Tokenizer Effect on Functional Material Prediction: Investigating Contextual Word Embeddings for Knowledge Discovery

Tong Xie
University of New South Wales
GreenDynamics Pty. Ltd
`tong.xie@unsw.edu.au`

Yuwei Wan[*]
City University of Hong Kong
GreenDynamics Pty. Ltd
`yuweiwan2-c@my.cityu.edu.hk`

Ke Lu
University of California, Los Angeles
GreenDynamics Pty. Ltd

Wenjie Zhang
University of New South Wales

Chunyu Kit[†]
City University of Hong Kong
`ctckit@cityu.edu.hk`

Bram Hoex[‡]
University of New South Wales
`b.hoex@unsw.edu.au`

Abstract

Exploring the predictive capabilities of natural language processing models in material science is a subject of ongoing interest. This study examines material property prediction, relying on models to extract latent knowledge from compound names and material properties. We assessed various methods for contextual embeddings and explored pre-trained models like BERT and GPT. Our findings indicate that using information-dense embeddings from the third layer of domain-specific BERT models, such as MatBERT, combined with the context-average method, is the optimal approach for utilizing unsupervised word embeddings from material science literature to identify material-property relationships. The stark contrast between the domain-specific MatBERT and the general BERT model emphasizes the value of domain-specific training and tokenization for material prediction. Our research identifies a "tokenizer effect", highlighting the importance of specialized tokenization techniques to capture material names effectively during the pretraining phase. We discovered that a tokenizer which preserves compound names entirely, while maintaining a consistent token count, enhances the efficacy of context-aware embeddings in functional material prediction.

1 Introduction

Materials science is an inherently versatile field with intricate connections and intersections with various domains like physics, chemistry, energy and engineering. In recent decades, innovative materials often emerge from integrating knowledge and ideas across multiple disciplines, resulting in advancements across various industries. This interdisciplinary nature raises a longstanding question

---

[*] Equal contribution

[†] Co-corresponding author

[‡] Corresponding author

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

in materials informatics: How can we efficiently identify materials with optimal properties for specific purposes from vast literature findings?

Currently, the majority of these advances are still largely attributed to a methodical process of trial-and-error experimentation base on researcher experience, entailing substantial investments in time and resources. This process relies to a significant extent on the interdisciplinary knowledge and practical experience of scientists. The contemporary landscape of global challenges amplifies the urgency for expediting research that is faster, more efficient, and cost-effective. Recent works [1, 2] have utilized an unsupervised natural language processing (NLP) technique to discover new material synthesize materials science knowledge and demonstrated that it could discover materials for specific applications for thermoelectric, photovoltaic materials. Their technique, word embedding, represents a word using a real-valued vector in a high-dimensional space, where words with similar meanings are mapped to nearby points. These vectors, called word embeddings, are often trained on large corpora to encapsulate the distributional properties of words in the text. Word embeddings can capture complex materials science concepts and structure-property relationships directly from text without needing explicit domain knowledge insertion [3].

While word embeddings trained by algorithms like Word2Vec [4] and GloVe [5] have gained significant popularity, more advanced models like Bidirectional Encoder Representations from Transformers (BERT) [6] and Generative Pre-trained Transformers (GPT) [7] have emerged, capable of capturing even richer contextual information. Given their remarkable performance in various NLP tasks, we hypothesize that these novel models may enhance domain-specific knowledge discovery by utilizing their intrinsic word representations.

This study aims to assess the feasibility of using contextualized word representations for a knowledge discovery task, material prediction. This task is designed to identify materials suitable for a specific functional application by ranking a list of potential materials from the corpus. We reproduced an existing thermoelectric material prediction dataset containing 84 materials ranked by their experimental zT value. On the methodological front, we conducted preliminary investigations into the efficacy of various contextual embeddings for this task, including pre-trained models like BERT and GPT and methods of obtaining embeddings. The results emphasized the importance of domain-specific pretraining and tokenization techniques in enhancing the performance of contextual language models for knowledge discovery. The study also identified the "tokenizer effect", which makes contextual embeddings favours shorter tokenized material names in material prediction, underscoring the need for improved tokenization mechanisms. The code used for datasets and experiments in our study are available at An anonymous link to the code/data

2 Related Works

Knowledge discovery was defined as 'the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable pattern or knowledge in data' in 1996 [8]. Word embedding technique, which has its root in distributional semantics [9], can be applied in knowledge discovery prediction directly [10]. It often rely on semantic-similarity measures between word representations to predict relationships, which are subsequently validated using domain-specific scientific methods [11]. Traditional approaches have leveraged static word embeddings (each word has fixed vector) such as Word2Vec and GloVe to uncover latent knowledge within domain-specific text data. For example, Tshitoyan et al. [1] proposed an embedding-based ranking prediction: Word2Vec embeddings of material names were tanked by their cosine similarity to the embedding of 'thermoelectric' and obtained a 59% rank correlation with experimental results. In contrast, the ranking predicted by density functional theory (DFT) calculation only exhibits a 31% rank correlation. The context analysis reveals that the direct relationship between the novel materials and 'thermoelectric' may be attributed to indirect connections involving the material names and related terms, such as 'chalcogenide' (many chalcogenides are good thermoelectrics) and 'band gap' (important to thermoelectric properties). This method was applied to solar materials and identified potential candidates such as As_2O_5 [12]. Shetty and Ramprasad also demonstrated that word embeddings trained on a corpus of polymer papers could encode materials science knowledge and used to identify novel polymers for certain applications [2]. In biomedical domain, Venkatakrishnan et al. [13] applied the same method to identify novel tissue-reservoirs of the ACE2 receptor used by SARS-CoV-2 to invade a host.

Several of above studies [1, 13] suggested in the discussion that dynamic embeddings (each word is assigned a distinct embedding based on its context) from pretrained models like BERT, might surpass those static embeddings in the realm of knowledge discovery. In a sentiment analysis study based on Twitter [14], dynamics embeddings demonstrate better ability of capturing hidden information and intricate relationship from context compared to static embeddings. However, there are relatively few efforts to tailor models like BERT for knowledge discovery prediction purpose. Panesar [11] obtained word embeddings based on sentences extracted from a biomedical corpus using domain-specific pretrained models like BioBERT [15], achieving superior performance on biomedical benchmarks compared to static embeddings. The benchmarks utilized in the paper involves directly comparing the cosine similarity of word representations from language models to human domain-expert ratings. This kind of task aims to measure ability of capturing semantic relationship, which differs from material prediction tasks based on experimental results, leaving room for further exploration.

3 Methods

BERT and GPT models generate distinct vectors for the same word in varying contexts, based on the surrounding words within the sentence. To investigate contextual word embedding in material prediction task, we employed various approaches to acquire BERT embeddings, which encompass methodologies involving the exclusion or inclusion of actual sentence contexts, different layers, and combinations of layers. This section provides a detailed exposition of two methods for acquiring BERT embeddings of material names in our study, namely, the context-free and context-average.

Context-free We use a single context [16] for each material name to obtain context-free BERT word embebddings. Initially, we utilized the pretrained BERT tokenizer to convert material names into tokens. For each material name word w, tokenizer decomposed w into x sub-word tokens ($x \geq 1$), resulting in tokenized text $\{t_1, ..., t_x\}$. Subsequently, we fed the tokenized text into the pretrained BERT model, resulting in output hidden states with four dimensions:

1. The layer number (13, including the initial layer and 12 output layers)
2. The batch number (always 1 corresponding to a single input string)
3. The token number (varying as the number of tokens in the string)
4. The hidden unit number (768 features)

As Figure 1 shows, each token t within our input generated 13 separate vectors, each of which possessed a length of 768. To form a token embedding $\boldsymbol{t}$, we directly extracted certain layer vector or combine some of the layer vectors by operations like summing, averaging, or concatenating. Given that there were no contextual words available to impart meaning to the material name in the input, we calculated the final representation by averaging across all token embeddings: $\boldsymbol{w} = \frac{1}{x}(\boldsymbol{t}_1 + ... + \boldsymbol{t}_x)$.

Context-average Inspired by Bommassani et al. [16], we also aggregated contextualized word representations over contexts to obtain static-equivalents. Utilizing a paper downloading tool (An anonymous link to the code/data), we collected scholarly literature related to the material names from Web of Science (English; published in 2000.01.01-2023.09.15; published by Elsevier or Springer Nature) and converted them into plain texts. As Figure 2 shows, for each material name w, we select n ($n \leq 100$) sentences containing it as contexts $\{s_1, ..., s_n\}$ (if the number of collected sentences is larger than 100, randomly sample 100 sentences from the collection). These sentences were then tokenized by pretrained BERT tokenizer. Assuming that the tokens of w (if w have more than 1 occurrences in s, only use the first one) are located within the tokens of s at the positions $[s_i, ..., s_j]$. For each sentence s, we obtained token embeddings for each token using same method of generateing token embeddings in context-free approach and extracted $\{\boldsymbol{t}_{s_i}, ..., \boldsymbol{t}_{s_j}\}$ from all token embeddings of s. The representation of w in the context s is calculated as:

$$\boldsymbol{w_s} = \frac{1}{j - i + 1}(\boldsymbol{t}_{s_i} + ... + \boldsymbol{t}_{s_j}) \tag{1}$$

After obtaining representation of w in all contexts, we calculated the arithmetic mean of n contextual representations:

$$\boldsymbol{w} = \frac{1}{n}(\boldsymbol{w}_{s_1} + ... + \boldsymbol{w}_{s_n}) \tag{2}$$

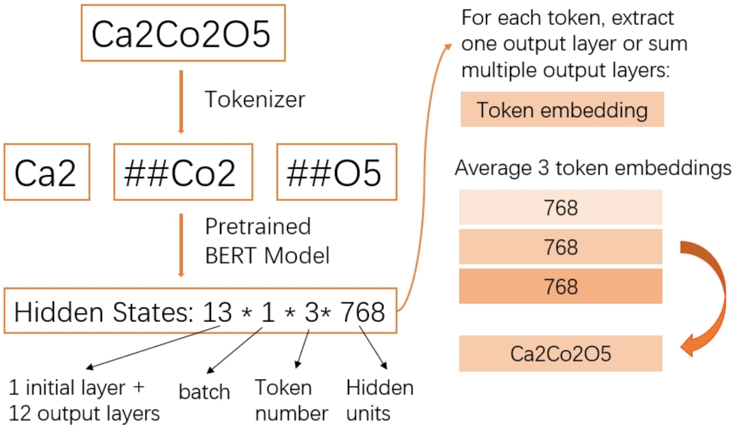
Ca2Co2O5
Tokenizer
Ca2
##Co2
##O5
Pretrained BERT Model
Hidden States: 13 * 1 * 3* 768
1 initial layer + 12 output layers
batch
Token number
Hidden units
For each token, extract one output layer or sum multiple output layers:
Token embedding
Average 3 token embeddings
768
768
768
Ca2Co2O5

Figure 1: A diagram of obtaining context-free embedding for Ca2Co2O5.

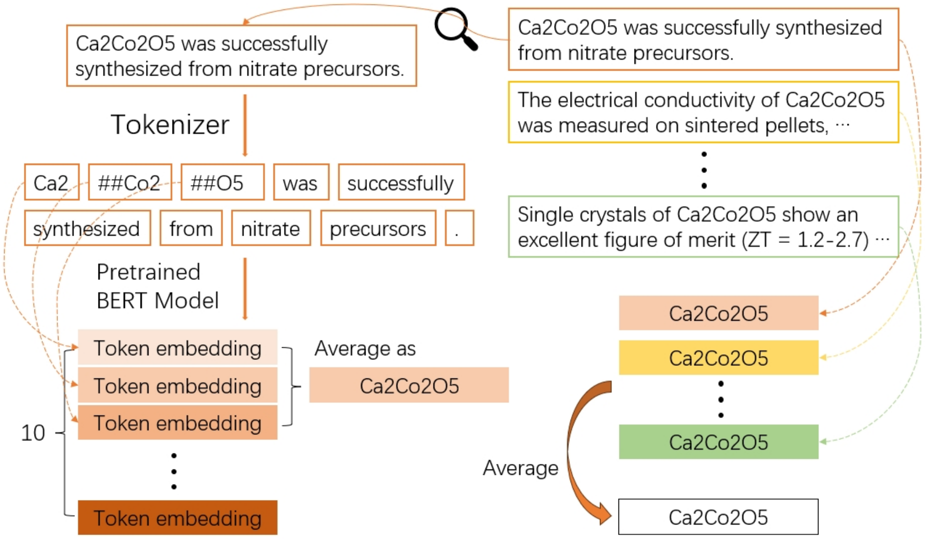
Ca2Co2O5 was successfully synthesized from nitrate precursors.
Tokenizer
Ca2
##Co2
##O5
was
successfully
synthesized
from
nitrate
precursors
.
Pretrained BERT Model
Token embedding
Token embedding
Token embedding
Token embedding
10
Average as
Ca2Co2O5
Ca2Co2O5 was successfully synthesized from nitrate precursors.
The electrical conductivity of Ca2Co2O5 was measured on sintered pellets, ⋯
Single crystals of Ca2Co2O5 show an excellent figure of merit (ZT = 1.2-2.7) ⋯
Ca2Co2O5
Ca2Co2O5
Ca2Co2O5
Average
Ca2Co2O5

Figure 2: A diagram of obtaining context-average embedding for Ca2Co2O5.

For both approaches, we also obtained the representation of a pre-selected word about a certain application, like ’thermoelectric’, to be the center embedding. The material names were ranked by cosine similarity between their embeddings and the center embedding.

4 Thermoelectrical Material Prediction

In order to assess the performance of contextual embeddings obtained through different methods and configurations in the context of a material prediction task, we conducted an initial exploration on an existing dataset.

4.1 Dataset

Tshitoyan et al. [1] constructed a small-scale dataset using 83 materials that appear both in their text corpus and the experimental set [17]. We reproduced a 84-material dataset ranked by zT value (an important component of the overall thermoelectric figure of merit) by using their released data in https://github.com/materialsintelligence/mat2vec.

4.2 Model

For our thermoelectric material prediction task, we evaluated the performance of contextual embeddings generated by several state-of-the-art pretrained models, including BERT [6], MatBERT [18], OpenAI Embeddings API [19]:

- **BERT** was pretrained on English language using a masked language modeling (MLM) and next sentence prediction (NSP) objective. In this study, we used a case-sensitive BERT, bert-base-cased. The texts used for training were tokenized using WordPiece [20] and reserved a vocabulary size of 28,996.
- **MatBERT** is a BERT model trained using scientific papers specifically from the field of materials science and only MLM objective. We used MatBERT-base-cased (available at https://github.com/lbnlp/MatBERT). The texts used for training were also tokenized using WordPiece and reserved a vocabulary size of 30,552.
- **OpenAI Embeddings API** can be directly used to directly retrieve contextual embeddings for text data. We used text-embedding-ada-002 in this study.

In addition, we also use ChatGPT (https://chat.openai.com) to re-rank the given materials as a supplementary trial. This chat mode does not involve embedding. Rather, it relies on interactive engagement with the model through prompts, enabling the model to re-rank the material using its inherent capabilities. The prompts used are available in Appendix **??**.

4.3 Results

To assess the performance of contextual embeddings, we compared the predicted rank with actual rank (ranked by the experimental results) using the Spearman's rank correlation coefficient [21]. Spearman's correlation ranges from -1 to 1 (values near 1 indicate similarity in two ranks, while values near -1 indicate dissimilarity in two ranks). The baseline of this 84-material dataset should be density functional theory (DFT) method with around 31% rank correlation [1]. For the sake of convenience in subsequent discussions, we shall denote the Spearman's correlation between the predicted rank of 84 materials and their actual rank as 'Correlation_1', and the Spearman's correlation between the predicted rank of 84 materials and their tokenized length as 'Correlation_2'. We selected 'thermoelectric' as the center word.

In our experiments, we encountered results that deviated somewhat from our initial expectations. However, we identified certain patterns that align with phenomena previously observed in related studies and we observed intriguing phenomena associated with the tokenizers of pretrained models.

Context-free As mentioned in context-free method in section 3, we directly tokenized the 84 material names and input them into both the BERT and MatBERT models. We calculated Correlation_1 and Correlation_2 using embeddings generated by each layers in two models (Table 1 shows part of the results). The results indicate that the BERT embeddings generated using the context-free method do not exhibit a meaningful correlation with the predicted and actual results. However, the output of the tokenizer does influence the ranking to some extent, suggesting that material names composed of fewer tokens tend to be ranked higher.

In contrast, embeddings obtained from the domain-specific pretrained model MatBERT demonstrate a very subtle correlation between the predicted and actual results. Comparing the results across layers (excluding the initial layer), we observe that the correlation values peak at the third layer and rapidly decline, with a resurgence towards the end. Additionally, it is notable that the impact of the MatBERT tokenizer is more pronounced than that of BERT. Interestingly, the third layer, which exhibits the weakest influence from the tokenizer, yields the most favorable results.

Regarding the impact of the tokenizer, we conducted several minor tests. First, we compare the length of tokenized 84 materials (see Appendix B). Compared to BERT, MatBERT segments a material's name more comprehensively, without being overly fragmented. Then we calculated the Spearman's correlation between the actual rank of 84 materials and their tokenized length, which is only 1.6%.

Table 1: Results of thermoelectrical material prediction using context-free method (part)

Model	Layer	Correlation_1	Correlation_2
BERT	2	0.0960	0.5397
	3	-0.029	0.1133
	6	-0.1946	0.52
	13	-0.082	-0.2672
MatBERT	2	0.0779	0.8130
	3	0.1707	0.4605
	6	0.018	0.7187
	13	0.067	0.7463

This indicates that the shorter tokenized length cannot imply better experimental performance. We also replaced the center word "thermoelectric" with unrelated words like "apple" or "hit." After this substitution, both BERT and MatBERT failed to demonstrate even the slight correlation between the predicted rank and the actual rank. Notably, the tokenizer effect at lower layers significantly diminished (for instance, MatBERT exhibited a Correlation_2 of -15% at the third layer). However, at upper layers, Correlation_2 remained relatively high, with both MatBERT and BERT reaching around 70%.

Context-average As mentioned in context-average method in section 3, we directly tokenized the context sentences of 79 material names (5 material names were not found in web of science with predefined conditions) and input them into both the BERT and MatBERT models to generate embeddings of material names. As shown in Figure 3, the performance of context-average BERT embeddings aligns with that of BERT when utilizing the context-free method, but MatBERT shows a substantial improvement. For BERT, even in the case of the best-performing layer, the Correlation_1 value is insufficient to establish a meaningful correlation between the predicted and actual results. In contrast, the embeddings from the third layer of MatBERT yield the highest correlation (Correlation_1=39.61%) with experimental results and exhibits the weakest influence from the tokenizer (Correlation_2=17.01%). Detailed ranking by the third layer of MatBERT is available in Appendix C).

The figure in Appendix B shows that the majority of the 84 material names tokenized by MatBERT tokenizer exhibit token lengths predominantly falling within the range of 7 to 11 tokens. By exclusively retaining this subset of material names that possess more closely aligned token lengths (39 material names), we re-ranked them using the embeddings derived from the third layer of MatBERT and the correlation with experimental results achieved 51.34%.

GPT embedding We employed the text-embedding-ada-002 model to generate embeddings for the 84 material names and subsequently ranked them based on their cosine similarity scores with center word. The results revealed a relatively lower correlation (23.55%) between the predicted ranks and the actual experimental results (detailed ranking is available in Appendix C). This performance metric fell notably short in comparison to the results achieved using context-average method and the MatBERT model. This further underscores the significance of domain-specific training and the provision of requisite context.

ChatGPT As a supplementary trial, we leveraged ChatGPT to re-rank the provided dataset comprising 84 distinct materials (see Appendix A. The model was tasked with considering various factors such as electrical and thermal conductivity, as well as the Seebeck coefficient, to assign novel rankings. The results demonstrated a moderate correlation of 24.03% with the actual experimental results. This performance still feww short of the correlation achieved by the MatBERT model using context-average method, but slightly surpass that achieved through the utilization of GPT embeddings. Through an analysis of the outcomes generated by ChatGPT, we deduced that ChatGPT itself possesses a degree of domain-specific knowledge, enabling it to discern and recommend materials commonly utilized in specific applications from the provided list. However, its proficiency in comprehending less conventional or less prevalent materials appears to be limited.

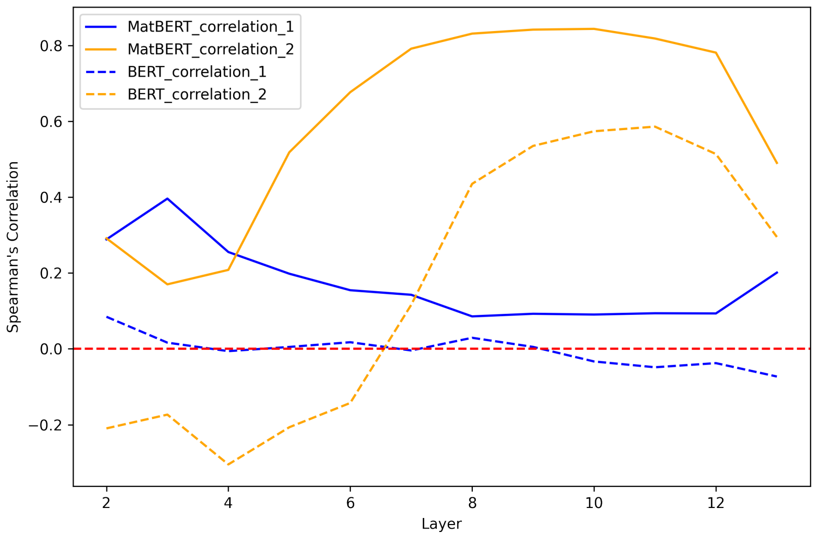

Figure 3: Results of thermoelectrical material prediction using context-average method

5 Discussion

In prior research [1, 13], there was a hypothesis that contextual embeddings would outperform traditional Word2Vec embeddings in material prediction task, and indeed, they have exhibited promise in some knowledge discovery tasks [11, 14]. However, our practical exploration in this study has unveiled nuanced findings. While contextual embeddings do possess the capability to predict material performance for a certain application to some extent (40%) and surpasses the correlation obtained by Density Functional Theory (DFT) predictions (31%), they fall short when compared to conventional Word2Vec methods (59%).

In our experiments, the substantial disparity in performance observed between MatBERT and the BERT model on the same method underscore the indispensability of domain-specific training when employing contextual embedding models for material prediction tasks. In other downstream domain-specific NLP tasks, such as Named Entity Recognition (NER), the availability of labeled data (there is usually annotated data for supervised fine-tuning), coupled with the inherent capabilities of BERT models, enables the narrowing of the knowledge gap. Consequently, the disparity in performance between BERT and MatBERT in these downstream tasks does not tend to be exceedingly pronounced [18]. In the contrast, material prediction is an unsupervised task without annotated data and primarily rely on the model's capacity to encapsulate latent knowledge embedded within material names, highlighting greater importance of domain-specific pretraining.

Through comparisons of different methods for obtaining contextual embeddings, we have found that utilizing the output embeddings from the third layer of contextual embedding models (BERT and MatBERT in our study), in conjunction with the context-average method (averaging embeddings of material names generated from various context sentences), is the most suitable approach for this specific task. Comparing the results across layers, the correlation values peak at the third layer and rapidly decline with a resurgence towards the end, which follows a similar trend in word similarity task [16]. Prior work has suggested that for most language models, the lower layers specialize in local syntactic relationships while the higher layers may be expressly encoding contextual semantic information [22, 23, 24]. In synthesizing these pieces of information, it can be inferred that the characterization of material names primarily stems from their intrinsic lexical-level information. While necessitating a certain degree of contextual information, this reliance on context is less pronounced compared to more advanced semantic-level tasks such as coreference identification. This observation indirectly lends support to the notion that the straightforward application of Word2Vec yields good performance in this task.

Additionally, our experiments have shed light on the "tokenizer effect", where contextual embeddings tend to prioritize material names with shorter tokenized lengths when ranking materials. We quantified the strength of this tokenizer effect using Spearman correlation and observed that good prediction performance is often associated with low tokenizer effect but low tokenizer effect does not have to be related with good prediction performance. In other words, low tokenizer effect is a necessary but not sufficient condition for good prediction performance of contextual embeddings.

The results of our experiments, along with a series of tokenizer-related tests, suggest that standard tokenization methods like WordPiece may not be suitable for capturing the terminology in material science and further improvements in the effectiveness of contextual embeddings for material prediction necessitate refining the current tokenizer mechanisms. Although MatBERT tokenizer was pretrained on domain-specific text, it cannot guarantee that longer material names are not overly segmented, potentially leading to loss of meaning when aggregating token embeddings to form the overall material name embedding. Therefore, future efforts should focus on enhancing tokenizer mechanisms to advance the utility of contextual embeddings in material science applications. One potential approach entails augmenting the vocabulary with as many complete material names as feasible, replacing subwords of material names, prior to domain-specific pretraining. The other approach is to use paragraph-level text as the source of representation of a material name, to bypass “tokenizer effect” mechanism. Comparing two approaches, the second one is operationally more feasible and has the ability to leverage contextual embedding for the encoding of higher-level information. The source text can be formed by different kinds of information, such as structure description, property data and experimental performance.

We also employed GPT embeddings and ChatGPT for the material prediction task. Intriguingly, ChatGPT exhibited a slightly better performance to GPT embeddings, autonomously providing limited insights without specific prompts. Their performance fell short of the MatBERT+context-average approach, but outperformed MatBERT+context-free approach. This suggests that there may be room for exploration in adopting a hybrid approach that leverages contextual information from sentences to enhance GPT’s comprehension of materials, potentially yielding improvements in predictive accuracy.

6 Conclusion

In conclusion, this study offers valuable insights into the challenges and opportunities of employing contextual embeddings in material prediction, with implications extending to other specialized domains. Future work will be required to explore novel tokenizer mechanisms tailored to material names. Combining the strengths of different language models, such as contextual embeddings and generative models, may also unlock new opportunities for further advancing our ability to unlock valuable knowledge from unstructured textual sources.

References

[1] Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, 2019.

[2] Pranav Shetty and Rampi Ramprasad. Automated knowledge extraction from polymer literature using natural language processing. *Iscience*, 24(1), 2021.

[3] Ghanshyam Pilania. Machine learning in materials science: From explainable predictions to autonomous design. *Computational Materials Science*, 193:110360, 2021.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[5] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[8] Usama M Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, et al. Knowledge discovery and data mining: Towards a unifying framework. In *KDD*, volume 96, pages 82–88, 1996.

[9] John Firth. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, pages 10–32, 1957.

[10] Pedro Aceves and James A Evans. Mobilizing conceptual spaces: How word embedding models can inform measurement and theory within organization science. *Organization Science*, 2023.

[11] Sandip S Panesar. Biomedical knowledge discovery from unstructured text corpora using contextual word embeddings. *Research Square preprint rs.3.rs-2286334*, 2022.

[12] Lei Zhang and Mu He. Unsupervised machine learning for solar cell materials from the literature. *Journal of Applied Physics*, 131(6), 2022.

[13] AJ Venkatakrishnan, Arjun Puranik, Akash Anand, David Zemmour, Xiang Yao, Xiaoying Wu, Ramakrishna Chilaka, Dariusz K Murakowski, Kristopher Standish, Bharathwaj Raghunathan, et al. Knowledge synthesis of 100 million biomedical documents augments the deep expression profiling of coronavirus receptors. *Elife*, 9:e58040, 2020.

[14] Sumona Deb and Ashis Kumar Chanda. Comparative analysis of contextual and context-free embeddings in disaster prediction from twitter data. *Machine Learning with Applications*, 7:100253, 2022.

[15] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[16] Rishi Bommasani, Kelly Davis, and Claire Cardie. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, 2020.

[17] Michael W Gaultois, Taylor D Sparks, Christopher KH Borg, Ram Seshadri, William D Bonificio, and David R Clarke. Data-driven review of thermoelectric materials: performance and resource considerations. *Chemistry of Materials*, 25(15):2911–2920, 2013.

[18] Amalie Trewartha, Nicholas Walker, Haoyan Huo, Sanghoon Lee, Kevin Cruse, John Dagdelen, Alexander Dunn, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns*, 3(4), 2022.

[19] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. Text and code embeddings by contrastive pre-training, 2022.

[20] Mike Schuster and Kaisuke Nakajima. Japanese and korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE, 2012.

[21] Charles Spearman. The proof and measurement of association between two things. 1961.

[22] Matthew E Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. Dissecting contextual word embeddings: Architecture and representation. *arXiv preprint arXiv:1808.08949*, 2018.

[23] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*, 2019.

[24] Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*, 2019.

A Material Prediction by ChatGPT

Prompt:

Rank the following 84 materials based on their thermoelectric performance, with 1 being the highest-performing material and 84 being the lowest-performing material. Consider factors such as electrical conductivity, thermal conductivity, and the Seebeck coefficient when determining their thermoelectric performance.
['WO2.9', 'In2O3', 'Ni0.02Zn0.98O', ..., 'Mn1.3Mo6S8']

Answer from ChatGPT:

Ranking materials based on their thermoelectric performance can be a complex task and depends on various factors, including electrical conductivity, thermal conductivity, and the Seebeck coefficient. Since I don't have access to the most up-to-date and specific thermoelectric performance data for these materials, I can provide a general ranking based on the properties of common thermoelectric materials. Keep in mind that this ranking may not be accurate for your specific application, as the performance depends on various parameters and conditions.

Here's a general ranking of the provided materials from 1 (highest thermoelectric performance) to 84 (lowest thermoelectric performance):

Bismuth Antimony Telluride (Bi2Te3)
Lead Telluride (PbTe)
Silicon-Germanium Alloy (Si0.8Ge0.2)
Zinc Antimonide (Zn4Sb3)
Strontium Titanium Niobium Oxide (SrTi0.8Nb0.2O3)
Bismuth Selenide (Bi2Se3)
Calcium Cobaltate (Ca3Co4O9)
Copper Selenide (Cu2Se)
Magnesium Silicide (Mg2Si)
Sodium Cobaltate (NaCo2O4)
Barium Strontium Titanate (Ba0.4Sr0.6TiO3)
Nickel-based Thermoelectric Materials (Ni0.02Zn0.98O, Ni2.0Mo6S8, Ni1.3Mo6S8)
Copper-based Thermoelectric Materials (Cu1.98Se, Cu4.0Mo6S8, Cu2Se)
Cerium-based Thermoelectric Materials (CeFe4Sb12, CeFe3CoSb12, CeFe3.5Co0.5Sb12)
Cobalt-based Thermoelectric Materials (Ca0.9Bi0.1MnO3, Ca0.9Ho0.1MnO3, Ca0.9Nd0.1MnO3, LaCoO3)
Tl-Bi-Te Compound (Tl9BiTe6, Tl2SnTe5, Tl2GeTe5)
Zinc Oxide (ZnO, Zn0.95Al0.05O, Zn0.98Al0.02O, Zn0.99Al0.01O)
Titanium-based Thermoelectric Materials (Zr0.25Hf0.25Ti0.5NiSn, Zr0.98Nb0.02NiSn, Zr0.5Hf0.5NiSn, ZrNiSn0.98Sb0.02, ZrNiSn, Zr0.35Hf0.35Ti0.3NiSn)
Copper-Chromium Oxides (CuCr0.97Mg0.03O2, CuCr0.98Mg0.02O2, CuCr0.95Mg0.05O2, CuCrO2)
Perovskite Oxides (La0.98Sr0.02CoO3, La0.8Sr0.2CoO3, Sr0.9La0.1TiO3, Sr0.61Ba0.39Nb2O6, Sr0.9Y0.1TiO3)
Rare Earth-based Thermoelectric Materials (Yb14MnSb11, Sm0.5Ca0.5MnO3, La1-xSrxCoO3, La0.95Sr0.05CoO3, LaFe3CoSb12, LaFe3CoSb12)
Other Oxides (BaPbO3, Ba8Ga16Ge30, Ca2.7Bi0.3Co4O9, Ca3AlSb3, Ca2Co2O5, Ca0.7Y0.3MnO3, CaMn0.98Nb0.02O3, CaMnO3)
Iron-based Thermoelectric Materials (Fe1.94Ti0.06O3, Fe1.3Mo6S8, Cr1.3Mo6S8, Mn1.3Mo6S8)
Other Compounds (Nd2CuO4, K2Bi8Se13, AgCrSe2, Mo3Te4)

Please note that this ranking is based on general knowledge of thermoelectric materials and properties as of my last update in September 2021. For specific applications or the most up-to-date information, it's essential to refer to recent research and literature on thermoelectric materials. Additionally, the actual performance may vary depending on factors like temperature, doping, and material preparation methods.

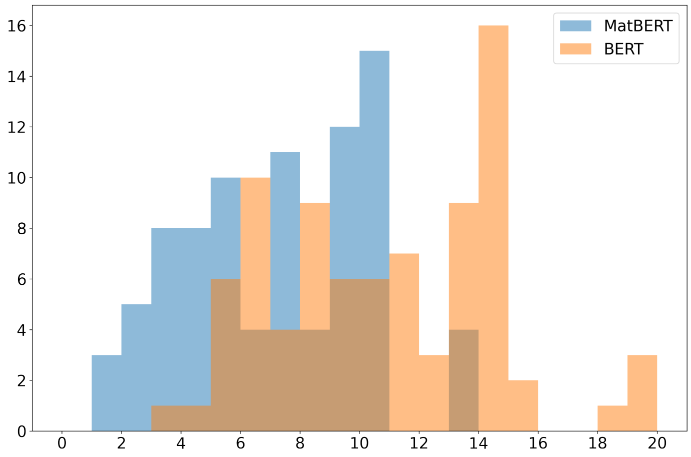

Figure 4: Comparison of length of tokenized 84 material names by different tokenizer

B Comparison of BERT tokenizer and MatBERT tokenizer

C Detailed Ranking Prediction

Table 2: Comparison of 79-material ranking by original experimental results and MatBERT (third-layer, context-average)

Rank	Experiment	MatBERT
1	Cu2Se	Bi2Te3
2	In0.25Co4Sb12	Sb2Te3
3	Zr0.25Hf0.25Ti0.5NiSn	Cu2Se
4	Ba8Ga16Ge30	Cu1.98Se
5	NaCo2O4	CeFe4Sb12
6	Tl0.02Pb0.98Te	Tl9BiTe6
7	In0.2Co4Sb12	LaFe3CoSb12
8	Na0.02Pb0.98Te	Tl2GeTe5
9	Zn4Sb3	Mo3Te4
10	Ag9TlTe5	Yb14MnSb11
11	Zr0.35Hf0.35Ti0.3NiSn	ZrNiSn
12	Cu1.98Se	Bi2Sr2Co2O8
13	Bi2Sr2Co2O8	CeFe3CoSb12
14	Ca2Co2O5	Na0.02Pb0.98Te
15	Tl9BiTe6	TiNiSn
16	Bi2Te3	ZrNiSn0.98Sb0.02
17	AgCrSe2	AgCrSe2
18	Ca3Co4O9	K2Bi8Se13
19	Sr0.61Ba0.39Nb2O6	Tl0.02Pb0.98Te
20	In0.1Co4Sb12	Zr0.98Nb0.02NiSn
21	Mg2Si0.98Bi0.02	In0.1Co4Sb12
22	Tl0.01Pb0.99Te	Ag9TlTe5
23	LaFe3CoSb12	In0.2Co4Sb12
24	Si0.8Ge0.2	Zr0.35Hf0.35Ti0.3NiSn
25	Yb14MnSb11	CeFe3.5Co0.5Sb12
26	CeFe3CoSb12	Tl0.01Pb0.99Te
27	CeFe4Sb12	NbCoSn

28	Zr0.98Nb0.02NiSn	SrTi0.8Nb0.2O3
29	Zr0.5Hf0.5NiSn	Zr0.5Hf0.5NiSn
30	Mg2Si	In0.25Co4Sb12
31	ZrNiSn	Zr0.25Hf0.25Ti0.5NiSn
32	Cu4.0Mo6S8	La0.95Sr0.05CoO3
33	TiNiSn	LaCoO3
34	Ba0.3Sr0.6La0.1TiO3	In1.8Ge0.2O3
35	SrTi0.8Nb0.2O3	Zn4Sb3
36	In1.8Ge0.2O3	KBi6.33S10
37	Ca2.7Bi0.3Co4O9	La0.8Sr0.2CoO3
38	In1.985Ge0.015O3	CaMn0.98Nb0.02O3
39	Zn0.95Al0.05O	CaMnO3
40	Zn0.98Al0.02O	Ca5Al2Sb6
41	Sr0.9La0.1TiO3	Si0.8Ge0.2
42	Cr1.3Mo6S8	Mg2Si0.98Bi0.02
43	Sb2Te3	Sr0.9La0.1TiO3
44	Sr0.9Y0.1TiO3	Fe1.3Mo6S8
45	CuRh0.9Mg0.1O2	In1.985Ge0.015O3
46	ZrNiSn0.98Sb0.02	Mn1.3Mo6S8
47	CaMn0.98Nb0.02O3	Ni2.0Mo6S8
48	Ca0.9Yb0.1MnO3	Sr0.9Y0.1TiO3
49	NbCoSn	In2O3
50	Mn1.3Mo6S8	Ba0.3Sr0.6La0.1TiO3
51	CeFe3.5Co0.5Sb12	NaCo2O4
52	WO2.9	CaMn0.96Ru0.04O3
53	Fe1.3Mo6S8	Nd2CuO4
54	Ca0.9Bi0.1MnO3	Sm0.5Ca0.5MnO3
55	CaMnO3	CuCr0.97Mg0.03O2
56	Ca0.9Ho0.1MnO3	Cr1.3Mo6S8
57	Zn0.99Al0.01O	Ca2.7Bi0.3Co4O9
58	In2O3	CuCr0.95Mg0.05O2
59	CaMn0.96Ru0.04O3	Ca0.9Yb0.1MnO3
60	WO2.722	Sr0.61Ba0.39Nb2O6
61	Ca5Al2Sb6	Cu4.0Mo6S8
62	La0.95Sr0.05CoO3	Mg2Si
63	CuCr0.97Mg0.03O2	Ca0.9Bi0.1MnO3
64	Ni2.0Mo6S8	CuCr0.98Mg0.02O2
65	CuCr0.95Mg0.05O2	Ca3Co4O9
66	Ca0.9Nd0.1MnO3	CuRh0.9Mg0.1O2
67	BaPbO3	Zn0.98Al0.02O
68	K2Bi8Se13	Ca0.9Nd0.1MnO3
69	Mo3Te4	Zn0.95Al0.05O
70	CuCr0.98Mg0.02O2	CuCrO2
71	Tl2GeTe5	Ni0.02Zn0.98O
72	Nd2CuO4	Ba8Ga16Ge30
73	La0.8Sr0.2CoO3	BaPbO3
74	CuCrO2	Ca0.9Ho0.1MnO3
75	Sm0.5Ca0.5MnO3	WO2.9
76	ZnO	WO2.722
77	LaCoO3	Zn0.99Al0.01O
78	KBi6.33S10	ZnO
79	Ni0.02Zn0.98O	Ca2Co2O5

Table 3: Comparison of 84-material ranking by original experimental results and GPT embedding

Rank	Experiment	GPT Embedding
1	Cu2Se	Tl0.01Pb0.99Te

2	In0.25Co4Sb12	Tl0.02Pb0.98Te
3	Zr0.25Hf0.25Ti0.5NiSn	Sb2Te3
4	Ba8Ga16Ge30	Tl2GeTe5
5	NaCo2O4	Tl2SnTe5
6	Tl0.02Pb0.98Te	Bi2Te3
7	In0.2Co4Sb12	SrTi0.8Nb0.2O3
8	Na0.02Pb0.98Te	TiNiSn
9	Zn4Sb3	Sr0.9Y0.1TiO3
10	Zr0.35Hf0.35Ti0.3NiSn	Mo3Te4
11	Cu1.98Se	Na0.02Pb0.98Te
12	Bi2Sr2Co2O8	Ni2.0Mo6S8
13	Ca2Co2O5	Zr0.5Hf0.5NiSn
14	Tl9BiTe6	Zr0.98Nb0.02NiSn
15	Bi2Te3	Cu1.98Se
16	AgCrSe2	Sr0.9La0.1TiO3
17	Ca3Co4O9	ZrNiSn
18	Sr0.61Ba0.39Nb2O6	Cu2Se
19	In0.1Co4Sb12	Mg2Si
20	Mg2Si0.98Bi0.02	Zr0.35Hf0.35Ti0.3NiSn
21	Tl0.01Pb0.99Te	In1.8Ge0.2O3
22	LaFe3CoSb12	Fe1.3Mo6S8
23	Si0.8Ge0.2	In2O3
24	Yb14MnSb11	Fe1.94Ti0.06O3
25	CeFe3CoSb12	ZrNiSn0.98Sb0.02
26	CeFe4Sb12	In1.985Ge0.015O3
27	Zr0.98Nb0.02NiSn	Zr0.25Hf0.25Ti0.5NiSn
28	Zr0.5Hf0.5NiSn	Si0.8Ge0.2
29	Mg2Si	K2Bi8Se13
30	ZrNiSn	Cr1.3Mo6S8
31	Cu4.0Mo6S8	Ba8Ga16Ge30
32	TiNiSn	Cu4.0Mo6S8
33	Ba0.3Sr0.6La0.1TiO3	Bi2Sr2Co2O8
34	SrTi0.8Nb0.2O3	CuCrO2
35	In1.8Ge0.2O3	Tl9BiTe6
36	Ca2.7Bi0.3Co4O9	BaPbO3
37	In1.985Ge0.015O3	Zn4Sb3
38	Zn0.95Al0.05O	Nd2Cu0.98Ni0.02O4
39	Zn0.98Al0.02O	Mn1.3Mo6S8
40	Sr0.9La0.1TiO3	Ca0.9Yb0.1MnO3
41	Cr1.3Mo6S8	Ba0.4Sr0.6PbO3
42	Sb2Te3	Ni0.02Zn0.98O
43	Sr0.9Y0.1TiO3	Zn0.99Al0.01O
44	CuRh0.9Mg0.1O2	Mg2Si0.98Bi0.02
45	ZrNiSn0.98Sb0.02	Ba0.3Sr0.6La0.1TiO3
46	CaMn0.98Nb0.02O3	In0.2Co4Sb12
47	Ba0.4Sr0.6PbO3	In0.25Co4Sb12
48	Ca0.9Yb0.1MnO3	In0.1Co4Sb12
49	NbCoSn	LaCoO3
50	Mn1.3Mo6S8	Nd2CuO4
51	CeFe3.5Co0.5Sb12	AgCrSe2
52	WO2.9	Ca0.7Y0.3MnO3
53	Fe1.3Mo6S8	Ca3AlSb3
54	Ca0.9Bi0.1MnO3	Sr1.6La0.4Nb2O7
55	Ca3AlSb3	CeFe4Sb12
56	CaMnO3	ZnO
57	Ca0.9Ho0.1MnO3	La0.98Sr0.02CoO3
58	Zn0.99Al0.01O	Yb14MnSb11
59	In2O3	Zn0.98Al0.02O
60	Tl2SnTe5	CeFe3CoSb12

61	WO2.722	WO2.9
62	Ca5Al2Sb6	Sr0.61Ba0.39Nb2O6
63	La0.95Sr0.05CoO3	CeFe3.5Co0.5Sb12
64	CuCr0.97Mg0.03O2	Zn0.95Al0.05O
65	Fe1.94Ti0.06O3	La0.8Sr0.2CoO3
66	Ni2.0Mo6S8	Ca5Al2Sb6
67	Sr1.6La0.4Nb2O7	LaFe3CoSb12
68	CuCr0.95Mg0.05O2	Sm0.5Ca0.5MnO3
69	Ca0.7Y0.3MnO3	CaMnO3
70	Ca0.9Nd0.1MnO3	Ca2Co2O5
71	BaPbO3	La0.95Sr0.05CoO3
72	K2Bi8Se13	Ca0.9Bi0.1MnO3
73	La0.98Sr0.02CoO3	NaCo2O4
74	Mo3Te4	NbCoSn
75	CuCr0.98Mg0.02O2	WO2.722
76	Tl2GeTe5	CaMn0.98Nb0.02O3
77	Nd2CuO4	Ca0.9Nd0.1MnO3
78	La0.8Sr0.2CoO3	Ca0.9Ho0.1MnO3
79	Nd2Cu0.98Ni0.02O4	CuRh0.9Mg0.1O2
80	CuCrO2	Ca2.7Bi0.3Co4O9
81	Sm0.5Ca0.5MnO3	CuCr0.97Mg0.03O2
82	ZnO	CuCr0.98Mg0.02O2
83	LaCoO3	CuCr0.95Mg0.05O2
84	Ni0.02Zn0.98O	Ca3Co4O9