

Prediction of Seebeck coefficient for compounds without restriction to fixed stoichiometry: A machine learning approach

Al'ona Furmanchuk,¹ James Saal², Jeff W. Doak², Gregory B. Olson², Alok Choudhary³, Ankit Agrawal³

Correspondence to: Al'ona Furmanchuk (E-mail: alona.furmanchuk@northwestern.edu)

¹ A. Furmanchuk

Center for Health Information Partnerships, Northwestern University, Feinberg School of Medicine, Chicago, IL 60611, USA.

² J.Saal, J.W. Doak, G.B. Olson

QuesTeck Innovations LLC, Evanston, IL 60201, USA

³ A. Choudhary, A. Agrawal

Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208, USA.

ABSTRACT

The regression model-based tool is developed for predicting the Seebeck coefficient of crystalline materials in the temperature range from 300K to 1000K. The tool accounts for the single crystal vs. polycrystalline nature of the compound, the production method, and properties of the constituent elements in the chemical formula. We introduce new descriptive features of crystalline materials relevant for the prediction the Seebeck coefficient. In order to address off-stoichiometry in materials, the predictive tool is trained on a mix of stoichiometric and non-stoichiometric materials. The tool is implemented into a web application (<http://info.eecs.northwestern.edu/SeebeckCoefficientPredictor>) in order to assist field scientists in the discovery of novel thermoelectric materials.

model capable of predicting the Seebeck coefficient for any arbitrary compound would dramatically improve the ability to discover and design new thermoelectric materials.

The success of this task relies on the possibility to find a universal set of descriptive features for any material with any charge transport mechanism. The underlying physics of charge transport in a material varies significantly with change in chemical bonding, the crystal structure and the reciprocal space features, as well as various scattering mechanisms. In this situation the descriptive feature set is at risk to be limited or specific to charge transport mechanism. Other important and far from being well understood contributors to thermoelectricity are feature describing the role⁵⁻⁷ of the production conditions. The task of collecting and unifying available information into a detailed and structured database is also complicated. The scientific practices have yet to come up with universal rules on reporting synthesis conditions. In addition, not all studies focused on thermoelectric properties of materials also report data on structural properties. As a result, the discovery process in the field mainly relies upon the chemical intuition of material scientists and affordability of experimental setup for material discovery routines.

Valuable contributions⁸⁻¹⁴ are emerging from the scientific community applying density functional theory (DFT) for prediction of Seebeck coefficient. Unfortunately,

Introduction

A temperature gradient applied to a material gives rise to a voltage difference across it. This fundamental electronic transport phenomenon plays a key-role in thermoelectric generators^{1, 2} and Peltier coolers.^{3, 4} The quantitatively the effect could be described with the coefficient of proportionality, the Seebeck coefficient, between the applied temperature gradient and resulting voltage difference. A generalized

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version record](#). Please cite this article as [doi: 10.1002/jcc.25067](https://doi.org/10.1002/jcc.25067).

This article is protected by copyright. All rights reserved.

underestimation of band gap is a well-known flaw of DFT, which usually results¹⁵⁻¹⁸ in overestimating bipolar conduction and thus underestimating the Seebeck coefficient at high temperatures and low doping levels. Another problem of DFT is failure to evaluate carrier's scattering due to the limits of constant relaxation time approximation. Thus, Seebeck can be under or overestimated, since the relaxation time is not a constant, and depends on scattering processes.¹⁹

The evolving field of materials informatics²⁰ has already provided machine learning-based solutions^{7, 21-25} for assessing several materials' properties in an efficient way. In the field of thermoelectric materials, only a few attempts have been made to apply machine learning to predict physical properties that constitute thermoelectric material figure-of-merit. They were limited to clustering analysis^{26, 27} and do not predict numerical values of thermoelectric properties.

Here we developed regression models that predict numerical value of the Seebeck coefficient in the temperature range between 300 K and 1000K. We trained our models on one-of-the-kind collection of experimental samples and their production methods (the UCSB database²⁸). The experimentally synthesized materials is a great choice for a training set since they provide possibility to link Seebeck coefficient with random structural alterations resulting from differences in production set up, and intentional or unintentional low level doping. Those are hard to capture with DFT analyses of perfect crystalline compounds. For convenience of using our predictive model in screening of novel materials, we implemented it into a web application *ThermoEI* (Part S1, SI).

Methods

Curation of the UCSB data set

The details on collection of the materials set used in this work was described elsewhere.^{28, 29} The original data set had multiple records for

selected compounds, as they were reported by different experimental groups. Few records reported variation within experimental uncertainty, while others varied significantly. Such duplicate records were withheld for further investigation (Table S1 in Part S2, SI). Initial curation, - reduced original data set (1082 compounds) by 130 compounds. The distribution of Seebeck coefficients from the UCSB dataset at different temperatures (Table S2 in Part S2, SI) suggested the presence of some outliers that could adversely affect the accuracy of the models. Such data points were discarded from all subsets in order to obtain a continuous distribution of data for each temperature (Fig.1). Since our target is thermoelectric applications, removing outliers with extremely large Seebeck coefficients could be an acceptable measure.

For each compound from the original UCSB database we extracted the following characteristics: chemical composition (formula), preparation method, crystallinity, temperature, Seebeck coefficient ($S@T$), figure-of-merit ($ZT@T$). All the properties, except for ZT , were used in attribute generation (see details below) for regression modelling. The experimental samples were reported for four temperature regimes: 300 K, 400 K, 700 K, and 1000 K. Not all finally selected 927 materials were tested at all temperatures. Authors did not attempt to impute missing values. Therefore, temperature specific subsets of varying sizes ($\sim 200-256$ compounds) were used for predictive modeling at different temperatures.

Attribute generation

The initial set of descriptive features provides a three-level characterization for the material. At the first level, we characterize sample features (labelled as Ψ^i) at the macro scale by considering the production method and crystallinity of the compound. The macro scale attributes are categorical.

At the second level we focus on properties of the elements within the compound. Elemental properties we considered (labelled as X^i in equations 1-8) fall into three categories: (i)

location of the element in the periodic table, (ii) fundamental properties of the elements, (iii) experimentally measured properties of pure elements in their crystalline states. Properties within category (i), the location of elements in the periodic table, include atomic number, period, and group, and whether or not the element can be classified as an alkali, alkaline earth, transition metal, post transition metal, metalloid, lanthanide, actinide, non-metal, halogen, or noble gas. Properties within category (ii), fundamental properties of the elements, include atomic weight, molar volume, Pauling's electronegativity, covalent radius, atomic radius,³⁰ ionic radius,³⁰⁻³² pseudo-potential radii sum of Zunger,³³ amount of valence electrons by Villars,³⁴ total number of valence electrons as well as specified by their s-, p-, d-, and f-character, and overall number of unfilled valence orbitals as well as those of s-, p-, d-, and f-character. Finally, properties within category (iii), experimentally measured properties of pure elements in their crystalline states,³⁵⁻³⁸ include crystal radius, melting point, boiling point, density, heat of vaporization, thermal conductivity, electron affinity, ionization energy, and ground-state crystal structure of the element.

At the third level we make use of simple features describing each compound: the number of elements in the compound under consideration, and the coefficients of elements in the chemical formula of the compound. The coefficients were

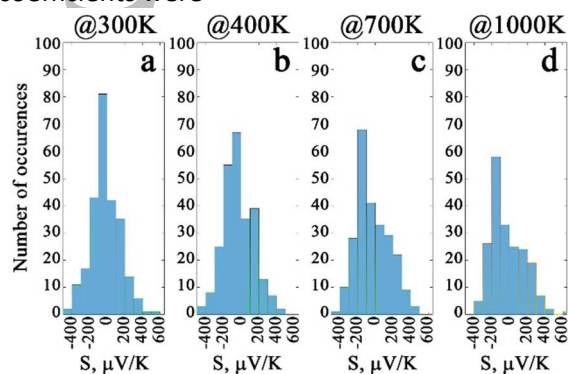


Figure 1. Seebeck coefficient distribution at (a) 300K, (b) 400K, (c) 700K, and (d) 1000K temperatures for pre-processed data subsets.

also used in deriving coefficient-weighted features.

For every property X^i we built ten features where the first five are minimum (ψ_{min}^i), maximum (ψ_{max}^i), sum (ψ_{SUM}^i), mean ($\bar{\psi}^i$), and mean absolute deviation from mean (MADFM) value (ψ_{MADFM}^i) of element's property X^i present in a material. The rest of five features are coefficient-weighted analogues of min, max, sum, MADFM, and mean. Equations 5 and 7 provide examples of coefficient-weighted mean and coefficient-weighted MADFM attributes, respectively.

$$\psi_{min}^k = \min(X_j^i), \quad (1)$$

$$\psi_{max}^k = \max(X_j^i), \quad (2)$$

$$\psi_{sum}^k = \sum_{j=1}^N X_j^i, \quad (3)$$

$$\bar{\psi}_{mean}^k = \frac{\sum_{j=1}^N X_j^i}{N}, \quad (4)$$

$$\bar{\psi}_{weighted_mean}^k = \frac{\sum_{j=1}^N x X_j^i}{N}, \quad (5)$$

$$\psi_{MADFM}^k = \frac{\sum_{j=1}^N \left| \frac{\sum_{j=1}^N \bar{\psi}^i}{N} - \bar{\psi}^i \right|}{N}, \quad (6)$$

$$\psi_{weighted_MADFM}^k = \frac{\sum_{j=1}^N \left| \frac{\sum_{j=1}^N \bar{\psi}_{weighted}^i}{N} - \bar{\psi}_{weighted}^i \right|}{N}, \quad (7)$$

where X^j is the property of interest (from the list described above), j is a chemical element in compound k ; N is the total number of elements in the formula of the compound k , and x is the coefficient next to the element symbol in the chemical formula. There are a total of 452 attributes.

Data mining

The aim of the regression analysis is to generate a statistical model that can predict a dependent variable (Seebeck coefficient) based on independent variables (attributes). Among all techniques we applied initially, methods belonging to lazy learners (IBk,³⁹ KStar⁴⁰) and ensemble of decision trees (Regression by discretization, Random subspace,⁴¹ Random

forest) categories produce strong correlation between measured and predicted Seebeck coefficient. We chose to use the Random Forest algorithm⁴² in our regression analysis, as it produced the most accurate cross-validated models among all four temperature regimes. The Random Forest algorithm consists of a collection of ensembles of simple tree predictors. Each tree is built on a random selection of 2/3 of the total training set. A fixed-size subset of attributes is sampled at random from the whole list of attributes during growing the tree at each level (so-called node). The best split on this subset is used to split the node in the tree. Each tree is capable of producing a prediction that is evaluated based on the remaining 1/3 of training data. As an ensemble learning algorithm, Random Forest generates predictions by combining multiple simple tree predictors. The outcome for any given new set of materials is predicted by running the feature records of materials through all the trees and averaging the predictions. This method is very robust to noise and over-fitting due to the presence of randomness and ensemble averaging in its routines. This robustness is especially important when training on small or noisy⁴³ datasets.

The Random Forest algorithm was used as implemented in the Scikit-Learn library in Python 2.7 language. Hyper parameters were optimized using the grid search as implemented in Scikit-Learn.⁴⁴ Generated models were evaluated using 5-fold cross-validation and leave-one-out cross-validation⁴⁵ error rate estimators.

The evaluation of the prediction accuracy of a regression model is characterized using the Pearson product moment correlation coefficient (R), coefficient of determination (R^2), mean absolute error (MAE), root mean squared error (RMSE), relative absolute error (RAE), and relative squared error (RSE):

$$R = \frac{N(\sum_{i=1}^N P_i \Psi_i) - (\sum_{i=1}^N P_i)(\sum_{i=1}^N \Psi_i)}{\sqrt{N(\sum_{i=1}^N P_i^2) - (\sum_{i=1}^N P_i)^2} \sqrt{N(\sum_{i=1}^N \Psi_i^2) - (\sum_{i=1}^N \Psi_i)^2}}, \quad (8)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (\Psi_i - P_i)^2}{\sum_{i=1}^N (\Psi_i - \bar{\Psi})^2}, \quad (9)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (P_i - \Psi_i)^2}, \quad (10)$$

$$RAE = \frac{\sum_{i=1}^N |P_i - \Psi_i|}{\sum_{i=1}^N |\Psi_i - \bar{\Psi}|}, \quad (11)$$

$$RSE = \frac{\sum_{i=1}^N (P_i - \Psi_i)^2}{\sum_{i=1}^N (\Psi_i - \bar{\Psi})^2}, \quad (12)$$

where P_i , Ψ_i , $\bar{\Psi}$ are the predicted, actual values for data point i , and the mean of all Ψ_i respectively.

The coefficient of determination is a measure of the proportion of variance in the predicted variable (here Seebeck coefficient). It is commonly assumed that when $R^2 > 64\%$, a strong correlation exists between the actual values and the values predicted by the regression. While many researches prefer coefficient of determination as a main measure for model accuracy, there is no commonly accepted statistical measure of accuracy of the model. Therefore, we use all mentioned measures (equations 8-12) in order to evaluate model according to all available standards. All graphical visualization was executed with MATLAB⁴⁶ programming language.

Results and Discussion

The grouping of experimentally synthesized materials into families was attempted with the use of unsupervised learning (Part S3, SI). All 452 attributes were used for the evaluation of materials similarity. Analysis produced the six materials families with a varying population. We found that good thermoelectric materials do not belong to a single family, rather populate all families identified by unsupervised learning. This finding motivated us to focus on a generalized predictive model rather than on separate models specific to materials group. Initial temperature-specific models were trained with all 452 attributes. The leave-one-out cross-validation was used at this stage to estimate model accuracy. Original publications for poorly predicted materials (outliers) were manually checked. For those, we frequently

found mistakes in the semi-automatic process of data extraction such as the wrong sign, mistyped values, wrong or missing information on chemical composition (see notes in table S2 in SI, Part S2).

The Seebeck coefficients for a subset of duplicate entries extracted earlier (Methods) was checked for text extraction mistakes in the same way (Table S1 in SI, Part S2). For the same compounds with similar experimental values, we averaged over all Seebeck coefficient records and added back to the training sets. Not all compounds from this list were moved to our curated training sets though. For example, we completely excluded compound Bi_2Te_3 at 300K. It is obvious that averaging over p- and n-doped materials ($162 \mu\text{V/K}$ and $-174 \mu\text{V/K}$) would lead to meaningless value $-6 \pm 237.588 \mu\text{V/K}$. In this case, one should avoid including averaged value as a ground truth for the regression model. We preferred to remove materials from analysis if authors did not provide enough details for conversion of doping information into explicit chemical formula notation (non-stoichiometric notation).

Upon review of some publications, doping does not appear as obvious reason for seemingly similar materials being reported. In many cases, finding real reasons is hard or even impossible task. Therefore, one should decide on reliability threshold for experimentally reported variations. In our case we chose to use the error of our initial regression models as extra threshold for material's exclusion from the curated data set. We only used compounds with experimental measurement variation less than the root mean squared error (RMSE) of our regression model. For example, RMSE of regression model at 1000K is $101.57 \mu\text{V/K}$. This value is smaller than variation of Seebeck coefficient for ZnO (from $-190.963 \mu\text{V/K}$ to $-436.926 \mu\text{V/K}$) compound measured at 1000K. Therefore, ZnO was not added to training data set at 1000K. The ZnO compound from Table S1 (Part S2, SI) was not included in data set at 700K for the same reason. For 300K model we did not include two CaMnO_3 compounds obtained by different production methods.

Our model does not rely on explicit carrier concentration since such information is not always available in the literature, and omitted in the UCSB database. Such information is partially contained within the chemical formula, as well as information on materials crystallinity and production method. Therefore, the method will be insensitive to difference between n- Bi_2Te_3 and p- Bi_2Te_3 unless it is stated explicitly as non-stoichiometric chemical formula. However, carrier concentration is indirectly available to our models via the exact coding of dopants in chemical formula of the material, its synthesis, and crystallinity. One might argue that in order for our approach to generate correct predictions, the chemical formula should include information on impurities as well. We not only agree with it, we believe such a step could possibly address existence of remaining unexplained outliers. However, accurately addressing situations with impurities is impossible with **currently available data**. For the sake of the study, we assume that presence of impurities is an inherent property of the production method. Readers also have to be aware that current reporting style on synthesis/production method varies in degree of provided details from article to article. Therefore, our list of 45 production methods could not be based on standardized protocols regulating set of synthesis variables.

After the extra curation step, we continued with further optimization of predictive model. The initial model was robust enough to predict erroneous records in the data set as outliers. However, it is impractical to use all 452 attributes when performing extensive screening for material of interest. In order to correctly reduce number of features in the models we combined domain knowledge with a stepwise backward elimination procedure. The domain knowledge is based on the idea that information about the carrier concentration could be encoded through the crystallinity and production method. The relevance of rest of the attributes is hard to guess based on chemical intuition since they are mostly based on properties of chemical elements constituting

the material (equations 1-7). For those we used stepwise backward elimination technique that ranks the importance of all attributes first, and then removes worst attributes at every step of optimization. At each step, we build a regression model and estimate its accuracy with the five statistical evaluation metrics.

The best model (Table 1) has the highest values for Pearson coefficient (R) and coefficient of determination (R^2), and the lowest values for the rest of error measures (MAE, RMSE, RAE, RSE). As one may see from Table 1, the best performing models are those with 187 and 87 attributes for lower (300K, 400K) and higher (700K, 1000K) temperature sets, correspondingly. We chose to go with models based on 187 attributes (Fig. 2).

Accuracy

As listed in Table 1, RMSE values for the best optimized models (Fig. 2) are below 84 $\mu\text{V/K}$. The range of Seebeck coefficients in the UCSB dataset is from -400 $\mu\text{V/K}$ to + 400 $\mu\text{V/K}$. A prediction error of 84 $\mu\text{V/K}$ corresponds to uncertainty within 11%. In order to check our models for transferability to other material spaces, an external test set of twenty materials was collected outside of the UCSB database (Table 2). These materials were recently manufactured and tested for their thermoelectric performance, including Seebeck coefficient. All four models were found to produce high accuracy predictions ($R^2 \geq 0.88$). The importance of current models accuracy could be appreciated if placed into context of materials screening for thermoelectric device applications. A material's performance in a thermoelectric device is characterized by the thermoelectric figure-of-merit, ZT, which heavily relies on Seebeck coefficient:

$$ZT = \frac{S^2 \sigma T}{k} \quad (13)$$

Together with the absolute temperature, T , Z describes the interplay between the Peltier cooling (given by the Seebeck coefficient, S), the

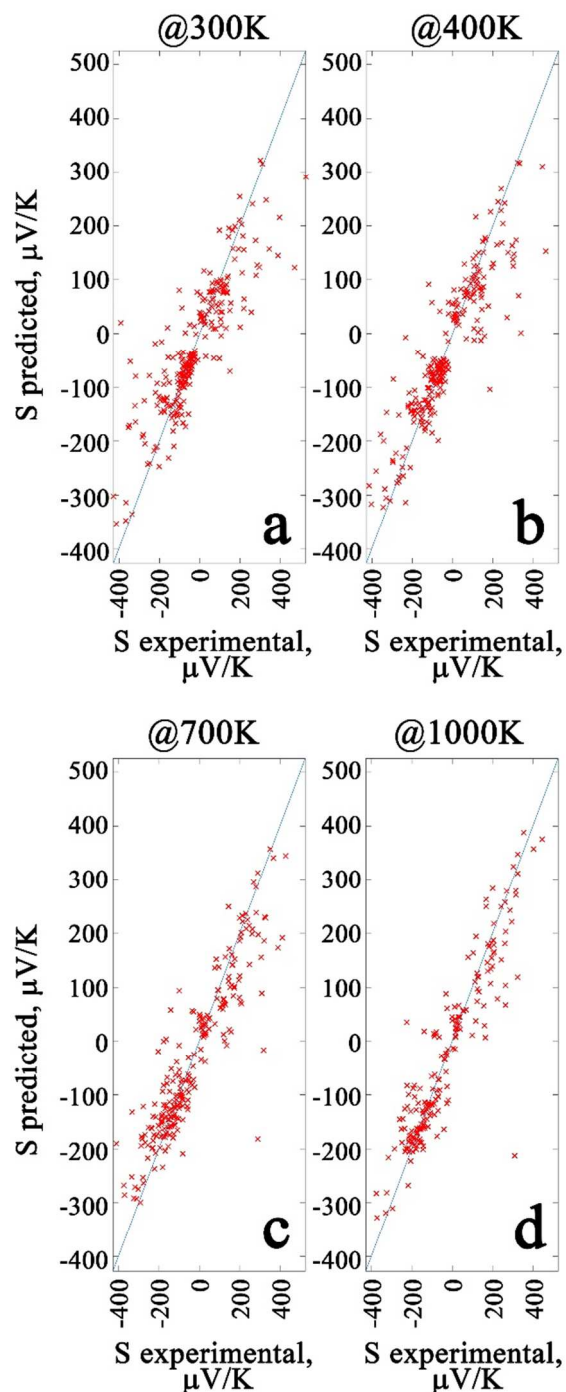


Figure 2. Plots show performance of the different models (all models are based on 187 attributes). Line in blue corresponds to condition when experimental value is equal to predicted at (a) 300K, (b) 400K, (c) 700K, and (d) 1000K temperatures.

Table 1. Accuracy of regression models predicting Seebeck coefficient at different steps of attribute reduction. Estimated through the correlation coefficient (R), coefficient

of determination (R^2), mean absolute error (MAE), root mean squared error (RMSE), relative absolute error (RAE), and relative squared error (RSE).

# of attributes in regression mode	R	R^2	MAE, $\mu\text{V/K}$	RMSE, $\mu\text{V/K}$	RAE, %	RSE, %
1000K						
452	0.90	0.81	49.22	77.52	33.36	19.33
387	0.91	0.82	47.29	73.53	32.05	17.40
287	0.90	0.82	48.91	75.42	33.15	18.30
187	0.91	0.84	44.63	71.33	30.25	16.37
87	0.92	0.84	43.88	69.67	29.74	15.62
50	0.86	0.74	45.77	79.34	40.36	26.35
700K						
452	0.90	0.81	51.26	80.46	34.45	20.37
387	0.90	0.80	52.70	80.97	35.42	20.63
287	0.90	0.80	53.15	81.41	35.72	20.86
187	0.91	0.82	51.33	77.26	34.50	18.78
87	0.90	0.82	51.22	77.10	34.42	18.71
50	0.87	0.75	60.34	89.59	40.55	25.26
400K						
452	0.88	0.78	50.42	78.74	38.86	23.09
387	0.88	0.77	51.72	81.54	39.86	24.76
287	0.88	0.77	53.74	81.18	41.42	24.54
187	0.89	0.79	49.27	77.52	37.97	22.38
87	0.89	0.79	49.51	76.18	38.16	21.61
50	0.86	0.74	53.71	83.84	41.40	26.18
300K						
452	0.85	0.73	55.42	85.05	44.84	28.33
387	0.85	0.73	55.20	85.66	44.66	28.74
287	0.84	0.71	57.17	87.33	46.26	29.87
187	0.86	0.74	53.73	83.15	43.47	27.08
87	0.85	0.72	55.34	84.66	44.78	28.07
50	0.83	0.68	56.45	90.13	45.67	31.82

“Joule” heating in the semiconductor (given by electrical conductivity, σ), and heat conduction from the hot to the cold side (given by thermal conductivity, k). All three parameters in the figure-of-merit are functions of charge carrier concentration.

The electrical and thermal conductivities increase as the carrier concentration increases, while Seebeck coefficient decreases. Optimization of electrical power factor, $S^2\sigma$, has

revealed^{47, 48} that in good thermoelectric materials the charge carrier concentration should be around 10^{25} m^{-3} . Such carrier concentrations correspond to heavily-doped semiconducting materials with Seebeck coefficients much higher than in metals.

Conventionally, materials possessing a $ZT > 1.0$ are regarded as good thermoelectric materials. Despite naïve expectations, real-world materials with large absolute values of Seebeck are typically not good thermoelectrics (Fig. 3).

During the screening for potential good candidates, one would like to focus on materials with Seebeck coefficient value around $\pm 300 \mu\text{V/K}$. It is important to avoid situations when material with almost zero Seebeck coefficient (bad thermoelectric) will be predicted as potentially good thermoelectric ($\pm 300 \mu\text{V/K}$), in other words false positives. Therefore, it is important that the method have a prediction error less than $100 \mu\text{V/K}$.

Despite overall good performance of our models we identified several outlier materials that have prediction error (difference between experimental and predicted values) equal or more than $200 \mu\text{V/K}$. The overall analysis show that wrong prediction of sign by our model is only observed in 5-6% of cases, and only few of them (Table 3) resulted in significant errors.

We would like to discuss few cases from Table 3 that might make reader think that the current set of descriptive features is not optimal. The $\text{Sr}_{0.61}\text{Ba}_{0.39}\text{Nb}_2\text{O}_6$ case suggests that features describing structural information are important. We discovered that slight variation in annealing conditions could result in presence of small amount of Nb^{4+}O_2 as a second phase in the $\text{Sr}_{0.61}\text{Ba}_{0.39}\text{Nb}_2\text{O}_6$ sample. This phase will alter the conduction mechanism and affect Seebeck coefficient at 300K and above. The original study⁴⁹ specified that the magnitude of the Seebeck coefficient is dependent on the crystal anisotropy. In other words, randomly oriented and textured polycrystalline samples had significant discrepancy in values of their Seebeck coefficient. The value of the coefficient at 300K

Table 2. External data set collected outside of UCSB database.

Formula	Production method	Seebeck coefficient, experiment				Seebeck coefficient, predicted			
		300 K	400 K	700 K	1000 K	300 K	400 K	700 K	1000 K
1.21(ZnO)x0.008(Zn ₂ TiO ₄ Al _{0.18})	Solid state reaction in air ⁵⁰	-144	-158	-184	-207	-146.89	-130.39	-156.82	-198.21
1.207(ZnOAl _{0.07})x0.008(Zn ₂ TiO ₄ Al _{0.16})	Solid state reaction in vacuum (experiment used in N ₂) ⁵⁰	-104	-117	-151	-174	-131.91	-132.42	-146.51	-191.17
1.205(ZnOAl _{0.12})x0.008(Zn ₂ TiO ₄ Al _{0.065})	Solid state reaction in vacuum (experiment used in N ₂ +CO) ⁵⁰	-53	-59	-76	-84	-138.29	-141.26	-149.8	-209.58
Fe _{0.9} Co _{0.1} Ga _{2.65} Ge _{0.35}	Spark plasma sintering ⁵¹	-70				-80.27			
Fe _{0.75} Co _{0.25} Ga _{2.65} Ge _{0.35}		-45				-80.72			
Fe _{0.5} Co _{0.5} Ga _{2.65} Ge _{0.35}		-45				-79.54			
Fe _{1.0} Ga _{2.85} Ge _{0.15}		-118				-86.86			
Fe _{1.0} Ga _{2.75} Ge _{0.25}		-85				-85			
Fe _{1.0} Ga _{2.65} Ge _{0.35}		-80				-88.09			
Ca ₃ Co ₄ O ₉	Spark plasma sintering ⁵²	133	132	151	165	91.99	104.72	130.96	177.04
Ca ₃ Co _{3.95} Cd _{0.05} O ₉		152	153	172	187	66.38	79.27	88.05	173.73
Ca ₃ Co _{3.9} Cd _{0.10} O ₉		141	141	184	210	68	81.2	94.6	172.29
Ca ₃ Co _{3.85} Cd _{0.20} O ₉		146	145	184	190	69.28	80.79	97.66	169.41
Yb ₁₄ MnSb ₁₁	Spark plasma sintering ⁵³	47	75	138	193	39.7	36.57	96.05	104.31
Yb _{13.8} Sc _{0.3} Mn _{0.98} Sb _{10.86}		43	75	137	203	9.15	18.33	53.98	77.48
Yb _{13.79} Y _{0.29} Mn _{0.99} Sb _{10.90}		30	58	130	192	6.62	16.66	48.11	72.02
Yb _{13.70} Y _{0.32} Mn _{1.01} Sb _{10.94}		42	76	129	192	9.76	15.7	50.1	71.8
Bi _{1.0} Cu _{1.0} Te _{1.0} O _{1.0}	Spark plasma sintering ^{23, 54, 55}	172.5	185	193.75		75.39	74.34	135.74	
Bi _{1.0} Cu _{1.0} Se _{1.0} O _{1.0}		178	175	185		84.17	80.54	146.8	
Bi _{1.0} Cu _{1.0} S _{1.0} O _{1.0}		213	202	160		74.61	78.14	150.21	
R						0.96	0.95	0.96	0.94
R ²						0.91	0.91	0.92	0.88
MAE, μ V/K						46.92	62.57	55.50	62.69
RMSE, μ V/K						60.03	69.79	63.02	80.21
RAE, %						47.24	74.05	55.43	45.60
RSE, %						28.71	41.04	25.21	26.09

for sample with randomly oriented grains was $\approx -150 \mu\text{V/K}$, and the values for textured grains varied from $-160 \mu\text{V/K}$ to $-450 \mu\text{V/K}$, depending on the orientation used for measurement. After averaging over all orientations for textured samples, one will end up with $\approx -220 \mu\text{V/K}$.

If one assumes that production method leads to similar chemical and phase compositions, observed difference of $70 \mu\text{V/K}$ can be attributed to grain size and alignment. Our

model predicts Seebeck coefficient to be $-19 \mu\text{V/K}$ without being aware of the presence of second phase, structural alignment/misalignment, and degree of granularity within the sample.

The Seebeck coefficient of undoped CaMnO_3 stands separately from trends recorded for its

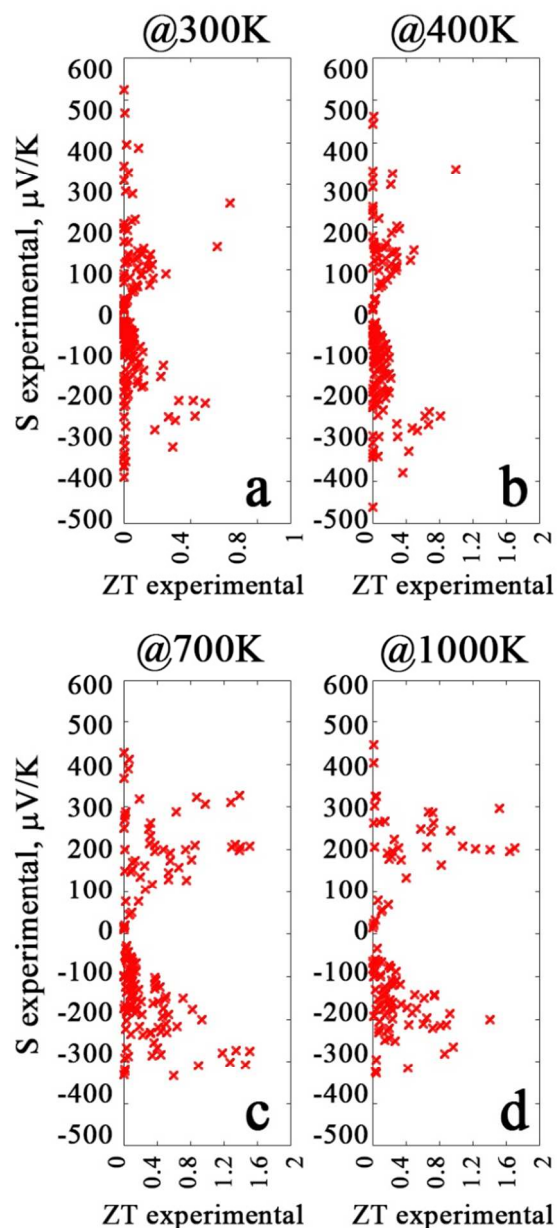


Figure 3. Relationship between Seebeck coefficients and thermoelectric materials' figure-of-merit measured at (a) 300K, (b) 400K, (c) 700K, and (d) 1000K temperatures.

doped forms at low-to-intermediate temperatures. It was speculated⁵⁶ that possible reasons could be difference in charge carrier concentration, and degree of contribution of hopping mechanism in CaMnO_3 and its perovskite-type oxides. The variation in ionic radii between Ca and a dopant ion could lead to a variation in the hopping inter-site distance

and, as a result, alter the carrier's mobility.⁵⁶ Despite the fact that ionic radius was included as one of attributes in our models, we were not able to correctly reproduce the type of carriers in CaMnO_3 . We suspect unspecified synthesis details might be affecting mechanism of charge transfer in the case of undoped CaMnO_3 .

Strontium doped perovskite-type LaCoO_3 showed very similar trends for the Seebeck coefficient in agreement with the trend for undoped LaCoO_3 , except for temperatures below 400K. Our model failed to predict the Seebeck coefficient for undoped LaCoO_3 at 300K as it is reported by experimental group.⁵⁷ Study of the literature⁵⁸ on various of undoped LaCoO_3 revealed the discrepancy in reports on the nature and concentration of charge carriers for undoped samples at low temperatures. Such disagreement below 400K is observed even if undoped samples are produced under the same conditions. There is no single opinion on the nature of such variation. Formation of oxygen defects, contaminations, and deviation from stoichiometric composition are believed to be responsible for this.

It is possible to come across publications that report alteration of carrier's concentration due to unknown reasons. One of such examples is sintered $\text{Bi}_{2-x}\text{Y}_x\text{Ru}_2\text{O}_7$ samples,⁵⁹ where $x=0.0, 0.5, 1.0, 1.5, 2.0$. The experimental sample with composition $x=2$ showed deviation from other compounds of the group once temperature drops below 873K. For this temperature zone, authors reported significant thermal activation of holes for $x=2$ sample only. Unfortunately, they were not able to determine the origin of the holes. Our models failed to predict Seebeck coefficient (errors $\sim 240 \mu\text{V/K}$) for this sample at 400K. At the same time, prediction errors for this compound at other temperatures are within $0.4 \mu\text{V/K} - 9 \mu\text{V/K}$.

Besides the fact that not many attempts exist toward creation of predictive models for Seebeck coefficient, the cases discussed above explain why it is hard to achieve accurate predictions. We see the current predictive model as an inexpensive tool for preliminary rapid estimation of promising materials.

Table 3. List of outliers. The differences (in $\mu\text{V/K}$) are given between predicted and experimental values of Seebeck coefficient. For cases when error is due to wrong sign prediction an alternative estimation based on absolute values is given in parentheses.

Chemical Formula	Data set @T			
	@1000K	@700K	@400K	@300K
Ba8Ga18Ge28	523.02 (96.98)	472.14 (107.86)	289.28 (80.72)	220.75 (79.24)
Ba8Ga16Sn30			225.59	
CuFe0.9Cr0.1O2	206.16			
Mg2Si0.98Bi0.02	259.30 (188.70)			
Ag9TiTe5		221.71	256.66	243.25
Ca0.9In0.1MnO3		220.25		
Ca3AlSb3		219.78		
Ca5Al2Sb6		216.65	310.58	347.69
Mg2Si0.6Ge0.4Ag0.02		335.04		
Ba8Au5.14Si39.51			209.59 (26.61)	
CaMnO3			366.92	
Nd2Cu0.98Zn0.02O4			208.87	291.10
Tl9BiTe6			336.58	216.07
Y2Ru2O7			240.37	
CuCr0.99Mg0.01O2				234.41
LaCoO3				409.82
Sr0.61Ba0.39Nb2O6				200.99
Zr0.3Hf0.3Ti0.4NiSn				236.30

The top features contributing to the Seebeck coefficient

One might want to know what features of complex experimental samples determine variation in Seebeck coefficient or the value itself. Due to the characteristics mentioned earlier (in the “Data mining” subsection of Method section), the Random Forest algorithm is known to be one of the best black-box regression methods. It is superb at finding hidden connections between predictor variables (attributes of material) and response variable (Seebeck coefficient) even if no linear correlation is present. It is also well-known for being one of the hardest methods for interpretation by humans since it generates no analytical equation as outcome. In our case the situation is complicated by the fact we are building a general model for samples delocalized in materials space. Different mechanisms could define Seebeck in different materials. Therefore, unique set of features could be specific to materials structure and mechanism types. The Random Forest model

would try to incorporate all those mechanism-specific features into single model. At the end of the day, one would not be able to tell exactly what feature set is specific to properties of perovskites or Zintl compounds.

It is still possible to get some sense of which features played the most important role in the construction of predictive model. This estimation comes from counting the number of times a training set passes through a node whose decision is based on a given attribute. Attributes that appear often and high up the tree are counted as more important since they would be frequently evaluated to make the predictions. Averaging those counts over several randomized trees reduces the variance of its estimate, and can be reported as the attribute importance. It should be clear that one must not think of attribute importance values in this case as direct analogue of coefficients in a linear regression model.

For the sake of future comparison and development, the list of all 187 attributes is provided for each model in Table S9 (Part S4, SI). Cross-correlation between features in all models is shown in figures S3-S5 (Part S5, SI). We observe no strong correlation between any single feature and Seebeck coefficient. We see benefits in discussion of top ten important attributes (Table 4), since we developed most attributes ourselves, and many of them have never been discussed before in relation to Seebeck coefficient. The notation used for attributes is AB, where A is the name of the element's property (X^i) used for calculation of attribute (ThermalConductivity means the thermal conductivity of each element in chemical formula); and B corresponds to attribute type (Min, Max, Sum, Mean, MADFM) defined in equations 1-4, 6. The corresponding coefficient-weighted attributes (see examples in equations 5 and 7) are marked as AB*.

At first, we would like to highlight the overall significance of coefficient-weighted attributes. They constitute 45.5%, 43.85%, 43.85%, 46.52% of attributes list (Table S9 in Part S4, SI) for models @300K, @400K, @700K, and @1000K, correspondingly. We hypothesize that they

might help to refine model sensitivity when it comes to effects of low concentrations of dopants. Attributes that appear within the top ten list in most models are ThermalConductivityMax*, ThermalConductivitySum*, and ThermalConductivityMean*. These features have weak positive correlation with Seebeck coefficient (not shown here). In general, coefficient-weighted and non-weighted attributes based on the thermal conductivity of

Table 4. Top ten important attributes as utilized by Random Forest model.

Attribute†	Importance,%
Model @300K	
ThermalConductivityMean	1.86
ThermalConductivitySum	1.82
ThermalConductivityMax*	1.71
ThermalConductivityMADFM	1.58
NValenceSum*	1.53
ThermalConductivitySum*	1.50
ThermalConductivityMax	1.36
ThermalConductivityMean*	1.26
ThermalConductivityMADFM*	1.24
AtomicNumberMax*	1.23
Model @400K	
ThermalConductivityMean	1.89
CrystalStructCenteredTetragonal	1.81
ThermalConductivityMax*	1.73
ThermalConductivitySum*	1.60
ThermalConductivityMADFM	1.58
ElectronegativityByMillarMin	1.55
ThermalConductivityMean*	1.48
ThermalConductivitySum	1.44
ElectronegativityByMillarMADFM	1.24
ElectronAffinitySum*	1.19
Model @700K	
ThermalConductivitySum	2.86
IonizationEnergySum*	2.18
ThermalConductivitySum*	2.15
ThermalConductivityMax*	1.70
NValenceSum*	1.65
ElectronegativityByMillarMin	1.61
ThermalConductivityMean	1.51
ThermalConductivityMax	1.46
DensityMADM*	1.41
DensityMax*	1.39
Model @1000K	
NUnfilledMean	9.09
ThermalConductivityMean*	7.43
ThermalConductivitySum*	7.29
ThermalConductivityMean	4.93
ThermalConductivityMADFM*	3.52
ThermalConductivityMax*	3.47
DensityMADM*	3.17
ElectronegativityByMillarMin	2.13
IonizationEnergySum*	1.92
ElectronAffinitySum*	1.87

chemical elements in their ground-state crystal structures turned out to be very important attributes for modelling the Seebeck coefficient

at all temperatures, which is not too surprising since it is a key thermoelectric property. Different from other features such as covalent radius, electron affinity, electronegativity, or various occupations of electronic orbitals, the thermal conductivity of chemical elements are used first time here in modeling of thermoelectric materials.

The thermal conductivity has two main contributions: electronic and lattice. Therefore, attributes based on mathematical operations with thermal conductivity of elements, especially coefficient-weighted ones, provide indirect information on the electronic transport properties of the compound. The other elemental-based attributes such as electron affinity, electronegativity together with information on occupancy of electronic orbitals, number of valence electrons, and similar features are connected with the type of charge carriers present in material. Different types of radii, such as ionic, covalent, and so on, should stipulate tree splitting for different types of bonding, and bond orders in the structure. Together with density attributes, these radii are expected to represent the degree of packing in the structure. Radii, density, production method and crystallinity are our only tools to address structural features of a material. It is interesting to note that some attributes more frequently define split of the nodes (present at the top ten list) at higher temperature regimes. For example, density-, electron affinity-electronegativity-, and ionization energy-based attributes and overall numbers of unfilled orbitals all seem to be quite important.

DensityMADM, the parameter indirectly characterizing heterogeneity of the density within the sample, reflects how much individual elements in the composition differ in density from mean value of density. Excess of such elements could happily form separate phases, or, in trace concentrations, create single point density fluctuation in the structure. Those are good phonon scatters elements. For low-temperature models we observe lower importance and lower positive correlation with Seebeck for this feature.

We see weak negative correlation between the minimum value of electronegativity of elements in material (ElectronegativityByMillarMin) and Seebeck. At the same time, sum of electron affinities (ElectronAffinitySum*), and mean absolute deviation from mean for the electronegativity,

(ElectronegativityByMillarMADM - the lesser importance contributor) have weak positive-to-no correlation with Seebeck coefficient. The overall insight here is that materials predominantly made of atoms strongly attracting electrons (high value of ElectronegativityByMillarMin) might not be good thermoelectric materials. This makes sense since such atoms could be playing the role of continuous traps for moving electrons. The preferable materials might be the ones made of elements with varying electron attraction powers (higher value of ElectronegativityByMillarMADM).

The correlation of the sum of ionization energies of elements (IonizationEnergySum*) in sample changes from weak positive to no correlation while moving from higher to lower temperature models. This could again be linked to the ease of ionization for certain elements at higher temperatures.

The alteration of attribute importance could also be partially explained by the variation in material types in each temperature training set. Selection of only certain type of materials for tests within a specified temperature regime could be attributed to *a priori* knowledge of materials property such as melting temperature or phase transformation, and so on. The reported measurements collected from the literature in this case will cover selective temperature regions only.

Let us assume that our set of attributes is good enough to indirectly sense differences in carrier concentrations, structural changes and mechanism changes due to alternative production. Then, because of the difference in sampling at lower and higher temperatures, certain mechanisms dominant at lower temperatures will not contribute much in higher temperature models, and vice-versa. We are

stressing here on non-contribution to the top ten list of features only. Such feature is still present in full list with lower rank of importance.

Some features were categorical, e.g., crystal structures of chemical elements, information on crystallinity or production method of sample. We found interesting that the model @ 400K very frequently uses attribute CrystalStructCenteredTetragonal for growing trees in all models regardless of number of attributes used. This attribute states that at least one of the elements in chemical formula forms centered tetragonal structure in its pure form. This attribute is used less frequently in other temperature models and, therefore, appears lower on the list of attribute importance.

As one might see from the full list of attributes (Table S9 in Part S4, SI), production methods are not refined enough to be used as key-attributes for building the trees. We are hesitant to go into analysis of their order according to importance because it is not clear to what degree the 45 production methods could capture intricacies of experimental protocols.

Finally, we would like to remind that our models were trained on data sets partially consisted of doped compounds. As a result, models are sensitive to doping information and sometimes provide “wrong” predictions if dopant information is omitted in the chemical formula. These outliers could be used by scientists for finding debatable measurements to be addressed with follow up experiments.

Conclusions

The study addresses the issue of fast property prediction for experimentally synthesized materials. We developed machine-learning model for predicting the numerical value of Seebeck coefficient of a material at four different temperatures (300K, 400K, 700K and 1000K). We have proposed the use of Random Forest (RF) as a higher accuracy alternative to some other popular machine learning methods. The five-fold cross-validation of our RF models

proved satisfactory prediction accuracy ($R^2=0.74-0.84$). All predictive models have been implemented in our web application ThermoEl ([http://info.eecs.northwestern.edu/SeebeckCo-efficientPredictor-SI, Part S4](http://info.eecs.northwestern.edu/SeebeckCo-efficientPredictor-SI,Part-S4)).

The scientific significance is in development of a novel material representation that has been used for the first time in prediction of Seebeck coefficient. Specifically, attributes based on the thermal conductivity of chemical elements were found to be very important for accurate model building. We also showed significance of using coefficient-weighted attributes, in which property of elements are multiplied by their coefficients in chemical formula. In our other module of ThermoEl toolkit, we have successfully implemented the same coefficient-weighted attributes for prediction of Bulk modulus of experimentally synthesized materials²⁴. Some of those materials (including non-stoichiometric) would be hard or impossible to address by DFT-based methods. We designed the “Seebeck” module of ThermoEl toolkit for field scientists seeking a fast and inexpensive screening tool. It will complement expensive and highly inefficient trial-and-error experimental high throughput campaigns.

Despite success, it is clear that machine learning methods with mostly atomic level structural features are not enough for accurate prediction of most thermoelectric properties. Details of synthesis, experimental conditions, material microstructure and phase diagram, carrier concentration will significantly improve accuracy of predictive models. With this in mind, a possible extension of the presented work lies in the exhaustive collection of such information for known thermoelectric materials.

Our outlier analysis also shows that machine learning models can help to some extent in tracing erroneous collection in the databases, and are thereby capable of addressing a growing public concern about the irreproducibility of experimental data and mistakes propagating through peer-reviewed scientific publications. However, as a

community, we must also take a systematic approach to ensure the quality of materials data through better data generation, reporting, and storage.

Acknowledgments

We thank Prof. Taylor Sparks for various technical discussions on collection of parameters comprising the UCSB database. The access to UCSB database was provided through Citration in accordance with requirements of the SIMPLEX project. All authors gratefully acknowledge thermoelectrics research at Northwestern University through the Center for Hierarchical Materials Design (CHiMaD) and primary financial support from the DARPA SIMPLEX program through SPAWAR (Contract #N66001-15-C-4036). In addition, AA and AC were supported in part by the following grants: NIST award 70NANB14H012; AFOSR award FA9550-12-1-0458; NSF award CCF-1409601; and DOE awards DE-SC0007456, DE-SC0014330.

Notes

‡ Meanings of some non-trivial abbreviations: NValence –number of valence electrons as listed in ref.⁶⁰; NUnfilled – Number of unfilled valence orbitals, where NUnfilled = 0 if shell is unoccupied, Maximum-Filled if occupied; CrystalStructCenteredTetragonal – at least one element in the compound forms centered tetragonal crystal structure in its pure state.

References

1. A. Date, L. Gauci, R. Chan and A. Date, *Renew. Energ.*, 2015, **83**, 256-269.
2. D. Madan, Z. Wang, P. K. Wright and J. W. Evans, *Appl. Energ.*, 2015, **156**, 587-592.
3. J. Sharp, J. Bierschenk and J. Lyon, H.B. , *Proc. IEEE*, 2006, **94**, 1602-1612.

4. W.-H. Chen, C.-C. Wang and C.-I. Hung, *Energy Convers. Manage.*, 2014, **87**, 566-575.
5. J. M. Simard, D. Vasilevskiy, F. Belanger, J. L'Ecuyer and S. Turenne, Beijing, China, 2001.
6. J. Tervo, A. Manninen, R. Ilola and H. Hänninen, *State-of-the-art of Thermoelectric Materials Processing, Properties and Applications*, Finland, 2009.
7. O. Isayev, D. Fourches, E. N. Muratov, C. Oses, K. Rasch, A. Tropsha and S. Curtarolo, *Chem. Mater.*, 2015, **27**, 735-743.
8. S. Azam, S. A. Khan, J. Minar, W. Khan, H. U. Din, R. Khenata, G. Murtaza, S. Bin-Omran and S. Goumri-Said, *Semicond. Sci. Technol.*, 2015, **30**, 105018.
9. S. Lemal, N. Nguyen, J. de Boor, P. Ghosez, J. Varignon, B. Klobes, R. P. Hermann and M. J. Verstraete, *Phys. Rev. B*, 2015, **92**, 205204.
10. P. D. Borges and L. Scolfaro, *J. Appl. Phys.*, 2014, **116**, 223706.
11. T. B. Nasr, H. Maghraoui-Meherzi and N. Kamoun-Turki, *J. Alloys Compd.*, 2016, **663**, 123-127.
12. J. Zhou, B. Liao, B. Qiu, S. Huberman, K. Esfarjani, M. S. Dresselhaus and G. Gang Chen, *Proc. Nat. Acad. Sci.*, 2015, **112**, 14777-14782.
13. N. Hirayama, T. Iida, S. Morioka, M. Sakamoto, K. Nishio, Y. Kogo, Y. Takanashi and N. Hamada, *J. mater. Res.*, 2015, **30**, 2564-2577.
14. X.-F. Yang, W.-Q. Zhou, X.-K. Hong, Y.-S. Liu, X.-F. Wang and J.-F. Feng, *J. Chem. Phys.*, 2015, **142**, 024706.
15. J. Yang, H. Li, T. Wu, W. Zhang, L. Chen and J. Yang, *Adv. Func. Mater.*, 2008, **18**, 2880-2888.
16. L. Bjerg, G. K. H. Madsen and B. B. Iversen, *Chem. mater.*, 2011, **23**, 3907-3914.
17. K. P. Ong, D. J. Singh and P. Ping Wu, *Phys. Rev. Lett.*, 2011, **83**, 115110.
18. D. J. Singh, *Phys. Rev. B.*, 2010, **81**, 195217.
19. P. D. Borges, D. E. S. Silva, N. S. Castro, C. R. Ferreira, F. G. Pinto, J. Tronto and L. Scolfaro, *J. Solid State Chem.*, 2015, **231**, 123-131.
20. A. Agrawal and A. Choudhary, *APL Mater.*, 2016, **4**, 1-10.
21. R. Jalem, T. Aoyama, M. Nakayama and M. Nogami, *Chem. Mater.*, 2012, **24**, 1357-1364.
22. A. Seko, T. Maekawa, K. Tsuda and I. Tanaka, *Phys. Rev. B.*, 2014, **89**, 054303.
23. O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo and A. Tropsha, *Nature Comm.*, 2017, **8**, 15679.
24. A. Furmanchuk, A. Agrawal and A. Choudhary, *RSC Adv.*, 2016, **6**, 95246-95251.
25. R. Liu, A. Kumar, Z. Chen, A. Agrawal, V. Sundararaghavan and A. Choudhary, *Nature Sci. Rep.*, 2015, **5**, 11551.
26. M. W. Gaultois, A. O. Oliynyk, A. Mar, T. D. Sparks, G. J. Mulholland and B. Meredig, *APL Mater.*, 2016, **4**, 053213.
27. Sparks T.D., Gaultois M.W., Oliynyk A, Brgoch J and M. B., *Scripta Materialia*, 2015, **111**, 10-15.
28. M. W. Gaultois, T. D. Sparks, C. K. H. Borg, R. Seshadri, W. D. Bonificio and D. R. Clarke, UCSB database. (2013) Available at: <http://www.mrl.ucsb.edu:8080/datamine/thermoelectrics.jsp> (Accessed: 13 August 2015)).
29. M. W. Gaultois, T. D. Sparks, C. K. H. Borg, R. Seshadri, W. D. Bonificio II and D. R. Clarke, *Chem. Mater.*, 2013, **25**, 2911-2920.
30. E. Clementi and D. L. Raimondi, *J. Chem. Phys.*, 1963, **38**, 2686-2689.
31. J. C. Slater, *J. Chem. Phys.*, 1964, **41**, 3199-3204.
32. J. C. Slater, *Quantum Theory of Molecules and Solids. Symmetry and Bonds in Crystals.*, McGraw-Hill, New York, 1965.

3. *Structure and Bonding in Crystals*, Academic Press, New York, 1981.
4. P. Villars, *J. Less Common Met.*, 1985, **109**, 93-115.
5. R. D. Shannon and C. T. Prewitt, *Acta Cryst.*, 1969, **B25**, 925-946.
6. R. D. Shannon, *Acta Cryst.*, 1976, **A23**, 751-761.
7. in *CRC Handbook of Chemistry and Physics*, ed. D. R. Lide, CRC Press, Boca Raton, Florida, 2003, ch. 10.
8. P. Villars and J. L. C. Daams, *J. Alloys Compd.*, 1993, **197**, 177-196.
9. D. Aha and D. Kibler, *Mach. Learn.*, 1991, **6**, 37-66.
0. J. G. Cleary and L. E. Trigg, Tahoe City, CA, 1995.
1. T. K. Ho, *IRRR Trans. Pattern Anal. Mach. Intelligence*, 1998, **20**, 832-844.
2. L. Breiman, *Mach. Learn.*, 2001, **45**, 5-32.
3. M. Fernandez-Delgado, E. Cernadas, S. Barro and D. Amorin, *J. Mach. Learn. Res.*, 2014, **15**, 3133-3181.
4. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825-2830.
5. S. Arlot, *Stat. Surv.*, 2010, **4**, 40-79.
6. MATLAB and Statistics Toolbox Release 2012b).
7. C. Wood, *Rep. Prog. Phys.*, 1988, **51**, 459-539.
8. A. Shakouri, *Annu. Rev. Matter. Res.*, 2011, **41**, 399-431.
9. L. Leea, S. Dursuna, C. Durana and C. A. Randall, *J. Mater. Res.*, 2011, **26**, 26-30.
0. T. Tian, L. Cheng, L. Zheng, J. Xing, H. Gu, S. Bernik, H. Zeng, W. Ruan, K. Zhao and G. Li, *Acta Materialia*, 2016, **119**, 136-144.
1. V. Ponnambalam and D. T. Morelli, *J. Appl. Phys.*, 2015, **118**, 245101.
52. S. Butt, W. Xu, W. Q. He, Q. Tan, G. K. K. Ren, Y. Lin and C.-W. Nan, *J. Mater. Chem. A*, 2014, **2**, 19479-19487.
53. J. H. Grebenkemper, S. Klemenz, B. Albert, S. K. Bux and S. M. Kauzlarich, *J. Solid. State. Chem.*, 2016, **242**, 55-61.
54. K. Wilayat, A. Sikander, K. M. Benali and G.-S. Souraya, *Solid State Sci.*, 2016, **58**, 86-93.
55. B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary and C. Wolverton, *Phys. Rev. B*, 2014, **89**.
56. M. Ohtaki, H. Koga, T. Tokunaga, K. Eguchi and H. Arai, *J. Solid State Chem.*, 1995, **120**, 105-111.
57. S. He'bert, D. Flahaut, C. Martin, S. Lemonnier, J. Noudem, C. Goupil a, A. Maignan a and J. Hejtmanek, *Prog. Solid State Chem.*, 2007, **35**, 457-467.
58. K. Iwasaki, T. Ito, T. Nagasaki, Y. Arita, M. Yoshino and T. Matsui, *J. Solid State Chem.*, 2008, **181**, 3145-3150.
59. M. Yasukawaa, S. Kuniyoshia and T. Konob, *Solid State Comm.*, 2003, **126**, 213-216.
60. Mathematica's ElementData function from Wolfram Research, Inc. (2007) Available at: <http://periodictable.com/Properties/A/ElectronConfigurationString.v.html> (Accessed: 20st August 2015)).

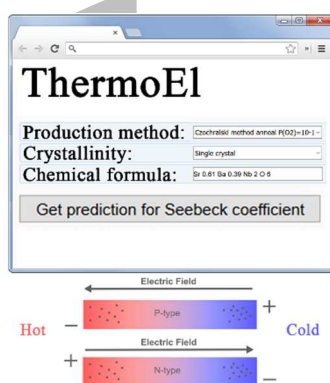
GRAPHICAL ABSTRACT

AUTHOR NAMES: Al'ona Furmanchuk, James Saal, Jeff W. Doak, Gregory B. Olson, Alok Choudhary, Ankit Agrawal

TITLE: Prediction of Seebeck coefficient for compounds without restriction to fixed stoichiometry: A machine learning approach

TEXT: The machine learning is employed for prediction of the Seebeck coefficient of crystalline materials in the temperature range from 300K to 1000K. In order to address off-stoichiometry in experimental samples, new descriptive features are introduced. The tool is implemented into a web application (<http://info.eecs.northwestern.edu/SeebeckCoefficientPredictor>) in order to assist field scientists in the discovery of novel thermoelectric materials.

GRAPHICAL ABSTRACT FIGURE



[Supporting Information (SI to accompany)]

Prediction of Seebeck coefficient for compounds without restriction to fixed stoichiometry: A machine learning approach.

Al'ona Furmanchuk^a, James Saal^b, Jeff Doak^b, Gregory B. Olson^b, Alok Choudhary^c, Ankit Agrawal^c

^aCenter for Health Information Partnerships, Northwestern University, Feinberg School of Medicine, Chicago, IL 60611, USA.

^bQuesTek Innovations LLC, Evanston, IL 60201, USA

^cDepartment of Electrical Engineering and Computer Science, Northwestern University, USA

Table of Contents

Part S1. ThermoEI web application	S1
Part S2. Extra curation step	S2
Part S3. Characterization of the compound space by means of unsupervised learning	S7
Part S4. Attributes importance list	S17
Part S5. Correlations between the top 30 most important attributes	S24

Part S1. ThermoEI web application

The optimized Random Forest algorithm-based models were implemented as separate module in our user-friendly web application ThermoEI.¹ User is asked to input expected crystallinity and production method of the material together with its chemical formula. Code does not accept information on type or concentration of charge carriers. The information about dopants must be typed explicitly into chemical formula as atom type and corresponding to it coefficient. Once request is sent to code, it checks for erroneous writes in chemical formula. If no errors were detected, application calculates 187 attributes, and predicts values for Seebeck coefficient at 300K, 400K, 700K, 1000K.

ThermoEI toolkit

Publications:
 • A. Furmanchuk, A. Agrawal, J. Saal, J. W. Doak, G. B. Olson, A. Choudhary, "Prediction of Seebeck coefficient for non-stoichiometric compounds: A machine learning approach", 2016, under review.

Developed by Alona Furmanchuk under guidance of Ankit Agrawal and Alok Choudhary
 Center for Ultrascable Computing and Information Security (CUSCIS), EECS Department, Northwestern University, Evanston, IL 60090, USA

Figure S1. Interface of ThermoEl application built based on our Random Forest-based regression models.

Accepted Article

Part S2. Extra curation step.

Table S1. Predictions done by Random Forest model vs variation in experimental data.

Chemical Formula	Seebeck coefficient actual, $\mu\text{V/K}$	Seebeck coefficient actual averaged, $\mu\text{V/K}$	Seebeck coefficient predicted, $\mu\text{V/K}$	Code in Fig. 4.	Crystallinity	Preparation_Method	ZT	Reference	Notes
1000K									
Ca _{0.9} Bi _{0.1} Mn _{0.9} Nb _{0.1} O ₃	-172.3	-92.67±0	-155.238	Ca _{0.9} Bi _{0.1} Mn _{0.9} Nb _{0.1} O ₃ /P/SSR	Polycrystalline	Solid_state_reaction	0.120992	http://dx.doi.org/10.1016/j.jallcom.2009.08.012	Actual should be -92.67
Ca _{0.9} Bi _{0.1} Mn _{0.9} Nb _{0.1} O ₃	-92.67				Polycrystalline	Solid_state_reaction	NAN	http://dx.doi.org/10.1016/j.jallcom.2009.08.012	
Ca _{0.9} Ho _{0.1} MnO ₃	-130.04	-122.31±10.932	-125.673	Ca _{0.9} Ho _{0.1} MnO ₃ /P/SSR	Polycrystalline	Solid_state_reaction	0.060407	http://dx.doi.org/10.1063/1.2362922	
Ca _{0.9} Ho _{0.1} MnO ₃	-114.58				Polycrystalline	Solid_state_reaction	0.077076	http://dx.doi.org/10.1016/0022-4596(91)90248-G	
Ca _{0.9} Tb _{0.1} MnO ₃	-130.04	-141.205±15.79	-119.439	Ca _{0.9} Tb _{0.1} MnO ₃ /P/SSR	Polycrystalline	Solid_state_reaction	0.049893	http://dx.doi.org/10.1063/1.2362922	
Ca _{0.9} Tb _{0.1} MnO ₃	-152.37				Polycrystalline	Solid_state_reaction	0.141281	http://dx.doi.org/10.1016/0022-4596(91)90248-G	
CaMnO ₃	-320.53	-322.346±59.318	-139.584	CaMnO ₃ /P/SSR	Polycrystalline	Solid_state_reaction	0.058868	http://dx.doi.org/10.1063/1.2362922	
CaMnO ₃	-282.213				Polycrystalline	Solid_state_reaction	0.045715	http://dx.doi.org/10.1109/ICT.2006.331291	
CaMnO ₃	-406.81				Polycrystalline	Solid_state_reaction	0.089159	http://dx.doi.org/10.1016/j.jallcom.2009.08.012	
CaMnO ₃	-279.83				Polycrystalline	Solid_state_reaction	NAN	http://www.jmst.org/EN/Y2009/V25/I04/0535	
Zn _{0.95} Al _{0.05} O	-164.443	-182.235±25.161	-182.096	Zn _{0.95} Al _{0.05} O/P/SSR_air	Polycrystalline	Solid_state_reaction__air	NAN	http://dx.doi.org/10.1016/j.jeurceramsoc.2006.04.012	
Zn _{0.95} Al _{0.05} O	-200.026				Polycrystalline	Solid_state_reaction__air	0.227869	http://dx.doi.org/10.1039/A602506D	
Zn _{0.98} Al _{0.02} O	-182.054	-178.527±4.988	-180.581	Zn _{0.98} Al _{0.02} O/P/SSR_air	Polycrystalline	Solid_state_reaction__air	NAN	http://dx.doi.org/10.1016/j.jeurceramsoc.2006.04.012	
Zn _{0.98} Al _{0.02} O	-175				Polycrystalline	Solid_state_reaction__air	0.212496	http://dx.doi.org/10.1039/A602506D	
ZnO	-190.963	-313.945±173.922	0.17927	ZnO/P/SSR_air	Polycrystalline	Solid_state_reaction__air	NAN	http://dx.doi.org/10.1016/j.jeurceramsoc.2006.04.012	
ZnO	-436.926				Polycrystalline	Solid_state_reaction__air	0.000749	http://dx.doi.org/10.1039/A602506D	
700K									
Ca _{0.9} Ho _{0.1} MnO ₃	-123.85	-112.66±15.825	-119.416	Ca _{0.9} Ho _{0.1} MnO ₃ /P/SSR	Polycrystalline	Solid_state_reaction	0.050077	http://dx.doi.org/10.1063/1.2362922	
Ca _{0.9} Ho _{0.1} MnO ₃	-101.47				Polycrystalline	Solid_state_reaction	0.04998	http://dx.doi.org/10.1016/0022-4596(91)90248-G	
Ca _{0.9} Tb _{0.1} MnO ₃	-123.85	-124.2±0.495	-117.691	Ca _{0.9} Tb _{0.1} MnO ₃ /P/SSR	Polycrystalline	Solid_state_reaction	0.039876	http://dx.doi.org/10.1063/1.2362922	
Ca _{0.9} Tb _{0.1} MnO ₃	-124.55				Polycrystalline	Solid_state_reaction	0.082914	http://dx.doi.org/10.1016/0022-4596(91)90248-G	
CaMnO ₃	-374.8	-421.314±61.188	-242.81	CaMnO ₃ /P/SSR	Polycrystalline	Solid_state_reaction	0.019481	http://dx.doi.org/10.1063/1.2362922	
CaMnO ₃	-362.325				Polycrystalline	Solid_state_reaction	0.041301	http://dx.doi.org/10.1109/ICT.2006.331291	
CaMnO ₃	-477.48				Polycrystalline	Solid_state_reaction	0.012122	http://dx.doi.org/10.1016/j.jallcom.2009.08.012	
CaMnO ₃	-470.65				Polycrystalline	Solid_state_reaction	NAN	http://www.jmst.org/EN/Y2009/V25/I04/0535	
CuRh _{0.9} Mg _{0.1} O ₂	203.027	173.9635±41.102	177.6391	CuRh _{0.9} Mg _{0.1} O ₂ /P/SSR_air	Polycrystalline	Solid_state_reaction__air	0.073516	http://dx.doi.org/10.1109/ICT.2006.331289	

CuRh0.9Mg0.1O2	144.9				Polycrystalline	Solid_state_reaction__air	NAN	http://dx.doi.org/10.1103/PhysRevB.80.115103
In0.2Co4Sb12	-281.75	-264.645±24.19	-281.258	In0.2Co4Sb12/P/SSR	Polycrystalline	Solid_state_reaction	1.334134	http://dx.doi.org/10.1021/cm052055b
In0.2Co4Sb12	-247.54				Polycrystalline	Solid_state_reaction	1.385962	http://dx.doi.org/10.1557/jmr.2011.163
Zn0.95Al0.05O	-112.738	-154.192±58.625	-139.733	Zn0.95Al0.05O/P/SSR_air	Polycrystalline	Solid_state_reaction__air	NAN	http://dx.doi.org/10.1016/j.jeurceramsoc.2006.04.012
Zn0.95Al0.05O	-195.646				Polycrystalline	Solid_state_reaction__air	0.089682	http://dx.doi.org/10.1039/A602506D
Zn0.98Al0.02O	-133.608	-152.544±26.78	-159.853	Zn0.98Al0.02O/P/SSR_air	Polycrystalline	Solid_state_reaction__air	NAN	http://dx.doi.org/10.1016/j.jeurceramsoc.2006.04.012
Zn0.98Al0.02O	-171.48				Polycrystalline	Solid_state_reaction__air	0.112229	http://dx.doi.org/10.1039/A602506D
Zn4Sb3	184.769	188.3173±3.521	167.4617	Zn4Sb3/P/SSR_Ar	Polycrystalline	Solid_state_reaction__Ar	0.962808	http://dx.doi.org/10.1039/c0jm02011g
Zn4Sb3	188.372				Polycrystalline	Solid_state_reaction__Ar	1.055897	http://dx.doi.org/10.1039/c0jm02011g
Zn4Sb3	191.811				Polycrystalline	Solid_state_reaction__Ar	1.044056	http://dx.doi.org/10.1039/c0jm02011g
ZnO	-159.904	-244.887±120.183	-5.29792	ZnO/P/SSR_air	Polycrystalline	Solid_state_reaction__air	NAN	http://dx.doi.org/10.1016/j.jeurceramsoc.2006.04.012
ZnO	-329.869				Polycrystalline	Solid_state_reaction__air	0.000291	http://dx.doi.org/10.1039/A602506D
Zr0.15Hf0.15Ti0.7NiSn	-260.541	-260.736±0.276	-229.419	Zr0.15Hf0.15Ti0.7NiSn/P/AM_Ar	Polycrystalline	Arc-melted__Ar	1.016372	http://dx.doi.org/10.1063/1.1868063
Zr0.15Hf0.15Ti0.7NiSn	-260.931				Polycrystalline	Arc-melted__Ar	NAN	http://dx.doi.org/10.1016/j.jallcom.2004.05.078
Zr0.25Hf0.25Ti0.5NiSn	-328.251	-331.075±3.993	-254.445	Zr0.25Hf0.25Ti0.5NiSn/P/AM_Ar	Polycrystalline	Arc-melted__Ar	1.419171	http://dx.doi.org/10.1063/1.1868063
Zr0.25Hf0.25Ti0.5NiSn	-333.898				Polycrystalline	Arc-melted__Ar	NAN	http://dx.doi.org/10.1016/j.jallcom.2004.05.078
Zr0.35Hf0.35Ti0.3NiSn	-348.046	-346.632±2.000	-252.051	Zr0.35Hf0.35Ti0.3NiSn/P/AM_Ar	Polycrystalline	Arc-melted__Ar	1.263615	http://dx.doi.org/10.1063/1.1868063
Zr0.35Hf0.35Ti0.3NiSn	-345.217				Polycrystalline	Arc-melted__Ar	NAN	http://dx.doi.org/10.1016/j.jallcom.2004.05.078
Zr0.4Hf0.4Ti0.2NiSn	-276.689	-281.97±7.468	-245.079	Zr0.4Hf0.4Ti0.2NiSn/P/AM_Ar	Polycrystalline	Arc-melted__Ar	0.949138	http://dx.doi.org/10.1063/1.1868063
Zr0.4Hf0.4Ti0.2NiSn	-287.251				Polycrystalline	Arc-melted__Ar	NAN	http://dx.doi.org/10.1016/j.jallcom.2004.05.078
Zr0.5Hf0.5NiSn	-236.062	-236.987±1.308	-214.064	Zr0.5Hf0.5NiSn/P/AM_Ar	Polycrystalline	Arc-melted__Ar	0.549896	http://dx.doi.org/10.1063/1.1868063
Zr0.5Hf0.5NiSn	-237.912				Polycrystalline	Arc-melted__Ar	0.539	http://dx.doi.org/10.1016/j.jallcom.2004.05.078
400K								
Ca0.9Ho0.1MnO3	-90.28	-86.035±6.003	-82.6807	Ca0.9Ho0.1MnO3/P/SSR	Polycrystalline	Solid_state_reaction	0.022586	http://dx.doi.org/10.1063/1.2362922
Ca0.9Ho0.1MnO3	-81.79				Polycrystalline	Solid_state_reaction	0.0234	http://dx.doi.org/10.1016/0022-4596(91)90248-G
Ca0.9Tb0.1MnO3	-90.28	-92.05±2.503	-86.1641	Ca0.9Tb0.1MnO3/P/SSR	Polycrystalline	Solid_state_reaction	0.0191	http://dx.doi.org/10.1063/1.2362922
Ca0.9Tb0.1MnO3	-93.82				Polycrystalline	Solid_state_reaction	0.035074	http://dx.doi.org/10.1016/0022-4596(91)90248-G
CaMnO3	-600.16	-549.859±61.843	-324.068	CaMnO3/P/SSR	Polycrystalline	Solid_state_reaction	0.010514	http://dx.doi.org/10.1063/1.2362922
CaMnO3	-462.97				Polycrystalline	Solid_state_reaction	NAN	http://www.jmst.org/EN/Y2009/V25/I04/0535
CaMnO3	-587.185				Polycrystalline	Solid_state_reaction	0.018604	http://dx.doi.org/10.1109/ICT.2006.331291
CaMnO3	-549.12				Polycrystalline	Solid_state_reaction	0.001247	http://dx.doi.org/10.1016/j.jallcom.2009.08.012
CuRh0.9Mg0.1O2	153.007	134.0035±26.875	149.8803	CuRh0.9Mg0.1O2/P/SSR_air	Polycrystalline	Solid_state_reaction__air	0.035498	http://dx.doi.org/10.1109/ICT.2006.331289

CuRh0.9Mg0.1O2	115				Polycrystalline	Solid_state_reaction__air	NAN	http://dx.doi.org/10.1103/PhysRevB.80.115103
In0.2Co4Sb12	-248.19	-231.29±23.9	-260.325	In0.2Co4Sb12/P/SSR	Polycrystalline	Solid_state_reaction	0.400778	http://dx.doi.org/10.1021/cm052055b
In0.2Co4Sb12	-214.39				Polycrystalline	Solid_state_reaction	0.380118	http://dx.doi.org/10.1557/jmr.2011.163
Zn4Sb3	148.361	154.8113±6.318	69.68974	Zn4Sb3/P/SSR_Ar	Polycrystalline	Solid_state_reaction__Ar	0.445642	http://dx.doi.org/10.1039/c0jm02011g
Zn4Sb3	155.085				Polycrystalline	Solid_state_reaction__Ar	0.5249	http://dx.doi.org/10.1039/c0jm02011g
Zn4Sb3	160.988				Polycrystalline	Solid_state_reaction__Ar	0.550828	http://dx.doi.org/10.1039/c0jm02011g
300K								
Bi2Te3	162	-6±237.588	40.44715	Bi2Te3/S/M	Single crystal	Melted	0.975	http://dx.doi.org/10.1201/9781420049718.ch19
Bi2Te3	-174				Single crystal	Melted	0.5025	http://dx.doi.org/10.1201/9781420049718.ch19
CaMnO3	-462.97	-561.973±94.374	-115.097	CaMnO3/P/SSR	Polycrystalline	Solid_state_reaction	NAN	http://www.jmst.org/EN/Y2009/V25/I04/0535
CaMnO3	-650.91				Polycrystalline	Solid_state_reaction	0.011795	http://dx.doi.org/10.1109/ICT.2006.331291
CaMnO3	-572.04				Polycrystalline	Solid_state_reaction	NAN	http://dx.doi.org/10.1016/j.jallcom.2009.08.012
CaMnO3	-357.564	-417.124±84.230	-114.891	CaMnO3/P/SSR_air	Polycrystalline	Solid_state_reaction__air	0.008282	http://dx.doi.org/10.1103/PhysRevB.60.14057
CaMnO3	-476.683				Polycrystalline	Solid_state_reaction__air	0.000327	http://dx.doi.org/10.1006/jssc.1995.1384
CuRh0.9Mg0.1O2	138.814	119.257±27.658	137.1018	CuRh0.9Mg0.1O2/P/SSR_air	Polycrystalline	Solid_state_reaction__air	NAN	http://dx.doi.org/10.1109/ICT.2006.331289
CuRh0.9Mg0.1O2	99.7				Polycrystalline	Solid_state_reaction__air	NAN	http://dx.doi.org/10.1103/PhysRevB.80.115103
In0.2Co4Sb12	-222.15	-207.18±21.171	-229.514	In0.2Co4Sb12/P/SSR	Polycrystalline	Solid_state_reaction	0.400778	http://dx.doi.org/10.1021/cm052055b
In0.2Co4Sb12	-192.21				Polycrystalline	Solid_state_reaction	0.380118	http://dx.doi.org/10.1557/jmr.2011.163
Sb2Te3	38	50.5±17.678	21.29163	Sb2Te3/S/M	Single crystal	Melted	0.0675	http://dx.doi.org/10.1201/9781420049718.ch19
Sb2Te3	63				Single crystal	Melted	0.1875	http://dx.doi.org/10.1201/9781420049718.ch19
Sr0.61Ba0.39Nb2O6	-92.4	-139.2±66.185	-156.048	Sr0.61Ba0.39Nb2O6/S/CM	Single crystal	Czocharalski_method__anneal_P(O2)=10 ¹⁴	0.05437	http://dx.doi.org/10.1063/1.3291563
Sr0.61Ba0.39Nb2O6	-186				Single crystal	Czocharalski_method__anneal_P(O2)=10 ¹⁴	0.005928	http://dx.doi.org/10.1063/1.3291563
Zn4Sb3	122.133	128.306±6.991	79.75912	Zn4Sb3/P/SSR_Ar	Polycrystalline	Solid_state_reaction__Ar	0.251704	http://dx.doi.org/10.1039/c0jm02011g
Zn4Sb3	126.887				Polycrystalline	Solid_state_reaction__Ar	0.298945	http://dx.doi.org/10.1039/c0jm02011g
Zn4Sb3	135.898				Polycrystalline	Solid_state_reaction__Ar	0.32555	http://dx.doi.org/10.1039/c0jm02011g
Zr0.15Hf0.15Ti0.7NiSn	-252.204	-251.458±1.056	-197.057	Zr0.15Hf0.15Ti0.7NiSn/P/AM_Ar	Polycrystalline	Arc-melted__Ar	0.116776	http://dx.doi.org/10.1063/1.1868063
Zr0.15Hf0.15Ti0.7NiSn	-250.711				Polycrystalline	Arc-melted__Ar	NAN	http://dx.doi.org/10.1016/j.jallcom.2004.05.078
Zr0.25Hf0.25Ti0.5NiSn	-311.579	-307.573±5.665	-224.725	Zr0.25Hf0.25Ti0.5NiSn/P/AM_Ar	Polycrystalline	Arc-melted__Ar	0.19989	http://dx.doi.org/10.1063/1.1868063
Zr0.25Hf0.25Ti0.5NiSn	-303.567				Polycrystalline	Arc-melted__Ar	NAN	http://dx.doi.org/10.1016/j.jallcom.2004.05.078
Zr0.35Hf0.35Ti0.3NiSn	-369.912	-349.818±28.417	-227.879	Zr0.35Hf0.35Ti0.3NiSn/P/AM_Ar	Polycrystalline	Arc-melted__Ar	0.150489	http://dx.doi.org/10.1063/1.1868063
Zr0.35Hf0.35Ti0.3NiSn	-329.724				Polycrystalline	Arc-melted__Ar	NAN	http://dx.doi.org/10.1016/j.jallcom.2004.05.078
Zr0.4Hf0.4Ti0.2NiSn	-188.659	-185.208±4.881	-215.106	Zr0.4Hf0.4Ti0.2NiSn/P/AM_Ar	Polycrystalline	Arc-melted__Ar	0.030958	http://dx.doi.org/10.1063/1.1868063

Zr _{0.4} Hf _{0.4} Ti _{0.2} NiSn	-181.756				Polycrystalline	Arc-melted__Ar	NAN	http://dx.doi.org/10.1016/j.jallcom.2004.05.078
Zr _{0.5} Hf _{0.5} NiSn	-178.765	-178.009±1.070	-197.397	Zr _{0.5} Hf _{0.5} NiSn/P/AM_Ar	Polycrystalline	Arc-melted__Ar	0.024064	http://dx.doi.org/10.1063/1.1868063
Zr _{0.5} Hf _{0.5} NiSn	-177.252				Polycrystalline	Arc-melted__Ar	0.0256	http://dx.doi.org/10.1016/j.jallcom.2004.05.078

Table S2. Analysis of outliers in Random Forest models based on 452 attributes.

Actual Seebeck, $\mu\text{V/K}$	Predicted Seebeck, $\mu\text{V/K}$	Chemical formula	Crystallinity	Preparation Method	URL reference	Note
Random Forest @1000K						
289	-168.378	Ba ₈ Ga ₁₆ Ge ₃₀	Polycrystalline	Arc_melting	http://dx.doi.org/10.1109/ICT.2002.1190269	Sign in actual should be negative; refer to Fig. 4
-266	100.7905	Ba ₈ Ga ₁₈ Ge ₂₈	Polycrystalline	Arc_melting	http://dx.doi.org/10.1109/ICT.2002.1190269	Sign in actual should be positive; refer to Fig. 4
224	-38.151	Mg ₂ Si _{0.6} Ge _{0.4} Ag _{0.02}	Polycrystalline	Solid_state_reaction_He/H ₂	http://dx.doi.org/10.1007/s11664-009-0735-1	
-212.3	75.306	Mg ₂ Si _{0.98} Bi _{0.02}	Polycrystalline	Solid_state_reaction_He/H ₂	http://dx.doi.org/10.1007/s11664-009-0735-1	Actual should be -224
642.62	-39.1193	NiO	Polycrystalline	Solid_state_reaction	http://dx.doi.org/10.1143/JJAP.38.L1336	Measurement is shown for compound doped with Li. No NiO is reported in the article.
248.114	-211.694	Si _{0.79936} Ge _{0.1998480.0008}	Polycrystalline	Vacuum_hot_pressed	http://dx.doi.org/10.1063/1.348408	Not clear source of coefficients in equation. P- and n-type Si _{0.8} Ge _{0.2} are compounds studied. No discussion of exact coefficients. Might be wrong reference.
Random Forest @700K						
288	-138.98	Ba ₈ Ga ₁₆ Ge ₃₀	Polycrystalline	Arc_melting	http://dx.doi.org/10.1109/ICT.2002.1190269	Sign in actual should be negative
-219	76.274	Ba ₈ Ga ₁₈ Ge ₂₈	Polycrystalline	Arc_melting	http://dx.doi.org/10.1109/ICT.2002.1190269	Sign in actual should be positive
390	120.1282	Ca ₅ Al ₂ Sb ₆	Polycrystalline	Solid_state_reaction__Ar	http://dx.doi.org/10.1002/adfm.201000970	
-332.067	-68.8963	CaMnO ₃	Polycrystalline	Solid_state_reaction__air	http://dx.doi.org/10.1006/jssc.1995.1384	
318.8	26.32075	Mg ₂ Si _{0.6} Ge _{0.4} Ag _{0.02}	Polycrystalline	Solid_state_reaction_He/H ₂	http://dx.doi.org/10.1007/s11664-009-0735-1	
483.47	-4.41576	NiO	Polycrystalline	Solid_state_reaction	http://dx.doi.org/10.1143/JJAP.38.L1336	Measurement is shown for compound doped with Li. No pure NiO is reported in article.
-230.877	122.6241	Si _{0.7956} Ge _{0.19890.0055}	Polycrystalline	Vacuum_hot_pressed	http://dx.doi.org/10.1063/1.348408	Not clear source of coefficients in equation. P- and n-type Si _{0.8} Ge _{0.2} is the compound studied.
209.823	-103.317	Si _{0.79936} Ge _{0.1998480.0008}	Polycrystalline	Vacuum_hot_pressed	http://dx.doi.org/10.1063/1.348408	No discussion of exact coefficients. Might be wrong reference.
-358.019	-90.8259	Zr _{0.3} Hf _{0.3} Ti _{0.4} NiSn	Polycrystalline	Arc-melted__Ar	http://dx.doi.org/10.1109/ICT.2002.1190269	
Random Forest @400K						
185	-77.86	Ba ₈ Ga ₁₆ Ge ₃₀	Polycrystalline	Arc_melting	http://dx.doi.org/10.1109/ICT.2002.1190269	Sign in actual should be negative, refer to Fig. 4
327	63.46365	Ag ₉ TiTe ₅	Polycrystalline	Melted	http://dx.doi.org/10.1063/1.2009828	
463	128.2506	Ca ₅ Al ₂ Sb ₆	Polycrystalline	Solid_state_reaction__Ar	http://dx.doi.org/10.1002/adfm.201000970	
-460.992	-69.0541	CaMnO ₃	Polycrystalline	Solid_state_reaction__air	http://dx.doi.org/10.1006/jssc.1995.1384	

337.9	23.32705	Tl9BiTe6	Polycrystalline	Melted__zone_refined	http://dx.doi.org/10.1103/PhysRevLett.86.4350
<hr/>					
470	115.6886	Ca5Al2Sb6	Polycrystalline	Random Forest @300K Solid_state_reaction__Ar	http://dx.doi.org/10.1002/adfm.201000970
-390	45.43803	LaCoO3	Polycrystalline	Solid_state_reaction	http://dx.doi.org/10.1016/j.jssc.2008.08.078
-343.22	-27.9614	Nd2Cu0.98Zn0.02O4	Polycrystalline	Solid_state_reaction__air	http://dx.doi.org/10.1002/chin.200318015
<hr/>					

Accepted Article

S3. Characterization of the compound space by means of unsupervised learning

The cleaned UCSB database consists of representatives from different families in the materials space. Their differences in structural features, carrier's concentration, and conduction mechanisms will be manifested in the Seebeck coefficient of materials. One might want to start from grouping materials into families. The goal is to find similarities between attributes in different materials that would result in assignment of materials to families. In other words, we are looking for numerosity reduction that would help one to build regression models specific to family. In our data set each material comes with 452 attributes. We used all 452 attribute to calculate similarity between materials for further assignment of materials to groups. Latter we employed the hierarchical dendrogram² for visualization and comprehension (Fig. S1).

Along the *x-axis* of Fig. S1, each discrete point represents a compound. From each compound leads a vertical line. At some height, these lines are all connected. The height at which they connect is determined by the similarity between the compounds. If some compounds are already connected, then the height represents the similarity between groups of compounds. The way in which this similarity is defined is called the linkage method.² One of the ways to decide the similarity between a group of compounds and another compound (or between two groups of compounds) is to compare the average set of attributes of all the members of the group with the attributes of the other compound. This is called the group average linkage method,² and it was the method used to produce the dendrogram in Fig. S1.

The process of grouping together the most similar compounds is then repeated, each time linking less similar compounds at higher levels on the dendrogram and each time reducing the number of groups by one. If one were to stop making groups at the threshold similarity level, then one would be left with several groups/families. If the data fall into well-separated families, then there should be some level in the dendrogram with a large step in the linkage height. At some point (cut level shown by red dashed line in Fig. S1), there will be a link whose height is clearly larger than all of the links below it. This is where one shall put cut level. In our case it resulted in six clusters we highlighted in red (61.1% of data set, Table S1), yellow (29.7% of data set, S1, Table S2), green (2.7% of data set, Table S3), blue (4.4% of data set, Table S4), cyan (0.3% of data set, Table S5), and magenta (1.7% of data set, Table S6).

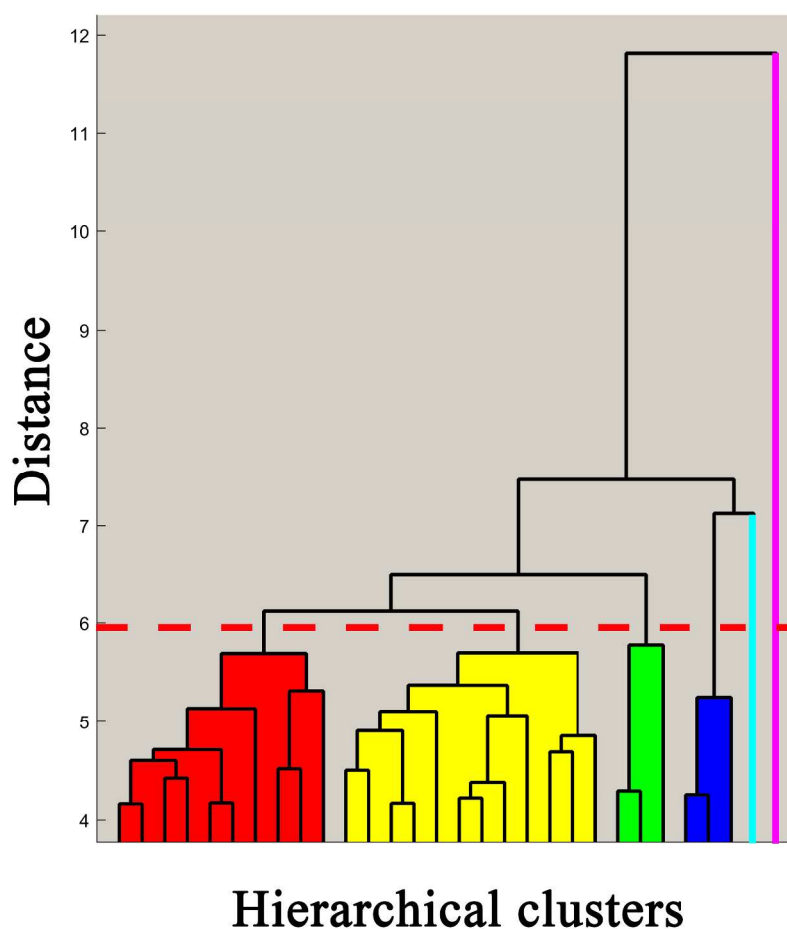


Figure S2. Clustering results employing hierarchical tree.

The reason we discuss dendrogram method in detail is to show that for analysis of multi-dimensional space of attributes, machine learning is probably more efficient approach compared to human trying to comprehend a set of approximately 300 compounds with 452 attributes each. We tried to make sense of machine-made family assignment. Therefore, we searched through publications in UCSB database and collected possibly useful bits of knowledge that would help us in comprehension of commonalities. At least those were our initial expectations.

The red cluster (181 compounds in Fig. S1) was assigned predominantly polycrystalline doped oxide thermoelectric materials, as well as Chevrel-phase compounds (molybdenum chalcogenides $MxMo_6X_8$ (M: Cu, Ag, Ni, Fe, rare earth elements, etc., X: S, Se, Te)). Several data points in the red cluster correspond to perovskite-type (Sr, La)TiO₃ oxides doped with rare earth elements. The large Seebeck coefficient in those is caused by the degenerated Ti-3d orbital in the conduction band. However, the high thermal conductivity decreases the figure of merit. It was observed that lattice scattering is predominant for electrons, while thermal conductivities decrease with rare earth doping according to the ionic rare earth radius. In the same cluster one could see delafossite-type oxides doped with transition metals. The dominant charge carriers of $CuCr_{1-x}Mg_xO_2$ are positive holes. Another representative is pyrochlore ruthenates $Ln_{2-x}BixRu_2O_7$ (Ln=Nd and Yb, x=1 and 0.6) susceptible to geometrical frustration. Their

electrical resistivity depends on the x values and the doped ionic sizes, while the Seebeck coefficient does not.³

There are also layered structures $\text{La}_{1.85}\text{Sr}_{0.15}\text{CuO}_{4+\delta}$ (La-214) as well as more complex $\text{La}_{2-x}\text{Sr}_x\text{Cu}_{0.94}\text{M}_{0.06}\text{O}_4$ ($M=\text{Ti}$, Mn and Ru) systems. Systematic investigation of the effects of high-valence substitution on the Cu-spin dynamics and superconductivity from the resistivity and magnetization measurements suggests that the superconductivity is attributed to compensation effect originating from the local electrons introduced by the Ti^{4+} doping compensated by the holes introduced by Sr substitution for La.

The last member of the red cluster is formed by doped polycrystalline ceramic $(\text{Ca}_{0.9}\text{M}_{0.1})\text{MnO}_3$ ($M = \text{Y}$, La , Ce , Sm , In , Sn , Sb , Pb , Bi) compounds. It was shown that substitution at the Ca site causes an increase in the electrical conductivity along with a moderate decrease in the absolute value of the Seebeck coefficient. There is indication that hopping conduction occurs in these oxides. The electrical conductivity values at room temperature increase with increasing ionic radii of the cation substituents, implying an increase in the carrier mobility due to larger inter-site distance for hopping.⁴

The yellow cluster consists of 88 structures shared by doped Zintl compounds (including binary and filled skutterudites), and half-Heusler compounds. Zintl compounds are anionic clusters that are generated by reduction of heavy main group p elements, mostly metals or semimetals, with alkali, alkaline-earth or rare earth metals. The structural complexity of Zintl compounds is implicated in their unusually low thermal conductivity values, which approaches the minimum thermal conductivity limit at high temperatures. Carrier concentration control by means of doping of Zintl compounds converts them into good candidates for thermoelectric power generation. Doped Zintl structures exhibit good thermoelectric efficiency, arising from either high electronic mobility ($\text{CeCoFe}_3\text{Sb}_{12}$, YbZn_2Sb_2) or low lattice thermal conductivity ($\text{Yb}_{14}\text{AlSb}_{11}$, and $\text{Ba}_8\text{Ga}_{16}\text{Ge}_{30}$). While Zintl compounds form covalent networks, half-Heusler compounds have more ionic character of their bonds. Ordinarily, YXZ , a half-Heusler alloy has C1_b structure, which can be derived from the L2_1 structure of a full-Heusler alloy (Y_2XZ). Y and X are transition-metal elements, while Z is a main group element.

The green cluster is represented by 8 compounds: polycrystalline $\text{PbTe}_{1-x}\text{Se}_x$ structures doped with Na, and Bi_2Q_3 ($\text{Q} = \text{S}$, Se , Te) compounds doped with Sb and Cs. In general, those are suspected to form some segregation to native defects, nanometer-scale particles, voids or other interfaces affecting thermoelectric properties. Specific to Bi/Sb chemistry we can highlight the stereochemical localization of ns^2 lone-pair electrons, and its influence on the structure type and the electronic properties of compounds. The alkali or alkaline earth metals introduced into the Bi_2Q_3 lattices rearrange the octahedrally coordinated Bi/Sb elements often causing the group 15 element to exhibit varying degrees of ns^2 lone-pair stereochemical activity. The application of dopants manipulates the electron density at the Fermi level and controls the conductivity type in these materials.

The blue cluster is composed of 13 structures represented by Si- and Ge-based type-I clathrates. One of the prominent features of these materials lies in their crystal structures composed of polyhedral cages, formed by covalently bonded host framework, in which guest atoms can be entrapped. The impact of a guest atom (alkali, alkaline-earth metals and europium) on the thermal transport might be related to an enhanced phonon-charge carrier coupling. It is known that an addition of Ga, Al, Au, and Ag to clathrates opens up a gap near the Fermi level, potentially enabling to reach higher ZT values.^{5, 6}

The cyan cluster consists only of $\text{TI}_{11.5}\text{Sb}_{11.5}\text{Cu}_8\text{Se}_{27}$ compound. It is known that differently from green cluster (alkali metal chalcogenides), higher electronegativity of TI is expected to result in less ionic character, smaller band gap, and higher carrier mobility of quaternary chalcogenide structures. The thermal conductivity of the TI compounds will be lower than that of the alkali metal compounds simply due to the larger mass of TI.⁷

The magenta cluster consists of five $\text{Bi}_{100-x}\text{Sb}_x$ ($x=8-17$) high-homogeneity alloys. Their thermoelectric properties change gradually with the Sb concentration x , which is attributed to the variation of the energy gap. The charge carrier mobility was enhanced by annealing, which leads to a small electrical resistivity and a large Seebeck coefficient.

After analysis of literature we ended up with very diverse family representatives as well as non-uniform collection of features researchers chose to report. It might be that scientific community did not accumulate enough of structured knowledge to come up with analogues to clustering assignment to families. From the statistical point of view, small population of some clusters does not allow one to create models specific to each family.

Table S3. List of compounds in red cluster.

Chemical Formula	Crystallinity	Production method
AgCrSe2	Polycrystalline	Solid_state_reaction__sealed
Ba0.3Sr0.6La0.1TiO3	Polycrystalline	Solid_state_reaction__Ar
Ba0.4Sr0.6PbO3	Polycrystalline	Solid_state_reaction_(under_oxygen)
Ba0.6Sr0.4PbO3	Polycrystalline	Solid_state_reaction_(under_oxygen)
Ba0.8Sr0.2PbO3	Polycrystalline	Solid_state_reaction_(under_oxygen)
BaPbO3	Polycrystalline	Solid_state_reaction_(under_oxygen)
Bi2Ru2O7	Polycrystalline	Solid_state_reaction
Ca0.7Ho0.3MnO3	Polycrystalline	Solid_state_reaction
Ca0.7Tb0.3MnO3	Polycrystalline	Solid_state_reaction
Ca0.7Y0.3MnO3	Polycrystalline	Solid_state_reaction
Ca0.92La0.08MnO3	Polycrystalline	Solid_state_reaction
Ca0.94La0.06MnO3	Polycrystalline	Solid_state_reaction
Ca0.96Bi0.04Mn0.96Nb0.04O3	Polycrystalline	Solid_state_reaction
Ca0.96La0.04MnO3	Polycrystalline	Solid_state_reaction
Ca0.96Sm0.04MnO3	Polycrystalline	Co-precipitation
Ca0.98Bi0.02Mn0.98Nb0.02O3	Polycrystalline	Solid_state_reaction
Ca0.98La0.02MnO3	Polycrystalline	Solid_state_reaction
Ca0.9Bi0.1Mn0.9Nb0.1O3	Polycrystalline	Solid_state_reaction
Ca0.9Bi0.1MnO3	Polycrystalline	Solid_state_reaction__air
Ca0.9Ce0.1MnO3	Polycrystalline	Solid_state_reaction__air
Ca0.9Ho0.1MnO3	Polycrystalline	Solid_state_reaction
Ca0.9In0.1MnO3	Polycrystalline	Solid_state_reaction__air
Ca0.9La0.1MnO3	Polycrystalline	Solid_state_reaction__air
Ca0.9Nd0.1MnO3	Polycrystalline	Solid_state_reaction
Ca0.9Pb0.1MnO3	Polycrystalline	Solid_state_reaction__air
Ca0.9Sb0.1MnO3	Polycrystalline	Solid_state_reaction__air
Ca0.9Sm0.1MnO3	Polycrystalline	Solid_state_reaction__air
Ca0.9Sn0.1MnO3	Polycrystalline	Solid_state_reaction__air
Ca0.9Tb0.1MnO3	Polycrystalline	Solid_state_reaction
Ca0.9Y0.1MnO3	Polycrystalline	Solid_state_reaction

Accepted Article

Ca0.9Y0.1MnO3	Polycrystalline	Solid_state_reaction__air
Ca0.9Yb0.1MnO3	Polycrystalline	Solid_state_reaction
Ca2.4Na0.3Bi0.3Co4O9	Polycrystalline	Solid_state_reaction__air
Ca2.7Bi0.3Co4O9	Polycrystalline	Solid_state_reaction__air
Ca2.7Na0.3Co4O9	Polycrystalline	Solid_state_reaction__air
Ca2Co2O5	Single crystal	Melted
Ca3Co4O9	Polycrystalline	Sol-gel__air
Ca3Co4O9	Polycrystalline	Solid_state_reaction__air
Ca3Co4O9	Single crystal	Flux_(SrCl2)__air
CaGd0.94Mn0.06O3	Polycrystalline	Co-precipitation
CaGd0.96Mn0.04O3	Polycrystalline	Co-precipitation
CaGd0.98Mn0.02O3	Polycrystalline	Co-precipitation
CaMn0.82Ru0.18O3	Polycrystalline	Solid_state_reaction
CaMn0.94Ru0.06O3	Polycrystalline	Solid_state_reaction
CaMn0.96Ru0.04O3	Polycrystalline	Solid_state_reaction
CaMn0.98Nb0.02O3	Polycrystalline	Ultrasonic_spray_pyrolysis
CaMn0.98Ru0.02O3	Polycrystalline	Solid_state_reaction
CaMn0.9Ru0.1O3	Polycrystalline	Solid_state_reaction
CaMn6.5Cu0.5O12	Polycrystalline	Solid_state_reaction__air
CaMn6CuO12	Polycrystalline	Solid_state_reaction__air
CaMn7O12	Polycrystalline	Solid_state_reaction__air
CaMnO3	Polycrystalline	Solid_state_reaction__air
CaYb0.05Mn0.95O3	Polycrystalline	Solid_state_reaction
CaYb0.15Mn0.85O3	Polycrystalline	Solid_state_reaction
CaYb0.1Mn0.9O3	Polycrystalline	Solid_state_reaction
CaYb0.4Mn0.6O3	Polycrystalline	Solid_state_reaction
Cr1.3Mo6S8	Polycrystalline	Solid_state_reaction__vacuum
Cu1.98Se	Polycrystalline	Melted__vacuum
Cu2Se	Polycrystalline	Melted__vacuum
Cu4.0Mo6S8	Polycrystalline	Solid_state_reaction__vacuum
CuCr0.95Mg0.05O2	Polycrystalline	Solid_state_reaction__air
CuCr0.96Mg0.04O2	Polycrystalline	Solid_state_reaction__air
CuCr0.97Mg0.03O2	Polycrystalline	Solid_state_reaction__air
CuCr0.98Mg0.02O2	Polycrystalline	Solid_state_reaction__air
CuCr0.98Mg0.02O2	Polycrystalline	Solid_state_reaction
CuCr0.99Mg0.01O2	Polycrystalline	Solid_state_reaction__air
CuCrO2	Polycrystalline	Solid_state_reaction__air
CuFe0.9Cr0.1O2	Polycrystalline	Solid_state_reaction
CuRh0.6Mg0.4O2	Polycrystalline	Solid_state_reaction__air
CuRh0.96Mg0.04O2	Polycrystalline	Solid_state_reaction__air
CuRh0.99Mg0.01O2	Polycrystalline	Solid_state_reaction__air
CuRh0.9Mg0.1O2	Polycrystalline	Solid_state_reaction__air
CuRh0.9Mg0.1O2	Polycrystalline	Solid_state_reaction
CuRhO2	Polycrystalline	Solid_state_reaction__air
Fe0.978Co0.00196Si1.96Y0.04O0.06	Polycrystalline	Arc-melted__Ar
Fe0.978Co0.00196Si1.96Y0.12O0.18	Polycrystalline	Arc-melted__Ar
Fe1.3Mo6S8	Polycrystalline	Solid_state_reaction__vacuum

Accepted Article

Fe _{1.94} Ti _{0.06} O ₃	Polycrystalline	Solid_state_reaction__air
Fe _{1.96} Sn _{0.04} O ₃	Polycrystalline	Solid_state_reaction__air
Fe _{1.96} Ti _{0.04} O ₃	Polycrystalline	Solid_state_reaction__air
Fe _{1.98} Sn _{0.02} O ₃	Polycrystalline	Solid_state_reaction__air
Fe _{1.98} Ti _{0.02} O ₃	Polycrystalline	Solid_state_reaction__air
In _{1.8} Ge _{0.2} O ₃	Polycrystalline	Solid_state_reaction
In _{1.94} Ge _{0.06} O ₃	Polycrystalline	Solid_state_reaction
In _{1.985} Ge _{0.015} O ₃	Polycrystalline	Solid_state_reaction
In _{1.994} Ge _{0.006} O ₃	Polycrystalline	Solid_state_reaction
In _{1.998} Ge _{0.002} O ₃	Polycrystalline	Solid_state_reaction
In _{1.9} Ge _{0.1} O ₃	Polycrystalline	Solid_state_reaction
In ₂ O ₃	Polycrystalline	Solid_state_reaction
La _{0.05} Ca _{2.85} Co _{3.8} O _{8.55}	Polycrystalline	Sol-gel__air
La _{0.3} Ca _{2.7} Co ₄ O ₉	Polycrystalline	Sol-gel__air
La _{0.45} Ca _{2.55} Co ₄ O ₉	Polycrystalline	Sol-gel__air
La _{0.85} Sr _{0.2} Co ₃	Polycrystalline	Solid_state_reaction
La _{0.95} Sr _{0.05} Co ₃	Polycrystalline	Solid_state_reaction
La _{0.98} Sr _{0.02} Co ₃	Polycrystalline	Solid_state_reaction
La _{0.99} Sr _{0.01} Co ₃	Polycrystalline	Solid_state_reaction
La _{0.9} Bi _{0.1} NiO ₃	Polycrystalline	Evaporate_nitrates_(1173_K__air)
La _{1.61} Sr _{0.39} Cu _{0.94} Ti _{0.06} O ₄	Polycrystalline	Solid_state_reaction
La _{1.67} Sr _{0.34} Cu _{0.94} Ti _{0.06} O ₄	Polycrystalline	Solid_state_reaction
La _{1.69} Sr _{0.31} Cu _{0.94} Ti _{0.06} O ₄	Polycrystalline	Solid_state_reaction
La _{1.725} Sr _{0.28} Cu _{0.4}	Polycrystalline	Solid_state_reaction_(O2_atmosphere)
La _{1.73} Sr _{0.27} Cu _{0.94} Ti _{0.06} O ₄	Polycrystalline	Solid_state_reaction
La _{1.85} Sr _{0.15} Cu _{0.94} Ti _{0.06} O ₄	Polycrystalline	Solid_state_reaction
La _{1.85} Sr _{0.15} Cu _{0.4}	Polycrystalline	Solid_state_reaction_(O2_atmosphere)
La _{1.95} Sr _{0.05} Cu _{0.4}	Polycrystalline	Solid_state_reaction_(O2_atmosphere)
La _{1.95} Sr _{0.1} Cu _{0.4}	Polycrystalline	Solid_state_reaction_(O2_atmosphere)
La ₂ CuO ₄	Polycrystalline	Solid_state_reaction
LaCoO ₃	Polycrystalline	Solid_state_reaction
LaNiO ₃	Polycrystalline	Evaporate_nitrates_(1173_K__air)
Li _{0.0024} Ni _{0.9976} O	Polycrystalline	Solid_state_reaction
Li _{0.0066} Ni _{0.9944} O	Polycrystalline	Solid_state_reaction
Li _{0.0184} Ni _{0.9816} O	Polycrystalline	Solid_state_reaction
Li _{0.0242} Ni _{0.9758} O	Polycrystalline	Solid_state_reaction
LiMn ₂ O ₄	Polycrystalline	Solid_state_reaction__air
LiMn ₂ O ₄	Polycrystalline	Solid_state_reaction_(oxalates)
Mn _{1.3} Mo ₆ S ₈	Polycrystalline	Solid_state_reaction__vacuum
Mo ₃ Te ₄	Polycrystalline	Solid_state_reaction__vacuum
Mo ₆ Te ₆ S ₂	Polycrystalline	Solid_state_reaction__vacuum
Mo ₆ Te ₇ S	Polycrystalline	Solid_state_reaction__vacuum
NaCo ₂ O ₄	Polycrystalline	Solid_state_reaction__air
NaCo ₂ O ₄	Single crystal	Flux_(NaCl)__air
Nd _{1.4} Bi _{0.6} Ru ₂ O ₇	Polycrystalline	Solid_state_reaction
Nd ₂ Cu _{0.98} Ni _{0.02} O ₄	Polycrystalline	Solid_state_reaction__air
Nd ₂ Cu _{0.98} Zn _{0.02} O ₄	Polycrystalline	Solid_state_reaction__air

Accepted Article

Nd2CuO4	Polycrystalline	Solid_state_reaction__air
Nd2Ru2O7	Polycrystalline	Solid_state_reaction
Nd2Ru2O7	Polycrystalline	Ultrasonic_spray_pyrolysis
NdBirRu2O7	Polycrystalline	Solid_state_reaction
Ni2.0Mo6S8	Polycrystalline	Solid_state_reaction__vacuum
NiO	Polycrystalline	Solid_state_reaction
Pr0.5Ca0.5MnO3	Polycrystalline	Solid_state_reaction__air
Sm0.5Ca0.5MnO3	Polycrystalline	Solid_state_reaction__air
Sm1.7Ca0.3MnO3	Polycrystalline	Solid_state_reaction__air
Sr0.61Ba0.39Nb2O6	Polycrystalline	Templated_grain_growth__air__anneal_P(O2)_=_10\$^{-(14)}\$
Sr0.61Ba0.39Nb2O6	Single crystal	Czochralski_method__anneal_P(O2)_=_10\$^{-(14)}\$
Sr0.8La0.2TiO3	Polycrystalline	Solid_state_reaction__Ar
Sr0.95La0.05TiO3	Polycrystalline	Solid_state_reaction__Ar
Sr0.9La0.1TiO3	Polycrystalline	Solid_state_reaction__Ar
Sr0.9V0.1TiO3	Polycrystalline	Solid_state_reaction__air
Sr1.6La0.4Nb2O7	Single crystal	Floating_zone_method__Ar
Sr2Ti0.8Nb0.2O4	Polycrystalline	Solid_state_reaction__Ar
Sr3Ti1.6Nb0.4O7	Polycrystalline	Solid_state_reaction__Ar
Sr4.5Nb4.5O15.5	Single crystal	Floating_zone_method__Ar
SrDy0.08Ti0.92O3	Polycrystalline	Solid_state_reaction__air
SrMn0.96Mo0.04O3	Polycrystalline	Solid_state_reaction__air
SrMn0.98Mo0.02O3	Polycrystalline	Solid_state_reaction__air
SrNb0.15Ti0.85O3	Polycrystalline	Solid_state_reaction__Ar
SrNd0.17Ti0.83O3	Polycrystalline	Solid_state_reaction__air
SrNd0.24Ti0.76O3	Polycrystalline	Solid_state_reaction__air
SrNd0.2Ti0.8O3	Polycrystalline	Solid_state_reaction__air
SrTi0.8Nb0.2O3	Polycrystalline	Solid_state_reaction__Ar
SrTi0.8Nb0.2O3	Polycrystalline	Solid_state_reaction__air
TiCr5Se8	Polycrystalline	Solid_state_reaction
TiTiP5S	Polycrystalline	Solid_state_reaction__vacuum
W0.95O2.95Co0.10O.15	Polycrystalline	Solid_state_reaction
W0.99O2.97Co0.02O0.03	Polycrystalline	Solid_state_reaction
W0.9O2.7Co0.20O.3	Polycrystalline	Solid_state_reaction
WO2.722	Polycrystalline	Spark_plasma_sintering
WO2.9	Polycrystalline	Spark_plasma_sintering
WO3	Polycrystalline	Solid_state_reaction
Y0.5Bi1.5Ru2O7	Polycrystalline	Solid_state_reaction
Y2Ru2O7	Polycrystalline	Solid_state_reaction
Yb1.4Bi0.6Ru2O7	Polycrystalline	Solid_state_reaction
YbBirRu2O7	Polycrystalline	Solid_state_reaction
YBirRu2O7	Polycrystalline	Solid_state_reaction
Zn0.95Al0.05O	Polycrystalline	Solid_state_reaction__air
Zn0.97Al0.015Ti0.015O	Polycrystalline	Solid_state_reaction__air
Zn0.97Al0.01Ti0.02O	Polycrystalline	Solid_state_reaction__air
Zn0.97Al0.025Ti0.005O	Polycrystalline	Solid_state_reaction__air
Zn0.97Al0.02Ti0.01O	Polycrystalline	Solid_state_reaction__air
Zn0.97Al0.03O	Polycrystalline	Solid_state_reaction__air

Zn0.97Al0.030	Polycrystalline	Solid_state_reaction__vacuum
Zn0.98Al0.020	Polycrystalline	Solid_state_reaction__air
Zn0.9925Al0.00750	Polycrystalline	Solid_state_reaction__vacuum
Zn0.995Al0.0050	Polycrystalline	Solid_state_reaction__air
Zn0.995Al0.0050	Polycrystalline	Solid_state_reaction__vacuum
Zn0.9975Al0.00250	Polycrystalline	Microwave_solvothral__air
Zn0.9975Al0.00250	Polycrystalline	Solid_state_reaction__vacuum
Zn0.99Al0.010	Polycrystalline	Solid_state_reaction__air
Zn0.99Al0.010	Polycrystalline	Solid_state_reaction__vacuum
ZnO	Polycrystalline	Solid_state_reaction__air

Table S4. List of compounds in yellow cluster.

Chemical formula	Crystallinity	Production method
Ag0.15Sb0.15Te1.15Ge0.85	Single crystal	Melted
Ag9TlTe5	Polycrystalline	Melted
Ca2.85Na0.15AlSb3	Polycrystalline	Solid_state_reaction__Ar
Ca2.94Na0.06AlSb3	Polycrystalline	Solid_state_reaction__Ar
Ca2.97Na0.03AlSb3	Polycrystalline	Solid_state_reaction__Ar
Ca3AlSb3	Polycrystalline	Solid_state_reaction__Ar
Ca4.75Na0.25Al2Sb6	Polycrystalline	Solid_state_reaction__Ar
Ca4.95Na0.05Al2Sb6	Polycrystalline	Solid_state_reaction__Ar
Ca5Al2Sb6	Polycrystalline	Solid_state_reaction__Ar
CeFe1.5Co2.5Sb12	Polycrystalline	Melted__vacuum
CeFe2.5Co1.5Sb12	Polycrystalline	Melted__vacuum
CeFe2Co2Sb12	Polycrystalline	Combination_of_melting_and_powder_metallurgy_techniques
CeFe2Co2Sb12	Polycrystalline	Melted__vacuum
CeFe3.5Co0.5Sb12	Polycrystalline	Combination_of_melting_and_powder_metallurgy_techniques
CeFe3.5Co0.5Sb13	Polycrystalline	Combination_of_melting_and_powder_metallurgy_techniques
CeFe3.5Co0.5Sb14	Polycrystalline	Combination_of_melting_and_powder_metallurgy_techniques
CeFe3CoSb12	Polycrystalline	Combination_of_melting_and_powder_metallurgy_techniques
CeFe3CoSb12	Polycrystalline	Melted__vacuum
CeFe4Sb12	Polycrystalline	Combination_of_melting_and_powder_metallurgy_techniques
CeFe4Sb12	Polycrystalline	Melted__vacuum
CeFeCo3Sb12	Polycrystalline	Melted__vacuum
Fe0.998Co0.002Si2	Polycrystalline	Arc-melted__Ar
In0.05Co4Sb12	Polycrystalline	Solid_state_reaction
In0.15Co4Sb12	Polycrystalline	Solid_state_reaction
In0.1Co4Sb12	Polycrystalline	Solid_state_reaction
In0.25Co4Sb12	Polycrystalline	Solid_state_reaction
In0.3Co4Sb12	Polycrystalline	Solid_state_reaction
La2.74Te4	Polycrystalline	Solid_state_reaction__Ar
La2.99Te4	Polycrystalline	Solid_state_reaction__Ar
La3Te3.35Bi0.65	Polycrystalline	Solid_state_reaction__Ar
La3Te3.35Sb0.65	Polycrystalline	Solid_state_reaction__Ar
La3Te3.65Sb0.35	Polycrystalline	Solid_state_reaction__Ar
La3Te3.8Sb0.2	Polycrystalline	Solid_state_reaction__Ar
LaFe3CoSb12	Polycrystalline	Melted__vacuum

Accepted Article

Mg1.95Ca0.05Si	Polycrystalline	Melted__Ar
Mg1.9Ca0.1Si	Polycrystalline	Melted__Ar
Mg2Si0.6Ge0.4Ag0.02	Polycrystalline	Solid_state_reaction__He/H2
Mg2Si0.6Ge0.4Bi0.02	Polycrystalline	Solid_state_reaction__He/H2
Mg2Si0.98Ag0.02	Polycrystalline	Solid_state_reaction__He/H2
Mg2Si0.98Bi0.02	Polycrystalline	Solid_state_reaction__He/H2
Mg2Si0.993Bi0.007	Polycrystalline	Mechanochemical__Ar
Mg2Si0.994Bi0.006	Polycrystalline	Mechanochemical__Ar
Mg2Si0.995Bi0.005	Polycrystalline	Mechanochemical__Ar
Mg2Si0.997Bi0.003	Polycrystalline	Mechanochemical__Ar
Mg2Si0.9985Bi0.0015	Polycrystalline	Mechanochemical__Ar
Mg2Si0.999Bi0.001	Polycrystalline	Mechanochemical__Ar
Mg2Si	Polycrystalline	Bridgman_method__Ar-H_gas
Mg2Si	Polycrystalline	Melted__Ar
NbCo1.05Sn	Single crystal	Floating_zone_melting
NbCo1.10Sn	Single crystal	Floating_zone_melting
NbCoSn	Single crystal	Floating_zone_melting
Si0.7956Ge0.1989P0.0055	Polycrystalline	Vacuum_hot_pressed
Si0.79936Ge0.19984B0.0008	Polycrystalline	Vacuum_hot_pressed
Si0.8Ge0.2	Polycrystalline	Ball_milling__hot-pressed_nanopowders
Sr0.145Ga0.302Ge0.553	Polycrystalline	Solid_state_reaction
Sr0.146Ga0.285Ge0.569	Polycrystalline	Solid_state_reaction
Sr0.147Ga0.298Ge0.555	Polycrystalline	Solid_state_reaction
Ti0.95Nb0.05NiSn	Polycrystalline	Arc-melted__vacuum
Ti0.98Nb0.02NiSn	Polycrystalline	Arc-melted__vacuum
Ti0.99Nb0.01NiSn	Polycrystalline	Arc-melted__vacuum
TiNiSn	Polycrystalline	Arc-melted__vacuum
TiNiSn	Polycrystalline	Magnetic_levitation_induction_furnace
Ti0.01Pb0.99Te	Polycrystalline	Melted
Ti0.02Pb0.98Te	Polycrystalline	Melted
Ti2Cu2SnTe4	Polycrystalline	Solid_state_reaction__vacuum
Ti2GeTe5	Polycrystalline	Melted
Ti2SnTe5	Polycrystalline	Melted
Ti2SnTe5	Polycrystalline	Melted__hotpressed
Ti2SnTe5	Single crystal	Flux_(TiTe2)
Ti9BiTe6	Polycrystalline	Melted__zone_refined
Yb14MnSb11	Single crystal	Flux_(Sn)__Ar
Zn4Sb3	Polycrystalline	Melted__inert
Zr0.15Hf0.15Ti0.7NiSn	Polycrystalline	Arc-melted__Ar
Zr0.25Hf0.25Ti0.5NiSn	Polycrystalline	Arc-melted__Ar
Zr0.35Hf0.35Ti0.3NiSn	Polycrystalline	Arc-melted__Ar
Zr0.3Hf0.3Ti0.4NiSn	Polycrystalline	Arc-melted__Ar
Zr0.4Hf0.4Ti0.2NiSn	Polycrystalline	Arc-melted__Ar
Zr0.5Hf0.5NiSn1.994Sb0.006	Polycrystalline	Arc-melted__Ar
Zr0.5Hf0.5NiSn1.998Sb0.002	Polycrystalline	Arc-melted__Ar
Zr0.5Hf0.5NiSn	Polycrystalline	Arc-melted__Ar
Zr0.94Y0.06NiSn0.96Sb0.04	Polycrystalline	Arc_melted__Ar

Zr _{0.95} Nb _{0.05} NiSn	Polycrystalline	Arc-melted__vacuum
Zr _{0.98} Nb _{0.02} NiSn	Polycrystalline	Arc-melted__vacuum
Zr _{0.99} Nb _{0.01} NiSn	Polycrystalline	Arc-melted__vacuum
ZrNi _{0.76} Co _{0.004} Cu _{0.2} Sn	Polycrystalline	Arc_melted__Ar
ZrNi _{1.98} Cu _{0.02} Sn	Polycrystalline	Arc_melted__Ar
ZrNiSn _{0.98} Sb _{0.02}	Polycrystalline	Arc_melted__Ar
ZrNiSn	Polycrystalline	Arc-melted__vacuum

Table S5. List of compounds in green cluster.

Chemical formula	Crystallinity	Production method
CsBi ₄ Te ₆	Single crystal	Melted__air
K ₂ Bi ₈ Se ₁₃	Polycrystalline	Solid_state_reaction__+_extra
K ₂ Bi ₈ Se ₁₃	Single crystal	Flux__(Bi_self-flux)
KBi _{6.33} Si ₁₀	Polycrystalline	Solid_state_reaction__vacuum
Na _{0.02} Pb _{0.98} Te _{0.75} Se _{0.25}	Polycrystalline	Melted
Na _{0.02} Pb _{0.98} Te _{0.85} Se _{0.15}	Polycrystalline	Melted
Na _{0.02} Pb _{0.98} Te	Polycrystalline	Melted
Sb _{0.005} IO _{0.015} Cs _{0.995} Bi _{3.98} Te _{5.97}	Single crystal	Melted__air

Table S6. List of compounds in blue cluster.

Chemical formula	Crystallinity	Production method
Ba ₇ SrAl ₁₆ Si ₃₀	Polycrystalline	Flux__(Al)__dynamic__vacuum
Ba ₈ Au _{5.14} Si _{39.51}	Polycrystalline	Melted__inert
Ba ₈ Au _{5.59} Si _{39.01}	Polycrystalline	Melted__inert
Ba ₈ Au _{6.10} Si _{38.97}	Polycrystalline	Melted__inert
Ba ₈ Ga ₁₆ Ge ₃₀	Polycrystalline	Arc_melting
Ba ₈ Ga ₁₆ Ge ₃₀	Polycrystalline	Melted__Ar
Ba ₈ Ga ₁₆ Ge ₃₀	Polycrystalline	Solid_state_reaction__Ar
Ba ₈ Ga ₁₆ Ge ₃₀	Single crystal	Czochralski_method__argon
Ba ₈ Ga ₁₆ Ge ₃₀	Single crystal	Czochralski_method__He
Ba ₈ Ga ₁₆ Si ₃₀	Polycrystalline	Melted__Ar
Ba ₈ Ga ₁₆ Sn ₃₀	Polycrystalline	Melted__Ar
Ba ₈ Ga ₁₈ Ge ₂₈	Polycrystalline	Arc_melting
Sr ₈ Ga ₁₆ Ge ₃₀	Polycrystalline	Melted__Ar

Table S7. List of compounds in cyan cluster

Chemical formula	Crystallinity	Production method
Ti _{11.5} Sb _{11.5} Cu ₈ Se ₂₇	Polycrystalline	Solid_state_reaction__vacuum

Table S8. List of compounds in magenta cluster

Chemical formula	Crystallinity	Production method
Bi ₈₃ Sb ₁₇	Polycrystalline	Melted__vacuum
Bi ₈₆ Sb ₁₄	Polycrystalline	Melted__vacuum
Bi ₈₈ Sb ₁₂	Polycrystalline	Melted__vacuum
Bi ₉₀ Sb ₁₀	Polycrystalline	Melted__vacuum
Bi ₉₂ Sb ₈	Polycrystalline	Melted__vacuum

Accepted Article

S18

John Wiley & Sons, Inc.

This article is protected by copyright. All rights reserved.

S4. Attributes importance list.

Table S9. List of attributes (in descending order of importance) used in each temperature regime model.

Order of importance	@1000K	@700K	@400K	@300K
1	NUnfilledMean	ThermalConductivitySum	ThermalConductivityMean	ThermalConductivityMean
2	ThermalConductivityMean*	IonizationEnergySum*	CrystalStructCenteredTetragonal	ThermalConductivitySum
3	ThermalConductivitySum*	ThermalConductivitySum*	ThermalConductivityMax*	ThermalConductivityMax*
4	ThermalConductivityMean	ThermalConductivityMax*	ThermalConductivitySum*	ThermalConductivityMADFM
5	ThermalConductivityMADFM*	NValenceSum*	ThermalConductivityMADFM	NValenceSum*
6	ThermalConductivityMax*	ElectronegativityByMillarMin	ElectronegativityByMillarMin	ThermalConductivitySum*
7	DensityMADFM*	ThermalConductivityMean	ThermalConductivityMean*	ThermalConductivityMax
8	ElectronegativityByMillarMin	ThermalConductivityMax	ThermalConductivitySum	ThermalConductivityMean*
9	IonizationEnergySum*	DensityMADFM*	ElectronegativityByMillarMADFM	ThermalConductivityMADFM*
10	ElectronAffinitySum*	DensityMax*	ElectronAffinitySum*	AtomicNumberMax*
11	DensityMax	ThermalConductivityMADFM*	atomic_weightMax*	DensityMax*
12	VE_by_VillarsSum*	MeltingPointMADFM	atomic_weightMean*	DensityMADFM*
13	NUnfilledMax	NsUnfilledSum	CrystalRadiusMean	NValenceMean*
14	ThermalConductivitySum	AtomicNumberSum*	ThermalConductivityMADFM*	DensitySum*
15	NUnfilledMADFM	NdValenceSum*	NUnfilledMean	NdValenceSum*
16	DensityMADFM	DensityMADFM	MeltingPointMADFM	atomic_weightSum*
17	ElectronAffinityMean	IonicRadiusSum*	DensityMADFM*	VE_by_VillarsMean*
18	ElectronAffinitySum	NUnfilledMean	DensityMax*	DensityMax
19	GroupSum*	ElectronegativityByMillarSum*	NpValenceMean*	NValenceMADFM*
20	CrystalRadiusMADFM	AtomicNumberMax*	CovalentRadiusMADFM	AtomicNumberMADFM*
21	BoilingPointMADFM*	ElectronAffinityMean	AtomicNumberMax*	CovalentRadiusMADFM
22	NValenceSum*	PeriodSum*	ElectronAffinitySum	ElectronAffinityMean*
23	ElectronegativityByMillarMADFM*	NValenceMean*	ThermalConductivityMax	VE_by_VillarsSum*
24	MeltingPointMADFM*	ElectronegativityByMillarMADFM	VE_by_VillarsSum*	BoilingPointMADFM*
25	NUnfilledSum	DensityMax	NsUnfilledSum	AtomicNumberSum*
26	ElectronegativityByMillarSum*	CovalentRadiusSum*	ElectronAffinityMADFM	DensityMean*
27	CovalentRadiusMADFM	BoilingPointMax	ElectronAffinityMADFM*	ElectronAffinitySum
28	DensityMax*	ThermalConductivityMean*	DensityMADFM	IonicRadiusSum*

ElectronAffinityMADFM	NpValenceMean*	IonizationEnergyMean*	atomic_weightMean*
ThermalConductivityMADFM CrystalStructCenteredTe	tragonal	AtomicNumberMADFM*	AtomicNumberMean*
ThermalConductivityMax	PeriodMean*	ElectronegativityByMillarMean*	IonicRadiusMADFM
NfValenceMean*	IonicRadiusMADFM	NsValenceMean	ElectronegativityByMillarMADFM*
ElectronAffinityMADFM*	DensitySum*	NValenceMean*	NpValenceMean*
transition_metalsMADFM*	NValenceMADFM*	AtomicNumberSum*	GroupMean*
DensitySum*	NsUnfilledMean*	NValenceMADFM*	AtomicRadiusMean*
NUnfilledMean*	NdValenceMean*	CrystalRadiusMADFM	VE_by_VillarsSum
CrystalRadiusSum*	NUnfilledMax*	NValenceSum*	GroupSum*
NUnfilledMADFM*	NsUnfilledMADFM	IonizationEnergySum*	CovalentRadiusSum*
CovalentRadiusMax	atomic_weightSum*	DensitySum*	NValenceSum
transition_metalsMax*	NpValenceMADFM*	CovalentRadiusSum*	HeatVaporizatioMADFM*
AtomicRadiusMean	ElectronegativityByMillarMax*	NpValenceMADFM*	VE_by_VillarsMADFM
AtomicNumberSum*	ThermalConductivityMADFM	NdValenceSum*	ElectronAffinityMADFM*
NfValenceSum*	HeatVaporizatioMADFM*	atomic_weightSum*	NpValenceSum*
BoilingPointMax*	IonizationEnergyMADFM*	ElectronAffinityMean*	NsUnfilledMean
MeltingPointMax*	ElectronegativityByMillarMean	CovalentRadiusMADFM*	NdValenceMean*
AtomicRadiusMax	NsValenceMean	GroupMin*	CrystalRadiusMean
IonizationEnergyMin	ElectronAffinityMean*	PeriodSum*	atomic_weightMADFM*
NsValenceSum*	GroupSum*	NsUnfilledMADFM	ZungerRadiusMADFM*
VE_by_VillarsMin	AtomicRadiusSum*	HeatVaporizatioSum*	IonicRadiusMean
IonizationEnergyMean	DensityMean*	AtomicRadiusMin*	NsValenceMean
ElectronegativityByMillarMean BoilingPointMADFM		HeatVaporizatioMADFM*	ElectronegativityByMillarMean*
HeatVaporizatioMADFM*	NpValenceSum*	ElectronegativityByMillarMADFM*	AtomicRadiusMADFM*
NUnfilledMax*	ElectronAffinitySum	MeltingPointMADFM*	MeltingPointMADFM*
NdUnfilledSum*	NUnfilledMADFM*	atomic_weightMADFM*	CrystalRadiusSum*
VE_by_VillarsMean	NpUnfilledMADFM*	NUnfilledSum*	IonicRadiusMax*
PeriodSum*	atomic_weightMax*	NdValenceMean*	CrystalRadiusMean*
NUnfilledSum*	ElectronAffinityMADFM*	NdValenceMADFM*	NValenceMean
atomic_weightMADFM*	CrystalRadiusSum*	NsUnfilledMax*	CrystalRadiusSum
HeatVaporizatioSum*	NpUnfilledMean*	CrystalRadiusMADFM*	IonizationEnergyMADFM*
NsUnfilledSum*	NpUnfilledSum*	ElectronAffinityMin	IonizationEnergySum*
AtomicNumberMADFM*	CrystalRadiusMean	ElectronegativityByMillarSum*	DensityMin*

MolarVolumeMin	CovalentRadiusMADFM	NpUnfilledMADFM*	ElectronAffinityMean
NsUnfilledMean	NsUnfilledMean	NsValenceMADFM	ZungerRadiusMean*
AtomicRadiusSum*	ElectronegativityByMillarMean*	CovalentRadiusMax	IonicRadiusMin*
NdUnfilledMADFM*	GroupMean*	NUnfilledMADFM	NValenceMax*
AtomicNumberMax*	ElectronAffinitySum*	DensityMax	NsUnfilledSum
atomic_weightMax*	GroupMADFM*	AtomicRadiusMean*	NpUnfilledMean*
NValenceMADFM	AtomicRadiusMax*	GroupSum*	NdValenceMax*
NValenceMax*	VE_by_VillarsMADFM*	AtomicRadiusMean	NdValenceMADFM*
NpValenceMADFM	NpUnfilledMADFM	MolarVolumeMADFM*	NsValenceMADFM*
ElectronAffinityMin	IonicRadiusMean*	VE_by_VillarsMean*	CovalentRadiusMax*
CrystalRadiusMADFM*	VE_by_VillarsMADFM	CovalentRadiusMin*	NUnfilledSum*
atomic_weightSum	VE_by_VillarsMean*	MeltingPointMean*	VE_by_VillarsMean
CovalentRadiusSum*	NsValenceSum*	AtomicRadiusMADFM*	AtomicRadiusMin*
IonicRadiusSum*	AtomicRadiusMax	CrystalRadiusSum*	HeatVaporizatioSum*
DensityMin*	NUnfilledSum*	IonicRadiusMax*	CovalentRadiusMean*
PeriodMADFM	NdValenceSum	NpUnfilledMADFM	DensityMean
MeltingPointSum*	NUnfilledMean*	NdUnfilledSum*	VE_by_VillarsMax
IonizationEnergyMADFM*	MolarVolumeMADFM*	CovalentRadiusMean*	DensityMADFM
ElectronegativityByMillarMean* BoilingPointMADFM	*	transition_metalsMean*	atomic_weightMin*
MeltingPointMin*	NUnfilledMADFM	BoilingPointMean*	MeltingPointMean*
CrystalRadiusMean	NValenceMin	HeatVaporizatioMean*	AtomicRadiusSum*
IonizationEnergyMax*	GroupMax*	NsUnfilledMean*	IonizationEnergyMean
NpValenceMean*	atomic_weightMean	IonicRadiusMean*	IonicRadiusMADFM*
VE_by_VillarsMean*	atomic_weightMean*	NsUnfilledMean	PeriodMADFM*
NdUnfilledMean*	HeatVaporizatioMax*	CrystalRadiusMax	CrystalRadiusMADFM
NdValenceMADFM*	VE_by_VillarsMean	AtomicNumberMin*	ElectronegativityByMillarMADFM
NsUnfilledSum	NpUnfilledMax	NsUnfilledSum*	NsValenceMADFM
GroupMean*	CrystalRadiusMean*	PeriodMADFM*	CovalentRadiusMean
MeltingPointMean*	CrystalRadiusMADFM*	IonicRadiusSum*	MeltingPointMin*
NpUnfilledSum*	MolarVolumeMean	AtomicRadiusMADFM	GroupMin*
atomic_weightSum*	AtomicRadiusMADFM	IonicRadiusMin*	NUnfilledMean
MolarVolumeSum	AtomicRadiusMean	GroupMADFM	NdUnfilledMean*
NdValenceSum*	VE_by_VillarsMin	AtomicNumberMean	ElectronAffinityMax*

atomic_weightMADFM	NsValenceMADFM	PeriodMean*	MeltingPointMax*
NpValenceSum*	NUnfilledSum	NsUnfilledMADFM*	IonicRadiusMean*
CovalentRadiusMADFM*	CrystalRadiusMADFM	ZungerRadiusMADFM*	atomic_weightSum
MolarVolumeSum*	DensityMin	ElectronegativityByMillarMean	NdValenceSum
NValenceMADFM*	ElectronAffinityMADFM	AtomicNumberMADFM	HeatVaporizatioMean*
NfValenceMADFM*	AtomicNumberMADFM*	VE_by_VillarsMax	BoilingPointMean*
NdUnfilledMax*	NdUnfilledMean*	NdValenceMADFM	BoilingPointMax*
CrystalRadiusSum	HeatVaporizatioMean*	NValenceMin*	AtomicRadiusMax*
BoilingPointMADFM	GroupMin*	atomic_weightMean	atomic_weightMax
GroupMin*	CovalentRadiusMax	VE_by_VillarsSum	NpUnfilledMADFM
alkaliMean	BoilingPointMin*	AtomicRadiusMax*	AtomicNumberMean
MeltingPointMADFM	NValenceMin*	ThermalConductivityMin*	atomic_weightMADFM
ElectronegativityByMillarMin* HeatVaporizatioMi	n*	VE_by_VillarsMean	NdValenceMADFM
alkaliSum*	MolarVolumeSum	BoilingPointMax	BoilingPointMean
AtomicRadiusMean*	transition_metalsMADFM*	MeltingPointSum*	NsValenceMean*
ElectronegativityByMillarMax* VE_by_VillarsMax*		BoilingPointMax*	PeriodMax*
VE_by_VillarsMADFM*	NdValenceMADFM*	NValenceMean	MeltingPointSum*
NfValenceMax*	NsUnfilledMADFM*	DensityMean	CovalentRadiusMax
HeatVaporizatioMADFM	DensityMin*	NdValenceMean	IonicRadiusMax
NsValenceMADFM*	IonizationEnergyMax*	NpUnfilledSum*	VE_by_VillarsMin*
NpUnfilledMADFM*	NdValenceMean	HeatVaporizatioMean	ZungerRadiusMax*
CrystalRadiusMin*	CovalentRadiusMean	CovalentRadiusMax*	AtomicRadiusMean
ZungerRadiusMADFM*	MeltingPointMax	MolarVolumeMean	ElectronegativityByMillarMean
BoilingPointMax	AtomicNumberSum	ElectronAffinityMin*	NpUnfilledSum*
AtomicRadiusMADFM	CrystalRadiusMax*	AtomicRadiusMax	MolarVolumeMADFM*
AtomicNumberSum	alkaliMean*	MolarVolumeMin*	CrystalStructCenteredTetragonal
NdValenceMADFM	NpUnfilledMean	DensitySum	NValenceMADFM
NsUnfilledMADFM*	IonizationEnergyMin*	transition_metalsMADFM*	IonizationEnergySum
CrystalRadiusMin	CovalentRadiusMADFM*	PeriodMin*	AtomicNumberMax
BoilingPointMin*	ElectronAffinityMin*	postTransition_metalsMADFM*	AtomicRadiusMax
alkaline_earth_metalsMADFM* HeatVaporizatioMax		NpValenceMax*	GroupMADFM
NsUnfilledMADFM	DensitySum	NValenceSum	BoilingPointMax
DensitySum	DensityMean	CrystalRadiusMin	MeltingPointMean

ElectronegativityByMillarSum NValenceMax*		BoilingPointMin*	HeatVaporizatioMADFM
NdValenceMax*	IonicRadiusMean	IonicRadiusMean	NUnfilledMADFM
transition_metalsMean*	ElectronAffinityMax	NpUnfilledMax*	postTransition_metalsMADFM*
MolarVolumeMADFM*	NdUnfilledMax*	NpUnfilledMax	MolarVolumeSum*
VE_by_VillarsMax	AtomicNumberMin*	NUnfilledSum	ElectronegativityByMillarMax*
ElectronAffinityMean*	alkaliSum*	MeltingPointMean	MolarVolumeMean*
alkaliMax*	BoilingPointSum	AtomicRadiusSum	alkaline_earth_metalsMax*
MeltingPointMax	GroupMADFM	MeltingPointSum	postTransition_metalsMean
AtomicRadiusMADFM*	IonicRadiusMin*	NValenceMADFM	IonizationEnergyMax*
alkaline_earth_metalsSum*	CovalentRadiusSum	GroupSum	transition_metalsSum*
CovalentRadiusSum	NpUnfilledSum	NUnfilledMax	HeatVaporizatioMax
AtomicNumberMean	IonicRadiusSum	NpUnfilledSum	NpValenceMADFM
NdUnfilledMean	NValenceMean	NsUnfilledMax	GroupMin
Solid_state_reaction	NsValenceMADFM*	Solid_state_reaction_air	NpUnfilledMax
Solid_state_reaction_air	Melted	alkaliMean	MolarVolumeMin
Solid_state_reaction_Ar	Solid_state_reaction	Solid_state_reaction	Solid_state_reaction_air
NpUnfilledMADFM	Solid_state_reaction_Ar	Melted	Solid_state_reaction_vacuum
Arc_melting	Melted__Ar	Solid_state_reaction__Ar	Solid_state_reaction_oxalates
Sol_gel__air	Solid_state_reaction_air	Melted_vacuum	Crystallinity
Microwave_solvothetmal__air Arc_melting		Crystallinity	Solid_state_reaction__Ar
Crystallinity	Crystallinity	Solid_state_reaction__vacuum	Arc_melted_vacuum
Solid_state_reaction__vacuum Solid_state_reacti	on__vacuum	Solid_state_reaction_oxalates	Bridgman_method__Ar_H_gas
Czochralski_method_He	Mechanochemical_Ar	Solid_state_reaction__He_H2	Sol_gel__air
Flux_SrCl2__air	Czochralski_method__He	Sol_gel__air	Melted__Ar
Czochralski_method__argon Sol_gel__air		Floating_zone_melting	Microwave_solvothetmal__air
Solid_state_reaction_oxalates Combination_of_me	lting_and_powder_metallurgy_techniques	Arc_melting	Arc_melting
Melted		Microwave_solvothetmal__air	Templated_grain_growth__air__anneal_P_O2_eq_10pow_neg14
Ultrasonic_spray_pyrolysis	Solid_state_reaction_oxalates	Arc_melted_vacuum	Melted__hotpressed
Flux_NaCl__air	Flux_SrCl2__air	Co_precipitation	Flux_TiTe2
Solid_state_reaction__He_H2 Microwave_solvothet	Solid_state_reaction__He_H2	Melted__zone_refined	Magnetic_levitation_induction_furnace
Evaporate_nitrates_1173_K__air Arc_melted_vacuu	mal__air	Melted__Ar	Ultrasonic_spray_pyrolysis
Flux_Sn__Ar	m	Melted__Ar	Melted_vacuum
Melted_vacuum	Melted_vacuum	Flux_SrCl2__air	Co_precipitation
	Czochralski_method__argon	Melted__inert	

Solid_state_reaction_under_oxygen Co_precipitat	ion	Czochralski_method__anneal_P_O2_eq_10pow_neg14	Flux_Bi_self_flux
Solid_state_reaction_O2_atmosphere Solid_state_	reaction_O2_atmosphere	Evaporate_nitrates_1173_K__air	Arc_melted_Ar
Arc_melted_vacuum	Evaporate_nitrates_1173_K__air	Magnetic_levitation_induction_furnace	Flux_SrCl2__air
Co_precipitation	Solid_state_reaction_under_oxygen	Solid_state_reaction_O2_atmosphere	Melted__zone_refined
Flux_Al__dynamic_vacuum	Magnetic_levitation_induction_furnace	Flux_Sn__Ar	CrystalStructFaceCenteredOrthorhombic
Magnetic_levitation_induction_furnace Melted__i	ner	Czochralski_method__He	Floating_zone_melting
Floating_zone_melting	Czochralski_method__anneal_P_O2_eq_10pow_neg14	Arc_melted_Ar	Mechanochemical_Ar
Arc_melted_Ar	Flux_Al__dynamic_vacuum	Ultrasonic_spray_pyrolysis	Solid_state_reaction__He_H2
Ball_milling__hot_pressed_nanopowders Bridgman_	method__Ar_H_gas	Templated_grain_growth__air__anneal_P_O2_eq_10pow_neg14	Combination_of_melting_and_powder_metallurgy_techniques
Bridgman_method__Ar_H_gas Flux_NaCl__air	chniques	Combination_of_melting_and_powder_metallurgy_techniques	Melted__inert
Combination_of_melting_and_powder_metallurgy_te	chniques	Czochralski_method__argon	Solid_state_reaction_pl_extra
Czochralski_method__anneal_P_O2_eq_10pow_neg14	Ultrasonic_spray_pyrolysis	Bridgman_method__Ar_H_gas	Flux_Al__dynamic_vacuum
Floating_zone_method__Ar	Floating_zone_melting	Solid_state_reaction_under_oxygen	Floating_zone_method__Ar
Flux_Bi_self_flux	Flux_Sn__Ar	Flux_NaCl__air	Solid_state_reaction_O2_atmosphere
Flux_TiTe2	Solid_state_reaction__sealed	Flux_Al__dynamic_vacuum	Melted__air
Mechanochemical_Ar	Arc_melted_Ar	Spark_plasma_sintering	Czochralski_method__He
Melted__air	Ball_milling__hot_pressed_nanopowders	Mechanochemical_Ar	Flux_NaCl__air
Melted__Ar	Floating_zone_method__Ar	Solid_state_reaction__sealed	Solid_state_reaction_under_oxygen
Melted_hotpressed	Flux_Bi_self_flux	Ball_milling__hot_pressed_nanopowders	Flux_Sn__Ar
Melted__inert	Flux_TiTe2	Floating_zone_method__Ar	Czochralski_method__argon
Melted__zone_refined	Melted__air	Flux_Bi_self_flux	Spark_plasma_sintering
Solid_liquid_vapor_reaction__vacuum Melted__hot	pressed	Flux_TiTe2	Solid_state_reaction__sealed
Solid_state_reaction_pl_extra Melted__zone_refi	ned	Melted__air	Ball_milling__hot_pressed_nanopowders
Solid_state_reaction__sealed Solid_liquid_vapor	__reaction__vacuum	Melted_hotpressed	Czochralski_method__anneal_P_O2_eq_10pow_neg14
Spark_plasma_sintering	Solid_state_reaction_pl_extra	Solid_liquid_vapor_reaction__vacuum	Evaporate_nitrates_1173_K__air
Templated_grain_growth__air__anneal_P_O2_eq_10pow_neg14	Spark_plasma_sintering	Solid_state_reaction_pl_extra	Solid_liquid_vapor_reaction__vacuum
Vacuum_hot_pressed	Vacuum_hot_pressed	Vacuum_hot_pressed	Vacuum_hot_pressed

Part S5. Correlations between the 30 most important attributes

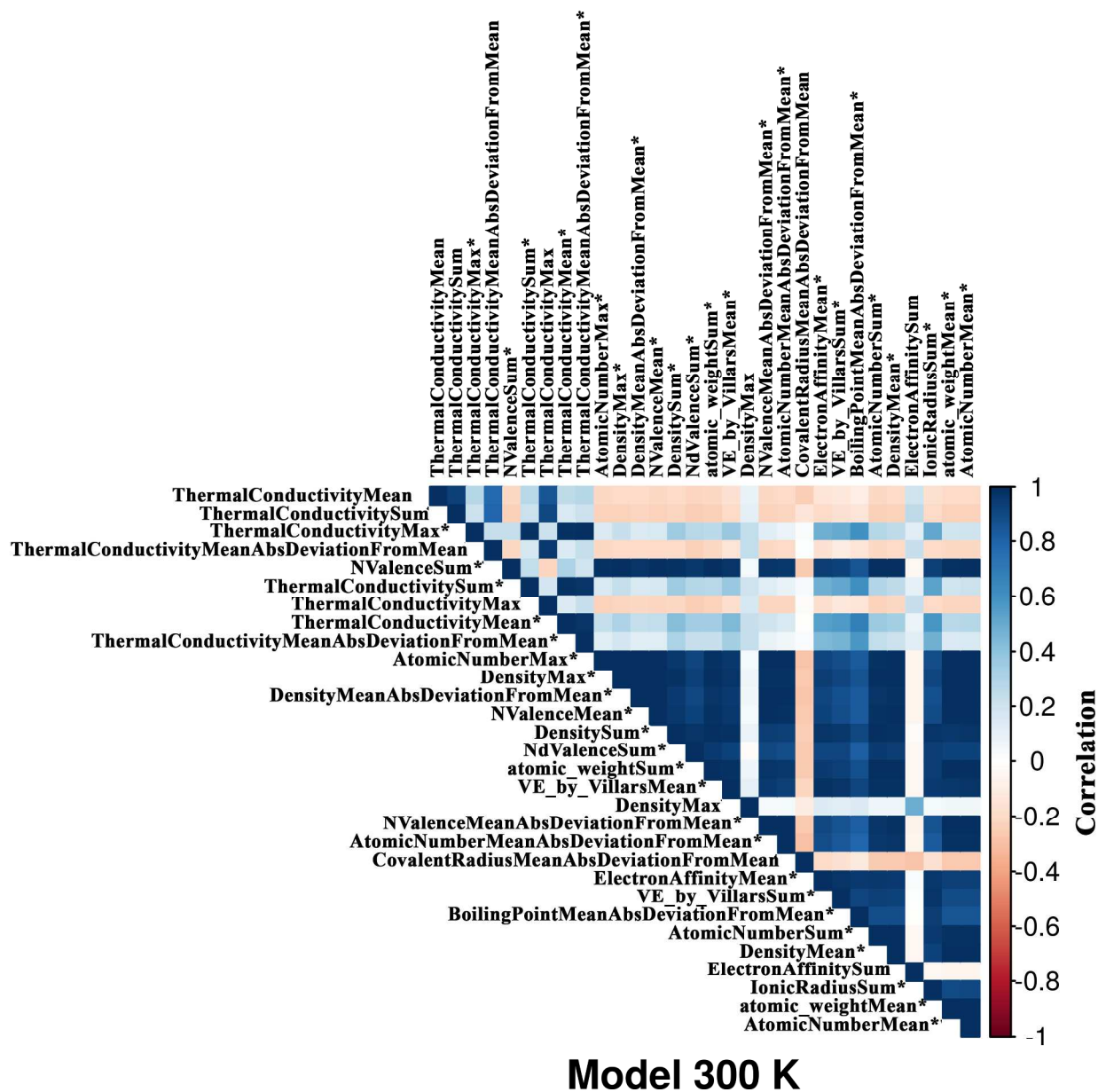


Figure S3. Cross-correlation for the top 30 important features in model at 300K temperature

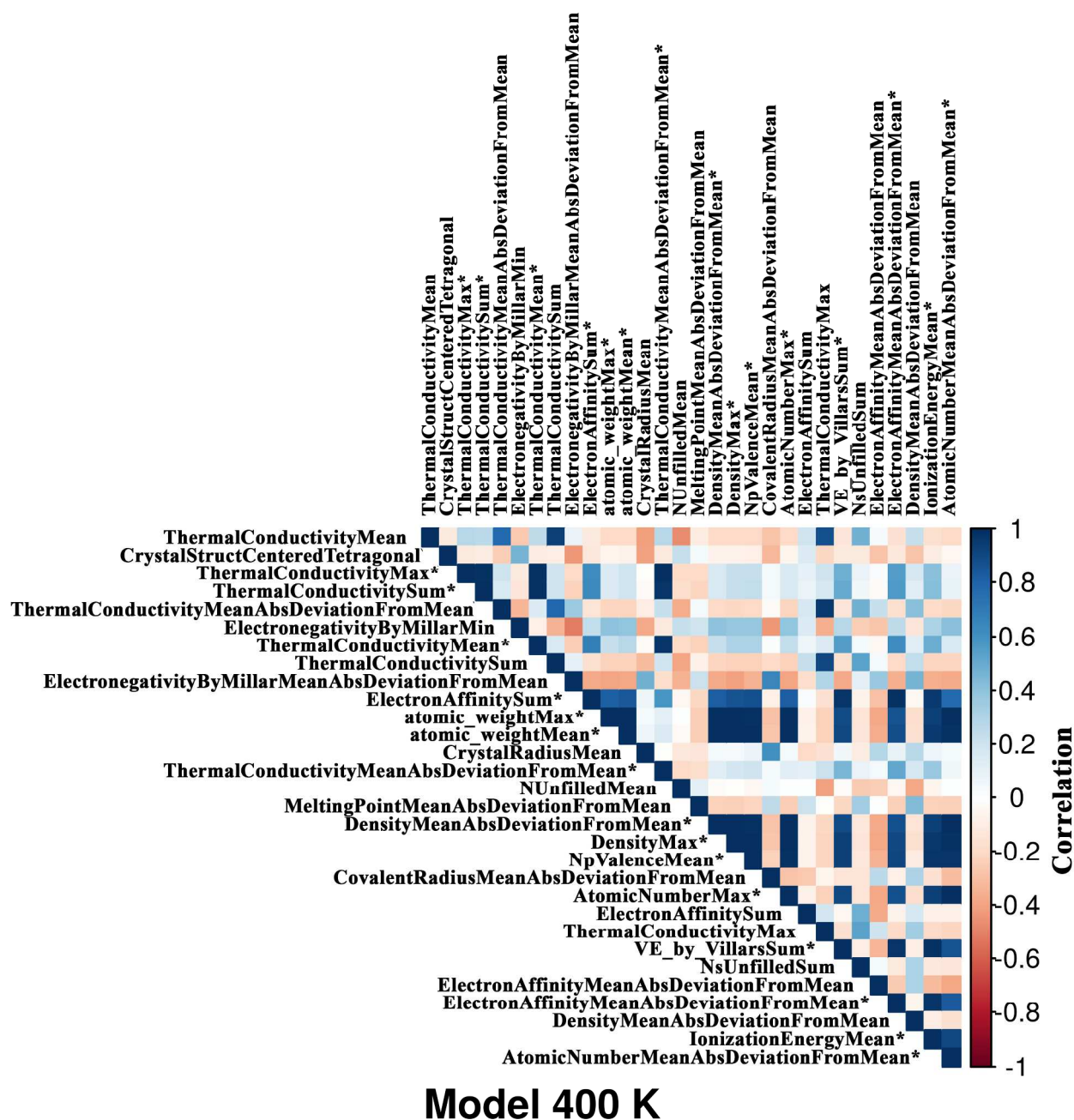
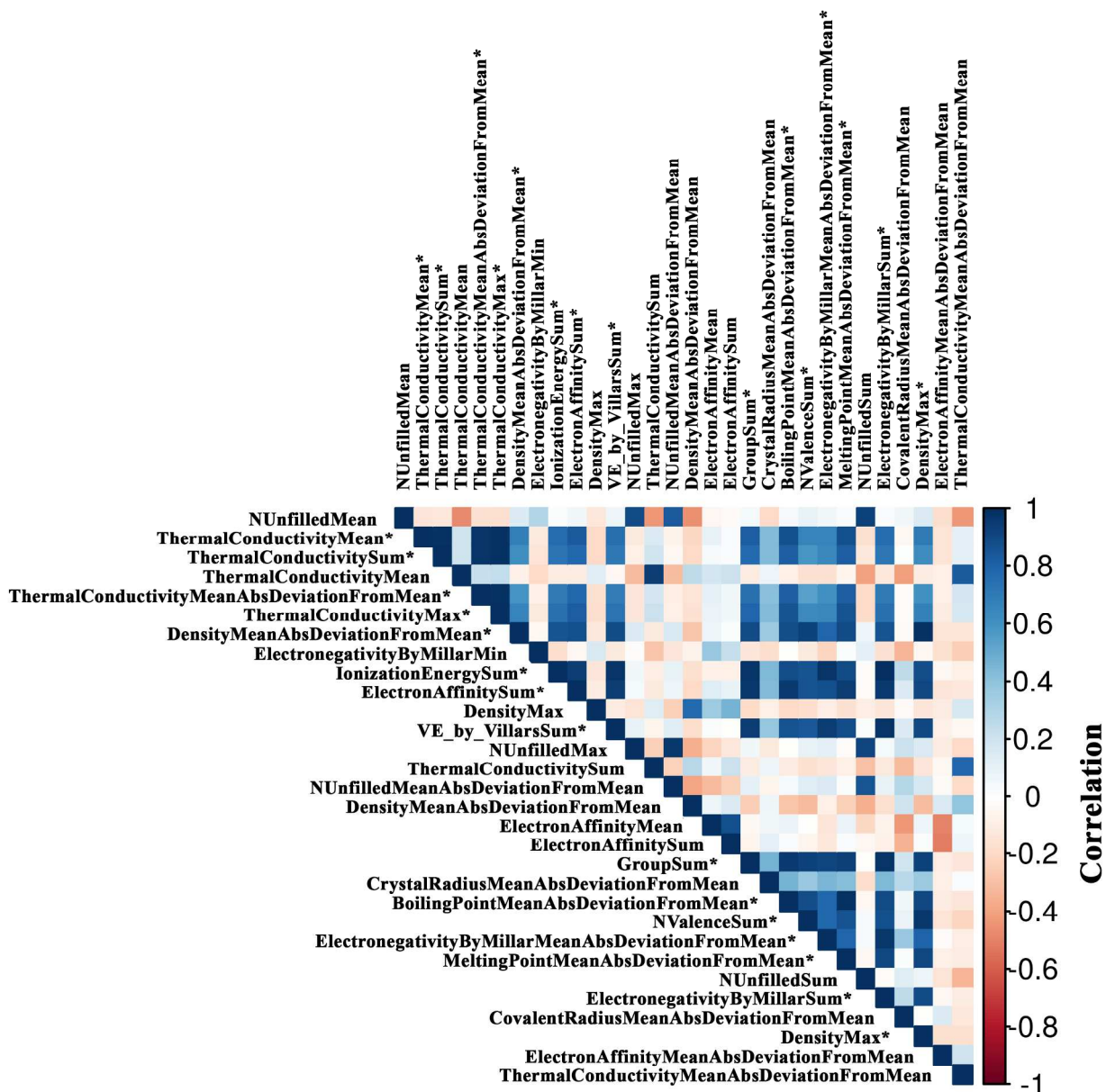


Figure S4. Cross-correlation for the top 30 important features in model at 400K temperature

AC



Model 1000 K

Figure S6. Cross-correlation for the top 30 important features in model at 1000K temperature

References

1. A. Furmanchuk, A. Agrawal, J. Saal, J. Doak, G. B. Olson and A. Choudhary, ThermoEl web tool.(2016) Available at: <http://info.eecs.northwestern.edu/ThermoEl> (Accessed: 24th August 2016)).
2. A. D. Gordon, *Classification*, Boca Raton, London CRC, 2nd edn., 1999.
3. Y. Zhou, I. Matsubara, R. Funahashi and S. Sodeoka, *Mater. Let.*, 2001, **51**, 347-350.
4. J. Lan, Y. Lin, A. Mei, C. Nan, Y. Liu, B. Zhang and J. Li, *J. Mater. Sci. Technol.*, 2009, **25**, 535-538.
5. E. S. Toberer, M. Christensen, B. B. Iversen and G. J. Snyder, *Phys. Rev. B* 2008, **77**, 075203.
6. C. Candolfi, U. Aydemir, M. Baitinger, N. Oeschler, F. Steglich and Y. and Yu . Grin, *J. Appl. Phys.* , 2012, **111**, 043706.
7. M. A. McGuire, T. K. Reynolds and F. J. DiSalvo, *Chem. Mater.*, 2005, **17**, 2875–2884.

Accepted Article