FLeW: Facet-Level and Adaptive Weighted Representation Learning of Scientific Documents

Zheng Dou[1], Deqing Wang[1,3(✉)], Fuzhen Zhuang[2,3], Jian Ren[1], and Yanlin Hu[4]

[1] SKLSDE, School of Computer Science, Beihang University, Beijing, China
{miracle_dz, dqwang, renjian}@buaa.edu.cn
[2] Institute of Artificial Intelligence, Beihang University, Beijing, China
zhuangfuzhen@buaa.edu.cn
[3] Zhongguancun Laboratory, Beijing, China
[4] National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing, China
yanlinhu@cert.org.cn

Abstract. Scientific document representation learning provides powerful embeddings for various tasks, while current methods face challenges across three approaches. 1) Contrastive training with citation-structural signals underutilizes citation information and still generates single-vector representations. 2) Fine-grained representation learning, which generates multiple vectors at the sentence or aspect level, requires costly integration and lacks domain generalization. 3) Task-aware learning depends on manually predefined task categorization, overlooking nuanced task distinctions and requiring extra training data for task-specific modules. To address these problems, we propose a new method that unifies the three approaches for better representations, namely FLeW. Specifically, we introduce a novel triplet sampling method that leverages citation intent and frequency to enhance citation-structural signals for training. Citation intents (background, method, result), aligned with the general structure of scientific writing, facilitate a domain-generalized facet partition for fine-grained representation learning. Then, we adopt a simple weight search to adaptively integrate three facet-level embeddings into a task-specific document embedding without task-aware fine-tuning. Experiments show the applicability and robustness of FLeW across multiple scientific tasks and fields, compared to prior models.

Keywords: Scientific Document Representation · Facet-Level Learning · Adaptive Weighted Embedding

1 Introduction

The rapid growth of scientific publications across diverse fields has spurred the need for high-quality document representations to support downstream tasks like classification, retrieval, and search [4, 15, 1]. Compared to general-purpose text, scientific documents exhibit unique relational structures and encapsulate more concentrated knowledge. Current approaches focus on these features of scientific documents for better representations while still facing challenges.

First, contrastive training with citation-structural signals has shown effectiveness in scientific document representation learning [4, 13]. However, they underutilize richer information from citation edges and generate single-vector representations, limiting their ability to capture fine-grained information. Second, fine-grained representation learning generates multiple vectors at the sentence or aspect level to capture detailed knowledge and information [12, 14]. However, these methods either require costly integration or lack generalization across different fields. Third, task-aware learning also generates multi-vector representations, with each vector tailored to a specific task category [18, 15]. But they need manually predefined task categorization and fail to capture nuanced differences within the same category. Additionally, such methods often require extra modules and additional training data to learn task-specific parameters.

In this work, we propose **F**acet-**L**evel and Adaptiv**e** **W**eighted representation learning of scientific documents, named **FLeW**, which unifies citation-structural contrastive training, fine-grained multi-vector representation, and task-aware learning into a single framework and address their challenges. Specifically, FLeW uses a novel triplet sampling method to utilize citation intent (background, method, result) and citation frequency for more informative contrastive training. The classification of citation intent aligns with the general structure of scientific writing, serving as a generalized facet partition for fine-grained multi-vector learning with more applicability across different fields. FLeW adopts a simple weight search to use the weighted sum of facet-level representations as the final document representation adaptive to multiple tasks instead of task-aware fine-tuning. By integrating task information into facet weights, FLeW effectively captures task-specific differences and improves the applicability across tasks. In summary, the contributions of this work are as follows:

- We propose a structural sampling method that uses citation intent and frequency information for enhanced citation-structural learning, along with a textual splitting method to divide triplet texts into faceted parts for better fine-grained textual representation.
- We propose to adopt a simple weight search and use the weighted sum of facet-level representations as the final document representation adaptive to multiple tasks instead of task-aware fine-tuning.
- Extensive experiments on a large-scale benchmark comprising 19 tasks and a citation recommendation dataset spanning 19 fields demonstrate that FLeW achieves superior performance compared to prior methods across multiple scientific tasks and fields.

2 Background

Document representation learning is the process of encoding a given input document $\mathcal{D}$ into a low-dimensional dense vector $\boldsymbol{V}$ by an *Encoder* model. The generated vector, also known as embedding or representation, contains rich semantic and context information of the document and can be used for vector

calculation in downstream tasks. BERT [5] is a pre-trained model with encoder-only architecture consisting of multiple layers of Transformer [16] to encode the tokens in a given input sequence. The final hidden state of special $[CLS]$ token is commonly used as an aggregate representation of the input sequence.

For a scientific document, the title provides a thorough overview of the paper's topic and the abstract is a comprehensive summary of the entire text. Following prior work [4, 13, 15], using the *title* and *abstract* separated by the $[SEP]$ token as input effectively balances input length constraints with high-quality representation. Since the great success of BERT on various NLP tasks especially text understanding, we use BERT-based model as the *Encoder* and also take the final hidden state of the $[CLS]$ token as output representation:

$$\boldsymbol{V} = Encoder(title[SEP]abstract)_{[CLS]} \tag{1}$$

Our goal is to obtain such pre-trained $Encoder(s)$ which can generate informative representation of any input scientific document (title and abstract) directly utilized for downstream tasks.

3 Methodology

3.1 Structural Sampling

A citation from a citing paper consists of one or several sentences referring to a specific aspect of the cited paper. This aspect, commonly referred to as citation intent, can be classified into three distinct categories: background, method, and result. Furthermore, the influence of cited papers within the same citing paper can vary significantly. An intuitive criterion is that the more frequently a cited paper is referenced within the citation contexts of a citing paper, the greater

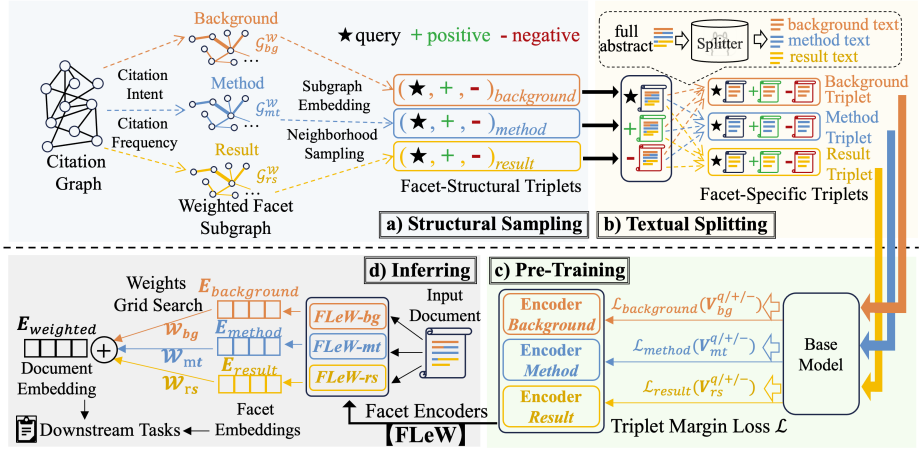

Fig. 1. Approach overview of FLeW.

its influence on the citing paper. Hence, citation edges can both be classified into three intents and be influence-weighted based on citation frequency. As shown in Fig.1(a), we propose a novel triplet sampling method utilizing this information to sample facet-structural triplets from citation graph for enhanced citation-structural contrastive learning.

Weighted Facet Subgraphs. We construct the original citation graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents the set of paper nodes, and $\mathcal{E}$ denotes the set of citation edges between citing and cited papers. Each edge $e \in \mathcal{E}$ is associated with citation contexts and annotated with citation intent labels ($\mathcal{T} = \{background\,(bg), method\,(mt), result\,(rs)\}$). Based on the intent labels, we extract three facet-specific subgraphs, $\mathcal{G}_{bg}, \mathcal{G}_{mt}, \mathcal{G}_{rs}$, where each subgraph captures citation edges corresponding to a specific intent. Formally, the facet subgraphs are defined as follows:

$$\mathcal{G}_x = (\mathcal{V}, \mathcal{E}_x), \quad \mathcal{E}_x = \{e \in \mathcal{E} \mid e^{\mathcal{T}=x}\}, \quad x \in \{bg, mt, rs\}. \tag{2}$$

To further incorporate the influence of citation frequency, we assign weights to the edges in each facet subgraph. According to our proposed criterion of influence, for the same citing paper, the cited paper with more citations is more influential than those with less citations. The weight $w(e)$ of an edge $e \in \mathcal{E}_x$ is determined by the number of citation contexts $|\mathcal{C}_x(e)|$ associated with the same intent x between citing and cited nodes. The resulting weighted facet subgraphs are formalized as:

$$\mathcal{G}_x^W = (\mathcal{V}, \mathcal{E}_x, \mathcal{W}_x), \quad \mathcal{W}_x(e) = |\mathcal{C}_x(e)|, \quad x \in \{bg, mt, rs\}. \tag{3}$$

Subgraph Embedding and Neighborhood Sampling. We utilize PyTorch-BigGraph (PBG) [9] to train and generate subgraph embeddings in the latent space derived from the citation-structural information. As PBG cannot be directly used to embed weighted graphs, we convert the weighted edges into repeated edges with a frequency equal to their weights. With the structural embedding results of papers in each facet subgraph respectively, distances between paper nodes become measurable. By conducting neighborhood sampling to obtain positive and negative papers based on queries, we finally get facet-structural triplets: $(query, positive, negative)_{background/method/result}$.

3.2 Textual Splitting

Our proposed structural sampling not only leverages the intent and frequency information on citation edges, but also samples three sets of facet-structural triplets from three weighted subgraphs as training data. Three types of triplets enable fine-grained multi-vector representations of scientific documents. The partition into three facets (background, method, and result) also aligns with the general structure of scientific writing, enhancing its applicability across fields.

As shown by Eq.(1), abstract texts serve as the primary content for representing scientific documents. But full abstract texts in our facet-structural triplets contain information from all three facets. To further learn facet-specific textual information, we propose splitting the full text of triplet abstracts into three faceted parts and retaining only the part corresponding to the type of triplet facet. The process of textual splitting is illustrated in Fig.1(b), where we use different colors to represent facet-specific texts.

Prior works [8, 3] have defined this task as *Sequential Sentence Classification* and train discriminative models on domain-specific labeled data. Considering multiple fields of our triplet papers and the issue of sentence boundary disambiguation in discriminative models, we propose to finish this task by a generative large language model (LLM) with instruction tuning as follows:

> **Instruction:** *Here is the abstract of a scientific paper. Your task is to split the text into three distinct sections based on the content of sentences:*
> *1.Background: The context or previous knowledge related to the topic.*
> *2.Method: The methodology, approach, or contribution proposed in the paper.*
> *3.Result: The findings, outcomes, or conclusions derived from experiments or analysis.*
> *Please return the output in a structured JSON format, as shown below:*
> {"*background*" : "*xxx*", "*method*" : "*xxx*", "*result*" : "*xxx*"}
> *Ensure that the original text remains intact in each section and that every sentence is categorized appropriately.*
> **Input:** "[*abstract*]"
> **Output:**{*"background":"abstract[bg]","method":"abstract[mt]","result":"abstract[rs]"*}

Following the template above, we first use GPT-4o to generate outputs of 12k papers as training data, and then conduct an instruction tuning on the Llama-3.1-8B-Instruct model [6] with LoRA [7] to adapt it to current task as *Splitter*. Finally, we get facet-specific triplets $(query_{[x]}, positive_{[x]}, negative_{[x]})_x$ enriched with both structural and textual information:

$$(query, positive, negative)_x \xrightarrow{Splitter} (query_{[x]}, positive_{[x]}, negative_{[x]})_x\ , \quad x \in \{background\,(bg), method\,(mt), result\,(rs)\} \tag{4}$$

3.3 Pre-Training

With facet-specific triplets from Structural Sampling and Textual Splitting, three encoders for representing facet-specific information are pre-trained on corresponding triplets respectively as shown in Fig.1(c). Triplet margin loss is used for contrastive learning to make queries and positives closer while queries and negatives further:

$$\mathcal{L}_x = \max\{f(\boldsymbol{V}_x^q, \boldsymbol{V}_x^+) - f(\boldsymbol{V}_x^q, \boldsymbol{V}_x^-) + \delta, 0\}, \quad x \in \{bg, mt, rs\}. \tag{5}$$

where f is a L2 norm distance function and δ is the margin hyperparameter ($\delta = 1$ as prior works). $\boldsymbol{V}_x$ is obtained following Eq.(1) except a slight difference that the abstract text used here is facet-specific $abstract_{[x]}$ from textual splitter.

The three facet encoders Encoder_x, optimized from the base model through their respective loss functions $\mathcal{L}_x$, serve as the final pre-trained models directly for scientific document representation, achieving the goal outlined in Section 2.

$$\begin{gathered}\text{Encoder}_x = f(\cdot;\theta_x^*), \quad \theta_x^* = \arg\min_{\theta} \mathcal{L}_x(\theta), \quad \theta^{(0)} = \theta_{base}\ , \\ x \in \{background\,(bg), method\,(mt), result\,(rs)\}\end{gathered} \tag{6}$$

3.4 Inferring

Considering the contextual information in the full abstract text helpful to document representation and the extra cost of splitting every input abstract into facet-based parts, we use the title and full abstract text as inputs during inference. Enhanced by the facet-textual $abstract_{[x]}$ during training, facet encoders can focus on representing facet-specific information even with the full abstract text. As shown in Fig.1(d), three pre-trained encoders, referred to as FLeW-bg, FLeW-mt, and FLeW-rs, generate three facet-level embeddings based on the title and full abstract of an input scientific document:

$$\begin{aligned}\boldsymbol{E_{background}} &= \text{FLeW-bg}(title[SEP]abstract)_{[CLS]} \\ \boldsymbol{E_{method}} &= \text{FLeW-mt}(title[SEP]abstract)_{[CLS]} \\ \boldsymbol{E_{result}} &= \text{FLeW-rs}(title[SEP]abstract)_{[CLS]}\end{aligned} \tag{7}$$

We use a simple weighted sum of three representations as the final document representation, where facet weights are adaptive to different downstream tasks and reflect the importance of each facet. $E_{weighted}$ is the document-level embedding of our proposed FLeW:

$$\begin{aligned}&\boldsymbol{E_{weighted}} = \mathcal{W}_{bg} * \boldsymbol{E_{background}} + \mathcal{W}_{mt} * \boldsymbol{E_{method}} + \mathcal{W}_{rs} * \boldsymbol{E_{result}} \\ &s.t. \qquad \mathcal{W}_{bg} + \mathcal{W}_{mt} + \mathcal{W}_{rs} = 1, \mathcal{W}_{bg} > 0, \mathcal{W}_{mt} > 0, \mathcal{W}_{rs} > 0\end{aligned} \tag{8}$$

Because of the limitation that the weight sum must be 1, there are actually only two free weight parameters ranging from 0 to 1. We use a simple grid search to iterate across all weight combinations and rank by evaluation score for each task on the validation dataset to select optimal weights.

4 Experiments

4.1 Implementation Details

We use the 2024-04-02 release version of Semantic Scholar Academic Graph (S2AG) [17] as our base corpus. In S2AG, we download *citations* dataset to build citation graph for structural sampling, and download *papers* and *abstracts* datasets to obtain metadata (title and abstract) of papers for textual splitting. Then we collect a query list of 216k papers for facet triplet sampling. By conducting neighborhood sampling across three facet subgraphs,we generate 2.16M

triplets with 2.95M, 2.81M and 2.94M metadata for training three facet encoders, respectively. During training, we initialize the model with SciBERT [2] and use Adam with weight decay as optimizer with the initial learning rate $\lambda = 2^{-5}$, following [4, 13]. Our facet encoders are trained on 2 NVIDIA GeForce RTX 3090 (24G) with batch size 10 for 2 epochs. We release our data and encoder models[5].

4.2 Baseline Models

We compare our method to the base model SciBERT [2] and other prior works with similar triplet sampling and contrastive learning approach: SPECTER [4], SciNCL [13] and SPECTER-2 [15]. Besides, to show the distinctiveness of scientific documents, there is also a comparison to SFR-2_R[6], the leader model (2024-06-14) on MTEB benchmark [11], which represents the performance of general-purpose text embedding models.

4.3 Experimental Results

Multiple Tasks Evaluation. Table 1 presents the document-level evaluation results across 19 tasks in four formats on SciRepEval benchmark [15]. Overall, our model FLeW achieves the best performance in 13 out of 19 tasks, and also obtains the second-best performance in 4 out of the remaining 6 tasks. The average performance across the total benchmark achieves the best result (60.81), surpassing the second-best by +0.60. FLeW also achieves the best score on average across all the four formats. For each format, FLeW performs best

Table 1. Document-level evaluation on SciRepEval benchmark across 19 tasks.

Format	Proximity [PRX]						Regression [RGN]					
Task	Relish	Reviewer	Same Author	HighInf Citation	SciDocs PRX	[PRX] Avg	Review Score	Max H-index	Tweet Mentions	Citation Count	Pub Year	[RGN] Avg
Metric	nDCG	Avg. P	MAP	MAP	nDCG		K Tau	K Tau	K Tau	K Tau	K Tau	
SciBERT	82.81	41.68	79.48	33.72	76.15	62.77	20.26	12.38	22.18	39.16	27.71	24.34
SPECTER	90.07	45.16	86.53	42.89	93.98	71.73	17.35	12.52	24.19	33.21	25.96	22.65
SciNCL	90.67	45.40	**87.47**	43.39	**95.01**	72.39	18.87	13.45	25.79	34.61	28.99	24.34
SPECTER-2	91.63	45.42	87.00	44.96	94.87	72.78	20.63	12.76	**27.11**	38.33	33.65	26.50
SFR-2_R	88.36	45.56	82.09	39.52	87.10	68.53	11.18	8.31	14.09	36.24	31.23	20.21
FLeW(ours)	**92.15**	**45.70**	87.03	**45.50**	94.46	**72.97**	**20.93**	**15.09**	27.01	**39.21**	**33.90**	**27.23**

Format	Query [QRY]				Classification [CLF]							Total AVG
Task	NFC-orpus	TREC Covid	Search	[QRY] Avg	Biomimicry	DRSM	MeSH	FoS	SciDocs MAG	SciDocs MeSH	[CLF] Avg	
Metric	nDCG	nDCG	nDCG		Wt. F1	Wt. F1	F1	F1	F1	F1		
SciBERT	53.34	79.73	71.49	68.19	50.00	64.01	76.71	42.67	79.50	79.97	65.48	54.37
SPECTER	64.90	86.53	73.25	74.89	**51.22**	**66.16**	85.46	43.00	79.75	87.80	68.90	58.42
SciNCL	70.85	87.67	73.46	77.33	50.22	65.10	86.17	43.53	81.11	89.13	69.21	59.52
SPECTER-2	72.03	**89.46**	73.76	78.42	50.53	62.96	86.76	42.16	81.03	89.00	68.74	60.21
SFR-2_R	67.76	86.97	70.89	75.21	47.33	63.40	81.96	38.52	78.63	86.68	66.09	56.10
FLeW(ours)	**72.64**	89.14	**74.13**	**78.64**	50.25	65.41	**87.17**	**44.61**	**81.71**	**89.39**	**69.76**	**60.81**

[5] https://huggingface.co/collections/Miracle-dz/flew-67b3074bce03f9a0573cd94d

[6] https://huggingface.co/Salesforce/SFR-Embedding-2_R

in more than half of the tasks ([PRX]$_{3/5}$,[RGN]$_{4/5}$,[QRY]$_{2/3}$,[CLF]$_{4/6}$), while the best model for the remaining tasks varies across four formats ([PRX]$_{\text{SciNCL}}$, [RGN]&[QRY]$_{\text{SPECTER-2}}$, [CLF]$_{\text{SPECTER}}$). The results show the stability and generalization ability of FLeW across multiple scientific tasks.

Multiple Fields Evaluation. Table 2 presents document-level evaluation results across 19 fields for the citation recommendation task on MDCR dataset [10]. Our model, FLeW, achieves the best performance in almost every field and on average. Compared to other baseline models, SPECTER-2 is more competitive and outperforms our model slightly in several fields. We find that these fields are more humanities-oriented, such as *Philosophy* and *Political Science*, which have fewer "scientific" attributes and lack a clear writing structure consisting of three facets (background, method, result). In contrast, FLeW shows significant improvements in more scientific fields which clearly follow the general writing structure, such as *Biology* and *Chemistry*. This demonstrates the effectiveness of our proposed method in capturing the facet-specific information of scientific documents and improving the applicability across multiple scientific fields.

Table 2. Document-level evaluation on MDCR dataset across 19 fields.

Metric	MAP						Recall@5					
Domain/Model	BM25	Sci BERT	SPEC TER	Sci NCL	SPEC TER-2	FLeW (ours)	BM25	Sci BERT	SPEC TER	Sci NCL	SPEC TER-2	FLeW (ours)
Art	38.2	22.4	34.1	34.7	43.4	**43.7**	32.3	16.6	28.8	29.2	**37.8**	37.5
Biology	38.3	20.4	34.6	36.8	39.9	**43.5**	33.6	14.0	30.0	32.3	33.3	**37.8**
Business	28.1	19.1	27.5	28.5	35.0	**35.1**	22.5	13.1	21.8	24.6	**30.5**	30.0
Chemistry	38.0	20.0	33.7	36.5	39.7	**43.0**	32.6	13.7	29.3	31.5	33.5	**38.2**
Computer Science	34.8	19.5	35.6	37.2	38.5	**39.9**	30.5	12.8	30.4	32.2	33.4	**36.2**
Economics	30.5	21.5	27.3	28.3	33.7	**34.7**	26.0	15.4	21.9	23.2	28.5	**29.7**
Engineering	34.6	20.5	31.3	34.2	35.4	**39.2**	29.3	13.9	27.3	28.0	30.3	**33.2**
Environmental Science	31.6	21.3	30.1	31.5	35.0	**37.2**	26.2	15.1	24.2	25.5	27.9	**32.0**
Geography	31.8	21.9	26.4	29.5	37.1	**40.1**	27.8	16.7	22.2	23.8	31.8	**35.0**
Geology	33.1	19.5	24.8	25.7	33.4	**35.2**	28.0	13.9	20.0	19.9	27.7	**29.1**
History	38.1	20.9	27.1	30.9	41.9	**43.2**	32.9	15.3	20.6	23.9	34.7	**37.6**
Materials Science	36.1	22.2	34.1	35.8	39.7	**40.6**	30.7	15.5	28.2	29.6	34.0	**35.4**
Mathematics	35.3	22.8	34.2	34.9	40.8	**41.2**	28.3	18.3	28.9	30.1	34.2	**35.0**
Medicine	38.6	22.0	41.4	42.7	43.8	**46.3**	32.5	16.4	36.3	36.5	39.0	**41.6**
Philosophy	30.2	19.2	27.1	29.9	**37.2**	36.6	25.7	13.3	21.1	23.5	**32.0**	31.5
Physics	35.1	23.9	30.8	34.5	37.6	**38.4**	30.2	18.2	26.3	30.3	32.5	**34.5**
Political Science	28.6	19.4	24.2	26.4	**35.7**	35.3	23.1	14.0	18.0	21.7	**31.6**	30.1
Psychology	32.5	20.3	32.3	34.2	38.8	**40.5**	28.9	16.2	28.1	30.5	33.2	**36.1**
Sociology	26.8	20.2	25.2	26.7	34.6	**34.8**	20.5	15.8	20.5	21.9	29.8	**30.2**
Avg	33.7	20.9	30.6	32.6	38.0	**39.4**	28.5	15.2	25.5	27.3	32.4	**34.2**

4.4 Ablation Study

Table 3 presents the results of ablation study on SciRepEval benchmark [15]. We train three additional facet encoders (FLeW-bg/mt/rs) with the full abstract texts of facet training triplets without textual splitting (*-w/o Textual*) compared to encoders trained with facet-textual triplets (*+w/ Textual*). And SPECTER-2 is used as the representative model from prior works (*-w/o Structural*).

For facet-level results (FLeW-xx), FLeW-bg outperforms other facet encoders when compared across facets. We attribute this phenomenon to the position bias, as the background facet often appears earlier in the full text, attracting more focus and avoiding the risk of truncation. Our proposed Textual Splitting eliminates this position bias, allowing each facet to be learned equally. The significant improvements of FLeW-rs (*+w/Textual*) provide strong support for this result. For document-level results (FLeW), the improvement of performance compared to FLeW-xx indicates the effectiveness of our facet weighted sum strategy, which emphasizes the different importance of each facet adaptive to multiple scientific tasks and fields. Besides, there is a stepwise increase of average performance in each format and total average in the last three rows (e.g. 60.21, 60.66, 60.81). This demonstrates the effectiveness of our proposed Structural Sampling and Textual Splitting, which help to capture the facet-specific information of scientific documents both structurally and textually.

Table 3. Ablation study on SciRepEval benchmark.

	Format	Proximity [PRX]						Regression [RGN]					
	Task	Relish	Reviewer	Same Author	HighInf Citation	SciDocs PRX	[PRX] Avg	Review Score	Max H-index	Tweet Mentions	Citation Count	Pub Year	[RGN] Avg
	Metric	nDCG	Avg. P	MAP	MAP	nDCG		K Tau	K Tau	K Tau	K Tau	K Tau	
+ w/ Textual	FLeW-bg	**92.10**	**45.02**	**86.54**	**45.45**	**93.64**	**72.55**	**19.05**	12.26	**26.20**	36.98	**31.36**	**25.17**
	FLeW-mt	**90.68**	44.96	**86.38**	**43.96**	**93.41**	**71.88**	**19.19**	12.95	**26.70**	**38.46**	32.13	**25.89**
	FLeW-rs	**89.82**	**45.14**	**85.53**	**42.17**	**93.30**	**71.19**	**20.79**	**14.83**	26.13	**36.70**	**32.76**	**26.24**
- w/o Textual	FLeW-bg	91.99	44.79	85.99	45.41	93.21	72.28	18.91	**13.24**	25.34	**37.16**	30.58	25.05
	FLeW-mt	90.50	**45.49**	86.00	43.95	93.14	71.82	18.80	**13.82**	26.17	35.11	**33.62**	25.50
	FLeW-rs	89.47	44.86	84.66	41.57	92.71	70.65	17.27	13.82	**27.42**	35.42	31.71	25.13
FLeW		**92.15**	45.70	**87.03**	45.50	94.46	**72.97**	**20.93**	**15.09**	27.01	**39.21**	33.90	**27.23**
- w/o Textual		92.09	**45.82**	86.65	**45.60**	94.22	72.88	20.23	14.78	**27.57**	38.60	**34.16**	27.07
- w/o Structural		91.63	45.42	87.00	44.96	**94.87**	72.78	20.63	12.76	27.11	38.33	33.65	26.50

	Format	Query [QRY]				Classification [CLF]							Total AVG
	Task	NFC-orpus	TREC Covid	Search	[QRY] Avg	Biomimicry	DRSM	MeSH	FoS	SciDocs MAG	SciDocs MeSH	[CLF] Avg	
	Metric	nDCG	nDCG	nDCG		Wt. F1	Wt. F1	F1	F1	F1	F1		
+ w/ Textual	FLeW-bg	71.63	**88.99**	**73.69**	**78.10**	**48.02**	**64.14**	**86.71**	41.31	**81.04**	87.70	**68.15**	**59.57**
	FLeW-mt	68.06	**87.51**	**73.34**	**76.30**	**48.92**	**64.28**	85.68	**43.35**	**80.95**	88.43	**68.60**	**59.44**
	FLeW-rs	67.31	85.23	**73.02**	75.19	**49.53**	**62.06**	**85.39**	42.39	**80.90**	**88.32**	**68.10**	**59.02**
- w/o Textual	FLeW-bg	**71.65**	88.64	73.68	77.99	47.82	62.44	86.19	**42.01**	80.35	**88.72**	67.92	59.37
	FLeW-mt	**68.87**	86.59	73.12	76.19	48.74	64.26	**86.37**	41.14	80.80	**89.01**	68.39	59.24
	FLeW-rs	**68.82**	**85.25**	72.65	**75.57**	46.59	59.96	84.52	**43.62**	79.83	87.48	67.00	58.30
FLeW		72.64	89.14	**74.13**	**78.64**	50.25	**65.41**	**87.17**	**44.61**	**81.71**	89.39	**69.76**	**60.81**
- w/o Textual		**72.89**	88.79	74.00	78.56	49.81	64.47	86.83	44.55	81.57	**89.83**	69.51	60.66
- w/o Structural		72.03	**89.46**	73.76	78.42	**50.53**	62.96	86.76	42.16	81.03	89.00	68.74	60.21

5 Conclusion

In this work, we propose FLeW to learn facet-level and adaptive weighted representations of scientific documents. Our approach unifies three key strategies for improved representation learning: citation-structural contrastive training, fine-grained multi-vector representation, and extra task-aware learning. We propose Structural Sampling, which leverages citation intent and frequency information

to enhance citation-structural signals for contrastive training. By aligning citation intent with a generalized facet partition of scientific documents, we further propose Textual Splitting to split full text of triplets into faceted parts for better fine-grained textual embedding. Three facet encoders are trained on facet-specific triplets enriched with structural and textual information. With three encoders for representation, FLeW adopts a simple weight search to use the weighted sum of three facet representations as adaptive document representation instead of task-aware fine-tuning. Experiments show the applicability and robustness of FLeW across multiple scientific tasks and domains, compared to prior methods.

Acknowledgments. This research work was supported by the National Natural Science Foundation of China (Grant No. 62276015).

References

1. Ban, Y., et al.: Pagerank bandits for link prediction. Advances in Neural Information Processing Systems **37**, 21342–21376 (2025)
2. Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. In: EMNLP-IJCNLP. pp. 3615–3620 (2019)
3. Cohan, A., et al.: Pretrained language models for sequential sentence classification. In: EMNLP-IJCNLP. pp. 3693–3699 (2019)
4. Cohan, A., et al.: Specter: Document-level representation learning using citation-informed transformers. In: ACL. pp. 2270–2282 (2020)
5. Devlin, J., et al.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL. pp. 4171–4186 (2019)
6. Dubey, A., et al.: The llama 3 herd of models. arXiv:2407.21783 (2024)
7. Hu, E.J., et al.: Lora: Low-rank adaptation of large language models. In: ICLR (2022)
8. Jin, D., Szolovits, P.: Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. In: EMNLP. pp. 3100–3109 (2018)
9. Lerer, A., et al.: Pytorch-biggraph: A large scale graph embedding system. In: Proceedings of Machine Learning and Systems. pp. 120–131 (2019)
10. Medić, Z., Šnajder, J.: Large-scale evaluation of transformer-based article encoders on the task of citation recommendation. In: ACL-sdp (2022)
11. Muennighoff, N., et al.: Mteb: Massive text embedding benchmark. arXiv preprint arXiv:2210.07316 (2022)
12. Mysore, S., Cohan, A., Hope, T.: Multi-vector models with textual guidance for fine-grained scientific document similarity. In: NAACL. pp. 4453–4470 (2022)
13. Ostendorff, M., et al.: Neighborhood contrastive learning for scientific document representations with citation embeddings. In: EMNLP. pp. 11670–11688 (2022)
14. Ostendorff, M., et al.: Specialized document embeddings for aspect-based similarity of research papers. In: JCDL. pp. 1–12 (2022)
15. Singh, A., et al.: Scirepeval: A multi-format benchmark for scientific document representations. In: EMNLP. pp. 5548–5566 (2023)
16. Vaswani, A., et al.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
17. Wade, A.D.: The semantic scholar academic graph (s2ag). In: Companion Proceedings of the Web Conference. pp. 739–739 (2022)
18. Zhang, Y., et al.: Pre-training multi-task contrastive learning models for scientific literature understanding. In: EMNLP (2023)