Title

Unsupervised word embeddings capture latent knowledge from materials science literature

Journal

Nature, 571(7763)

ISSN

0028-0836

Authors

Tshitoyan, Vahe
Dagdelen, John
Weston, Leigh
et al.

Publication Date

2019-07-01

DOI

10.1038/s41586-019-1335-8

Peer reviewed

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]*

The overwhelming majority of scientific knowledge is published as text, which is difficult to analyse by either traditional statistical analysis or modern machine learning methods. By contrast, the main source of machine-interpretable data for the materials research community has come from structured property databases[1,2], which encompass only a small fraction of the knowledge present in the research literature. Beyond property values, publications contain valuable knowledge regarding the connections and relationships between data items as interpreted by the authors. To improve the identification and use of this knowledge, several studies have focused on the retrieval of information from scientific literature using supervised natural language processing[3–10], which requires large hand-labelled datasets for training. Here we show that materials science knowledge present in the published literature can be efficiently encoded as information-dense word embeddings[11–13] (vector representations of words) without human labelling or supervision. Without any explicit insertion of chemical knowledge, these embeddings capture complex materials science concepts such as the underlying structure of the periodic table and structure–property relationships in materials. Furthermore, we demonstrate that an unsupervised method can recommend materials for functional applications several years before their discovery. This suggests that latent knowledge regarding future discoveries is to a large extent embedded in past publications. Our findings highlight the possibility of extracting knowledge and relationships from the massive body of scientific literature in a collective manner, and point towards a generalized approach to the mining of scientific literature.

Assignment of high-dimensional vectors (embeddings) to words in a text corpus in a way that preserves their syntactic and semantic relationships is one of the most fundamental techniques in natural language processing (NLP). Word embeddings are usually constructed using machine learning algorithms such as GloVe[13] or Word2vec[11,12], which use information about the co-occurrences of words in a text corpus. For example, when trained on a suitable body of text, such methods should produce a vector representing the word 'iron' that is closer by cosine distance to the vector for 'steel' than to the vector for 'organic'. To train the embeddings, we collected and processed approximately 3.3 million scientific abstracts published between 1922 and 2018 in more than 1,000 journals deemed likely to contain materials-related research, resulting in a vocabulary of approximately 500,000 words. We then applied the skip-gram variation of Word2vec, which is trained to predict context words that appear in the proximity of the target word as a means to learn the 200-dimensional embedding of that target word, to our text corpus (Fig. 1a). The key idea is that, because words with similar meanings often appear in similar contexts, the corresponding embeddings will also be similar. More details about the model are included in the Methods and in Supplementary Information sections S1 and S2, where we also discuss alternative algorithm options such as GloVe. We find that, even though no chemical information or interpretation is added to the algorithm, the obtained word embeddings behave consistently with chemical intuition when they are combined using various vector operations (projection, addition, subtraction). For example, many words in our corpus represent chemical compositions of materials, and the five materials most similar to $LiCoO_2$ (a well-known lithium-ion cathode compound) can be determined through a dot product (projection) of normalized word embeddings. According to our model, the compositions with the highest similarity to $LiCoO_2$ are $LiMn_2O_4$, $LiNi_{0.5}Mn_{1.5}O_4$, $LiNi_{0.8}Co_{0.2}O_2$, $LiNi_{0.8}Co_{0.15}Al_{0.05}O_2$ and $LiNiO_2$—all of which are also lithium-ion cathode materials.

Similar to the observation made in the original Word2vec paper[11], these embeddings also support analogies, which in our case can be domain-specific. For instance, 'NiFe' is to 'ferromagnetic' as 'IrMn' is to '?', where the most appropriate response is 'antiferromagnetic'. Such analogies are expressed and solved in the Word2vec model by finding the nearest word to the result of subtraction and addition operations between the embeddings. Hence, in our model,

$$\text{ferromagnetic} - \text{NiFe} + \text{IrMn} \approx \text{antiferromagnetic}$$

To better visualize such embedded relationships, we projected the embeddings of Zr, Cr and Ni, as well as their corresponding oxides and crystal structures, onto two dimensions using principal component analysis (Fig. 1b). Even in reduced dimensions, there is a consistent operation in vector space for the concepts 'oxide of' ($Zr - ZrO_2 \approx Cr - Cr_2O_3 \approx Ni - NiO$) and 'structure of' ($Zr - HCP \approx Cr - BCC \approx Ni - FCC$). This suggests that the positions of the embeddings in space encode materials science knowledge such as the fact that zirconium has a hexagonal close packed (HCP) crystal structure under standard conditions and that its principal oxide is ZrO_2. Other types of materials analogies captured by the model, such as functional applications and crystal symmetries, are listed in Extended Data Table 1. The accuracies for each category are close to 50%—similar to the baseline set in the original Word2vec study[12]. We stress that Word2vec treats these entities simply as strings, and no chemical interpretation is explicitly provided to the model; rather, materials knowledge is captured through the positions of the words in scientific abstracts. Notably, we also found that embeddings of chemical elements are representative of their positions in the periodic table when projected onto two dimensions (Extended Data Fig. 1a, b, Supplementary Information sections S4 and S5) and can serve as effective feature vectors in quantitative machine learning models such as formation energy prediction—outperforming several previously reported curated feature vectors (Extended Data Fig. 1c, d, Supplementary Information section S6).

The main advantage and novelty of this representation, however, is that application keywords such as 'thermoelectric' have the same representation as material formulae such as 'Bi_2Te_3'. When the cosine similarity of a material embedding and the embedding of 'thermoelectric' is high, one might expect that the text corpus necessarily includes abstracts reporting on the thermoelectric behaviour of this material[14,15]. However, we found that a number of materials that have relatively high cosine similarities to the word 'thermoelectric' never

[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

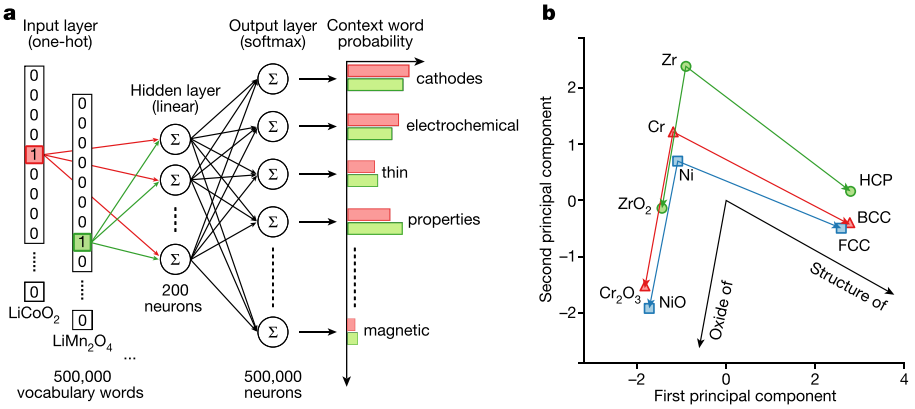


Fig. 1 | Word2vec skip-gram and analogies. a, Target words '$LiCoO_2$' and '$LiMn_2O_4$' are represented as vectors with ones at their corresponding vocabulary indices (for example, 5 and 8 in the schematic) and zeros everywhere else (one-hot encoding). These one-hot encoded vectors are used as inputs for a neural network with a single linear hidden layer (for example, 200 neurons), which is trained to predict all words mentioned within a certain distance (context words) from the given target word. For similar battery cathode materials such as $LiCoO_2$ and $LiMn_2O_4$, the context words that occur in the text are mostly the same (for example, 'cathodes', 'electrochemical', and so on), which leads to similar hidden layer weights after the training is complete. These hidden layer weights are the actual word embeddings. The softmax function is used at the output layer to normalize the probabilities. **b**, Word embeddings for Zr, Cr and Ni, their principal oxides and crystal symmetries (at standard conditions) projected onto two dimensions using principal component analysis and represented as points in space. The relative positioning of the words encodes materials science relationships, such that there exist consistent vector operations between words that represent concepts such as 'oxide of' and 'structure of'.

appeared explicitly in the same abstract with this word, or any other words that unequivocally identify materials as thermoelectric (Fig. 2a). Rather than dismissing these instances as spurious, we investigated whether such cases could be usefully interpreted as predictions of novel thermoelectric materials.

As a first test, we compared our predicted thermoelectric compositions with available computational data. Specifically, we identified compounds mentioned in our text corpus more than three times that are also present in a dataset[16] that reports the thermoelectric power factors (an important component of the overall thermoelectric figure of merit, zT) of approximately 48,000 compounds calculated using density functional theory (DFT)[17,18] (see Methods). A total of 9,483 compounds overlap between the two datasets, of which 7,663 were never mentioned alongside thermoelectric keywords in our text corpus and can be considered candidates for prediction. To obtain the predictions, we ranked each of these 7,663 compounds by the dot product of their normalized output embedding with the word embedding of 'thermoelectric' (see Supplementary Information sections S1 and S3 regarding the use of output versus word embeddings). This ranking can be interpreted as the likelihood that that material will co-occur with the word 'thermoelectric' in a scientific abstract despite this never occurring explicitly in the text corpus. The distributions of DFT maximum power factor values for all 9,483 materials (separated into known thermoelectrics and candidates) are plotted in Fig. 2b, and the values of the 10 highest ranked candidates from the word embedding approach are indicated with dashed lines. We find that the top ten predictions all exhibit computed power factors significantly greater than the average of candidate materials (green), and even slightly higher than the average of known thermoelectrics (purple). The average maximum power factor of 40.8 $\mu W\,K^{-2}\,cm^{-1}$ for these top ten predictions is 3.6 times larger than the average of candidate materials (11.5 $\mu W\,K^{-2}\,cm^{-1}$) and 2.4 times larger than the average of known thermoelectrics (17.0 $\mu W\,K^{-2}\,cm^{-1}$). Moreover, the three highest power factors from the top ten predictions are at the 99.6th, 96.5th and 95.3rd percentiles of known thermoelectrics. We note that in contrast to supervised methods, our embeddings are based only on the text corpus and are not trained or modified in any manner using the DFT data.

Next, we compared the same model directly against experimentally measured power factors and zTs[19]. Because our approach does not provide numerical estimations of these quantities, we compared the relative ranking of candidates through the Spearman rank correlation[20] for the 83 materials that appear both in our text corpus and the experimental set. We obtained a 59% and 52% rank correlation of experimental results with the embedding-based ranking for maximum power factor and maximum zT, respectively. Unexpectedly, our model outperformed the DFT dataset of power factors used in the previous paragraph, which exhibits only a 31% rank correlation with the experimental maximum power factors.

Finally, we tested whether our model—if trained at various points in the past—would have correctly predicted thermoelectric materials reported later in the literature. Specifically, we generated 18 different 'historical' text corpora consisting only of abstracts published before cutoff years between 2001 and 2018. We trained separate word embeddings for each historical dataset, and used these embeddings to predict the top 50 thermoelectrics that were likely to be reported in future (test) years. For every year past the date of prediction, we tabulated the cumulative percentage of predicted thermoelectric compositions that were reported in the literature alongside a thermoelectric keyword. Figure 3a depicts the result from each such 'historical' dataset as a thin grey line. For example, the light grey line labelled '2015' depicts the percentage of the top 50 predictions made using the model trained only on scientific abstracts published before 1 January 2015, and that were subsequently reported in the literature alongside a thermoelectric keyword after one, two, three or four years (that is, the years 2015–2018). Overall, our results indicate that materials from the top 50 word embedding-based predictions (red line) were on average eight times more likely to have been studied as thermoelectrics within the next five years as compared to a randomly chosen unstudied material from our corpus at that time (blue) and also three times more likely than a random material with a non-zero DFT bandgap (green). The use of larger corpora that incorporate data from more recent years improved the rate of successful predictions, as indicated by the steeper gradients for later years in Fig. 3a.

To examine these results in more detail, we focus on the fate of the top five predictions determined using only abstracts published before the year 2009. Figure 3b plots the evolution of the prediction rank of these top five compounds as more abstracts are added in subsequent years. One of these compounds, $CuGaTe_2$, represents one of the best present-day thermoelectrics and would have been predicted as a top five compound four years before its publication in 2012[21]. Two of the other predictions, ReS_2 and $CdIn_2Te_4$, were suggested in the literature to be good thermoelectrics[22,23] only approximately 8–9 years after the point at which they would have first appeared in the top five list from our algorithm. We note that the sharp increase in the rank of layered ReS_2 in 2015 coincides with the discovery of a record zT for

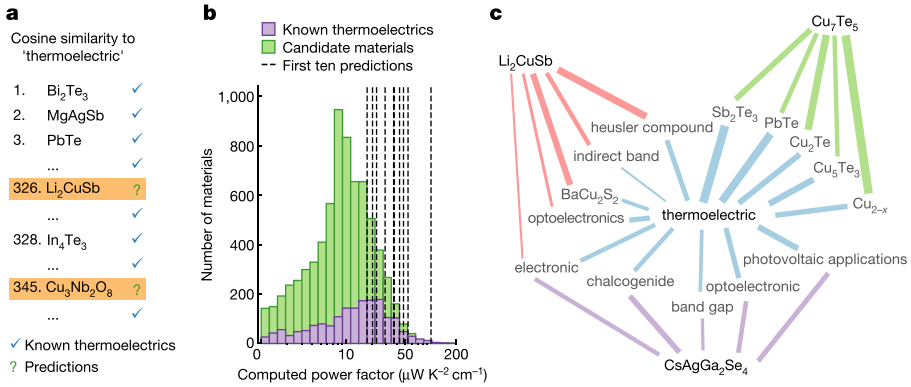


Fig. 2 | Prediction of new thermoelectric materials. a, A ranking of thermoelectric materials can be produced using cosine similarities of material embeddings with the embedding of the word 'thermoelectric'. Highly ranked materials that have not yet been studied for thermoelectric applications (do not appear in the same abstracts as words 'ZT', 'zT', 'seebeck', 'thermoelectric', 'thermoelectrics', 'thermoelectrical', 'thermoelectricity', 'thermoelectrically' or 'thermopower') are considered to be predictions that can be tested in the future. **b**, Distributions of the power factors computed using density functional theory (see Methods) for 1,820 known thermoelectrics in the literature (purple) and 7,663 candidate materials not yet studied as thermoelectric (green). Power factors of the first ten predictions not studied as thermoelectrics in our text corpus and for which computational data are available (Li_2CuSb, $CuBiS_2$, $CdIn_2Te_4$, $CsGeI_3$, $PdSe_2$, KAg_2SbS_4, $LuRhO_3$, MgB_2C_2, Li_3Sb and $TlSbSe_2$) are shown with black dashed lines. **c**, A graph showing how the context words of materials predicted to be thermoelectrics connect to the word thermoelectric. The width of the edges between 'thermoelectric' and the context words (blue) is proportional to the cosine similarity between the word embeddings of the nodes, whereas the width of the edges between the materials and the context words (red, green and purple) is proportional to the cosine similarity between the word embeddings of context words and the output embedding of the material. The materials are the first (Li_2CuSb), third ($CsAgGa_2Se_4$) and fourth (Cu_7Te_5) predictions. The context words are top context words according to the sum of the edge weights between the material and the word 'thermoelectric'. Wider paths are expected to make larger contributions to the predictions. Examination of the context words demonstrates that the algorithm is making predictions on the basis of crystal structure associations, co-mentions with other materials for the same application, associations between different applications, and key phrases that describe the material's known properties.

$SnSe$[24]—also a layered material. The final two predictions, HgZnTe and $SmInO_3$, contain expensive (Sm, In) or toxic (Hg) elements and have not been studied yet, and $SmInO_3$ has dropped appreciably in ranking with the addition of more data. The top 10 predictions for each year between 2001 and 2018 are available in Supplementary Table S3.

To illustrate how materials never mentioned next to the word 'thermoelectric' are identified as thermoelectrics with high expected probability, we investigated the series of connections that can lead to a prediction. In Fig. 2c, we present three materials from our top five predictions (Extended Data Table 2) alongside some of the key context words that connect these materials to 'thermoelectric'. For instance, $CsAgGa_2Se_4$ has high likelihood of appearing next to 'chalcogenide', 'band gap', 'optoelectronic' and 'photovoltaic applications': many good thermoelectrics are chalcogenides, the existence of a bandgap is crucial for the majority of thermoelectrics, and there is a large overlap between optoelectronic, photovoltaic and thermoelectric materials (see Supplementary Information section S8). Consequently, the correlations between these keywords and $CsAgGa_2Se_4$ led to the prediction. This direct interpretability is a major advantage over many other machine learning methods for materials discovery. We also note that several predictions were found to exhibit promising properties despite not being in any well known thermoelectric material classes (see Supplementary Information section S10). This demonstrates that word embeddings go beyond trivial compositional or structural similarity and have the potential to unlock latent knowledge not directly accessible to human scientists.

As a final step, we verified the generalizability of our approach by performing historical validation of predictions for three additional keywords—'photovoltaics', 'topological insulator' and 'ferroelectric'. We

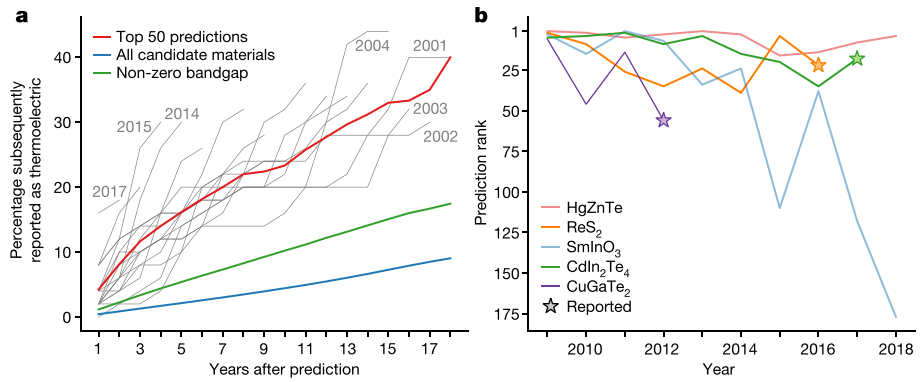


Fig. 3 | Validation of the predictions. a, Results of prediction of thermoelectric materials using word embeddings obtained from various historical datasets. Each grey line uses only abstracts published before that year to make predictions (for example, predictions for 2001 are performed using abstracts from 2000 and earlier). The lines plot the cumulative percentage of predicted materials subsequently reported as thermoelectrics in the years following their predictions; earlier predictions can be analysed over longer test periods, resulting in longer grey lines. The results are averaged (red) and compared to baseline percentages from either all materials (blue) or non-zero DFT bandgap[27] materials (green). **b**, The top five predictions from the year 2009 dataset, and evolution of their prediction ranks as more data are collected. The marker indicates the year of first published report of one of the initial top five predictions as a thermoelectric.

emphasize that the word embeddings used for these predictions are the same as those for thermoelectrics predictions; we have simply modified the dot product to be with a different target word. Notably, with almost no change in procedure, we find trends similar to the ones in Fig. 3a for all three functional applications, with the results summarized in Extended Data Fig. 2 and Extended Data Table 3.

The success of our unsupervised approach can partly be attributed to the choice of the training corpus. The main purpose of abstracts is to communicate information in a concise and straightforward manner, avoiding unnecessary words that may increase noise in embeddings during training. The importance of corpus selection is demonstrated in Extended Data Table 4, where we show that discarding abstracts unrelated to inorganic materials science improves performance, and models trained on the set of all Wikipedia articles (about ten times more text than our corpus) perform substantially worse on materials science analogies. Contrary to what might seem like the conventional machine learning mantra, throwing more data at the problem is not always the solution. Instead, the quality and domain-specificity of the corpus determine the utility of the embeddings for domain-specific tasks.

We suggest that the methodology described here can also be generalized to other language models, such that the probability of an entity (such as a material or molecule) co-occurring with words that represent a target application or property can be treated as an indicator of performance. Such language-based inference methods can become an entirely new field of research at the intersection between natural language processing and science, going beyond simply extracting entities and numerical values from text and leveraging the collective associations present in the research literature. Substitution of Word2vec with context-aware embeddings such as BERT[25] or ELMo[26] could lead to improvements for functional material predictions, as these models are able to change the embedding of the word based on its context. They substantially outperform context-independent embeddings such as Word2vec or GloVe across all conventional NLP tasks. Also, in addition to co-occurrences, these models can capture more complex relationships between words in the sentence, such as negation. In the current study, the effects of negation are somewhat mitigated because scientific abstracts often emphasize positive relationships. However, a natural extension of this work is to parse the full texts of articles. We expect the full texts will contain more negative relationships and in general more variable and complex sentences, and will therefore require more powerful methods.

Scientific progress relies on the efficient assimilation of existing knowledge in order to choose the most promising way forward and to minimize re-invention. As the amount of scientific literature grows, this is becoming increasingly difficult, if not impossible, for an individual scientist. We hope that this work will pave the way towards making the vast amount of information found in scientific literature accessible to individuals in ways that enable a new paradigm of machine-assisted scientific breakthroughs.

1. Hill, J. et al. Materials science with large-scale data and informatics: unlocking new opportunities. *MRS Bull*. **41**, 399–409 (2016).
2. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
3. Friedman, C., Kra, P., Yu, H., Krauthammer, M. & Rzhetsky, A. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* **17**, S74–S82 (2001).
4. Müller, H. M., Kenny, E. E. & Sternberg, P. W. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*. **2**, e309 (2004).
5. Swain, M. C. & Cole, J. M. ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model*. **56**, 1894–1904 (2016).
6. Eltyeb, S. & Salim, N. Chemical named entities recognition: a review on approaches and applications. *J. Cheminform*. **6**, 17 (2014).
7. Kim, E. et al. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater*. **29**, 9436–9444 (2017).
8. Leaman, R., Wei, C. H. & Lu, Z. TmChem: a high performance approach for chemical named entity recognition and normalization. *J. Cheminform*. **7**, S3 (2015).
9. Krallinger, M., Rabal, O., Lourenço, A., Oyarzabal, J. & Valencia, A. Information retrieval and text mining technologies for chemistry. *Chem. Rev*. **117**, 7673–7761 (2017).
10. Spangler, S. et al. Automated hypothesis generation based on mining scientific literature. In *Proc. 20th ACM SIGKDD Intl Conf. Knowledge Discovery and Data Mining* 1877–1886 (ACM, 2014).
11. Mikolov, T., Corrado, G., Chen, K. & Dean, J. Efficient estimation of word representations in vector space. Preprint at https://arxiv.org/abs/1301.3781 (2013).
12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed representations of words and phrases and their compositionality. Preprint at https://arxiv.org/abs/1310.4546 (2013).
13. Pennington, J., Socher, R. & Manning, C. GloVe: global vectors for word representation. *Proc. 2014 Conf. Empirical Methods in Natural Language Processing (EMNLP)* 1532–1543 (Association for Computational Linguistics, 2014).
14. Liu, W. et al. New trends, strategies and opportunities in thermoelectric materials: a perspective. *Materials Today Physics* **1**, 50–60 (2017).
15. He, J. & Tritt, T. M. Advances in thermoelectric materials research: looking back and moving forward. *Science* **357**, eaak9997 (2017).
16. Ricci, F. et al. An ab initio electronic transport database for inorganic materials. *Sci. Data* **4**, 170085 (2017).
17. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev*. **136**, B864–B871 (1964).
18. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev*. **140**, A1133–A1138 (1965).
19. Gaultois, M. W. et al. Data-driven review of thermoelectric materials: performance and resource onsiderations. *Chem. Mater*. **25**, 2911–2920 (2013).
20. Spearman, C. The proof and measurement of association between two things. *Am. J. Psychol*. **15**, 72–101 (1904).
21. Plirdpring, T. et al. Chalcopyrite $CuGaTe_2$: a high-efficiency bulk thermoelectric material. *Adv. Mater*. **24**, 3622–3626 (2012).
22. Tian, H. et al. Low-symmetry two-dimensional materials for electronic and photonic applications. *Nano Today* **11**, 763–777 (2016).
23. Pandey, C., Sharma, R. & Sharma, Y. Thermoelectric properties of defect chalcopyrites. *AIP Conf. Proc*. **1832**, 110009 (2017).
24. Zhao, L.-D. et al. Ultralow thermal conductivity and high thermoelectric figure of merit in SnSe crystals. *Nature* **508**, 373–377 (2014).
25. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. Preprint at https://arxiv.org/abs/1810.04805 (2018).
26. Peters, M. E. et al. Deep contextualized word representations. Preprint at https://arxiv.org/abs/1802.05365 (2018).
27. Jain, A. et al. The materials project: a materials genome approach to accelerating materials innovation. *APL Mater*. **1**, 011002 (2013).

METHODS

Data collection and processing. We obtained approximately 3.3 million abstracts, primarily focused on materials science, physics, and chemistry, through a combination of Elsevier's Scopus and Science Direct application programming interfaces (APIs) (https://dev.elsevier.com/), the Springer Nature API (https://dev.springernature.com/), and web scraping. Parts of abstracts (or full abstracts) that were in foreign languages were removed using text search and regular expression matching, as were articles with metadata types corresponding to 'Announcement', 'BookReview', 'Erratum', 'EditorialNotes', 'News', 'Events' and 'Acknowledgement'. Abstracts with titles containing keywords 'Foreword', 'Prelude', 'Commentary', 'Workshop', 'Conference', 'Symposium', 'Comment', 'Retract', 'Correction', 'Erratum' and 'Memorial' were also selectively removed from the corpus. Some abstracts contained leading or trailing copyright information, which was removed using regular expression matching and heuristic rules. Leading words and phrases such as 'Abstract:' were also removed using similar methods. We further retained only abstracts related to inorganic materials according to a binary classifier (see 'Abstract classification' below). We tuned the classifier for high recall to guarantee the presence of the majority of relevant abstracts at the expense of retaining some irrelevant ones. Removing irrelevant abstracts substantially improved the performance of our algorithm, as discussed in more detail in Supplementary Information section S2. The 1.5 million abstracts that were classified as relevant were tokenized using ChemDataExtractor[5] to produce the individual words. The tokens that were identified as valid chemical formulae using pymatgen[28] combined with regular expression and rule-based techniques were normalized such that the order of elements and common multipliers did not matter (NiFe is the same as $Fe_{50}Ni_{50}$). Valence states of elements were split into separate tokens (for example, Fe(III) becomes two separate tokens, Fe and (III)). We also performed selective lower-casing and deaccenting. If the token was not a chemical formula or an element symbol, and if only the first letter was uppercase, we lower-cased the word. Thus, chemical formulae and abbreviations stayed in their common form, whereas words at the beginning of sentences and proper nouns were lower-cased. Numbers with units were often not tokenized correctly by ChemDataExtractor. We addressed this in the processing step by splitting the common units from numbers and converting all numbers to a special token < nUm>. This reduced the vocabulary size by approximately 20,000 words. We found that correct preprocessing, especially the choice of phrases to include as individual tokens, substantially improved the results. The code used for preprocessing is available at https://github.com/materialsintelligence/mat2vec.

Abstract classification. This work focuses on inorganic materials science. However, our corpus contained some abstracts that fell outside this scope (for example, articles on polymer science). We removed articles outside our targeted area of research literature by training a binary classifier that could label abstracts as 'relevant' or 'not relevant'. We annotated 1,094 randomly selected abstracts; of these, 588 were labelled as 'relevant' and 494 were labelled 'not relevant'. The labelled abstracts were used as data to train a classifier; we used a linear classifier based on logistic regression, where each document is described by a term frequency–inverse document frequency (tf–idf) vector. The classifier achieved an accuracy (f1-score) of 89% using fivefold cross-validation.

Word2vec training. We used the Word2vec implementation in gensim (https://radimrehurek.com/gensim/) with a few modifications. We found that skip-gram with negative sampling loss ($n = 15$) performed best (see Supplementary Information section S2 for comparison between models). The vocabulary consisted of all words that occurred more than five times as well as normalized chemical formulae, independent of the number of mentions. The phrases were generated using a minimum phrase count of 10, score threshold of 15 (ref. [12]) and phrase depth of 2. The latter meant that we repeated the process twice, allowing generation of up to four grams. We also included common terms such as '-', 'of', 'to', 'a' and 'the', which in exceptional cases led to phrases with more tokens. For example, 'state-of-the-art thermoelectric' is one of the five 8-token phrases in our vocabulary. At the end of each phrase generation cycle, we removed phrases that contained punctuation and numbers. The size of the vocabulary approximately doubled after phrase generation. The rest of the hyperparameters were as follows: we used 200-dimensional embeddings, a learning rate of 0.01 decreasing to 0.0001 in 30 epochs, a context window of 8 and subsampling with a 10^{-4} threshold, which subsamples approximately the 400 most common words. Hyperparameters were optimized for performance on approximately 15,000 grammatical and 15,000 materials science analogies, with the score defined as the percentage of correctly 'solved' analogies from the two sets. Hyperparameter optimization and the choice of the corpus are also discussed in more detail in Supplementary Information section S2. The code used for the training and the full list of analogies used in this study are available at https://github.com/materialsintelligence/mat2vec.

Thermoelectric power factors. Each materials structure optimization and band structure calculation was performed with density functional theory (DFT) using the projector augmented wave (PAW)[29] pseudopotentials and the Perdew–Burke–Ernzerhof (PBE)[30] generalized-gradient approximation (GGA), implemented in the Vienna Ab initio Simulation Package (VASP)[31,32]. A $+U$ correction was applied to transition metal oxides[16]. Seebeck coefficient (S) and electrical conductivity (σ) were calculated using the BoltzTraP package[33] using a constant relaxation time of 10^{-14} s at simulated temperatures between 300 K and 1,300 K and for carrier concentrations (doping) between 10^{16} cm^{-3} and 10^{22} cm^{-3}. A 48,770-material subset of the calculations was taken from a previous work[16]; the remaining calculations were performed in this work using the software atomate[34]. All calculations used the pymatgen[28] Python library within the FireWorks[35] workflow management framework. To more realistically evaluate the thermoelectric potential of a candidate material, we devised a simple strategy to condense the complex behaviour of the S and σ tensors into a single power factor metric. For each semiconductor type $\eta \in \{n, p\}$, temperature T, and doping level c, the S and σ tensors were averaged over the three crystallographic directions, and the average power factor, PF_{avg}, was computed. PF_{avg} is a crude estimation of the polycrystalline power factor from the power factor of a perfect single crystal. To account for the complex behaviour of S and σ with T, c, and η, we then took the maximum average power factor over T, c, and η constrained to a maximum cutoff temperature T_{cut} and maximum cutoff doping c_{cut}. Formally, this is $PF^{T_{cut}, c_{cut}}_{avg, max} \equiv \max PF(\eta, T, c)$ such that $T \leq T_{cut}, c \leq c_{cut}$. We chose $T_{cut} = 600$ K and $c_{cut} = 10^{20}$ cm^{-3} because these values resulted in better correspondence with the experimental dataset than more optimistic values, owing to the limitations of the constant relaxation time approximation. The resulting power factor, $PF^{600\,K,\,10^{20}}_{avg, max}$, is equated with 'computed power factor' in this study. To rank materials according to experimental power factors (or zT), we used the maximum value for a given stoichiometry across all experimental conditions present in the dataset from Gaultois et al.[19].

Data availability

The scientific abstracts used in this study are available via Elsevier's Scopus and Science Direct API's (https://dev.elsevier.com/) and the Springer Nature API (https://dev.springernature.com/). The list of DOIs used in this study, the pretrained word embeddings and the analogies used for validation of the embeddings are available at https://github.com/materialsintelligence/mat2vec. All other data generated and analysed during the current study are available from the corresponding authors on reasonable request.

Code availability

The code used for text preprocessing and Word2vec training are available at https://github.com/materialsintelligence/mat2vec.

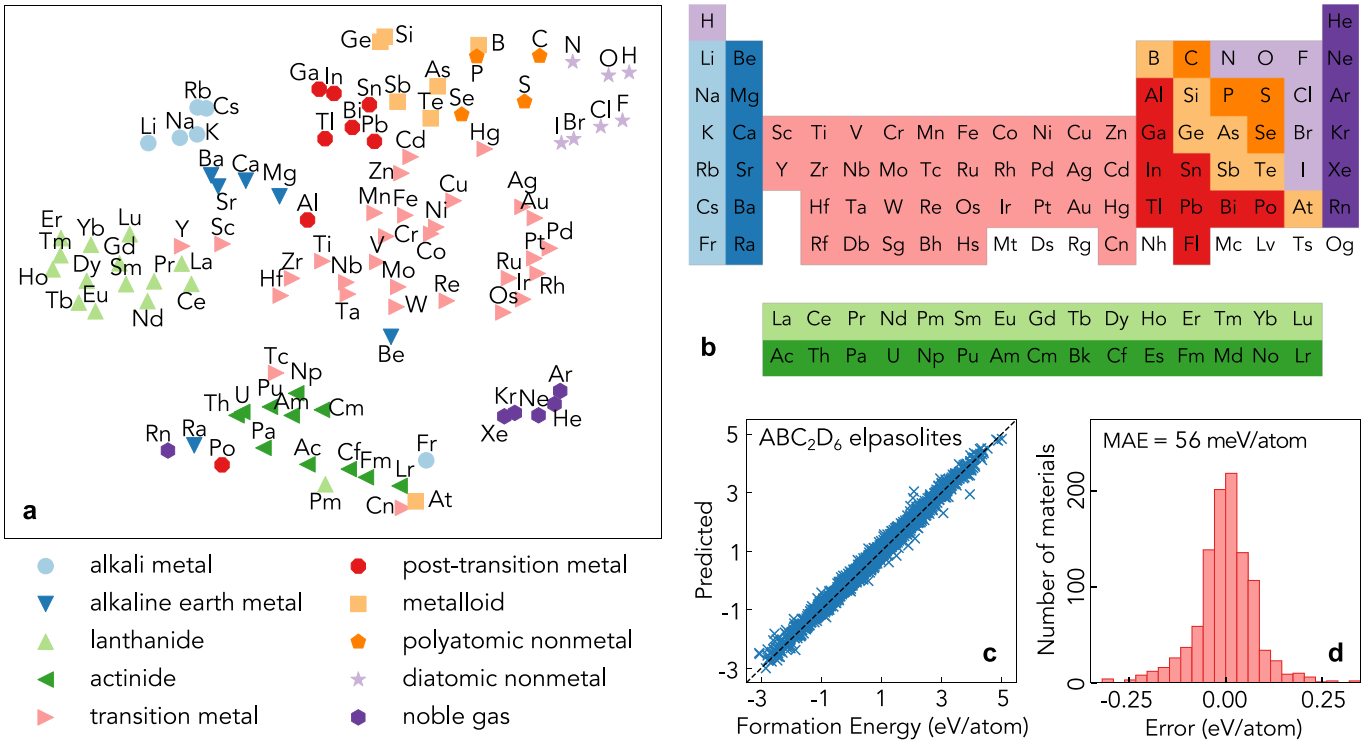
28. Ong, S. P. et al. Python Materials Genomics (pymatgen): a robust, open-source Python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
29. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B Condens. Matter Mater. Phys.* **59**, 1758–1775 (1999).
30. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
31. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B Condens. Matter* **54**, 11169–11186 (1996).
32. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15–50 (1996).
33. Madsen, G. K. & Singh, D. J. Boltztrap. A code for calculating band-structure dependent quantities. *Comput. Phys. Commun.* **175**, 67–71 (2006).
34. Mathew, K. et al. Atomate: a high-level interface to generate, execute, and analyze computational materials science workflows. *Comput. Mater. Sci.* **139**, 140–152 (2017).
35. Jain, A. et al. Fireworks: a dynamic workflow system designed for high-throughput applications. *Concurr. Comput.* **27**, 5037–5059 (2013).
36. Yang, X., Dai, Z., Zhao, Y., Liu, J. & Meng, S. Low lattice thermal conductivity and excellent thermoelectric behavior in Li_3Sb and Li_3Bi. *J. Phys. Condens. Matter* **30**, 425401 (2018).
37. Wang, Y., Gao, Z. & Zhou, J. Ultralow lattice thermal conductivity and electronic properties of monolayer 1T phase semimetal $SiTe_2$ and $SnTe_2$. *Physica E* **108**, 53–59 (2019).
38. Mukherjee, M., Yumnam, G. & Singh, A. K. High thermoelectric figure of merit via tunable valley convergence coupled low thermal conductivity in $A^{II}B^{IV}C_2^V$ chalcopyrites. *J. Phys. Chem. C* **122**, 29150–29157 (2018).
39. Kim, E. et al. Machine-learned and codified synthesis parameters of oxide materials. *Sci. Data* **4**, 170127 (2017).
40. Faber, F. A., Lindmaa, A., Von Lilienfeld, O. A. & Armiento, R. Machine learning energies of 2 million elpasolite (ABC_2D_6) crystals. *Phys. Rev. Lett.* **117**, 135502 (2016).
41. Zhou, Q. et al. Learning atoms for materials discovery. *Proc. Natl Acad. Sci. USA* **115**, E6411–E6417 (2018).

Acknowledgements This work was supported by Toyota Research Institute through the Accelerated Materials Design and Discovery program. We thank T. Botari, M. Horton, D. Mrdjenovich, N. Mingione and A. Faghaninia for discussions.

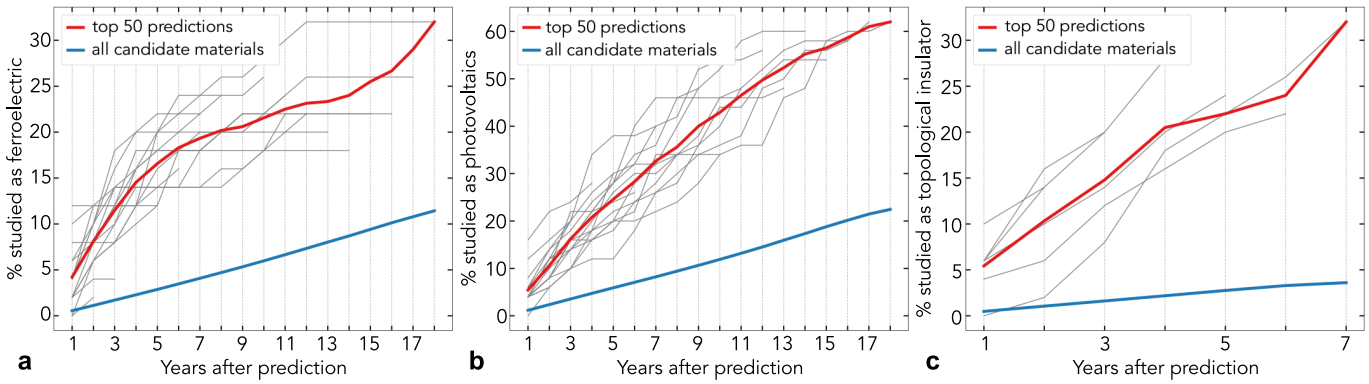
Author contributions All authors contributed to the conception and design of the study, as well as writing of the manuscript. V.T. developed the data processing pipeline, trained and optimized the Word2vec embeddings, trained the machine learning models for property predictions and generated the thermoelectric predictions. V.T., J.D. and L.W. analysed the results and developed the software infrastructure for the project. J.D. trained and optimized the GloVe embeddings and developed the data acquisition infrastructure. L.W. performed the abstract classification. A.D. performed the DFT calculation of thermoelectric power factors. Z.R. contributed to data acquisition. O.K. developed the code for normalization of material formulae. A.D., Z.R. and O.K. contributed to the analysis of the results. K.A.P., G.C. and A.J. supervised the work.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-019-1335-8.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1335-8.
Correspondence and requests for materials should be addressed to V.T., G.C. and A.J.



Extended Data Fig. 1 | Chemistry is captured by word embeddings. **a**, Two-dimensional *t*-distributed stochastic neighbour embedding (t-SNE) projection of the word embeddings of 100 chemical element names (for example, 'hydrogen') labelled with the corresponding element symbols and grouped according to their classification. Chemically similar elements are seen to cluster together and the overall distribution exhibits a topology reminiscent of the periodic table itself (compare to **b**). Arranged from top left to bottom right are the alkali metals, alkaline earth metals, transition metals, and noble gases while the trend from top right to bottom left generally follows increasing atomic number (see Supplementary Information section S4 for a more detailed discussion). **b**, The periodic table coloured according to the classification shown in **a**. **c**, Predicted versus actual (DFT) values of formation energies of approximately 10,000 ABC_2D_6 elpasolite compounds[40] using a simple neural network model with word embeddings of elements as features (see Supplementary Information section S6 for the details of the model). The data points in the plot use fivefold cross-validation. **d**, Error distribution for the 10% test set of elpasolite formation energies. With no extensive optimization, the word embeddings achieve a mean absolute error (MAE) of 0.056 eV per atom, which is substantially smaller than the 0.1 eV per atom error reported for the same task in the original study using hand-crafted features[40] and the 0.15 eV per atom achieved in a recent study using element features automatically learned from crystal structures of more than 60,000 compounds[41].



d

Target word / phrase	Indicator words of a potentially existing study
ferroelectric	ferroelectric, antiferroelectric, ferroelectrics, ferroelectricity, ferro-electricity, relaxor, paraelectric, para-electric, multiferroics, multiferroic, anti-ferroelectric, paraelectricity, ferrolectric, para-electric, ferro-electric, piezoelectric, PZT, pyroelectric, piezo-electric, pyro-electric, magnetoelectric, magnetoelectricity
photovoltaics	solar, photovoltaic, PV, photodevices, photoelectronics, optoelectronic, optoelectronics, nano-optoelectronics, nano-optoelectronic, opto-electronic, opto-electronics, photodiodes, photodiode, photodetectors, photodetector, photosensor, photosensors, photosensing, LED, LEDs
topological insulator	topological, topologically

Extended Data Fig. 2 | Historical validations of functional material predictions. a–c, Ferroelectric (**a**), photovoltaic (**b**) and topological insulator predictions (**c**) using word embeddings obtained from various historical datasets, similar to Fig. 3a. For ferroelectrics and photovoltaics, the range of prediction years is 2001–2018. The phrase 'topological insulator' obtained its own embedding in our corpus only in 2011 (owing to count and vocabulary size limits), so it is possible to analyse the results only over a shorter time period (2011–2018). Each grey line uses only abstracts published before a certain year to make predictions. The lines show the cumulative percentage of predicted materials studied in the years following their predictions; earlier predictions can be analysed over longer test periods. The results are averaged in red and compared to baseline percentages from all materials. **d**, The target word or phrase used to rank materials for each application (based on cosine similarity), and the corresponding words used as indicators for a potentially existing study.

Extended Data Table 1 | Materials science analogies

Relationship	Example vector operation	Answer	Validation pairs	Accuracy (%)
Chemical element names	helium - He + Fe	= iron	8372	71.4
Crystal symmetries	cubic - GaAs + CdSe	= hexagonal	2034	35.4
Crystal structure names	zincblende - GaP + GaN	= wurtzite	556	18.7
Elemental crystal structures	dhcp - La + Cr	= bcc	1198	48.6
Principal oxides	Al_2O_3 - Al + Si	= SiO_2	650	48.8
Units	pressure - Pa + Hz	= frequency	452	35.4
Magnetic properties	ferromagnetic - NiCo + IrMn	= antiferromagnetic	622	41.0
Applications	thermoelectric - PbTe + LiFePO4	= cathode materials	-	-
Grammar	structures - structure + energy	= energies	15162	61.6
Total			29046	60.1

Examples of verified word analogies corresponding to various materials science concepts. The first column lists the types of tested analogies. The second column is an example vector operation for the corresponding analogy type, with the observed answer listed in the third column. The fourth column gives the number of pairs used for scoring the corresponding analogy task, with the resulting score of our model shown in the fifth column. Application analogies were not tested quantitatively and the example is for demonstration purposes only. The full list of tested analogies is available at https://github.com/materialsintelligence/mat2vec.

Extended Data Table 2 | Top 50 thermoelectric predictions

Top 50 thermoelectric predictions				
1. Li_2CuSb	11. MgB_2C_2	21. $CuIn_5S_8$	31. Ag_3SbS_3	41. Eu_2CuSi_3
2. $Cu_3Nb_2O_8$	12. AlGaSb	22. $AlFe_2B_2$	32. $(CH_3NH_3)_3Bi_2I_9$	42. Cu_2ZnSiS_4
3. $CsAgGa_2Se_4$	13. Li_3Sb*	23. CeTe	33. $Ba_4Ga_4SnSe_{12}$	43. Bi_4Br_4
4. Cu_7Te_5	14. $Ba_{24}Si_{100}$	24. $Pb_{0.902}Sn_{0.098}Se$	34. In_3Se_2	44. $(YbS)_{1.25}CrS_2$
5. $Ge_{15}Sb_{47}Te_{38}$	15. GaNAsP	25. $Bi_{0.95}La_{0.05}FeO_3$	35. Ag_2PbGeS_4	45. KCu_2SbS_3
6. $CsGeI_3$	16. $ZnGa_2Te_4$	26. $CdSnP_2$*	36. $AgCrO_2$	46. Cu_2GeTe_3
7. KAg_2SbS_4	17. Cu_3TaS_4	27. PdTe	37. $TlCu_2Se_2$	47. $NaLaS_2$
8. $SnTe_2$*	18. HgMnTe	28. HgZnTe	38. AgGa	48. $Hg_{0.78}Cd_{0.22}Te$
9. Ni_2Te_3	19. $MnBi_2Se_4$	29. $Pr_{0.7}Ca_{0.3}Mn_{0.95}Co_{0.05}O_3$	39. BSb	49. InSn
10. $Yb_{11}AlSb_9$	20. $Ag_6Si_2O_7$	30. Cd_4GeSe_6	40. $CdIn_2S_2Se_2$	50. ReSSe

The top 50 thermoelectric predictions using the full text corpus available at the time of writing. Some of these have practical limitations (for example, the presence of air-sensitive species or toxic and expensive elements), but others appear to be experimentally testable candidates. An exhaustive manual literature search revealed that, from the first 150 predictions using the full corpus of collected abstracts published through 2018, 48 materials (32%) had already been studied as thermoelectrics in papers that were not represented in our corpus, many of which were published within the last two years. In the top 50 listed here we have excluded any predictions for which we could find thermoelectric reports outside our corpus.
*Materials reported as good thermoelectrics while this manuscript was being prepared and reviewed[36–38].

Extended Data Table 3 | Top five functional material predictions and context words

Prediction	Top 10 most contributing context words
	Topological Insulator
1. Sc_2CF_2	armchair direction, zigzag direction, phosphorene, $Sc_2C(OH)_2$, MXene, semiconducting, armchair, semiconductor, Sc_2AlC, strongly anisotropic
2. LaCuOSe	layered oxychalcogenides, oxychalcogenides, oxychalcogenide, degenerate semiconductor, semiconductor, (LaO)CuS, LaAgSeO, p - type, ZrCuSiAs, LaCuOTe
3. Co_2FeAl	heusler alloy, $Co_2Cr_{0.6}Fe_{0.4}Al$, full - heusler, $Co_2FeAl_{1-x}Si_x$, heusler, half - metallic, spin polarized, heusler compound, MFTJs, ferromagnet
4. $Ca_5In_2Sb_6$	$Ca_5Al_2Sb_6$, $Ca_5Ga_2Sb_6$, zintl compound, zintl compounds, $Sr_5In_2Sb_6$, thermoelectric, thermoelectric properties, carrier concentration, carrier mobility, effective mass
5. $AgBiP_2Se_6$	atomically thin, ferroelectricity, semiconductor, two - dimensional, ferroelectric, monolayer, devices, band edge, ground state, plane
	Photovoltaics
1. MoOHCF	electrochromic window, electrochromic, electrochromic device, vivid color, counter electrode, visible wavelengths, prussian blue, transmittance, optical transmittance, thin film
2. MoN_2	renewable energy, appealing, applications, NIBs, great potential, realization, ion batteries, dinitride, promising, MoN_3
3. $Ni_{0.4}Co_{0.6}(OH)_2$	flexible, supercapacitors, peony - like, fabrication, step hydrothermal, strategy, CFC, carbon fiber, cloth, superior
4. NiFeS	NiVS, highly efficient, low cost, $FeNiS_2$, [NiFe], NiFeVS, efficient electrocatalyst, Ni-Fe-V, OER, promising
5. Si_2BN	graphenelike, have attracted, NB_2Si, anode material, much attention, nanostructures, hydrogen storage, battery anode, $Si_3B_3N_7$, buckled
	Ferroelectric
1. $BaTiSi_2O_7$	spontaneous polarization, dielectric, lead - free, fresnoite, ceramics, $[TiO_5]$, difficult to prepare, glass, phase, properties
2. $GdTiO_3$	$BaTiO_3$, BTO, $SmTiO_3$, $SrTiO_3$, gate dielectric, ferrimagnetic, mott insulator, perovskite, pyrochlore, magnetic ordering
3. $TlGaTe_2$	dielectric, dependences of the permittivity, dielectric constant, $TlInSe_2$, $TlInTe_2$, relaxors, permittivity, $x(TlGaTe_2)x$, ε(T), dielectric relaxation
4. $Ba_5NdTi_3Ta_7O_{30}$	dielectric, $Bi_4Ti_3O_{12}$, tungsten – bronze, dielectrics, $Ba_4Nd_2Ti_4Ta_6O_{30}$, co-substitution, ceramics, tetragonal, x, co-modification
5. $Pb_3Mn_7O_{15}$	dielectric, $Pb_2Te_3O_8$, magnetic ordering, Zn_2PbO_4, antiferromagnetic, $Mn_2Te_3O_8$, $Pb_3Rh_7O_{15}$, Pb_2ZnTeO_6, orthorhombic, manganite

The top five predictions and top ten most important context words leading to the prediction for topological insulators, photovoltaics and ferroelectrics using the full text corpus. A list of context words that could indicate prior study in the target domain have already been excluded in the process of making the predictions, as mentioned in Extended Data Fig. 2d. Furthermore, we have excluded any predictions for which we could find reports outside our corpus for the target application.

Extended Data Table 4 | Importance of the text corpus

Text corpus	Materials	Grammar	All	Corpus size
Wikipedia	2.6	72.8	51.0	2.81B words
Wikipedia elements	2.7	72.1	41.4	1.08B words
Wikipedia materials	2.2	72.8	41.3	781M words
All abstracts	43.3	58.3	51.0	643M words
Relevant abstracts	48.9	54.9	52.0	290M words
Pre-trained model from Kim et al[39]	10.4	47.1	30.8	640k papers

The top analogy scores in per cent for materials science and grammatical analogy tasks for different corpora. All models except Kim et al.[39] were trained using CBOW—continuous bag of words, the other variant of Word2vec, alongside skip-gram—with the same hyper-parameters (negative sampling loss with 15 samples, 10^{-4} downsampling, window 8, size 200, initial learning rate 0.01, 30 training epochs, minimum word count 5) and no phrases. We used the English Wikipedia dump from March 1, 2018. 'Wikipedia elements' corresponds to a subset of articles that mention a chemical element name (for example, 'gold'), whereas 'Wikipedia materials' corresponds to a subset that mention at least one material formula. The smallest corpus on which we train our model has the best performance on materials-related analogies, whereas the largest corpus has the best performance for grammar. We believe this is due to the highly specialized nature of the relevant abstracts, suitable for the tested analogy pairs. We used the 'Relevant abstracts' corpus throughout this study.