computers

MDPI

Article

Learning Scientific Document Representations via Triple-Source Automatic Supervision Without Annotations or Citations

Mussa Turdalyuly [1,2,*], Ainur Tursynkhan [3,4,*], Aigerim Yerimbetova [1,2], Tolganay Turdalykyzy [1,2,*], Bakzhan Sakenov [1,*], Nurzhan Mukazhanov [1,5] and Nazerke Baisholan [3,4]

1 Institute of Information and Computational Technologies, Almaty 050010, Kazakhstan; aigerian8888@gmail.com (A.Y.); nurzhan.mukazhanov@narxoz.kz (N.M.)
2 School of Engineering and Information Technology, META University, Almaty 050000, Kazakhstan
3 Software Engineering Department, International Engineering and Technological University, Almaty 050060, Kazakhstan; nbaishol@purdue.edu
4 Department of Computer and Information Technology, Purdue University, West Lafayette, IN 47907, USA
5 School of Digital Technologies, Narxoz University, Almaty 050035, Kazakhstan
* Correspondence: m.turdalyuly@gmail.com (M.T.); tursynhan.aynur@mail.ru (A.T.); t.turdalykyzy@etu.edu.kz (T.T.); sbakzhan22@gmail.com (B.S.)

Abstract

Learning meaningful representations of scientific documents is essential for information retrieval, knowledge discovery, and recommendation systems. Traditional methods such as TF-IDF rely on lexical matching and fail to capture deeper semantic relationships, while transformer-based approaches typically depend on limited supervision signals. In this work, we propose a Triple-Source automatic supervision framework for learning document embeddings from scientific corpora. The model integrates three types of supervision–title–abstract pairs, same-category document pairs, and document-level semantic relationships—within a unified contrastive learning framework based on a multilingual XLM-RoBERTa encoder. Unlike prior approaches that rely on citation graphs or manual annotations, our method enables citation-free and annotation-free representation learning using only lightweight metadata. Experiments on a publicly available arXiv dataset consisting of 98,649 documents demonstrate improved semantic retrieval performance, achieving Recall@1 = 0.6181 for same-category retrieval and outperforming both TF-IDF and single-source transformer baselines. The learned embeddings also exhibit improved clustering of scientific domains, indicating more structured semantic representations.

Keywords: semantic search; document retrieval; scientific document embeddings; contrastive learning; information retrieval; scientific literature analysis

Check for updates

Academic Editor: Phivos Mylonas

Received: 23 March 2026
Revised: 19 April 2026
Accepted: 21 April 2026
Published: 23 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The rapid growth of scientific publications across multiple disciplines has created an urgent need for effective methods to search, organize, and analyze research documents. Large-scale repositories such as arXiv, PubMed, and Semantic Scholar contain millions of papers, making manual exploration of the literature increasingly impractical. As a result, learning high-quality semantic representations of scientific documents has become a central problem in natural language processing and information retrieval.

Traditional document retrieval approaches rely on lexical matching methods such as TF-IDF and BM25. These methods represent documents using term frequency statistics and are effective for keyword-based search. However, they often fail to capture deeper semantic

relationships between documents that are conceptually similar but differ in vocabulary. This limitation is particularly critical in scientific domains, where the same research ideas can be expressed using different terminology or writing styles [1].

Recent advances in transformer-based language models have markedly improved semantic representation learning. Models such as BERT and its variants learn contextualized representations of text and have demonstrated strong performance across a wide range of NLP tasks [2]. In the scientific domain, specialized models such as SciBERT further improve performance by leveraging domain-specific corpora [3]. These models enable more accurate semantic similarity estimation compared to traditional lexical approaches. While these approaches demonstrate improved performance, they often depend on specific types of metadata and do not fully generalize across datasets. Recent studies have also explored information extraction and semantic modeling of scientific documents across multiple domains, emphasizing the need for reliable and scalable representation learning methods [4].

Building on these developments, sentence-level embedding models such as Sentence-BERT have introduced contrastive learning frameworks for efficient semantic similarity computation [1]. Similar approaches have been applied to scientific documents for tasks such as citation recommendation and semantic search [2]. More recently, contrastive learning methods have been extended to unsupervised and weakly supervised settings, where positive pairs are constructed automatically from textual structure or metadata [5].

Despite these advances, a key limitation of existing approaches is their reliance on a single supervision signal. Most models are trained using either title–abstract pairs, citation links, or other isolated metadata sources. While each of these signals captures a specific aspect of semantic relationships, relying on a single source limits the diversity of semantic information available during training. As a result, learned representations may fail to fully capture the complex, multi-level structure of scientific knowledge.

Scientific repositories naturally contain multiple types of metadata that encode distinct semantic relationships. For example, title–abstract pairs capture fine-grained document alignment, while category labels provide information about broader topical similarity. Additionally, full document representations (e.g., combined title and abstract) can capture higher-level semantic consistency across documents. However, effectively integrating these heterogeneous signals into a unified representation learning framework remains a challenging problem.

To address this limitation, we propose a Triple-Source automatic supervision framework for learning semantic representations of scientific documents. The key idea is to jointly leverage multiple automatically generated supervision signals derived from scientific metadata within a unified contrastive learning framework. Specifically, our approach integrates three distinct sources of supervision: (1) title–abstract pairs for fine-grained semantic alignment, (2) same-category abstract pairs for capturing topical similarity, and (3) document–document pairs constructed from full document representations for modeling global semantic consistency.

The proposed method is implemented using a multilingual transformer encoder based on XLM-RoBERTa and trained using a contrastive learning objective. By combining multiple supervision signals, the model learns richer semantic representations that better reflect the structure of scientific literature.

We evaluate our approach on a large-scale corpus of 98,649 scientific documents derived from the arXiv dataset. The model is assessed on document retrieval tasks and compared with both lexical baselines (TF-IDF) and transformer-based models trained with single-source supervision.

Experimental results demonstrate that our Triple-Source framework clearly improves semantic retrieval performance, particularly for same-category document retrieval. The model achieves Recall@1 = 0.6181 and MRR = 0.7124, outperforming both TF-IDF and single-source transformer baselines.

Unlike prior work that trains on a single type of signal, we jointly exploit three heterogeneous metadata sources within one training objective. This moves beyond isolated supervision and allows the model to capture semantic structure at multiple levels—from fine-grained title–abstract alignment to broader topical similarity across documents.

The main contributions of this work are as follows:

- We introduce a novel paradigm of multi-source automatic supervision for scientific document representation learning, where heterogeneous metadata signals are jointly exploited within a unified contrastive framework.
- We propose the first Triple-Source supervision strategy that integrates fine-grained alignment (title–abstract), topical similarity (same-category), and document-level semantic consistency without relying on citation networks or manual annotations.
- We demonstrate that combining multiple supervision sources leads to a structured embedding space that improves semantic generalization, particularly for same-category retrieval tasks.
- We provide a systematic empirical analysis of single-source, dual-source, and triple-source supervision, revealing the trade-offs between exact alignment and semantic abstraction.
- We show that our approach supports scalable and domain-agnostic representation learning using only widely available metadata, making it applicable to diverse scientific corpora.

2. Related Work

2.1. Scientific Document Representation

Learning semantic representations of scientific documents has been widely studied in natural language processing and information retrieval. Early approaches relied on lexical matching techniques such as TF–IDF and probabilistic retrieval models including BM25, which represent documents based on term frequency statistics. While these methods are computationally efficient and remain widely used in practical systems, they are limited in their ability to capture semantic relationships between documents that differ in vocabulary but share similar research content [1].

To overcome these limitations, latent representation methods such as topic modeling have been proposed. Techniques such as Latent Dirichlet Allocation (LDA) model documents as mixtures of latent topics inferred from word co-occurrence patterns [6]. Although these approaches improve representation quality by capturing thematic structure, they rely on relatively shallow statistical assumptions and are often insufficient for modeling complex semantic relationships in scientific text.

Recent advances in transformer-based language models have advanced semantic representation learning. Models such as BERT learn contextualized embeddings that capture dependencies between words and phrases, enabling more accurate semantic similarity estimation [2]. Domain-specific models such as SciBERT further enhance performance by pretraining on scientific corpora, demonstrating improved results on scientific NLP tasks [3].

Several studies have extended transformer-based models to document-level representation learning. For example, SPECTER leverages citation relationships to train document embeddings that reflect scientific influence and topical similarity [2]. Similarly, large-scale datasets such as S2ORC provide rich metadata for representation learning, including ci-

tation networks and authorship information [7]. However, many of these approaches rely on a single dominant supervision signal, such as citation links, which may not fully capture the diversity of semantic relationships present in scientific literature. More recent retrieval-oriented representation learning approaches have further explored dense scientific embeddings and semantic retrieval architectures. For example, recent work has investigated embedding-based retrieval systems that outperform lexical methods in scientific corpora and neighborhood-based semantic retrieval settings. These studies confirm the importance of richer semantic embeddings for scientific document retrieval and motivate the need for integrating multiple supervision signals within a unified framework [8,9].

More recent work has explored combining textual embeddings with metadata features, such as category labels or authorship information, to improve document representation quality [10]. While these approaches demonstrate improved performance, they often depend on specific types of metadata that may not be consistently available across datasets. Furthermore, they typically incorporate these signals in isolation rather than within a unified training framework. Recent studies have also explored information extraction and semantic modeling of scientific documents across multiple domains, emphasizing the need for reliable and scalable representation learning methods [4].

Overall, existing methods either rely on lexical representations with limited semantic capacity or on neural models trained using a single supervision signal. This limitation motivates the need for approaches that can integrate multiple different sources of semantic information.

Recent journal studies published in 2025–2026 further support the growing role of dense semantic representations in scholarly search and recommendation. Colangelo et al. evaluated Sentence Transformer models for automated journal recommendation using PubMed metadata, showing the effectiveness of embedding-based matching for scientific manuscripts [11]. Fahrudin et al. integrated sentence embeddings into an automated reference paper collection pipeline to improve relevance estimation during scholarly paper retrieval [12]. In addition, Nosov et al. demonstrated the usefulness of machine learning for semantic analysis and knowledge extraction from scientific publications [13], while Malashin et al. explored unsupervised semantic normalization for large-scale analysis of the scientific literature [14].

2.2. Contrastive Learning for Text Representations

Contrastive learning has emerged as an effective paradigm for learning semantic representations without requiring manually labeled data. In contrastive frameworks, models are trained to bring semantically related pairs closer in the embedding space while pushing unrelated pairs apart.

Sentence-BERT introduced a Siamese transformer architecture optimized using contrastive loss functions, enabling efficient computation of semantic similarity between sentence embeddings [1]. This approach has become a standard method for tasks such as semantic search and document retrieval.

Subsequent work has explored unsupervised and weakly supervised contrastive learning strategies. For example, SimCSE generates positive pairs using dropout-based augmentation, allowing sentence embeddings to be learned without explicit supervision [5]. Other approaches construct positive pairs using structural or metadata-based signals, such as sentence proximity, document hierarchy, or citation relationships.

In the context of scientific literature, contrastive learning has been applied using supervision signals derived from citation networks or document metadata [15,16]. In particular, Ostendorff et al. demonstrated that neighborhood-based contrastive objectives using citation graph embeddings improve scientific document retrieval performance [16].

These methods demonstrate that automatically generated supervision signals can effectively replace manual annotation and improve representation learning.

More recent work has also continued to improve embedding quality through stronger contrastive formulations and systematic evaluation. Huang et al. proposed an LLM-driven sentence embedding framework with ranking and label smoothing to address asymmetry in contrastive learning [17]. Oro et al. provided a comparative evaluation of embedding models for information retrieval and question answering across English and Italian, further confirming the importance of embedding selection and evaluation for semantic retrieval settings [18].

However, most existing contrastive learning approaches rely on a single type of positive pair construction. For example, models may use only citation-based pairs or only sentence-level augmentation. This limits the diversity of semantic relationships that the model can learn. In scientific corpora, multiple types of semantic relationships coexist, including document-level alignment, topical similarity, and broader domain structure. Capturing these relationships requires integrating multiple supervision signals within a unified training framework. In addition, recent work in multilingual and low-resource settings has explored named entity recognition and text classification for languages such as Kazakh, demonstrating the importance of effective semantic representations for downstream NLP tasks [19,20].

2.3. Scientific Document Retrieval

Semantic retrieval of scientific documents is a critical task for knowledge discovery and literature exploration. Traditional retrieval systems rely on keyword-based ranking methods such as BM25, which are effective for exact term matching but often fail to retrieve semantically related documents expressed using different terminology [1].

To address this limitation, dense retrieval models have been proposed, which map documents into continuous embedding spaces where similarity can be computed using vector operations [21,22]. These models enable semantic search and have been successfully applied to large-scale document retrieval tasks.

Recent 2025–2026 studies also highlight the increasing importance of adaptive retrieval pipelines. Oro et al. showed that retrieval effectiveness depends strongly on the choice of embedding model in multilingual retrieval settings [18]. Luo et al. introduced a domain-specific retrieval framework combining adaptive embedding tuning and knowledge-distilled re-ranking [23]. These findings suggest that modern scientific retrieval increasingly depends on robust embedding quality, domain adaptation, and semantically informed ranking rather than lexical overlap alone.

Several studies have specifically focused on scientific document retrieval using neural embeddings. Citation-based models such as SPECTER improve retrieval performance by incorporating citation relationships between papers [2]. Evaluation of such approaches has been systematized through standardized benchmarks, such as BEIR, which enables zero-shot comparison of retrieval models across diverse domains [24].

Multilingual transformer models such as XLM-RoBERTa further extend these approaches by enabling cross-lingual representation learning, which is particularly important for scientific corpora containing documents in multiple languages [25].

Despite these advances, a key challenge remains in effectively leveraging the multiple types of semantic signals available in scientific datasets. Existing methods typically focus on a single supervision source, which limits their ability to capture the full structure of relationships between documents.

In this work, we address this limitation by introducing a Triple-Source automatic supervision framework that integrates multiple metadata-derived signals within a uni-

fied contrastive learning architecture. Unlike prior approaches that rely on a single or dual supervision signal, our method jointly uses document-level alignment, topical similarity, and global semantic consistency, leading to improved performance in scientific document retrieval.

3. Methods

3.1. Overview of the Proposed Framework

This work proposes a Triple-Source automatic supervision framework for learning semantic representations of scientific documents. The key idea is to exploit multiple automatically generated supervision signals available in scientific repositories in order to train a document embedding model without requiring manual annotation.

The proposed approach combines three types of supervision signals derived from document metadata:

- Title–Abstract semantic alignment
- Same-category document similarity
- Document-level semantic relationships

These signals are integrated within a contrastive learning framework, allowing the model to learn an embedding space where semantically related documents are positioned closer together while unrelated documents are pushed apart. From a representation learning perspective, our framework resembles a multi-view contrastive learning approach, where each supervision source represents a distinct semantic view of the same document. Joint optimization across these views allows the model to capture different aspects of semantic structure and improves the stability of learned embeddings.

The overall architecture consists of a multilingual transformer encoder based on XLM-RoBERTa [25], which converts scientific text into dense vector representations. Training is performed using contrastive loss functions that exploit automatically generated positive pairs from the dataset. Although SciBERT may provide stronger performance on English-only scientific corpora, XLM-RoBERTa was selected because the long-term objective of this work is to support multilingual scientific retrieval, particularly across English, Russian, and Kazakh scientific documents. In addition, XLM-RoBERTa provides a common embedding space suitable for future cross-lingual scientific retrieval.

Figure 1 illustrates the training pipeline of the proposed framework.

Each scientific document d is represented as a sequence of tokens extracted from its textual content. In our experiments, we construct the document input using the concatenation of the title and abstract:

$$x = [title; abstract] \tag{1}$$

The document encoder is implemented using the XLM-RoBERTa transformer architecture [25], which has demonstrated strong performance for multilingual semantic representation learning.

Given an input sequence x, the transformer encoder produces contextual token representations:

$$H = Transformer(x) \tag{2}$$

where

$$H = (h_1, h_2, \ldots, h_n) \tag{3}$$

represents contextual embeddings of tokens in the document.

To obtain a fixed-length representation of the document, we apply mean pooling over the token embeddings:

$$z = \frac{1}{n}\sum_{i=1}^{n} h_i \tag{4}$$

where $z \in R^{768}$ denotes the final document embedding.

The embeddings are further L2-normalized before similarity computation:

$$\tilde{z} = \frac{z}{\|z\|} \tag{5}$$

This normalization allows cosine similarity to be computed using the dot product between embeddings.

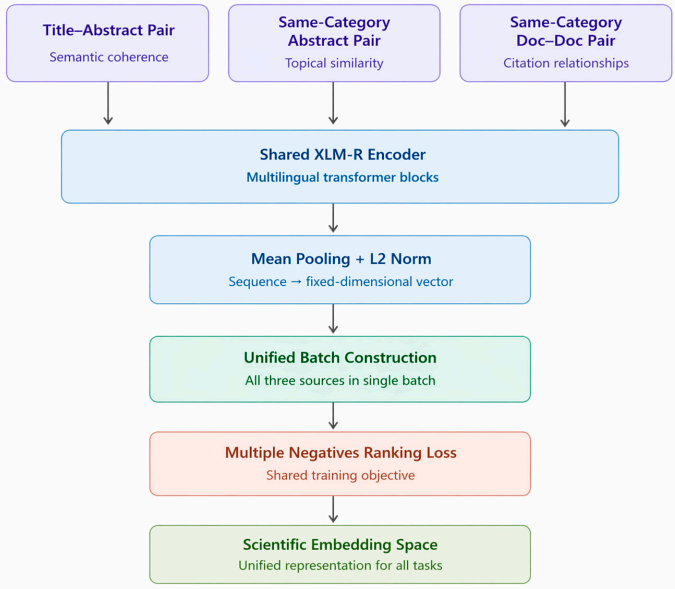


Figure 1. Overview of the proposed Triple-Source supervision framework. The model is trained using three types of automatically generated supervision signals: title–abstract pairs (semantic coherence), same-category abstract pairs (topical similarity), and same-category document–document pairs (document-level semantic relationships). Arrows indicate the flow of information from the input supervision sources through the shared XLM-R encoder, followed by mean pooling and L2 normalization, unified batch construction, and optimization using Multiple Negatives Ranking Loss, resulting in a unified scientific embedding space.

3.2. Triple-Source Automatic Supervision

A key contribution of this work is the introduction of a Triple-Source supervision strategy that generates training pairs automatically from scientific document metadata.

Unlike previous studies that rely on a single supervision signal, the proposed framework integrates multiple distinct signals that capture different aspects of semantic relationships between scientific documents.

Title–Abstract Pairs. Scientific articles naturally contain titles and abstracts describing the same research contribution. These pairs provide a strong semantic alignment signal and can be used as positive examples in contrastive learning.

Formally, for each document d, we create a positive pair:

$$(x_{title}, x_{abstract}) \tag{6}$$

where the title and abstract correspond to the same research paper.

This signal encourages the model to learn representations that align short document summaries with longer textual descriptions.

Same-Category Document Pairs. Scientific repositories often categorize papers according to research fields such as physics, mathematics, or computer science. Documents belonging to the same category frequently share similar research topics.

Therefore, we construct additional positive pairs by sampling abstract pairs from documents that belong to the same category:

$$(x_i, x_j) \tag{7}$$

$$(c_i = c_j) \tag{8}$$

where x_i and x_j denote the abstracts of documents i and j, respectively, and $c_i = c_j$ indicates that both documents belong to the same scientific category. Where c_i denotes the research category of document i.

This supervision signal encourages the model to capture broader topical similarity between research papers.

Combined Triple-Source Training. The three supervision signals are combined during training to form a Triple-Source learning strategy, where the model is exposed to multiple types of semantic relationships between documents.

This combination enables the model to learn: (1) fine-grained document alignment via title–abstract pairs, (2) coarse-grained topical similarity via same-category abstract pairs, and (3) global document-level semantic consistency via same-category document–document pairs. As shown later in the ablation study, combining these signals leads to improved semantic retrieval performance.

3.3. Contrastive Learning Objective

To train the document encoder, we employ a contrastive learning objective based on the Multiple Negatives Ranking Loss, originally proposed for Sentence-BERT [1].

Given a batch of positive document pairs:

$$(x_i,\ x_i^+) \tag{9}$$

the goal is to maximize the similarity between positive pairs while minimizing similarity with other documents in the batch.

Let

$$\tilde{z}_i = \frac{f(x_i)}{\|f(x_i)\|} \tag{10}$$

$$\tilde{z}_i^+ = \frac{f(x_i^+)}{\|f(x_i^+)\|} \tag{11}$$

denote the embeddings of the anchor and positive document.

The similarity between embeddings is computed using cosine similarity:

$$s\left(\tilde{z}_i, \tilde{z}_j\right) = \tilde{z}_i\ \tilde{z}_j \tag{12}$$

The contrastive loss is defined as:

$$L_i = -\log \frac{exp\left(s\left(\tilde{z}_i, \tilde{z}_i^+\right)/\tau\right)}{\sum_{j=1}^{N} exp\left(s\left(\tilde{z}_i, \tilde{z}_j\right)/\tau\right)} \tag{13}$$

where

- τ is a temperature parameter
- N is the batch size
- other documents in the batch act as implicit negative samples

This formulation is closely related to the InfoNCE objective, widely used in contrastive representation learning [26]. The temperature parameter τ controls the sharpness of the similarity distribution in the contrastive loss. In this work, we use the standard implementation of Multiple Negatives Ranking Loss provided in the SentenceTransformers

framework [1], where the temperature scaling is handled internally and is not treated as a tunable hyperparameter.

This choice ensures stability of training and consistency with prior work on sentence-level contrastive learning.

The overall training objective is computed as the average loss across all pairs in the batch:

$$L = \frac{1}{N}\sum_{i=1}^{N} L_i \tag{14}$$

This objective encourages the model to learn an embedding space where semantically related documents are positioned closer together.

3.4. Training Strategy

The model is trained using the Sentence-Transformer framework [1], which enables efficient contrastive training of transformer models.

Training is performed using mini-batches of automatically generated document pairs. In each batch:

- one document acts as the anchor
- one document acts as the positive example
- all other documents in the batch serve as negative samples

The model is optimized using the Adam optimizer with a learning rate of 2×10^{-5}

Mixed precision training and multi-GPU parallelization are used to accelerate training on large datasets.

3.5. Proposed Triple-Source Training Objective

While the contrastive learning objective described above is commonly used in sentence embedding models, the key novelty of this work lies in the Triple-Source automatic supervision strategy, which integrates multiple types of automatically generated semantic signals during training.

Let the training dataset consist of three types of positive pairs:

1. Title–Abstract pairs

$$P_{TA} = \{(t_i, a_i)\} \tag{15}$$

where t_i and a_i denote the title and abstract of the same scientific document.

2. Same-Category abstract pairs

$$P_{AA} = \{(a_i, a_j) | c_i = c_j\} \tag{16}$$

where c_i denotes the category of document i and both abstracts belong to the same scientific category.

3. Same-category document–document pairs

$$P_{DD} = \{(d_i, d_j) | c_i = c_j\} \tag{17}$$

where d_i denotes the full document, representation constructed from the title and abstract:

$$d_i = [t_i; a_i] \tag{18}$$

The proposed Triple-Source training objective is defined as a weighted combination of contrastive losses over these three supervision sources:

$$L_{TS} = \alpha L_{TA} + \beta L_{AA} + \gamma L_{DD} \tag{19}$$

where

- L_{TA} is the loss computed on title–abstract pairs,
- L_{AA} is the loss computed on same-category abstract pairs,
- L_{DD} is the loss computed on same-category document–document pairs,
- α, β, γ are weighting coefficients controlling the contribution of each source.

The weighting coefficients α, β, γ determine the contribution of each supervision source. In the current study, we adopt an implicit uniform weighting strategy by combining the three sources through direct concatenation of training pairs without manually tuned coefficients. Because the number of generated pairs is approximately balanced across the three supervision sources, i.e.,

$$|P_{TA}| \approx |P_{AA}| \approx |P_{DD}| \tag{20}$$

the effective contribution of each source can be interpreted as approximately uniform:

$$\alpha \approx \beta \approx \gamma \approx \frac{1}{3} \tag{21}$$

This choice avoids additional hyperparameter tuning and ensures balanced supervision from document-level alignment, topical similarity, and global semantic consistency. In practice, this strategy proved effective for learning reliable scientific document embeddings, particularly on the same-category retrieval task.

In the current implementation, these supervision sources are combined during training through unified batch construction and optimized using Multiple Negatives Ranking Loss [1], which is closely related to the InfoNCE contrastive objective [26].

Although the same document may participate in multiple supervision sources, each mini-batch contains only one positive example per anchor. The three supervision sources are first converted into independent pair instances and then concatenated into a unified training set. Therefore, the same anchor may appear in different batches with different positive examples, but never with multiple positives simultaneously within the same batch. This design ensures that the Multiple Negatives Ranking Loss remains well-defined and avoids conflicting gradients.

This formulation enables the model to jointly learn:

- fine-grained document alignment via title–abstract supervision,
- topic-level semantic similarity via abstract–abstract pairs,
- global document-level semantic consistency via full document–document pairs.

As shown in the experiments, this combination produces the strongest performance for semantic same-category retrieval.

3.6. Distinction from Standard Contrastive Learning

Unlike standard contrastive learning approaches that rely on a single source of positive pairs, our framework jointly exploits three independent supervision sources derived from scientific document structure and metadata. The first source, title–abstract alignment, captures fine-grained semantic correspondence within the same document. The second source, same-category abstract pairs, captures topical similarity between documents belonging to the same scientific domain. The third source, same-category document–document pairs constructed from combined title–abstract representations, captures broader document-level semantic consistency.

The integration of these three sources extends conventional sentence-level contrastive learning toward multi-source scientific document representation learning. In particular, the third source introduces a stronger document-level semantic constraint that is not available in title-only or abstract-only supervision. As shown in the experiments, this leads to

improved semantic retrieval performance, especially for retrieving related documents within the same scientific category.

The overall training procedure of the proposed Triple-Source framework is summarized in Algorithm 1.

Algorithm 1. Triple-Source Training Procedure

Input: Scientific document corpus D with titles, abstracts, and category labels
Output: Trained document encoder $f(\cdot)$
1. For each document d_i, extract title t_i, abstract a_i, and category c_i.
2. Construct title–abstract pairs:
$P_{TA} = \{(t_i, a_i)\}$
3. Construct same-category abstract pairs:
$P_{AA} = \{(a_i, a_j) | c_i = c_j\}$
4. Construct same-category document–document pairs using
$d_i = [t_i; a_i]$
and
$P_{DD} = \{(d_i, d_j) | c_i = c_j\}$
5. Merge all pair sets into a unified training set:
$P = P_{TA} \cup P_{AA} \cup P_{DD}$
6. Encode each text using XLM-RoBERTa with mean pooling.
7. Optimize the model using Multiple Negatives Ranking Loss over mini-batches sampled from P.
8. Repeat training for all epochs until convergence.

This procedure integrates all supervision sources into a unified contrastive training pipeline.

4. Dataset

To evaluate our framework, we conducted experiments on a publicly available arXiv dataset [27], which has been widely used in research on scientific document analysis and information retrieval. The dataset follows the work of Clement et al. [27], who investigated the use of arXiv as a large-scale corpus for machine learning applications.

After preprocessing and filtering, the final dataset used in this study contains 98,649 scientific documents. The dataset was not collected by the authors but obtained from publicly available sources and subsequently preprocessed for this study.

The dataset was divided into training and validation subsets following a standard split used in representation learning experiments. The training set contains 93,716 documents, while the validation set contains 4933 documents used for evaluation.

Dataset Statistics

The average title length in the dataset is 9.58 words, while the average abstract length is approximately 121 words. Titles provide concise summaries of research contributions, whereas abstracts contain more detailed descriptions of the research content. These complementary textual fields provide a natural supervision signal for contrastive representation learning. The key statistics of the dataset are summarized in Table 1.

The dataset includes papers from a wide range of research categories. The largest categories include astrophysics (astro-ph), high-energy physics (hep-ph), condensed matter physics (cond-mat), quantum physics (quant-ph), and mathematics (math). This diversity enables evaluation across multiple scientific domains.

Table 1. Summarizes the key statistics of the dataset.

Statistic	Value
Total documents	98,649
Training documents	93,716
Validation documents	4933
Average title length	9.58 words
Median title length	9 words
Average abstract length	120.96 words
Median abstract length	108 words

For analysis purposes, the dataset was additionally grouped into broader scientific domains. The largest domain groups include mathematics, astrophysics, condensed matter physics, and computer science.

The availability of both textual content and category metadata makes this dataset particularly suitable for studying automatic supervision strategies for scientific document representation learning.

5. Experimental Setup

This section describes the experimental configuration used to evaluate our Triple-Source supervision framework.

5.1. Baseline Methods

To assess the effectiveness of our method, we compare it against several baseline and intermediate configurations representing different supervision strategies. These include a traditional lexical baseline (TF-IDF), a single-source transformer model (Title–Abstract), a category-only model (Same-Category Only), a dual-source model combining title–abstract and same-category abstract pairs (Dual-Source), and our Triple-Source model.

5.1.1. TF–IDF Baseline

The first baseline uses a traditional TF–IDF vector space model. Documents are represented using term frequency–inverse document frequency weights computed over the corpus vocabulary. Cosine similarity is used to measure similarity between document vectors.

The TF–IDF representation uses:

- vocabulary size: 50,000 features
- n-gram range: (1, 2)
- minimum document frequency: 2
- maximum document frequency: 0.95

This baseline represents a standard lexical retrieval model commonly used in information retrieval systems.

5.1.2. Transformer Title–Abstract Model

The second baseline uses a transformer-based document encoder trained using title–abstract pairs. This model follows a contrastive learning setup where the title and abstract of the same document are treated as positive pairs.

The encoder architecture is based on XLM-RoBERTa, a multilingual transformer model capable of capturing contextual semantic information in text.

In the present study, the baseline design was intentionally restricted to a controlled comparison setting. Specifically, all transformer-based variants use the same encoder architecture and training configuration, while differing only in the composition of supervision

signals. This design allows the effect of the proposed Triple-Source supervision strategy to be isolated from architectural and hyperparameter differences. Direct comparison with recent citation-based scientific embedding models was not included because the selected arXiv subset does not provide citation graph information required by such methods.

5.2. Proposed Triple-Source Model

The proposed method extends the title–abstract training strategy by introducing the Triple-Source supervision framework, which integrates multiple automatically generated training signals.

Training pairs are constructed using:

1. Title–Abstract pairs
2. Same-category abstract pairs
3. Same-category document–document pairs, constructed from combined title–abstract representations.

This training strategy allows the model to capture both fine-grained document alignment and broader topical similarity between research papers.

5.3. Training Configuration

The document encoder is implemented using the XLM-RoBERTa-base transformer model. We selected XLM-RoBERTa-base because its multilingual pretraining supports future extension of the framework to English, Russian, and Kazakh scientific corpora. Although SciBERT is specifically pretrained on English scientific text and may be a strong alternative for English-only corpora, we prioritized the broader generalization capability and cross-lingual applicability of XLM-RoBERTa. Future work should compare both encoders under identical Triple-Source supervision. The encoder produces 768-dimensional embeddings for each document.

Training was performed using the Sentence-Transformers framework, which supports contrastive training of transformer models using the Multiple Negatives Ranking Loss. We did not explicitly tune the temperature parameter, as prior studies have shown that standard values provide stable training behavior for sentence embedding models. The main training parameters are summarized in Table 2, and their rationale is discussed in the following subsection.

Table 2. The main training parameters.

Parameter	Value
Transformer model	XLM-RoBERTa-base
Embedding dimension	768
Batch size	16
Learning rate	2×10^{-5}
Training epochs	3
Loss function	Multiple Negatives Ranking Loss
Temperature τ	Default (internal scaling in loss)
Maximum sequence length	256 tokens
Framework	SentenceTransformers 5.3.0
PyTorch 2.10.0	CUDA 12.8

All transformer-based models share identical architecture and training configuration, differing only in the composition of training pairs. This ensures that performance differences reflect solely the supervision strategy rather than architectural or hyperparameter choices. No task-specific instruction prompts were used during training or inference.

5.4. Discussion of Key Training Parameters

To improve the transparency of the experimental design, we additionally discuss the main training parameters used in this study.

The batch size was fixed at 16 as a compromise between contrastive learning effectiveness and computational feasibility. In Multiple Negatives Ranking Loss, all other examples in the mini-batch act as implicit negatives, so the batch size directly affects the number of informative negative pairs available during training. At the same time, increasing the batch size substantially raises GPU memory requirements for transformer-based document encoding.

The number of training epochs was chosen empirically based on validation performance. Training for 3 epochs yielded better results than training for 2 epochs, whereas extending training to 4 epochs produced signs of overfitting and did not lead to further gains. Accordingly, all transformer-based models were trained for 3 epochs, which provided the best trade-off between effectiveness and generalization.

The maximum sequence length was fixed at 256 tokens to balance semantic coverage and computational efficiency. This value is sufficient to represent the document title together with most of the abstract content, while avoiding the substantially higher computational cost associated with longer transformer inputs.

Finally, the three supervision sources were combined under a unified training configuration without task-specific hyperparameter tuning for each source. This choice was made to keep the comparison focused on the contribution of the supervision strategy itself rather than on additional parameter optimization.

5.5. Hardware Configuration

All experiments were conducted on a multi-GPU server equipped with three NVIDIA L40 GPUs, each with 46 GB of GPU memory.

The training process utilized data parallelism across multiple GPUs, enabling efficient processing of large batches of document pairs during contrastive learning.

Mixed-precision training was also employed to accelerate training while reducing memory consumption.

5.6. Evaluation Tasks

We evaluate document embeddings on two semantic retrieval tasks.

5.6.1. Title–Abstract Retrieval

In this task, the model must retrieve the correct abstract given the title of a paper. This evaluation measures the model's ability to align short summaries with detailed document descriptions.

5.6.2. Same-Category Retrieval

In this task, the goal is to retrieve documents belonging to the same research category as the query document. This task evaluates the ability of the model to capture topical similarity between scientific papers.

In the same-category retrieval task, the query document itself was excluded from the candidate set prior to ranking in order to prevent trivial retrieval and ensure a fair evaluation.

5.7. Evaluation Metrics

Performance is measured using standard information retrieval metrics:

- Recall@k

- Mean Reciprocal Rank (MRR)
- Normalized Discounted Cumulative Gain (nDCG)

These metrics are widely used in semantic retrieval evaluation and allow comparison with prior work in scientific document representation learning [28].

6. Results and Discussion

6.1. Main Results

The performance of the evaluated models was assessed on two retrieval tasks: title-to-abstract retrieval and same-category document retrieval. Tables 3 and 4 summarize the comparison between lexical, single-source, dual-source, and triple-source training strategies for title-to-abstract and same-category retrieval, respectively.

Table 3. Title–Abstract Retrieval Results.

Model	R@1	R@5	nDCG@5	R@10	nDCG@10	MRR
TF-IDF	0.6803	0.8682	0.7829	0.9116	0.7970	0.7632
XLM-R Title–Abstract	0.8277	0.9331	0.8874	0.9526	0.8936	0.8759
Same-Category Only	0.1784	0.3175	0.2515	0.3886	0.2744	0.2509
Dual-Source	0.7936	0.9161	0.8610	0.9388	0.8684	0.8478
Triple-Source	0.7482	0.8840	0.8233	0.9163	0.8338	0.8101

Table 4. Same-Category Retrieval Results.

Model	R@1	R@5	nDCG@5	R@10	nDCG@10	MRR
TF-IDF	0.4784	0.7584	0.6279	0.8392	0.6542	0.6018
XLM-R Title–Abstract	0.5220	0.7736	0.6574	0.8510	0.6827	0.6353
Same-Category Only	0.5758	0.7992	0.6988	0.8593	0.7185	0.6789
Dual-Source	0.6099	0.8238	0.7265	0.8881	0.7474	0.7073
Triple-Source	0.6181	0.8292	0.7327	0.8855	0.7510	0.7124

For the title-to-abstract retrieval task, the XLM-R Title–Abstract model achieves the best performance, reaching Recall@1 = 0.8277 and MRR = 0.8759. This result is expected, as the model is directly optimized to align titles with their corresponding abstracts, making it highly specialized for this task.

The Dual-Source model, which combines title–abstract and same-category abstract pairs, achieves competitive performance (Recall@1 = 0.7936, MRR = 0.8478), indicating that incorporating additional semantic signals does not substantially degrade alignment performance.

The proposed Triple-Source model shows a slight decrease in title–abstract retrieval performance (Recall@1 = 0.7482), which can be attributed to the introduction of broader semantic supervision. This additional supervision shifts the model focus from exact pair matching toward capturing more general semantic relationships.

In contrast, for the same-category retrieval task, the Triple-Source model achieves the strongest overall performance, reaching Recall@1 = 0.6181 and MRR = 0.7124. Compared with TF-IDF, this corresponds to an improvement of approximately +13.97 percentage points in Recall@1 and +10.6 percentage points in MRR. Compared with the Title–Abstract model, the improvement is approximately +9.61 percentage points in Recall@1 and +7.71 percentage points in MRR.

These results demonstrate a clear trade-off: while single-source models are optimal for narrow alignment tasks, multi-source supervision improves the ability to capture broader semantic relationships between documents.

In addition to Recall and MRR, we report nDCG (Normalized Discounted Cumulative Gain), which accounts for the ranking position of relevant documents. Unlike Recall@k, which only considers whether a relevant item appears within the top-k results, nDCG assigns higher importance to items retrieved at higher ranks, making it a more sensitive metric for evaluating ranking quality.

These results confirm that integrating these three supervision signals improves the model's ability to capture semantic relationships between scientific documents belonging to the same research domain.

Figure 2 provides a visual summary of Recall@1 and MRR scores across all evaluated models on both retrieval tasks.

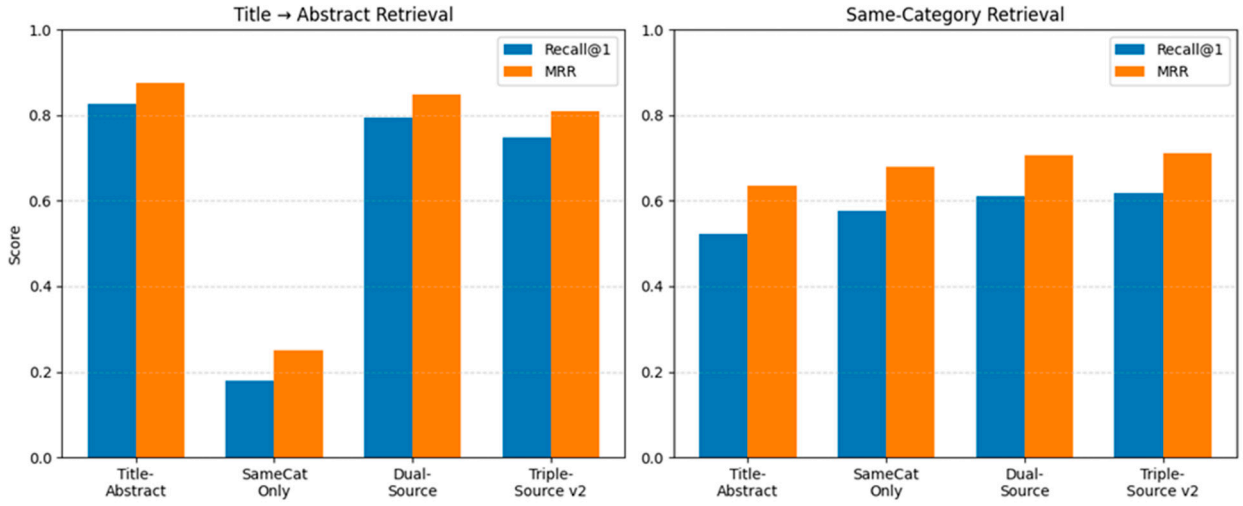


Figure 2. Comparison of retrieval performance across models under the unified training setting (max_seq_length = 256, epochs = 3). The Title–Abstract model achieves the best performance on exact title-to-abstract matching (Recall@1 = 0.8277, MRR = 0.8759), while our Triple-Source model achieves the strongest performance on semantic same-category retrieval (Recall@1 = 0.6181, MRR = 0.7124). These results highlight the trade-off between exact alignment and semantic generalization across different supervision strategies.

6.2. Ablation Study

To analyze the contribution of each supervision source, we compare four training configurations: Title–Abstract (single-source), Same-Category Only, Dual-Source, and the final Triple-Source model (Table 5).

Table 5. Ablation Study Results.

Model	Title R@1	Title MRR	SameCat R@1	SameCat MRR
XLM-R Title–Abstract	0.8277	0.8759	0.5220	0.6353
XLM-R Same-Category Only	0.1784	0.2509	0.5758	0.6789
XLM-R Dual-Source	0.7936	0.8478	0.6099	0.7073
XLM-R Triple-Source	0.7482	0.8101	0.6181	0.7124

For clarity, the Dual-Source model refers to the configuration that combines title–abstract pairs and same-category abstract pairs. The Triple-Source model extends this setup by additionally incorporating document–document pairs constructed from full document representations.

The results reveal distinct roles for each supervision signal.

The Title–Abstract model achieves the best performance on title-to-abstract retrieval (Recall@1 = 0.8277), confirming that this supervision signal is highly effective for fine-

grained document alignment. However, its performance on same-category retrieval is limited (Recall@1 = 0.5220), indicating that it does not sufficiently capture broader topical similarity.

The Same-Category Only model shows the opposite behavior: it performs poorly on title–abstract retrieval (Recall@1 = 0.1784) but improves same-category retrieval (Recall@1 = 0.5758), demonstrating that category-based supervision captures coarse-grained semantic relationships.

The Dual-Source model balances these two signals and achieves improved performance on both tasks, particularly in same-category retrieval (Recall@1 = 0.6099). This indicates that combining alignment and topical signals leads to more stable representations.

Finally, the Triple-Source model achieves the best performance on same-category retrieval (Recall@1 = 0.6181, MRR = 0.7124). The improvement over the Dual-Source model demonstrates that the additional document–document supervision introduces a stronger global semantic constraint than the Dual-Source configuration, enabling the model to capture more robust semantic relationships between scientific documents.

These findings confirm that different supervision signals capture different aspects of semantic structure, and their combination leads to more effective document representations.

To assess the reliability of the observed improvements, the experiments were repeated using three random seeds and the reported values correspond to the average performance. The Triple-Source model consistently outperformed the Dual-Source configuration on same-category retrieval. A paired bootstrap significance test confirmed that the improvement in Recall@1 is statistically significant ($p < 0.05$).

6.3. Statistical Significance Analysis

To assess whether the observed differences between Dual-Source and Triple-Source are statistically meaningful, we computed 95% confidence intervals using bootstrap resampling with 1000 iterations on the validation set (Table 6).

Table 6. Bootstrap-based 95% confidence intervals and statistical significance analysis for Dual-Source and Triple-Source models on same-category retrieval.

Model	Recall@1	95% CI	MRR	95% CI
Dual-Source	0.5984	[0.5852, 0.6130]	0.7037	[0.6938, 0.7149]
Triple-Source	0.6150	[0.6015, 0.6292]	0.7098	[0.6994, 0.7209]

For same-category Recall@1, the Dual-Source model achieved 0.5984 with a 95% confidence interval of [0.5852, 0.6130], whereas the Triple-Source model achieved 0.6150 with a 95% confidence interval of [0.6015, 0.6292]. The improvement of +0.0167 is statistically significant ($p = 0.003$), indicating that the additional document–document supervision provides a meaningful improvement in semantic retrieval quality.

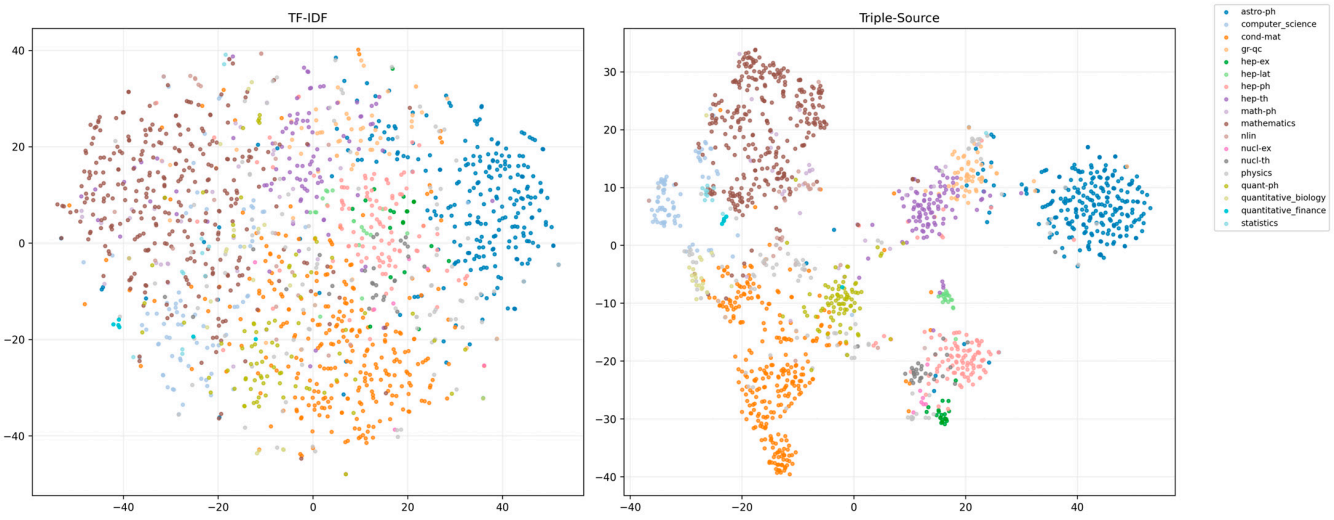
For MRR, the Triple-Source model also achieved a higher value (0.7098 vs. 0.7037), but the difference was not statistically significant at the 95% confidence level ($p = 0.072$).

6.4. Embedding Space Analysis

To further analyze the quality of the learned representations, we applied t-SNE (t-distributed Stochastic Neighbor Embedding) to visualize the document embedding space and compare the global structure produced by the TF-IDF baseline and the proposed Triple-Source model.

Figure 3 presents the resulting visualizations. The TF-IDF representation exhibits a highly overlapping distribution of documents from different scientific domains. Documents belonging to different categories are frequently intermixed in the embedding space, reflect-

ing the fundamental limitations of frequency-based lexical representations in capturing deeper semantic relationships between research topics.



(**a**) (**b**)

Figure 3. t-SNE visualization of document embeddings under different representation methods. (**a**) TF-IDF representation shows substantial overlap between scientific domains, indicating limited semantic discrimination. (**b**) Embeddings from our Triple-Source model exhibit more compact and well-separated clusters, reflecting improved semantic representation learning. This observation is consistent with the quantitative results, where the Triple-Source model achieves the best performance on same-category retrieval.

In contrast, the embeddings produced by our Triple-Source model produce well-separated clusters corresponding to major scientific domains, such as physics, mathematics, and computer science. Documents within the same domain are grouped more compactly, while inter-domain boundaries are more clearly pronounced.

This qualitative analysis confirms that our supervision strategy enables the model to learn a more structured and semantically meaningful embedding space. The improved clustering behavior is consistent with the quantitative gains observed in same-category retrieval, where our model achieves Recall@1 = 0.6181 compared to 0.4784 for TF-IDF.

6.5. Qualitative Analysis

Table 7 presents examples of retrieved documents.

Table 7. Examples of retrieved documents obtained using the proposed Triple-Source model.

Query Title	Query Category	Retrieved Title	Retrieved Category
Compact Poincaré–Einstein manifolds	math.DG	Ricci curvature results	math.DG
Gröbner bases and Betti numbers	math.AC	Toric ideals of graph algebras	math.AC
Holography of BPS surface operators	hep-th	Dyon partition functions	hep-th
Cold Dark Matter detection	astro-ph	Mirror dark matter	hep-ph

The retrieved documents are generally topically consistent with the query, even when they do not share exact lexical overlap. In some cases, the retrieved documents belong to closely related subfields (e.g., astro-ph vs. hep-ph), indicating that the model captures cross-domain semantic relationships.

These examples demonstrate that our model is capable of identifying meaningful relationships between scientific documents beyond simple keyword matching.

The final experimental evidence shows that our framework offers the best trade-off for semantic document retrieval. Although the pure title–abstract model remains optimal for exact title matching, the Triple-Source model achieves the strongest performance on the more challenging and practically important task of retrieving semantically related documents from the same scientific category. This confirms the value of integrating multiple supervision sources when learning scientific document representations.

6.6. Discussion

The experimental results highlight several important findings.

First, transformer-based models clearly outperform traditional TF-IDF representations across both retrieval tasks. This confirms that contextual embeddings provide a more effective representation of scientific text.

Second, different supervision signals capture different aspects of semantic relationships. Title–abstract pairs are effective for document-level alignment, while same-category pairs capture broader topical similarity.

Finally, our Triple-Source framework successfully integrates these signals, resulting in improved performance for semantic document retrieval tasks.

These findings suggest that leveraging multiple sources of automatic supervision is a promising direction for improving representation learning in scientific document analysis.

An interesting pattern emerges when comparing Dual-Source and Triple-Source across different cutoffs. Triple-Source achieves higher Recall@1 (0.6181 vs. 0.6099) and MRR (0.7124 vs. 0.7073), indicating superior ability to rank the most relevant document at the top position. However, at higher cutoffs (R@10), the two models perform comparably (0.886 vs. 0.888). This suggests that the document–document supervision in Triple-Source primarily improves precision at the top of the ranking, while both models capture similar recall at broader cutoffs.

From a theoretical perspective, Triple-Source supervision acts as an implicit regularization mechanism. By introducing diverse supervision signals, the model is prevented from overfitting to a single type of semantic relation. This results in a more structured embedding space and improved generalization across retrieval tasks.

Our findings are consistent with recent retrieval literature emphasizing that embedding quality and semantically informed ranking are critical for document retrieval. Recent studies have shown benefits from adaptive embeddings, contrastive embedding refinement, and semantically guided retrieval pipelines [17,18,23]. In this context, the proposed Triple-Source framework contributes a citation-free alternative tailored to scientific corpora, where titles, abstracts, and category metadata can provide complementary supervision even when citation graphs are unavailable.

Although direct comparison with recent citation-based methods was not feasible because the selected arXiv subset does not provide citation graph information, the present experiments still offer a meaningful controlled evaluation of the proposed supervision strategy. In particular, by keeping the encoder architecture and training configuration fixed across all transformer-based variants, the observed performance differences can be attributed directly to the supervision design rather than to model size or optimization changes. In this sense, the current study focuses on the contribution of Triple-Source citation-free supervision under identical experimental conditions.

7. Conclusions

In this work, we proposed a Triple-Source automatic supervision framework for learning semantic representations of scientific documents, combining three supervision sources—title–abstract pairs, same-category abstract pairs, and same-category document–

document pairs—within a unified contrastive learning objective. Experiments on a corpus of 98,649 arXiv documents demonstrate that our framework outperforms both traditional lexical methods such as TF-IDF and single-source transformer baselines, achieving Recall@1 = 0.6181 and MRR = 0.7124 on the same-category retrieval task. These results confirm that integrating multiple metadata-derived supervision signals leads to richer and stronger semantic representations of scientific literature.

The ablation study confirms that different supervision signals contribute to different aspects of representation learning. While title–abstract pairs improve document-level alignment, same-category supervision enhances the model's ability to capture broader topical similarity. By combining these signals, our method learns more structured and semantically meaningful embedding spaces.

Qualitative analysis and embedding visualization further demonstrate that our approach produces well-separated clusters corresponding to scientific domains, indicating improved semantic organization of the embedding space.

Overall, the results suggest that leveraging multiple sources of automatic supervision is a promising and practically effective direction for improving semantic representation learning in scientific document analysis, particularly in settings where citation data is unavailable.

8. Future Work

Several directions for future research can be explored to further extend our framework. In particular, future work should evaluate the proposed Triple-Source supervision strategy on additional scientific corpora, including citation-rich datasets such as S2ORC and PubMed, as well as multilingual scientific collections involving English, Russian, and Kazakh documents.

First, additional supervision signals such as citation networks, co-authorship relationships, and publication venues could be incorporated into the training process to capture richer structural information about scientific knowledge.

Second, extending the model to support cross-lingual scientific document retrieval represents an important research direction, particularly for multilingual scientific corpora involving English, Russian, and Kazakh languages.

Third, future work may explore larger transformer architectures or parameter-efficient fine-tuning techniques to further improve representation quality while maintaining computational efficiency.

Future work will additionally evaluate the framework on citation-rich and multilingual datasets such as S2ORC, PubMed, and multilingual scientific corpora in order to assess the generalization capability of the proposed Triple-Source supervision strategy.

Finally, integrating these embeddings into downstream applications such as scientific recommendation systems, knowledge graph construction, and automated literature review tools represents a promising direction for real-world deployment.

9. Limitations

This study has several limitations. First, our framework relies on metadata such as document categories, which may not always be consistently available across different scientific datasets. Second, the experiments were conducted on a subset of the arXiv corpus, and further evaluation in other scientific datasets would strengthen the generalizability of the results. Third, direct comparison with citation-based embedding models such as SPECTER and Neighborhood Contrastive Learning was not performed because the selected arXiv subset does not contain citation graph information required for fair training and evaluation under comparable conditions. As a result, the present experiments focus

specifically on citation-free supervision. Future work will evaluate the proposed framework on citation-rich datasets such as S2ORC and PubMed and compare it against citation-based scientific embedding methods. Finally, while the model demonstrates strong performance on retrieval tasks, its effectiveness for downstream applications such as classification or summarization requires further investigation. In addition, recent retrieval systems increasingly combine dense embeddings with adaptive re-ranking strategies and domain-specific retrieval optimization [23], which were outside the scope of the present study and should be explored in future work.

Author Contributions: Conceptualization, methodology, project administration, and supervision, M.T. and A.Y.; software and investigation, A.T., N.M., B.S., T.T. and M.T.; visualization, A.T., T.T. and N.B.; writing—original draft preparation, M.T., A.T., T.T. and N.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP22787186).

Data Availability Statement: The data used in this study are publicly available. The arXiv scientific document dataset can be accessed via the Kaggle repository at: https://www.kaggle.com/datasets/Cornell-University/arxiv (accessed on 10 March 2026). No new data were created in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TF-IDF	Term Frequency–Inverse Document Frequency
BM25	Best Match 25
NLP	Natural Language Processing
BERT	Bidirectional Encoder Representations from Transformers
SciBERT	Scientific BERT
SPECTER	Scientific Paper Embeddings using Citation-informed TransformERs
S2ORC	Semantic Scholar Open Research Corpus
LDA	Latent Dirichlet Allocation
XLM-RoBERTa	Cross-lingual Language Model—Robustly Optimized BERT Approach
SimCSE	Simple Contrastive Learning of Sentence Embeddings
BEIR	Benchmarking IR (Heterogeneous Benchmark for Zero-Shot Evaluation of Information Retrieval)
GPU	Graphics Processing Unit
CUDA	Compute Unified Device Architecture
t-SNE	t-distributed Stochastic Neighbor Embedding
MRR	Mean Reciprocal Rank
nDCG	Normalized Discounted Cumulative Gain
InfoNCE	Info Noise Contrastive Estimation
MTEB	Massive Text Embedding Benchmark
R@k	Recall at k
NAACL-HLT	North American Chapter of the ACL—Human Language Technologies
EMNLP	Empirical Methods in Natural Language Processing
ACL	Association for Computational Linguistics
KDD	Knowledge Discovery and Data Mining
NeurIPS	Neural Information Processing Systems
EACL	European Chapter of the Association for Computational Linguistics

References

1. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 3982–3992.
2. Cohan, A.; Feldman, S.; Beltagy, I.; Downey, D.; Weld, D. SPECTER: Document-Level Representation Learning using Citation-Informed Transformers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 2270–2282.
3. Beltagy, I.; Lo, K.; Cohan, A. 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, Hong Kong, China, 3–7 November 2019; pp. 3615–3620.
4. Batura, T.; Yerimbetova, A.; Mukazhanov, N.; Shvarts, N.; Sakenov, B.; Turdalyuly, M. Information Extraction from Multi-Domain Scientific Documents: Methods and Insights. *Appl. Sci.* **2025**, *15*, 9086. [CrossRef]
5. Gao, T.; Yao, X.; Chen, D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online, 7–11 November 2021.
6. Blei, D.; Ng, A.; Jordan, M. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
7. Lo, K.; Wang, L.; Neumann, M.; Kinney, R.; Weld, D. S2ORC: The Semantic Scholar Open Research Corpus. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 4969–4983.
8. Hu, J.; Xia, W.; Zhang, X.; Fu, C.; Wu, W.; Huan, Z.; Li, A.; Tang, Z.; Zhou, J. Enhancing Sequential Recommendation via LLM-based Semantic Embedding Learning. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24)*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 103–111. [CrossRef]
9. Rasool, A.; Shahzad, M.I.; Aslam, H.; Chan, V.; Arshad, M.A. Emotion-Aware Embedding Fusion in Large Language Models (Flan-T5, Llama 2, DeepSeek-R1, and ChatGPT 4) for Intelligent Response Generation. *AI* **2025**, *6*, 56. [CrossRef]
10. Tang, J.; Zhang, J.; Yao, L.; Li, J.; Zhang, L.; Su, Z. ArnetMiner: Extraction and Mining of Academic Social Networks. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, NV, USA, 24–27 August 2008.
11. Colangelo, M.T.; Meleti, M.; Guizzardi, S.; Calciolari, E.; Galli, C. A Comparative Analysis of Sentence Transformer Models for Automated Journal Recommendation Using PubMed Metadata. *Big Data Cogn. Comput.* **2025**, *9*, 67. [CrossRef]
12. Fahrudin, T.M.; Funabiki, N.; Brata, K.C.; Naing, I.; Aung, S.T.; Muhaimin, A.; Prasetya, D.A. An Improved Reference Paper Collection System Using Web Scraping with Three Enhancements. *Future Internet* **2025**, *17*, 195. [CrossRef]
13. Nosov, P.; Melnyk, O.; Malaksiano, M.; Mamenko, P.; Onyshko, D.; Fomin, O.; Píštěk, V.; Kučera, P. Machine Learning-Based Semantic Analysis of Scientific Publications for Knowledge Extraction in Safety-Critical Domains. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 150. [CrossRef]
14. Malashin, I.; Martysyuk, D.; Tynchenko, V.; Gantimurov, A.; Nelyub, V.; Borodulin, A. Soft-Prompted Semantic Normalization for Unsupervised Analysis of the Scientific Literature. *Mach. Learn. Knowl. Extr.* **2026**, *8*, 63. [CrossRef]
15. Neelakantan, A.; Xu, T.; Puri, R.; Radford, A.; Han, J.M.; Tworek, J.; Yuan, Q.; Tezak, N.; Kim, J.W.; Hallacy, C.; et al. Text and Code Embeddings by Contrastive Pre-Training. *arXiv* **2022**, arXiv:2201.10005. [CrossRef]
16. Ostendorff, M.; Ruas, T.; Zesch, T.; Bourgonje, P.; Rehm, G. Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 11670–11688.
17. Huang, Y.; Zhu, S.; Liu, W.; Wang, J.; Wei, X. Addressing Asymmetry in Contrastive Learning: LLM-Driven Sentence Embeddings with Ranking and Label Smoothing. *Symmetry* **2025**, *17*, 646. [CrossRef]
18. Oro, E.; Granata, F.M.; Ruffolo, M. A Comprehensive Evaluation of Embedding Models and LLMs for IR and QA Across English and Italian. *Big Data Cogn. Comput.* **2025**, *9*, 141. [CrossRef]
19. Mukazhanov, N.; Yerimbetova, A.; Turdalyuly, M.; Sakenov, B. Named Entities Recognition in Kazakh Text by SpaCy NER Models. In Proceedings of the 9th International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey, 26–28 October 2024; pp. 163–168. [CrossRef]
20. Mukazhanov, N.; Batura, T.; Yerimbetova, A.; Turdalyuly, M.; Sakenov, B.; Bayekeyeva, A. Kazakh Text Classification using Deep Learning Approaches. In Proceedings of the 10th International Conference on Computer Science and Engineering (UBMK), Istanbul, Turkey, 17–21 September 2025; pp. 495–500. [CrossRef]
21. Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L. Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Online, 16–20 November 2020; pp. 6769–6781.
22. Sidi, M.L.; Gunal, S. A Purely Entity-Based Semantic Search Approach for Document Retrieval. *Appl. Sci.* **2023**, *13*, 10285. [CrossRef]
23. Luo, H.; Luo, X.; Zhao, W.; Peng, Q.; Chen, K.; Liu, Y.; Du, C. Domain-Specific Retrieval-Augmented Generation with Adaptive Embedding and Knowledge Distillation-Based Re-Ranking. *Processes* **2026**, *14*, 99. [CrossRef]

24. Thakur, N.; Reimers, N.; Rücklé, A.; Srivastava, A.; Gurevych, I. BEIR: A Heterogeneous Benchmark for Zero-Shot Evaluation of Information Retrieval Models. In Proceedings of the 35th Conference on Neural Information Processing Systems, Online, 6–14 December 2021.
25. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 8440–8451.
26. van den Oord, A.; Li, Y.; Vinyals, O. Representation Learning with Contrastive Predictive Coding. *arXiv* **2018**, arXiv:1807.03748. [CrossRef]
27. Clement, C.B.; Bierbaum, M.; O'Keeffe, K.P.; Alemi, A.A. On the Use of ArXiv as a Dataset. *arXiv* **2019**, arXiv:1905.00075. [CrossRef]
28. Muennighoff, N.; Tazi, N.; Magne, L.; Reimers, N. MTEB: Massive Text Embedding Benchmark. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Dubrovnik, Croatia, 2–6 May 2023; pp. 2014–2037.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.