Incorporating Domain Knowledge into Materials Tokenization

Yerim Oh[1] **Jun-Hyung Park**[2] **Junho Kim**[1] **SungHo Kim**[1] **SangKeun Lee**[1,3]
[1]Department of Artificial Intelligence, Korea University
[2]Division of Language & AI, Hankuk University of Foreign Studies
[3]Department of Computer Science and Engineering, Korea University
{yerim0210, monocrat, sungho3268, yalphy}@korea.ac.kr, jhp@hufs.ac.kr

Abstract

While language models are increasingly utilized in materials science, typical models rely on frequency-centric tokenization methods originally developed for natural language processing. However, these methods frequently produce excessive fragmentation and semantic loss, failing to maintain the structural and semantic integrity of material concepts. To address this issue, we propose MATTER, a novel tokenization approach that integrates material knowledge into tokenization. Based on MatDetector trained on our materials knowledge base and a re-ranking method prioritizing material concepts in token merging, MATTER maintains the structural integrity of identified material concepts and prevents fragmentation during tokenization, ensuring their semantic meaning remains intact. The experimental results demonstrate that MATTER outperforms existing tokenization methods, achieving an average performance gain of 4% and 2% in the generation and classification tasks, respectively. These results underscore the importance of domain knowledge for tokenization strategies in scientific text processing.[1]

1 Introduction

Recent advances in language models have expanded their applications in materials science (Pilania, 2021; Olivetti et al., 2020). However, typical language models for materials science utilize frequency-centric subword tokenization methods originally developed for general natural language processing (NLP) tasks (Trewartha et al., 2022; Gupta et al., 2022; Huang and Cole, 2022). These methods prioritize high-frequency words in tokenization, resulting in misrepresentation of low-frequency words (Yuan et al., 2024; Lee et al., 2024; Liang et al., 2023), which is particularly problematic in material corpora.

[1]Our code is available at https://github.com/yerimoh/MATTER

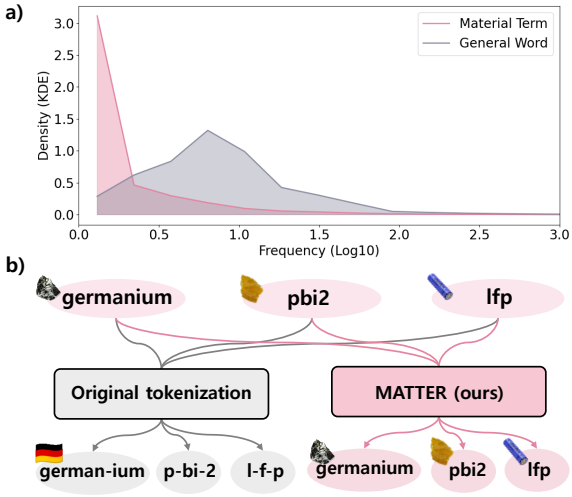

Figure 1: (a) Frequency histograms of material concepts and general words on 150K materials-related scientific papers. (b) Tokenization results of material concepts using conventional tokenization and MATTER (ours).

Material concepts—such as material names and chemical formulas—tend to appear infrequently in materials-related scientific papers as shown in Figure 1(a). This can lead to the oversight of material concepts in frequency-centric tokenization methods, whereas high-frequency general words dominate the subword vocabulary. As a result, material concepts are indeed fragmented into semantically unrelated subwords. For example, as shown in Figure 1(b), the word *germanium*, which means a chemical element, is split into semantically unrelated subwords *german* and *-ium*. Such fragmentation may cause language models to misinterpret the meaning of material concepts, resulting in performance degradation in materials science tasks. Several previous studies have also shown that preserving domain-specific subwords is crucial for maintaining model effectiveness (Gutiérrez et al., 2023; Gu et al., 2021; Hofmann et al., 2021), but how to identify and preserve such words remains unexplored in the materials science domain.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9623–9644

To address this issue, we propose **MATTER** (**Mat**erials **T**ok**e**nization F**r**amework), a novel approach that integrates material knowledge into tokenization. MATTER involves carefully designed frequency computation and merging processes to effectively capture the material concepts. We present MatDetector, a material concept identifier that scores each concept by its relevance to the materials domain, trained on a corpus of material knowledge that we carefully constructed. Subsequently, jointly considering the relevance scores and statistics of words, MATTER re-ranks the score of multiple possible merged tokens, prioritizing material-related subwords to be preserved. By integrating material knowledge into frequency computation and restructuring token merging, MATTER addresses the limitations of standard frequency-centric tokenization and enhances the representation of material concepts.

To verify the efficacy of MATTER, we conduct comprehensive experiments across diverse downstream tasks in materials science, including both generation and classification. The results demonstrate that MATTER significantly enhances performance on material-specific tasks while preserving the unique characteristics of material terminology. By integrating material knowledge into tokenization training, MATTER enables more precise learning of domain-specific concepts, underscoring the effectiveness of this tailored approach. In summary, this paper presents the following key contributions:

- We introduce MATTER, a novel domain-specific tokenization framework that integrates material knowledge into the tokenization process.
- We develop a novel scheme for materials tokenization based on MatDetector trained on our materials knowledge corpus integrated into our re-ranked token merging process.
- We demonstrate that MATTER outperforms existing tokenization methods, achieving an average improvement of 4% on generation tasks and 2% on classification tasks through extensive experiments.

2 Related Work

2.1 Subword Tokenization

Tokenization plays a crucial role in the performance of language models (Rust et al., 2021; Singh and Strouse, 2024; Wang et al., 2024a). One significant advancement in this area is subword tokenization, a pivotal approach in NLP. Various subword tokenization techniques exist, among which frequency-centric methods, such as Byte Pair Encoding (BPE; Gage 1994; Sennrich et al. 2016) and WordPiece (Wu, 2016), construct subword vocabularies by merging frequently co-occurring character sequences. Recent studies have explored integrating additional linguistic and contextual signals into tokenization. SAGE (Yehezkel and Pinter, 2023) introduces contextual embeddings to guide token segmentation, while PickyBPE (Chizhov et al., 2024) refines intermediate "junk" tokens.

However, while these methods effectively preserve high-frequency words, they often fragment low-frequency words, obscuring their meaning (Schmidt et al., 2024; Wu, 2016; Sennrich, 2015; Mikolov et al., 2012). Additionally, they are designed for general-domain corpora and fail to account for specialized terminology in the materials domain, where key concepts are both semantically significant and infrequent. As a result, conventional tokenization methods frequently split material concepts into unrelated subwords, disrupting their meaning. In contrast, MATTER is designed to address these domain-specific challenges in materials science.

2.2 Language Models in Materials Science

The discovery and practical application of materials is a time-intensive process, often spanning decades (Science and , US; Jain et al., 2013). To accelerate this process, leveraging the wealth of knowledge captured in textual datasets has become essential. NLP-based approaches have potential in materials informatics, enabling advancements in extracting and utilizing domain-specific knowledge (Wang et al., 2024b; Friedrich et al., 2020; Weston et al., 2019; Mysore et al., 2019). Tshitoyan et al. (2019) introduced embedding-based unsupervised methods, effectively capturing chemical knowledge and understanding chemical properties. Building on this foundation, Trewartha et al. (2022) introduced pre-trained language models trained on a materials science corpus, utilizing BERT (Devlin et al., 2019). Further extending the capabilities of BERT-based models, SciBERT (Beltagy et al., 2019), trained on material and battery-specific corpora, was adapted into MatSciBERT (Gupta et al., 2022) and BatteryBERT (Huang and Cole, 2022), respectively.

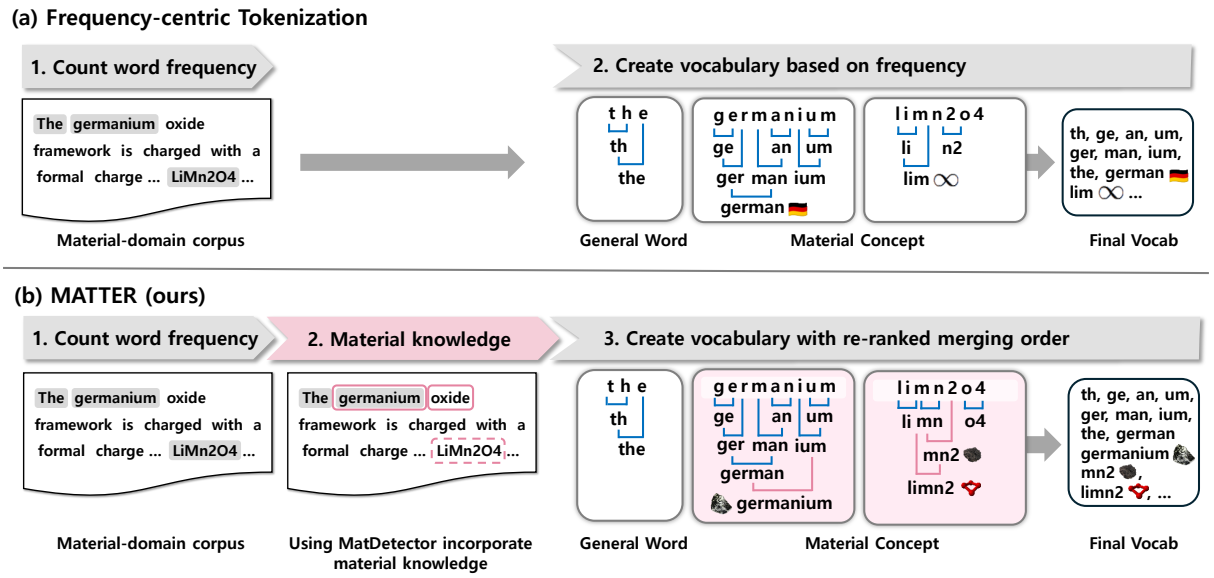

Figure 2: Comparison of the overall methodology between the existing frequency-centric tokenization and MATTER: (a) The existing frequency-centric tokenization creates the vocabularies based on word frequency. (b) In contrast, our approach, MATTER, incorporates material knowledge from MatDetector into subword vocabularies.

However, these models rely on tokenization strategies originally designed for general NLP tasks, which can be suboptimal for material specialized terminology. In contrast, MATTER introduces a tokenization approach tailored to the unique linguistic characteristics of the materials domain.

3 MATTER

We propose a materials-aware tokenization approach that integrates material knowledge into tokenization training and re-ranks token merging order. The overall procedure is illustrated in Figure 2.

3.1 Word Frequency Calculation

MATTER incorporates the WordPiece algorithm, a frequency-centric tokenization method, with material domain knowledge. The standard WordPiece algorithm first computes the frequency of each word in the corpus. Then, it tokenizes words into sequences of characters or byte units and iteratively merges the most frequent pair of tokens. Similarly, MATTER also computes word frequencies, denoted as $\text{freq}_{\text{origin}}(w)$ for a word w:

$$\text{freq}_{\text{origin}}(w) = \text{count}(w)$$

where $\text{count}(w)$ represents the number of occurrences of word w in the corpus. However, MATTER further incorporates material knowledge (§ 3.2) and re-ranks the token merging order (§ 3.3) to better preserve domain-specific terminology.

3.2 Material Knowledge Incorporation

To integrate material knowledge into MATTER, we adjust word frequencies (§ 3.1) by assigning weights to material concepts. Therefore, precise identification of material concepts is crucial. Traditionally, ChemDataExtractor (Kumar et al., 2024) has been widely used in materials science for this purpose, but since it was trained on biomedical data, its accuracy in identifying material concepts is limited (Kim et al., 2024; Kumar et al., 2024; Tran et al., 2024; Xu et al., 2023).

To address this, we introduce MatDetector, a material-agnostic tool that detects material concepts in a target corpus and assigns probability scores to each concept. Developed using the architecture of Trewartha et al. (2022), MatDetector is optimized for material concept detection. The dataset creation process is as follows:

Material Concept Extraction The MatDetector searches the PubChem database (Kim et al., 2019) for material-related concepts, extracting 80K material concepts (chemical names, IUPAC names, synonyms, and molecular formulas).

Material Corpus Crawling Using these concepts extracted from PubChem, we crawl Semantic Scholar, collecting around 42K scientific papers.

Crawled Data Tagging The collected corpus is tagged with PubChem material concepts, creating a NER material dataset with labels "material name",

Algorithm 1 MATTER Tokenization Training

Input: Corpus C, Vocabulary size V, MatDetector MD, Material importance factor λ
Output: Vocabulary $\mathcal{V}$ of size V (ordered)

procedure MATTER(C, V, MD, λ)
  $\mathcal{V} \leftarrow \{c \mid c \in C\}$ ▷ Unique characters
  $\text{freq}_{\text{origin}}(w) \leftarrow$ word frequency for all C
  $\hat{y}_{\text{mat}}(w) \leftarrow MD(w)$
  for all $w \in C$ **do** ▷ Re-ranking
    if $\hat{y}_{\text{mat}}(w) \neq \emptyset$ **then**
      $\text{freq}_{\text{mat}}(w) \leftarrow \text{freq}_{\text{origin}}(w) + \lambda \cdot \frac{\hat{y}_{\text{mat}}(w)}{1-\hat{y}_{\text{mat}}(w)}$
    else
      $\text{freq}_{\text{mat}}(w) \leftarrow \text{freq}_{\text{origin}}(w)$
    end if
  end for
  Compute Score(t_L, t_R) for all token pairs
  while $|\mathcal{V}| < V$ **do** ▷ Merge tokens
    $\langle t_L, t_R \rangle \leftarrow \arg\max_{(t_L,t_R)} \text{MatScore}(t_L, t_R)$
    $t_{\text{new}} \leftarrow t_L \oplus t_R$ ▷ Create new token
    $\mathcal{V} \leftarrow \mathcal{V} \cup \{t_{\text{new}}\}$
    C.ReplaceAll$(\langle t_L, t_R \rangle, t_{\text{new}})$ ▷ Update corpus
    Recompute Score for updated token set
  end while
  return $\mathcal{V}$
end procedure

"material formula", and "other". Labels other than "other" are treated as "material concept".

Data Augmentation While Semantic Scholar offers relatively clean data, material-related datasets from journals and repositories often contain formatting inconsistencies, OCR errors, and structural variations. To address this, we standardized common noise and expanded the dataset fourfold to enhance model robustness. Details in Appendix B.

Using the MatDetector, we can detect material concepts and compute their probability. Specifically, for Given a word w that is split into n subword tokens $\{t_1, t_2, ..., t_n\}$, the label for the word is determined as follows:

$$\hat{y}(w) = \arg\max_{c \in C} \frac{1}{n} \sum_{i=1}^{n} P(t_i, c) \quad (1)$$

where C is the set of all possible labels, and $P(t_i, c)$ denotes the probability of subword token t_i being classified as label c. If the predicted label $\hat{y}(w)$ falls under "material concept," we denote it as $\hat{y}_{\text{mat}}(w)$. The equation is as follows:

$$\hat{y}_{\text{mat}}(w) = \begin{cases} \hat{y}(w), & \text{if } \hat{y}(w) \in \{\text{material}\} \\ \emptyset, & \text{otherwise} \end{cases} \quad (2)$$

Ultimately, material concepts identified within the vocabulary are assigned $\hat{y}_{\text{mat}}(w)$, representing the likelihood of a word being relevant to the material domain. A higher probability value indicates stronger relevance to material concepts, ensuring that domain-specific concepts are effectively distinguished from general words.

3.3 Vocab Creation with Re-ranked Order

To integrate $\hat{y}_{\text{mat}}(w)$ into tokenization, we adjust word frequency computations by weighting material concepts based on their assigned probability scores. This adjustment prevents material concepts from being underrepresented, preserving their structural and semantic integrity during tokenization. To incorporate material information, we assign weighted frequencies to material concepts as follows:

Using this $\hat{y}_{\text{mat}}(w)$, MATTER adjusts the original frequency to prioritize material concepts. The adjusted frequency is computed as follows:

$$\text{freq}_{\text{mat}}(w) = \text{freq}_{\text{origin}}(w) + \lambda \cdot \frac{\hat{y}_{\text{mat}}(w)}{1 - \hat{y}_{\text{mat}}(w)} \quad (3)$$

With the adjusted frequency incorporating material knowledge, MATTER re-ranks the merging order based on incorporated material knowledge. Words are initially decomposed into sequences of characters or byte units, and the algorithm iteratively merges token pairs according to the re-ranked order guided by material relevance. The detailed algorithm is provided in Algorithm 1.

4 Experiments

4.1 Experimental Setups

Baselines To verify the efficacy of MATTER, we mainly compare ours with strong tokenization method: BPE (Sennrich et al., 2016), WordPiece (Wu, 2016), SAGE (Yehezkel and Pinter, 2023), PickyBPE (Chizhov et al., 2024). More hyperparameters are detailed in Appendix A.1. The detailed experimental setups are described in Appendix A.2.

Pre-training To evaluate the impact of tokenization on performance, we trained models using both baseline and MATTER specifically for the domain of materials science. Consistent with prior methodology (Gupta et al., 2022), we adopt SciBERT (Beltagy et al., 2019) as the encoder backbone for all experiments, due to its widespread use in materials-specific language modeling (Gupta et al., 2022; Huang and Cole, 2022; Kim et al., 2024). All models are trained with a fixed vocabulary size of 31,090 and a corpus of 150K materials science

Tokenization	Metric	Generation Task							
		NER	RC	EAE	PC	SAR	SC	SF	Overall
BPE	Micro-F1	$55.7_{\pm 0.4}$	$49.3_{\pm 0.2}$	$48.3_{\pm 0.8}$	$67.3_{\pm 0.1}$	$61.1_{\pm 1.8}$	$90.7_{\pm 2.4}$	$36.3_{\pm 1.4}$	$63.5_{\pm 0.5}$
(Sennrich et al., 2016)	Macro-F1	$47.1_{\pm 0.5}$	$47.2_{\pm 0.9}$	$36.3_{\pm 0.3}$	$40.2_{\pm 0.0}$	$41.8_{\pm 1.3}$	$47.6_{\pm 0.0}$	$16.7_{\pm 1.6}$	$42.0_{\pm 0.9}$
WordPiece	Micro-F1	$76.6_{\pm 0.2}$	$80.9_{\pm 0.3}$	$48.5_{\pm 0.2}$	$73.1_{\pm 0.5}$	$81.9_{\pm 0.4}$	$90.0_{\pm 0.1}$	$57.4_{\pm 0.2}$	$72.6_{\pm 0.1}$
(Wu, 2016)	Macro-F1	$56.1_{\pm 0.2}$	$58.5_{\pm 0.6}$	$29.4_{\pm 0.3}$	$58.9_{\pm 1.0}$	$74.6_{\pm 0.9}$	$60.3_{\pm 0.8}$	$32.6_{\pm 0.2}$	$52.9_{\pm 0.2}$
SAGE	Micro-F1	$77.0_{\pm 0.2}$	$82.3_{\pm 0.4}$	$47.3_{\pm 0.1}$	$68.3_{\pm 0.8}$	$77.1_{\pm 0.4}$	$90.9_{\pm 0.1}$	$57.1_{\pm 0.3}$	$71.4_{\pm 0.2}$
(Yehezkel and Pinter, 2023)	Macro-F1	$57.0_{\pm 0.3}$	$61.6_{\pm 0.4}$	$28.3_{\pm 0.3}$	$59.6_{\pm 1.3}$	$67.4_{\pm 0.9}$	$61.6_{\pm 0.8}$	$35.0_{\pm 0.3}$	$52.9_{\pm 0.3}$
PickyBPE	Micro-F1	$55.4_{\pm 0.1}$	**92.1**$_{\pm 0.1}$	$47.9_{\pm 0.4}$	$67.2_{\pm 0.0}$	$75.7_{\pm 0.2}$	$90.7_{\pm 0.0}$	$43.6_{\pm 0.1}$	$67.5_{\pm 0.1}$
(Chizhov et al., 2024)	Macro-F1	$41.7_{\pm 0.1}$	**65.1**$_{\pm 0.2}$	$36.5_{\pm 0.6}$	$40.2_{\pm 0.0}$	$66.1_{\pm 0.7}$	$47.6_{\pm 0.0}$	$23.1_{\pm 0.1}$	$45.8_{\pm 0.1}$
MATTER (ours)	Micro-F1	**80.0**$_{\pm 0.0}$	$83.8_{\pm 0.1}$	**53.1**$_{\pm 0.2}$	**73.7**$_{\pm 0.2}$	**85.5**$_{\pm 0.3}$	**91.2**$_{\pm 0.1}$	**61.9**$_{\pm 0.3}$	**75.6**$_{\pm 0.1}$
	Macro-F1	**59.3**$_{\pm 0.2}$	$59.1_{\pm 0.5}$	**36.9**$_{\pm 0.3}$	**67.6**$_{\pm 0.6}$	**79.3**$_{\pm 0.7}$	**64.9**$_{\pm 0.5}$	**38.0**$_{\pm 0.3}$	**57.9**$_{\pm 0.1}$

Table 1: Evaluation results on MatSci-NLP (generation tasks): The tasks encompass Named Entity Recognition (NER), Relation Classification (RC), Event Argument Extraction (EAE), Paragraph Classification (PC), Synthesis Action Retrieval (SAR), Sentence Classification (SC), and Slot Filling (SF). The best-performing results are highlighted in **boldface**.

papers. All training conditions—including model architecture, optimizer, and learning rate—are held constant across tokenizers to ensure fair comparison. In MATTER, the weighting parameter λ was set to 1 based on empirical analysis (see § 4.6), and further implementation details are provided in Appendix A.2.

Downstream Tasks and Datasets To comprehensively evaluate the performance of MATTER, we compare models trained with different tokenization methods on both generation and classification tasks. For generation tasks, we assess each baseline on the MatSci-NLP dataset (Song et al., 2023a), which includes seven materials-related tasks. We follow the MatSci-NLP benchmark protocol, which evaluates domain-specific encoders using a transformer-based schema decoder tailored for generation-based tasks. For classification tasks, we adopt four distinct benchmarks from prior work (Gupta et al., 2022), including named entity recognition (Weston et al., 2019; Friedrich et al., 2020), paragraph classification (Venugopal et al., 2021), and slot filling (Friedrich et al., 2020). These classification models are evaluated under standard encoder-only settings as used in prior work (Gupta et al., 2022). Detailed descriptions of evaluation metrics are provided in Appendix A.3.

4.2 Main Results

Generation Tasks Table 1 shows that MATTER outperforms existing tokenization methods, boosting Micro-F1 and Macro-F1 by 3% and 5% on average. These gains highlight MATTER's broad applicability across materials science tasks. Notably, SAGE and PickyBPE, which introduce non-material-specific signals, perform worse than WordPiece, emphasizing the need for domain-specific knowledge in tokenization. To further examine the generalizability of MATTER, we additionally evaluate its performance on materials-domain QA tasks using decoder-based and encoder-decoder models.[2]

Classification Tasks Similar to the generation tasks (Table 2), classification results confirm MATTER's superiority, with an average Micro-F1 and Macro-F1 improvement of 1.6% and 1.8%, respectively. These consistent gains highlight its robustness and ability to generalize across diverse materials science contexts, reinforcing its impact on materials informatics. To rigorously verify that these improvements are not attributed to random variation, we conducted paired t-tests for both generation and classification tasks. The detailed statistical analysis is presented in Appendix D, confirming

[2] See Appendix C for full details and results. MATTER outperforms other tokenization methods on the MaScQA benchmark, showing consistent gains in both model types.

Tokenization	Metric	Classification Task									
		NER_{SOFC}		$NER_{Matscholar}$		SF		RC		PC*	
		val	test	val	test	val	test	val	test	val	test
BPE	Micro-F1	$81.6_{\pm 0.2}$	$81.4_{\pm 0.1}$	$86.4_{\pm 0.3}$	$84.3_{\pm 0.5}$	$68.1_{\pm 0.5}$	$68.3_{\pm 0.6}$	$90.2_{\pm 0.4}$	$89.9_{\pm 0.0}$	$95.5_{\pm 0.0}$	$95.6_{\pm 0.0}$
(Sennrich et al., 2016)	Macro-F1	$80.7_{\pm 0.2}$	$78.9_{\pm 0.1}$	$85.0_{\pm 0.6}$	$82.9_{\pm 0.7}$	$65.5_{\pm 0.4}$	$59.3_{\pm 0.8}$	$86.4_{\pm 0.1}$	$85.5_{\pm 0.1}$		
WordPiece	Micro-F1	$82.0_{\pm 0.6}$	$80.9_{\pm 0.4}$	$88.8_{\pm 0.2}$	$86.1_{\pm 0.3}$	$67.4_{\pm 0.5}$	$\mathbf{60.4}_{\pm 0.7}$	$90.6_{\pm 0.2}$	$91.0_{\pm 0.7}$	$95.2_{\pm 0.1}$	$95.2_{\pm 0.1}$
(Wu, 2016)	Macro-F1	$83.0_{\pm 0.2}$	$83.0_{\pm 0.4}$	$87.6_{\pm 0.3}$	$85.8_{\pm 0.2}$	$69.2_{\pm 0.4}$	$69.6_{\pm 0.4}$	$86.3_{\pm 0.3}$	$87.5_{\pm 0.1}$		
SAGE	Micro-F1	$82.0_{\pm 0.2}$	$79.7_{\pm 0.4}$	$88.4_{\pm 0.3}$	$86.7_{\pm 0.4}$	$67.9_{\pm 0.5}$	$60.3_{\pm 0.4}$	$89.8_{\pm 0.4}$	$90.6_{\pm 0.3}$	$95.3_{\pm 0.0}$	$95.6_{\pm 0.2}$
(Yehezkel and Pinter, 2023)	Macro-F1	$82.7_{\pm 0.2}$	$82.5_{\pm 0.8}$	$87.6_{\pm 0.2}$	$86.1_{\pm 0.1}$	$\mathbf{69.7}_{\pm 0.3}$	$69.5_{\pm 0.6}$	$86.4_{\pm 0.7}$	$87.1_{\pm 0.0}$		
PickyBPE	Micro-F1	$77.3_{\pm 0.3}$	$78.8_{\pm 0.6}$	$84.1_{\pm 0.4}$	$83.4_{\pm 0.6}$	$62.0_{\pm 0.3}$	$60.2_{\pm 0.4}$	$88.6_{\pm 0.1}$	$85.8_{\pm 0.2}$	$95.7_{\pm 0.3}$	$95.8_{\pm 0.2}$
(Chizhov et al., 2024)	Macro-F1	$78.6_{\pm 0.4}$	$81.0_{\pm 0.7}$	$86.1_{\pm 0.3}$	$84.7_{\pm 0.5}$	$67.1_{\pm 0.1}$	$55.4_{\pm 0.2}$	$88.8_{\pm 0.6}$	$87.0_{\pm 0.2}$		
MATTER (ours)	Micro-F1	$\mathbf{83.1}_{\pm 0.2}$	$\mathbf{82.0}_{\pm 0.4}$	$\mathbf{89.6}_{\pm 0.1}$	$\mathbf{87.8}_{\pm 0.4}$	$\mathbf{68.4}_{\pm 0.1}$	$\mathbf{60.4}_{\pm 0.4}$	$\mathbf{90.9}_{\pm 0.2}$	$\mathbf{92.6}_{\pm 0.6}$	$\mathbf{96.9}_{\pm 0.1}$	$\mathbf{96.2}_{\pm 0.2}$
	Macro-F1	$\mathbf{84.3}_{\pm 0.2}$	$\mathbf{84.4}_{\pm 0.3}$	$\mathbf{88.6}_{\pm 0.2}$	$\mathbf{86.3}_{\pm 0.3}$	$\mathbf{69.7}_{\pm 0.4}$	$\mathbf{70.1}_{\pm 0.3}$	$\mathbf{87.3}_{\pm 0.4}$	$\mathbf{87.9}_{\pm 0.9}$		

Table 2: Evaluation results are presented across five classification tasks. Here, PC* represents accuracy, while the remaining metrics are reported as Micro-F1 and Macro-F1 scores. The best-performing results are highlighted in **boldface**.

	Train	Dev	Test	Total
English Set	458,692	57,371	57,755	573,818
Material Subset	16,286	2,010	2,173	20,469

Table 3: Statistics for the SIGMORPHON 2022 morpheme segmentation dataset and the material dataset, as described in Section 4.3.

Tokenization for MatSciBERT	Segmentation F1
WordPiece (Wu, 2016)	44.3
SAGE (Yehezkel and Pinter, 2023)	43.4
PickyBPE (Chizhov et al., 2024)	36.2
MATTER (ours)	**59.9**

Table 4: Material morpheme segmentation performance for different tokenization of the MatSciBERT model. The best-performing results are highlighted in **boldface**.

Tool	Recall	Precision	F1 Score
ChemDataExtractor	18%	57%	27%
MatDetector (ours)	**57%**	**69%**	**63%**

Table 5: Average performance of two material concept extraction tools on external materials NER datasets across all evaluation metrics.

that MATTER's performance gains are statistically significant across all major benchmarks.

4.3 Material Morpheme Segmentation

To validate MATTER's ability to segment material concepts into meaningful subwords, we evaluated its performance on the material subset of the SIGMORPHON dataset (Batsuren et al., 2022). The SIGMORPHON 2022 Shared Task provides a reliable benchmark for assessing whether words are segmented into morphologically meaningful units. For this analysis, we identified material concepts shared between SIGMORPHON, PubChem, and MatKG (Venugopal et al., 2022). The resulting subset, as shown in Table 3, revealed that approximately 20% of annotated words are relevant material concepts.

Using this subset, we evaluated the morpheme segmentation. As shown in Table 4, MATTER achieved an average improvement of 18.6% in segmentation accuracy compared to other tokenization algorithms. These results confirm that MATTER tokenization, effectively incorporates the characteristics of material corpora, enabling it to segment material concepts into meaningful subwords.

4.4 Extracted Material Concepts

Validation on Training Corpus To validate MatDetector on the training corpus, we constructed a reference lexicon of 100K material-related entries from PubChem and MatKG, including names, formulas, and synonyms. These were decomposed into 1.6M normalized tokens for broader coverage. Entities extracted from 150K materials papers were matched to the lexicon, and considered valid if found in the lexicon. MatDetector extracted 6×

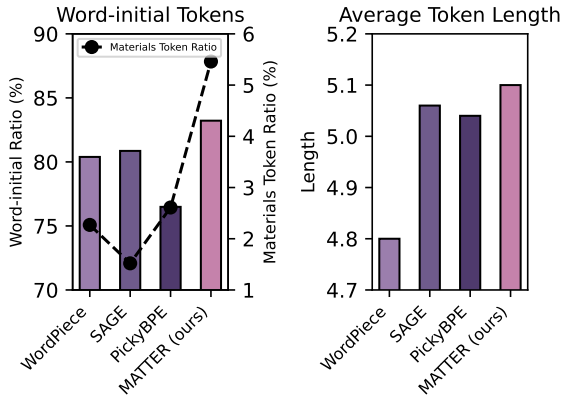

Figure 3: Comparison tokenization methods by word-initial token ratio (bar), materials token ratio (line), and average token length.

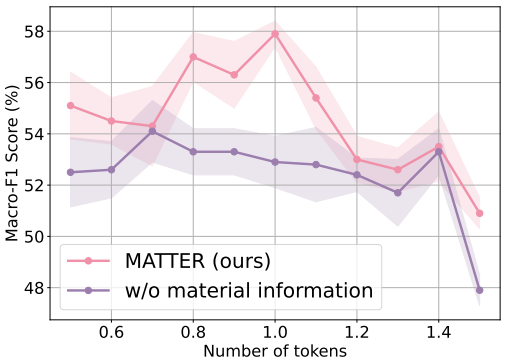

Figure 4: Comparison of Macro-F1 scores for MATTER and w/o material knowledge across during tokenization training different number of tokens.

more material concepts than ChemDataExtractor and achieved 64% higher match rate, confirming its precision and suitability for identifying material concepts in materials science corpus.

Validation on Materials NER To quantify absolute performance, we additionally evaluate both tools on two external materials NER datasets: MatScholar (Weston et al., 2019) and SOFC (Friedrich et al., 2020). As shown in Table 5, which reports the average performance across the two datasets, MatDetector consistently outperforms ChemDataExtractor across precision, recall, and F1 score. Notably, it achieves over twice the F1 score on average, highlighting its effectiveness not only in coverage but also in accurately identifying material entities. Detailed per-dataset results are provided in Appendix E. These results further validate MatDetector's ability to accurately and comprehensively detect material concepts in domain-specific NER tasks.

4.5 Token Qualities

To assess material token quality, we extract material-related tokens using MatDetector and compare tokenization methods.More hyperparameters are detailed in Appendix A.4.

Word-Initial Token One key aspect of token quality is the proportion of word-initial tokens, which help preserve word structure and meaning (Yehezkel and Pinter, 2023; Chizhov et al., 2024). For example, in tokenizing *"germanium"* into *"german"* and *"-ium"*, the word-initial token is "german". As shown in the left part of Figure 3, MATTER preserves a higher proportion of word-initial tokens (bar) compared to other methods. To evaluate this more rigorously, we further measured the materials-related word-initial token ratio (line) using a manually annotated set of approximately 9,000 material concepts, curated for downstream evaluation only (details in Appendix F). While this represents a small fraction of the full corpus, the results consistently demonstrate that MATTER achieves a significantly higher proportion of materials-related word-initial tokens, even on unseen datasets. This indicates that its vocabulary is enriched with material-specific terms, enabling better preservation of the semantic integrity of materials-related concepts.

Token Length According to Bostrom and Durrett (2020), longer mean token length reflects gold-standard morphologically-aligned tokenization, which enhances token quality. Based on this, we also measure mean token length. As shown in the right part of Figure 3, our method achieves a higher mean token length. Notably, it surpasses even SAGE and PickyBPE, which deliberately eliminate shorter intermediate tokens through compression at the cost of increased computational expense. This demonstrates that our approach not only maintains morphological alignment for material concepts but also preserves higher-quality tokenization.

Number of Tokens Figure 4 presents the experimental results comparing MATTER with a tokenizer trained without material knowledge during tokenization training. The number of tokens was varied from 0.5x to 1.5x of the original size. The results show that MATTER consistently outper-

Tokenization	Concept	Word Embedding	Sim.	Formula	Word Embedding	Sim.	Abbr	Word Embedding	Sim.
WordPiece	germanium	agilent	90.6	PbI2	nowak	81.8	LFP	inlet	95.5
	(german-ium)	fri	85.9	(pib-2)	10c	81.8	(lf-p)	chattopadhyay	93.7
SAGE	germanium	lot	83.0	PbI2	-gen	43.9	LFP	occupation	95.8
	(german-ium)	segregation	82.8	(p-ib-2)	pounds	43.6	(lf-p)	multiphonon	95.2
PickyBPE	germanium	nomin	81.2	PbI2	gaussian	63.1	LFP	her,	75.8
	(g-erman-ium)	inex	81.0	(p-bi-2)	p	62.8	(l-f-p)	consideration	75.0
MATTER (ours)	germanium	**dithiocarbamate**	81.5	PbI2	**pb5**	89.9	LFP	**zrf7**	90.9
	(germanium)	**ammonium**	81.4	(pbi2)	**pbf2**	89.2	(lfp)	**acyclohex**	90.8

Table 6: Comparison of subword embedding averaging results across different tokenization methods. The table presents the five nearest neighbor words based on subword embedding averages for each method. The similarity scores (Sim.) indicate the relevance of the nearest neighbors to the target material concept. **Boldface** highlights words that are directly related to materials.

forms the tokenizer trained without material knowledge in all cases. This demonstrates that providing material-specific information during tokenization training is crucial, regardless of the token count.

Subword Embedding Analysis Table 6 presents the two nearest neighbors of material concepts using cosine similarity. The results show that the nearest neighbors of MATTER are more material-specific and semantically relevant compared to other methods. For instance, while WordPiece and SAGE generate less relevant neighbors (*fri*, *segregation* for *germanium*), our method produces material concepts such as *dithiocarbamate* and *ammonium* for *germanium* . This indicates our tokenizer better preserves material-specific meanings, improving representation quality for scientific text.

Further inspection reveals that the learned subword embeddings capture a variety of chemically meaningful relationships. For example, pairs such as PbI_2 and PbF_2 belong to the same chemical family of lead halides, while *germanium* and *dithiocarbamate* co-occur as known compound pairs in Ge–S coordination complexes. Other relationships reflect compositional connections, such as the co-existence of *germanium* and *ammonium* in ammonium tris(oxalato)germanate, or functional similarity, as seen in *LFP* and ZrF_7, both of which are used in energy storage and sensing applications.

These findings support the claim that the embedding space goes beyond capturing surface-level co-occurrence, instead reflecting deeper, domain-relevant semantics. A more comprehensive analysis and additional examples can be found in Appendix G.2.

4.6 Ablation Study

Comparison of Detectors To confirm whether using MatDetector to extract material concepts and assign weights is more suitable for providing accurate and domain-relevant signals in the material domain compared to the widely used ChemDataExtractor, we performed ablation studies. Specifically, we replaced MatDetector with ChemDataExtractor to assign weights. While ChemDataExtractor is capable of partially extracting material concepts, it lacks the ability to assess the importance of the extracted concepts within the material domain. Consequently, all material concepts extracted by ChemDataExtractor were assigned the highest signal weight of 0.99.

Table 7 show that using MatDetector outperforms ChemDataExtractor, achieving a 2% higher average Micro-F1 score and a 2.7% higher Macro-F1 score. This confirms that MatDetector is more effective in providing material domain-relevant signals. Additionally, when examining the performance of ChemDataExtractor, we observed that it achieved a 1.1% higher Micro-F1 score and a 2.3% higher Macro-F1 score compared to the baseline method, which did not incorporate any material signals. This underscores the importance of incorporating material signals into tokenization.

However, as evidenced by the performance gap between ChemDataExtractor and MatDetector, it is clear that the accuracy of the material signals plays a critical role. These results highlight the necessity of not only incorporating material signals but also ensuring that accurate material concepts and their respective significance are properly consid-

Tokenization	Metric	MatSci-NLP							
		NER	RC	EAE	PC	SAR	SC	SF	Overall
w/o material knowledge (WordPiece)	Micro-F1	$76.6_{\pm 0.2}$	$80.9_{\pm 0.3}$	$48.5_{\pm 0.2}$	$73.1_{\pm 0.5}$	$\underline{81.9}_{\pm 0.4}$	$90.0_{\pm 0.1}$	$57.4_{\pm 0.2}$	$72.6_{\pm 0.1}$
	Macro-F1	$56.1_{\pm 0.2}$	$58.5_{\pm 0.6}$	$29.4_{\pm 0.3}$	$58.9_{\pm 1.0}$	$\underline{74.6}_{\pm 0.9}$	$60.3_{\pm 0.8}$	$32.6_{\pm 0.2}$	$52.9_{\pm 0.2}$
ChemDataExtractor (Swain and Cole, 2016)	Micro-F1	$\underline{77.1}_{\pm 1.1}$	$\underline{81.5}_{\pm 0.7}$	$\underline{53.1}_{\pm 3.5}$	$\underline{73.6}_{\pm 0.6}$	$80.6_{\pm 3.3}$	$\underline{91.2}_{\pm 1.0}$	$\underline{58.8}_{\pm 2.5}$	$\underline{73.7}_{\pm 1.3}$
	Macro-F1	$\underline{56.4}_{\pm 1.3}$	$\underline{58.9}_{\pm 0.7}$	$\underline{35.0}_{\pm 4.2}$	$\underline{67.6}_{\pm 1.2}$	$68.0_{\pm 9.5}$	$\underline{64.8}_{\pm 0.1}$	$\underline{35.6}_{\pm 1.6}$	$\underline{55.2}_{\pm 2.8}$
MatDetector (ours)	Micro-F1	$\mathbf{80.0}_{\pm 0.0}$	$\mathbf{83.8}_{\pm 0.1}$	$\mathbf{53.1}_{\pm 0.2}$	$\mathbf{73.7}_{\pm 0.2}$	$\mathbf{85.5}_{\pm 0.3}$	$\mathbf{91.2}_{\pm 0.1}$	$\mathbf{61.9}_{\pm 0.3}$	$\mathbf{75.6}_{\pm 0.1}$
	Macro-F1	$\mathbf{59.3}_{\pm 0.2}$	$\mathbf{59.1}_{\pm 0.5}$	$\mathbf{36.9}_{\pm 0.3}$	$\mathbf{67.6}_{\pm 0.6}$	$\mathbf{79.3}_{\pm 0.7}$	$\mathbf{64.9}_{\pm 0.5}$	$\mathbf{38.0}_{\pm 0.3}$	$\mathbf{57.9}_{\pm 0.1}$

Table 7: Ablation results on different detectors for the MatSci-NLP dataset across multiple tasks. w/o material knowledge represents frequency-centric tokenization without any additional signal. ChemDataExtractor and MatDetector incorporate additional signals using their respective tools. **Bold** values indicate the highest scores for each metric-task pair, while underline represent the second-highest scores.

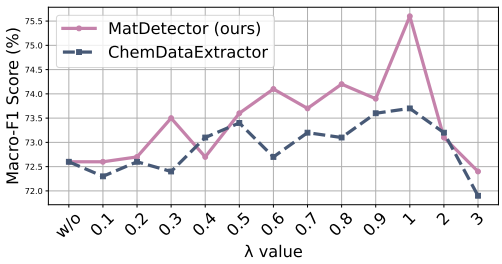

Figure 5: Comparison of Macro-F1 scores for ChemDataExtractor and MatDetector across λ values.

ered. The use of MatDetector effectively addresses both aspects, demonstrating its suitability for enhancing performance in the material domain. Both detectors achieved their highest performance at a λ value of 1 as show in Figure 5.

Comparison of Lambda Figure 5 demonstrates that adding material signals, regardless of the weighting method used, consistently yields better performance compared to the baseline where no material signals were incorporated (λ =0). This observation aligns with previous findings and further substantiates that the inclusion of material Knowledge is beneficial. Moreover, it emphasizes the necessity of using appropriate tools to effectively assign these signals for optimal performance.

Notably, both ChemDataExtractor and MatDetector achieved their highest performance at $\lambda = 1$. Based on this consistent observation across models, all preceding experiments in this study were conducted using this optimal setting.

5 Conclusion

We proposed MATTER, a novel tokenization approach that incorporates material knowledge derived from material corpora into the tokenization process. MATTER has enabled the creation of vocabularies tailored to the material domain, effectively maintaining the structure and semantics of material concepts. Our extensive experiments have demonstrated that MATTER tokenization significantly improves performance across a wide range of material generation and classification tasks, outperforming conventional tokenization methods. Our work has provided a strong, adaptable foundation components for materials NLP, empowering future research on materials science.

Limitations

While we have demonstrated that MATTER effectively enhances tokenization for pretrained language models in the materials science domain. Nevertheless, our work also opens several valuable opportunities for further improvements and exploration.

Hyperparameter Selection. MATTER introduces a tunable hyperparameter (λ) to balance frequency statistics with material-specific signals during vocabulary construction. While we observed stable improvements across a range of λ values, the method still requires manual selection of this parameter. Although $\lambda = 1$ was found to be effective in our experiments, identifying an optimal value for different domains or corpora may require

additional tuning. This reliance on hyperparameter selection may affect general usability in practice.

Further Analysis on Corpus The current experiments were conducted following the prior methodology outlined in (Gupta et al., 2022), which emphasizes the use of material-specialized corpora. Although this ensures consistency and relevance to domain-specific evaluation, future work may benefit from expanding the diversity of training corpora to test MATTER's generalizability across subdomains and heterogeneous sources.

NER Dependency and Scalability Our approach relies on the identification of material concepts through NER-based classification. To support this, we constructed a high-quality NER dataset using a curated materials knowledge base, ensuring accurate detection of domain-specific terminology essential for effective vocabulary construction in materials science. However, this reliance on supervised signals may introduce challenges in scalability, particularly when applied to broader or less-structured corpora. Addressing this limitation remains an important direction for future work.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.RS-2025-00517221 and No.RS-2024-00415812) and Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2024-00439328, Karma: Towards Knowledge Augmentation for Complex Reasoning (SW Starlab), No.RS-2024-00457882, AI Research Hub Project, and No.RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University)).

References

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, et al. 2018. Construction of the literature graph in semantic scholar. *arXiv preprint arXiv:1805.02262*.

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtskỳ, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, et al. 2022. The sigmorphon 2022 shared task on morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 103–116.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. *arXiv preprint arXiv:2004.03720*.

Yizhou Chen, Seira Yamaguchi, Atsushi Sato, Dong Xue, and Kazuhiro Marumoto. 2025. Operando spin observation elucidating performance-improvement mechanisms during operation of ruddlesden–popper sn-based perovskite solar cells. *npj Flexible Electronics*, 9(1):1.

Pavel Chizhov, Catherine Arnett, Elizaveta Korotkova, and Ivan Yamshchikov. 2024. Bpe gets picky: Efficient vocabulary refinement during tokenizer training. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16587–16604.

Qingchen Deng, Jiangen Li, Xiang Li, Xuye Du, Lanlan Wu, Junrui Wang, and Xinlong Wang. 2024. Incorporating nano-znco-zif particles in the electrospinning polylactide membranes to improve their filtration and antibacterial performances. *Polymer Bulletin*, 81(15):14067–14081.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Maruscyk, and Lukas Lange. 2020. The sofc-exp corpus and neural approaches to information extraction in the materials science domain. *arXiv preprint arXiv:2006.03039*.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.

Darren S Gray, Joe Tien, and Christopher S Chen. 2004. High-conductivity elastomeric electronics. *Advanced Materials*, 16(5):393–397.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Tanishq Gupta, Mohd Zaki, NM Anoop Krishnan, and Mausam. 2022. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102.

Bernal Jiménez Gutiérrez, Huan Sun, and Yu Su. 2023. Biomedical language models are robust to sub-optimal tokenization. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 350–362.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves bert's interpretation of complex words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3594–3608.

Shu Huang and Jacqueline M Cole. 2022. Batterybert: A pretrained language model for battery database enhancement. *Journal of chemical information and modeling*, 62(24):6365–6377.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. 2013. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1).

Junho Kim, Yeachan Kim, Jun-Hyung Park, Yerim Oh, Suho Kim, and SangKeun Lee. 2024. Melt: Materials-aware continued pre-training for language model adaptation to materials science. *In Findings of the Association for Computational Linguistics: EMNLP. Association for Computational Linguistics*.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. 2019. Pubchem 2019 update: improved access to chemical data. *Nucleic acids research*, 47(D1):D1102–D1109.

Pankaj Kumar, Saurabh Kabra, and Jacqueline M Cole. 2024. a database of stress-strain properties autogenerated from the scientific literature using chemdataextractor. *Scientific Data*, 11(1):1273.

Jungseob Lee, Hyeonseok Moon, Seungjun Lee, Chanjun Park, Sugyeong Eo, Hyunwoong Ko, Jaehyung Seo, Seungyoon Lee, and Heui-Seok Lim. 2024. Length-aware byte pair encoding for mitigating oversegmentation in korean machine translation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2287–2303.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. Xlm-v: Overcoming the vocabulary bottleneck in multilingual masked language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13142–13152.

Tomáš Mikolov, Ilya Sutskever, Anoop Deoras, Hai-Son Le, Stefan Kombrink, and Jan Cernocky. 2012. Subword language modeling with neural networks. *preprint (http://www. fit. vutbr. cz/imikolov/rnnlm/char. pdf)*, 8(67).

Sheshera Mysore, Zach Jensen, Edward Kim, Kevin Huang, Haw-Shiuan Chang, Emma Strubell, Jeffrey Flanigan, Andrew McCallum, and Elsa Olivetti. 2019. The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures. *arXiv preprint arXiv:1905.06939*.

Elsa A Olivetti, Jacqueline M Cole, Edward Kim, Olga Kononova, Gerbrand Ceder, Thomas Yong-Jin Han, and Anna M Hiszpanski. 2020. Data-driven materials research enabled by natural language processing and information extraction. *Applied Physics Reviews*, 7(4).

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

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135.

Craig W Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. 2024. Tokenization is more than compression. *arXiv preprint arXiv:2402.18376*.

National Science and Technology Council (US). 2011. *Materials genome initiative for global competitiveness*. Executive Office of the President, National Science and Technology Council.

Rico Sennrich. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Aaditya K Singh and DJ Strouse. 2024. Tokenization counts: the impact of tokenization on arithmetic in frontier llms. *arXiv preprint arXiv:2402.14903*.

Yu Song, Santiago Miret, and Bang Liu. 2023a. Matscinlp: Evaluating scientific language models on materials science language tasks using text-to-schema modeling. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). Association for Computational Linguistics*.

Yu Song, Santiago Miret, Huan Zhang, and Bang Liu. 2023b. Honeybee: Progressive instruction finetuning of large language models for materials science. *arXiv preprint arXiv:2310.08511*.

Matthew C Swain and Jacqueline M Cole. 2016. Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature. *Journal of chemical information and modeling*, 56(10):1894–1904.

Huan Tran, Rishi Gurnani, Chiho Kim, Ghanshyam Pilania, Ha-Kyung Kwon, Ryan P Lively, and Rampi Ramprasad. 2024. Design of functional and sustainable polymers assisted by artificial intelligence. *Nature Reviews Materials*, pages 1–21.

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

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

Vineeth Venugopal, Sumit Pai, and Elsa Olivetti. 2022. The largest knowledge graph in materials science-entities, relations, and link prediction through graph representation learning. In *AI for Accelerated Materials Design NeurIPS 2022 Workshop*.

Vineeth Venugopal, Sourav Sahoo, Mohd Zaki, Manish Agarwal, Nitya Nand Gosvami, and NM Anoop Krishnan. 2021. Looking through glass: Knowledge discovery from materials science literature using natural language processing. *Patterns*, 2(7).

Dixuan Wang, Yanda Li, Junyuan Jiang, Zepeng Ding, Guochao Jiang, Jiaqing Liang, and Deqing Yang. 2024a. Tokenization matters! degrading large language models through challenging their tokenization. *arXiv preprint arXiv:2405.17067*.

Lei Wang, Fei Wu, Xiaoqing Liu, Chong Wang, Wanxin Wang, Mingshi Cui, and Zhaoyang Qu. 2024b. A joint extraction method for fault text entity relationships in smart grid considering nested entities and complex semantics. *Energy Reports*, 11:6150–6159.

Leigh Weston, Vahe Tshitoyan, John Dagdelen, Olga Kononova, Amalie Trewartha, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *Journal of chemical information and modeling*, 59(9):3692–3702.

Yonghui Wu. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Pengcheng Xu, Xiaobo Ji, Minjie Li, and Wencong Lu. 2023. Small data machine learning in materials science. *npj Computational Materials*, 9(1):42.

Shaked Yehezkel and Yuval Pinter. 2023. Incorporating context into subword vocabularies. *In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 623—-635.

Fei Yuan, Shuai Yuan, Zhiyong Wu, and Lei Li. 2024. How vocabulary sharing facilitates multilingualism in llama? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12111–12130.

Mohd Zaki, NM Anoop Krishnan, et al. 2024. Mascqa: investigating materials science knowledge of large language models. *Digital Discovery*, 3(2):313–327.

Lin Zhang, Zonghui Lu, Zhe Su, Ye Zhang, and Hui He. 2025. Efficiency of carbothermal reduction in treating norm waste containing ba (226ra) so4. *Journal of Radioanalytical and Nuclear Chemistry*, pages 1–8.

A Implementation Details and Setups

A.1 Tokenization baseline

We compared our tokenization approach against baseline methods, including the widely used frequency-centric tokenization, WordPiece, as well as more recent and strong tokenization methods, SAGE and PickyBPE. To ensure a fair comparison, all tokenization methods adhered to the vocabulary size of 31,090, as defined in the prior methodology (Gupta et al., 2022). The implementation details are as follows:

WordPiece (Wu, 2016) being one of the most widely used and fundamental frequency-centric tokenization methods, was configured with a min frequency of 2 and a limit alphabet of 6,000.

SAGE (Yehezkel and Pinter, 2023) enhances frequency-centric tokenization by incorporating contextual signals into the process. The implementation of SAGE included several key parameters: vocabulary schedule progressively reducing from 32,000 to the target size of 31,090; an embedding schedule synchronized with the vocabulary schedule; a maximum token length of 17 bytes; and the use of skip-gram embedding training with a vector size of 50, a context window size of 5, and 15 negative samples. To ensure reproducibility, the random seed was set to 692,653.

PickyBPE (Chizhov et al., 2024) was employed to construct the vocabulary, with a desired vocabulary size of 31,090 and an IoS (Importance of

Symbols) threshold set to 0.9. The initial vocabulary ensured comprehensive coverage with a relative symbol coverage of 0.9999. During training, the frequency of merges was logged at intervals of 200 merges to monitor the tokenization process effectively.

A.2 Hyper-parameters

Pre-Train. We follow the previous work (Gupta et al., 2022). The detailed configuration of the main model and training hyperparameters is summarized as follows:

Parameter	Value
Encoder Layers	12
Embedding Dim	768
Hidden Dim	768
Attention Heads	12
Max Tokens in a Batch	128
Optimizer	Adam
Weight Decay	0.01
Learning Rate (LR)	2e-5
LR Scheduler	Linear with Warmup
Warmup Strategy	Linear
Precision	FP16

A.3 Evaluation metrics

Classification task. We follow the previous work (Gupta et al., 2022). The detailed configuration of the main model and training hyperparameters is summarized in Table 8.

Generation task. We follow the previous work (Song et al., 2023a). The detailed configuration of the main model and training hyperparameters is summarized as follows:

Parameter	Value
Decoder Layers	3
Embedding Dim	768
Hidden Dim	768
Attention Heads	8
Max Tokens in a Batch	4
Optimizer	Adam
Learning Rate (LR)	2e-5
Precision	FP32
Training Epochs	Up to 20 (early stopping)

We evaluate using metrics from MatSciNLP (Song et al., 2023a) and MatSciBERT (Gupta et al., 2022). Generation tasks use Micro-F1 and Macro-F1, averaged over five seeds. Classification tasks report Macro-F1 (SOFC-NER, SOFC-Filling), Micro-F1 (MatScholar), and accuracy (Glass Science), with cross-validation over five folds and three seeds.

A.4 Token Qualities Details

Among 31,090 vocabulary entries, we extract material-related tokens using MatDetector and compare tokenization methods.

Tokenization	#material token
WordPiece	10,420
SAGE	9,602
PickyBPE	9,203
MATTER (ours)	**10,633**

B Details of MatDetector Construction

MatDetector is a domain-specific Named Entity Recognition (NER) tool designed to extract material concepts from scientific texts. The detailed steps for its construction are as follows:

Crawling Material Corpus To construct the training dataset for the MatDetector, we first extract chemical names, IUPAC names, synonyms, and molecular formulas from PubChem (Kim et al., 2019), obtaining 80K material concepts. The number of concepts by category is provided in Table (Kim et al., 2019). Using these extracted concepts as keywords, we collect 42K scientific papers from Semantic Scholar (Ammar et al., 2018), focusing on titles and abstracts that contain high-density material knowledge, with detailed comparative information in Table 9.

Creating Train Dataset While Semantic Scholar provides relatively clean data, most material-related data is collected from various journals and repositories, where formatting inconsistencies, OCR errors, and structural variations introduce significant noise. To address this, we construct a Noisy NER Dataset, improving model robustness and expanding the dataset to be four times larger than the original. The details of noise augmentation are as follow:

- **Material Name Noise**: This includes capitalization errors in element symbols, misplaced

Parameter	Value				
	NER_{SOFC}	$NER_{Matscholar}$	SF	RC	PC
Max Tokens in a Batch	32	32	32	64	64
Training Epochs	20	15	40	10	10
Optimizer	Adam				
Learning Rate (LR)	[2e-5, 3e-5, 5e-5]				
LR Scheduler	Linear with Warmup				
Warmup Strategy	Warmup ratio of 0.1				
Precision	FP32				

Table 8: Detailed configuration of the main model and training hyperparameters for classification task.

Tool		ChemDataExtractor (Swain and Cole, 2016)	MatDetector (ours)
Entity type		chemical mention	chemical concepts, formulas
Train data	domain	BioMedical	Material
	annotated method	manually annotated	manually & automatic annotated
	number of abstract	10,000	404,262

Table 9: Comparison of Extractable Entity Types and Training Data in ChemDataExtractor and MatDetector.

or duplicated digits, reordering of elements, and insertion of unnecessary characters or special symbols. These modifications reflect common errors found in chemical names and mimic the inconsistencies in scientific documents.

- **Material Formula Noise**: Common formatting inconsistencies in formulas are simulated by adding spaces around special symbols such as `(`, `)`, `[`, and `]`, or by replacing digits with placeholders. Combined patterns are also introduced to replicate multiple error types.

Using this dataset, we generate a material NER dataset by tagging the collected corpus with material concepts extracted from PubChem, ensuring precise identification of material-related terminology. In this tagging process, *Material Name*, IUPAC Name, and Synonym of Material Name are categorized as Material Concept, while Material Formula is tagged separately as *Material Formula*. This approach maintains a clear distinction between conceptual material entities and their chemical formulas, enabling more accurate entity recognition in materials science applications.

Training the MatDetector We train the MatDetector using the material NER dataset constructed in the previous step and the Trewartha et al. (2022) model architecture. The model achieves high accuracy in detecting material concepts, even in noisy corpora, and provides NER tagging probabilities, estimating the likelihood that a concepts belongs to materials science.

C Additional QA Experiments on MaScQA

To evaluate the generalizability of MATTER beyond classification tasks, we conducted additional experiments on the MaScQA (Zaki et al., 2024) benchmark, which focuses on materials-domain question answering.

Decoder-based setup. We fine-tuned two decoder-based models— Llama-3.2-1B-Instruct and SciBERT—on the HoneyBEE (Song et al., 2023b) instruction dataset and evaluated their performance on MaScQA. MATTER consistently achieved higher accuracy compared to other tokenizations:

Encoder-decoder setup. Following the setup in MatSciNLP (Song et al., 2023a), we used MatSciBERT as the encoder and a transformer-based decoder. We trained on 10% of the HoneyBEE QA data and evaluated on the remaining 90%, simulat-

Tokenization	Accuracy (%)
BPE	7.1
PickyBPE	7.3
MATTER (ours)	**8.9**

Table 10: MaScQA benchmark accuracy using decoder-based models.

ing a low-resource QA scenario. MATTER again yielded the best performance:

Tokenization	Accuracy (%)
BPE	20.01
WordPiece	22.74
SAGE	22.93
PickyBPE	21.01
MATTER (ours)	**23.96**

Table 11: MaScQA benchmark performance with encoder-decoder model.

These results confirm MATTER's effectiveness in enhancing QA performance across diverse model architectures and reinforce its generalizability to downstream materials tasks.

D Statistical Significance

Generation task To quantitatively assess the statistical significance of performance improvements introduced by MATTER, we conducted paired t-tests on the *average F1 scores* across eight generation tasks (NER, RC, EAE, PC, SAR, SC, SF, Overall), comparing MATTER against four widely-used tokenization baselines: BPE, WordPiece, SAGE, and PickyBPE. The average F1 score was computed as the arithmetic mean of the Micro-F1 and Macro-F1 values for each task.

The paired t-test evaluates whether the mean difference in Avg-F1 scores between MATTER and a baseline is statistically significant. The t-statistic is given by:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \tag{4}$$

where $\bar{d}$ is the mean of the differences between MATTER and a baseline across tasks, s_d is the standard deviation of those differences, and $n = 8$ is the number of generation tasks.

Tokenization	Avg-F1 (p)	Significant
BPE	0.0009	Yes
WordPiece	0.0001	Yes
SAGE	0.0066	Yes
PickyBPE	0.0155	Yes

Table 12: Paired t-test results comparing the average F1 score between MATTER and each baseline across generation tasks.

As shown in Table 12, MATTER achieves statistically significant improvements over all four baselines in terms of average F1 score. All comparisons yield $p < 0.05$, confirming that MATTER's performance gains are unlikely due to random variation. These results reinforce the effectiveness of MATTER's domain-aware tokenization strategy in improving generation performance across diverse material-related tasks.

Classification task We conducted the same analysis for classification tasks to evaluate whether MATTER's improvements generalize to discriminative settings. Paired t-tests were performed on the *average F1 scores* across five classification tasks (SOFC-NER, MatScholar-NER, SF, RC, PC), using the same computation.

Tokenization	Avg-F1 (p)	Significant
BPE	0.0001	Yes
WordPiece	0.0001	Yes
SAGE	0.0021	Yes
PickyBPE	0.0009	Yes

Table 13: Paired t-test results comparing the average F1 score between MATTER and each baseline across classification tasks.

As shown in Table 13, all comparisons again yield statistically significant results ($p < 0.005$), confirming that MATTER consistently outperforms all baselines in overall classification performance. This aggregated F1-based analysis further demonstrates the robustness of MATTER's tokenization advantages in both generation and classification tasks, effectively balancing frequency-weighted and class-balanced evaluation perspectives.

E Details of validation on materials NER

Tool	MatScholar	SOFC	Overall
ChemDataExtractor	12%	24%	18%
MatDetector (ours)	**53%**	**60%**	**57%**

Table 14: Recall of two material concept extraction tools on external materials NER datasets—MatScholar and SOFC.

Tool	MatScholar	SOFC	Overall
ChemDataExtractor	52%	61%	57%
MatDetector (ours)	**63%**	**75%**	**69%**

Table 15: Precision of two material concept extraction tools on external materials NER datasets—MatScholar and SOFC.

Tool	MatScholar	SOFC	Overall
ChemDataExtractor	20%	34%	27%
MatDetector (ours)	**58%**	**67%**	**63%**

Table 16: F1 Score of two material concept extraction tools on external materials NER datasets—MatScholar and SOFC.

F Details of the Word-Initial Token Analysis

To validate the effectiveness of our tokenization and avoid any potential circularity in evaluation, we perform an additional analysis using external and independent sources of material-related terms, separate from those used to construct the tokenization. Specifically, we collect named entities from two manually annotated materials NER datasets used in the paper:

NER Dataset	#Material Entity
MatScholar (Weston et al., 2019)	8,660
SOFC (Friedrich et al., 2020)	1,201
Total	9,861

G Case Study: Tokenization Robustness

G.1 Analysis in Material Science Papers

In this section, we applied WordPiece, SAGE, PickyBPE, and the proposed method, MATTER, to tokenization results from real materials science papers. As shown in Table 18, existing tokenization methods such as WordPiece, SAGE, and PickyBPE tend to overtokenize important material concepts. For instance, the chemical formula for Lead, "Pb", is split into "p-b", while "dimethylsiloxane" is divided into "dimethyl-sil-oxane or d-imethyl-sil-oxane". Such overtokenization distorts the semantic integrity of material concepts and can degrade the performance of downstream natural language processing tasks.

In contrast, our proposed MATTER method effectively prevents the overtokenization of material concepts. When applying MATTER, essential material concepts such as "Pb", "dimethylsiloxan"e, and "barium sulfate" remain intact, preserving their contextual meaning. Notably, complex material concepts such as "perovskite" and "ethylenediaminetetraacetic acid" are properly maintained, demonstrating that MATTER provides a more suitable tokenization approach for materials science texts.

G.2 Subword Embedding Analysis

To evaluate the impact of different tokenization methods on word representations in materials science, we analyze the nearest neighbors of material concepts based on subword embedding averaging. This experiment is conducted in conjunction with the tokenization results presented in Figure 1 and Table 18, allowing us to assess how subword segmentation affects semantic consistency in word embeddings. We compare four tokenization strategies—WordPiece, SAGE, PickyBPE, and our proposed method, MATTER—by computing word embeddings as the mean of their constituent subword embeddings. The similarity between words is measured using cosine similarity, and the five nearest neighbors (5-NN) for each concept are retrieved. The retrieved neighbors allow us to assess whether the tokenization method preserves materials science semantics or introduces artifacts from suboptimal subword segmentation. The dataset used for evaluation includes materials science terminology, chemical formulas, and domain-specific abbreviations, ensuring a realistic assessment of tokenization impact.

The results, presented in Table 19, indicate that WordPiece and SAGE exhibit a strong tendency to retrieve words that share surface-level subword structures rather than those with true material relevance. For instance, 'germanium' is tokenized as german-ium in WordPiece, leading to nearest neigh-

Type of material concept	#material concept
Material name	22,482
IUPAC name	22,482
Synonym of material name	719,885
Material formula	22,479

Table 17: Summary of approximately 80K extracted material concepts from PubMed, categorized by concepts type.

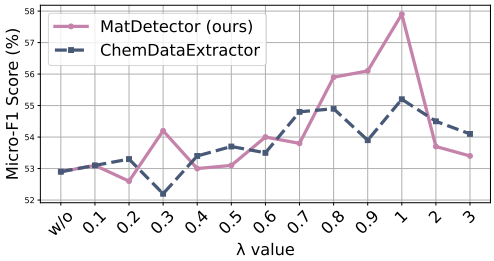

Figure 6: Comparison of Micro-F1 scores for ChemDataExtractor and MatDetector across different λ values.

bors such as 'german' and '-ium', which lack meaningful chemical associations. PickyBPE partially alleviates this issue by merging frequent subwords, but still retrieves words that reflect tokenization artifacts rather than conceptually related material concepts. In contrast, our MATTER method significantly improves semantic alignment by retrieving chemically relevant words. For example, the nearest neighbors of 'germanium' include 'dithiocarbamate', 'ammonium', and 'borohydride', demonstrating a stronger connection to materials science concepts. Similarly, 'ethylenediaminetetra-acetic' acid retrieves '-oxycarb' and -'sulfanyl', which accurately reflect its chemical properties. These results suggest that MATTER effectively mitigates tokenization-induced distortions, leading to more precise materials science word representations that enhance performance in downstream NLP tasks such as entity linking, material property prediction, and knowledge graph construction.

G.3 Comparison of Lambda Details

The Macro-F1 scores for ChemDataExtractor and MatDetector were compared across different λ values to evaluate their performance. The specific numerical values are detailed in Table 20 and Table 21, while Figure 6 provides a visual representation for easier interpretation.

Method	Tokenized Output
Origin	poly (dimethylsiloxane) (pdms) waschosen to form the elastomeric circuit board, mechanically protecting and electrically insulating the wires, based on itsdurability, adjustable stiffness, **biocompatibility**, and commer-cial availability as an insulating compound. (Gray et al., 2004)
WordPiece	poly (dimethyl-sil-oxane) (pd-ms) was-cho-sen to form the elast-omeric circuit board , mechanically protecting and electrically ins-ulating the wires , based on its-du-rab-ility , adjustable stiffness , **bioc-omp-ati-bility** , and comme-r-ci-al availability as an ins-ulating compound.
SAGE	poly (dimethyl-siloxane) (pd-ms) was-cho-sen to form the elastomer-ic circuit board , mechanically protecting and electrically insulating the wires , based on its-du-rab-ility , adjustable stiffness , **biocompatibility** , and comme-r-ci-al availability as an insulating compound.
PickyBPE	p-oly (d-imethyl-sil-xane) (p-d-ms) was-ch-osen to form the elast-omeric circuit board , mechanically protecting and electrically insulating the wires , based on its-d-urability , adjustable stiffness , **biocompatibility** , and comm-er-c-ial availability as an insulating compound.
MATTER (ours)	poly (dimethyl-siloxane) (pdms) was-cho-sen to form the elastomeric circuit board , mechanically protecting and electrically insulating the wires , based on its-du-rab-ility , adjustable stiffness , **biocompatibility** , and comme-r-ci-al availability as an insulating compound.
Origin	t he waste was solubilized using ethylenediaminetetraacetic acid, and its constituents were determined employing x-ray diffraction and inductively coupled **plasma-atomic** emission **spectrometry**, identifying **barium** sulfate (baso4) as the predominant component at a weight percentage of 67.13%. (Zhang et al., 2025)
WordPiece	the waste was solubil-ized using ethylenedi-amine-tetr-aa-ce-tic acid , and its constituents wer determined employing x - ray diffraction and inductively coupled **plasma - atomic** emission **spectrometry** , identifying **bari-um** sulfate (bas-o-4) as the predominant component at a weight percentage of 67 . 13 %.
SAGE	the waste was solub-ilized using ethylene-diam-inet-etra-ace-tic acid , and its constituents were determined employing x - ray diffraction and inductively coupled **plasma - atomic** emission **spectrometry** , identifying **barium** sulfate (bas-o-4) as the predominant component at a weight percent-ag-e of 67 . 13 %.
PickyBPE	The waste was solub-ilized using ethyl-ened-i-amin-et-etra-acetic acid , and its constituents were determined employing x - ray diffraction and inductively coupled **plasma - atomic** emission **spectrometry** , identifying **barium** sulfate (b-as-o-4) as the predominant component at a weight percentage of 67.13 %.
MATTER (ours)	the waste was solub-ilized using ethylenediaminetetra-acetic acid , and its constituents were determined employing x - ray diffraction and inductively coupled **plasma - atomic** emission **spectrometry** , identifying **barium** sulfate (baso4) as the predominant component at a weight percentage of 67 . 13 %.

Origin	it should be described that the **ers** signals derived from **perovskite** layers cannot be observed because of their **pauli paramagnetism** nature, resulting in low **ers** intensity, and because of the heavy-atom effects of **pb** or **sn**, leading to short spin-lattice relaxation time and broad **ers** linewidths. (Chen et al., 2025)
WordPiece	it should be described that the **esr** signals derived from **perovsk-ite** layers cannot be observed because of their **paul-i param-agne-tism** nature , resulting in low **ers** intensity , and because of the heavy - atom effects of **p-b** or **sn** , leading to short spin - lattice relaxation time and broad **ers** line-width-s.
SAGE	it should be described that the **e-sr** signals derived from **per-o-v-skite** layers cannot be ob-served because of their **paul-i param-agnetism** nature , resulting in low **e-sr** intens-ity , and because of the heavy - atom effects of **p-b** or **sn** , leading to short spin - lattice relaxation time and broad **e-sr** linewidth-s.
PickyBPE	it should be described that the **es-r** signals derived from **perovskite** layers cannot be observed because of their **pa-uli param-agnetism** nature , resulting in low **es-r** intensity , and because of the heavy - atom effects of **p-b** or **sn** , leading to short spin - lattice relaxation time and broad **es-r** linewidths.
MATTER (ours)	it should be described that the **esr** signals derived from **perovskite** layers cannot be observed because of their **pauli paramagnetism** nature , resulting in low **esr** intensity , and because of the heavy - atom effects of **pb** or **sn** , leading to short spin - lattice relaxation time and broad **esr** linewidths.
Origin	in this study, the **nanoparticles** of the zinc and cobalt **imidazolate** framework (**znco-zif**) were synthesized and directly incorporated into **polylactide** (**pla**) to prepare **pla / znco-zif fibrous** membranes through **electrospinning** methodology. (Deng et al., 2024)
WordPiece	in this study , the **nanoparticles** of the zinc and cobalt **im-ida-zol-ate** framework (**zn-co** - **zi-f**) were synthesized and directly incorporated into **poly-lact-ide** (**pl-a**) to prepare **pla / zn-co - zi-f fibrous** membranes through **electros-pin-ning** methodology.
SAGE	in this study , the **nanoparticles** of the zinc and cobalt **imid-azol-ate** framework (**z-nc-o** - **z-if**) were synthesized and directly incorporated into **polyl-act-ide** (**pl-a**) to prepare **pl-a / z-nc-o - z-if fibrous** membranes through **electrospinning** methodology.
PickyBPE	in this study, the **nanoparticles** of the zinc and cobalt **imid-az-olate** framework (**z-n-co** - **z-if**) were synthesized and directly incorporated into **pol-yl-actide** (**pl-a**) to prepare **pl-a / z-n-co - z-if fibrous** membranes through **electrospinning** methodology.
MATTER (ours)	in this study , the **nanoparticles** of the zinc and cobalt **imidazol-ate** framework (**zn-co** - **zif**) were synthesized and directly incorporated into **polylactide** (**pla**) to prepare **pla / zn-co - zif fibrous** membranes through **electrospinning** methodology.

Table 18: **Boldface** and **pink** concepts are important material concepts extracted using MatDetector. **Boldface** concepts are correctly tokenized in both the baseline and our method, indicating no issues. In contrast, **pink** concepts are highly important but are often split into unrelated subwords or overtokenized in conventional tokenization. However, as shown in this table, our method, MATTER, effectively prevents the overtokenization of important material concepts, preserving their semantic integrity.

Tokenization Method	Concept	Word Embedding 5-NN	Sim.	Formula	Word Embedding 5-NN	Sim.	Abbr	Word Embedding 5-NN	Sim.
WordPiece	germanium (german-ium)	agilent	90.6	PbI2 (pib-2)	nowak	81.8	LFP (lf-p)	inlet	95.5
		fri	85.9		10c	81.8		chattopadhyay	93.7
		stephan	85.3		-agnetically	81.0		1263.0	92.7
		valley	85.2		colouring	79.6		foreland	92.7
		galvanic	86.1		**quasicrystal**	79.6		-rink	92.6
SAGE	germanium (german-ium)	lot	83.0	PbI2 (p-bi-2)	-gen	43.9	LFP (lf-p)	occupation	95.8
		segregation	82.8		pounds	43.6		multiphonon	95.2
		100	82.9		-pt	43.5		-l	95.2
		segregation	82.9		-uck	43.2		-circ	95.1
		agi	82.0		-8	43.2		multiphonon	95.1
PickyBPE	germanium (g-erman-ium)	lot	83.0	PbI2 (p-bi-2)	gaussian	63.1	LFP (l-f-p)	her,	75.8
		segregation	82.8		p	62.8		consideration	75.0
		-inov	81.3		total	62.3		-ermany	75.0
		-w,	81.1		-'s	62.0		-sd102	75.0
		compatibility	81.0		-ories	61.0		{[}40{]}	75.0
MATTER (ours)	germanium (germanium)	**dithiocarbamate**	81.5	PbI2 (pbi2)	**pb5**	89.9	LFP (lfp)	**zrf7**	90.9
		monium	81.4		**pbf2**	89.2		**acyclohex**	90.8
		-orib	81.3		**-anesulfonic**	88.9		**dodecane**	90.0
		borohydride	81.2		**-ob2o3**	88.7		**-acyclohex**	90.0
		-stannyl	81.2		**-dithiocarbamate**	88.5		**-azobenzene**	90.0
Tokenization Method	**Concept**	**Word Embedding 5-NN**	**Sim.**	**Formula**	**Word Embedding 5-NN**	**Sim.**	**Abbr**	**Word Embedding 5-NN**	**Sim.**
WordPiece	ethylenediaminetetra-acetic acid (ethylenedi-amine-tetr-aa-ce-tic acid)	-oreg	92.4	BaSo4 (bas-o-4)	**bas**	91.3	PDMS (pd-ms)	-ms	87.6
		sulphates	92.3		**-o4**	87.3		pd	82.9
		consistence	92.2		**adopts**	87.0		drilled	75.2
		crop	92.1		somehow	85.3		gilbert	75.1
		-ulos	92.1		**reflect**	85.1		connect	74.3
SAGE	ethylenediaminetetra-acetic acid (ethylenedi-amine-tetr-aa-ce-tic acid)	**-ilent**	94.8	BaSo4 (bas-o-4)	bas	85.2	PDMS (pd-ms)	p	93.1
		athermal	94.7		**nanobelts**	75.3		others	91.5
		-stoichi	94.6		interv	75.2		heas	91.4
		thermogravimetric	94.6		).(	75.1		ellips	91.3
		-true	94.5		-rino	74.8		-rac	91.2
PickyBPE	ethyl-ened-i-amin-et-etra-acetic acid	contrast,	89.6	BaSo4 (b-as-o-4)	sliding	71.6	PDMS (p-d-ms)	**ppe**	76.5
		represents	89.2		charged	69.4		);	69.9
		sophistic	89.1		-adi	67.7		**prem**	69.5
		zn(II)	89.1		2013.0	60.2		inductively	66.6
		distribution	89.0		**mocvd**	59.8		bat	66.6
MATTER (ours)	ethylenediaminetetra-acetic acid	**ethylenediaminetetra**	93.7	BaSo4 (baso40	**bas**	91.4	PDMS (pdms)	**perfluoroalkyl**	85.7
		-acetic	93.6		**-o4**	87.5		**trimethoxysilyl**	85.6
		-oxycarb	91.8		**bast**	84.4		**-yloxy**	85.5
		-sulfanyl	91.7		**cyclohexyl**	82.0		**-obenzoic**	85.4
		agre	91.6		**-cyclopentadienyl**	81.9		**borohyd**	85.4

Table 19: Comparison of subword embedding averaging results across different tokenization methods, including WordPiece, SAGE, PickyBPE, and our proposed method, MATTER. The table presents the five nearest neighbor words based on subword embedding averages for each method, illustrating how different tokenization strategies impact semantic similarity in word embeddings. The similarity scores (Sim.) indicate the relevance of the nearest neighbors to the target material concept. **Boldface** highlights words that are directly related to materials.

MatDetector (ours)		MatSci-NLP							
		NER	RC	EAE	PC	SAR	SC	SF	Overall
w/o material knowledge	Micro-F1	76.6	80.9	48.5	73.1	81.9	90.0	57.4	72.6
	Macro-F1	56.1	58.5	29.4	58.9	74.6	60.3	32.6	52.9
0.1	Micro-F1	76.4	78.6	47.4	74.1	79.5	91.0	61.2	72.6
	Macro-F1	54.3	54.4	32.6	69.5	62.7	61.1	37.0	53.1
0.2	Micro-F1	78.3	78.7	49.7	74.5	76.2	**91.2**	60.4	72.7
	Macro-F1	58.5	53.1	30.2	68.8	63.3	58.0	36.7	52.6
0.3	Micro-F1	78.5	80.2	51.2	76.6	73.9	91.5	**62.7**	73.5
	Macro-F1	56.2	55.7	35.7	69.2	58.9	62.5	**41.3**	54.2
0.4	Micro-F1	75.4	80.7	52.9	73.3	77.2	90.6	59.1	72.7
	Macro-F1	54.5	56.5	33.4	68.4	59.9	63.5	35.0	53.0
0.5	Micro-F1	78.7	82.4	53.0	74.2	76.2	90.8	60.2	73.6
	Macro-F1	55.2	58.4	32.6	66.5	63.9	61.5	33.5	53.1
0.6	Micro-F1	77.8	83.6	49.6	**77.3**	78.3	**91.2**	61.2	74.1
	Macro-F1	57.2	**60.7**	30.7	68.9	63.1	59.3	38.0	54.0
0.7	Micro-F1	78.3	81.5	48.3	74.5	82.4	90.9	59.8	73.7
	Macro-F1	56.7	58.0	34.9	69.7	60.0	59.6	37.4	53.8
0.8	Micro-F1	76.9	80.4	**54.0**	75.2	80.4	91.1	61.5	74.2
	Macro-F1	54.9	55.7	37.0	**71.4**	74.7	59.5	37.8	55.9
0.9	Micro-F1	79.0	81.3	53.1	74.2	78.6	90.2	60.8	73.9
	Macro-F1	58.8	59.0	**37.3**	69.5	64.1	**66.4**	37.7	56.1
1.0	Micro-F1	**80.0**	**83.8**	53.1	73.7	**85.5**	**91.2**	61.9	**75.6**
	Macro-F1	**59.3**	59.1	36.9	67.6	**79.3**	64.9	38.0	**57.9**
2.0	Micro-F1	77.3	79.2	52.1	75.1	75.7	91.1	61.1	73.1
	Macro-F1	55.7	55.9	36.6	66.6	62.6	60.1	38.0	53.7
3.0	Micro-F1	76.1	79.2	50.5	71.6	77.9	90.2	61.5	72.4
	Macro-F1	54.2	57.6	34.2	65.9	63.8	59.7	38.6	53.4

Table 20: Specific numerical results of MatDetector's Macro-F1 and Micro-F1 scores across different λ values.

ChemDataExtractor (Swain and Cole, 2016)		MatSci-NLP							
		NER	RC	EAE	PC	SAR	SC	SF	Overall
w/o material knowledge	Micro-F1	76.6	80.9	48.5	73.1	81.9	90.0	57.4	72.6
	Macro-F1	56.1	58.5	29.4	58.9	**74.6**	60.3	32.6	52.9
0.1	Micro-F1	75.5	81.0	52.7	72.8	77.9	90.4	55.8	72.3
	Macro-F1	52.4	61.1	34.4	63.0	66.6	62.2	31.9	53.1
0.2	Micro-F1	76.4	83.2	52.1	70.5	76.7	91.0	58.6	72.6
	Macro-F1	56.3	61.0	32.8	64.8	63.8	60.3	34.2	53.3
0.3	Micro-F1	75.4	82.3	52.3	73.1	76.4	90.4	56.8	72.4
	Macro-F1	53.2	61.0	29.8	65.2	64.4	58.9	33.0	52.2
0.4	Micro-F1	73.4	84.0	**55.1**	71.9	79.5	90.6	57.1	73.1
	Macro-F1	51.4	58.4	**37.9**	**68.8**	66.6	60.0	30.5	53.4
0.5	Micro-F1	77.0	82.3	53.8	72.2	79.8	91.0	57.7	73.4
	Macro-F1	**56.4**	61.3	35.8	67.7	62.0	61.8	30.8	53.7
0.6	Micro-F1	76.5	84.1	54.1	67.3	78.2	**91.2**	57.8	72.7
	Macro-F1	55.1	61.8	36.6	59.6	66.2	61.9	33.6	53.5
0.7	Micro-F1	75.4	82.4	52.5	71.8	**82.3**	90.2	58.0	73.2
	Macro-F1	54.5	60.7	33.3	65.8	68.1	63.2	**37.9**	54.8
0.8	Micro-F1	76.0	84.0	53.4	71.1	78.1	91.1	58.1	73.1
	Macro-F1	55.5	62.8	34.7	64.7	65.8	**65.0**	35.5	54.9
0.9	Micro-F1	75.6	81.9	52.8	73.3	82.0	91.1	58.4	73.6
	Macro-F1	54.1	59.0	37.5	67.5	65.5	58.0	35.7	53.9
1.0	Micro-F1	**77.1**	81.5	53.1	**73.6**	80.6	**91.2**	**58.8**	**73.7**
	Macro-F1	**56.4**	58.9	35.0	67.6	68.0	64.8	35.6	**55.2**
2.0	Micro-F1	77.0	**84.7**	52.3	69.4	80.7	91.1	57.3	73.2
	Macro-F1	55.3	**64.2**	34.1	64.3	68.2	60.3	35.4	54.5
3.0	Micro-F1	76.1	83.2	52.0	67.7	75.6	90.2	58.7	71.9
	Macro-F1	55.5	60.6	34.0	65.8	67.8	60.6	34.5	54.1

Table 21: Specific numerical results of ChemDataExtractor's Macro-F1 and Micro-F1 scores across different λ values.