nature

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41586-019-1335-8

In the format provided by the authors and unedited.

Unsupervised word embeddings capture latent knowledge from materials science literature

Vahe Tshitoyan[1,3]*, John Dagdelen[1,2], Leigh Weston[1], Alexander Dunn[1,2], Ziqin Rong[1], Olga Kononova[2], Kristin A. Persson[1,2], Gerbrand Ceder[1,2]* & Anubhav Jain[1]*

[1]Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [2]Department of Materials Science and Engineering, University of California, Berkeley, CA, USA. [3]Present address: Google LLC, Mountain View, CA, USA. *e-mail: vahe.tshitoyan@gmail.com; gceder@lbl.gov; ajain@lbl.gov

Unsupervised word embeddings capture latent knowledge from materials science literature

Supplementary Information

Vahe Tshitoyan[1], John Dagdelen[1,2], Leigh Weston[1], Alexander Dunn[1,2], Ziqin Rong[1], Olga Kononova[2], Kristin A. Persson[1,2], Gerbrand Ceder[1,2] & Anubhav Jain[1]

[1]*Lawrence Berkeley National Laboratory, Berkeley 94720, California, USA*

[2]*Department of Materials Science and Engineering, University of California, Berkeley, California 94720, USA*

S1 Word2vec Skip-gram

Skip-gram, one of the two variants of Word2vec, is explained schematically in fig. S1a. Assume we have $\mathrm{V} = 500,000$ unique words in the vocabulary with each word assigned an arbitrarily index, so that it can be represented as a V-dimensional vector with zeros everywhere except that index. This representation is called one-hot encoding. Word2vec skip-gram loops through all words in the training text and uses its one-hot encoding as an input for a neural network. The task of the network is to predict all words within a certain

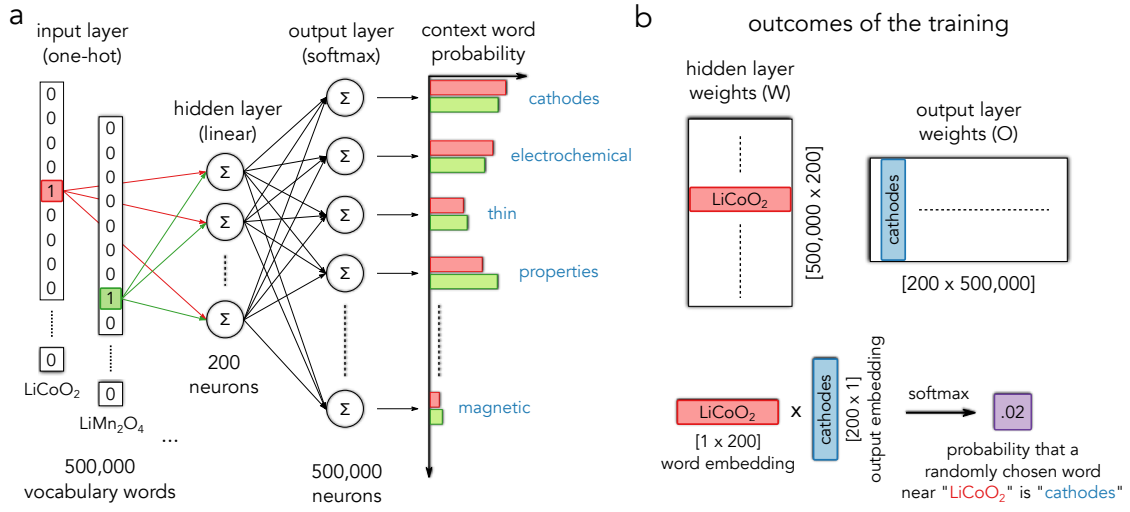


Figure S1: Word2vec skip-gram. a. A neural network with a single linear hidden layer learns to predict context words for every word in the vocabulary. For battery cathode materials $LiCoO_2$ and $LiMn_2O_4$ the network has to predict mostly the same context words. This results in similar hidden layer weights and therefore similar word embeddings. The softmax function is used at the output to produce normalized probabilities. **b.** Matrices W and O are the outcomes of the training, corresponding to the weights of the hidden and the output layers. Rows of W are called word embeddings, whereas columns of O are called output embeddings. The product of the two types of embeddings is the probability of the corresponding words to be used in close proximity in the text.

distance from this center word (usually ranging from 2 - 10 words away)*. While there is no single correct answer - every word occurs alongside 100s or 1000s of other words - the

*Larger word window often captures semantic relationships better, whereas smaller windows capture the syntactic relationships. We chose a relatively large window size of 8 to focus on semantic relationships, since this is more relevant for materials science relationships such as oxides of materials or common crystal structures.

end goal is not to correctly predict all neighbours but to learn compressed representations for the words. This representation is encoded in the weights of the single linear hidden layer of the neural network at the end of the training. The weights of the hidden layer are given by a [V x n] dimensional matrix W (fig. S1b), where n is the size of the space we set to "embed" the words in (200 in our case). When the one-hot encoded vector of the center word is fed into the network, all it does is select the corresponding row from matrix W. Then the output layer uses this row as an input for the softmax classifier to predict one of the neighbouring words. The classifier has to predict the same words for the words that occur in the same context, therefore, the network will adjust the corresponding rows of the matrix W to optimize this task. These row vectors are referred to as word vectors or word embeddings. Similarly, columns of the [n x V] matrix O of output weights are called output embeddings. In this notation, the task of the neural network is reduced to multiplying the row w of matrix W with the columns of matrix O and applying a softmax function, producing the probabilities of every word in the vocabulary to be next to the word w (fig. S1b).

The other variation of Word2vec is called continuous bag of words (CBOW). The neural network architecture is very similar, except instead of using the center word to predict the context words it uses the average embedding of the context words (hence, bag of words) to predict the center word. In the next section we demonstrate that Skip-

gram generally works better than CBOW for our application, therefore, we use Skip-gram throughout this work.

S2 Word2vec optimization

We tuned hyper-parameters of Word2vec to optimize its performance on the combined materials science and grammatical analogies. The full list of categorized analogies is available with supplementary materials. We found that including phrases as described in the Methods section of the main text improves the performance by approximately 4% for both CBOW and Skip-gram architectures, as shown in Table S1. We also find that Skip-gram performs approximately 4% better than CBOW both with and without phrases. We used negative sampling loss since it is faster to train. The rest of the hyperparameter optimization is summarized in table S2. We also trained GloVe embeddings[1] resulting in slightly worse performance compared to Word2vec (Table S1).

We check if analogy-based optimization leads to better performance for materials predictions using two additional metrics - one to quantify the quality of the predictions and the other for the quality of the ranking. For predictions, we use the average power factor of the first 10 predicted thermoelectrics. For the ranking, we compute the Spearman rank correlation[5] of our ranking versus approximately 80 experimental thermoelectric figures

Algorithm	Materials	Grammar	All
Default	38.0	50.4	44.4
CBOW	48.9	54.9	52.0
CBOW + phrases	54.2	58.0	56.2
Skip-gram	54.7	58.2	56.5
Skip-gram + phrases	**58.9**	**61.6**	**60.3**
GloVe + phrases	53.8	56.0	55.0

Table S1: Algorithm choice. Top 1 analogy scores in % for materials science and grammatical analogy tasks. Each task consists of approximately 15,000 analogy pairs. The answer is considered correct only if the first nearest word matches the expected analogy. The default algorithm uses the original hyperparameters of the Word2vec code[7], whereas the other four Word2vec algorithms use the optimized hyper-parameters. The GloVe algorithm uses the recommended parameters from the original paper[1], found to perform the best after trying to optimize the context window and the parameter alpha.

initial learning rate:	0.001	0.003	**0.01**	0.03	0.1
	50.6	54.6	56.8	55.1	52.6
downsampling:	10^{-3}	$\mathbf{10^{-4}}$	10^{-5}	10^{-6}	
	56.8	58.2	56.5	50.6	
dimension:	100	**200**	300	400	
	54.7	60.4	60.5	59.0	
negative samples:	5	8	10	12	**15**
	59.3	59.5	59.8	59.8	60.3

Table S2: Hyper-parameter optimization. Top 1 analogy score in % for various hyper-parameter choices. Only one parameter is varied while the rest are kept the same.

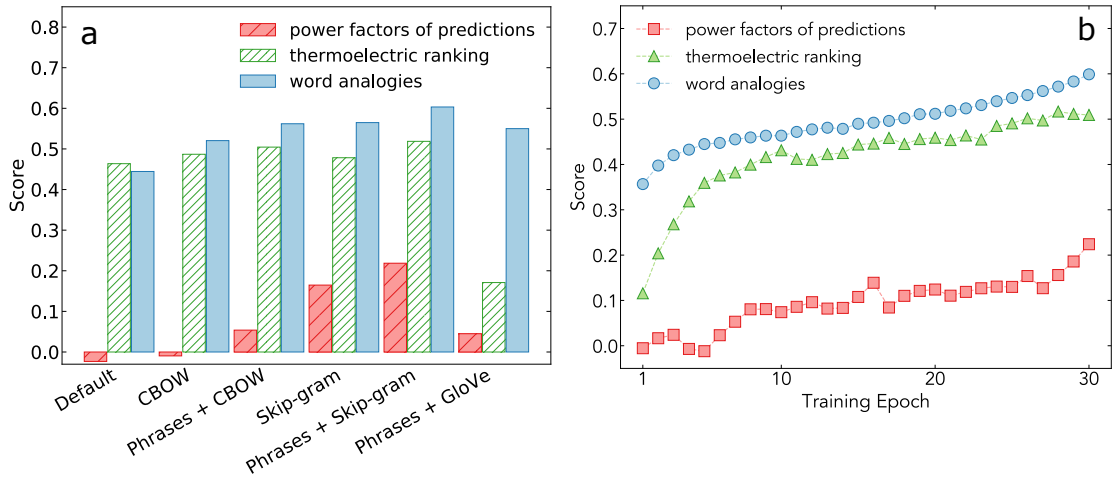


Figure S2: Accuracy of predictions. a. Performance metrics for different algorithms and parameters. Word analogies (blue) are analogy scores based on materials science and grammatical analogy tasks[2–4]. Thermoelectric ranking score (green) is the Spearman rank correlation coefficient[5] between the rank of our predictions and the experimentally measured thermoelectric figures of merit for approximately 80 materials[6]. For comparison, the correlation between the DFT and the experimental power factors from the same dataset is 0.31. The power factor score (red) is defined as $\frac{\mathrm{PF}_{\mathrm{pred10}}-\mathrm{PF}_{\mathrm{mean}}}{\mathrm{PF}_{\mathrm{best10}}-\mathrm{PF}_{\mathrm{mean}}}$, where $\mathrm{PF}_{\mathrm{mean}}$ is the average power factor of all candidates, $\mathrm{PF}_{\mathrm{pred10}}$ is the average power factor of the first 10 predictions and $\mathrm{PF}_{\mathrm{best10}}$ is the average of the 10 highest power factors. The default algorithm uses the original hyperparameters of the Word2vec code[7]. The CBOW and Skip-gram use optimized hyperparameters with or without the common phrases. The GloVe model uses the recommended hyperparameters from the original paper[1]. We found that hyper-parameter tuning changed the analogy scores for GloVe by less than a percent, however, we did not perform an extensive optimization similar to Word2vec. **b.** Evolution of the scores in a. for the "Phrases + Skip-gram" model over 30 training epochs. The learning rate decreases linearly from 10^{-2} to 10^{-4}.

of merit[6]. Fig. S2a shows the scores after 30 training epochs[†] for different models and parameters. Similar to the analogy scores, we see that Skip-gram performs better than CBOW, and that the inclusion of phrases results in performance gains. Additionally, all of the Word2vec models outperform GloVe at ranking the thermoelectrics. We attribute this to the predictive nature of Word2vec and the use of output embeddings for ranking and predictions (see the next section). GloVe is count-based and does not provide an additional set of output embeddings. Fig. S2b shows these evaluation metrics as functions of training epochs for the Skip-gram model with phrases. Until after 5 training epochs the predictions are not better than a random guess (power factor score of 0). The scores begin to improve following this initialization, and a substantial gain is made during the last few epochs of fine-tuning the embeddings. A similar trend is seen for all the metrics.

S3 Word versus output embeddings for predictions

The ranking (and consequently predictions) are performed by multiplying the embedding of the application keyword (e.g. "thermoelectric") with the embeddings of all materials (with some count threshold, more than 3 in our case). For the application keyword we always use the normalized word embedding. However, for the materials we attempt to use either the word or the output embedding (fig. S1b). If we use word embeddings, the

[†]An epoch corresponds to a single full pass over the corpus.

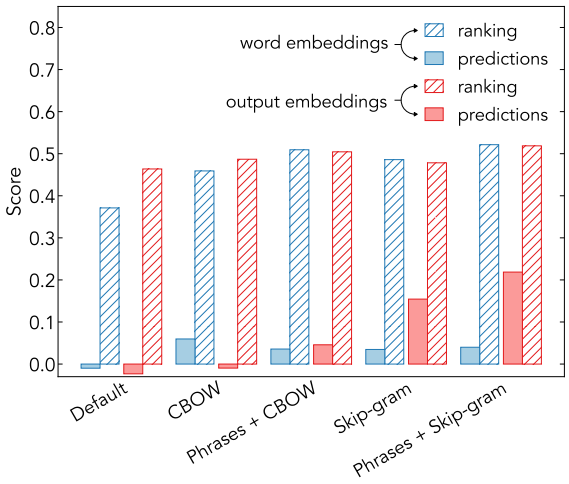


Figure S3: Word vs output embeddings.. Word embeddings corresponds to using word embeddings both for the application keyword and the material formula, whereas output embeddings corresponds to using the output embedding of the formula and the word embedding of the application keyword. The definitions of the scores are the same as in fig. S2a.

ranking is based on similarity of the application keyword and the material word. One can think of this as their interchangeability in text. If instead we use the normalized output embedding of the material, the predictions are based on the likelihood of the application keyword and the material formula being mentioned next to each other, if all materials were mentioned equal number of times in the text[‡]. This second approach generally yields better results as shown in fig. S3 and is used throughout this work.

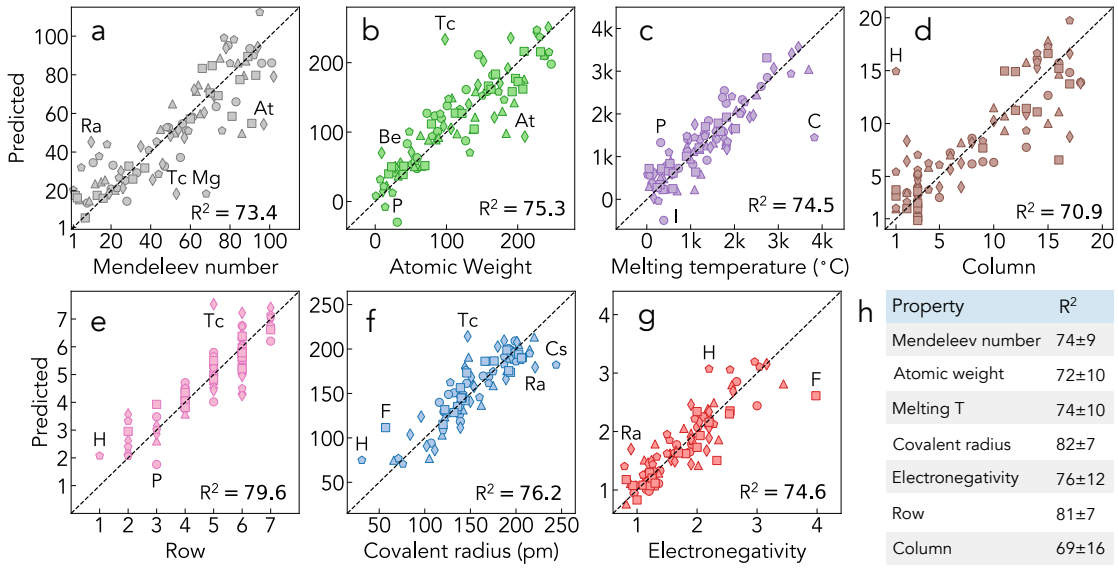
[‡]The norm of the embedding was shown to depend on the number of mentions - with more common words usually having longer embedding vectors[8].

S4 Word2vec element clustering versus periodic table

It is remarkable that using only relative positions of words in scientific text the algorithm learns a high dimensional representation for elements that is very similar to the periodic table when projected onto a plane. However, not all of the structure of our t-SNE projected word embeddings match well with the periodic table. Given that this is a context-based representation, it is unsurprising that the inert noble gases are far removed from the rest of the elements whereas post-transition metals, metalloids, and alkali metals, which are often used with each other in various applications, group closer together. The astute reader may observe that hydrogen is clustered with oxygen, nitrogen, and carbon; we attribute this to the fact that these elements are the main components of organic compounds. Similarly, Radon (Rn), radium (Ra) and polonium (Po), all radioactive elements, are found in closer proximity to uranium (U) and thorium (Th) in the plot than to their neighbors in the periodic table. Some elements, nevertheless, are completely out of place compared to the periodic table for what we believe to be non-physical reasons. We note that these elements' symbols overlap with common words that have the same spelling, such as "be" for beryllium, "at" for astatine or "Tc" for technetium which is also used to denote critical temperatures. Despite this, the high dimensionality of the embeddings enables relationships such as "being" - "Be" + "measure" $\approx$ "measuring" and "BeO" - "Be" + "Mg" $\approx$ "MgO" to be captured simultaneously, therefore, preserving both the chemical and the

syntactic relationships.

S5 Linear regression for elemental properties



h

Property	R^2
Mendeleev number	74±9
Atomic weight	72±10
Melting T	74±10
Covalent radius	82±7
Electronegativity	76±12
Row	81±7
Column	69±16

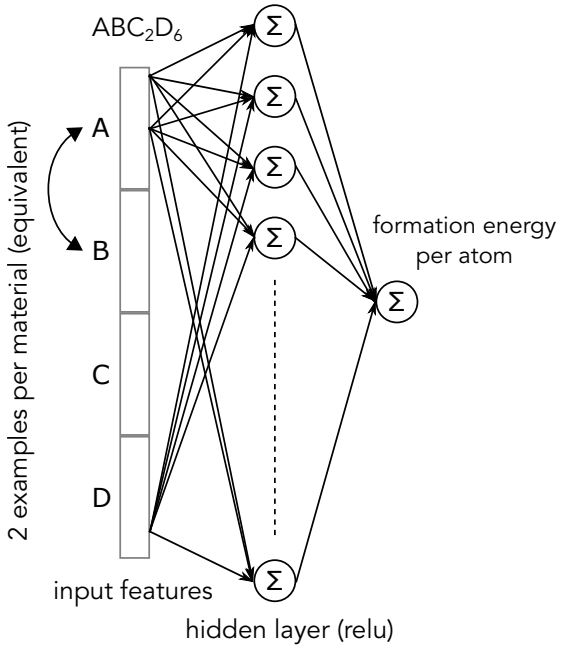
Figure S4: Predictions of Elemental Properties. a-g. 5-fold cross-validated predictions of 7 elemental properties using linear regression. The first 15 principal components of word embeddings of element names (e.g. "hydrogen") were used as features. The 5 different shapes indicate the exact cross-validation splitting, such that each shape (e.g. square) represents a set of validation elements predicted using the training elements represented by the 4 other shapes (e.g. triangles, diamonds, circles, pentagons). The splitting was determined randomly. **h.** Means and standard deviations of validation R^2 scores (in percent) from 20 random 80% (training) / 20% (validation) splits.

We can determine whether there exist directions in the embedding space that correlate with elemental properties by fitting a linear regression to predict each property using

embeddings as features. We test on the following 7 elemental properties: Mendeleev number, atomic weight, melting temperature, covalent radius, electronegativity, as well as row and column in the periodic table. Since there are 200 features but only around 100 elements, even a model as simple as linear regression will overfit. To avoid this, we reduce the dimensionality to 15 by applying principal component analysis (PCA) to the normalized word embeddings. The new features are linear combinations of the original 200 and explain 65% of the total variance. Sample plots of predicted versus actual values using 5-fold cross-validation are shown in fig. S4a-g. The mean and standard deviations of R^2 for all tested properties are shown in fig. S4h. We do not perform model selection and there are no hyper-parameters to optimize, therefore, there is no need for a test set outside of the cross-validation.

S6 Formation energies of $\mathrm{ABC_2D_6}$ elpasolites

We were able to predict formation energies of elpasolites with mean absolute errors as low as $55.7\,\mathrm{meV/atom}$ using only word embeddings (both Word2vec and Glove were tested) of their constituent elements as features. We use a dataset with approximately 10,000 $\mathrm{ABC_2D_6}$ materials available from reference [9]. We use one of the simplest neural network architectures - a single fully connected hidden layer with ReLU (rectified linear unit) activation and a single output neuron (fig. S5a) - the same as reference [10]. For



a

Features	Hidden neurons	MAE (meV/atom) validation	test
Word2vec embeddings	10	153.1±6.7	156.3
	50	94.1±3.1	87.6
	200	74.3±2.1	69.8
	800	66.4±1.7	62.0
	1600	67.0±4.6	59.7
	3200	67.5±3.6	55.7
GloVe embeddings	10	162.3±38.2	145.4
	50	86.4±4.7	85.8
	200	76.0±6.1	65.0
	800	65.0±1.9	64.0
	1600	64.9±3.8	59.6
	3200	63.3±3.6	56.4
Fundamental properties	10	373.3±24.3	337.6
	50	304.3±8.9	297.0
	200	280.4±10.2	274.0
	800	268.5±3.8	254.3
	1600	254.2±10.0	288.7
	3200	267.3±9.6	255.5
One-hot encoding	10	646.3±10.8	606.0
	50	818.7±16.8	760.4
	200	864.4±15.3	836.4
	800	825.3±19.6	795.8
	1600	808.7±12.8	775.1
	3200	790.9±18.1	761.2

b

Figure S5: Formation energies of ABC_2D_6 elpasolites. a. The architecture of the neural network used for predictions. **b.** Validation and test scores for 4 different feature choices as well as different hidden layer sizes. The performance for word embeddings does not improve much above 800 hidden neurons.

the input we concatenate embeddings of A, B, C and D elements, and also augment the data by creating 2 training examples for each material because A and B are equivalent. It is important to perform this data augmentation after splitting the data into training, validation and test sets to make sure every distinct material occurs only in one of the

sets. The mean absolute error on the test set of the best performing model decreases from $69.2\,\mathrm{meV/atom}$ to $55.7\,\mathrm{meV/atom}$ if we use this augmentation scheme. We also test alternative feature vectors with the same neural network architecture, such as one-hot encoding of elements and min-max scaled (all feature values between 0 and 1) vectors composed of the 7 elemental properties from the previous section. The performances for different features as well as different sizes of the hidden layer are summarized in fig. S5b, with word embeddings clearly performing the best. The displayed validation scores are the mean absolute errors (MAE) for 5-fold cross-validation. The test score is reported for a 10% test set separated before the training.

S7 Zero versus non-zero band gap classification

There are 1544 materials in our text corpus that have experimental band gaps in reference [11], with 603 materials having zero band gap and 941 materials having a non-zero band gap. Using 200-dimensional word embeddings of materials normalized to unit length as features, we trained a support vector machines (SVM) classifier with radial basis function (RBF) kernel to differentiate between zero vs non-zero band gap materials. Hyper-parameter optimization for parameters C (regularization) and γ (inverse of the standard deviation of the kernel) was performed using grid search. An average f1-score over 20 random train / validation splits of 80% / 20% was used for scoring. The highest f1-score

of $90.8 \pm 1.0\%$ was obtained for $\gamma = 2.34$ and $\mathrm{C} = 1.83$. In fig. S6a we plot a confusion matrix corresponding to a single 5-fold cross-validation applied to a re-shuffled (test) dataset using the optimal hyper-parameters. In fig. S6b we plot the distribution of the decision functions for these predictions, showing a good separation.

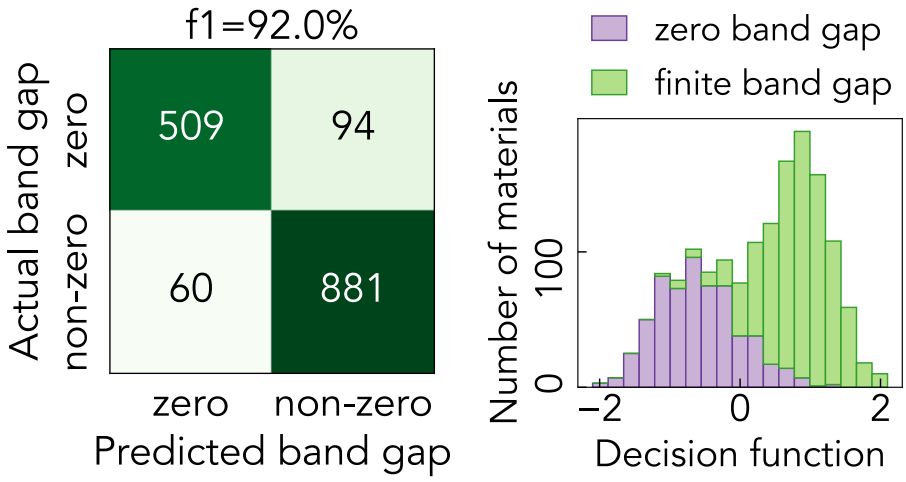


Figure S6: Prediction of zero vs non-zero band gaps. a. The confusion matrix of 5-fold cross-validation using hyper-parameters optimized on 20 other randomized train / validation splits. **b.** Distribution of decision function values of the 5-fold cross-validation shown in a. Values below 0 are classified as zero band gap, whereas above 0 as non-zero band gap.

S8 Material maps

Similar to chemical elements, one can visualize word embeddings of material formulas in 2D as shown in fig. S7a. We highlight a few large clusters using an unsupervised clustering

algorithm called DBSCAN[12]. We also mark the most "connected" material within each cluster using PageRank[13], an algorithm often used by search engines to rank web pages. Implementation details of DBSCAN and PageRank are discussed in the next section. If we zoom into the cluster with PbTe (fig. S7b), we see that these are all thermoelectric chalcogenides. Similarly, the cluster with $\mathrm{LiFePO_4}$ contains mostly lithium-ion battery materials, the cluster with CdS is made up of materials used predominantly for solar cells, etc. We summarize the common elements in each cluster in fig. S7c and find that these elements correspond to those typically used within a particular functional application.

In addition to groups of similar materials, the long range order is also meaningful. Applications like thermoelectrics and photovoltaics merge into one another since both typically involve intermediate band semiconductors that need to be highly doped. Fig. S7c illustrates that the cluster marked as photovoltaics is composed mostly of sulfides and selenides - chalcogenides also used as thermoelectrics. Interestingly, the cluster with III-V semiconductors containing GaAs that can also be used for photovoltaics is far from CdS (II-VI semiconductor), since is not only the application that determines the position on the map but also the similarity of chemical compositions. This can be directly encoded when the name of the material is mentioned next to the formula in text, for example "gallium arsenide" next to "GaAs". It turns out that many materials at the top of the map are oxides, bottom right are metallic, the ones stretching from the center to bottom are semiconductors

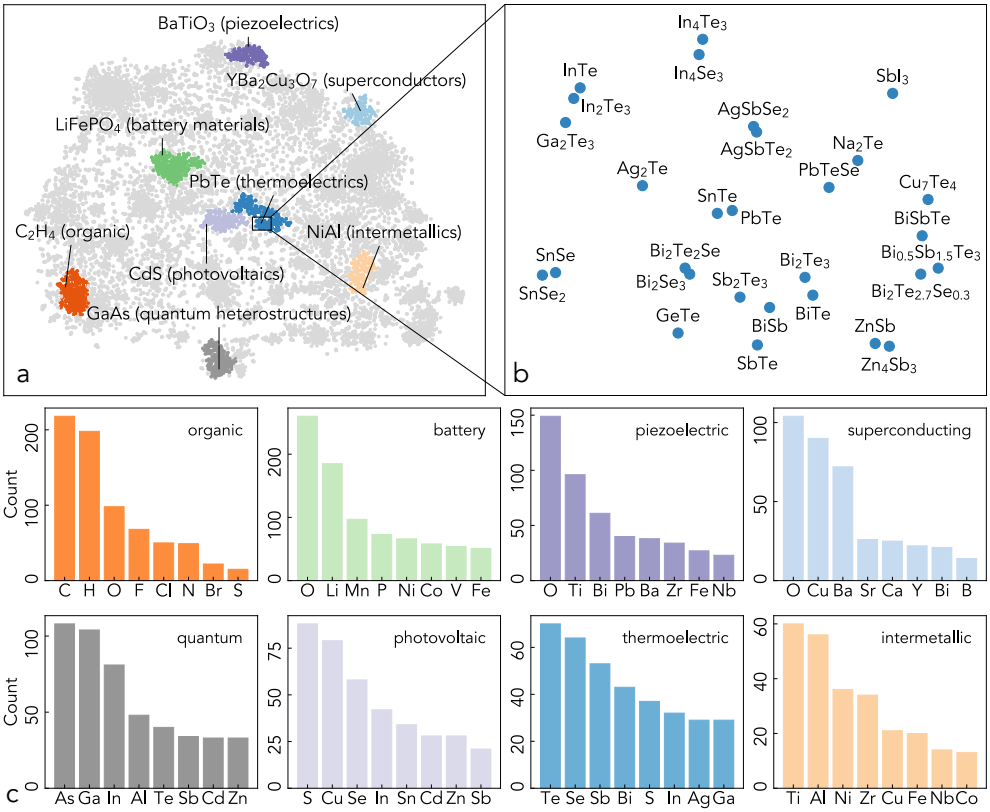


Figure S7: Material maps. a. t-SNE projection of 12,340 word embeddings corresponding to materials mentioned at least 10 times in the corpus. Each point represents a unique stoichiometry. The relative distance of materials can be interpreted as their context-based similarity. The materials are clustered in an unsupervised manner using DBSCAN[12], which groups together high density areas. The labeled material in each cluster corresponds to the "most connected" material within that cluster. This is determined using PageRank[13] within each cluster, with weights corresponding to cosine similarities of word embeddings. An interactive version of the map can be found at reference [14]. **b.** A region of the map in a. in the vicinity of PbTe – one of the most common thermoelectric materials. **c.** Counts of the eight most common elements from each cluster in a., counted one per material independent on their stoichiometric ratios.

whereas the ones on the bottom left are organic. In fact, using only word embeddings of materials as features without any explicit knowledge of the compositions, we can predict if a material has a band-gap with 90.8% accuracy (f1-score), similar to a reported 91.4% score using a composition-based representation[11] (see Supplementary Information for the details).

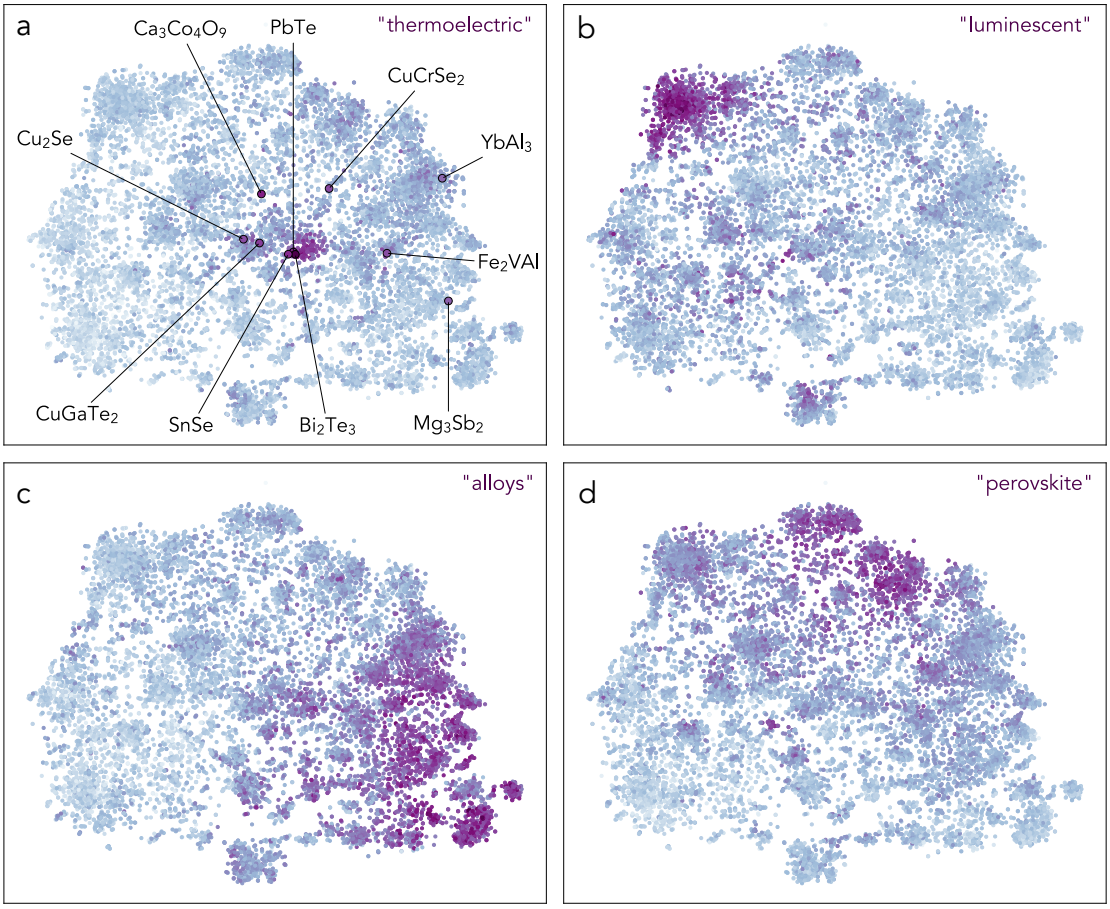


Figure S8: Dynamic material maps. Material maps highlighted according to various keywords. Darker colors correspond to more similarity.

We can create dynamic visualizations of material / keywords similarities by coloring each material on a 2D map according to its cosine similarity to that keyword - be that an application word (e.g. "thermoelectric"), a class of materials (e.g. "alloys") or a crystal structure (e.g. "perovskite"). As an example, in fig. S8a we show the map highlighted according to the word "thermoelectric" - with darker colors corresponding to higher similarity. There are many types of thermoelectrics, hence, one should not expect all of them to cluster together. Some materials are used for other applications in other contexts, so they are further away from the main cluster containing conventional thermoelectrics such as Bi_2Te_3 and $PbTe$. $SnSe$, a recently discovered thermoelectric with a record power factor in 2014[15] is also in this cluster. $CuGaTe_2$ is a well-known semiconductor also considered as a promising candidate for thin film solar cells[16]. Mg_3Sb_2 is a thermoelectric with Zintl structure[17], Cu_2Se is an recently discovered ion-liquid like thermoelectric[18], $Ca_3Co_4O_9$ is an oxide thermoelectric[19], Fe_2VAl is a heusler-type nonmagnetic semimetal[20], $CuCrSe_2$ is a layered antiferromagnet and a superionic conductor[21] whereas $YbAl_3$ is an intermetallic compound with a record power factor[22]. More extensive reviews of different types of thermoelectrics can be found at references [23] and [24]. Examples for a few other keywords are plotted in fig. S8b-d.

S9 Projection, clustering and ranking of materials embeddings.

Projection. In fig. S7a, the 200-dimensional embeddings of the 12,340 materials mentioned more than 10 times in our corpus were reduced to 2 dimensions using t-SNE[25,26]. We used cosine distance between the embeddings as a metric, perplexity 30, learning rate 200, early exaggeration 12.0 and 10,000 iterations - with coordinates initialized using PCA.

Clustering. To group high density areas of the 2D projection for easier visualization, we used an unsupervised clustering technique called DBSCAN[12,26]. We used neighbour distance cutoff $\epsilon = 2.75$ and a minimum count of 8, producing well separated clusters. Clusters with less than 120 materials were ignored. For the final visualization, we chose 8 clusters from the remaining 18.

Ranking. To find a representative material within each cluster, we use the implementation of PageRank[13] available via igraph[27] software package. Each node of the undirected graph corresponds to a material, with the weights of the edges corresponding to cosine similarities between the materials. We used the default damping value of 0.85 to compute the ranks within each cluster, with the highest ranked materials labelled in fig. S7a. Globally, the five most connected materials in our corpus (excluding chemical elements) were $\mathrm{TiO_2}$,

ZnO, SiO_2, Al_2O_3 and SiC, which are all used for a large spectrum of applications.

S10 Unconventional thermoelectric predictions

In addition to well known thermoelectric material classes, we observe predictions such as KAg_2SbS_4 (see Table 2 of extended data) that do not have strong similarity to known thermoelectrics. This particular compound has recently been suggested as a candidate photovoltaic material[28]. Another example is $BiOCl$, an atypical oxychloride which was the top prediction in the 2010 historical corpus (Supplementary Table S5) and has a computed p-type power factor of $25.4\,\mu\mathrm{W/K^2 \cdot cm}$ (calculated using the constraints described in the Methods section of the main text) – ranking in the 93rd percentile of our dataset's power factors. Potential issues with the oxychloride chemistry might be large band gaps and dopability. However, this material contains desirable band structure features, including two doubly-degenerate valence band peaks aligned at the off-symmetry points X and R[29] that are responsible for the high computed power factor. $ZnSiP_2$ was #3 in 2005 and has a similarly high p-type power factor of $33.2\,\mu\mathrm{W/K^2 \cdot cm}$ (95th percentile) and n-type power factor of $29.5\,\mu\mathrm{W/K^2 \cdot cm}$ (96th percentile among n-types power factors). Phosphides are typically thought to have high thermal conductivities not desirable for thermoelectric applications, however, this is not strictly true[30] and it is unclear if that is the case for this compound. This material is also interesting due to its band structure features, which include

a triply degenerate valence band peak and doubly-degenerate conduction band pocket at the gamma point[31], as well as a degenerate conduction band pocket between Z and Σ_1. Another notable example is $\mathrm{Nd_{0.5}Sr_{0.5}MnO_3}$ (#5 in 2004). The computed power factor for this compound is missing from our dataset since it exhibits site disorder that is more difficult to model with density functional theory methods. However, it is known experimentally to have a relatively low thermal conductivity ($< 3\ \mathrm{W/K \cdot m}$ at 300 K)[32], high dopability, and high electrical conductivity ($> 300\ \mathrm{S/cm}$ at 300 K)[33] – all promising indicators for a high zT material. Each of these examples may require additional synthesis and optimization work to overcome potential limitations in doping and thermal conductivity due to their unconventional chemistries, nevertheless, they are viable thermoelectric candidates that are not closely related to any mainstream thermoelectrics.

Year	Top 10 thermoelectric predictions	Total potential predictions	Total abstracts in corpus
2001	HgMnTe, HgZnTe, $EuLiH_3$, $CdGeP_2$, $La_{0.5}Sr_{0.5}MnO_3$, VB_2, $CoCr_2S_4$, CdSeTe, $Bi_2Sr_2CuO_6$, $AgInS_2$	13221	288178
2002	Mo_3Te_4, HgMnTe, ZrB_2, $ZrSi_2$, $La_{0.5}Sr_{0.5}MnO_3$, Mo_5Si_3, $Ge_{22}Se_{78}$, TmSb, BaLaCuO, $Nd_{0.5}Sr_{0.5}MnO_3$	14181	331414
2003	EuB_6, $CdGeP_2$, HgMnTe, $ReSe_2$, $Cd_{0.8}Zn_{0.2}Te$, Yb_4As_3, HgZnTe, ReS_2, $CoCr_2S_4$, CuNb	15042	375079
2004	HgMnTe, V_2Ga_5, HgZnTe, Yb_4As_3, $Nd_{0.5}Sr_{0.5}MnO_3$, CoS_2, EuB_6, $CdGeP_2$, ReS_2, $Ge_{22}Se_{78}$	15906	422439
2005	V_2Ga_5, $BaSi_2$, $ZnSiP_2$, HgZnTe, HgMnTe, $CoCr_2S_4$, EuB_6, Sb_2O_5, ReS_2, SbSI	16824	473567

2006	ReS_2, $BaSi_2$, TiSi, $SmInO_3$, $ReSe_2$, $Na_{0.9}Mo_6O_{17}$, HgMnTe, HgZnTe, $CeOs_4Sb_{12}$, $CoCr_2S_4$	17595	523433
2007	ReS_2, HgZnTe, $BaSi_2$, $SmInO_3$, $ReSe_2$, EuB_6, LaOAgS, $CeOs_4Sb_{12}$, CdP_2, Sn_4P_3	18510	580323
2008	ReS_2, HgZnTe, $ReSe_2$, SbSI, GeI_2, $SmInO_3$, $FeIn_2Se_4$, Yb_4As_3, $TeCl_4$, $CdIn_2Te_4$	19320	639825
2009	HgZnTe, ReS_2, $SmInO_3$, $CdIn_2Te_4$, $CuGaTe_2$, $ReSe_2$, HgMnTe, $TlSbSe_2$, Co_2FeGa, $(YbS)_{1.25}CrS_2$	20177	702186
2010	BiOCl, HgZnTe, Co_2FeGa, $CdIn_2Te_4$, HgMnTe, $(YbS)_{1.25}CrS_2$, $La_{0.9}Sr_{0.1}MnO_3$, ReS_2, $NiTe_2$, NiP_3	21037	766690
2011	$SmInO_3$, $CdIn_2Te_4$, $(YbS)_{1.25}CrS_2$, $FeIn_2Se_4$, HgZnTe, NiP_3, $CdGa_2O_4$, $AgInSe_2$, $La_{0.9}Sr_{0.1}MnO_3$, ZrNCl	21679	831227
2012	$FeIn_2Se_4$, $(YbS)_{1.25}CrS_2$, HgZnTe, $CdGa_2O_4$, YbTe, $TlCrS_2$, $SmInO_3$, $HoCu_2$, $CdIn_2Te_4$, $SrNb_{0.01}Ti_{0.99}O_3$	22446	904141

2013	HgZnTe, $Zn_{0.7}Cd_{0.3}Se$, $CaYb_2S_4$, $CdIn_2Te_4$, $FeIn_2Se_4$, $CuSbS_2$, $CdGa_2O_4$, $TeCl_4$, CeTe, $(YbS)_{1.25}CrS_2$	23179	979040
2014	$TlCrS_2$, YbTe, HgZnTe, $FeIn_2Se_4$, $SnSb_2Te_4$, $La_{0.7}Ca_{0.2}Sr_{0.1}MnO_3$, $CaYb_2S_4$, $HoCu_2$, $Bi_{0.95}La_{0.05}FeO_3$, $CdSnO_3$	24029	1067395
2015	$TlCrS_2$, YbTe, $FeIn_2Se_4$, ReS_2, $Zn_{0.7}Cd_{0.3}Se$, Co_2FeGa, $ReSe_2$, NiP_3, CoCrFeNi, $LiGaSe_2$	24749	1159529
2016	YbTe, $TeCl_4$, $FeIn_2Se_4$, $TlCrS_2$, $La_{0.7}Ca_{0.2}Sr_{0.1}MnO_3$, $Hf_{0.2}Zr_{0.8}O_2$, $CdSnP_2$, MoSe, $Bi_{0.95}La_{0.05}FeO_3$, $Pb_{0.902}Sn_{0.098}Se$	25469	1257788
2017	YbTe, $CdSnP_2$, In_3Se_2, $Hf_{0.2}Zr_{0.8}O_2$, $InFeZnO_4$, $TlSbSe_2$, Sc_2CF_2, HgZnTe, Ag_3AuSe_2, $TlCrS_2$	26184	1358468
2018	YbTe, In_3Se_2, $ZnSnP_2$, HgZnTe, $TlSbSe_2$, $CdSnP_2$, CuTe, $TlCu_2Se_2$, SbSI, MoSe	26804	1470230

Table S3: Top 10 thermoelectric predictions from each year. The list of materials is ordered from prediction #1 to prediction #10. Total candidates is the number of materials considered for the prediction, which includes all materials mentioned more than 3 times but not studied as thermoelectric before. Total abstracts is the number of (relevant) abstracts used to train the word embeddings.

1. Pennington, J., Socher, R. & Manning, C. Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 1532–1543 (2014).

2. Mikolov, T., Corrado, G., Chen, K. & Dean, J. Efficient Estimation of Word Representations in Vector Space. In *International Conference on Learning Representations* (2013).

3. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In *International Conference on Neural Information Processing Systems*, 3111–3119 (2013).

4. `http://www.materialsintelligence.com/materials-analogies`.

5. Spearman, C. The proof and measurement of association between two things. *The American Journal of Psychology* **15**, 72–101 (1904).

6. Gaultois, M. W. *et al.* Data-driven review of thermoelectric materials: Performance and resource onsiderations. *Chemistry of Materials* **25**, 2911–2920 (2013).

7. `https://code.google.com/archive/p/word2vec/`.

8. Schakel, A. M. & Wilson, B. J. Measuring word significance using distributed representations of words. *arXiv preprint arXiv:1508.02297* (2015).

9. Faber, F. A., Lindmaa, A., Von Lilienfeld, O. A. & Armiento, R. Machine Learning Energies of 2 Million Elpasolite (ABC2D6) Crystals. *Physical Review Letters* **117**, 2–7 (2016). 1508.05315.

10. Zhou, Q. *et al.* Learning atoms for materials discovery. *Proceedings of the National Academy of Sciences* **115**, E6411–E6417 (2018).

11. Zhuo, Y., Mansouri Tehrani, A. & Brgoch, J. Predicting the Band Gaps of Inorganic Solids by Machine Learning. *Journal of Physical Chemistry Letters* **9**, 1668–1673 (2018).

12. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD'96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–231 (1996).

13. Brin, S. & Page, L. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems* **30**, 107 – 117 (1998). Proceedings of the Seventh International World Wide Web Conference.

14. http://www.materialsintelligence.com/materials-map.

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

16. Plirdpring, T. *et al.* Chalcopyrite CuGaTe2: A high-efficiency bulk thermoelectric material. *Advanced Materials* **24**, 3622–3626 (2012).

17. Condron, C. L., Kauzlarich, S. M., Gascoin, F. & Snyder, G. J. Thermoelectric properties and microstructure of Mg3Sb2. *Journal of Solid State Chemistry* **179**, 2252–2257 (2006).

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

19. Funahashi, R. *et al.* An oxide single crystal with high thermoelectric performance in air. *Japanese Journal of Applied Physics* **39**, L1127–L1129 (2000).

20. Xu, B. *et al.* The structural, elastic and thermoelectric properties of Fe2VAl at pressures. *Journal of Alloys and Compounds* **565**, 22–28 (2013).

21. Bhattacharya, S. *et al.* CuCrSe2: A high performance phonon glass and electron crystal thermoelectric material. *Journal of Materials Chemistry A* **1**, 11289–11294 (2013).

22. Liang, J., Fan, D., Jiang, P., Liu, H. & Zhao, W. First-principles study of the thermoelectric properties of intermetallic compound YbAl3. *Intermetallics* **87**, 27–30 (2017).

23. Liu, W. *et al.* New trends, strategies and opportunities in thermoelectric materials: A perspective. *Materials Today Physics* **1**, 50–60 (2017).

24. He, J. & Tritt, T. M. Advances in thermoelectric materials research: Looking back and moving forward. *Science* **357** (2017).

25. van der Maaten, L. & Hinton, G. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).

26. `http://scikit-learn.org/stable/`.

27. `http://igraph.org`.

28. Nhalil, H. *et al.* Optoelectronic properties of candidate photovoltaic cu2pbsis4, ag2pbges4 and kag2sbs4 semiconductors. *Journal of Alloys and Compounds* **746**, 405–412 (2018).

29. `https://www.materialsproject.org/materials/mp-22939/`.

30. Pöhls, J.-H. *et al.* Metal phosphides as potential thermoelectric materials. *Journal of Materials Chemistry C* **5**, 12441–12456 (2017).

31. `https://www.materialsproject.org/materials/mp-4763/`.

32. Fujishiro, H., Sugawara, S. & Ikebe, M. Anomalous phonon transport enhancement at first-order ferromagnetic transition in (gd, sm, nd) 0.55 sr0. 45mno3. *Physica B: Condensed Matter* **316**, 331–334 (2002).

33. Cui, C., Tyson, T. A., Chen, Z. & Zhong, Z. Transport and structural study of pressure-induced magnetic states in nd 0.55 sr 0.45 mno 3 and nd 0.5 sr 0.5 mno 3. *Physical Review B* **68**, 214417 (2003).