

CHRONOS: Chronology-Grounded Continued Pre-training for Language Model Adaptation to Materials Science

Sangyun Kim^{1*}, Junho Kim^{1*}, SangKeun Lee^{1,2}

¹Department of Artificial Intelligence, Korea University, Seoul, Republic of Korea

²Department of Computer Science and Engineering, Korea University, Seoul, Republic of Korea
{silky, monocrat, yalphy}@korea.ac.kr

Abstract

Continued pre-training on domain-specific corpora has become the prevailing approach for adapting pre-trained language models (PLMs) to materials science. However, existing strategies typically treat scientific literature as independent samples, thereby overlooking the cumulative nature of scientific knowledge where novel discoveries systematically build upon prior research. To address this limitation, we propose **Chronology**-grounded continued pre-training for materials science (CHRONOS¹), which explicitly considers the chronological knowledge in terms of data construction and training schemes. Specifically, our framework traces the relevant prior articles by using citation-based information and constructs a chronological curriculum to progressively learn the materials knowledge. The comprehensive experiments demonstrate that CHRONOS outperforms state-of-the-art (SOTA) baselines on various materials science benchmarks and backbone models, achieving improvements of 4.7% on the Macro-F1 score. In-depth analysis shows that considering the cumulative nature of knowledge enhances both the quality of the corpus and the effectiveness of training schemes.

1 Introduction

Materials science comprises an interdisciplinary scientific field that requires research on the behaviors, properties, and applications of diverse materials. Regarding the vast search space in materials science, natural language processing (NLP) holds significant promise in advancing the discovery of novel materials by providing informative knowledge extraction tools, such as entity recognition (Weston et al., 2019; Friedrich et al., 2020) or synthesis action retrieval (Wang et al., 2022). However, the scarce and limited datasets in materials sciences pose substantial challenges for developing models

that generalize well across a broad range of materials sciences (Song et al., 2023; Kim et al., 2024).

Continued pre-training has become the prevailing approach to address such challenges by adapting general-domain PLMs on materials science corpora. These methods usually focus on building domain-specific corpora by collecting materials science-related articles (Gupta et al., 2022; Huang and Cole, 2022; Oh et al., 2025) or material-specific training strategies, such as masked language modeling based on materials knowledge graph (Kim et al., 2024). These domain-specific continued pre-training methods can significantly improve the performance of PLMs across a wide range of low-resource materials science downstream tasks (Song et al., 2023; Zaki et al., 2024).

However, previous continued pre-training approaches neglect the cumulative nature of knowledge in materials science. As shown in Figure 1, scientific research is usually built upon prior discoveries (Kuhn, 1962; della Briotta Parolo et al., 2020). For example, without understanding the basic concepts of *silicon solar cells* and *perovskite structures*, it is difficult to understand the advanced techniques in *monolithic tandem solar cells* that combine both materials to surpass single-junction efficiency limits. By tracing these semantic relationships backward from current materials research papers, PLMs can learn informative foundational prior works that underpin current research during pre-training. Hence, considering the nature of cumulative knowledge in pre-training can lead to better materials science understanding and accordingly benefit diverse materials-driven applications.

For incorporating the cumulative knowledge into the pre-training process, there are two main challenges: (1) **Chronological corpus construction**: since most previous works rely on randomly collected domain corpus for pre-training, it is important to consider how to construct corpus that entails both temporal and semantic relations among vari-

* These authors contributed equally to this work.

¹Our code is available at [Official CHRONOS GitHub](#)

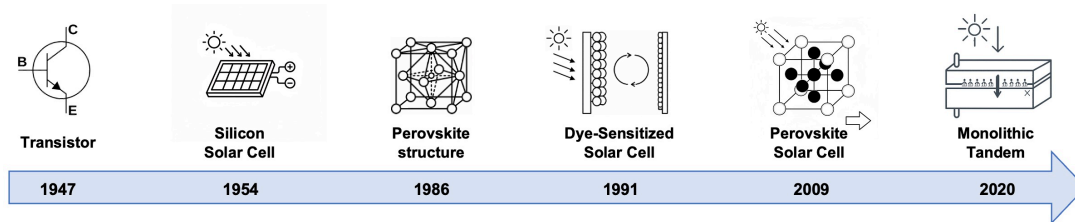


Figure 1: Illustration of the cumulative nature of scientific knowledge in materials science. Recent advances such as monolithic perovskite/silicon tandem solar cells (Al-Ashouri et al., 2020) build upon foundational semiconductor and photovoltaic discoveries (Bardeen and Brattain, 1948; Chapin et al., 1954; Bednorz and Müller, 1986) and subsequent solar cell platform innovations (O’Regan and Grätzel, 1991; Kojima et al., 2009).

ous articles; (2) **Chronological training scheme:** current pre-training processes for materials science usually treat each article as independent samples, neglecting the directional prerequisite relations where foundational knowledge underpin contemporary, high-level research. Thus, these challenges pose a necessity to design a specific pre-training framework that reflects the temporal relevance and importance among articles.

In this paper, we introduce **CHRONOS** (**Chronology-grounded continued pre-training for materials science**), which incorporates coherent cumulative materials knowledge in respects of both corpus construction and training schemes: (1) Chronological data construction that traces the relevant prior articles from contemporary research backward by leveraging the diverse citation-based information, including references and citation numbers; (2) Chronological curriculum learning that organizes corpora in temporal order to progressively train the cumulated knowledge from the prior foundation works. By utilizing citation links, CHRONOS entails not merely temporal signals, but also encodes more semantically related lineage between prior discoveries and subsequent advances in continued pre-training.

To validate the efficacy of our framework, we evaluate CHRONOS on a broad range of models and materials science benchmarks, including generation and classification tasks. Our comprehensive results clearly demonstrate that CHRONOS outperforms other strong continued pre-training baselines and achieves the SOTA performance by a large margin of up to about 2.9% and 4.7% in Micro-F1 and Macro-F1 scores, respectively. Moreover, we verify that our CHRONOS not only provides diverse and informative corpus but also represents generalizability of curriculum learning designs compared to previous approaches.

To summarize the contributions of our paper:

- We investigate the effectiveness of explicitly incorporating the cumulative knowledge during continued pre-training for language model adaptation in materials science.
- We propose CHRONOS, a novel framework that incorporates the cumulative knowledge for pre-training through constructing a chronological corpus and curriculum learning.
- We demonstrate CHRONOS significantly improves the performance compared to previous SOTA baselines in a broad range of materials science downstream tasks and models.

2 Related Work

2.1 Materials Science Continued Pre-training

Building domain-specific PLMs from scratch requires substantial computational costs. Thus, continued pre-training has been standard practice to adapt general PLMs by performing language modeling on the target domain corpus (Gururangan et al., 2020). Early efforts like SciBERT (Beltagy et al., 2019) demonstrated the value of continued pre-training on a broad range of scientific corpora. Subsequently, materials-specific approaches have proposed further specialized to materials science domain by training on randomly collected corpus by using materials-specific terms (Gupta et al., 2022; Huang and Cole, 2022; Mishra et al., 2024). Building on this foundation, recent research has advanced the adaptation process by investigating domain-specific data construction or training schemes. ETS-DACP (Xie et al., 2024) highlighted the efficiency of targeted domain data selection based on information density, while MELT (Kim et al., 2024) incorporated materials science knowledge through semantic graph-based curriculum

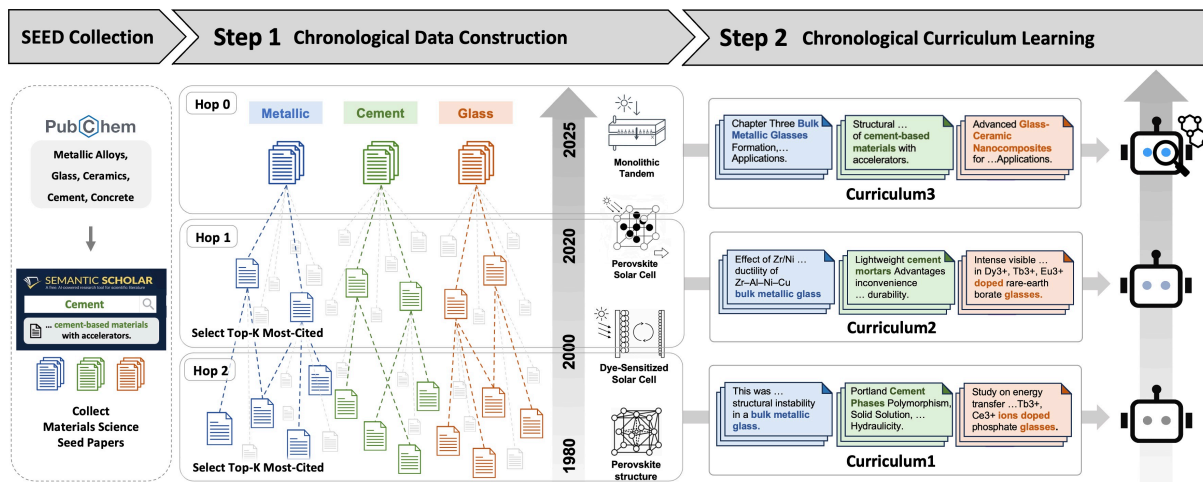


Figure 2: Overview of the CHRONOS framework. The framework consists of chronology-grounded data construction that traces semantic relevance backward through citations from articles, followed by chronological curriculum learning that reorganizes selected documents in forward temporal order.

learning. While recent works (Mishra et al., 2024; Xie et al., 2025; Park et al., 2025) emphasized the importance of instruction tuning datasets for materials science, the impact of data quality and diversity during pre-training has not been well explored.

Previous studies have demonstrated promising results in adapting PLMs to materials science. However, previous approaches typically treat all scientific literature as a static collection in pre-training, overlooking the cumulative nature of scientific knowledge. In contrast, CHRONOS explicitly incorporates the cumulative nature of scientific knowledge in the pre-training process with respect of both data construction and training schemes.

2.2 Curriculum Learning for PLMs

Curriculum learning is a training strategy that organizes training into a progression from easier to harder examples, rather than learning all data simultaneously (Bengio et al., 2009). The effectiveness of curriculum learning critically depends on how difficulty is measured (Kumar et al., 2010). In the NLP area, previous studies measure difficulty commonly estimated from model-based signals in training dynamics (Graves et al., 2017; Xu et al., 2020) or heuristic linguistic measures such as sentence length or word rarity (Platanios et al., 2019; Zhang et al., 2018). However, previous works have usually focused on PLM fine-tuning for improving optimization speed or generalizability in downstream tasks (Chen et al., 2023; Pouransari et al., 2024) due to the challenges of designing an effective difficulty measure for large-scale training (Soviany

et al., 2022). Several studies have introduced curriculum designs for language model pre-training by utilizing the relations in knowledge graphs as difficulty measures. Lee et al. (2022) first proposed a knowledge graph-based masking strategy by utilizing the depth of hops in ConceptNet (Speer et al., 2017), which reflects the connectivity of language concepts. In the materials science domain, Kim et al. (2024) constructs materials-aware semantic graphs and employs the node degree of the entities to build the multi-stage curriculum learning.

However, previous approaches primarily exploit intra-document relations among entities and do not explicitly capture inter-document knowledge relations. Different from the previous works, we focus on exploring the effectiveness of the chronological relations among inter-documents as a difficulty measure in continued pre-training for materials science domain adaptation.

3 CHRONOS

We elaborate our CHRONOS, which incorporates the nature of scientific knowledge to pre-training process. Our key idea is to leverage the citation-based information for both chronological datasets and a curriculum. To this end, we first collect a chronologically structured corpus by constructing reference-based graphs from expanding the recent publications. We then organize the corpus in chronological order, building a curriculum where models first learn foundational materials knowledge, and progressively advance to recent research. The overall framework is illustrated in Figure 2.

3.1 Chronological Data Construction

We begin by constructing a corpus that captures the chronological evolution of scientific knowledge in materials science. Since the citation information entails both relevance and significance of each article, we leverage the reference relations and the number of citations to filter out important articles. Specifically, we suggest a multi-hop expansion strategy, where each hop represents progressively earlier papers in the reference relations, naturally yielding a knowledge-cumulative dataset. Moreover, we filter out the articles by utilizing citation counts and temporal constraints, ensuring the dataset is composed of core foundational articles while maintaining alignment with the chronological progression.

Contemporary seed collection. We first identify recent research papers that represent the latest research topics in materials science. To ensure domain specificity, we extract materials science keywords from the PubChem database (Kim et al., 2023; Oh et al., 2025), a comprehensive resource for chemical compounds and materials. These keywords capture a broad range of core concepts, including specific compounds (e.g., lithium-ion, graphene), material classes (e.g., perovskites, semiconductors), and domain-specific terminology².

Using these materials-specific keywords, we query Semantic Scholar to retrieve materials science candidate papers based on their titles and abstracts. We then query the Scopus API (National Center for Science and Engineering Statistics and Science-Metrix and Elsevier, 2025) by DOI to obtain publication-year and citation-count metadata and retain publications from 2020 onward as contemporary seed papers. This temporal cutoff ensures that the seed set reflects recent research directions and advances in materials science.

Citation-based corpus expansion. From the seed set (i.e., $\mathcal{H}_0$), we iteratively expand the corpus through citation networks following the reference relations in the papers. For each parent paper $p \in \mathcal{H}_i$ in hop i , we collect its reference list and use it as the candidate pool for the next-hop expansion:

$$\mathcal{C}_p = \mathcal{R}(p), \quad (1)$$

where $\mathcal{R}(p)$ denotes the reference list of p . To ensure data quality and coverage, we employ a multi-source API strategy. In this pipeline, we

use Semantic Scholar for keyword-based discovery, query Scopus by DOI to retrieve and validate metadata for recency and citation-based filtering, and retrieve and validate reference relations through OpenCitations (Peroni and Shotton, 2020). We collect abstracts through a sequential source pipeline (e.g., OpenAlex and Europe PMC), with Semantic Scholar used as a fallback source³. We report the one-time corpus-construction cost in Appendix A.

We then apply two-stage filtering strategies to ensure temporal coherence and quality. For temporal coherence, we enforce a relative temporal constraint on citation gaps. Since research papers often cite very old foundational studies that are not closely relevant or have multi-hop relations (e.g., as background literature), we constrain each candidate to be published no later than its parent, within a Δt -year window, and no earlier than $y_{\min}$:

$$\mathcal{C}'_p = \{c \in \mathcal{C}_p \mid \tau_p \leq y_c \leq y_p\}, \quad (2)$$

where $\tau_p = \max(y_{\min}, y_p - \Delta t)$ is the per-parent temporal lower bound, and y_p, y_c are publication years of the parent p and a candidate c , respectively⁴. This process preserves temporal proximity to each parent paper while progressively expanding toward earlier foundational studies. We compare the hop-level publication-year distributions with and without this constraint in Appendix A.5.

After temporal filtering, we apply citation-based quality filtering to prioritize influential works. Since citation count reflects scholarly influence, we rank the valid candidates for each parent paper by citation count and select the top- K papers:

$$\mathcal{S}_p = \text{Top-}K(\mathcal{C}'_p, c \mapsto \text{cites}(c)), \quad (3)$$

where $\text{Top-}K(A, s)$ returns the K elements of A with the largest values of the score function s , and $\text{cites}(c)$ denotes the citation count of candidate c .

The constructed hop $i + 1$ aggregates the per-parent selections with deduplication against earlier hops:

$$\mathcal{H}_{i+1} = \left(\bigcup_{p \in \mathcal{H}_i} \mathcal{S}_p \right) \setminus \bigcup_{j=0}^i \mathcal{H}_j. \quad (4)$$

This guarantees each paper appears only in its first discovered hop, preserving a clean chronological structure for the curriculum. Finally, we

²We collect 2.6k keywords for queries

³Detailed settings are provided in Appendix A

⁴We set $y_{\min} = 1980$ and $\Delta t = 10$ years.

Algorithm 1 CHRONOS Framework

```
1: Input: Seed papers  $\mathcal{H}_0$ , max hops  $N$ , temporal window  $\Delta t$ , top- $K$ , min year  $y_{\min}$ , PLM parameters  $\theta$ , epoch number  $E$ 
2: Output: PLM parameters  $\theta$ 
3:
4: // Chronological Data Construction
5: for  $i = 0$  to  $N - 1$  do
6:   for each parent  $p \in \mathcal{H}_i$  do
7:     // Candidate pool from references
8:      $\mathcal{C}_p \leftarrow \mathcal{R}(p)$ 
9:     // Per-parent temporal filtering
10:     $\mathcal{C}'_p \leftarrow \{c \in \mathcal{C}_p \mid \max(y_{\min}, y_p - \Delta t) \leq y_c \leq y_p\}$ 
11:    // Impact-based top- $K$  selection
12:     $\mathcal{S}_p \leftarrow \text{Top-}K(\mathcal{C}'_p, \text{cites})$ 
13:  end for
14:   $\mathcal{H}_{i+1} \leftarrow \left( \bigcup_{p \in \mathcal{H}_i} \mathcal{S}_p \right) \setminus \bigcup_{j=0}^i \mathcal{H}_j$ 
15: end for
16:
17: // Chronological Curriculum Learning
18: for  $\text{epoch} = 1$  to  $E$  do
19:   for stage  $i = N$  down to  $0$  do
20:     Train PLM  $\theta$  on  $\sigma(\mathcal{H}_i)$ 
21:   end for
22: end for
23: return PLM  $\theta = 0$ 
```

preserve citation relations, which enable the analysis of knowledge flow patterns and the validation of semantic coherence across temporal boundaries.

3.2 Chronological Curriculum Learning

The corpus construction naturally yields a hop-based structure where each hop represents progressively earlier papers. Since this structure reflects the knowledge accumulation, we directly leverage it to construct the multi-stage curriculum.

Formally, let $\mathcal{H}_i$ denote papers in hop i , where hop-0 contains recent papers and higher hop numbers indicate earlier foundational works. We denote the random shuffle operation in each hop as $\sigma(\cdot)$ and define the stage-wise curriculum schedule as:

$$(\sigma(\mathcal{H}_N), \sigma(\mathcal{H}_{N-1}), \dots, \sigma(\mathcal{H}_0)) \quad (5)$$

where each stage is trained sequentially from foundational works ($\mathcal{H}_N$) to contemporary research ($\mathcal{H}_0$), iterating from $i = N$ down to 0. Within each stage, papers are randomly shuffled before mini-batch construction.

This stage-wise curriculum organization enables our method to utilize all datasets in the corpora while maintaining clear chronological progression. Our curriculum design aligns with how scientific understanding naturally accumulates (Kuhn, 1962; della Briotta Parolo et al., 2020). For instance, fundamental studies on *solid oxide fuel cells* from the

1990s establish core principles that inform recent work on advanced *electrode materials*. By following such cumulative progression, models can learn contemporary advanced material concepts with informative prior knowledge. The overall process is shown in Algorithm 1.

4 Experiments

We conduct comprehensive experiments to verify the efficacy of CHRONOS. Specifically, we address the following research questions:

Q1. (Generalizability) Does CHRONOS enable better adaptation compared to existing continued pre-training methods across diverse materials science tasks and models? (§4.2, §4.5, §4.6)

Q2. (Explainability) How do the chronological data construction and curriculum learning contribute to the overall performance? (§4.3, §4.4)

Q3. (Efficiency) Does the chronological curriculum learning enable more efficient learning dynamics and better knowledge accumulation? (§4.7)

4.1 Experimental Setups

Baselines. Consistent with the previous works settings (Kim et al., 2024), we compare with strong continued pre-training baselines, including both general domain adaptations: DSP (Gururangan et al., 2020), ETS-DACP (Xie et al., 2024), and materials-specific domain adaptation baselines: EntityBERT (Lin et al., 2021), MELT (Kim et al., 2024). Detailed baseline configurations are provided in Appendix B.

Data construction. For our model, we construct the training corpus through citation network traversal as described in §3.1. Starting from 10,888 post-filtered hop-0 seed papers, the corpus spans five hop levels (hop-0 through hop-4), with a 10-year temporal gap and top-5 citation-based selection, yielding 183,202 unique papers (approximately 183k). For the baselines, we construct materials science corpora of matched scale using the term-based collection strategy adopted in prior work (Gupta et al., 2022; Oh et al., 2025). Detailed corpus statistics are in Appendix A.

Pre-training configuration. To demonstrate the broad applicability of our framework, we adapt the SciBERT (Beltagy et al., 2019) as an encoder-based PLM and Llama-3.2-1B (Grattafiori et al., 2024) as a decoder-based PLM. We perform language modeling with a chronological curriculum

Method	Metric	MatSci-NLP							Overall
		NER	RC	EAE	PC	SAR	SC	SF	
SciBERT (Beltagy et al., 2019)	Micro-F1	0.738	0.818	0.458	0.671	0.701	0.901	0.500	0.693
	Macro-F1	0.517	0.600	0.290	0.568	0.528	0.612	0.258	0.482
DSP (Gururangan et al., 2020)	Micro-F1	0.769	0.852	<u>0.503</u>	0.713	0.827	<u>0.904</u>	0.568	0.734
	Macro-F1	0.556	0.637	0.327	0.576	0.729	0.604	0.308	0.534
EntityBERT (Lin et al., 2021)	Micro-F1	0.789	0.856	0.489	0.744	0.852	0.903	0.584	0.745
	Macro-F1	0.575	<u>0.652</u>	0.318	<u>0.683</u>	<u>0.756</u>	0.643	<u>0.352</u>	0.568
ETS-DACP (Xie et al., 2024)	Micro-F1	0.784	<u>0.866</u>	0.479	0.732	0.816	0.834	0.590	0.729
	Macro-F1	<u>0.577</u>	0.643	0.310	0.686	0.735	0.569	0.327	0.550
MELT (Kim et al., 2024)	Micro-F1	<u>0.797</u>	0.856	0.499	<u>0.787</u>	0.806	0.895	<u>0.593</u>	<u>0.748</u>
	Macro-F1	0.572	0.651	<u>0.339</u>	0.666	0.766	<u>0.616</u>	0.353	<u>0.571</u>
CHRONOS (ours)	Micro-F1	0.811	0.877	0.523	0.815	<u>0.839</u>	0.908	0.621	0.770
	Macro-F1	0.611	0.659	0.341	0.765	0.747	0.569	0.498	0.598

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

Method	MatSci-NLP (Overall)	
	Micro-F1	Macro-F1
CHRONOS (ours)	0.770	0.598
Components in CHRONOS		
CHRONOS w/o CCL	0.745	0.560
CHRONOS w/o CDC, CCL	0.715	0.523
Curriculum Learning in CHRONOS		
CHRONOS w/ CCL (ours)	0.770	0.598
CHRONOS w/ Reverse	0.738	0.542
CHRONOS w/ MELT	0.748	0.568
CHRONOS w/ Masking Ratio	0.741	0.564
CHRONOS w/ Non-curriculum	0.745	0.560

Table 2: Ablation results for each component in CHRONOS on MatSci-NLP. CCL and CDC indicate chronological curriculum learning and chronological data construction, respectively.

where papers from higher hops are presented first, followed by progressively more recent hops, with random shuffling within each hop. More detailed training settings are in Appendix C.

Downstream tasks and evaluation. To demonstrate effectiveness across diverse applications, we evaluate on three different benchmark categories. For comprehensive assessment, we adopt MatSci-NLP (Song et al., 2023), comprising seven tasks including entity recognition, relation classification, and synthesis action retrieval. For classification-specific evaluation, we use three tasks: NER and

slot filling on SOFC (Friedrich et al., 2020), and NER on MatScholar (Weston et al., 2019). For decoder-based models, we conduct fine-tuning experiments on MatSci-NLP and zero-shot evaluation on MaScQA (Zaki et al., 2024), a materials science question-answering benchmark. We report average scores for all metrics across five random seeds. More detailed settings are in Appendix B.

4.2 Main Results

For a fair comparison to the previous works on materials science (Kim et al., 2024; Gupta et al., 2022), we mainly investigate the performance of the encoder-based model. Specifically, we evaluate CHRONOS by comparing its performance on the MatSci-NLP benchmark. The results are shown in Table 1⁵. We observe that our CHRONOS yields the best performance on various materials science tasks compared to the previous baselines. Specifically, CHRONOS enhances the performance of the backbone models by an average of 2.9% and 4.7% in terms of Micro-F1 and Macro-F1 scores compared to the SOTA method MELT, respectively. These results demonstrate the importance of considering the nature of cumulative knowledge when adapting PLMs to the materials science domain.

⁵The results for same corpus are shown in Appendix D

4.3 Ablation Study

We perform ablation studies to assess the contributions of chronological data construction and curriculum learning in CHRONOS⁶.

Data construction and curriculum learning As shown in the upper part of Table 2, we observe that omitting chronological data construction (CDC) and curriculum learning (CCL) adversely affects both Micro-F1 and Macro-F1 scores. For CDC, we randomly collect the papers following the previous work (Beltagy et al., 2019). Specifically, eliminating both curriculum learning and data construction significantly degrades performance by 7.6% and 14% on Micro-F1 and Macro-F1 scores, respectively. These results indicate that both data construction and the training scheme in our framework are essential to adopt PLMs in materials science.

Curriculum design comparisons To further investigate the design choices for curriculum learning, we compare our method with four curriculum learning methods: 1) Reverse, which trains the model with the reverse order of the curriculum in CHRONOS, 2) MELT, which uses the node degree of a materials knowledge graph (Kim et al., 2024), 3) Masking Ratio, which gradually increases the masking ratio in the pre-training corpus, and 4) Non-curriculum, which trains the model with a random order. As shown in the lower part of Table 2, the chronological curriculum in CHRONOS achieves the strongest aggregate performance among the evaluated strategies. The reverse and non-curriculum variants use the same CHRONOS corpus and therefore provide direct evidence for the contribution of training order.

4.4 Effect of Chronological Data Selection

Comparison to other data construction To verify the efficacy of our data construction, we compared to three strong baselines: 1) API-based random selection (Beltagy et al., 2019), 2) materials term-based selection (Gupta et al., 2022; Oh et al., 2025), and 3) ETS-DACP that selects the informative subset for each downstream task from corpora (Xie et al., 2024). For a fair comparison, we utilize the same corpus scale for all data construction methods⁷. The results are shown in Table 3. While the material-related corpus construction baselines achieved better than random selection,

Method	MatSci-NLP (Overall)	
	Micro-F1	Macro-F1
Random	0.715	0.523
Term-based	0.734	0.534
ETS-DACP	0.729	0.550
CHRONOS (ours)	0.745	0.560

Table 3: Comparison of different data construction strategies with CHRONOS. The best results are highlighted in **boldface**.

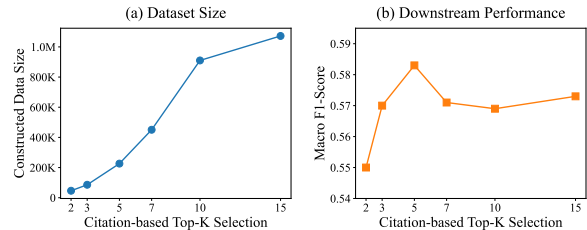


Figure 3: Analysis of citation-based selection. Panel (a) shows the number of papers, and Panel (b) shows the MatSci-NLP performance for different values of top- K .

our proposed CHRONOS achieve the best performance compared to other baselines on both F1 scores. These results highlight that it is crucial to consider chronological relations for building the pre-training corpus in materials science domain⁸.

Effects of citation-based selection To investigate the effectiveness of our citation-based document selection, we evaluate the performance of our methods by varying the number of top- K citations. The results are shown in Figure 3. We observe that performance increases as the number of top- K documents grows, but it saturates at over 5 documents and begins to decline. These results demonstrate that including all documents for pre-training is not always helpful and highlight the necessity of selecting informative document sets. This may be due to the noisy and irrelevant documents, which align with previous data selection methods in pre-training (Xie et al., 2024). Overall, the results clearly demonstrate that our citation-based top- K document selection can not only enhance the quality of the corpus, but also improve the efficiency for collecting corpora in the materials science domain.

⁶Further ablations are provided in Appendix D.

⁷Detailed settings are shown in Appendix A.6

⁸Corpus impact and diversity analysis in Appendix F.

Method	MatSci-NLP (Overall)		MaScQA
	Micro-F1	Macro-F1	
Llama-3.2-1B	0.207	0.135	18.15
DSP	<u>0.229</u>	<u>0.165</u>	<u>20.77</u>
ETS-DACP	0.194	0.153	19.69
CHRONOS (ours)	0.265	0.186	21.69

Table 4: Evaluation results of Llama-3.2-1B backbones on MatSci-NLP and MaScQA benchmarks. The best and second best results are highlighted in **boldface** and underline, respectively.

4.5 Analysis on Decoder-based Models

To further investigate the adaptability of CHRONOS, we apply our framework to the decoder-based PLM Llama-3.2-1B (Grattafiori et al., 2024). Most previous works on materials science adaptation focus on encoder-based models through domain-specific masked language modeling (Lin et al., 2021; Kim et al., 2024). Thus, we compare CHRONOS with two strong general domain adaptation baselines (Gururangan et al., 2020; Xie et al., 2024) on generation and question answering tasks (Song et al., 2023; Zaki et al., 2024). As shown in Table 4, CHRONOS consistently outperforms the baselines on both MatSci-NLP and MaScQA benchmarks. Specifically, our CHRONOS enhance the performance by an average on 15.7% and 12.7% for Micro- and Macro-F1 scores. Moreover, we observe a similar trend in the Llama-3.1-8B MaScQA setting, where CHRONOS achieves 31.38 compared with 30.46 for DSP⁹. These results not only demonstrate the adaptability of our CHRONOS on decoder-based PLMs, but also highlight the scalability on the larger models.

4.6 Evaluation on Classification Tasks

To evaluate the generality of our CHRONOS, we additionally compare to each baseline on three materials science classification tasks, including two different named entity recognition and slot filling tasks (Weston et al., 2019; Friedrich et al., 2020), following the settings of previous works settings (Gupta et al., 2022; Kim et al., 2024)¹⁰. The results are shown in Table 5. We observe that CHRONOS achieved the SOTA performance, similar to the generation task. Specifically, CHRONOS consistently outperforms SOTA baselines by 0.6 aver-

⁹Detailed results are provided in Appendix H

¹⁰Detailed settings are in Appendix B.2

Method	NER _{SOFC}	NER _{MS}	SF	Avg.
SciBERT	80.3	85.3	59.1	74.9
DSP	80.4	84.6	59.8	74.9
EntityBERT	<u>80.8</u>	85.2	60.8	75.6
ETS-DACP	79.7	<u>85.5</u>	58.4	74.5
MELT	80.3	85.4	<u>61.7</u>	<u>75.8</u>
CHRONOS (ours)	81.2	86.0	61.9	76.4

Table 5: Evaluation results on three classification tasks. The best and second best results are highlighted in **boldface** and underline, respectively.

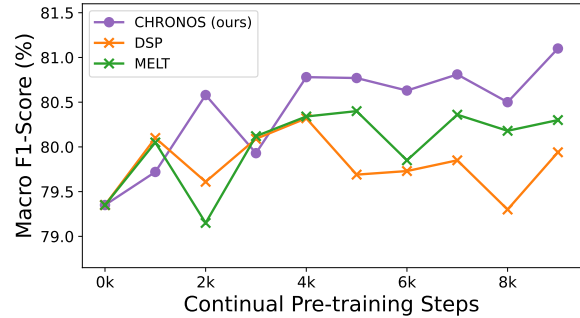


Figure 4: Training dynamics of performance on SF_{SOFC} test set during continued pre-training.

age points on three different information extraction tasks. These results underscore the effectiveness of CHRONOS across diverse types of materials science downstream tasks.

4.7 Analysis of Training Dynamics

To examine how chronological curriculum learning affects training dynamics, we evaluate the performance of CHRONOS according to the steps during pre-training. Figure 4 shows Macro-F1 score on the SF_{SOFC} test set for CHRONOS compared to SOTA baselines. We observe that CHRONOS converges faster and yields more stable performance improvements across the pre-training process. Specifically, CHRONOS achieves higher performance at 2k steps compared to other baseline’s final scores by only utilizing 25% of the training steps. Moreover, while other baselines saturated early in the steps, CHRONOS continues to improve steadily, and reaches its best performance in the later steps. These results demonstrate that considering chronological orders in the pre-training process not only further improves the generalizability but also the training efficiency of continued pre-training on materials science.

5 Conclusion

We have proposed CHRONOS, a novel chronology-grounded continued pre-training framework to effectively adapt PLMs for materials science. To this end, we have introduced a chronological data construction and curriculum learning to reflect the cumulative nature of scientific knowledge in the pre-training process. Specifically, CHRONOS has traced foundational prior works through citation networks and organized them into a chronological curriculum to progressively learn materials science regarding the priority of knowledge. Through extensive experiments, we have demonstrated that CHRONOS has significantly outperformed the previous SOTA baselines across various materials science benchmarks and backbone models. Moreover, we have verified the effectiveness of considering the cumulative knowledge on the pre-training stage for both corpora construction and training schemes.

Limitations

While we have demonstrated the effectiveness of CHRONOS for both data construction and training schemes on materials science, we identify several limitations that present valuable opportunities for future research.

Potential Risk on citation-based selection

While we have clearly demonstrated that citation-based data selection improves data quality in the pre-training corpus construction, relying only on citation-based metrics might cause a potential issue in limiting the diversity of the topics or journals for corpora. To mitigate this, we have introduced materials-term-based seed data collection methods for recent works, which first collect papers covering diverse topics and current issues in materials science (§F). Nevertheless, regarding the metric-agnostic designs of our framework, adopting the advanced citation-based metrics, such as impact factor or field-weighted citation impact, can further improve our framework. We will consider the adaptations for other fine-grained metrics as promising future research directions.

Exploration on Diverse Domain Adaptation

We have mainly focused on the materials science domain, which is characterized by a strong cumulative knowledge structure following previous materials-specific works for PLMs (Gupta et al., 2022; Kim et al., 2024; Oh et al., 2025). However, the applicability of our framework to other

scientific domains, such as biomedical science and computer science, remains underexplored in this work. We leave the exploration of PLM adaptations for diverse academic domains as a future research direction in this field.

Exploration on Larger Models Although we have demonstrated the generalizability of CHRONOS on diverse PLMs, including both encoder-based and decoder-based models, the applicability of our framework on larger PLMs, remains under-explored in this work. Nevertheless, considering the consistent performance improvements of CHRONOS on the 1B and 8B Llama models (Grattafiori et al., 2024), we believe that CHRONOS is expected to work well within other larger PLMs. We leave the exploration of the scalability as a promising future research direction.

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

- Amran Al-Ashouri, Eike Köhnen, Bor Li, Artiom Magomedov, Hannes Hempel, Pietro Caprioglio, José A. Márquez, Anna Belen Morales Vilches, Ernestas Kasparavicius, Joel A. Smith, Ngoc Phung, Dietrich Menzel, Max Grischek, Lukas Kegelmann, Dieter Skroblin, Christian Gollwitzer, Torbjörn Malm, Christopher Rieger, Bernd Stannowski, and 3 others. 2020. [Monolithic perovskite/silicon tandem solar cell with >29% efficiency by enhanced hole extraction](#). *Science*, 370(6522):1300–1309.
- Udo Bach, Donald Lupo, Pascal Comte, Jacques-E. Moser, Frank Weissörtel, Josef Salbeck, Heinz Spreitzer, and Michael Grätzel. 1998. [Solid-state dye-sensitized mesoporous TiO₂ solar cells with high photon-to-electron conversion efficiencies](#). *Nature*, 395(6702):583–585.
- John Bardeen and Walter H. Brattain. 1948. [The transistor, a semi-conductor triode](#). *Physical Review*, 74(2):230–231.

- J. G. Bednorz and K. A. Müller. 1986. [Possible high \$T_c\$ superconductivity in the Ba-La-Cu-O system](#). *Zeitschrift für Physik B*, 64:189–193.
- 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, Hong Kong, China. Association for Computational Linguistics.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. [Curriculum learning](#). In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML 2009)*, pages 41–48, Montreal, Quebec, Canada. ACM.
- Kevin A. Bush, Axel F. Palmstrom, Zheng Li, Mathieu Boccard, Soraya Cheacharoen, Kenneth M. Bush, Michael Hodgson, Carrie M. Beal, T. A. Leijmakers, Samuel Graham, Zachary C. Holman, and Michael D. McGehee. 2017. [23.6%-efficient monolithic perovskite/silicon tandem solar cells with improved stability](#). *Nature Energy*, 2(4):17009.
- Daryl M. Chapin, Calvin S. Fuller, and Gerald L. Pearson. 1954. [A new silicon p-n junction photocell for converting solar radiation into electrical power](#). *Journal of Applied Physics*, 25(5):676–677.
- Mayee F. Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. 2023. [Skill-it! a data-driven skills framework for understanding and training language models](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 36000–36040. Curran Associates, Inc.
- Pietro della Briotta Parolo, Rainer Kujala, Kimmo Kaski, and Mikko Kivelä. 2020. [Tracking the cumulative knowledge spreading in a comprehensive citation network](#). *Physical Review Research*, 2(1):013181.
- Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Maruszczyk, and Lukas Lange. 2020. [The SOFC-exp corpus and neural approaches to information extraction in the materials science domain](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1255–1268, Online. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and Llama Team. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Michael Grätzel. 2001. [Photoelectrochemical cells](#). *Nature*, 414(6861):338–344.
- Alex Graves, Marc G. Bellemare, Jacob Menick, Rémi Munos, and Koray Kavukcuoglu. 2017. [Automated curriculum learning for neural networks](#). In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, pages 1311–1320, Sydney, NSW, Australia. PMLR.
- Tanishq Gupta, Mohd Zaki, N. M. Anoop Krishnan, and Mausam. 2022. [MatSciBERT: A materials domain language model for text mining and information extraction](#). *npj Computational Materials*, 8(1):102.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. [Don’t stop pretraining: Adapt language models to domains and tasks](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Shu Huang and Jacqueline M. Cole. 2022. [Battery-BERT: A pretrained language model for battery database enhancement](#). *Journal of Chemical Information and Modeling*, 62(24):6365–6377.
- 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 2024*, pages 10690–10703, Miami, Florida, USA. 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, Leonid Zaslavsky, Jian Zhang, and Evan E. Bolton. 2023. [PubChem 2023 update](#). *Nucleic Acids Research*, 51(D1):D1373–D1380.
- Akihiro Kojima, Kenjiro Teshima, Yasuo Shirai, and Tsutomu Miyasaka. 2009. [Organometal halide perovskites as visible-light sensitizers for photovoltaic cells](#). *Journal of the American Chemical Society*, 131(17):6050–6051.
- Thomas S. Kuhn. 1962. [The Structure of Scientific Revolutions](#). University of Chicago Press, Chicago, IL.
- M. Pawan Kumar, Benjamin Packer, and Daphne Koller. 2010. [Self-paced learning for latent variable models](#). In *Advances in Neural Information Processing Systems*, volume 23, pages 1189–1197. Curran Associates, Inc.
- Mingyu Lee, Jun-Hyung Park, Junho Kim, Kang-Min Kim, and SangKeun Lee. 2022. [Efficient pre-training of masked language model via concept-based curriculum masking](#). In *Proceedings of the 2022 Conference*

- on Empirical Methods in Natural Language Processing, pages 7417–7427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Chen Lin, Timothy A. Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2021. [EntityBERT: Entity-centric masking strategy for model pretraining for the clinical domain](#). In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 191–201, Online. Association for Computational Linguistics.
- Mingzhen Liu, Michael B. Johnston, and Henry J. Snaith. 2013. [Efficient planar heterojunction perovskite solar cells by vapour deposition](#). *Nature*, 501(7467):395–398.
- Vaibhav Mishra, Somaditya Singh, Mohd Zaki, Hargun Singh Grover, Santiago Miret, Mausam, and N. M. Anoop Krishnan. 2024. [Llamat: Large language models for materials science](#). In *AI for Accelerated Materials Design-Vienna 2024*.
- David B. Mitzi, C. A. Feild, W. T. A. Harrison, and A. M. Guloy. 1994. [Conducting tin halides with a layered organic-based perovskite structure](#). *Nature*, 369(6480):467–469.
- National Center for Science and Engineering Statistics and Science-Metrix and Elsevier. 2025. [Science and engineering indicators](#). Technical report, National Science Foundation.
- Yerim Oh, Jun-Hyung Park, Junho Kim, SungHo Kim, and SangKeun Lee. 2025. [Incorporating domain knowledge into materials tokenization](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9623–9644. Association for Computational Linguistics.
- Brian O’Regan and Michael Grätzel. 1991. [A low-cost, high-efficiency solar cell based on dye-sensitized colloidal TiO₂ films](#). *Nature*, 353(6346):737–740.
- Hyuntae Park, Yeachan Kim, and SangKeun Lee. 2025. [Bridging the gap between molecule and textual descriptions via substructure-aware alignment](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23459–23479. Association for Computational Linguistics.
- Silvio Peroni and David Shotton. 2020. [OpenCitations, an infrastructure organization for open scholarship](#). *Quantitative Science Studies*, 1(1):428–444.
- Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. [Competence-based curriculum learning for neural machine translation](#). 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 1162–1172, Minneapolis, Minnesota. Association for Computational Linguistics.
- Hadi Pouransari, Chun-Liang Li, Jen-Hao Rick Chang, Pavan Kumar Anasosalu Vasu, Cem Koc, Vaishaal Shankar, and Oncel Tuzel. 2024. [Dataset decomposition: Faster llm training with variable sequence length curriculum](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 36121–36147. Curran Associates, Inc.
- Yu Song, Santiago Miret, and Bang Liu. 2023. [MatSci-NLP: Evaluating scientific language models on materials science language tasks using text-to-schema modeling](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3621–3639. Association for Computational Linguistics.
- Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. [Curriculum learning: A survey](#). *Int. J. Comput. Vis.*, 130(6):1526–1565.
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. [Conceptnet 5.5: An open multilingual graph of general knowledge](#). In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4444–4451.
- Samuel D. Stranks, Giles E. Eperon, Giulia Grancini, Christopher Menelaou, Marcelo J. P. Alcocer, Tomas Leijtens, Laura M. Herz, Annamaria Petrozza, and Henry J. Snaith. 2013. [Electron-hole diffusion lengths exceeding 1 micrometer in an organometal trihalide perovskite absorber](#). *Science*, 342(6156):341–344.
- Ching W. Tang. 1986. [Two-layer organic photovoltaic cell](#). *Applied Physics Letters*, 48(2):183–185.
- Ching W. Tang and Steven A. VanSlyke. 1987. [Organic electroluminescent diodes](#). *Applied Physics Letters*, 51(12):913–915.
- 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):100488.
- 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.
- Zheren Wang, Kevin Cruse, Yuxing Fei, Ann Chia, Yan Zeng, Haoyan Huo, Tanjin He, Bowen Deng, Olga Kononova, and Gerbrand Ceder. 2022. [ULSA: Unified language of synthesis actions for the representation of inorganic synthesis protocols](#). *Digital Discovery*, 1(3):313–324.
- Dieter Weber. 1978. [CH₃NH₃PbX₃, ein Pb\(II\)-System mit kubischer Perowskitstruktur](#). *Zeitschrift für Naturforschung B*, 33(12):1443–1445.

- 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.
- Tong Xie, Yuwei Wan, Yixuan Liu, Yuchen Zeng, Shaozhou Wang, Wenjie Zhang, Clara Grazian, Chunyu Kit, Wanli Ouyang, Dongzhan Zhou, and Bram Hoex. 2025. [DARWIN 1.5: Large language models as materials science adapted learners](#). In *AI4X 2025 International Conference*.
- Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. 2024. [Efficient continual pre-training for building domain specific large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10184–10201, Bangkok, Thailand. Association for Computational Linguistics.
- Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. [Curriculum learning for natural language understanding](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6095–6104, Online. Association for Computational Linguistics.
- Yanpeng Ye, Jie Ren, Shaozhou Wang, Yuwei Wan, Imran Razzak, Bram Hoex, Haofeng Wang, Tong Xie, and Wenjie Zhang. 2024. [Construction and application of materials knowledge graph in multidisciplinary materials science via large language model](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 56878–56897. Curran Associates, Inc.
- Mohd Zaki, Jayadeva, Mausam, and N. M. Anoop Krishnan. 2024. [MaScQA: Investigating materials science knowledge of large language models](#). *Digital Discovery*, 3(2):313–327.
- Xuan Zhang, Gaurav Kumar, Huda Khayrallah, Kenton Murray, Jeremy Gwinnup, Marianna J. Martindale, Paul McNamee, Kevin Duh, and Marine Carpuat. 2018. [An empirical exploration of curriculum learning for neural machine translation](#). *arXiv preprint arXiv:1811.00739*.
- Huanping Zhou, Qi Chen, Gang Li, Song Luo, Tze-bing Song, Hsin-Sheng Duan, Ziruo Hong, Jingbi You, Yongsheng Liu, and Yang Yang. 2014. [Interface engineering of highly efficient perovskite solar cells](#). *Science*, 345(6196):542–546.

Appendix

A Details for Data Construction

A.1 Seed Collection

We construct our hop-0 seed set following previous work (Oh et al., 2025). Specifically, we extract 2.6k materials-related concepts from the PubChem database (Kim et al., 2023), including chemical names, IUPAC names, synonyms, and molecular formulas. Using these concepts as keywords, we query Semantic Scholar for materials science candidate papers based on their titles and abstracts. We then query the Scopus API by DOI to obtain publication-year and citation-count metadata and retain publications from 2020 onward up to the time of corpus construction. This process yields 10,888 post-filtered papers, which form the hop-0 seed set. This temporal cutoff ensures that the seed set reflects recent research directions in materials science.

A.2 Citation-based Expansion

We expand the corpus through citation network traversal with two-stage filtering. For each paper in hop i , we retrieve and validate its reference list through the OpenCitations API (Peroni and Shotton, 2020) and apply temporal coherence filtering: candidate papers must satisfy $y_c \geq \max(y_{\min}, y_p - \Delta t)$, where y_p and y_c denote publication years of parent and candidate papers, respectively. We set $y_{\min} = 1980$ and $\Delta t = 10$ years to maintain relevance while ensuring comprehensive historical coverage, aligning with established materials corpora (Tshitoyan et al., 2019; Gupta et al., 2022). For quality filtering, we rank candidates by citation count and select the top- k cited references per parent paper, prioritizing influential foundational works. We set $k = 5$ in our main experiments, with ablation studies presented in §4.4. This process produces five hop levels (i.e., hop-0 through hop-4), with deduplication ensuring each paper appears only in its first discovered hop.

A.3 Dataset Statistics

Table 7 shows the detailed distribution of papers across hops used in our main experiments (i.e., top-5 configuration). The exponential growth pattern reflects the cumulative nature of scientific knowledge, where each paper references multiple earlier works. Considering computational efficiency, we

Hop	Papers	Cumulative
Hop 0 (≥ 2020)	10,888	10,888
Hop 1	21,683	32,571
Hop 2	26,946	59,517
Hop 3	29,516	89,033
Hop 4	31,360	120,393
Total	120,393	-

Table 6: Distribution of papers across hops in the CHRONOS corpus (Top-3 configuration).

Hop	Papers	Cumulative
Hop 0 (≥ 2020)	10,888	10,888
Hop 1	34,034	44,922
Hop 2	47,363	92,285
Hop 3	44,761	137,046
Hop 4	46,156	183,202
Total	183,202	-

Table 7: Distribution of papers across hops in the CHRONOS corpus (Top-5 configuration, main).

Hop	Papers	Cumulative
Hop 0 (≥ 2020)	10,888	10,888
Hop 1	69,731	80,619
Hop 2	172,674	253,293
Hop 3	298,903	552,196
Hop 4	576,124	1,128,320
Total	1,128,320	-

Table 8: Distribution of papers across hops in the CHRONOS corpus (Top-10 configuration).

select the top-5 configuration, containing approximately 183k papers, for our main experiments. For reference, we also provide the statistics of top-3 and top-10 configurations used for analysis in Table 8 and Table 6, respectively.

A.4 API Implementation

We employ a multi-source API strategy to maximize data coverage and quality. The three sources have distinct roles in corpus construction: we use Semantic Scholar for keyword-based candidate discovery, query Scopus (National Center for Science and Engineering Statistics and Science-Metrix and Elsevier, 2025) in batches of 25 DOIs for publication-year and citation-count metadata, and retrieve and validate reference relations through OpenCitations (Peroni and Shotton, 2020). For abstract collection, we implement a sequential pipeline prioritizing OpenAlex, Europe PMC, Elsevier, Semantic Scholar, CORE, and PubMed. The pipeline selects the longest available abstract (pre-

Citation-year gap	Hop	≥ 2020	2010–2019	2000–2009	1990–1999	1980–1989
10-year limit	H0	100.0%	0.0%	0.0%	0.0%	0.0%
	H1	13.8%	86.2%	0.0%	0.0%	0.0%
	H2	1.1%	47.7%	51.2%	0.0%	0.0%
	H3	0.2%	12.3%	66.1%	21.4%	0.0%
	H4	0.1%	2.8%	36.9%	54.0%	6.2%
No gap limit	H0	100.0%	0.0%	0.0%	0.0%	0.0%
	H1	9.6%	49.0%	30.2%	8.2%	3.1%
	H2	0.7%	20.0%	42.4%	25.6%	11.2%
	H3	0.1%	7.6%	30.8%	34.9%	26.7%
	H4	0.0%	2.1%	18.4%	34.1%	45.4%

Table 9: Distribution of publication-year ranges across citation hops with and without the 10-year citation-gap limit. Values denote percentages within each hop and may not sum to 100 due to rounding.

	Term-based	CHRONOS
API calls	346,618	674,113
Collection time	17 h	36 h

Table 10: One-time corpus-construction cost under our collection setup.

ferring >100 characters) to ensure text quality, achieving approximately 98% coverage across the corpus.

Construction overhead To assess the additional cost of citation-graph construction, we compare the number of API calls and collection time with term-based collection. Both methods share keyword-based candidate discovery and same-scale abstract collection, while citation-graph resolution is unique to CHRONOS. As shown in Table 10, CHRONOS requires 674,113 API calls and 36 hours, compared with 346,618 calls and 17 hours for term-based collection. This one-time overhead is confined to corpus construction, and the resulting corpus can be reused across subsequent pre-training runs.

A.5 Effect of the Citation-Year Gap Limit

To assess the effect of limiting the publication-year gap along each citation edge, we compare the publication-year distributions across citation hops with and without the 10-year gap limit. As shown in Table 9, removing the citation-gap limit produces substantial overlap between adjacent hops, whereas applying the 10-year limit yields clearer separation across hop-specific publication periods. This result demonstrates that the citation-gap limit provides clearer chronological separation across

hops.

A.6 Data Construction Baselines

To demonstrate the efficacy of our data construction, we adopt two different data construction baselines, consisting of random (Beltagy et al., 2019) and term-based methods (Gupta et al., 2022; Oh et al., 2025), which are widely used for materials science corpus construction.

Random To collect the broad range of scientific papers, previous works utilize the category restriction for API-based random search, such as the materials science topic in Semantic Scholar (Beltagy et al., 2019). Following the previous work, we leverage our three-tier API implementation for collecting diverse types of materials science papers.

Term-based To collect papers that are more related to the materials science domain, previous works utilize the materials-specific terms, such as glass, cement, or inorganic as a query for searching papers (Gupta et al., 2022; Oh et al., 2025). Consistent with the prior works, we leverage the materials terms that we utilize to construct seed datasets (i.e., 2.6k keywords) and randomly collect the same scale with a temporal restriction.

B Baselines and Evaluation

B.1 Baseline Configurations

We compare CHRONOS with four strong continued pre-training methods for domain adaptation. All baselines use the same backbone models (i.e., SciBERT (Beltagy et al., 2019) and Llama-3.2-1B (Grattafiori et al., 2024)) and are trained on

matched-scale materials science corpora of approximately 183k papers each to ensure a fair comparison.

Baseline Corpus Following the standard domain-adaptive pre-training for materials science (Gupta et al., 2022; Oh et al., 2025), we mainly utilize term-based collected corpora for the baseline dataset.

DSP (Gururangan et al., 2020) performs domain-adaptive pre-training using standard masked language modeling (MLM) with random token masking. Following the previous work, we utilize the whole word masking and original implementation with 15% masking ratio.

EntityBERT (Lin et al., 2021) employs an entity-centric masking strategy where identified domain entities are preferentially masked. To compare more strong baseline, we use the large scale chemical entity datasets from MELT (Kim et al., 2024), consisting of 71k entities. We utilize them for pre-training with 15% masking ratio.

ETS-DACP (Xie et al., 2024) applies continued pre-training with efficient task-aware data selection. Following the original settings, we utilize the cosine similarity between the embeddings of the target task data (i.e., MatSci-NLP) and the domain corpus to select relevant subsets.

MELT (Kim et al., 2024) performs materials-aware curriculum masked language modeling guided by a materials semantic graph. Following the previous settings, we utilize the three-stage curriculum, which progressively masks entities based on their node degrees in the semantic graph.

B.2 Evaluation Benchmarks

MatSci-NLP. We evaluate generation performance on MatSci-NLP (Song et al., 2023), a low-resource multitask materials science benchmark comprising seven task types across 13 datasets. It covers 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). Following the original settings in MatSci-NLP, we evaluate domain-specific encoders using a transformer-based schema decoder tailored for generation-based tasks.

Classification Tasks. For encoder-only evaluation, we adopt standard materials-domain classification benchmarks following prior work (Gupta et al.,

2022; Kim et al., 2024), focusing on NER_{SOFC} , NER_{MS} (MatScholar NER), and SF (SOFC slot filling). These tasks have been widely used to assess domain-adapted encoders under standard fine-tuning settings in prior work.

MaScQA. We additionally evaluate materials-domain question answering on MaScQA (Zaki et al., 2024), which contains 650 undergraduate-level materials science questions curated from the GATE examination and categorized into four structural types (MCQ, MATCH, MCQN, NUM) across 14 domain categories. We report accuracy following the evaluation protocol described in Zaki et al. (2024).

C Training Configuration

All experiments are conducted on a single NVIDIA RTX A6000 GPU (48GB). Unless otherwise specified, we use AdamW optimization and keep the backbone model and overall training budget fixed across comparisons, varying only the data construction and training strategy.

C.1 Continual Pre-training

Encoder We continually pre-train SciBERT (Beltagy et al., 2019) on our chronologically ordered corpus. Following standard materials science domain-adaptation settings (Gupta et al., 2022; Kim et al., 2024), we train for 3 epochs using a hop-4 to hop-0 curriculum and apply whole-word masking with a 15% ratio.

Decoder We continually pre-train Llama-3.2-1B (Grattafiori et al., 2024) with Low-Rank Adaptation (LoRA) (Hu et al., 2022). LoRA modules are applied to both attention and MLP layers. Consistent with the encoder setting, training follows the hop-4 $\rightarrow$ hop-0 curriculum. We use bf16 mixed-precision training and gradient accumulation.

C.2 Fine-tuning

Encoder We followed the settings for generation tasks on MatSci-NLP (Song et al., 2023). We adopt an encoder-decoder architecture: the encoder is initialized from our continually pre-trained checkpoint, while the decoder (a 3-layer Transformer) is trained from scratch. We fine-tune for 20 epochs and train on 1% of the available training data. We report the average score over five different random seeds. For classification tasks, we fine-tune the encoder with task-specific classification heads. We

Method	Metric	MatSci-NLP							Overall
		NER	RC	EAE	PC	SAR	SC	SF	
SciBERT (Beltagy et al., 2019)	Micro-F1	0.738	0.818	0.458	0.671	0.701	0.901	0.500	0.693
	Macro-F1	0.517	0.600	0.290	0.568	0.528	0.612	0.258	0.482
DSP (Gururangan et al., 2020)	Micro-F1	0.792	<u>0.868</u>	0.482	0.725	0.837	0.904	<u>0.606</u>	0.745
	Macro-F1	<u>0.595</u>	0.646	0.304	0.632	0.745	<u>0.616</u>	<u>0.379</u>	0.560
EntityBERT (Lin et al., 2021)	Micro-F1	0.796	0.854	0.516	<u>0.784</u>	0.796	0.908	0.591	0.749
	Macro-F1	<u>0.595</u>	0.637	0.357	<u>0.709</u>	0.689	0.596	0.377	0.566
MELT (Kim et al., 2024)	Micro-F1	<u>0.780</u>	0.862	<u>0.529</u>	0.753	0.818	0.912	0.581	<u>0.748</u>
	Macro-F1	0.558	<u>0.655</u>	0.324	0.688	<u>0.698</u>	0.676	0.348	<u>0.568</u>
CHRONOS (ours)	Micro-F1	0.811	0.877	0.523	0.815	<u>0.839</u>	<u>0.908</u>	0.621	0.770
	Macro-F1	0.611	0.659	<u>0.341</u>	0.765	0.747	0.569	0.498	0.598

Table 11: Evaluation results on MatSci-NLP for the CHRONOS corpus. The best and the second best results are highlighted in **boldface** and underline, respectively.

Method	MatSci-NLP (Overall)		MaScQA
	Micro-F1	Macro-F1	
Year-ordered curriculum	0.243	0.178	20.92
CHRONOS (ours)	0.265	0.186	21.69

Table 12: Comparison between the citation-hop curriculum in CHRONOS and the publication-year-ordered curriculum on the same CHRONOS corpus.

select the learning rate from a small candidate set and choose epoch counts per task. For SOFC tasks, we report results averaged over five-fold cross-validation and three different seeds.

Decoder For decoder models, we fine-tune the decoder using LoRA targeting the attention layers. To simulate low-resource settings, we train on 1% of the available training data for 5 epochs, and constrain generation to a maximum of 64 output tokens. We report the average score over five different random seeds.

D Further Ablations

Results on CHRONOS corpus To further verify the effectiveness of our curriculum learning, we compared the other baselines on the same corpus. The results are shown in Table 11. Similar to the main results, we observe that our CHRONOS consistently achieves the best performance even in the same corpus. These results demonstrate that our chronological curriculum learning can further enhance the effectiveness of continued pre-training for materials science by incorporating cumulative knowledge.

Configuration	Micro-F1	Macro-F1
5 hops ($\Delta t = 10$)	0.770	0.598
3 hops ($\Delta t = 15$)	0.743	0.563
7 hops ($\Delta t = 7$)	0.756	0.583

Table 13: Aggregate MatSci-NLP performance under jointly varied hop counts and temporal gaps. All settings cover the same 1980–2025 period.

Comparison with Year-Ordered Curriculum

To compare two ways of organizing a chronological curriculum, we evaluate CHRONOS against a publication-year-ordered curriculum on the same CHRONOS corpus. The year-ordered curriculum uses the same backbone model, corpus, and training budget as CHRONOS, while organizing the curriculum stages according to publication years with stage sizes matched to the hop-based curriculum. As shown in Table 12, the citation-hop curriculum consistently achieves better performance than the year-ordered curriculum across all evaluation metrics on the same corpus. These results demonstrate the effectiveness of organizing the chronological curriculum based on citation hops rather than publication years.

Sensitivity analysis for time gap and hops To assess sensitivity to temporal schedule granularity, we compare three configurations that cover the same 1980-2025 period by jointly varying the hop count and temporal gap. As shown in Table 13, the main five-hop setting achieves the strongest aggregate performance among the evaluated con-

Method	Metric	MatSci-NLP							Overall
		NER	RC	EAE	PC	SAR	SC	SF	
MatSciBERT (Gupta et al., 2022)	Micro-F1	0.769	0.845	0.483	0.712	0.814	0.906	<u>0.614</u>	0.735
	Macro-F1	0.596	0.650	0.321	0.672	0.715	<u>0.645</u>	0.378	0.568
MatBERT (Trewartha et al., 2022)	Micro-F1	<u>0.806</u>	<u>0.870</u>	<u>0.493</u>	<u>0.731</u>	<u>0.832</u>	<u>0.907</u>	0.610	<u>0.750</u>
	Macro-F1	<u>0.605</u>	0.671	0.317	0.655	0.710	0.654	0.391	0.572
CHRONOS (ours)	Micro-F1	0.811	0.877	0.523	0.815	0.839	0.908	0.621	0.770
	Macro-F1	0.610	<u>0.659</u>	0.341	0.765	0.747	0.569	0.498	0.598

Table 14: Comparison with domain-specific PLMs (MatSciBERT, MatBERT) on MatSci-NLP. The best and the second best results are highlighted in **boldface** and underline, respectively.

figurations.

E Further Analysis

Comparison with Domain-Specific PLMs

To further contextualize the effectiveness of CHRONOS relative to existing domain-specific pre-trained models, we additionally compare CHRONOS with MatSciBERT (Gupta et al., 2022) and MatBERT (Trewartha et al., 2022) on the MatSci-NLP benchmark. Both baselines are SciBERT-based PLMs further pre-trained on materials science corpora that are larger than the corpus used for CHRONOS. As shown in Table 14, CHRONOS achieves better overall performance than both MatSciBERT and MatBERT on the MatSci-NLP benchmark, despite these baselines being pre-trained on larger materials science corpora. These results further demonstrate the effectiveness of incorporating chronological knowledge into PLMs in continued pre-training for the materials science domain.

Structured Information Inference (SII) To evaluate the model’s capability in extracting structured relational knowledge, we tested the Llama-3.2-1B models on SII tasks (Ye et al., 2024). Tables 15 and 16 show that CHRONOS consistently outperforms the DSP baseline in precision, recall, F1-score, and accuracy across various relational tasks.

F Corpus Impact and Diversity Analysis

To examine whether citation-based selection in CHRONOS introduces venue or topic bias, we compare the impact and diversity profiles of our corpus against the term-based random selection baseline at the same scale. We jointly use paper-level impact (FWCI, citation percentile, Semantic Scholar cita-

Task	Method	Precision	Recall	F1
NER	CHRONOS	91.21	89.50	90.28
	DSP	89.56	85.81	87.60
RE (A-B)	CHRONOS	82.60	83.21	82.78
	DSP	74.70	71.15	72.70
RE (A-C)	CHRONOS	80.53	74.66	76.79
	DSP	78.12	68.65	72.65
RE (ABC-D)	CHRONOS	71.78	74.66	69.82
	DSP	68.01	60.02	63.34

Table 15: NER and RE results on the SII benchmark.

Task	Method	Support	Accuracy
II	CHRONOS	76	82.89
	DSP	76	82.89
EN_U	CHRONOS	42	71.43
	DSP	42	66.67
EN_T	CHRONOS	167	79.64
	DSP	167	71.86

Table 16: II and EN results on the SII benchmark.

tions, SNIP), venue concentration, and field/topic breadth metrics, since impact scores alone cannot distinguish high-quality selection from prestige-driven narrowing.¹¹

As shown in Table 17, CHRONOS achieves a substantially higher impact profile than the term-based baseline across FWCI, citation percentile, Semantic Scholar citations, and SNIP (e.g., FWCI median 12.50 vs. 1.46), without inducing venue or topic collapse: CHRONOS retains a lower top-10 venue share (13.9% vs. 16.9%), a larger effective number of venues (1,321 vs. 952) and OpenAlex topics (1,147 vs. 685), and a lower top-10 topic share (4.3% vs. 9.9%).

¹¹Metadata coverage exceeds 96% across all reported metrics, except for Elsevier SNIP on the term-based baseline (88.4%).

Aspect	Metric	CHRONOS	Term-based
Scale	# papers	183,202	183,133
Impact profile	OpenAlex FWCI median / p90	12.50 / 55.16	1.46 / 9.92
	OpenAlex citation percentile median	0.992	0.815
	Semantic Scholar citationCount median	377	13
	Elsevier SNIP median	1.64	1.00
Venue concentration	Unique OpenAlex sources	7,973	10,353
	Effective # OpenAlex sources	1,321	952
	Top-10 source share	13.9%	16.9%
Broad field breadth	Effective # Semantic Scholar fields	7.09	5.81
	Top Semantic Scholar field share	37.0%	32.7%
Fine-grained topic breadth	Unique OpenAlex topics	3,000	2,727
	Effective # OpenAlex topics	1,147	685
	Top-10 OpenAlex topic share	4.3%	9.9%
High-impact subset	OpenAlex FWCI top-10% threshold	55.16	9.92
	Unique sources in FWCI top-10%	2,079	1,572
	Unique topics / top-10 topic share in FWCI top-10%	1,835 / 7.3%	1,307 / 22.9%

Table 17: Impact and diversity comparison with the term-based baseline.

Method	Micro-F1	Macro-F1
Term-based	0.728	0.553
CHRONOS (ours)	0.743	0.567

Table 18: Overall MatSci-NLP results for the FWCI-matched corpus control. Both corpora contain 9,701 papers and have a rounded median FWCI of 16.87.

To test whether the impact gain is concentrated in a narrow set of prestigious sources, we further inspect the high-impact subset (top 10% by FWCI) within each corpus. Even within this subset, CHRONOS retains broader coverage (2,079 vs. 1,572 venues, 1,835 vs. 1,307 topics, top-10 topic share 7.3% vs. 22.9%). These results demonstrate that the citation-based selection in CHRONOS improves corpus impact without sacrificing diversity, supporting the effectiveness of incorporating chronological knowledge into PLMs in continued pre-training for the materials science domain. Our framework is metric-agnostic and can readily incorporate finer-grained citation metrics (e.g., FWCI, impact factor) as the selection criterion in future extensions.

FWCI-matched corpus control To control for differences in scholarly impact, we construct two corpora matched in size and FWCI distribution without using downstream results. We shuffle the hop-0 papers using a fixed random seed and select a root-paper subset that satisfies the corpus-size and FWCI-matching criteria. Expanding this sub-

Method	8-gram	13-gram
CHRONOS (ours)	1.71%	0.05%
Random	3.11%	0.28%

Table 19: N-gram overlap ratio on MatSci-NLP tasks.

Method	Micro-F1 <i>p</i> -value	Macro-F1 <i>p</i> -value
MELT	0.03	0.02
ETS-DACP	0.009	0.007
EntityBERT	0.03	0.05
DSP	0.0005	0.0006

Table 20: Paired t-test *p*-values comparing CHRONOS with each continued-pretraining baseline on aggregate MatSci-NLP performance across five random seeds.

set through the citation links used in the original CHRONOS corpus construction yields 9,800 unique papers, of which 9,701 have valid FWCI, publication year, title, and abstract. We use these papers as the CHRONOS subset. After removing duplicate and overlapping DOIs from the term-based corpus, we select the same number of papers without replacement based on FWCI. Both corpora have a rounded median FWCI of 16.87. We require the difference in median FWCI to be at most 0.05 and the standardized mean difference (SMD), Kolmogorov-Smirnov (KS) statistic, and normalized Wasserstein distance computed on $\log(1 + \text{FWCI})$ to be at most 0.10. Both corpora use the same backbone model, continued pre-training budget, and downstream evaluation setup.

As shown in Table 18, CHRONOS achieves

Method	Metric	MatSci-NLP							Overall
		NER	RC	EAE	PC	SAR	SC	SF	
Llama-3.2-1B (Grattafiori et al., 2024)	Micro-F1	<u>0.006</u>	0.104	0.166	0.500	0.119	<u>0.483</u>	0.070	0.207
	Macro-F1	0.003	0.020	0.067	<u>0.459</u>	0.054	0.333	0.010	0.135
DSP (Gururangan et al., 2020)	Micro-F1	0.012	0.072	0.300	0.357	<u>0.210</u>	0.473	0.177	<u>0.229</u>
	Macro-F1	0.018	<u>0.038</u>	0.183	0.318	0.096	<u>0.441</u>	<u>0.057</u>	<u>0.165</u>
ETS-DACP (Xie et al., 2024)	Micro-F1	0.003	0.081	<u>0.303</u>	0.500	0.184	0.170	<u>0.118</u>	0.194
	Macro-F1	0.004	0.066	<u>0.199</u>	0.499	<u>0.092</u>	0.170	0.037	0.153
CHRONOS (ours)	Micro-F1	0.012	0.054	0.366	<u>0.454</u>	0.217	0.575	0.177	0.265
	Macro-F1	<u>0.017</u>	0.028	0.229	0.348	0.085	0.526	0.069	0.186

Table 21: Evaluation results on MatSci-NLP for the Llama-3.2-1B backbone. The best and the second best results are highlighted in **boldface** and underline, respectively.

Method	MaScQA
Llama-3.1-8B	30.00
DSP	30.46
CHRONOS (ours)	31.38

Table 22: Zero-shot MaScQA results for methods based on Llama-3.1-8B.

higher overall Micro-F1 and Macro-F1 scores than the term-based control under the matched corpus setting. These results demonstrate the effectiveness of CHRONOS after controlling for corpus size and scholarly impact. Together with the results on the same corpus in Table 2, this comparison further demonstrates the effectiveness of our data construction and curriculum learning.

G Further Analysis on Stability

Data Leakage Analysis To verify that the downstream performance gains are not a result of data leakage between the pre-training corpus and MatSci-NLP evaluation sets, we measured the N -gram overlap ratio. As shown in Table 19, our corpus exhibits equal or lower overlap compared to the random baseline.

Statistical Significance To assess the statistical reliability of the aggregate improvements, we conduct paired t-tests over the same five random seeds. As shown in Table 20, the reported p -values are at or below 0.05 for all four continued-pretraining baselines. These results support the aggregate performance advantage of CHRONOS over the evaluated continued-pretraining baselines.

H Detailed results for Decoder

The detailed results for Llama-3.2-1B are shown in Table 21. We observe that CHRONOS achieves better performance than the baselines on most evaluated materials science tasks. These results support the applicability of our method to the evaluated decoder setting. To evaluate CHRONOS beyond the 1B decoder, we additionally use the Llama-3.1-8B base model as a common backbone under the same zero-shot MaScQA setting. As shown in Table 22, CHRONOS achieves 31.38, compared with 30.46 for DSP and 30.00 for the base model. This result further demonstrates the effectiveness of CHRONOS on the larger model.

I Case Studies

To validate that our citation-based multi-hop data construction captures meaningful cumulative knowledge, we examine a representative case from the collected corpus. Specifically, we trace the citation network starting from a recent high-impact paper on perovskite/silicon tandem solar cells (Al-Ashouri et al., 2020) and analyze whether the collected papers across hops provide essential prior knowledge for understanding the target research.

Table 23 presents the representative papers collected through our hop-based expansion. The target paper (Hop 0) achieves >29% efficiency in monolithic tandem solar cells through enhanced hole extraction at the perovskite interface. We observe that papers collected in earlier hops systematically provide the foundational knowledge required to understand this achievement.

In Hop 1, the collected papers establish direct technical prerequisites: Bush et al. (2017) demon-

strates the monolithic tandem architecture that Hop 0 improves upon, while [Zhou et al. \(2014\)](#) introduces the interface engineering principles central to the target paper’s contribution. Moving to Hop 2, [Kojima et al. \(2009\)](#) represents the origin of perovskite photovoltaics, without which the entire field would not exist. Hop 3 captures the platform technology—the dye-sensitized solar cell ([O’Regan and Grätzel, 1991](#))—whose mesoporous electrode architecture directly influenced early perovskite device designs. Finally, Hop 4 contains the foundational discoveries: [Mitzi et al. \(1994\)](#) first demonstrated that halide perovskites exhibit semiconducting properties, and [Weber \(1978\)](#) established the crystal structure definition itself.

These results demonstrate that our citation-based data construction effectively captures the chronological evolution of knowledge, from fundamental material properties to state-of-the-art device engineering. By including such semantically coherent prior works in the training corpus, PLMs can progressively learn the cumulative materials science knowledge rather than encountering disconnected facts, thereby facilitating more effective domain adaptation.

Hop	Paper Title / Contribution
Hop 0	Monolithic perovskite/silicon tandem solar cell with >29% efficiency by enhanced hole extraction (Al-Ashouri et al., 2020) Contribution: SAM-based hole extraction; Monolithic tandem architecture
Hop 1	23.6%-efficient monolithic perovskite/silicon tandem solar cells with improved stability (Bush et al., 2017) Contribution: Monolithic tandem architecture; Direct predecessor to Hop 0 paper
	Interface engineering of highly efficient perovskite solar cells (Zhou et al., 2014) Contribution: Interface recombination suppression; Selective charge extraction
	Electron-hole diffusion lengths exceeding 1 micrometer in organometal trihalide perovskites (Stranks et al., 2013) Contribution: Long carrier diffusion length; Thick absorber feasibility
Hop 2	Efficient planar heterojunction perovskite solar cells by vapour deposition (Liu et al., 2013) Contribution: Planar device architecture; Scaffold-free structure
	Organometal halide perovskites as visible-light sensitizers for photovoltaic cells (Kojima et al., 2009) Contribution: First perovskite solar cell; Origin of the field
	Photoelectrochemical cells (Grätzel, 2001) Contribution: Photoconversion framework; Charge separation principles
Hop 3	A low-cost, high-efficiency solar cell based on dye-sensitized colloidal TiO ₂ films (O'Regan and Grätzel, 1991) Contribution: DSSC platform; Mesoporous electrode concept
	Solid-state dye-sensitized mesoporous TiO ₂ solar cells with high photon-to-electron conversion efficiencies (Bach et al., 1998) Contribution: Spiro-OMeTAD as HTL; Solid-state hole transport
Hop 4	Conducting tin halides with a layered organic-based perovskite structure (Mitzi et al., 1994) Contribution: Halide perovskite as semiconductor; Electronic conductivity discovery
	Organic electroluminescent diodes (Tang and VanSlyke, 1987) Contribution: Foundation of organic optoelectronics; Organic heterojunction principles
	Two-layer organic photovoltaic cell (Tang, 1986) Contribution: Bilayer heterojunction PV; Thin-film stacking concept

Table 23: Representative papers collected through citation-based multi-hop expansion from the target paper (Hop 0). Each hop contains foundational works whose knowledge directly contributed to understanding and developing the target research.