npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-024-01231-8

# Leveraging language representation for materials exploration and discovery



Jiaxing Qu [1,5] ✉, Yuxuan Richard Xie [2,5], Kamil M. Ciesielski[3], Claire E. Porter[3], Eric S. Toberer[3] & Elif Ertekin [1,4] ✉

Data-driven approaches to materials exploration and discovery are building momentum due to emerging advances in machine learning. However, parsimonious representations of crystals for navigating the vast materials search space remain limited. To address this limitation, we introduce a materials discovery framework that utilizes natural language embeddings from language models as representations of compositional and structural features. The contextual knowledge encoded in these language representations conveys information about material properties and structures, enabling both similarity analysis to recall relevant candidates based on a query material and multi-task learning to share information across related properties. Applying this framework to thermoelectrics, we demonstrate diversified recommendations of prototype crystal structures and identify under-studied material spaces. Validation through first-principles calculations and experiments confirms the potential of the recommended materials as high-performance thermoelectrics. Language-based frameworks offer versatile and adaptable embedding structures for effective materials exploration and discovery, applicable across diverse material systems.

The goal of inorganic materials discovery and design is to efficiently navigate the materials space and identify candidates that exhibit targeted and desirable properties. However, search and identification remain persistent challenges due to (i) rapidly growing complexity with structural and chemical diversity, and (ii) varied and complicated mappings from material space to objective space. The lack of a universal technique for exploring the vast and mostly unlabeled materials space is a significant bottleneck that limits search efficiency. Ab-initio methods play a crucial role in materials science by providing accurate and predictive insights into the electronic structure, properties, and behaviors of materials. High-throughput simulations of material properties through ab-initio simulations have greatly facilitated material discovery for many functional applications[1–3]. The rapid growth of data in materials science has led to the promising application of machine learning (ML) to also overcome these obstacles and expedite materials discovery workflows[4–6].

A key challenge in the widespread adoption of ML for materials search lies in defining universal model input representations. An ideal representation should enable the conversion of inorganic crystals into a machine-readable format and facilitate the encoding of inorganic materials into features that capture complexities such as defects, alloying, and disorder. In early ML models, material representation involved hand-crafted descriptors that contained information about composition and structure[7–9]. These descriptors relied heavily on physical and chemical intuition. In more recent approaches, material atomic structures have been treated as graphs, where convolution operations are used to extract features from local chemical environments[10–14], enabling more accurate property predictions. However, both hand-crafted features and specialized structural models have limitations in providing universal and task-agnostic representations within the vast material space. For instance, representations obtained by tailoring graph neural networks to energy prediction regression tasks may perform poorly on conductivity predictions.

In this work, we assess the effectiveness of language representations in tackling general materials discovery tasks. Advances in natural language processing have made it possible to extract valuable information from the extensive corpus of materials science literature. Further, these advances have made it easier to encode domain knowledge into compact and information-rich vector representations. A pioneering study demonstrated that word embeddings have the ability to capture underlying knowledge in materials science, and effectively applied these embeddings for tasks such as materials search and ranking[15]. However, word embeddings alone do not capture the

[1]Department of Mechanical Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [2]Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [3]Department of Physics, Colorado School of Mines, Golden, CO 80401, USA. [4]Materials Research Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61802, USA. [5]These authors contributed equally: Jiaxing Qu, Yuxuan Richard Xie. ✉e-mail: jiaxing6@illinois.edu; ertekin@illinois.edu

contextual meaning that is present within a sentence or paragraph. Progress in contextual embeddings has been greatly enhanced by transformer models using masked language modeling. This advance has enabled the creation of information-rich contextual embeddings within the latent materials science domain[16–18]. Here, we leverage pretrained transformer models to capture contextual embeddings, and then utilize them for language representations to enable materials exploration and discovery.

In the context of materials exploration and discovery, some essential factors should be included in the recommendation pipeline: (i) effective representations of both chemical and structural complexity, (ii) successful recall of relevant candidates to the query material or property of interest, and (iii) accurate candidate ranking based on multiple desired functional properties. Previously, recommender-like systems for materials were developed to filter by identifying materials for which predicted confidence levels of target properties fall within a desirable range for thermoelectrics[19], to predict chemically relevant compositions for pseudo-ternary systems[20,21], and to propose experimental synthesis conditions[22] or by similarity analysis from word embedding models[15]. However, a systematic and generalizable recommendation approach, which incorporates the three factors mentioned above for representation, recall, and ranking, could accelerate discovery of desirable material candidates across diverse applications.

We present a materials recommendation framework that utilizes language representations to explore the vast materials space. This framework allows for the identification of similar candidates given a query material with targeted properties. The framework invokes a funnel architecture consisting of two steps: candidate generation, referred to as “recall”, and property evaluation, referred to as “ranking” (Fig. 1a). This architecture helps narrow down the search space by initially generating potential candidates and then evaluating their properties to identify the most relevant ones. Through the evaluation of different embedding methods across various downstream tasks, our findings demonstrate that language representations are highly potent in recalling relevant material candidates. Additionally, these representations are effective in predicting material properties, achieving performance levels comparable to state-of-the-art specialized machine learning models. In order to enhance the ranking performance, we introduce a multi-gate mixture-of-experts (MMoE) model, which is a multi-task learning strategy that leverages correlations between material property prediction tasks (Fig. 1b). By incorporating MMoE, we demonstrate that pre-existing knowledge contained in the latent space can be effectively transferred to new tasks, resulting in faster and more effective learning.

As a demonstration of the framework’s materials discovery capabilities, we apply it to search for and recommend high-performance thermoelectric (TE) materials. The framework successfully identifies structurally-diversified thermoelectric (TE) candidates that are relevant to several query materials. Furthermore, the framework identifies and subsequently searches several under-explored materials spaces that host promising TE candidates. The effectiveness of this framework is evaluated using first-principles calculations and experimental validation on the recommended materials, resulting in several promising thermoelectrics.

## Results

### A language-based framework enables material recommendations and discovery

Machine learning-based recommender systems leverage a large corpus of training data to provide precise suggestions when querying for items among a large candidate pool[23,24]. During the recommendation process, a funnel-based architecture is typically applied for initial screening, followed by more fine-grained ranking steps. Inspired by the standard design of recommender systems, we adapted the framework to materials science to effectively search a large space and recommend relevant materials with similar functional performance to a query material. Specifically, we designed a funnel-based architecture that can be decoupled into a recall step and a ranking step (Fig. 1a). To enable candidate recall for a query material, we converted each material into text-based descriptions that include both compositional and structural information. Using language models[16,17] pretrained on materials science literature, we then obtained output embeddings on these text-based material descriptions. These embeddings encode contextual representations to capture compositional and structural features with high-level interactions arising from self-attention[25]. In the recall step ("candidate generation”), candidates can be searched via cosine similarity against the query material in the representation space (Fig. 1b). In the ranking step, recalled candidate materials are evaluated and ranked using a multi-objective scoring function trained on the encoded representations to simultaneously predict multiple material properties through neural networks. For this work, we exploited task correlations between predicting five TE properties by training multi-task learning MMoE models, which provided improved accuracy compared to models trained on single tasks.

To obtain compositional and structural level representations for the database consisting of 116K materials (Data Preparation), we embedded all material formulae (e.g., “PbTe") and sentence descriptors automatically generated (Robocrystallographer[26]) from the structures (e.g., “PbTe is Halite, Rock Salt structured and crystallizes in the cubic $Fm\bar{3}m$ space group...") as the input to pretrained language models. Embedding each formula or structure generates a dense vector output from the model’s hidden layer, which contains latent material-specific knowledge learnt during unsupervised pretraining. In Fig. 2, we demonstrate that recalled candidates in the representation space are not only compositionally and structurally related to the query material, but also can exhibit similar functional performance to a query material. Starting with known materials with favorable properties for TEs such as PbTe, we analyzed the top recalled candidates and found significantly different predicted figure-of-merit $zT$

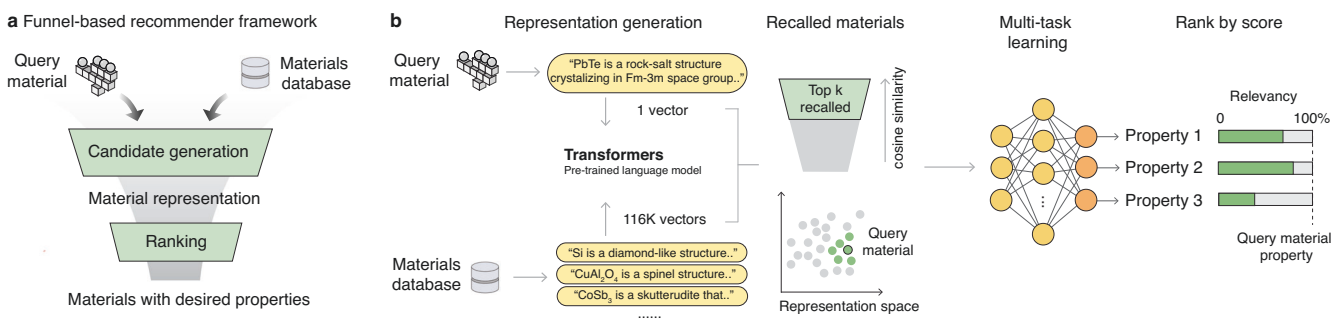


**Fig. 1 | Recommender framework for material discovery. a** The proposed funnel-based recommender framework in which candidate materials are recalled, and ranked based on similarity to the query material. **b** The schematic workflow to screen candidate materials including constructing language representations, recalling candidates, and multi-task prediction for ranking.

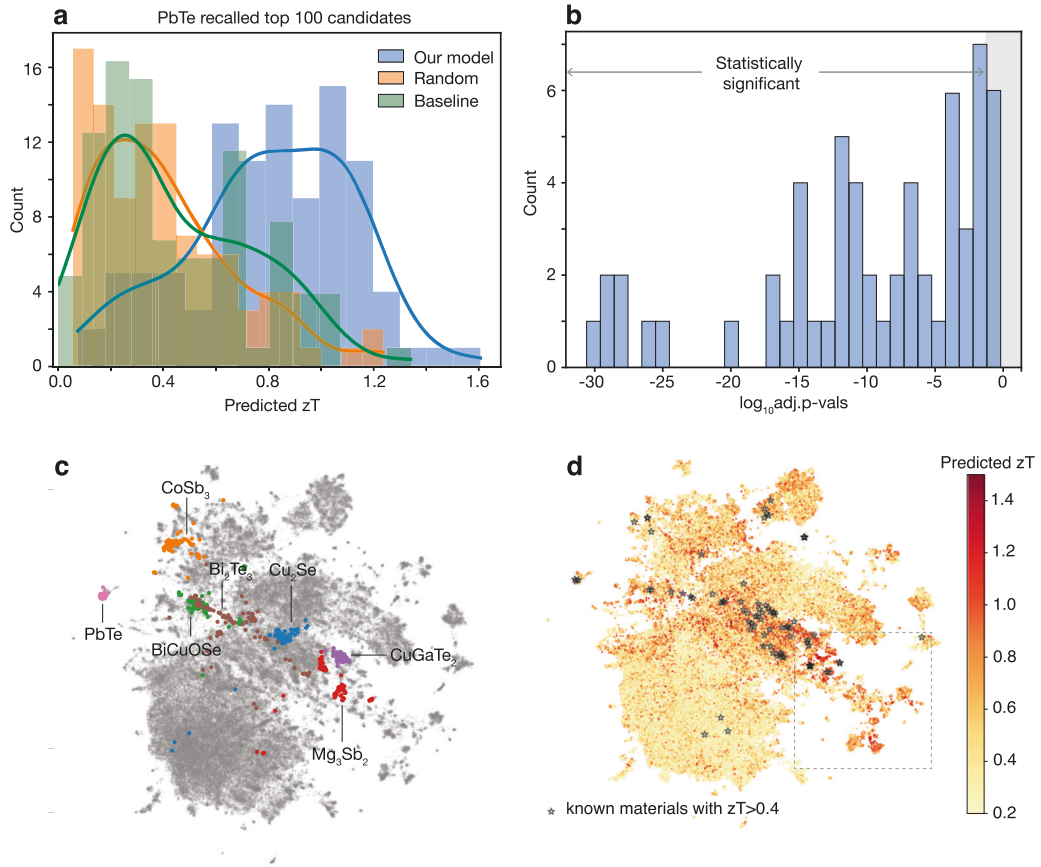


**Fig. 2 | Overview of recommender framework performance. a** Distributions of predicted *zT* of the top-100 recalled candidates for PbTe as the query material predicted by MatBERT, baseline model with fingerprint[28] representation (Material language representations) and randomly sampled 100 materials. The distributions are also visualized in kernel density estimation. Predicted *zT* are obtained from the MMoE models. **b** Adjusted *p*-values of the candidates for the top-100 highest *zT* materials from experimental UCSB and ESTM datasets. Out of 100 materials, 94 show significantly different *zT* distributions from random. **c** Recall results of seven high-performing TE materials are highlighted on the UMAP projection of 116K material representations. Each color corresponds to first 100 materials recalled via cosine similarity. The UMAP projection is obtained on the best performing embeddings (MatBERT, structure embedding). **d** UMAP overlaid with the predicted *zT*. All known materials with *zT*>0.4 from the experimental datasets are indicated by stars.

distributions from selected baseline representations (see details in, Material language representations) and random sampling, as shown in Fig. 2a. The differences in distributions are quantified by *p*-values. The baseline model distribution exhibits a significant degree of overlap with random sampling, as indicated by a high *p*-value (0.88), while our model with contextual language representation shows statistical significance with $p<0.05$. We repeated this experiment for 100 distinct materials with the known highest *zT*; 94 of these exhibit statistical significance (Fig. 2b), indicating that recalled materials show distinct distributions from the baseline representation and random sampling. Low-dimensional Uniform Manifold Approximation and Projection (UMAP)[27] of the material representations display latent signatures of seven high-performing TE materials along with their top-100 recalled materials, each indicated by a different color (Fig. 2c). We further observed a distinct clustering pattern, in which known materials with good *zT* (> 0.4) form a "band" in the projection (Fig. 2d). Additionally, the observed "band" overlaps with the MMoE-predicted high *zT* (also Fig. 2d). The distribution of *zT* in the representation space provides opportunities to explore under-explored material spaces, such as the region enclosed in the grey box with high predicted *zT*.

To understand how individual steps contribute to the performance of the material recommendation framework, in the following we assess the effectiveness of different representation strategies, recall ability, and property prediction via multi-task learning. Further, we demonstrate the framework to search, ranking, and exploration tasks for TE materials.

## Language models offer effective representations of material composition, structure, and properties

Effective representations require rendering information about material design principles and intrinsic properties. We evaluated several strategies for material representation, focusing on unsupervised generation of features to convey diverse chemical and structural information. In total, we investigated six embedding methods. For composition level representation, we embed the material formula using pretrained word embedding Mat2Vec[15] and contextualized word embedding from MatSciBERT[17] and MatBERT[16]. For structural level representation, we obtained local environment based structure fingerprints[28] and sentence embeddings of text-based material descriptions from MatSciBERT and MatBERT. Note that for BERT models, we constructed embedding vectors from entire passages of text consisting of human-readable crystal structure characteristics[26], as described in Material language representations. In the following analysis, we choose Mat2Vec and fingerprint as baselines for compositional embedding and structural embedding respectively.

To assess whether the embedding models have encoded material knowledge in the representations, we projected the six different material embedding vectors into low-dimensional spaces with UMAP, as visualized in Fig. 3. Embedded materials consisting of groups 15 (pnictogen), 16 (chalcogen), and 17 (halogen) on the periodic table are indicated by color (Fig. 3a). Overall, structure level representations exhibit more distinct separation (well-defined domains) by material groups, apart from fingerprints which are solely

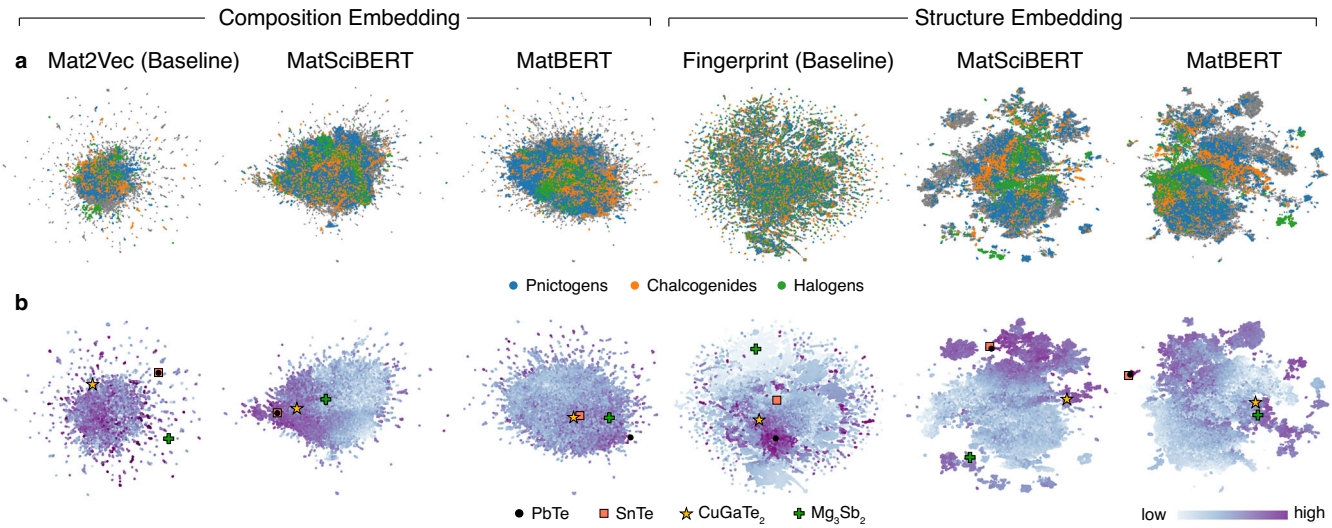


**Fig. 3 | UMAP projections of 116K materials using different embedding models.** Materials are colored (**a**) by anionic groups, and (**b**) by cosine similarity to PbTe under the indicated embedding model.

**Table 1 | Benchmarking six different embedding models on six regression property prediction tasks with MAEs and $R^2$-scores**

| Composition embedding | | | | | Structure embedding | | |
|---|---|---|---|---|---|---|---|
| Property | Metric | Mat2Vec (Baseline) | MatSciBERT | MatBERT | Fingerprint (Baseline) | MatSciBERT | MatBERT |
| *E*/atom | MAE | 0.47 ± 0.02 | 0.42 ± 0.01 | 0.37 ± 0.01 | 1.13 ± 0.02 | 0.32 ± 0.02 | 0.29 ± 0.03 |
| | $R^2$ | 0.81 ± 0.02 | 0.86 ± 0.01 | 0.88 ± 0.01 | 0.283 ± 0.02 | 0.95 ± 0.01 | 0.96 ± 0.01 |
| $E_g$ | MAE | 0.15 ± 0.01 | 0.20 ± 0.02 | 0.19 ± 0.01 | 0.54 ± 0.03 | 0.25 ± 0.01 | 0.23 ± 0.01 |
| | $R^2$ | 0.92 ± 0.02 | 0.88 ± 0.02 | 0.88 ± 0.01 | 0.45 ± 0.04 | 0.88 ± 0.01 | 0.89 ± 0.01 |
| log_*K* | MAE | 0.18 ± 0.01 | 0.18 ± 0.01 | 0.17 ± 0.01 | 0.45 ± 0.01 | 0.16 ± 0.01 | 0.15 ± 0.01 |
| | $R^2$ | 0.83 ± 0.01 | 0.83 ± 0.03 | 0.85 ± 0.02 | 0.26 ± 0.02 | 0.90 ± 0.01 | 0.93 ± 0.01 |
| log_*G* | MAE | 0.20 ± 0.01 | 0.23 ± 0.01 | 0.22 ± 0.01 | 0.48 ± 0.01 | 0.24 ± 0.01 | 0.23 ± 0.01 |
| | $R^2$ | 0.82 ± 0.01 | 0.80 ± 0.01 | 0.81 ± 0.02 | 0.29 ± 0.03 | 0.83 ± 0.01 | 0.84 ± 0.01 |
| $\log_{10}$_θ | MAE | 0.06 ± 0.01 | 0.07 ± 0.01 | 0.06 ± 0.01 | 0.13 ± 0.01 | 0.07 ± 0.01 | 0.06 ± 0.01 |
| | $R^2$ | 0.81 ± 0.02 | 0.82 ± 0.03 | 0.84 ± 0.02 | 0.34 ± 0.05 | 0.85 ± 0.03 | 0.88 ± 0.02 |
| $\log_{10}$_α | MAE | 0.07 ± 0.01 | 0.07 ± 0.01 | 0.07 ± 0.01 | 0.15 ± 0.01 | 0.07 ± 0.01 | 0.06 ± 0.01 |
| | $R^2$ | 0.78 ± 0.03 | 0.81 ± 0.02 | 0.81 ± 0.02 | 0.19 ± 0.02 | 0.87 ± 0.03 | 0.90 ± 0.01 |

E/atom Energy per atom (eV), $E_g$ Band gap (eV), *K* Bulk modulus (GPa), *G* Shear modulus (GPa), θ Debye temperature (K), α Coefficient of thermal expansion ($K^{-1}$).

determined by structural similarity and include information only about local but not semi-local and global environments. By contrast, composition level representations retain the expected chemical differences, but form more disperse and heterogeneous clusters.

To better interpret the embedding results, we picked three well-studied TE materials, including SnTe – a rock-salt structural analog of PbTe with highest reported *zT* of ~ 1.8[29], $CuGaTe_2$ – a diamond-like semiconductor in chalcopyrite structure that achieves a *zT* of 1.5[30], as well as $Mg_3Sb_2$ – a layered Zintl phase with the highest *zT* of ~ 1.65[31,32], and visualized their proximity to PbTe in the representation space (Fig. 3b). All three materials have demonstrate high *zT* around 1.5, but the high performance arises from different combinations of properties relevant to TEs (i.e. electronic and thermal transport) due to their different structures. Embedded materials in the representation space follow our anticipated similarity (PbTe ≈ SnTe > $CuGaTe_2$ > $Mg_3Sb_2$), apart from fingerprints and composition embeddings from MatBERT. Although the embedding analysis helps with the visualization of chemical trends under different embedding models, it is possible that embedding models could show different clustering patterns when using alternative measures of similarity.

For further evaluation, we quantitatively evaluated material embedding performance on downstream property prediction tasks. We applied a feature-based approach to train regression models directly on the derived embeddings, instead of optimizing BERT parameters on the task-specific loss, i.e., fine-tuning[25]. This approach is more computationally efficient due to fixed features, and grants flexibility to adapt task-specific architectures or combine features of various sources across different models. We list the cross validation performance on predicting six material properties for 5700 materials in Table 1. The task models were multi-layer perceptrons (MLPs) with mean-absolute-error (MAE) training loss. The tasks consisted of band gap, energy per atom, bulk modulus, shear modulus, Debye temperature, and coefficient of thermal expansion from AFLOW dataset[33]. Performance metrics of models trained using several embeddings, such as structure embeddings extracted from MatBERT, achieved accurate performance. In addition, we performed predictions on 7 different material benchmarks from the MatBench[34] dataset as shown in Supplementary Fig. 15. The results indicate reasonable performances using the language representation as the model input for downstream predictive tasks. By leveraging latent materials science knowledge embeddings from pretrained large language models, the language

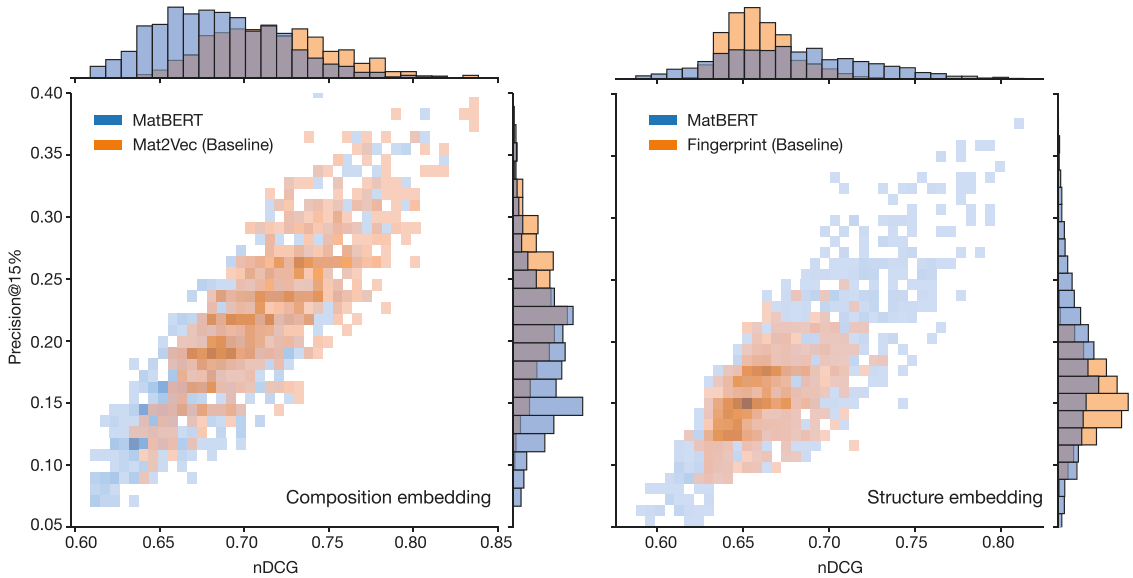


**Fig. 4 | Evaluation metric of candidate generation stage for 116K materials.** Each square represents a different query material. The *x* and *y*-axis represent precision at 15% and nDCG, which measure the accuracy and relative ranking order respectively. Evaluations are performed for composition embedding and structure embedding.

representation supports learning in the face of data scarcity, a ubiquitous challenge in applying ML to materials science. For small data with only 200 training materials, models trained using these embeddings outperform graph neural networks (CGCNN[10] when tested on 100 independent materials (Supplementary Fig. 1). These results suggest that pretrained language models, in combination with text-based structure descriptions, provide a competitive avenue to generate features for material representations.

### Unsupervised candidate recall extracts highly relevant materials

For the recall (candidate generation) step, we use an unsupervised approach, as the models need to generalize to unseen query materials without further training while correctly recalling relevant candidates. In the unsupervised context, each query material is an individual prediction task, where the goal is to find a set of related materials. While material recall is strictly based on cosine similarity in embedding space from text-based descriptions of composition and/or structure, we hypothesize that these embeddings contain latent materials science knowledge, in which recalled materials will also share some similarities in properties to the query material.

For commercial recommender systems, online learning[35,36] makes data collection and model evaluation straightforward. For this framework, we evaluate recalled TE materials in an offline setting with predefined 'relevancy' (Ranking score and exploratory analysis) as a measure of the composite differences in TE properties. We considered five TE properties – power factor, Seebeck coefficient, electrical conductivity, thermal conductivity, and $zT$ for 826 unique host materials (Data preparation) For each query material, the relevancy is obtained as the summation of absolute percentage differences of these five properties, i.e., candidates with similarity across all properties are considered most relevant.

For evaluation, precision and normalized discounted cumulative gain (nDCG) were used as recall performance metrics (see Evaluation: unsupervised recall of relevant materials). Specifically, we calculated precision@15% to assess the recall accuracy by defining the top 15% of 826 materials (124 materials) as 'relevant' to the query material based on experimental TE properties. We then evaluated the overlap between the top 124 recalled and relevant materials. A higher precision@15% score indicates that more relevant materials have been recalled. On the other hand, nDCG evaluates the ranking from the perspective of relative positions of items in the similarity-based list. The two metrics are jointly visualized in Fig. 4 and analyzed separately for composition and structure embeddings. Each scatter point denotes performance for one queried material. Ideally, candidates should have high precision@15% and nDCG (top right corner). Using composition embeddings, Mat2Vec exhibited overall better performance than MatBERT on both precision@15% and nDCG, indicating the effectiveness of Mat2Vec word embedding in capturing latent material science knowledge. For single-word inputs like material composition, the word embedding Mat2Vec model performs well where contextual information is less critical, while the sentence embedding MatBERT model shows suboptimal performance limited by complicated model architectures for simple tasks. For structure embeddings, however, MatBERT recalls significantly more relevant materials than using fingerprint. This performance is not surprising, since fingerprints only contain information about local structure at the motif level, but lack information at the semi-local level (i.e., motif connectivity) and global level (e.g. space group). Leveraging the contextual information from the sentence embedding model derived from robocrystallographer descriptions, MatBERT embeddings achieve better performance than the baseline. In general, MatBERT embeddings have the overall best performance among all embedding models. From both composition and structure MatBERT embeddings, a considerable number of materials achieved precision > 0.25 and nDCG > 0.7, suggesting that the representations extracted similarity preserving signals which could be utilized for unsupervised search for similarly performing materials.

### Multi-task learning exploits cross-task correlations for improved property predictions

For a more accurate candidate material ranking, in the second stage of the funnel approach of Fig. 1 we improved multi-property predictions through multi-task learning. Learning from multiple related tasks provides superior performance over single-task learning by modeling task-specific objectives and cross-task relationships[37,38]. Multi-task learning is thus ideal to learn the underlying commonalities across different yet correlated material properties, improving performance for each task.[39] showed that joint-training on several material properties leads to better model performance in prediction tasks. A mixture-of-experts framework[40] demonstrates transferability between models trained on different material properties, thereby improving task performance.

To this aim, we introduce multi-task learning with the MMoE model, which contains a set of expert networks and gating networks (Fig. 5a). Through task-specific tower networks, the gating network for each property prediction allows the model to learn mixture contributions from different experts, thus exploiting the interconnections between tasks (MMoE and TE prediction). In the approach adopted here, the input representations for MMoE models, discussed later in this section, are concatenated composition

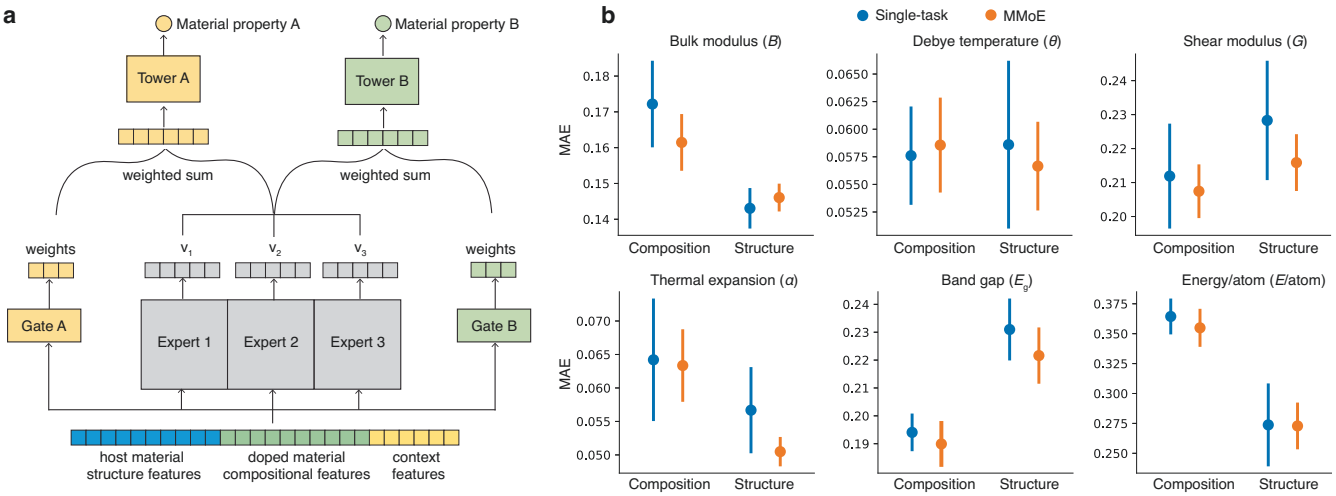


**Fig. 5 | Multi-task learning framework for material property prediction.** **a** Schematic of the MMoE model. The gating networks learn contributions from different experts through task-specific tower networks. **b** Comparison of model performance for 6 material property prediction tasks between single-task models and MMoE using composition or structure embeddings.

**Table 2 | MAEs and $R^2$-scores of four representation methods as the input for MatBERT model to predict the thermoelectric properties on the UCSB and ESTM dataset**

| Property | Evaluation | Host composition | Host structure | Doped composition | Host structure + Doped composition |
|---|---|---|---|---|---|
| log_*PF* | MAE | 0.580 ± 0.035 | 0.566 ± 0.045 | 0.471 ± 0.039 | 0.433 ± 0.024 |
| | $R^2$ | 0.584 ± 0.080 | 0.624 ± 0.095 | 0.740 ± 0.060 | 0.778 ± 0.063 |
| *S* | MAE | 52.3 ± 6.1 | 53.3 ± 8.0 | 36.8 ± 5.1 | 35.4 ± 3.4 |
| | $R^2$ | 0.741 ± 0.069 | 0.753 ± 0.032 | 0.862 ± 0.070 | 0.872 ± 0.046 |
| log_σ | MAE | 1.151 ± 0.076 | 1.157 ± 0.063 | 0.693 ± 0.040 | 0.654 ± 0.074 |
| | $R^2$ | 0.576 ± 0.080 | 0.585 ± 0.076 | 0.813 ± 0.036 | 0.832 ± 0.044 |
| log_κ | MAE | 0.270 ± 0.020 | 0.272 ± 0.029 | 0.237 ± 0.014 | 0.221 ± 0.022 |
| | $R^2$ | 0.779 ± 0.051 | 0.772 ± 0.049 | 0.824 ± 0.018 | 0.841 ± 0.025 |
| *zT* | MAE | 0.098 ± 0.009 | 0.099 ± 0.009 | 0.094 ± 0.007 | 0.088 ± 0.010 |
| | $R^2$ | 0.678 ± 0.068 | 0.668 ± 0.055 | 0.708 ± 0.034 | 0.741 ± 0.065 |

Note that for all representations, context features for temperature are also included. (Power factor – *PF* ($S^2m^{-2}$), Seebeck coefficient – *S* ($\mu VK^{-1}$), Electrical conductivity – σ ($Sm^{-1}$), Thermal conductivity – κ ($Wm^{-1}K^{-1}$), figure-of-merit – *zT*).

and structure embeddings, as well as context features for growth conditions (Fig. 5a). We first benchmarked MMoE with single-task prediction to predict the six properties shown in Table 1. As shown in Fig. 5b, the MMoE results are within error of the single-task results, but show modest improvement by around 5-10% for most cases. MMoE does show notably better model stability, indicated by lower variance in cross-validation performance. The complete single-task and MMoE performance can be found in Supplementary Figs. 2, 3.

Next, we purposed MMoE for multi-task learning of thermoelectric properties. The efficiency of TE energy conversion is given by figure of merit (*zT*) as: $zT = S^2\sigma T/\kappa$, where *S* is Seebeck coefficient, $\sigma$ is electrical conductivity, $\kappa$ is thermal conductivity, and *T* is the temperature. A high *zT* indicates a good thermoelectric, however, the properties that lead to high *zT* are inter-dependent and often conflicting[41]. For example, thermal conductivity increases with electrical conductivity as carrier concentrations approach the degenerate regime. Optimizing for TE performance is thus a challenging task that requires a balance of several properties. For this reason, we speculate that multi-task learning can naturally leverage the TE task correlations for better model performance. We found moderate Pearson correlation ranging from 0.15 to 0.5 between the five TE properties considered here (Supplementary Figure 4), which is considered ideal for multi-task learning. Interestingly, we found that multi-task learning significantly enhances the predictive performance of Seebeck coefficient by 71% compared with single-task prediction, with close performance for the other four tasks within variance from cross-validation (Supplementary Fig. 5).

The accuracy of the property predictions is rooted in the quality of the data representations. In addition to the embeddings derived from language models, we added further information based on context features as model input. Materials science optimization techniques including doping/defect-engineering[42], alloying and phase-boundary mapping[32,43] are widely utilized and critical to enhance the performance of TE materials. The composition of a material after optimization (e.g., doping) is different from the original composition of the host material via the introduction of dopants and other defects. A small degree of doping can substantially affect TE performance. For example, in the experimental database the reported *zT* of PbTe can vary from as low as 0.10 to as high as 1.56 depending on doping/alloying strategy according to the UCSB dataset[44]. Material properties resulting from different synthesis conditions (especially temperature) can vary substantially.

For these reasons, we devised different material representations that can include up to three components: (i) host material structural features, (ii) composition features accounting for doping, phase boundary mapping, and alloying (on normalized chemical formulae), and (iii) context features (one-

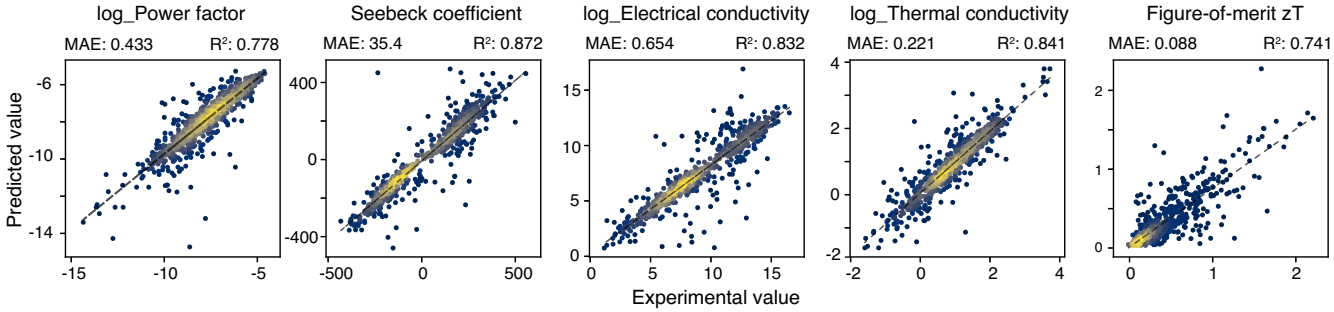


**Fig. 6 | Multi-task prediction results of five TE properties from the best performing MMoE model with 5-fold cross-validation on the experimental dataset.** Composition embeddings, structure embeddings, and context features are concatenated as the model input. Color indicates density.

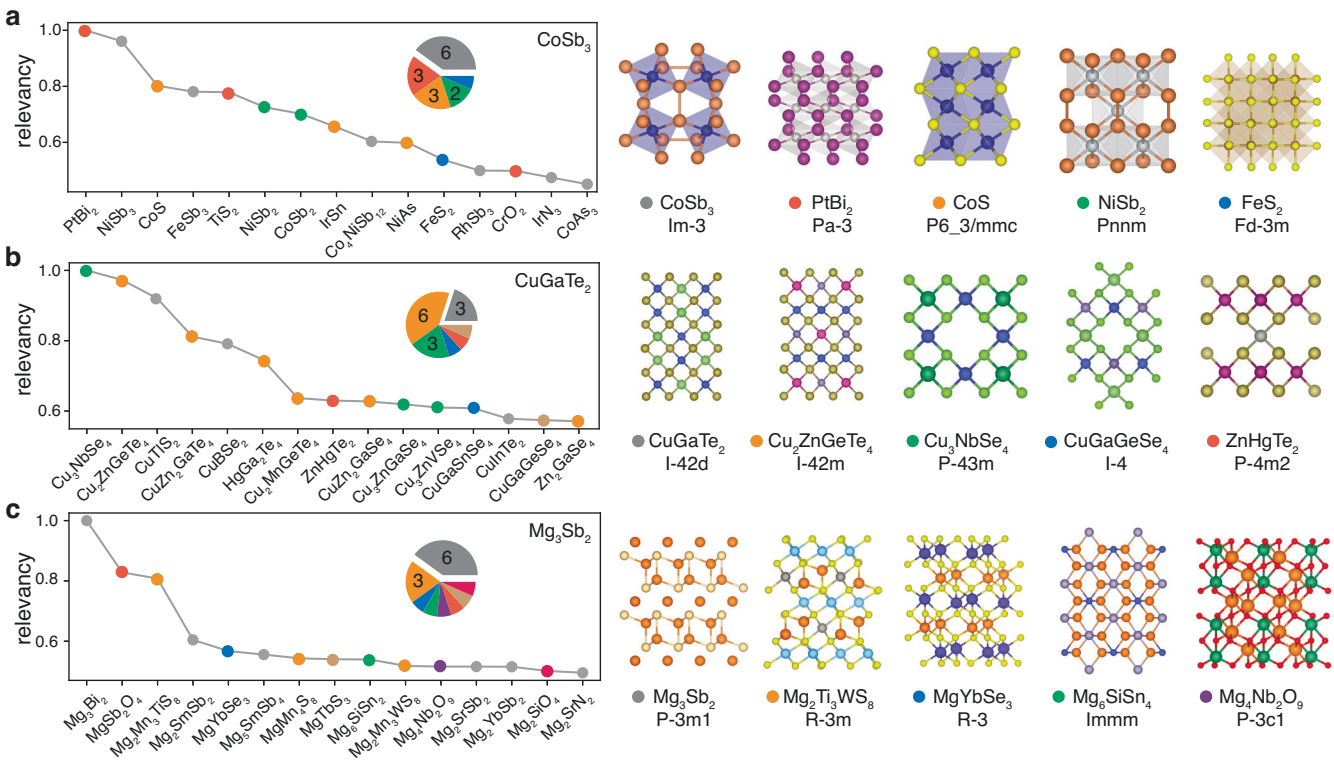


**Fig. 7 | Ranking results of top 15 materials based on similarity.** Materials that exhibit most similar TE potential to (**a**), $CoSb_3$, (**b**) $CuGaTe_2$, and (**c**) $Mg_3Sb_2$ are shown. The color of each data point denotes the structure prototype as shown on the right panel for each query material. The recommended structure prototypes share similar structural features with the query material.

hot encoded temperature). For property prediction, structure embeddings carry important information regarding host material structure, while composition embeddings bring in information about off-stoichiometry. We summarize how different input material representations affect MMoE performance for TE tasks in Table 2. With context features being included, the best performing multi-task TE model was achieved by concatenating doped composition, host structure, embeddings (Table 2), whereas host composition embeddings alone gave the worst performance metrics. While doping and alloying often present significant challenges for first-principles modeling, the language representation accounts for such material complexity naturally, through the contextual knowledge contained in the embedding. In general, models trained with both structure and composition representations perform consistently better than those with only composition embeddings. Therefore, modeling TE properties requires accurate representations of both structural and doped compositions, which can be effectively extracted through BERT-based language models. The multi-task learning results from the best-performing material representation and MMoE is shown in Fig. 6. In all five prediction tasks, MMoE accurately predicts the TE properties for the input material under each one-hot encoded temperature category with $R^2 > 0.7$. Despite being trained directly on general representations of crystals, this model achieves comparable accuracy to recent domain-specific models in the TE field[45,46]. It additionally shows significant enhancement over the MMoE model architecture on baseline representation fingerprints (Supplementary Fig. 6).

### Search ranking of TE materials with similar potential

To interpret and evaluate the ranking performance, we demonstrated the ranking outcomes from the recommendation framework on seven representative TE materials. Candidates were ranked by their *relevancy score* (Ranking score and exploratory analysis), which is defined as the reciprocal of the summed absolute percent difference of five properties from the query material. Figure 7 shows the ranking results for $CoSb_3$, $CuGaTe_2$, and

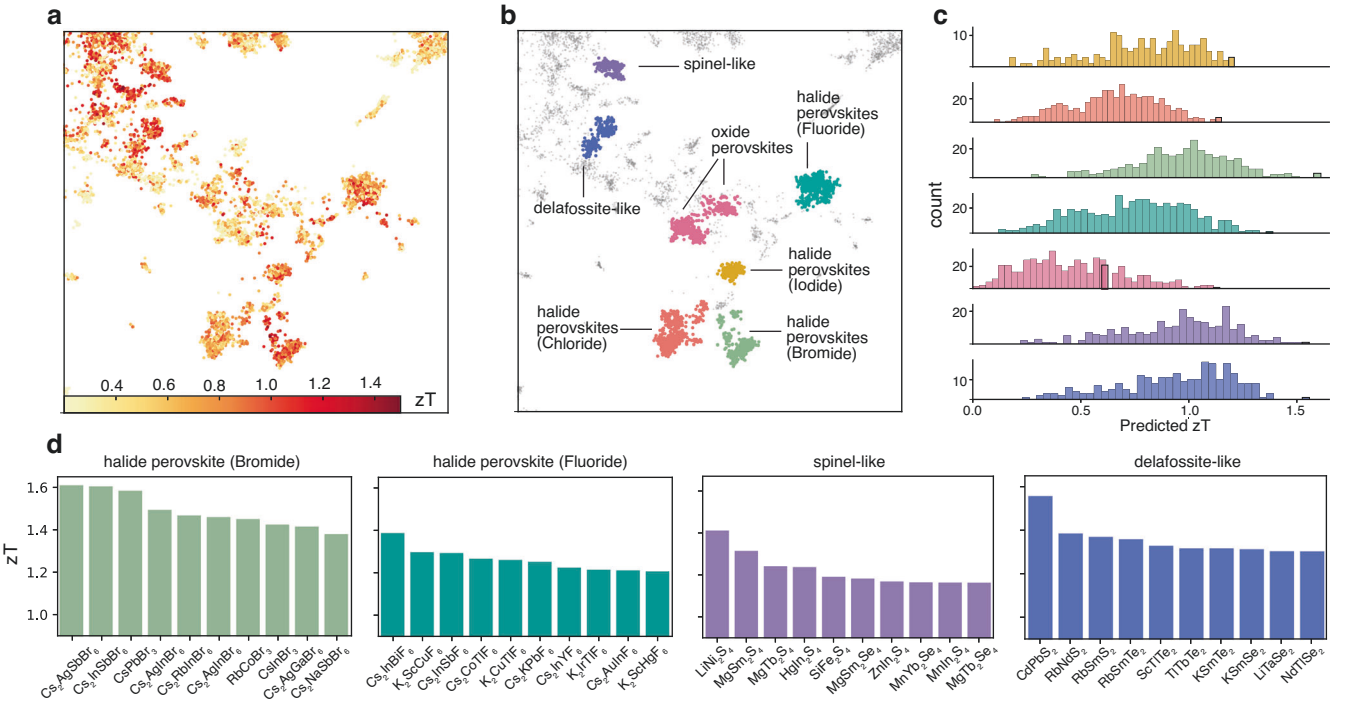


**Fig. 8 | Thermoelectric material exploration and predicted performance.** **a** Zoomed-in region of UMAP projected embedding space as shown by the dashed line in Fig. 3, colored by predicted *zT* from the MMoE model. **b** Groups of materials that exhibit high predicted *zT* are colored in the UMAP. **c** Distributions of the predicted *zT* from each material group in (**b**). **d** Top 10 candidates ranked by their predicted *zT* from each material group.

$Mg_3Sb_2$, representing skutterudite, diamond-like semiconductors (DLS), and Zintl phases. For each query material, the top 15 ranked materials that exhibit the most similar TE potential are shown. A full list of the search ranking results for the other materials (PbTe, BiCuSeO, $Cu_2Se$, $Bi_2Te_3$) can be found in Supplementary Figure 7.

In Fig. 7, each candidate is colored by its structure prototype to visualize the structural diversity. The distribution of prototype structures is shown by the pie chart. For skutterudite $CoSb_3$, the top 15 recommendations consist of 5 prototype structures, and 9 out of the 15 top are different from the space group of $CoSb_3$ ($I\bar{m}3$, No. 204). As expected, several $AX_3$ skutterudites (grey in Fig. 7a) appear in the list, sharing the same prototype structure with $CoSb_3$. Several novel structure prototypes also appear, including pyrite ($P\bar{a}3$, No. 205, red in Fig. 7a) and marcasite structures (*Pnnm*, No. 58, green in Fig. 7a), as close relevant TE materials to $CoSb_3$. Two more prototype structures – covellite ($P6_3/mmc$, No. 194, orange) and carrollite-like $AX_2$ structure ($F\bar{d}3m$, No. 227, blue) are also recommended, both of which have received limited attention historically but may warrant further investigation[47]. Note that all above structure prototypes have corner-sharing octahedral motifs, a local structural feature shared with query material $CoSb_3$ that may correlate to similar TE properties. The recommendations based on querying of diamond-like chalcopyrite material $CuGaTe_2$ ($I\bar{4}2d$, No. 122, grey in Fig. 7b) render diversified outcomes with 5 different structure prototypes. In addition to four more $ABX_2$ chalcopyrites, the framework selected quaternary stannite ($I\bar{4}2m$, No. 121, orange in Fig. 7b), sulvanite ($P\bar{4}3m$, No. 215, green in Fig. 7b), defect kesterite ($\bar{I}4$, No. 79, blue in Fig. 7b), and chalcopyrite-like ($P\bar{4}m2$, No. 115, red in Fig. 7b) structures. For Zintl phase $Mg_3Sb_2$, the top 15 recommendations comprise 8 unique prototype structures (5 of which are shown in Fig. 7c). Interestingly, the prototypes do not exhibit the layered structure of query material $Mg_3Sb_2$. Instead, the common local structural feature of octahedral motifs is present throughout the recommended prototypes. Unlike other computational materials discovery strategies which generate candidate materials by applying chemical substitutions to a single prototype structure[48,49], the framework is able to suggest candidates with diversified structures that are different from, but still related to, the prototype. Such capability can offer insights and understanding of structural similarity between different prototypes and structure-to-property mappings for ML tasks.

To evaluate the performance of the ranking tasks, we performed first-principles calculations on the TE properties of top recommended candidates (see computational details in 4.7). As shown in Supplementary Figure 11, the calculated properties of the recommended materials resemble those of the query material. For example, both $CuGaTe_2$ and its top ranked candidates exhibit high *p*-type TE performance that outperforms the *n*-type counterparts. Upon experimental evaluation of several top ranked candidates, we identified $CuZn_2GaTe_4$ as a *p*-type TE material with high Seebeck coefficient ($S = 250\,\mu VK^{-1}$ at 575 K) and relatively low thermal conductivity ($\kappa = 1.82\ Wm^{-1}K^{-1}$ at 575 K), see Supplementary Fig. 12. This immediate positive result arose from self-doping that yielded a Hall carrier concentration near $4.5 \times 10^{19}\ cm^{-3}$ at 473 K. Compared with query material $CuGaTe_2$ ($S = 320\,\mu VK^{-1}$ [50], $\kappa = 2.5\ Wm^{-1}K^{-1}$ [50] at 575 K), $CuZn_2GaTe_4$ exhibits overall more favorable thermal transport and comparable Seebeck coefficients. Preliminary experimental measurements on other candidates, while not demonstrating good performance immediately, revealed individual features that are beneficial to TEs and have the potential to achieve good performance upon further optimization. The most important features are, e.g., strongly suppressed thermal conductivities at room temperature of $HgGa_2Te_4$ ($0.36\ Wm^{-1}K^{-1}$) and $CuGaGeSe_4$ ($0.62\ Wm^{-1}K^{-1}$). $Cu_2ZnGeTe_4$ also belongs to the group of $CuGaTe_2$-like materials and was previously reported to show decent mobility (ca. 30 $cm^2V^{-1}s^{-1}$ at 550 K,[51]. Lastly, $TiS_2$ suggested as relevant to $CoSb_3$, has previously been reported with large Seebeck and $zT = 0.3$ at 700 K[52]. For future study, we recommend further investigation on the candidates with the highest predicted figure of merit *zT* from our ab initio analysis. Alternatively, one may also consider ab initio calculations on other compounds that emerged from the recommendation approach.

### Exploration of under-studied materials in the representation space

Through material language representations, we noticed that the distributions of materials for both known and predicted high *zT* appear within the

same "band"-like region in the UMAP (Fig. 3d). Despite their high predicted $zT$, materials at the bottom right corner of the "band" (shown by the dashed grey box in Fig. 3) are under-explored with no records from the experimental datasets (Fig. 8a). In that region, we have identified high-$zT$ clusters composed of halide perovskites (fluoride, chloride, bromide, and iodide), oxide perovskites, spinel-like, and delafossite-like structures, as labeled in Fig. 8b. The distributions of predicted $zT$ for all materials in each cluster are visualized in Fig. 8c. Among halide perovskites, bromides have the highest predicted $zT$ with a mean above 1.0, while fluorides, chlorides, and iodides are close to each other in the predicted $zT$ distributions. The top 10 highest $zT$ candidates from bromide and fluoride perovskite clusters (Fig. 8d) are mostly Cs- and K-containing double perovskites $A_2BB'X_6$, with a few single perovskites $ABX_3$. The high TE performance of halide perovskites can likely be attributed to low thermal conductivity. Ref. [53] revealed that inorganic halide perovskites exhibit ultra-low thermal conductivity due to a unique cluster rattling mechanism, resulting in thermal conductivities comparable to the amorphous limit. One of the top predicted candidates $CsPbBr_3$, in particular, has attracted wide attention in the TE field[54]. Recent first-principles calculations[55,56] support the findings of high TE performance for several double perovskites. It is worth noting that all of the recommended candidates are lead-free and have high-temperature stability[57], good oxidation resistance, and lower processing costs. With recent experimental advances in improving the stability of perovskites[58,59], halide perovskites may become more appealing TE candidates. Oxide perovskites, interestingly, show inferior TE performance to halide counterparts. This observation aligns with chemical intuition, since oxides are in general more ionic and insulating materials, rendering them hard to dope to optimize the TE power factor.

Delafossite-like and spinel-like structures are also under-explored structural spaces with potential to host TE candidates. Unlike perovskites which form isolated clusters in the representation space, these two material groups neighbor the more well-explored chalcopyrites and $AB_2X_2$ Zintl phases. Delafossite-like structures, also known as caswellsivlerites, refer to $ABX_2$ (X = O, S, Se, Te) materials crystallizing in the trigonal structure, with $CuFeO_2$ ($R\bar{3}$ m, No.166) being the prototype. So far, extensive research efforts have focused on delafossite-type oxides as TE materials[60,61], while a recent high-throughput computational study[62] revealed that sulfide, selenide and telluride delafossite-like structures are also thermodynamically stable. For the top 10 predicted $zT$ candidates from the delafossite group (Fig. 8d), all candidates are sulfides and selenides that we recommend for further investigation. Similarly, less attention has been focused on sulfide, selenide, and telluride spinels compared to oxide counterparts, while all top 10 spinels from the recommended list (Fig. 8d) are sulfides and selenides. Recent theoretical works have suggested several high-performance spinel sulfides $Tm_2MgS_4$ ($zT \sim 0.8$[63]), $Y_2CdS_4$ ($zT \sim 0.8$,[64]), $MgIn_2Se_4$ ($zT \sim 0.7$,[65]), suggesting that the discovery framework is able to select good candidates from a large and diverse search space. As evaluation of the recommended under-explored material groups, we performed first-principles calculations on the top candidates from each group in Fig. 8d, which are summarized in Supplementary Fig. 11. These calculations corroborated the promising TE potential of several candidates, e.g. delafossite-like $CdPbS_2$ ($n$-type, $zT_{max}$=1.7 at 800 K), halide perovskite $Cs_2InSbF_6$ ($n$-type, $zT_{max}$=1.0 at 800 K), etc.

## Discussion

While representation learning has facilitated extraction of more meaningful features from large unlabeled data, methods for learning material representations have also gained substantial momentum for predition tasks[66–68]. On the other hand, language-based models have achieved remarkable outcomes in prediction and generation tasks across an extensive array of domain areas. In this work, we demonstrated the use of language representations in the inorganic crystalline materials domain. Specifically, we introduced a language-based framework to extract composition and structure embeddings as material representations via pretrained language models. The discovery framework is designed to be task-agnostic. We anticipate that it can be expanded upon and utilized to search and explore vast chemical and structural spaces, towards functional materials design and discovery.

Representing materials in the format of natural language enables effective utilization of materials science knowledge learnt from ever-growing unstructured scientific texts. Indeed, the extracted embeddings form a chemically meaningful representation space without task-specific supervision. We find that knowledge can be extracted from representations by unsupervised recall on embedding vectors and supervised neural networks, together enabling the funnel-based approach of Fig 1. In particular, the recall step allows reliable recommendation by constraining the ranking on candidates that are similar to the query material in the representation space. A benefit of such pre-screening is the avoidance of common pitfalls where materials exhibit similar properties for inherently different reasons, i.e. far from each other in the representation space. For the use case of thermoelectrics, for example, high $zT$ can arise from either high power factor or low thermal conductivity. Another strength of language representations is that they can effectively handle off-stoichiometric material compositions to account for alloying and doping, which typically require complicated computational techniques (e.g., disorder modelling) for accurate predictions in first-principles simulations.

An interesting consideration is model performance for embedding models trained on materials science specific language models, compared to those trained on general Large Language Models (LLMs). While both encoder-only and decoder-only large language models (LLMs) can generate language representations, we believe that BERT-based models may outperform them for the following reasons. Unlike decoder-only models (e.g. GPT) which are tailored for sequence generation tasks, encoder-only models (e.g., BERT) are generally more versatile for understanding and encoding information due to the bi-directional context and masked language modeling training technique. In the context of recommendation tasks, encoder-only LLMs can generate more effective representations for downstream tasks like material property prediction or similarity analysis. As has been verified by Trewartha et al.[16], language models with extended materials science knowledge tend to perform better on materials science-related tasks, with the simpler model consistently outperforming models with more parameters. Even so, GPT-based models show promise for certain chemistry applications in the small data limit[69].

Exploitation and exploration trade-off has been a common phenomenon in recommender systems[70,71]. For the recommendation framework, while exploitation refers to seeking maximum reward, exploration may be thought of as consideration of new structural prototypes present in the top-ranked candidates that share structural features with, but are distinct from, the query material. A reasonable balance between exploitation and exploration, which can be tuned by the number of candidates recalled from the candidate generation step, will diversify the recommendation while still proposing structurally-related materials. For example, the top-15 ranked materials for both $CoSb_3$ and $CuGaTe_2$ each contain 5 different prototype structures when 100 recalled materials from the candidate generation step are considered for ranking, while the number of prototypes increases to 14 and 9 respectively if the number of recalled materials considered is increased to 1000 (Supplementary Fig. 8).

As future directions for language representation for crystals, we suggest to enrich the material representations by diversifying both text-based input and structures. The automatically generated text descriptions from Robocrystallographer are repetitive with subtle differences for similar materials. The descriptions are monotonic with little variation between descriptive words/phrasing[72]. We suggest that including more detailed structural descriptions (off-centering, bond distortions, etc) and including different writing styles from multiple sources could help improve performance. For instance, these descriptions can possibly be diversified via paraphrasing or developing structure to sentence machine translation models to describe crystal structures in text or using crystal structure descriptions from

published papers. The structural complexity in the representation space can be diversified via generative models, e.g., diffusion models[73,74], to design new prototype structures beyond simple lattice decoration of known crystals.

## Methods

### Data preparation

The training dataset was collected from the Materials Project[75] to include 116,216 materials that are possible to be thermodynamically stable. Using decomposition enthalpy < 0.5 eV as a query criteria, we utilized Materials Project API[76] and Pymatgen[77] library to collect materials for use in this study.

In this work we considered five different property datasets, all of which include properties relevant to thermoelectric materials; UCSB dataset[44] – an experimental dataset from Materials Research Laboratory (MRL) about 1092 materials (500 unique materials) with their thermoelectric properties; ESTM dataset[46] – an experimental dataset containing 5205 materials (880 unique materials) with their thermoelectric properties; ChemExtracter dataset[78] – a mixture of experimental and theory dataset by auto-generation from the scientific literature spanning 10,641 unique chemical names; TEDesignLab dataset[79] – a theory dataset containing lattice thermal conductivity for 3278 materials; Citrine dataset[80] – an experimental dataset from Matminer[80] containing thermal conductivity records for 871 materials. In all five datasets, 826 materials that have records for five TE properties are used for evaluation of recall performance in Evaluation: unsupervised recall of relevant materials. We calculated the numeric mean for materials with repeated entry for certain properties and properties at different temperatures. For MMoE model training and testing, UCSB and ESTM datasets are utilized as ground-truth labels. During training, the TE properties are matched to corresponding temperature range via one-hot encoding.

### Embedding models

Three model-based and one model-free embedding methods were used in this work. For the model-based approach, we obtained pretrained weights for Mat2Vec[15], MatsciBERT[17], and MatBERT[16]. For Mat2Vec, it was trained similarly as Word2vec training through skip-gram with negative sampling. Each word is embedded into a 200-dimensional vector. For the BERT-based models, MatsciBERT was pretrained on whole sections of more than 1 million materials science articles, whereas MatBERT was trained by sampling 50 million paragraphs from 2 million articles. Both models were trained with masked language modeling (15% dynamic whole word masking) and next-sentence prediction as the unsupervised training objectives. Both models are uncased, and have maximum 512 input token size with 768 hidden dimensions. The vocabulary size for the tokenizer is 30,522. For the fingerprint generation, it was generated using CrystalNN[28] algorithm as implemented in Matminer[80] package. The fingerprint contains statistical information about local motifs with a size dimension of 122.

### Material language representations

We acquired compositional and structural level representations for 116K materials in total. To acquire structural level representations for each individual material, we applied robocrystallographer[26], an open-source toolkit that converts the material structure into a human-readable text passage describing local, semi-local and global structural features of the given material. We used robocrystallographer descriptions from[72]. Similar to material descriptions found in literature, such material passage encodes naturally interpretable structural information. The whole passage is processed by tokenizers and fed into the pretrained BERT models (MatsciBERT and MatBERT) for output embeddings from hidden layers. The output embeddings are $L$ by 768 dimensional matrix, where $L \in (0512]$ is the total number of tokens within the passage. We partitioned passages with more than 512 tokens to fit the maximum input token size. The final embeddings for each material are constructed by averaging output embeddings across all tokens, resulting in a fixed length of vector representations with 768 dimensions.

For the compositional level representations, Mat2Vec embeddings are directly obtained as the 200-dimensional word embedding vectors of the material formulas. With BERT models, we performed same tokenization and embedding procedures on material formulas only. This results in the same number of 768-dimensional embedding vectors but only contains information related to the material composition. For composition embeddings of the doped material formulas (UCSB dataset), we normalized the compositions to the element with the most number of atoms in the unit cell. The output embeddings are obtained on the normalized formulas.

To quantify the strength of contextual language representation used in this work, it is essential to choose reasonable baseline representations that capture the structural and/or compositional features of solid-state materials for comparison. Despite the existence of various representations of solid-state materials[81], not all can serve as the baseline representations for this study because: (1) the language representation utilized in this work contains latent features that are directly obtained from pretrained models, (2) the language representation serves as the input to unsupervised downstream tasks (e.g., recall) in our recommendation framework. For the reasons above, we select fingerprint as the baseline for structural representation which characterizes the local environment of each atom and considers the global structure as a combination of local representations. On the other hand, we employ Mat2Vec as the chosen baseline for compositional representation.

### MMoE and TE property prediction

A shared-bottom multi-task network was first introduced by[38] and widely applied for multi-task learning. The basic network formulation is:

$$y_k = h^k(f(x)) \tag{1}$$

where $k = 1, 2, 3 \ldots K$ for $K$ number of tasks, $f$ is the shared-bottom network, $h^k$ is the tower network for task $k$, and $y_k$ is the output for task $k$. The key difference in MMoE network is to substitute the shared-bottom $f$ with MoE layer $f^k(x)$ for a specific task $k$, which is defined as:

$$f^k(x) = \sum_{i=1}^{n} g^k(x)_i f_i(x) \tag{2}$$

$$g^k(x) = \mathrm{softmax}(W_{gk}x) \tag{3}$$

where $i = 1, 2, 3 \ldots n$ for $n$ number of experts, $g^k(x)$ is the gating network for each task $k$, and $W_{gk}$ is the trainable matrix. In this implementation, all expert network is a three-layered MLP with 128, 64, and 32 dimensions. The gating network is a two-layered MLP with 32 and 16 dimensions. In all experiments, networks are trained for 500 epochs with learning rate = $10^{-3}$, weight decay = $10^{-5}$, and batch size = 64. We used k-fold cross-validation method to train and evaluate the model performance. For all datasets, we employed 5-fold cross validation by splitting the dataset into 5 nonoverlapping portions. The number of experts is set to 8 for both AFLOW benchmark dataset and TE dataset.

Recent works[45,46] reported that doping and alloying information, as well as context features greatly enhance the model performance for TE predictions. As for context features for MMoE, we first sorted the continuous temperatures into four ranges (0, 300], (300, 600], (600, 900], (900, + ∞], which were one-hot encoded into sparse feature vectors and passed to embedding layers of the MMoE model. Since the structure embeddings are restricted to the host materials, dopant or alloying information will be derived from composition embeddings to delineate the compositional effect. To match doped materials to their hosts, we encoded the normalized doped formulas into composition vectors (sparse vector with number of corresponding elements at each site), followed by mapping to existing host composition vectors via cosine similarity. Host materials with the highest cosine similarities were selected.

### Ranking score and exploratory analysis

Once candidates are recalled for the query, their predicted properties are used to compute total absolute percent difference (TAPD) defined as:

$$\text{TAPD} = \sum_{k=1}^{K} \left( \frac{|y_k^c - y_k^q|}{y_k^q} \right) \tag{4}$$

where $K$ is the total number of material properties, $y^c$ and $y^q$ are the candidate and query properties respectively. This measures the composite deviation of candidate properties from the query properties. All properties need to be close to those of the query to have a low TAPD. We define *relevancy score* as the reciprocal of TAPD:

$$\text{relevancy} = \frac{1}{\text{TAPD}} \tag{5}$$

In these experiments, 100 candidates were recalled per query material. We ranked the candidates based on their *relevancy score*. The scores presented in the figure were normalized by the maximum score within the recalled list. For the exploratory analysis, clusters were hand-selected based on localization of materials with high predicted $zT$. Within each selected cluster, we extracted and ranked the materials according to their $zT$. All predictions were made at high temperature $(900, +\infty]$ as the context features.

### Evaluation: unsupervised recall of relevant materials

Recalling relevant material candidates is an unsupervised process which does not require training labels. First, candidates are searched in the representation space by computing cosine similarities between the embedding vector of the query and the rest of the embedding vectors. The similarity-sorted top candidates are returned as the relevant materials. Metrics including Precision@$k$ and Normalized Discounted Cumulative Gain (nDCG) are used to evaluate the recall performance. Such evaluation metrics are common for recommender system, where the goal is to maximize the number of relevant items in the recalled list, i.e., the top@k items with $k$ being the size of the list, as well as the relative order of recalled items. Precision@$k$ measures the percentage of the relevant materials in the first $k$ recalled materials:

$$\text{precision@}k = \frac{\text{relevant items@}k}{k} \tag{6}$$

while nDCG is an evaluation method which compares the ideal ranking of a test set (iDCG), with the ranking assigned by the recommendation algorithm (DCG – Equ. (7)).

$$\text{DCG} = \sum_{i=1}^{n} \frac{\text{relevance}}{\log_2(i+1)} \tag{7}$$

$$\text{nDCG} = \frac{\text{DCG}}{\text{iDCG}} \tag{8}$$

### Evaluation: First-principles calculations

The ab initio scattering and transport (AMSET[82] software package was used to estimate scattering rates (or lifetime) and transport properties based on momentum relaxation time approximation (MRTA), which has been shown to give comparable results to state-of-art EPW code[83]. The carrier mobility was simulated by considering three scattering processes, including acoustic phonon scattering (ADP), polar optical phonon scattering (POP), and ionized impurity scattering (IMP). Each component of carrier lifetime was evaluated by Fermi's golden rule, with total characteristic scattering time following Matthiessen's rule. The associated Seebeck coefficient, electrical conductivity, and electronic component of the thermal conductivity were were calculated by solving the Boltzmann transport equation (BoltzTraP) using Onsager transport coefficients. All ab initio inputs are computed from density functional theory (DFT) using the GGA-PBE[84] exchange-correlation functional. Lattice thermal conductivity ($\kappa_L$) was calculated using a semi-empirical model based on a modified Debye-Callaway model[85] which captures anharmonicity. Bulk modulus ($B$) was determined by fitting the Birch-Murnaghan equation of state to a set of total energies computed at different volumes that were expanded and contracted around the equilibrium volume. Other parameters of the semi-empirical model are directly accessible from the relaxed structures, including density, average atomic mass, volume per atom, average coordination number, and number of atoms in the primitive cell. The expression for lattice thermal conductivity is give by

$$\kappa_{\text{L,ac}} = A_2 \frac{\bar{M} v_s^y}{T V^{2/3} n^{1/3}} + A_3 \frac{v_s}{V^{2/3}} \left( 1 - \frac{1}{n^{2/3}} \right), \tag{9}$$

where $A_1$ and $A_2$ are fitted parameters, $\bar{M}$ is the average atomic mass, $v_s$ is the speed of sound, $T$ is the temperature, $V$ is the volume per atom, and $n$ is the number of atoms in the primitive cell. $v_s$ is approximated as $v_s \sim (B/d)^{\frac{1}{2}}$.

### Evaluation: experiments

$CuZn_2GaTe_4$, $CuGaGeSe_4$, and $HgGa_2Te_4$ samples were prepared from elements: Cu (99.9%), Hg (99.999%), Ga (99.999%), Zn (99.999%), Ge (99.999%), and Te (99.999%), Se (99.999%). The stoichiometric weights were first sealed in evacuated silica ampoules and melted at 1000° C for several hours. Next, the ingots were milled in high-energy mechanical mill Spex 8000D for 90 min in an inert environment. The powders were consolidated in an induction heating hot press at 500° C, 40 MPa for at least 2 hours. Electrical resistivity and Hall coefficient were studied under vacuum on a home-built apparatus with van-der Pauw geometry[86]. Seebeck coefficient measurements were carried out using a custom-built device[87] in 300 Torr of nitrogen gas. Diffusivity coefficient ($D$) measurements were performed on Netzsch LFA 467 apparatus. To obtain thermal conductivity ($\kappa$), we used formula $\kappa = DC_p d_{\text{exp}}$, where $C_p$ is heat capacity and $d_{\text{exp}}$ is experimental density. Values of $C_p$ were obtained from Dulong-Petit law, while density of the samples was measured with geometric method. For all obtained materials $d_{\text{exp}}$ was ca. 90% of the theoretical value or higher.

## Data availability

The preprocessed AFLOW and thermoelectric datasets used for training and testing the models, as well as material embeddings obtained in this work, are available at https://doi.org/10.6084/m9.figshare.22718668.v1.

## Code availability

The code and the model weights are available under the MIT license at: https://github.com/ertekin-research-group/Material_Recommender.

Received: 31 August 2023; Accepted: 17 February 2024;
Published online: 21 March 2024

## References

1. Meng, Y. S. & Arroyo-de Dompablo, M. E. Recent advances in first principles computational research of cathode materials for lithium-ion batteries. *Acc. Chem. Res.* **46**, 1171–1180 (2013).
2. Hautier, G., Jain, A. & Ong, S. P. From the computer to the laboratory: materials discovery and design using first-principles calculations. *J. Mater. Sci.* **47**, 7317–7340 (2012).
3. Brunin, G., Ricci, F., Ha, V.-A., Rignanese, G.-M. & Hautier, G. Transparent conducting materials discovery using high-throughput computing. *Npj Comput. Mater.* **5**, 63 (2019).
4. Aykol, M., Herring, P. & Anapolsky, A. Machine learning for continuous innovation in battery technologies. *Nat. Rev. Mater.* **5**, 725–727 (2020).

5. Wang, T., Zhang, C., Snoussi, H. & Zhang, G. Machine learning approaches for thermoelectric materials research. *Adv. Funct. Mater.* **30**, 1906041 (2020).
6. Mahmood, A. & Wang, J.-L. Machine learning for high performance organic solar cells: current scenario and future prospects. *Energy Environ. Sci.* **14**, 90–105 (2021).
7. Schmidt, J., Marques, M. R., Botti, S. & Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. *Npj Comput. Mater.* **5**, 83 (2019).
8. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
9. Isayev, O. et al. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* **8**, 15679 (2017).
10. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
11. Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **31**, 3564–3572 (2019).
12. Choudhary, K. & DeCost, B. Atomistic line graph neural network for improved materials property predictions. *Npj Comput. Mater.* **7**, 185 (2021).
13. Yan, K., Liu, Y., Lin, Y. & Ji, S. Periodic graph transformers for crystal material property prediction. *Adv. Neural. Inf. Process. Syst.* **35**, 15066–15080 (2022).
14. Jha, D. et al. Element: Deep learning the chemistry of materials from only elemental composition. *Sci. Rep.* **8**, 1–13 (2018).
15. Tshitoyan, V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
16. Trewartha, A. et al. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns* **3**, 100488 (2022).
17. Gupta, T., Zaki, M. & Krishnan, N. A. Matscibert: A materials domain language model for text mining and information extraction. *Npj Comput. Mater.* **8**, 102 (2022).
18. Wang, Z. et al. Dataset of solution-based inorganic materials synthesis procedures extracted from the scientific literature. *Sci. Data* **9**, 231 (2022).
19. Gaultois, M. W. et al. Perspective: Web-based machine learning models for real-time screening of thermoelectric materials properties. *APL Mater.* **4**, 053213 (2016).
20. Seko, A., Hayashi, H. & Tanaka, I. Compositional descriptor-based recommender system for the materials discovery. *J. Chem. Phys.* **148**, 241719 (2018).
21. Seko, A., Hayashi, H., Kashima, H. & Tanaka, I. Matrix-and tensor-based recommender systems for the discovery of currently unknown inorganic compounds. *Phys. Rev. Mater.* **2**, 013805 (2018).
22. Hayashi, H., Hayashi, K., Kouzai, K., Seko, A. & Tanaka, I. Recommender system of successful processing conditions for new compounds based on a parallel experimental data set. *Chem. Mater.* **31**, 9984–9992 (2019).
23. Covington, P., Adams, J. & Sargin, E. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, 191–198 (2016).
24. Gomez-Uribe, C. A. & Hunt, N. The Netflix recommender system: Algorithms, business value, and innovation. *ACM Transac. Manag. Info. Syst. (TMIS)* **6**, 1–19 (2015).
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. Ganose, A. M. & Jain, A. Robocrystallographer: automated crystal structure text descriptions and analysis. *MRS Commun.* **9**, 874–881 (2019).
27. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *Preprint at* https://arxiv.org/abs/1802.03426 (2018).
28. Zimmermann, N. E. & Jain, A. Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity. *RSC Adv.* **10**, 6063–6081 (2020).
29. Tang, J. et al. Manipulation of band structure and interstitial defects for improving thermoelectric snte. *Adv. Funct. Mater.* **28**, 1803586 (2018).
30. Wu, M. et al. Significantly enhanced thermoelectric performance achieved in cugate2 through dual-element permutations at cation sites. *ACS Appl. Mater. Interfaces* **14**, 30046–30055 (2022).
31. Zhang, J. et al. Discovery of high-performance low-cost n-type mg3sb2-based thermoelectric materials with multi-valley conduction bands. *Nat. Commun.* **8**, 13901 (2017).
32. Ohno, S. et al. Phase boundary mapping to obtain n-type mg3sb2-based thermoelectrics. *Joule* **2**, 141–154 (2018).
33. Curtarolo, S. et al. Aflow: An automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **58**, 218–226 (2012).
34. Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *Npj Comput. Mater.* **6**, 138 (2020).
35. Xiao, J., Wang, M., Jiang, B. & Li, J. A personalized recommendation system with combinational algorithm for online learning. *J. Ambient. Intell. Humaniz. Comput.* **9**, 667–677 (2018).
36. Song, L., Tekin, C. & Van Der Schaar, M. Online learning in large-scale contextual recommender systems. *IEEE Trans. Serv. Comput.* **9**, 433–445 (2014).
37. Ma, J. et al. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 1930–1939 (2018).
38. Caruana, R. Multitask learning. *Mach. Learn.* **28**, 41–75 (1997).
39. Sanyal, S. et al. Mt-cgcnn: Integrating crystal graph convolutional neural network with multitask learning for material property prediction. *Preprint at* https://arxiv.org/abs/1811.05660 (2018).
40. Chang, R., Wang, Y.-X. & Ertekin, E. Towards overcoming data scarcity in materials science: unifying models and datasets with a mixture of experts framework. *Npj Comput. Mater.* **8**, 242 (2022).
41. Snyder, G. J. & Toberer, E. S. Complex thermoelectric materials. *Nat. Mater.* **7**, 105–114 (2008).
42. Toriyama, M. Y., Qu, J., Snyder, G. J. & Gorai, P. Defect chemistry and doping of bicuseo. *J. Mater. Chem. A* **9**, 20685–20694 (2021).
43. Ortiz, B. R. et al. Carrier density control in cu 2 hggete 4 and discovery of hg 2 gete 4 via phase boundary mapping. *J. Mater. Chem. A* **7**, 621–631 (2019).
44. Gaultois, M. W. et al. Data-driven review of thermoelectric materials: performance and resource considerations. *Chem. Mater.* **25**, 2911–2920 (2013).
45. Na, G. S., Jang, S. & Chang, H. Predicting thermoelectric properties from chemical formula with explicitly identifying dopant effects. *Npj Comput. Mater.* **7**, 106 (2021).
46. Na, G. S. & Chang, H. A public database of thermoelectric materials and system-identified material representation for data-driven discovery. *Npj Comput. Mater.* **8**, 214 (2022).
47. Mukherjee, B., Tarachand, T., Hussain, S. & Okram, G. S. Effect of different surfactants on thermoelectric properties of cus nanoparticles. In *AIP Conf.*, vol. 2100, 020091 (AIP Publishing LLC, 2019).
48. Wang, H.-C., Botti, S. & Marques, M. A. Predicting stable crystalline compounds using chemical similarity. *Npj Comput. Mater.* **7**, 12 (2021).
49. Qu, J., Stevanovic, V., Ertekin, E. & Gorai, P. Doping by design: finding new n-type dopable abx 4 zintl phases for thermoelectrics. *J. Mater. Chem. A.* **8**, 25306–25315 (2020).

50. Plirdpring, T. et al. Chalcopyrite cugate2: a high-efficiency bulk thermoelectric material. *Adv. Mater.* **24**, 3622–3626 (2012).
51. Ortiz, B. R. et al. Ultralow thermal conductivity in diamond-like semiconductors: selective scattering of phonons from antisite defects. *Chem. Mater.* **30**, 3395–3409 (2018).
52. Bourges, C. et al. Thermoelectric properties of tis2 mechanically alloyed compounds. *J. Eur. Ceram. Soc.* **36**, 1183–1189 (2016).
53. Lee, W. et al. Ultralow thermal conductivity in all-inorganic halide perovskites. *Proc. Natl. Acad. Sci. USA* **114**, 8693–8697 (2017).
54. Yan, L., Wang, M., Zhai, C., Zhao, L. & Lin, S. Symmetry breaking induced anisotropic carrier transport and remarkable thermoelectric performance in mixed halide perovskites cspb (i1–x br x) 3. *ACS Appl. Mater. Interfaces* **12**, 40453–40464 (2020).
55. Mahmood, Q. et al. Study of lead-free double perovskites halides cs2ticl6, and cs2tibr6 for optoelectronics, and thermoelectric applications. *Mater. Sci. Semicond.* **137**, 106180 (2022).
56. Saeed, M. et al. First-principles prediction of the ground-state crystal structure of double-perovskite halides cs2agcrx6 (x= cl, br, and i). *J. Phys. Chem. Solids* **160**, 110302 (2022).
57. Gao, Z. et al. Screening for lead-free inorganic double perovskites with suitable band gaps and high stability using combined machine learning and dft calculation. *Appl. Surf. Sci.* **568**, 150916 (2021).
58. Niu, G., Guo, X. & Wang, L. Review of recent progress in chemical stability of perovskite solar cells. *J. Mater. Chem. A* **3**, 8970–8980 (2015).
59. Tiep, N. H., Ku, Z. & Fan, H. J. Recent advances in improving the stability of perovskite solar cells. *Adv. Energy Mater.* **6**, 1501420 (2016).
60. Hayashi, K., Sato, K.-i, Nozaki, T. & Kajitani, T. Effect of doping on thermoelectric properties of delafossite-type oxide cucro2. *Jpn. J. Appl. Phys.* **47**, 59 (2008).
61. Hoang, D. V. et al. Effects of multi-scale defects on the thermoelectric properties of delafossite cucr1-xmgxo2 materials. *J. Alloys Compd.* **844**, 156119 (2020).
62. Shi, J. et al. High-throughput search of ternary chalcogenides for p-type transparent electrodes. *Sci. Rep.* **7**, 43179 (2017).
63. Nazar, M. et al. First-principles calculations to investigate structural, magnetic, optical, electronic and thermoelectric properties of x2mgs4 (x= gd, tm) spinel sulfides. *J. Phys. Chem. Solids* **166**, 110719 (2022).
64. Yakhou, H., Maachou, A., Riane, H. & Sahnoun, M. Theoretical investigation of electronic and thermoelectric properties of spinel sulfides a2bs4 (a= sc and y; b= cd and zn). *Comput. Condens. Matter* **21**, e00417 (2019).
65. Mahmood, Q. et al. Opto-electronic and thermoelectric properties of mgin2x4 (x= s, se) spinels via ab-initio calculations. *J. Mol. Graph.* **88**, 168–173 (2019).
66. Xu, M., Wang, H., Ni, B., Guo, H. & Tang, J. Self-supervised graph-level representation learning with local and global structure. In *ICML*, 11548–11558 (PMLR, 2021).
67. Gupta, V. et al. Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data. *Nat. Commun.* **12**, 6595 (2021).
68. Na, G. S. & Kim, H. W. Contrastive representation learning of inorganic materials to overcome lack of training datasets. *Chem. comm.* **58**, 6729–6732 (2022).
69. Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A. & Smit, B. Leveraging large language models for predictive chemistry. *Nat. Mach. Intell.* **6**, 161–169 (2024).
70. Gao, C., Lei, W., He, X., de Rijke, M. & Chua, T.-S. Advances and challenges in conversational recommender systems: A survey. *AI Open* **2**, 100–126 (2021).
71. Vanchinathan, H. P., Nikolic, I., De Bona, F. & Krause, A. Explore-exploit in top-n recommender systems via gaussian processes. In *Proceedings of the 8th ACM Conference on Recommender systems*, 225–232 (2014).
72. Sayeed, H. M., Baird, S. G. & Sparks, T. D. Structure feature vectors derived from robocrystallographer text descriptions of crystal structures using word embeddings. *Preprint at* https://chemrxiv.org/engage/chemrxiv/article-details/640acf476642bf8c8f462235 (2023).
73. Xie, T., Fu, X., Ganea, O.-E., Barzilay, R. & Jaakkola, T. Crystal diffusion variational autoencoder for periodic material generation. https://openreview.net/forum?id=03RLpj-tc_ (International Conference on Learning Representations (ICRL) conference, 2022).
74. Lyngby, P. & Thygesen, K. S. Data-driven discovery of 2d materials by deep generative models. *Npj Comput. Mater.* **8**, 232 (2022).
75. Jain, A. et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
76. Ong, S. P. et al. The materials application programming interface (api): A simple, flexible and efficient api for materials data based on representational state transfer (rest) principles. *Comput. Mater. Sci.* **97**, 209–215 (2015).
77. 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).
78. Sierepeklis, O. & Cole, J. M. A thermoelectric materials database auto-generated from the scientific literature using chemdataextractor. *Sci. Data* **9**, 648 (2022).
79. Gorai, P. et al. Te design lab: A virtual laboratory for thermoelectric material design. *Comput. Mater. Sci.* **112**, 368–376 (2016).
80. Ward, L. et al. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **152**, 60–69 (2018).
81. Damewood, J. et al. Representations of materials for machine learning. *Annu. Rev. Mater. Res.* **53**, 399–426 (2023).
82. Ganose, A. M. et al. Efficient calculation of carrier scattering rates from first principles. *Nat. Commun.* **12**, 1–9 (2021).
83. Poncé, S., Margine, E. R., Verdi, C. & Giustino, F. Epw: Electron–phonon coupling, transport and superconducting properties using maximally localized wannier functions. *Comput. Phys. Commun.* **209**, 116–133 (2016).
84. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
85. Miller, S. A. et al. Capturing anharmonicity in a lattice thermal conductivity model for high-throughput predictions. *Chem. Mater.* **29**, 2494–2501 (2017).
86. Borup, K. A. et al. Measurement of the electrical resistivity and hall coefficient at high temperatures. *Rev. Sci. Instrum.* **83**, 123902 (2012).
87. Iwanaga, S., Toberer, E. S., LaLonde, A. & Snyder, G. J. A high temperature apparatus for measurement of the seebeck coefficient. *Rev. Sci. Instrum.* **82**, 063905 (2011).


## Acknowledgements

This work was funded with support from the U.S. National Science Foundation (NSF) via Grant No. 2118201 (HDR Institute for Data-Driven Dynamical Design) and Grant No. 1922758 (DIGI-MAT). This work was also funded in part by the IBM-Illinois Discovery Accelerator Institute. This work used PSC Bridges-2 at the Pittsburgh Supercomputing Center through allocation MAT220011 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants No. 2138259, No. 2138286, No. 2138307, No. 2137603, and No. 2138296.

## Author contributions

J.Q., Y.R.X., E.S.T., and E.E. conceived and designed this research project. J.Q. and Y.R.X. contributed equally to project conceptualization, investigation, visualization, and formal analysis. J.Q. performed the first-principles DFT calculations. C.E.P. and K.M.C. contributed to experimental

methodology and formal analysis of transport data. E.S.T. and E.E. contributed to project administration. All authors participated in preparing and editing the manuscript.

## Competing interests
The authors declare no competing interests.

## Additional information
**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41524-024-01231-8.

**Correspondence** and requests for materials should be addressed to Jiaxing Qu or Elif Ertekin.