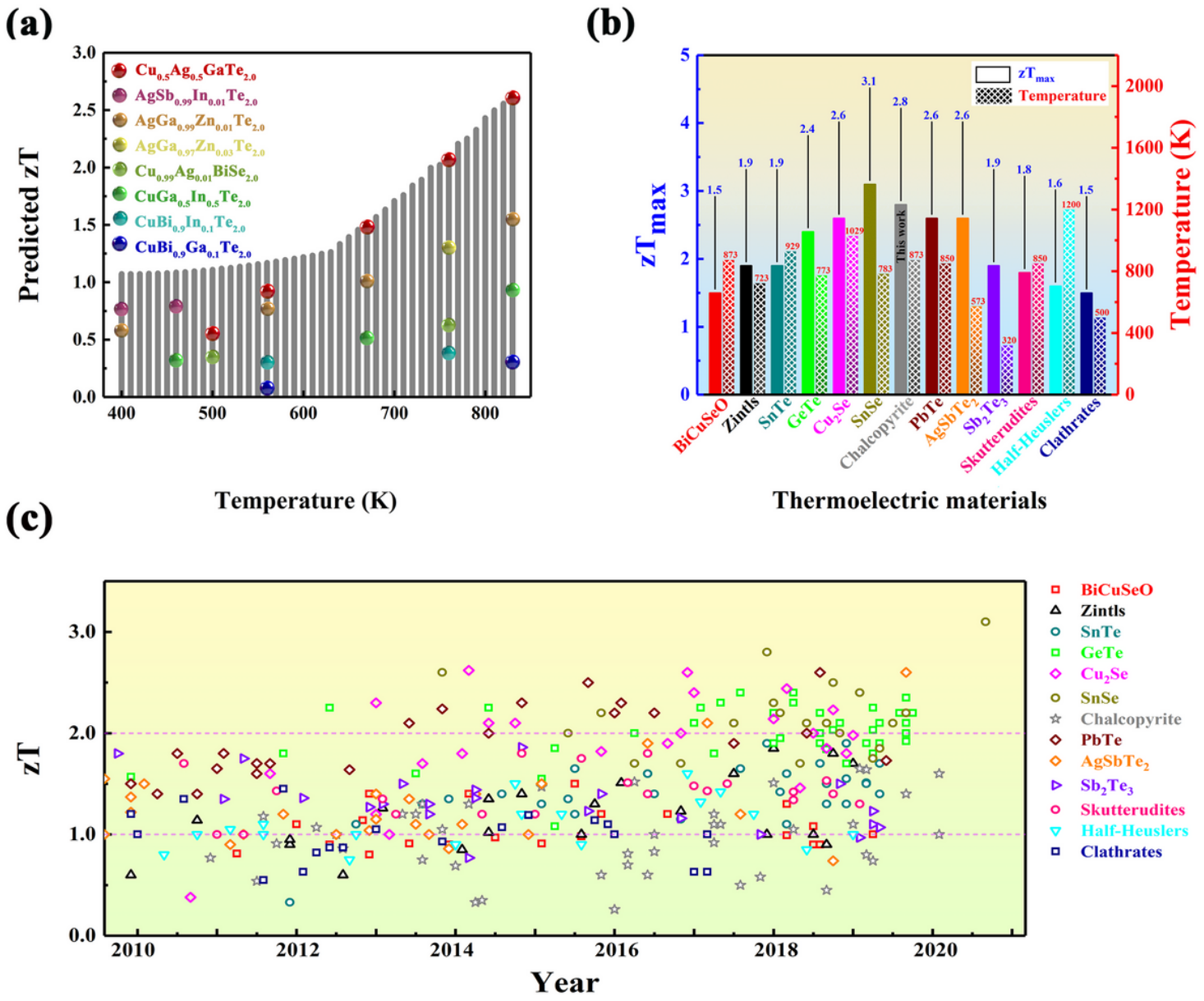

Figure 3

(a) Results of high-throughput search for new thermoelectric materials using the SISSO model. The height of the gray bar represents the maximum predicted zT value from the highthroughput screening. (b) Comparison of the experimental maximum zT values in other types of thermoelectric materials. (c) Experimental zT values of established systems since 2009 (the zT values and original literatures are collected in Supplementary file).

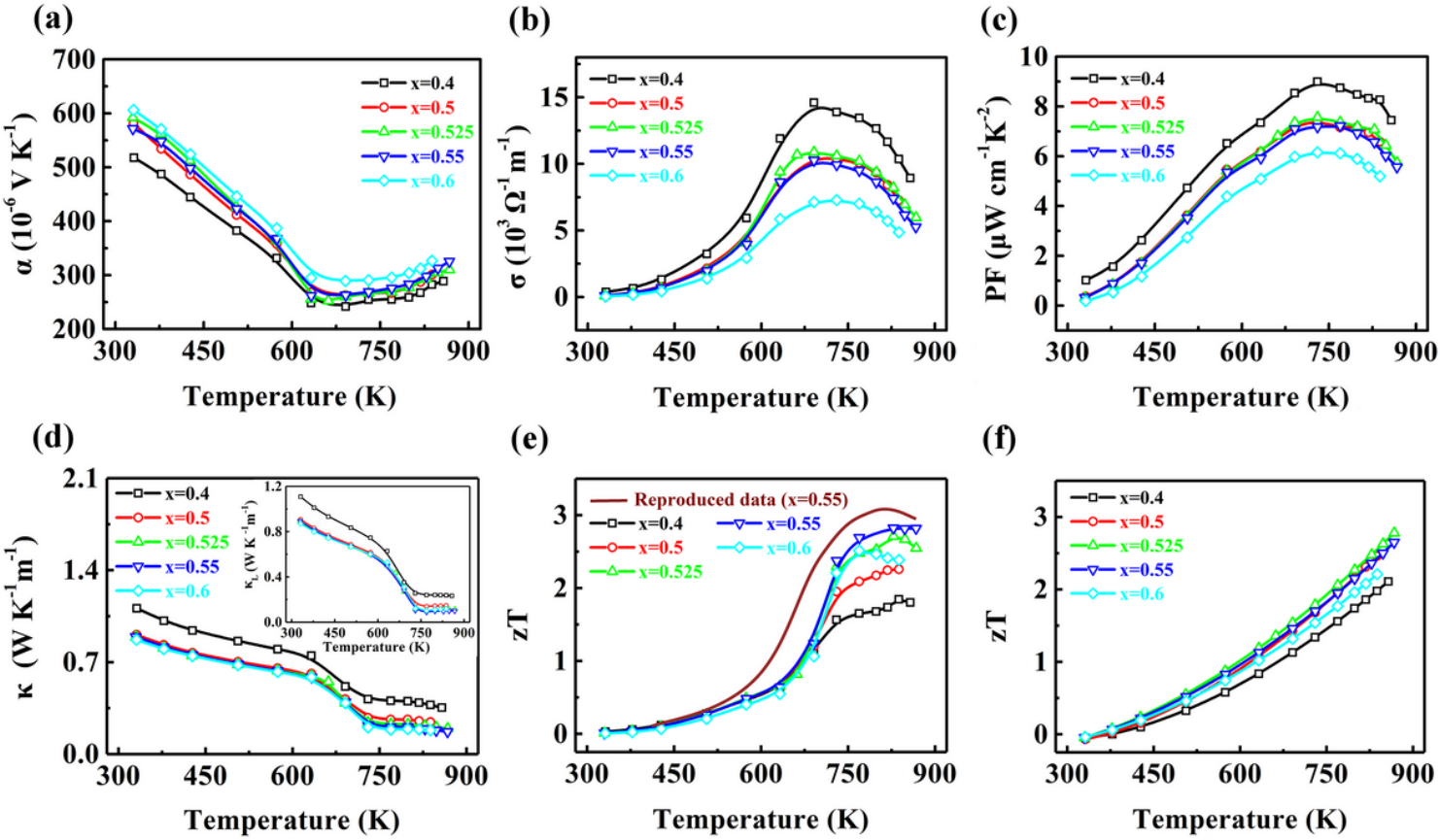

Figure 4

Measured parameters of the bulk Cu1-xAgxGaTe2 (x = 0.4-0.6) as a function of temperature: (a) Seebeck coefficients (α); (b) Electrical conductivities (σ); (c) Power factors (PF); (d) Total thermal conductivities (κ) and lattice thermal conductivities (κL); (e) Experimentally measured figure of merit (zT); (f) SISSO-predicted figure of merit (zT).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFile.docx
- SupplementaryInformation.docx