A language-based recommendation system for material discovery

Jiaxing Qu [*] [1] **Yuxuan Richard Xie** [*] [2] **Elif Ertekin** [3]

Abstract

Data-driven approaches for material discovery have been accelerated by emerging efforts in machine learning. We introduce a material discovery framework that uses natural language embeddings derived from pretrained language models as generalized representations of inorganic materials. The discovery framework consists of a joint scheme that first recalls relevant candidates, and next ranks the candidates based on multiple target properties. Leveraging the contextual knowledge encoded in language representations, the discovery framework enables both representational similarity analysis for candidate generation, and multi-task learning to share information across related properties for ranking. Our language-based framework provides a generalized means of embedding structure for effective material recommendation, which is task-agnostic and can be applied to various material systems.

1. Introduction

Rapid growth of data in materials science has opened a data-centric paradigm (Hey et al., 2009) for discovery of novel materials. In this paradigm, machine learning (ML) models trained on large material data sets can computationally screen candidates for field-specific applications. The key objective of the model-driven approach is to identify candidates that exhibit targeted, desirable material properties. Extracting representative features of materials to capture attributes is therefore a key to success of accurate model performance and property prediction. Conventionally, material feature extraction has consisted of hand-crafted descriptors that contain essential information related to composition and crystal structure, relying on physical and mathematical intuition (Schmidt et al., 2019; Behler & Parrinello, 2007; Isayev et al., 2017). Until recently, materials' atomic structures have been treated as graphs, where convolution operations extract features from local chemical environments for accurate property predictions (Xie & Grossman, 2018; Chen et al., 2019). An outstanding challenge, however, is to identify a universal and task-agnostic representation that can enables generalized efficient search of the vast and largely unlabeled material space to recommend desirable material candidates.

Previously, recommender-like systems for materials search were developed to filter by identifying materials for which predicted confidence levels of target properties fall within a desirable range for thermoelectrics(Gaultois et al., 2016), to predict chemically relevant compositions for pseudo-ternary systems(Seko et al., 2018b;a), and to propose experimental synthesis conditions(Hayashi et al., 2019). However, a systematic and generalizable recommendation approach, which incorporates general material representation, recall, and ranking, could accelerate discovery of desirable material candidates across diverse applications.

Here we present a material recommendation framework that leverages language representations to explore a large space and identify similar candidate materials, given a query material with targeted desired properties. The framework invokes a funnel-based architecture comprising a candidate generation ("recall") step and a subsequent property evaluation ("ranking") step (Figure 1a). We first constructed representations for ∼116,000 materials using text description as the input to the transformer based language models. By evaluating different embedding methods on various downstream tasks, we found that material language representations are both highly potent in recalling relevant material candidates, and capable of predicting properties with comparable performance to state-of-the-art specialized ML models. For improved ranking, we introduced a multi-gate mixture-of-experts (MMoE) model, a multi-task learning strategy, to exploit correlations between material property prediction tasks (Figure 1b). As a demonstration example of material discovery, we applied our framework to search and recommend high-performance thermoelectrics (TEs) – materials that convert waste heat into electricity.

[*]Equal contribution [1]Department of Mechanical Science and Engineering [2]Beckman Institute for Advanced Science and Technology [3]Materials Research Laboratory, University of Illinois Urbana-Champaign, Urbana, IL. Correspondence to: Jiaxing Qu <jiaxing6@illinois.edu>.

Accepted after peer-review at the 1st workshop on Synergy of Scientific and Machine Learning Modeling, SynS & ML ICML, Honolulu, Hawaii, USA. July, 2023. Copyright 2023 by the author(s).

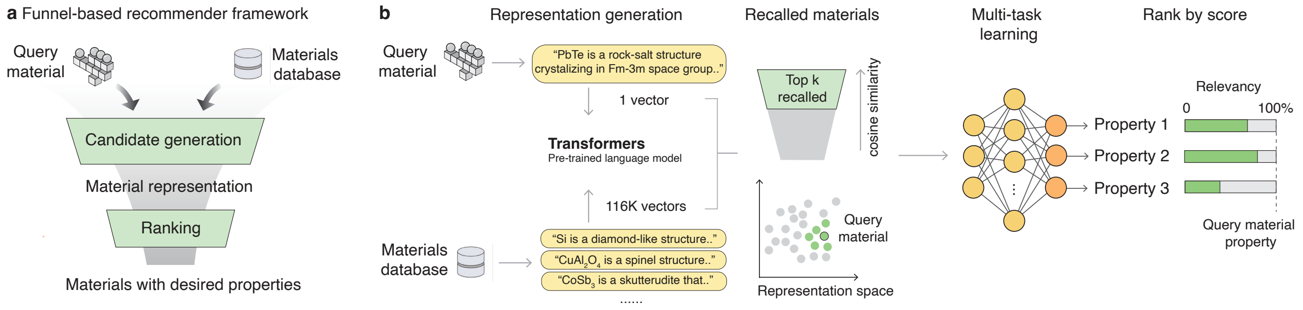

Figure 1. **a**, The proposed funnel-based recommender framework in which candidate materials are recalled, and ranked based on similarity to the query material. **b**, The schematic workflow to screen candidate materials including constructing language representations, recalling candidates, and multi-task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

2. Related work

Advances in natural language processing have allowed information mining from the large corpus of material science related literature in an unsupervised fashion. A pioneering work utilizes word embeddings trained on a large material text corpus to encode material science knowledge into information-dense vector representations (Tshitoyan et al., 2019). Given a context word for technological application, e.g. "thermoelectrics", candidate materials are ranked by similarity to the word embedding of the context word. Word embeddings obtained on material compositions have also shown competitive performance on material property prediction tasks (Wang et al., 2021). However, word embedding, such as in Word2Vec (Tshitoyan et al., 2019), does not capture the contextual meaning of the word that is present in a sentence. Progress on contextual embedding models has been enabled by masked language modeling to train Transformer-based language models (MatBERT) (Trewartha et al., 2022), (MatSciBERT) (Gupta et al., 2022) for material discovery and knowledge extraction from millions of unstructured material science literatures. By employing pretrained BERT models, latent knowledge learnt from the material science text corpus can be encoded into the representation and then subjected to a number of subsequent prediction tasks.

3. Methodology

3.1. Datasets

The training dataset was collected from the Materials Project (Jain et al., 2013) to include 116,216 materials that are thermodynamically stable. We considered five different datasets for material TE properties as training labels, including UCSB (Gaultois et al., 2013), ESTM (Na & Chang, 2022), ChemExtracter (Sierepeklis & Cole, 2022), TEDesignLab (Gorai et al., 2016), and Citrine (Ward et al., 2018) datasets. In all five datasets, 826 materials that have records for five TE properties are used for evaluation of recall performance . 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 dataset, which are experimental values, are utilized as ground-truth labels.

3.2. Material representations

We first established a "baseline" method to represent materials using structure fingerprints to quantify the crystal structural similarity. For the fingerprint generation, it was generated using CrystalNN (Zimmermann & Jain, 2020) algorithm as implemented in Matminer (Ward et al., 2018) package. The fingerprint contains statistical information about local motifs with a size dimension of 122. To acquire the representations for each individual material, we applied robocrystallographer (Ganose & Jain, 2019), 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 embedded all material formulae (e.g., "PbTe") and sentence descriptors automatically generated 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. 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 ∈(0,512] 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.

3.3. MMoE and TE property prediction

A shared-bottom multi-task network was first introduced by (Caruana, 1997) and widely applied for multi-task learning. The basic network formulation is:

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

where $k = 1, 2, 3...K$ 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:

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

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

where i = 1, 2, 3... 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 our 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 of our 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.

3.4. Ranking score

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

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

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:

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

In our experiments, 100 candidates were recalled per query material. The scores presented in the figure were normalized by the maximum score within the recalled list.

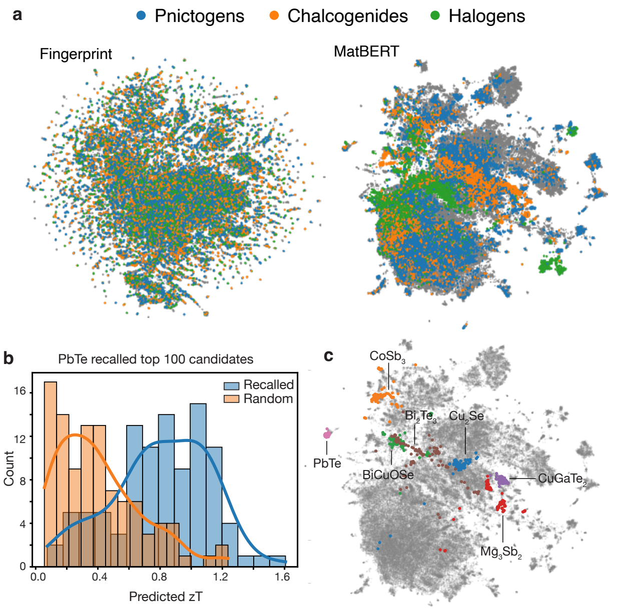

Figure 2. **a**, UMAP projections of 116K materials using different embedding models. Materials are colored by anionic groups. **b**, Distributions of predicted zT of the top-100 recalled candidates for PbTe as the query material and randomly sampled 100 materials. Predicted zT are obtained from the MMoE models. **c**, Recall results of seven high-performing TE materials are highlighted on the UMAP projection of 116K material representations obtained through MatBERT. Each color corresponds to first 100 materials recalled via cosine similarity.

4. Results

4.1. A language-based framework enables material recommendations and discovery.

Inspired by the standard design of recommender systems, we designed a framework for material 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 (Figure 1a). To enable candidate recall for a query material, we embedded each material into a dense vector output from the pretrained language models, which contains latent material-specific knowledge learnt during unsupervised pretraining. In Figure 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 distributions from random sampling as indicated by p-

values (Figure 2a). We repeated this experiment for 100 materials with the known highest zT; 94 of these show statistical significance with p <0.05, showing that recalled materials show distributions that are distinct from random. Moreover, low-dimensional Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) of the material representations display latent signatures of seven high-performing TE materials along with top-100 recalled materials, each indicated by a different color (Figure 2b).

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

Effective representations require rendering information about material design principles and intrinsic properties. We utilized local environment based structure fingerprints (Zimmermann & Jain, 2020) as the baseline, and compare it with sentence embeddings of text-based material descriptions from MatSciBERT and MatBERT by quantitatively evaluated material embedding performance on downstream property prediction tasks. 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 (Curtarolo et al., 2012). Performance metrics of models trained using structure embeddings extracted from MatBERT, achieved most accurate performance (Table 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.

4.3. 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 Figure 1 we improved multi-property predictions through multi-task learning. To this aim, we introduce multi-task learning with the MMoE model, which contains a set of expert networks and gating networks. 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. The input representations for MMoE models are natural language embeddings for structural and compositional features. We first benchmarked MMoE with single-task prediction to predict the properties, as shown in Figure 3a. The MMoE results are within error of the single-task results, but show modest improvement by around 5-10%. MMoE does show notably better model stability, indicated by lower variance in cross-validation performance.

For five TE properties as learning tasks, we found moderate Pearson correlation ranging from 0.15 to 0.5 between the five TE properties, which is considered ideal for multi-task learning. Interestingly, we found that multi-task learning significantly enhances the predictive performance of one learning task (Seebeck coefficient) by 71% compared with single-task prediction, with close performance for the other four tasks within variance from cross-validation. The multi-task learning results from our best-performing material representation and MMoE is shown in Figure 3b. In all five prediction tasks, MMoE accurately predicts the TE properties for the input material 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 (Na et al., 2021; Na & Chang, 2022).

Table 1. Benchmarking six different embedding models on six regression property prediction tasks with MAEs. (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})).

Property	Fingerprint	MatSciBERT	MatBERT
E/atom	1.13±0.02	0.32±0.02	**0.29±0.03**
E_g	0.54±0.03	0.25±0.01	0.23±0.01
log_K	0.45±0.01	0.16±0.01	**0.15±0.01**
log_G	0.48±0.01	0.24±0.01	0.23±0.01
log$_{10}$_θ	0.13±0.01	0.07±0.01	**0.06±0.01**
log$_{10}$_α	0.15±0.01	0.07±0.01	**0.06±0.01**

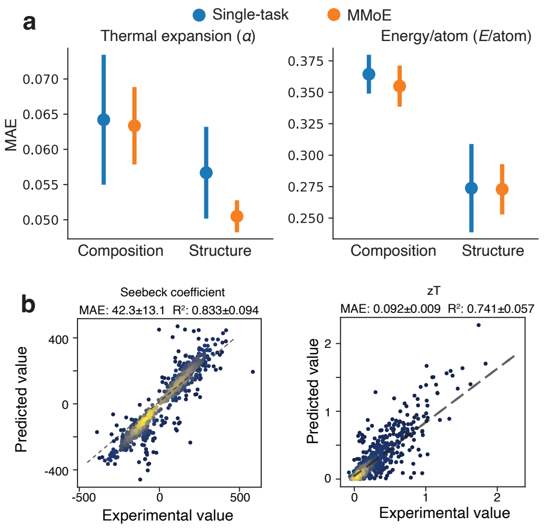

Figure 3. **a**, Comparison of model performance for material properties prediction between single-task models and MMoE. **b**, Multi-task prediction results of TE properties from the best performing MMoE model with 5-fold cross validation.

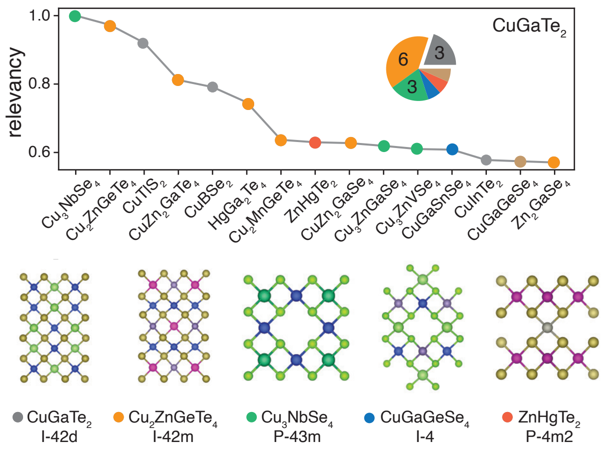

Figure 4. Ranking results of top 15 materials that exhibit most similar TE potential to CuGaTe$_2$. The color of each data point denotes the structure prototype as shown on the right panel for each query material. The recommended structure prototypes share similar structural features with the query material.

4.4. Search ranking of materials with similar potential

To interpret and evaluate the ranking performance, we demonstrated the ranking outcomes from our recommendation framework on a state-of-the-art TE material – CuGaTe$_2$ (Figure 4). Candidates were ranked by *relevancy score* (defined in section 3.4) and the top 15 ranked materials that exhibit the most similar TE potential are shown. In Figure 4, each candidate is colored by its structure prototype to visualize the structural diversity. The distribution of prototype structures is shown by the pie chart. Insterestingly, the recommendations based on querying of CuGaTe$_2$ render diversified outcomes with 5 different structure prototypes. Is it shown that our 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. Moreover, the recommended materials from the framework are further corroborated by our first-principles simulations and experiments, proving the effectiveness of our framework.

Broader impact

While representation learning has facilitated extraction of more meaningful features from large unlabeled data, methods for learning material representations have also gained substantial momentum (Xu et al., 2021; Gupta et al., 2021; Na & Kim, 2022). On the other hand, language-based models have achieved remarkable outcomes in prediction and generation tasks across an extensive array of domain areas. Through language representations in the inorganic crystalline materials domain, we demonstrated a recommendation framework encompassing (i) effective representations of both chemical and structural complexity in the large material space, (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. The framework is designed to be task-agnostic. We anticipate that it can be expanded upon and utilized to search and explore vast chemical spaces, towards functional materials and drug design.

References

Behler, J. and Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters*, 98(14):146401, 2007.

Caruana, R. Multitask learning. *Machine learning*, 28: 41–75, 1997.

Chen, C., Ye, W., Zuo, Y., Zheng, C., and Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9): 3564–3572, 2019.

Curtarolo, S., Setyawan, W., Hart, G. L., Jahnatek, M., Chepulskii, R. V., Taylor, R. H., Wang, S., Xue, J., Yang, K., Levy, O., et al. Aflow: An automatic framework for high-throughput materials discovery. *Computational Materials Science*, 58:218–226, 2012.

Ganose, A. M. and Jain, A. Robocrystallographer: automated crystal structure text descriptions and analysis. *MRS Communications*, 9(3):874–881, 2019.

Gaultois, M. W., Sparks, T. D., Borg, C. K., Seshadri, R., Bonificio, W. D., and Clarke, D. R. Data-driven review of thermoelectric materials: performance and resource considerations. *Chemistry of Materials*, 25(15):2911–2920, 2013.

Gaultois, M. W., Oliynyk, A. O., Mar, A., Sparks, T. D., Mulholland, G. J., and Meredig, B. Perspective: Web-based machine learning models for real-time screening of thermoelectric materials properties. *Apl Materials*, 4 (5):053213, 2016.

Gorai, P., Gao, D., Ortiz, B., Miller, S., Barnett, S. A., Mason, T., Lv, Q., Stevanovic, V., and Toberer, E. S. Te design lab: A virtual laboratory for thermoelectric material design. *Computational Materials Science*, 112: 368–376, 2016.

Gupta, T., Zaki, M., and Krishnan, N. A. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8 (1):102, 2022.

Gupta, V., Choudhary, K., Tavazza, F., Campbell, C., Liao, W.-k., Choudhary, A., and Agrawal, A. Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data. *Nature communications*, 12(1):6595, 2021.

Hayashi, H., Hayashi, K., Kouzai, K., Seko, A., and Tanaka, I. Recommender system of successful processing conditions for new compounds based on a parallel experimental data set. *Chemistry of Materials*, 31(24):9984–9992, 2019.

Hey, A. J., Tansley, S., Tolle, K. M., et al. *The fourth paradigm: data-intensive scientific discovery*, volume 1. Microsoft research Redmond, WA, 2009.

Isayev, O., Oses, C., Toher, C., Gossett, E., Curtarolo, S., and Tropsha, A. Universal fragment descriptors for predicting properties of inorganic crystals. *Nature communications*, 8(1):15679, 2017.

Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1):011002, 2013.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Na, G. S. and Chang, H. A public database of thermoelectric materials and system-identified material representation for data-driven discovery. *npj Computational Materials*, 8(1):214, 2022.

Na, G. S. and Kim, H. W. Contrastive representation learning of inorganic materials to overcome lack of training datasets. *Chemical Communications*, 58(47):6729–6732, 2022.

Na, G. S., Jang, S., and Chang, H. Predicting thermoelectric properties from chemical formula with explicitly identifying dopant effects. *npj Computational Materials*, 7(1): 106, 2021.

Schmidt, J., Marques, M. R., Botti, S., and Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, 5(1):83, 2019.

Seko, A., Hayashi, H., Kashima, H., and Tanaka, I. Matrix- and tensor-based recommender systems for the discovery of currently unknown inorganic compounds. *Physical Review Materials*, 2(1):013805, 2018a.

Seko, A., Hayashi, H., and Tanaka, I. Compositional descriptor-based recommender system for the materials discovery. *The Journal of chemical physics*, 148(24): 241719, 2018b.

Sierepeklis, O. and Cole, J. M. A thermoelectric materials database auto-generated from the scientific literature using chemdataextractor. *Scientific Data*, 9(1):648, 2022.

Trewartha, A., Walker, N., Huo, H., Lee, S., Cruse, K., Dagdelen, J., Dunn, A., Persson, K. A., Ceder, G., and Jain, A. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns*, 3(4):100488, 2022.

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., and Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, 2019.

Wang, A. Y.-T., Kauwe, S. K., Murdock, R. J., and Sparks, T. D. Compositionally restricted attention-based network for materials property predictions. *Npj Computational Materials*, 7(1):77, 2021.

Ward, L., Dunn, A., Faghaninia, A., Zimmermann, N. E., Bajaj, S., Wang, Q., Montoya, J., Chen, J., Bystrom, K., Dylla, M., et al. Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, 152:60–69, 2018.

Xie, T. and Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120 (14):145301, 2018.

Xu, M., Wang, H., Ni, B., Guo, H., and Tang, J. Self-supervised graph-level representation learning with local and global structure. In *International Conference on Machine Learning*, pp. 11548–11558. PMLR, 2021.

Zimmermann, N. E. and Jain, A. Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity. *RSC advances*, 10(10):6063–6081, 2020.