Expert Systems With Applications 298 (2026) 129794

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A hybrid graph and LLM approach for measuring scientific novelty via knowledge recombination and propagation[⋆]

Zhongyi Wang[a], Zeren Wang[a], Guangzhao Zhang[a], Jiangping Chen[b], Markus Luczak-Roesch[c], Haihua Chen[d,e,*]

[a] School of Information Management, Central China Normal University, Wuhan, 430079, Hubei, China
[b] School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, 61820, Illinois, USA
[c] School of Information Management, Victoria University of Wellington, Wellington, 6140, Wellington, New Zealand
[d] Anuradha and Vikas Sinha Department of Data Science, University of North Texas, Denton, 76203, Texas, USA
[e] Intelligent Data Engineering and Analytics Lab, University of North Texas, Denton, 76203, Texas, USA

ARTICLE INFO

Keywords:
Scientific novelty
Novelty measures
Graph network
Large language model
Knowledge recombination
Knowledge propagation

ABSTRACT

Scientific novelty constitutes a fundamental catalyst for both disciplinary innovation and interdisciplinary progress. Nevertheless, prevailing approaches to novelty assessment predominantly emphasize a single analytical dimension–either the semantic content of the focal paper or its cited references. Content-based methodologies frequently fail to incorporate the foundational knowledge cited by the target publication, whereas reference-based strategies tend to disregard the intrinsic conceptual contributions of the focal work itself. To address this limitation, the present study introduces a hybrid graph and large language model approach to jointly capture and integrate knowledge embedded in both the focal paper and its cited literature. The proposed method, which integrates knowledge recombination and propagation, is structured into four primary stages. First, prompt-based extraction techniques using general LLMs are applied to extract knowledge. Second, a Reference Knowledge Combination Network (RKCN) is constructed to model the knowledge referenced by the focal paper. Third, the RKCN is initialized with representations generated by SciDeBERTa(CS), and a graph attention network is employed to propagate knowledge across the network. Finally, the novelty of the focal paper is quantified by aggregating the novelty scores of all internal knowledge combinations based on the propagated representations. Experimental evaluation in the domain of artificial intelligence (AI) demonstrates that the proposed method significantly outperforms existing baseline approaches in quantifying scientific novelty. Additional ablation studies further validate the contribution of the knowledge propagation module. A case study illustrates the interpretability of our approach, and a cross-field validation in Biomedical Engineering (BME) domain highlights its robustness and cross-domain generalizability. A multi-dimensional comparative analysis between award-winning and non-award papers further reveals that the former generally incorporate a larger volume of knowledge and exhibit greater diversity in knowledge combinations. Moreover, while both groups encompass knowledge combinations spanning a wide range of novelty, award-winning papers display a stronger concentration at higher novelty levels, in contrast to the more uniform distribution observed in non-award papers. Data, code, and more detailed results are publicly available at: https://github.com/haihua0913/graphLLM4ScientificNovelty.

1. Introduction

Scientific research has been serving as a fundamental driving force for the advancement of human civilization. Through the relentless exploration of natural laws, discovery of new knowledge, and innovation of technological methods, scientific research continually expands the boundaries of human cognition, providing a powerful impetus for progress across economic, social, and technological domains (Mormina, 2019). Within the trajectory of scientific development, scientific novelty, especially in its radical form, serves as a pivotal role in advancing scientific systems (Min et al., 2021; Sun et al., 2022). A notable example is the elucidation of the double-helix structure of DNA, which

[⋆] Funding: This work was supported by the Fundamental Research Funds for the Central Universities [grant number CCNU24ZZ132].

[*] Corresponding author.

E-mail addresses: wangzhy@ccnu.edu.cn (Z. Wang), w2499041033@mails.ccnu.edu.cn (Z. Wang), ricky@mails.ccnu.edu.cn (G. Zhang), jpchen@illinois.edu (J. Chen), markus.luczak-roesch@vuw.ac.nz (M. Luczak-Roesch), haihua.chen@unt.edu (H. Chen).

https://doi.org/10.1016/j.eswa.2025.129794
Received 3 May 2025; Received in revised form 6 August 2025; Accepted 18 September 2025
Available online 24 September 2025

not only redefined the theoretical foundations of molecular biology but also catalyzed technological advances in medicine, agriculture, and forensic science (Hood & Galas, 2003; Macgregor & Poon, 2003). Such novel scientific discoveries do not merely accelerate progress within individual disciplines but also generate far-reaching impacts across interdisciplinary domains. In recent years, the identification and quantification of scientific novelty have attracted increasing scholarly attention, emerging as a prominent area of investigation.

Contemporary scholarship has proposed a range of methodologies for assessing the novelty of scientific publications, which can be broadly classified into two categories: reference-based and content-based approaches. Reference-based methods evaluate novelty of a paper by examining the uniqueness of journal combinations cited within the paper (Foster et al., 2015; Lee et al., 2015; Uzzi et al., 2013; Wang et al., 2017a). While these approaches can partially identify scientific novelty, they primarily focus on the journal level rather than the detailed content of individual papers. As a result, they may fail to capture the novel combinations of fine-grained knowledge within the references. In contrast, content-based methods leverage textual information, such as titles, abstracts, and the main body of scientific articles, to measure novelty. For example, some studies have employed Medical Subject Headings (MeSH) terms to determine whether a publication introduces new concepts or exhibits unique term combinations (Azoulay et al., 2011; Ruan et al., 2023). Through bibliometric frequency analyses, these techniques calculate the proportion of innovative terminology presented. Nonetheless, such methods often fail to account for deeper semantic relationships among knowledge. Moreover, since MeSH is a domain-specific vocabulary developed for the biomedical and life sciences, methods that rely on it are inherently constrained in their generalizability. The absence of similarly comprehensive and standardized knowledge frameworks in many other scientific domains limits the applicability and effectiveness of these approaches beyond their original context.

To overcome the limitations of previous methods, this research combines the strengths of both reference-based and content-based approaches. Building on the theory of knowledge recombination and utilizing knowledge propagation model, we introduce a fine-grained method to quantitatively evaluate the novelty of scientific papers, implemented through a hybrid Graph and LLM framework.

Innovation fundamentally arises from the processes of knowledge combination and recombination (Schilling & Phelps, 2007; Weitzman, 1998). Innovation emerges when existing knowledge is restructured in novel ways (Wang et al., 2024a). For instance, the core of artificial intelligence–neural networks–was inspired by biological neural systems. One of the field's pioneers, Geoffrey Hinton, was awarded the 2024 Nobel Prize in Physics for his contributions. Likewise, genetic algorithms (Holland, 1992), mimic natural evolutionary processes to search for optimal solutions, with a single paper on the topic accumulating over 7000 citations. The way knowledge is combined plays a crucial role in the likelihood of scientific novelty. Atypical, complementary, and heterogeneous knowledge combinations are more likely to give rise to disruptive advancements (Fleming, 2001; Ma et al., 2023; Schilling & Green, 2011), whereas typical, homogeneous combinations tend to yield only incremental improvements.

References form the foundation upon which researchers filter, integrate, and recombine existing knowledge (Lubis et al., 2023; Shrivastava & Shrivastava, 2022). In writing a paper, researchers draw upon relevant literature to construct a knowledge base, which they then recombine to generate novel solutions to the problems they investigate. Accordingly, assessing the novelty of a focal paper's knowledge combination requires not only examining the content of the paper itself but also accounting for the prior knowledge embedded in its references. In this context, we refer to the paper whose novelty is being evaluated as the focal paper, following the terminology used in prior work (Wu et al., 2019).

Building on this foundation, the present study proposes a systematic approach to quantifying scientific novelty by evaluating the novelty of knowledge combinations within focal papers. The method comprises four main steps. First, key knowledge is extracted from the abstracts of both the focal paper and its cited references. Second, relationships among the extracted knowledge are identified by analyzing their cooccurrence within the reference abstracts. Based on these relationships, a knowledge association network is constructed from the cited references. This network is also referred to as the Reference Knowledge Cooccurrence Network (RKCN). Third, Graph Attention Networks (GATs) are employed to simulate the propagation of knowledge within the RKCN. This enables the model to learn the latent relationships between knowledge, positioning more strongly associated knowledge closer together in the embedding space. Finally, the similarity between each pair of knowledge in the focal paper is calculated using these embeddings. Lower similarity scores indicate more novel and less conventional combinations. By aggregating these novelty scores, the proposed method provides a quantitative framework for assessing the novelty of a paper.

The rest of paper is organized as follows: Section 2 reviews the related work in scientific novelty. Section 3 introduces the methodology to quantify scientific novelty with knowledge recombination and propagation. Section 4 details the experimental design and settings. Section 5 presents the results and discusses the findings. Section 6 summarizes the paper, discusses its limitations, and outlines future research directions.

2. Related work

2.1. Scientific novelty

Novelty is a widely recognized concept in scientific research and serves as a key criterion for assessing the academic value and contribution of publications (Zhao & Zhang, 2025). Consequently, evaluating the novelty of academic papers has become a central concern in scientometrics and research assessment (Hou et al., 2022). Research in this area generally follows two principal approaches: one focuses on the sources of novelty, while the other adopts a content-oriented interpretation.

The first line of research focuses on the intrinsic nature of scientific novelty, aiming to uncover the underlying mechanisms that drive novel developments. This approach emphasizes the processes of knowledge combination and reconfiguration that lead to transformative novelty (Jang et al., 2023; Mukherjee et al., 2017; Uzzi et al., 2013). It investigates how scientists draw upon diverse knowledge domains and creatively integrate them to generate novel insights. highlighting the critical role of cognitive mechanisms, heterogeneous knowledge combinations, and related factors in fostering scientific novelty. This theoretical perspective underpins a range of reference-based evaluation methods.

Conversely, the second line of research centers on content-level assessments of scientific novelty, with an emphasis on the novelty of the knowledge introduced within the focal paper itself. In his seminal work *The Structure of Scientific Revolutions* (1970), Thomas Kuhn proposed that scientific advancement unfolds through a cyclical process involving periods of normal science, crisis, revolutionary change, and the establishment of new paradigms. This conceptualization offers a foundational lens through which transformative developments can be understood as signaling the transition of a field into a novel phase of intellectual evolution, characterized by the emergence of new paradigms (Ahuja & Morris Lampert, 2001; Casadevall & Fang, 2016). Building on this view, content-based evaluation approaches seek to assess the intrinsic novelty embedded within a paper.

In summary, the intrinsic nature perspective emphasizes the proactive investigation of the generative mechanisms that give rise to scientific novelty, conceptualizing novelty as the outcome of recombining existing knowledge in novel configurations. In contrast, the content-level approach centers on evaluating the novelty inherent in the content of the paper itself. Both frameworks contribute valuable theoretical foundations for understanding and assessing scientific novelty. Integrating insights from these two perspectives, this study proposes a hybrid graph and LLM framework that provides a systematic approach to predict and measure scientific novelty.

2.2. Metrics for quantifying scientific novelty

Section 2.1 outlines the theoretical foundations underlying various metrics developed to quantify scientific novelty. Based on these theoretical distinctions, existing approaches can be broadly categorized into two types: reference-based evaluation methods and content-based evaluation methods.

Conceptually, reference-based evaluation methods emphasize the recombination of pre-existing knowledge. When these knowledge elements are combined in distant or unconventional ways, they are considered novel, a notion commonly referred to as "relative novelty" (Berlyne, 1960; Wang et al., 2025). In such frameworks, journals are typically treated proxies for knowledge elements (Zhao & Zhang, 2025). Early foundational work in this area can be traced to Uzzi et al. (2013), who introduced the Z-score to assess the novelty of scientific publications by identifying atypical combinations of cited journals. However, this approach is computationally intensive, both in terms of time and resource consumption. Subsequent studies sought to enhance this methodology. Lee et al. (2015) refined the computational strategy originally proposed by Uzzi et al. (2013), while Wang et al. (2017a) further optimized its computational complexity. These methods primarily assess whether the journal combinations referenced in a study are rare or unconventional. Despite their contributions, several limitations persist. First, their granularity is limited due to reliance on journal level analysis. As journals often encompass a wide range of topics and interdisciplinary publications have become increasingly common, assessing novelty at the journal level can be imprecise. Moreover, the relatively small number of journals compared to the vast and nuanced spectrum of domain-specific knowledge restricts the capacity of these methods to detect fine-grained knowledge recombination. Second, such approaches concentrate exclusively on the novelty of referenced journal pairs, while neglecting the internal knowledge structure of the focal paper. Given that scientific novelty frequently arises from the innovative integration of knowledge within a focus study, overlooking this dimension can significantly undermine the accuracy of novelty assessments.

Content-based approaches assess scientific novelty by leveraging textual features from the focal paper, including the title, abstract, and main body (Jeon et al., 2023; Shibayama et al., 2021; Wang et al., 2024b; Wu et al., 2025). A prominent subset of these methods employs Medical Subject Headings (MeSH), a standardized vocabulary system widely used in the biomedical and life sciences. These approaches quantify novelty by comparing the knowledge elements and their combinations within a given paper to those prevalent across the broader disciplinary landscape (Azoulay et al., 2011; Mishra & Torvik, 2016; Ruan et al., 2023). However, it is important to note that the application of MeSH terminology is primarily restricted to a single discipline–biomedical and life sciences. these methods necessitate access to all knowledge within a given field, making it challenging to apply in other disciplines. To address this limitation, subsequent research has adopted author-provided keywords as proxies for knowledge components within articles (Verhoeven et al., 2016; Yan et al., 2020). Nonetheless, a central challenge remains: the exhaustive collection of all prior knowledge and knowledge combinations within a given domain is highly demanding and often infeasible. Additionally, some scholars have explored sentence-level features to identify novelty, particularly through the analysis of contribution sentences that may signal scientific breakthroughs. For instance, Chen and Fang (2019) proposed a method to extract novel ideas directly from abstracts. While such approaches can partially capture elements of novelty, their fundamental limitation lies in its binary classification approach to novelty. Specifically, this dichotomous assessment fails to reflect the continuous and nuanced spectrum of novelty, thereby excluding many valuable papers that exhibit high degrees of novelty. Moreover, these methods typically disregard the relationship between the focal paper and its cited references, failing to evaluate how the paper diverges from or extends prior knowledge.

In summary, reference-based methods predominantly assess novelty by analyzing the combinations of journals associated with cited references, resulting in a relatively coarse level of granularity. Additionally, these approaches generally disregard the specific content of the focal paper itself. In contrast, content-based methods emphasize the intrinsic novelty of the focal paper's textual content but often fail to account for the foundational role of its referenced knowledge. Consequently, they overlook the combinatorial structure and interrelationships among the cited knowledge components, which are crucial for a comprehensive understanding of scientific novelty.

To address the aforementioned limitations, this study introduces a fine-grained approach that integrates the advantages of both reference-based and content-based methods while mitigating their respective shortcomings. Instead of treating the journal in which a paper is published as the fundamental unit of knowledge, our method considers the knowledge within the paper itself. Unlike reference-based methods that focus on journal combinations, we construct a knowledge association network based on the referenced knowledge combinations and leverage advanced graph representation learning techniques to model the propagation of knowledge. Through unsupervised learning, our approach enables knowledge with stronger associations to learn representations that bring them closer in the embedding space. Finally, we center our analysis on the focal paper, extracting knowledge from its abstract and identifying the corresponding knowledge combinations. Using the previously learned knowledge representations, we compute the similarity between each knowledge combination within the focal paper. A high similarity score indicates a lower degree of novelty for the combination, whereas a low similarity score suggests a higher level of novelty. By aggregating the novelty scores of all knowledge combinations within the focal paper, this approach provides a quantitative and systematic framework for evaluating scientific novelty.

3. Methodology

This paper proposes a comprehensive framework for quantifying novelty by analyzing the disparity between knowledge combinations within focal papers. The overall structure of this framework is illustrated in Fig. 1, which consists of four key components: Knowledge Extraction, **R**eference **K**nowledge **C**o-occurrence **N**etwork (RKCN) Construction, Knowledge Propagation on RKCN, and Focal Paper Novelty Computation.

The knowledge extraction module is designed to extract existing knowledge from the abstracts of the references cited by the focal paper, as well as the knowledge within the focal paper itself that require novelty assessment. The reference knowledge co-occurrence network (RKCN) construction module identifies knowledge combinations within the cited references and constructs a co-occurrence network that captures the structural relationships among the referenced knowledge. The knowledge propagation on RKCN module utilizes Graph Attention Networks (GATs) to simulate the propagation dynamics of knowledge within the network. Through unsupervised learning, the model captures the latent relationships between knowledge, enabling knowledge with stronger associations to learn representations that are closer in the embedding space. This ensures that knowledge combinations with high relevance exhibit greater representational similarity, while more novel combinations maintain a higher degree of differentiation. Finally, the focal paper novelty computation module quantifies the degree of novelty in the focal paper by computing the similarity between each pair of internal knowledge combinations based on the learned embeddings. A lower similarity score suggests a weaker prior association between the knowledge combination, indicating that the focal paper has combined the two knowledge elements that were previously weakly associated, thereby establishing a new link. Building on similarity scores, we

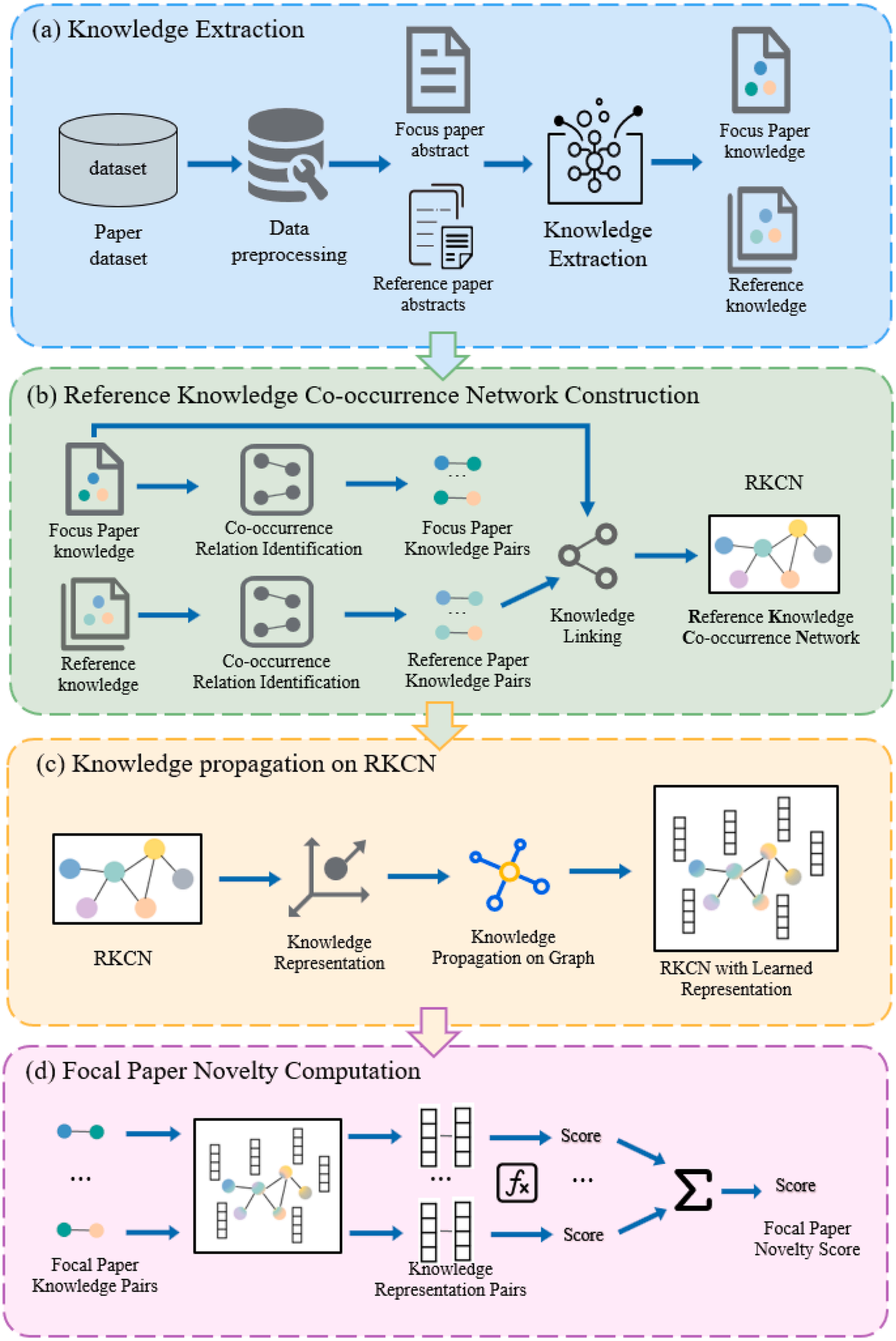

(a) Knowledge Extraction
dataset
Paper dataset
Data preprocessing
Focus paper abstract
Reference paper abstracts
Knowledge Extraction
Focus Paper knowledge
Reference knowledge
(b) Reference Knowledge Co-occurrence Network Construction
Focus Paper knowledge
Co-occurrence Relation Identification
Focus Paper Knowledge Pairs
Reference knowledge
Co-occurrence Relation Identification
Reference Paper Knowledge Pairs
Knowledge Linking
RKCN
Reference Knowledge Co-occurrence Network
(c) Knowledge propagation on RKCN
RKCN
Knowledge Representation
Knowledge Propagation on Graph
RKCN with Learned Representation
(d) Focal Paper Novelty Computation
Focal Paper Knowledge Pairs
Knowledge Representation Pairs
Score
Score
Score
Focal Paper Novelty Score

Fig. 1. Framework of measuring the novelty of paper based on knowledge recombination and propagation.

propose a method for measuring novelty, which quantifies the degree of novelty exhibited by the focal paper.

3.1. Knowledge extraction

Knowledge extraction plays a crucial role in academic research, facilitating the identification of knowledge utilized within studies (Nasar et al., 2018). It enables scholars to swiftly recognize and acquire key insights from vast volumes of research literature. This technique has been widely applied in various domains, including the construction of academic knowledge graphs (Al-Zaidy & Giles, 2017; Zhang et al., 2024), automated literature summarization (Bao et al., 2025; Hernández-Castañeda et al., 2022), topic analysis (Eshima et al., 2024; Lu et al., 2021), innovation detection (Wang et al., 2023; Zhao et al., 2024), and scientific recommendation (Bai et al., 2019; Zhang & Zhu, 2022).

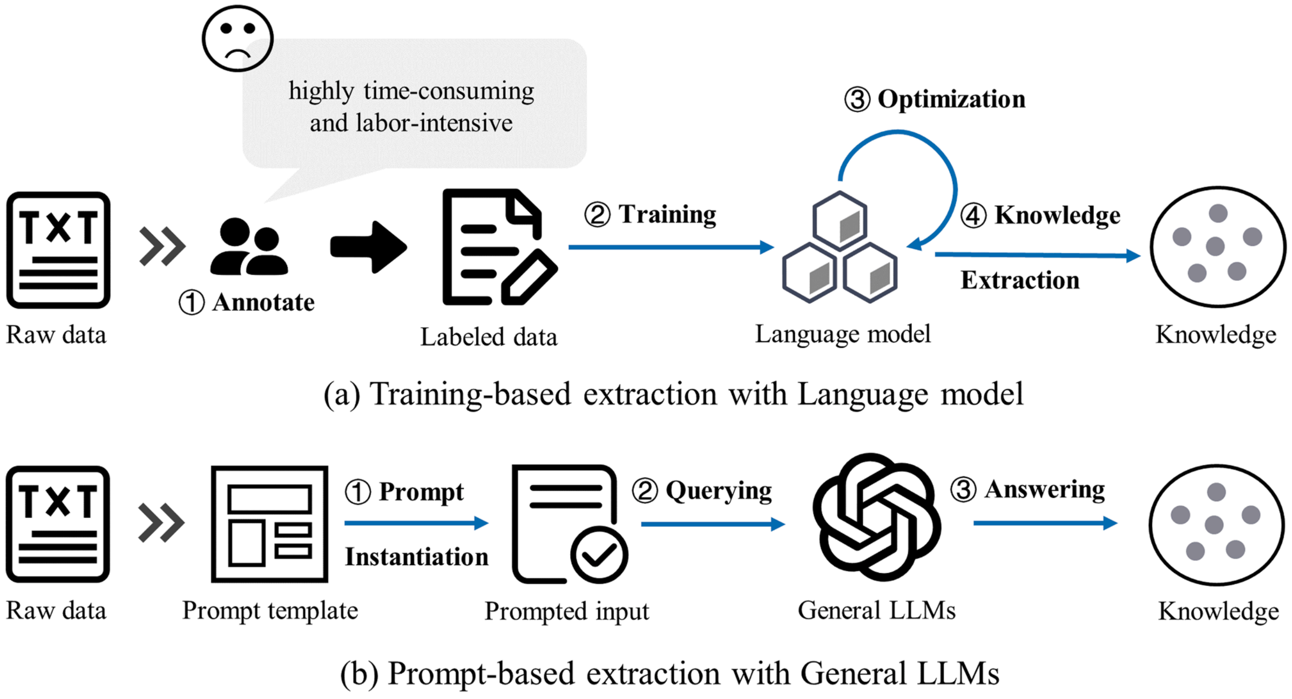
The extraction of key knowledge from academic literature generally follows two primary approaches: extraction from abstracts and extraction from full texts. Studies have shown that extracting knowledge from abstracts rather than full texts can enhance extraction accuracy and efficiency (Popova & Danilova, 2014). Although full-text documents offer comprehensive information, they often contain extensive background material and ancillary content, which can obscure the identification of core scientific insights. In contrast, abstracts function as succinct summaries that distill the critical elements of a paper, including the research context, methodologies, key findings, and conclusions (śauperl et al., 2008; Weil, 1970). Their highly condensed logical structure makes them an excellent source for identifying core academic knowledge. Knowledge extracted from abstracts more accurately reflects the principal research themes, minimizes redundancy, and reduces the risk of misinterpretation. Consequently, abstract-based extraction presents a more targeted and reliable approach for identifying the core scientific contributions of academic work, relative to full-text extraction.

Knowledge extraction typically follows one of two main approaches. The first is training-based extraction using language models (Fig. 2a), which requires a substantial amount of domain-specific annotated data for model training or fine-tuning. This approach is both time-consuming and labor-intensive, presenting significant challenges for practical implementation. The second approach involves prompt-based extraction using general LLMs (Fig. 2b). This method eliminates the need for model retraining by relying solely on the adaptation of prompt templates, thereby offering greater flexibility and efficiency. In this study, we adopt the prompt-based extraction paradigm. Specifically, we employ GPT-4o as our primary LLM and include OLMo2:13b as a baseline for comparative evaluation. The complete set of prompt templates is provided in Appendix A, and the overall extraction workflow is illustrated in Fig. 3. Formally, given an input abstract x, the extraction process consists of three stages. First, the abstract is embedded into the predefined prompt template via a mapping function $\mathcal{P}(\cdot)$, yielding a formatted prompt suitable for model inference. Second, this prompt is passed to the LLM, denoted as $\mathrm{LLM}(\cdot)$, which generates the corresponding output in natural language form. Third, the model output is parsed by a deterministic function $g(\cdot)$ based on comma separation, resulting in a set of extracted textual knowledge $\{T_k\}_{k=1}^{K}$, formally defined as follow:

$$\{T_k\}_{k=1}^{K} = g\Big(\mathrm{LLM}\big(\mathcal{P}(x)\big)\Big). \tag{1}$$

3.2. Reference knowledge co-occurrence network construction

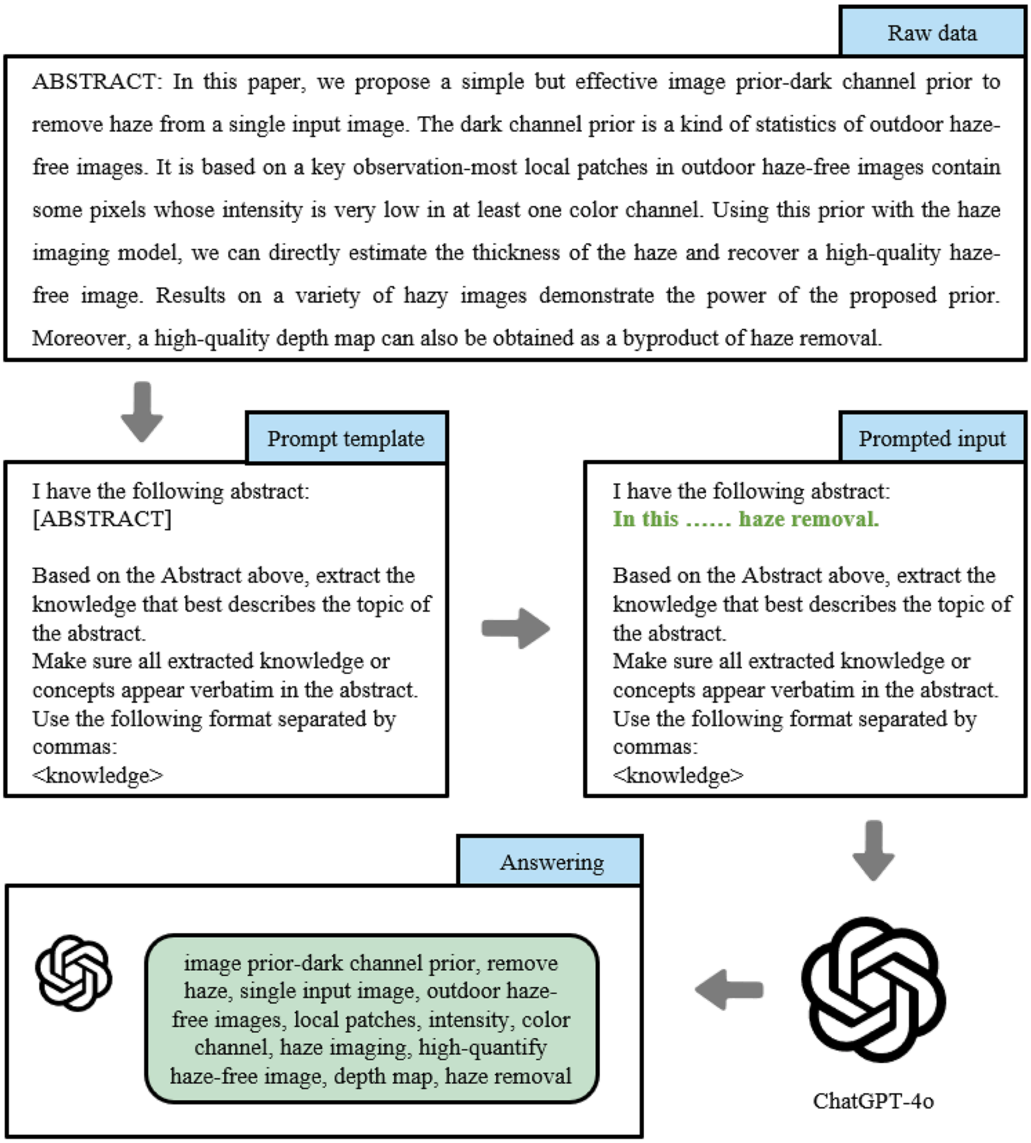
References constitute foundational sources of knowledge upon which authors build their research, forming the intellectual basis for scientific novelty. The primary objective of constructing a co-occurrence network is to uncover the intrinsic relationships among existing knowledge elements (Zhu et al., 2015; Zong et al., 2013). By establishing a reference knowledge co-occurrence network (RKCN), we can not only effectively identify established knowledge combinations but also reveal latent associations between knowledge items that may not be directly linked, yet are connected through indirect pathways within the network structure. In this study, the co-occurrence window is precisely defined as encompassing the current sentence, the preceding sentence, and the subsequent sentence. For instance, if a knowledge element appears in sentence i of an abstract, the knowledge found in sentence $i-1$ and sentence $i+1$ is considered to co-occur "directly" with it, indicating a



(a) Training-based extraction with Language model

(b) Prompt-based extraction with General LLMs

Fig. 2. Comparison of knowledge extraction methods: training-based method and prompt-based method.



ABSTRACT: In this paper, we propose a simple but effective image prior-dark channel prior to remove haze from a single input image. The dark channel prior is a kind of statistics of outdoor haze-free images. It is based on a key observation-most local patches in outdoor haze-free images contain some pixels whose intensity is very low in at least one color channel. Using this prior with the haze imaging model, we can directly estimate the thickness of the haze and recover a high-quality haze-free image. Results on a variety of hazy images demonstrate the power of the proposed prior. Moreover, a high-quality depth map can also be obtained as a byproduct of haze removal.

I have the following abstract:
[ABSTRACT]

Based on the Abstract above, extract the knowledge that best describes the topic of the abstract.
Make sure all extracted knowledge or concepts appear verbatim in the abstract.
Use the following format separated by commas:
<knowledge>

I have the following abstract:
In this ...... haze removal.

Based on the Abstract above, extract the knowledge that best describes the topic of the abstract.
Make sure all extracted knowledge or concepts appear verbatim in the abstract.
Use the following format separated by commas:
<knowledge>

image prior-dark channel prior, remove haze, single input image, outdoor haze-free images, local patches, intensity, color channel, haze imaging, high-quantify haze-free image, depth map, haze removal

Fig. 3. Examples of knowledge extraction leveraging prompt-based method with GPT-4o.

strong semantic association. This design is motivated by the observation that in concise texts such as abstracts, semantic information is typically concentrated in adjacent sentences, with minimal semantic extension to more distant parts of the text. Even when semantically related knowledge appears in non-adjacent sentences, these weaker associations can still be captured through the knowledge propagation mechanism described in Subsection 3.3. Thus, by restricting the co-occurrence window to a span of three sentences, the model effectively captures salient semantic relationships between knowledge units. For example, if sentence i contains knowledge A, sentence $i+1$ contains knowledge B, and sentence $i+2$ contains knowledge C, then according to the co-occurrence principle, A and B as well as B and C are directly co-occurring, reflecting strong semantic ties. Although A and C do not directly co-occur, they remain connected in the co-occurrence network, allowing their indirect relationship to be inferred through propagation. This is referred to as “indirect” co-occurrence. This strategy of leveraging local contextual windows has been widely adopted in various natural language processing models, including n-gram models (Brown et al., 1992), convolutional neural networks (Kim, 2017), long short-term memory networks (Hochreiter & Schmidhuber, 1997) and graph convolutional networks (Kipf & Welling, 2017), all of which have achieved notable success in capturing meaningful linguistic and semantic patterns.

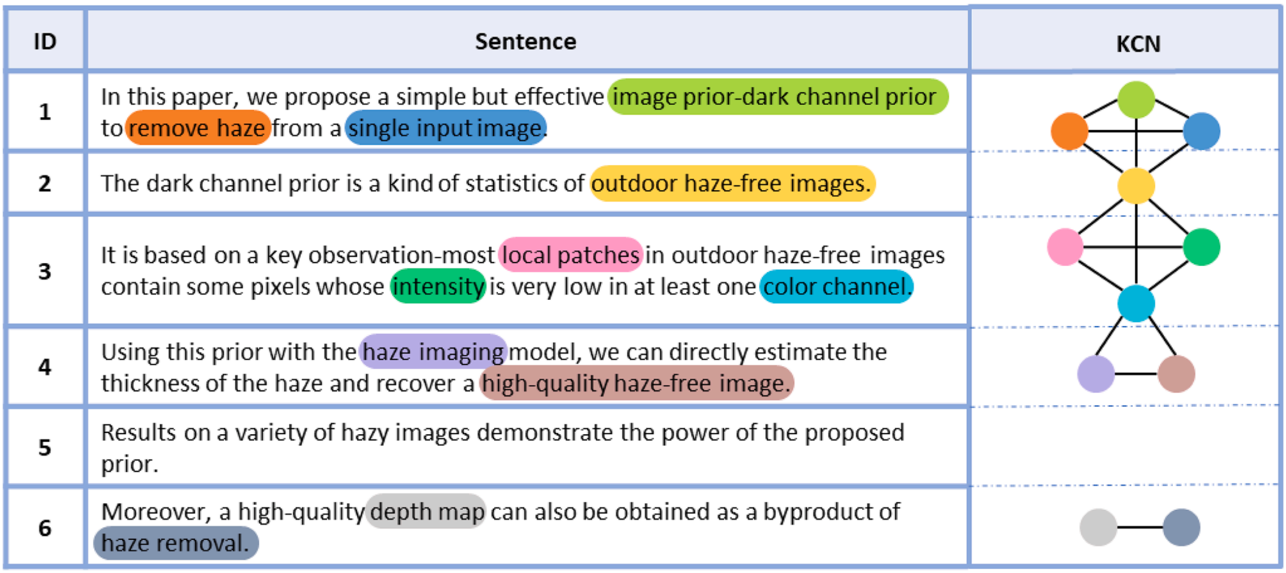
Formally, given an input abstract segmented into sentences $(s_1, s_2, \ldots, s_L)$ and a set of extracted knowledge elements $\{T_k\}_{k=1}^{K}$, each annotated with its corresponding sentence position, we define two knowledge elements T_i and T_j as co-occurring if the absolute difference between their sentence indices is at most one. The co-occurrence indicator is formally defined as:

$$\text{Cooccur}(T_i, T_j) = \begin{cases} 1, & \text{if } |\text{pos}(T_i) - \text{pos}(T_j)| \leq 1, \quad i \neq j, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Based on the defined co-occurrence relation, we construct a knowledge co-occurrence network (KCN) for the abstract of each reference, as illustrated in Fig. 4. In this network, each node represents an extracted knowledge element, and edges denote co-occurrence within adjacent sentences. The KCN is formally defined in Eq. 3. By subsequently merging all KCNs derived from the abstracts of the reference papers, we construct the reference knowledge co-occurrence network (RKCN) for the focal paper, as described in Eq. 4.

$$\text{KCN} = (V, E) \quad \text{where} \quad \begin{cases} V = \{T_1, T_2, \ldots, T_K\}, \\ E = \{(T_i, T_j) \mid \text{Cooccur}(T_i, T_j) = 1\}. \end{cases} \tag{3}$$

$$\text{RKCN} = (V', E', W') \quad \text{where}$$



ID	Sentence	KCN
1	In this paper, we propose a simple but effective image prior-dark channel prior to remove haze from a single input image.	
2	The dark channel prior is a kind of statistics of outdoor haze-free images.	
3	It is based on a key observation-most local patches in outdoor haze-free images contain some pixels whose intensity is very low in at least one color channel.	
4	Using this prior with the haze imaging model, we can directly estimate the thickness of the haze and recover a high-quality haze-free image.	
5	Results on a variety of hazy images demonstrate the power of the proposed prior.	
6	Moreover, a high-quality depth map can also be obtained as a byproduct of haze removal.	

Fig. 4. An example of constructing KCN from a single paper's abstract.

$$\begin{cases} V' = \bigcup V, \\ E' = \bigcup E, \\ W'(T_i, T_j) = \text{number of times } (T_i, T_j) \text{ appears across abstracts.} \end{cases} \tag{4}$$

3.3. Knowledge propagation on RKCN

A key prerequisite for conducting knowledge propagation on the reference knowledge co-occurrence network is obtaining high-quality semantic representations of the knowledge nodes. Since these nodes are expressed in natural language, they must first be transformed into structured embeddings to support subsequent propagation using Graph Neural Networks. To accurately capture the semantic content of knowledge within the RKCN, this study employs SciDeBERTa(CS) for embedding representation. SciDeBERTa(CS) is a pre-trained language model (PLM) tailored specifically for the computer science domain (Jeong & Kim, 2022). Compared to other domain-specific models such as SciBERT and S2ORC-SciBERT, SciDeBERTa(CS) consistently delivers superior performance because of its deeper and more targeted pre-training on a large corpus of computer science literature. It has demonstrated state-of-the-art results on multiple domain-relevant benchmarks. We therefore adopt SciDeBERTa(CS) to generate embeddings for the knowledge nodes in the RKCN, thereby ensuring semantically rich and contextually accurate representations for effective graph-based propagation.

In practice, each knowledge node in the RKCN corresponds to a textual description, denoted T_k. To obtain its semantic representation, we employ SciDeBERTa(CS) to encode each text sequence into a dense vector embedding. Specifically, given a textual input T_k, composed of a token sequence $w_1, w_2, \dots, w_n$, SciDeBERTa(CS) processes the sequence and outputs the corresponding embedding representation, defined as follows:

$$X_k = \text{SciDeBERTa}(T_k) = \{x_1, x_2, \dots, x_n\}, \quad x_i \in \mathbb{R}^{768} \tag{5}$$

where x_i represents the 768-dimensional hidden state of the i-th token.

To obtain a fixed-length embedding for each knowledge node, we apply mean pooling over all token embeddings, as follows:

$$X_k = \frac{1}{n} \sum_{i=1}^{n} x_i \in \mathbb{R}^{768} \tag{6}$$

where, the resulting 768-dimensional vector X_k serves as the initial feature representation for the knowledge node k in the reference knowledge co-occurrence network.

In this section, knowledge propagation is formulated as an adaptive message passing process, in which each node aggregates information from its neighbors using a multi-head attention mechanism, as

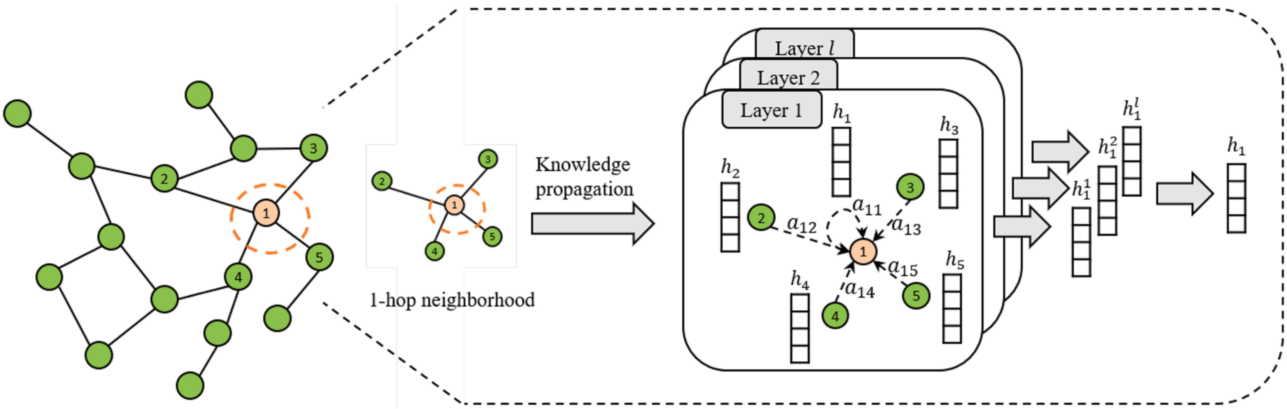


Fig. 5. GAT-based knowledge propagation from neighbors to target knowledge node.

illustrated in Fig. 5. Given the RKCN graph G, each knowledge node $k \in K$ is initialized with an embedding representation X_k obtained using SciDeBERTa(CS). The propagation process is then carried out using a Graph Attention Network (GAT), which enables the model to dynamically attend to the most relevant neighboring nodes during information aggregation.

First, for each knowledge node k, we compute an attention score between k and each of its neighbors v based on their current feature representations. The attention score for the i-th attention head is defined as:

$$e_{kv}^{(i)} = \text{LeakyReLU}\left(\mathbf{a}^T \left[W^{(i)}\mathbf{h}_k \parallel W^{(i)}\mathbf{h}_v\right]\right) \tag{7}$$

where:

- i denotes the index of the attention head.
- $W^{(i)} \in \mathbb{R}^{d\times d'}$ is a trainable weight matrix that projects the input features into a lower-dimensional space of dimension d'.
- $\mathbf{a} \in \mathbb{R}^{2d'}$ is a trainable attention vector used to compute the raw attention coefficients.
- $\parallel$ denotes vector concatenation, enabling the model to jointly consider the features of both nodes k and v.
- $e_{kv}^{(i)}$ represents the unnormalized attention coefficient, indicating the relative importance of node v to node k under the i-th attention head.

To ensure that each knowledge node aggregates information proportionally from its neighbors, the raw attention scores are normalized using the softmax function, defined as:

$$\alpha_{kv}^{(i)} = \frac{\exp(e_{kv}^{(i)})}{\sum\limits_{j\in\mathcal{N}(k)} \exp(e_{kj}^{(i)})}, \quad \alpha_{kv}^{(i)} \in [0,1] \tag{8}$$

where:

- $\alpha_{kv}^{(i)}$ denotes the normalized attention coefficient between knowledge node k and its neighbor v under the i-th attention head.
- $\mathcal{N}(k)$ denotes the set of neighboring nodes of node k.

Finally, using the computed attention coefficients, information is propagated from neighboring knowledge nodes to the target node through a weighted aggregation mechanism. The update rule is defined as:

$$\mathbf{h}_k^{(l+1)} = \parallel_{i=1}^{I} \sigma\left(\sum_{v\in\mathcal{N}(k)} \alpha_{kv}^{(i)} \mathbf{W}^{(i)} \mathbf{h}_v^{(l)}\right) \tag{9}$$

where:

- $\mathbf{h}_v^{(l)}$ denotes the feature vector of knowledge node v at layer l, with $\mathbf{h}_v^{(0)}$ corresponding to the initial input feature representation $\mathbf{X}_v$ of knowledge node v.
- $W^{(i)}$ is a learnable weight matrix that performs a linear transformation under the i-th attention head.
- $\alpha_{kv}^{(i)}$ represents the normalized attention coefficient determining how much knowledge is propagated from knowledge node v to knowledge node k.
- σ is a non-linear activation function that introduces model expressiveness and non-linearity.
- I denotes the total number of attention heads, and $\parallel$ indicates the concatenation of outputs from all heads.

To capture multi-scale relational dependencies, we adopt a multi-layer GAT, progressively reducing the dimensionality of node feature representations across layers. Each layer refines the node embeddings by aggregating both local neighborhood information and higher-order relational structures. To improve training stability and prevent overfitting, batch normalization and dropout are applied after each layer.

While GAT-based models effectively enable knowledge propagation, they face two key challenges: (i) ensuring effective propagation, i.e., verifying whether each knowledge node adequately integrates information from its neighbors; and (ii) maintaining knowledge diversity and global structural distinctiveness, as over-propagation during multi-hop aggregation can cause node representations to become excessively similar. This dilutes the uniqueness of individual nodes and obscures important global patterns. To address these issues and ensure high-quality knowledge propagation on the RKCN, we propose a dual-objective loss function, composed of:

1. **Neighborhood aggregation loss**: A local consistency term that encourages semantically similar representations among neighboring nodes, thus directly measuring propagation effectiveness. It is defined as:

   $$\mathcal{L}_{\text{neigh}} = \frac{1}{|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} \left(1 - \cos(\mathbf{h}_i, \mathbf{h}_j)\right), \tag{10}$$

   where
   - $\mathcal{E}$ is the set of edges in the graph, and $|\mathcal{E}|$ is its cardinality.
   - $\mathbf{h}_i$ and $\mathbf{h}_j$ are the embeddings of nodes i and j, respectively.
   - $\cos(\mathbf{h}_i, \mathbf{h}_j) = \frac{\mathbf{h}_i\cdot\mathbf{h}_j}{\|\mathbf{h}_i\|\|\mathbf{h}_j\|}$ denotes the cosine similarity between the two node embeddings.

   Although minimizing $\mathcal{L}_{\text{neigh}}$ promotes coherence among neighboring nodes, over-minimization can result in over-smoothing, where, after multiple training epochs, node representations converge to similar values, losing their unique characteristics and the overall structural distinctions of the graph. To mitigate over-smoothing and preserve global-level distinctiveness, we introduce Structural Entropy Loss, which minimizes the structural entropy of the entire graph.
2. **Structural entropy loss**: A global diversity term that preserves the distinctiveness of node representations, mitigating the risk of over-smoothing and excessive similarity. It is defined as:

   $$\mathcal{L}_{\text{entropy}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d_{\text{out}}} p_{ij}\log\left(p_{ij} + \epsilon\right), \tag{11}$$

   where
   - N is the number of nodes in the graph.
   - d_{out} is the output dimensionality of node embeddings.
   - p_{ij} is the softmax-normalized value of the j-th feature of node i.
   - ϵ is a small constant added to ensure numerical stability.

To achieve a structured and balanced knowledge propagation process, we integrate the above two objectives into a joint optimization function:

$$\mathcal{L} = \mathcal{L}_{\text{neigh}} + \lambda\,\mathcal{L}_{\text{entropy}}, \tag{12}$$

where λ is a tunable hyperparameter that balances local consistency and global distinctiveness.

3.4. Focal paper novelty computation

Novel knowledge is inherently built upon existing knowledge (Brockman & Morgan, 2003). In academic research, the novelty of a paper is often manifested through its unconventional combinations of pre-existing knowledge. Prior studies have shown that the introduction of novel ideas frequently arises from reconnecting weakly associated components within the current knowledge system (Uzzi et al., 2013; Xiao et al., 2022; Yan et al., 2020). In essence, when a paper successfully integrates knowledge that were previously only weakly connected, it may offer fresh perspectives and potentially catalyze theoretical breakthroughs. Building on this insight, we propose a quantitative approach for evaluating the novelty of focal papers. Specifically, we utilize the embedding representations derived through knowledge propagation to calculate the similarity between every pair of knowledge (k_1, k_2) within the focal paper. A lower similarity score indicates that the two knowledge elements were weak association in the existing literature, thereby suggesting that

Table 1
Award-winning and non-award papers by year and conference.

Year	Award-winning Papers							Non-award Papers							Total
	AAAI	ACL	CVPR	ICCV	ICML	IJCAI	NeurIPS	AAAI	ACL	CVPR	ICCV	ICML	IJCAI	NeurIPS	
1996	3	-	-	-	-	-	-	285	59	138	-	67	-	152	704
1997	4	-	-	-	-	3	-	210	74	174	-	49	236	157	907
1998	3	-	-	2	-	-	-	204	247	145	166	67	-	152	986
1999	1	-	-	2	1	2	-	193	84	193	176	54	202	151	1059
2000	4	-	1	-	-	-	-	230	80	229	-	152	-	154	850
2001	-	2	1	2	-	1	-	-	69	273	219	81	200	198	1046
2002	1	1	-	-	-	-	-	180	65	-	-	88	-	208	543
2003	-	2	1	3	-	2	-	-	70	206	196	118	295	199	1092
2004	1	1	1	-	-	-	-	194	88	257	-	118	-	208	868
2005	1	1	1	1	1	3	-	325	134	107	246	134	348	208	1510
2006	2	1	1	-	1	-	-	384	307	166	-	140	-	205	1207
2007	2	1	1	1	1	3	-	367	204	538	389	150	477	218	2352
2008	2	2	2	-	1	-	-	355	118	505	-	157	-	251	1393
2009	-	3	1	1	1	2	-	-	119	382	308	180	343	263	1603
2010	2	1	1	-	1	-	-	312	160	461	-	159	-	293	1390
2011	2	1	1	1	1	3	-	318	164	439	339	152	488	307	2216
2012	2	1	1	-	1	-	-	352	188	466	-	243	-	371	1625
2013	2	1	1	1	2	2	3	250	328	471	454	282	495	358	2650
2014	1	1	1	-	1	-	2	474	287	540	-	310	-	410	2027
2015	1	2	1	1	2	2	2	674	318	602	526	269	648	402	3450
2016	1	1	1	-	3	1	1	691	329	643	-	320	658	569	3218
2017	1	1	2	1	1	1	3	786	303	782	621	434	781	677	4394
2018	1	3	1	-	2	7	4	1102	437	979	-	620	864	1007	5027
2019	1	1	1	1	2	1	1	1343	703	1294	1075	773	965	1427	7588
2020	1	1	1	-	2	2	3	1864	778	1465	-	1084	777	1896	7874
2021	2	1	1	1	1	3	8	1960	713	1660	1611	1183	719	2327	10190
2022	1	4	1	-	10	1	12	1624	700	2072	-	1225	863	2823	9336
2023	1	3	2	2	6	3	2	2021	1073	2355	497	1823	844	3539	12171
total	43	36	26	20	41	42	41	16,698	8199	17,542	6823	10,432	10,203	19,130	89276

the paper has established a novel connection between them. To quantify the novelty introduced by such combinations, we adopt the method proposed by Liu et al. (2022), which defines the novelty contribution of each pair of knowledge combinations based on 1 − cosine similarity. Since 1 − cosine similarity theoretically ranges from 0 to 2, we normalize this value by dividing it by 2, ensuring that the final novelty score falls within the range $[0, 1]$. Formally, the novelty contribution of a knowledge pair is defined as:

$$\mathcal{N}(k_1, k_2) = \frac{1 - \frac{\sum_{t=1}^{n} h_{1,t} h_{2,t}}{\sqrt{\sum_{t=1}^{n} h_{1,t}^2} \times \sqrt{\sum_{t=1}^{n} h_{2,t}^2}}}{2} \tag{13}$$

where $\mathcal{N}(k_1, k_2) \in [0, 1]$. A higher value of $\mathcal{N}(k_1, k_2)$ indicates a greater degree of novelty in the knowledge combination, whereas a lower value suggests a more conventional or previously established relationship.

Finally, the overall novelty score of a focal paper is computed by summing the novelty contributions of all knowledge pairs it contains:

$$\mathcal{N} = \sum_{(k_1,k_2)\in P} \mathcal{N}(k_1, k_2) \tag{14}$$

where P denotes the set of all knowledge combinations within the paper.

This novelty metric reflects the contribution of the focal paper in forming novel knowledge combinations. When a paper successfully integrates knowledge that was previously weakly associated, the resulting low similarity scores yield higher novelty contributions, thereby increasing the overall novelty score. Through this quantitative framework, we introduce a systematic and interpretable metric for assessing scientific novelty.

4. Experimental design

4.1. Data

One major challenge in studying scientific novelty is the lack of a universally accepted and operational definition of novelty, as well as the absence of publicly available benchmark datasets that directly label the novelty level of individual scientific papers, making it difficult to objectively validate novelty-related methods. As a result, most existing studies rely on indirect proxies. While early approaches often used citation counts, which suffer from time delays and confounding factors, more recent efforts have turned to award-winning status as a more timely and expert-endorsed indicator. In many prestigious venues, novelty is explicitly listed as a key evaluation criterion for awards, making this proxy scientifically meaningful. Against this backdrop, the rapid advancement of artificial intelligence (AI) has profoundly transformed both industrial production and everyday life, positioning AI as a compelling domain for studying the dynamics of technological and scientific innovation. Due to the fast-paced evolution of AI, conferences have overtaken traditional journals as the primary venues for disseminating cutting-edge research. Consequently, in this study, we selected the proceedings of seven premier AI conferences–AAAI, ACL, NeurIPS, CVPR, ICCV, ICML and IJCAI–as our data sources. These conferences are internationally recognized as top-tier forums within the AI research community and are classified as Class A conferences by the China Computer Federation (CCF)[1]. In most editions, these conferences designate a subset of accepted papers as award-winning. Such selections are typically based on rigorous peer evaluation, with research novelty, originality, and potential impact serving as core criteria in the decision-making process, making them a valuable benchmark for evaluating scientific novelty.

Our dataset includes all papers published in these seven conferences from 1996 to 2023. Table 1 presents the statistics for award-winning and non-award papers over this period. Paper titles were obtained from the DBLP computer science bibliography database[2], and we also collected the corresponding abstracts. The control sample was constructed in two steps. First, for each award-winning paper, we randomly selected four non-award papers from the same conference and publication year, provided that their abstracts were available. Restricting the selection to the same venue and year helps mitigate potential confounding effects

[1] https://www.ccf.org.cn/Academic_Evaluation/By_category/
[2] https://dblp.org/

Table 2
Descriptive statistical analysis of award-winning and non-award papers.

Conf.	Metrics	Reference Count		Citation Count		Influential Citation Count	
		Non-award	Award-winning	Non-award	Award-winning	Non-award	Award-winning
AAAI	Mean	22.37	26.22	53.55	241.83	4.85	23.39
	Max	107	70	751	3103	108	458
	Min	2	6	0	2	0	0
	Median	20	21	20	56	1	5
	Std	15.45	16.03	96.11	540.40	12.19	72.98
ACL	Mean	31.23	35.57	81.17	352.46	9.44	33.91
	Max	119	133	808	3414	119	272
	Min	3	11	0	9	0	0
	Median	28	29	33	168	2	13
	Std	19.20	23.71	136.69	622.70	20.34	58.84
CVPR	Mean	31.44	51.08	114.24	9220.69	11.49	1386
	Max	142	154	1053	179,774	188	28678
	Min	5	5	0	11	0	2
	Median	29	45	37.5	402.5	2.5	58
	Std	18.37	35.68	187.67	35420.52	25.86	5633.14
ICCV	Mean	29.59	48	176.76	2751.05	11.83	393.15
	Max	93	120	2668	25,132	165	3793
	Min	2	19	0	3	0	0
	Median	25.5	40	38.5	408.5	2	25
	Std	17.73	26.89	398.90	6526.48	25.92	1003.76
ICML	Mean	40.45	50.05	184.91	500.13	17.76	64.44
	Max	105	103	9761	3500	770	439
	Min	11	14	0	0	0	0
	Median	37.5	48	35	107	3	11
	Std	21.44	23.67	810.38	900.97	65.82	121.05
IJCAI	Mean	28.63	31.33	43.01	83.28	4.08	7.39
	Max	107	91	648	492	146	76
	Min	2	6	0	0	0	0
	Median	26	29	15	43	1	2
	Std	16.90	18.40	96.76	106.62	14.81	14.14
NeurIPS	Mean	47.28	65.11	66.24	1344.55	7.72	165.32
	Max	100	168	589	34,537	116	3817
	Min	16	20	0	10	0	1
	Median	44	62.5	31	129	3	15.5
	Std	19.53	35.22	97.41	5627.60	15.11	631.81
Total	Mean	33.36	43.3	97.84	1652.35	9.42	233.06
	Max	142	168	9761	179,774	770	28678
	Min	2	5	0	0	0	0
	Median	29	36	28	125	2	11
	Std	20.19	28.84	374.52	12213.67	32.43	1919.55

arising from shifts in research focus or changes in peer-review standards over time. Second, we balanced statistical power with practical feasibility in determining the control group size. Using only one or two non-award papers per award-winning paper may lead to insufficient variance and reduce estimation precision. Conversely, using too many controls may introduce imbalance between the treatment (award-winning) and control groups, undermining comparability. Drawing on established practices in matched control study designs, we adopted a 1:4 matching ratio to ensure an appropriate level of representativeness without introducing excessive imbalance between the award and control groups. To evaluate the robustness of this choice, we conducted sensitivity analyses by varying the number of control papers per award-winning paper from 1:1 to 6:1. The results remained largely consistent across different ratios, suggesting that the model's performance is stable and not substantially affected by the matching ratio. Further experimental details and comparative results are reported in subsection 5.1.

Based on the collected paper titles, we constructed the final dataset for empirical analysis through the following steps. First, We utilized the Semantic Scholar API[3] to retrieve key metadata for each paper, including the fields: “paper_id”, “title”, “abstract”, “year”, “referenceCount”, “citationCount”, “influentialCitationCount”, and “fieldsOfStudy”. All papers, along with their references, were stored in csv format for further processing. Among these fields, influentialCitationCount represents the number of Highly Influential Citations, which refer to citations where the cited publication has a significant impact on the citing publication. Although it is not possible to rule out cases where certain non-award papers exhibit a high degree of novelty or where some award-winning papers demonstrate relatively low novelty, overall, award-winning papers tend to exhibit a higher degree of novelty compared to non-award papers presented at the same conference in the same year (Runhui et al., 2025).

To further compare the characteristics of award-winning and non-award papers, we conducted a descriptive statistical analysis of their reference counts, citation counts and influential citation counts. The summary statistics are presented in Table 2. Key observations include:

- **Reference count:** The statistical results indicate that award-winning papers exhibit slightly higher mean, median, maximum, and minimum reference counts compared to non-award papers. This suggests that award-winning papers tend to draw upon a broader and more diverse knowledge base. However, the standard deviation of reference counts is relatively large for award-winning papers, indicating

[3] https://www.semanticscholar.org/product/api

greater variability in the number of references among them compared to non-award papers.
- **Citation count:** In terms of citation impact, award-winning papers demonstrate significantly higher mean, maximum, and median citation counts than non-award papers. This trend aligns with the general expectation that award-winning papers tend to receive greater academic recognition, reflecting their higher scholarly influence. The relatively large standard deviation and the presence of low minimum values warrant attention, as they suggest that while most award-winning papers are widely cited, some receive considerably fewer citations, even approaching the citation levels of non-award papers.
- **Influential citation count:** When assessing the impact of papers on subsequent research, award-winning papers exhibit substantially higher mean, median, and maximum influential citation counts compared to non-award papers. This further supports the notion that award-winning papers are more likely to be highly innovative and serve as foundational research within their respective fields. However, similar to the citation count, the large standard deviation in influential citation counts suggests that while some award-winning papers achieve groundbreaking influence and become pivotal references for future studies, others have a more limited impact within the academic community.

4.2. Experimental design

We conducted six experiments on the aforementioned datasets to evaluate the effectiveness, interpretability and robustness of our proposed method and to analyze and compare the characteristics of award-winning and non-award papers. The experiments are outlined as follows:

- To evaluate the robustness and reliability of our model across different control group sizes, we conducted sensitivity analysis by altering the control-to-treatment ratio and tracking variations in key performance metrics such as AUC, accuracy, precision, recall, and F1-score.
- To comprehensively assess the effectiveness and robustness of our approach, we conducted comparative evaluations against baseline models and analyzed the sources of performance differences, statistical significance tests to verify whether our novelty metric meaningfully distinguishes award-winning papers, and robustness checks on the aggregation method using Tukey's HSD Test to evaluate the impact of potential outliers.
- To further verify the effectiveness of each module, particularly the knowledge propagation module, we performed an ablation study by systematically removing or replacing specific components and analyzing the resulting changes in performance.
- To explore and compare the characteristics of award-winning and non-award papers, we conducted an analysis focusing on: (1) the distribution characteristics of knowledge and knowledge combinations in award-winning and non-award papers. (2) the distribution characteristics of knowledge combination similarity and novelty in award-winning and non-award papers.
- To evaluate the interpretability of the proposed combinatorial novelty, we conducted a small-scale qualitative case study assessing both award-winning and non-award papers.
- To evaluate the generalizability of the proposed method beyond the AI domain, we conducted a cross-field validation using data from the biomedical engineering domain.

4.3. Evaluation metrics

This study employs the ROC curve to evaluate the predictive performance of various metrics. The ROC curve is a widely used performance evaluation method for tasks involving quantification, particularly adept at assessing a model's discriminative capacity across varying decision thresholds. Its fundamental concept involves adjusting the classification threshold to compute the true positive rate (TPR) and false positive rate (FPR) at each level, subsequently plotting their relationship to evaluate the model's overall discriminatory prowess. Specifically, the formulas for TPR and FPR are as follows:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{15}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{16}$$

where, TP (True Positives) denotes the number of positive instances correctly identified, while FP (False Positives) represents the number of negative instances incorrectly classified as positive. TN (True Negatives) refers to the number of negative instances accurately identified, and FN (False Negatives) indicates the number of positive instances mistakenly classified as negative. The Area Under the ROC Curve (AUC) serves as a comprehensive metric for evaluating the model's performance. A higher AUC indicates stronger discriminative power. Specifically, an AUC of 0.5 suggests that the model's performance is equivalent to random guessing. AUC values within the range of 0.5 to 0.6 are considered poor, while those between 0.6 and 0.7 are classified as fair. AUC scores ranging from 0.7 to 0.8 indicate good performance, whereas values between 0.8 and 0.9 are deemed excellent. An AUC greater than 0.9 signifies outstanding classification capability, with an AUC of 1.0 representing a perfect classifier (Luo et al., 2024; Mandrekar, 2010). Following the identification of the optimal threshold through ROC curve analysis, we further evaluate the classification performance using macro-averaged precision, recall, and f1-score. Unlike micro-averaging, which aggregates contributions from all classes and may be dominated by the majority class, macro-averaging computes metrics independently for each class and takes their unweighted mean. This approach is particularly suitable under class imbalance, as it ensures a more balanced and equitable assessment across categories. The corresponding formulas are presented below:

$$\text{Macro-Precision} = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FP_i} \tag{17}$$

$$\text{Macro-Recall} = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FN_i} \tag{18}$$

$$\text{Macro-F1} = \frac{1}{C}\sum_{i=1}^{C}\frac{2 \cdot P_i \cdot R_i}{P_i + R_i} \tag{19}$$

4.4. Baseline methods

To evaluate the performance of our method, we compare our model with a series of state-of-the-art models, including the ED_s described in Wang et al. (2023), as well as Novel^T and Novel^A proposed by Shibayama et al. (2021), and other methods proposed by Uzzi et al. (2013), Lee et al. (2015), Foster et al. (2015), Savov et al. (2020), Jeon et al. (2023) and Wang et al. (2017b). Among these baselines, ED_s, Novel^T, Novel^A, Savov et al. (2020), and the method proposed by Jeon et al. (2023) are content-based approaches, which rely on the content of publications to assess novelty. In contrast, Uzzi et al. (2013), Lee et al. (2015), Foster et al. (2015), and Wang et al. (2017b) are reference-based approaches, focusing on combination patterns of the journals to which the references in the focal paper belong.

5. Results and discussions

In this section, we first perform a sensitivity analysis by varying the control-to-treatment ratio from 1:1 to 6:1 to assess the robustness of our findings with respect to the number of matched non-award papers. Subsequently, we compare the performance of our proposed method with ten state-of-the-art baselines, conduct an in-depth analysis of the factors contributing to underperformance in several baselines, and, to further

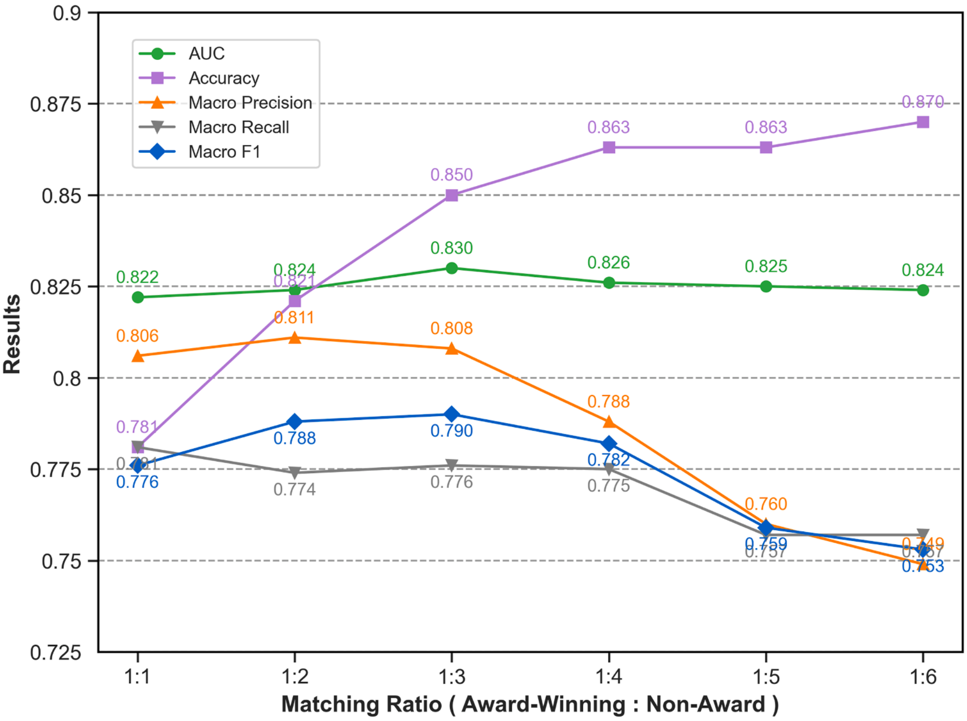


Fig. 6. Results of sensitivity analysis on matching ratios.

evaluate the effectiveness of the proposed novelty metric, perform statistical significance tests, including Tukey's Honest Significant Difference (HSD) test. Then, we conduct ablation experiments to assess the contribution of each module in our approach. Finally, we analyze and compare the distribution characteristics of knowledge in award-winning and non-award papers, including the quantity of knowledge, the number of knowledge combinations as well as the distribution characteristics of similarity and novelty. Furthermore, we conduct a more comprehensive analysis of the limited instances of unexpected novelty distribution observed in non-award papers.

5.1. Sensitivity analysis on the number of control papers

To evaluate the robustness of our findings with respect to the number of matched non-award papers, we conducted sensitivity analyses by varying the control-to-treatment ratio from 1:1 to 6:1. In addition to the evaluation metrics introduced in Section 4.3, we also assessed the accuracy, defined in Eq. 20, to comprehensively evaluate robustness. For each setting, non-award papers were drawn from the same conference and year as their award-winning counterparts, and all analyses followed the same procedures as in our main experiments. The model performance under varying control-to-treatment ratios is presented in Fig. 6. From Fig. 6, AUC remains essentially unchanged across matching ratios (hovering around 0.82), underscoring the model's stable discriminative power regardless of control-group size. Accuracy increases from 0.781 at a 1:1 ratio to 0.863 at 1:4, beyond which it plateaus (0.863 at 1:5 and 0.870 at 1:6). This suggests that increasing the number of controls up to four per award paper significantly improves overall performance, while further expansion yields only marginal benefits. Precision remains relatively stable at lower ratios but declines to 0.760 at 1:5 and 0.749 at 1:6, indicating that large control groups may introduce more false positives. Recall remains relatively stable as well, ranging from 0.781 (1:1) to 0.775 (1:4), before dropping slightly to 0.757 at both 1:5 and 1:6. The f1-score peaks at 0.790 at a 1:3 ratio, then gradually decreases to 0.782 (1:4), 0.759 (1:5) and 0.753 (1:6). These modest variations suggest that both recall and f1-score are only minimally affected by the matching ratio. Overall, aside from a slight precision decline at higher ratios, the other metrics show only minor fluctuations, indicating that model performance is largely robust to changes in control-group size. The 1:4 ratio, in particular, maintains strong performance across AUC, accuracy, recall, and f1-score, while avoiding a drop in precision, making it a balanced and reliable choice. Accordingly, we adopt the 1:4 ratio for all subsequent analyses in this study.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

5.2. Model comparison

Fig. 7 presents the ROC curves of various baseline models in distinguishing between award-winning and non-award papers. The curves reveal notable variations in model performance. Among them, our proposed approach achieves the highest AUC of 0.826, outperforming all other baselines and demonstrating its effectiveness in quantifying novelty for identifying award-winning papers. In contrast, $Novel^T$ yields the lowest AUC of just 0.49, highlighting the limitations of relying solely on paper titles for novelty detection. Moreover, the ROC trends indicate that our method achieves a relatively high true positive rate (TPR) in the low false positive rate (FPR) region, effectively capturing award-winning papers with minimal false alarms. Other models exhibit a more gradual increase in TPR. While Savov's model attains a slightly higher TPR in the high-FPR region, suggesting its ability to identify more award-winning papers at the cost of increased false positives, our method maintains a clear overall advantage. Despite this localized improvement, Savov's model records a lower overall AUC, reinforcing the superior robustness and effectiveness of our approach across the entire evaluation spectrum. Based on the optimal threshold derived from ROC curve analysis, we further computed macro-averaged precision, recall, and f1-Score for each method, as summarized in Table 3. The proposed hybrid methods consistently outperform traditional content-based and reference-based baselines. Among different configurations, methods using prompt-1 (P1) generally outperform those employing prompt-2 (P2). Furthermore, regarding language model choice, approaches leveraging GPT-4o for knowledge extraction surpass those based on OLMo2:13b. For knowledge representation, mod-

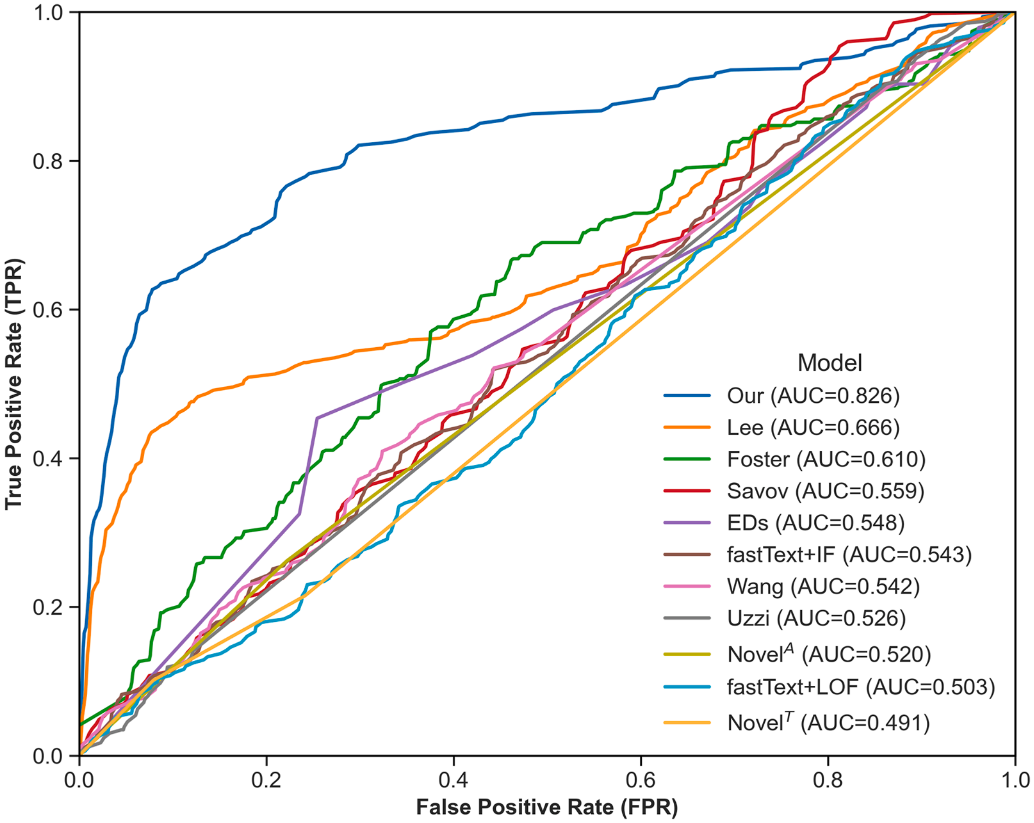


Fig. 7. ROC curves of various baseline models.

Table 3
Evaluation results of different measures.

Category	Metrics	Macro Precision	Macro Recall	Macro F1-Score
Content-based Methods	ED_s (Wang et al., 2023)	0.5938	0.6240	0.5981
	FastText + LOF (Jeon et al., 2023)	0.5209	0.5105	0.5011
	FastText + IF (Jeon et al., 2023)	0.5263	0.5258	0.5261
	Novel^T (Shibayama et al., 2021)	0.5231	0.5110	0.5006
	Novel^A (Shibayama et al., 2021)	0.5181	0.5200	0.5184
	Savov et al. (2020)	0.5232	0.5301	0.5199
Reference-based Methods	Wang et al. (2017b)	0.5291	0.5269	0.5276
	Foster et al. (2015)	0.5834	0.5666	0.5716
	Lee et al. (2015)	0.7274	0.6798	0.6976
	Uzzi et al. (2013)	0.5156	0.5106	0.5072
Hybrid Methods	P1 + GPT-4o + SciDeBERTa-cs *	**0.7883**	**0.7753**	**0.7815**
	P2 + GPT-4o + SciDeBERTa-cs	0.7053	0.6857	0.6943
	P1 + GPT-4o + SciBERT	0.7444	0.7764	0.7579
	P2 + GPT-4o + SciBERT	0.6733	0.7205	0.6873
	P1 + OLMo2:13b + SciDeBERTa-cs	0.7288	0.6925	0.7072
	P2 + OLMo2:13b + SciDeBERTa-cs	0.6815	0.6514	0.6631

⋆ indicates our main method, which uses prompt-1, GPT-4o and SciDeBERTa(CS) as the default configuration. The detailed prompt-1 (P1) and prompt-2 (P2) are provided in Appendix A.

els incorporating SciDeBERTa-cs yield better performance than those using SciBERT. Notably, the combination of P1 + GPT-4o + SciDeBERTa-cs achieves the highest macro-averaged f1-score of 0.7815, underscoring the benefits of leveraging a more powerful large language model for knowledge extraction and a domain-adaptive language model for knowledge representation initialization. These results suggest that both model selection and prompt design play critical roles in enhancing the performance of hybrid architectures for classification tasks.

To better understand the relatively poor performance of several baseline methods, we conduct an in-depth analysis of their underlying design assumptions, representational granularity, and the characteristics of the evaluation dataset. Rather than attributing the suboptimal results solely to the dataset, we argue that these outcomes reflect inherent methodological limitations when such approaches are applied to a domain-specific and topically homogeneous corpus.

For example, models such as Uzzi et al. (2013) and Wang et al. (2017a) rely on journal-level features to assess novelty. The method proposed by Uzzi et al. (2013) identifies novel papers based on rare journal co-citation patterns using a fixed global percentile, while Wang et al. (2017a) constructs vector representations of journals and quantifies novelty through cumulative divergence in the reference journal space. While effective in large-scale, interdisciplinary corpora with diverse citation behaviors, these approaches are less suited to our dataset–composed of top-tier AI conference papers–where referenced venues are

Table 4
Significance testing of novelty differences between award-winning and non-award papers.

				t-test			Mann–Whitney *U* test			
Variable	Group	Mean	Std.	*t*	df	*p*	Mean Rank	*U*	*Z*	*p*
Novelty	non-award	7.63	2.68	−16.49	277.50	< 0.001	515.73	39088.5	−15.54	< 0.001
	award-winning	12.75	4.59				902.07			

highly concentrated and journal overlap across papers is substantial. This lack of diversity reduces the discriminative power of journal-level novelty signals and limits these models' ability to identify genuinely novel contributions.

Similarly, methods such as FastText + LOF and FastText-IF (Jeon et al., 2023) also performed poorly. These approaches represent each paper using vectors derived solely from titles and apply anomaly detection techniques to identify outliers in the embedding space. However, paper titles–especially in AI–often use generic phrasing that fails to capture the conceptual specificity of the work. Additionally, the embeddings were trained on general-purpose corpora (e.g., Wikipedia), which inadequately reflect the nuanced semantics of scientific language in specialized domains. This mismatch likely contributed to the low expressiveness and poor separability of the resulting vectors.

Likewise, the Novel^T and Novel^A methods (Shibayama et al., 2021), which compute the average vector of the titles (T) or abstracts (A) of cited references, also suffer from coarse representational strategies. Such pooling techniques tend to overemphasize common functional words (e.g., "approach," "result," "method") and underrepresent semantically distinctive content. In a domain like AI–where many papers share similar lexical patterns–this leads to homogenized representations that obscure meaningful novelty.

The topic-based method proposed by Savov et al. (2020) also faces structural limitations. It models novelty through divergence in topic distributions–a strategy that assumes topical shifts are the primary signal of novelty. However, the AI field exhibits a relatively stable thematic landscape over time, with enduring themes such as classification, generation, and optimization. As a result, true novelty in this domain often stems from methodological advances within existing topical boundaries–a nuance that topic models are generally ill-equipped to detect.

Interestingly, the ED$_s$ method (Wang et al., 2023), which measures novelty via the imbalance between new and inherited knowledge units, demonstrated comparatively better performance. This may be attributed to its finer-grained approach, which analyzes the presence or absence of specific knowledge elements in the focal paper relative to its references. Such knowledge-level representations appear more effective for capturing localized novelty in specialized, non-interdisciplinary research domains.

Taken together, these findings suggest that coarse-grained or generalized models–while effective in broad, heterogeneous datasets–struggle to capture novelty in tightly focused, technically consistent domains. Approaches that leverage finer-grained representations or incorporate domain-specific mechanisms are likely to yield more reliable and meaningful assessments of scientific novelty in such contexts.

To further assess the effectiveness of the proposed novelty metric generated by our model, we conducted statistical significance tests to evaluate whether the scores meaningfully differentiate between award-winning and non-award papers. Before performing these tests, we examined the normality of the predicted novelty scores for both groups. The scores for award-winning papers passed the Shapiro-Wilk test ($p > 0.05$), indicating normal distribution, while those for non-award papers showed a significant deviation from normality ($p < 0.001$). However, given the relatively large sample size of the non-award group ($n = 948$), the Central Limit Theorem justifies that the sampling distribution of the mean approximates normality. Visual diagnostics from SPSS, including histograms and normal Q-Q plots, further supported this assumption. In particular, the Q-Q plot for the non-award group (see Appendix B) showed that most points closely followed the diagonal line, with only minor deviations in the tails. To further ensure the robustness of our findings in light of the non-normal distribution in the non-award group, we supplemented Welch's *t*-test with the non-parametric Mann-Whitney *U* test. As shown in Table 4, both tests yielded statistically significant results ($p < 0.001$), with award-winning papers exhibiting significantly higher novelty scores. These findings provide strong evidence for the discriminative validity of the proposed novelty metric.

Moreover, to assess the validity of summing pairwise knowledge novelty scores, as proposed in Eq. 14, and determine whether the aggregation is susceptible to outliers, we applied Tukey's Honest Significant Difference (HSD) Test. Tukey's Test is a statistical method used to identify significant differences between group means and to detect outliers in a dataset. In our study, we used this method to identify pairwise novelty scores that significantly deviate from the others, indicating potential outliers. For each paper, we calculated the proportion of outliers among the pairwise novelty scores. The distribution of these proportions is shown in Fig. 8, where the median is 0.00, indicating that most papers have negligible outlier proportions. The mean outlier proportion is 0.07, which reflects the slight influence of a few papers with higher outlier proportions. The Top 5 % cutoff is 0.21, indicating that only the top 5 % of papers have outlier proportions exceeding this threshold. Almost all outlier proportions were below 0.25, demonstrating that the vast majority of papers are not significantly affected by extreme values. This confirms that the aggregation method remains robust, as the novelty metric is not disproportionately influenced by a small number of outliers. Additionally, all pairwise novelty values are normalized to the range [0, 1], which inherently limits the extent of any individual score's influence, further ensuring the stability of the metric.

5.3. Ablation experiment

To assess the contribution of each module to the overall performance of our model, we conducted two sets of ablation experiments. First, we evaluated the impact of removing the knowledge propagation module on model performance. Based on this ablated variant, we further replaced SciDeBERTa(CS), a domain-specific language model trained on computer science literature, with the general-purpose BERT in the knowledge representation module to examine the performance gains provided by domain-specific language models. In addition to module-level ablations, we also evaluated the model's performance across different subsets of the dataset. Each subset corresponds to a top-tier AI conference, and this analysis aims to assess whether the model maintains consistent identification effectiveness across these conferences.

Table 5 presents the results of the ablation experiments and journal-wise analysis. Here "only award" refers to evaluation metrics calculated solely for the award category (positive class), whereas "macro" accounts for both award (positive class) and non-award (negative class) categories by averaging the metrics across both classes. In the first ablation, removing the knowledge propagation module led to a decrease in precision for the award category from 0.6682 to 0.5475, and a drop in F1-score from 0.6478 to 0.6162. This indicates that the knowledge propagation module plays a pivotal role in improving both precision and overall model performance. Although recall increased slightly from 0.6287 to 0.7046, the gain was insufficient to compensate for the decline in precision. Additionally, macro precision and macro f1-score decreased

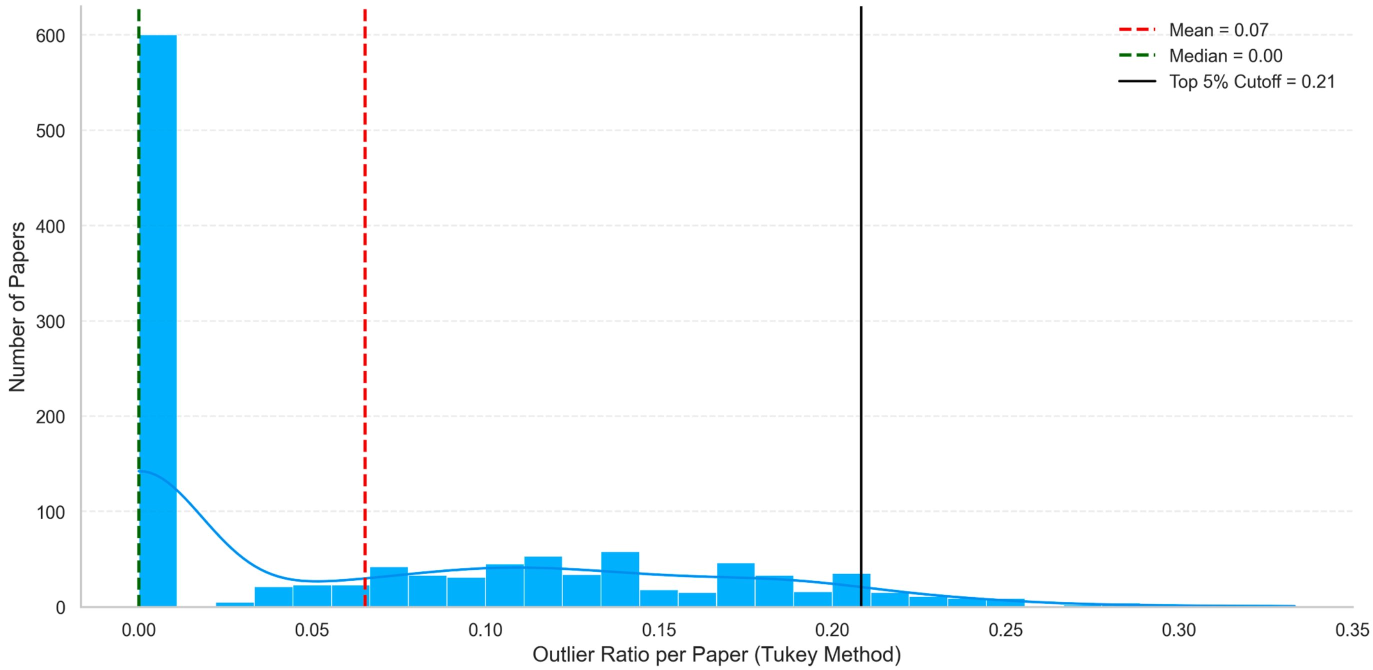


Fig. 8. Distribution of outlier ratios across papers.

Table 5
Performance evaluation results from the ablation study and journal-wise analysis.

	only award			macro		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Full Model	**0.6682**	0.6287	**0.6478**	**0.7883**	0.7753	**0.7815**
w/o knowledge propagation	0.5475	**0.7046**	0.6162	0.7340	**0.7795**	0.7512
SciDeBERTa-cs → Bert	0.5093	0.6962	0.5882	0.7128	0.7642	0.7303
AAAI	**0.7273**	0.5854	0.6486	**0.8142**	0.7652	0.7856
ACL	0.6750	0.7714	**0.7200**	0.8079	**0.8393**	**0.8218**
CVPR	0.5405	0.7692	0.6349	0.7380	0.8029	0.7591
ICCV	0.4186	**0.9000**	0.5714	0.6918	0.7938	0.6872
ICML	0.4198	0.8293	0.5574	0.6817	0.7713	0.6849
IJCAI	0.5088	0.8056	0.6237	0.7259	0.8056	0.7463
NeurIPS	0.5263	0.7895	0.6316	0.7331	0.8059	0.7544

by approximately 0.0543 and 0.0303, respectively, suggesting that the knowledge propagation module contributes not only to the prediction of award-winning papers but also to balanced classification across both categories.

In the second ablation, replacing the domain-specific SciDeBERTa(CS) with the general-purpose BERT in the knowledge representation module resulted in a substantial drop in precision, recall, and f1-score for the award category, falling to 0.5093, 0.6962, and 0.5882, respectively. This finding highlights the superiority of SciDeBERTa(CS) over general-purpose language models in capturing domain-specific semantic representations, particularly in the computer science fields. Additionally, the macro precision and macro f1-score also declined by approximately 0.02, further reinforcing the importance of domain-specific language models in enhancing overall performance.

Third, beyond the aforementioned module-level ablation studies, we conducted a subgroup analysis to further evaluate the robustness of the proposed model across different top-tier AI conferences. This analysis aimed to determine whether the model maintains consistent effectiveness when applied to different distributions of scientific literature. In addition to the numerical results summarized in Table 5, we further provide a visualization of the performance variation across conferences in Fig. 9. Notably, under both “only award” and “macro” metrics, recall consistently remains relatively high and exceeds precision across all subgroups, except for the AAAI conference, where precision slightly surpasses recall. This pattern suggests that the model demonstrates strong sensitivity in identifying potentially innovative papers. However, this heightened sensitivity is accompanied by a tendency to assign high novelty scores to some non-award papers, a trade-off that is more pronounced under the “only award” setting. In contrast, the “macro” metrics yields more balanced performance. Importantly, the consistency of these trends across conferences indicates that the model maintains stable performance across diverse conferences.

In summary, all modules contribute meaningfully to the model’s effectiveness: the knowledge propagation module enhances prediction accuracy, while the domain-specific knowledge representation module improves semantic understanding in specialized domains. Furthermore, the model demonstrates robust and consistent performance across different conference subgroups, thereby validating rationality and effectiveness of the proposed modular architecture.

5.4. Analysis and comparison of the characteristics of award-winning and non-award papers

5.4.1. Statistical distribution of knowledge and knowledge combinations

To further analyze the differences in the statistical distribution of knowledge and knowledge combinations between award-winning and non-award paper, as well as to characterize their distributional differences, we performed a statistical comparison of the number of knowledge extracted through Section 3.1 and the number of knowledge combinations identified via Section 3.2. As illustrated in Fig. 10, the figure presents the distributional differences in the number of knowledge and knowledge pairs between award-winning and non-award papers.

We first examine the distribution of knowledge count in award-winning and non-award papers. Here, Knowledge count refers to the number of distinct knowledge units addressed in a paper, which reflects the breadth of the study. This measure represents the extent of coverage of existing knowledge and the richness of the theoretical background. As shown in the violin plots, award-winning papers (red) exhibit significantly higher knowledge counts than non-award papers (blue), with differences evident across multiple distributional aspects. First, the median

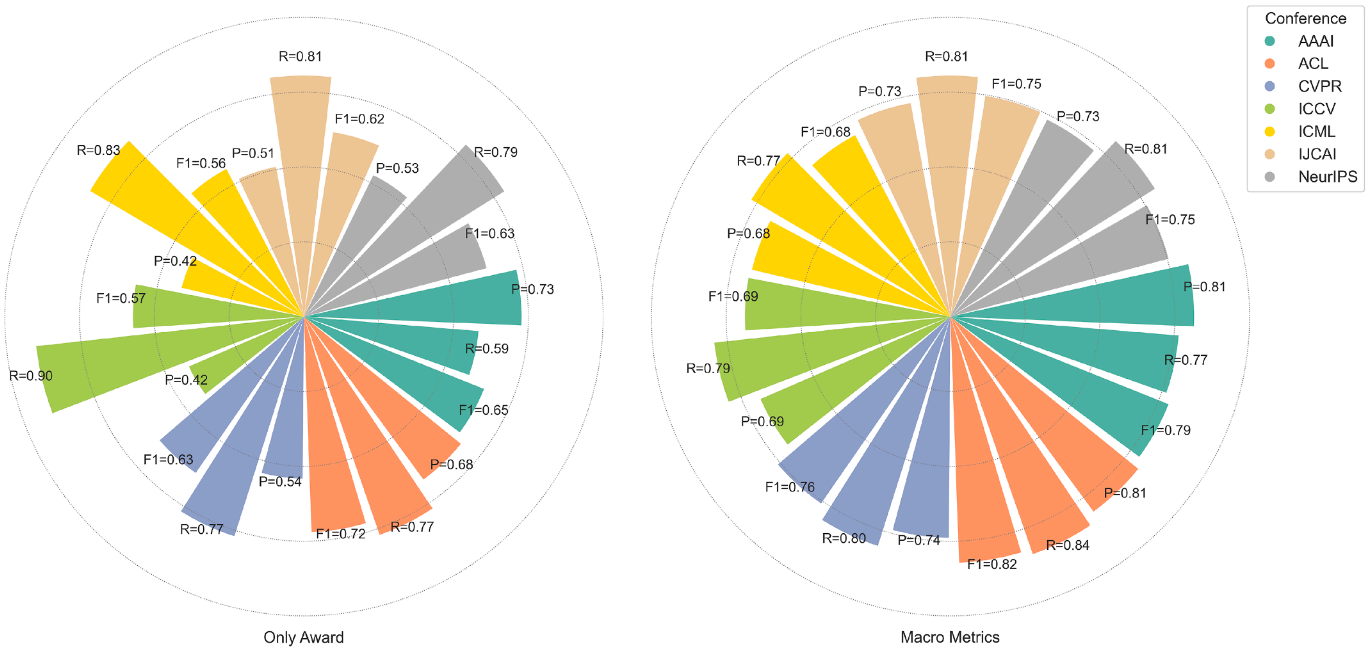


Fig. 9. Visualization of metric variations across journals.

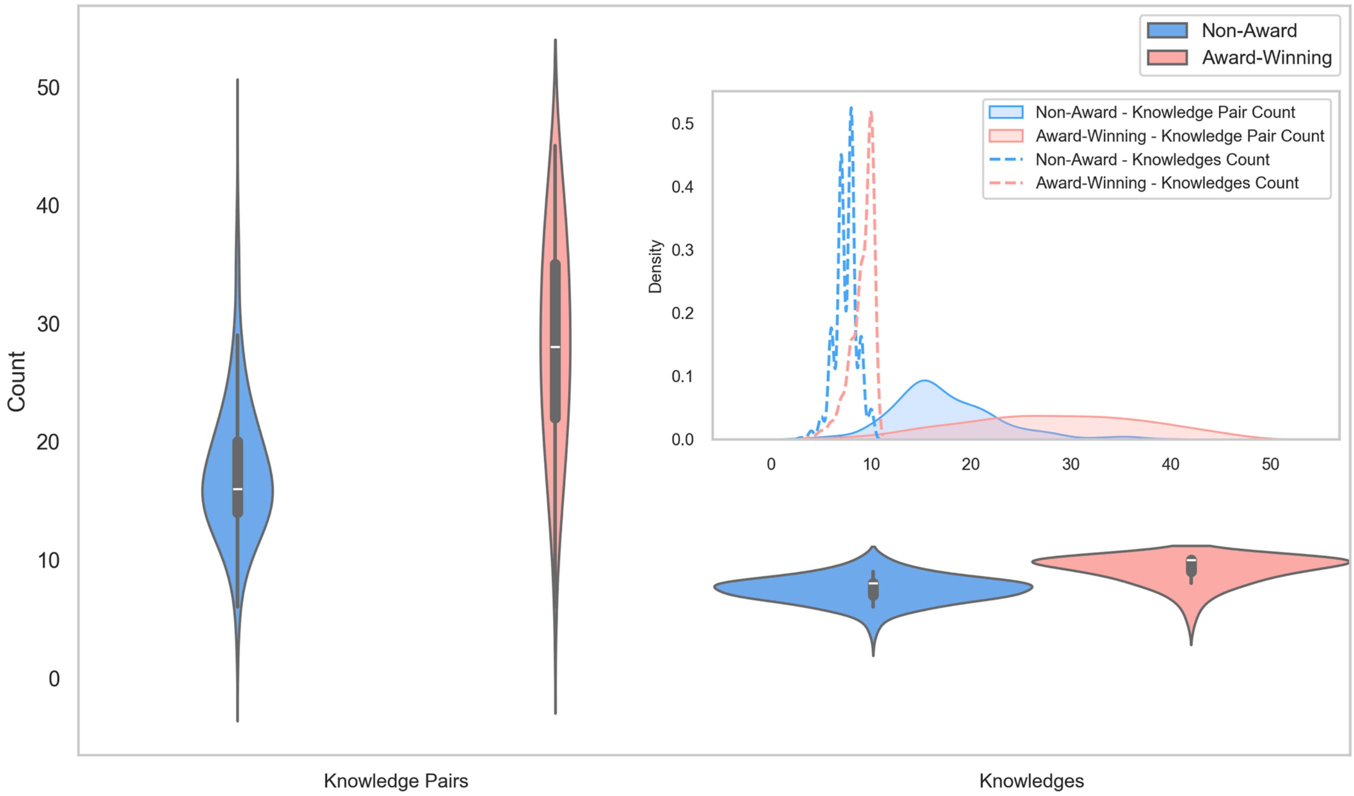


Fig. 10. Distribution of knowledge and knowledge pairs: non-award vs. award-winning papers.

knowledge count is notably higher in award-winning papers, suggesting that these works typically incorporate a broader range of independent knowledge units. Second, the distribution of knowledge counts in award-winning papers is more balanced and displays an extended upper tail, whereas that of non-award papers is skewed toward lower values, with few instances of high knowledge counts. The probability density curves further reveal that non-award papers exhibit a pronounced peak in the low knowledge count range (blue dashed line), reflecting a relatively limited coverage of knowledge in most cases. Conversely, the density curve for award-winning papers (red dashed line) is flatter and possesses a significantly longer tail, indicating that a majority of award-winning papers include a large number of knowledge units, suggesting a richer and more diverse knowledge framework and potentially broader research domains.

Papers with a higher knowledge count typically indicate that they address research questions by involving multiple related fields and may even integrate knowledge from different disciplines. This multidisciplinary integration renders the research content more comprehensive and robust, thereby facilitating deeper exploration based on existing knowledge. In contrast, papers with a lower knowledge count tend to focus on a specific issue within a relatively narrow knowledge scope, which may restrict the overall breadth of the study and subsequently influence the paper's overall academic novelty.

The above findings demonstrate that award-winning papers generally contain a higher number of knowledge units with a broader distribution, whereas non-award papers exhibit relatively fewer knowledge units, with their distribution concentrated in the lower range. The quantity of knowledge incorporated in a paper may influence its novelty level, as a higher knowledge count potentially provides stronger theoretical support, a richer research background, and even a higher degree of interdisciplinary integration.

We next analyze the distributional characteristics of knowledge pair counts. Knowledge pair count denotes the number of associations forged among distinct knowledge units within a paper. This metric not only reflects the depth of the investigation but also the extent of interconnection among its constituent knowledge. Analysis of the violin plots reveals that although the overall range (i.e., maximum and minimum values) of knowledge pair counts for award-winning and non-award papers is relatively similar, the distribution in award-winning works demonstrates a pronounced superiority.

In particular, the median knowledge pair count in award-winning papers is markedly higher than that observed in non-award counterparts. This finding suggests that such papers not only reference a greater array of knowledge but also explore their interrelations more thoroughly, thereby cultivating more intricate knowledge networks. Furthermore, a careful inspection of the violin plot widths indicates that the knowledge pair counts in award-winning papers are predominantly situated in the higher range, whereas those in non-award papers are largely confined to the mid to lower echelons, with extreme values appearing less frequently. The corresponding probability density curves further reveal that non-award papers exhibit a pronounced peak at the lower end, signifying a limited degree of knowledge integration. In contrast, the density curve for award-winning papers is flatter with an elongated tail, suggesting that certain award-winning works achieve a highly complex and tightly interwoven knowledge structure.

In summary, the probability density curves imply that non-award papers tend to combine relatively few knowledge, indicative of a weaker interconnection and a lack of deep integration. Conversely, the flatter density curve and extended tail characteristic of award-winning papers reflect significantly higher knowledge pair counts, emblematic of more sophisticated integration and the potential emergence of tightly interconnected knowledge networks. This comparative analysis of knowledge units and knowledge pairs intimates that award-winning papers not only encompass a greater number of knowledge units but also

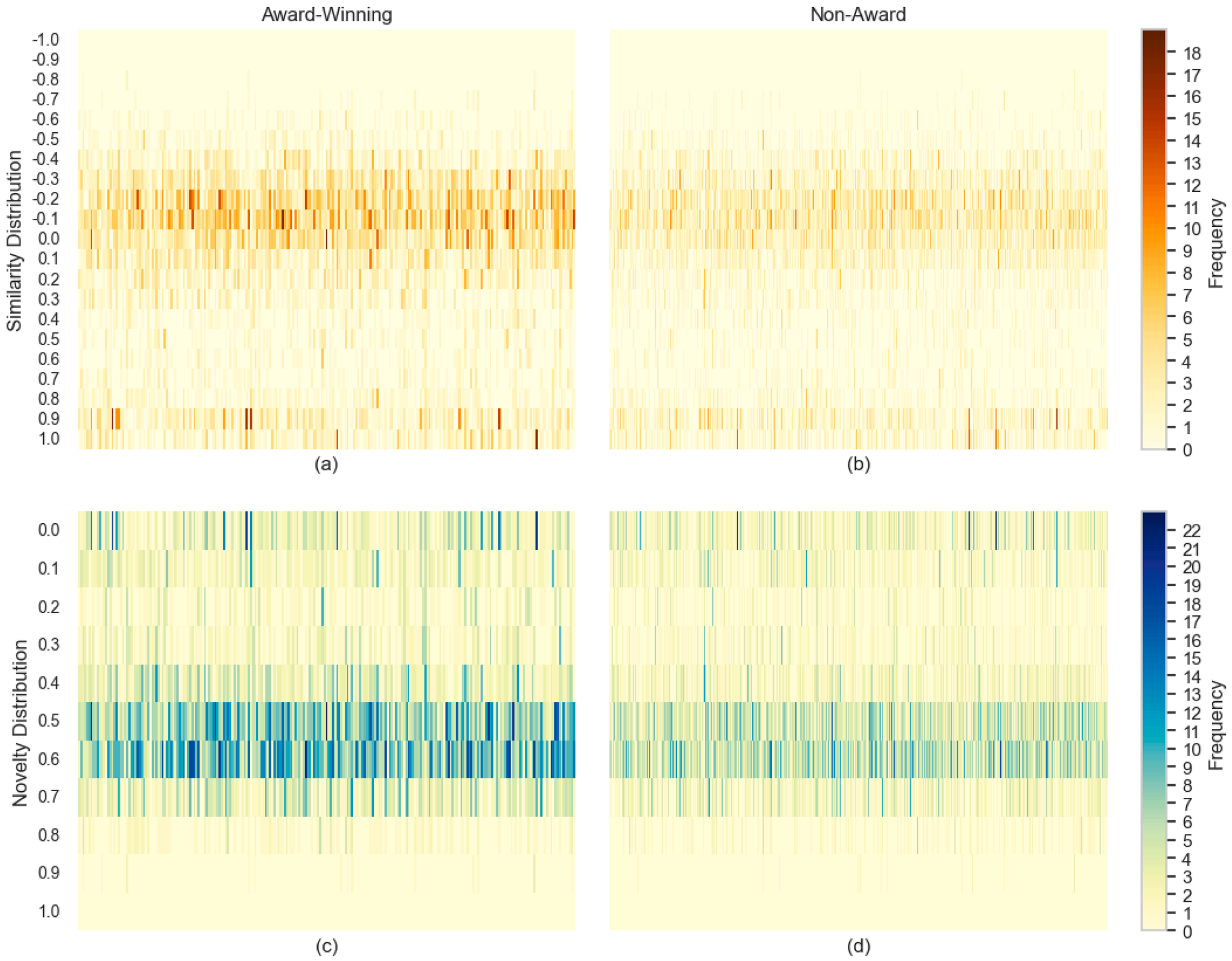


Fig. 11. The heatmap of similarity and novelty distributions of award-winning and non-award papers.

exhibit richer interrelationships among them. Such comprehensive knowledge associations may pave the way for breakthroughs in theoretical contributions and academic novelty, as novel insights often arise from the recombination and profound synthesis of existing knowledge. In contrast, the relatively knowledge pair counts in non-award papers suggest a tendency to remain entrenched in established paradigms, with insufficient cross-disciplinary integration, which may, in turn, limit their overall novelty and scholarly impact.

5.4.2. *Characteristics of knowledge combination novelty: award-winning vs non-award papers*

To compare the distributional characteristics of knowledge combination novelty between award-winning and non-award papers, we constructed heatmaps illustrating their similarity and novelty distributions of knowledge combinations, as shown in Fig. 11. In these heatmaps, each vertical column represents a paper, each horizontal row corresponds to a specific similarity or novelty value, and each small cell denotes the count of knowledge combinations within the paper that exhibit a given similarity or novelty value. The darker the color, the greater the number of knowledge combinations with that particular similarity or novelty value. In terms of distribution range, the similarity scores of award-winning papers are primarily concentrated between −0.4 and 0.2, while their novelty scores are predominantly distributed between 0.5 and 0.8.

From the color intensity distribution depicted in Fig. 11(a), it is evident that the similarity scores of intra-paper knowledge combinations in award-winning papers are predominantly concentrated within the range of -0.4 to 0.2, forming a high-frequency interval. This observation indicates that most of the knowledge combinations within these papers demonstrate relatively low similarity. Such heterogeneity enhances the overall novelty of the paper, reflecting a higher level of originality. In contrast, the color distribution in Fig. 11(b) appears lighter. Although some non-award papers include knowledge combinations with low similarity, the overall distribution is more dispersed and uniform, without a distinct high-frequency interval. This observation suggests that the similarity scores of knowledge combinations in non-award papers are more evenly distributed. While some combinations exhibit low similarity, a substantial proportion reveals relatively high similarity, implying that the novel contributions of these papers are primarily reflected in incremental improvements or applications rather than radical breakthroughs (Veugelers & Wang, 2019). This pattern is further reflected in Fig. 11(c) and (d), where the color intensity in the high-frequency region of the novelty score is noticeably deeper for award-winning papers compared to non-award papers, particularly within the range of 0.5 to 0.8. Overall, most knowledge combinations in award-winning papers exhibit significant heterogeneity, as evidenced by their relatively low similarity scores. This may indicate that these papers contain a larger number of highly novel knowledge combinations. Conversely, the similarity distribution in non-award papers is more balanced, potentially indicating that these works contribute to various fields through incremental advancements, albeit without achieving substantial radical breakthroughs.

In Fig. 11, we observe that a small subset of non-award papers exhibits a lower similarity distribution in their knowledge combinations, with cells in the lower similarity score range appearing darker. To facilitate a more comprehensive analysis of this unexpected non-award novelty distribution, we selected non-award papers characterized by low similarity, and analyzed these unexpected distribution features, as presented in Table 6.

Table 6
Top 20 non-award papers by novelty and similarity (median).

Paper Index	Similarity (Median)	Similarity (Average)	Combination Number	Influential Citation Count	Citation Count	Paper Novelty
532	-0.0547	0.0137	45	27	440	22.1919
482	-0.0779	0.0055	41	188	926	20.3876
632	-0.1174	-0.0192	39	15	571	19.8749
554	-0.0983	-0.0051	37	67	1345	18.5951
678	-0.0884	0.1101	40	69	1168	17.7990
717	-0.1299	0.0276	36	69	341	17.5033
550	-0.1441	-0.0217	34	165	2668	17.3683
319	0.0051	0.1252	39	85	374	17.0583
1040	-0.1083	0.0384	35	116	575	16.8279
649	-0.1042	-0.0161	33	172	1046	16.7650
386	-0.1174	0.0692	36	1	15	16.7548
336	-0.0559	0.0813	34	48	279	15.6172
703	-0.1365	0.0468	32	54	416	15.2517
623	-0.1649	0.0772	33	86	757	15.2266
311	-0.1511	-0.0124	30	110	674	15.1861
624	0.0491	0.1599	36	65	667	15.1222
263	0.0212	0.1360	35	58	300	15.1207
589	-0.1758	-0.0047	30	32	287	15.0700
911	-0.0207	0.0885	33	146	640	15.0390
626	-0.2388	-0.0700	28	26	352	14.9800

Table 6 presents the distribution of knowledge combination similarity for non-award papers with lower similarity scores and higher novelty scores. We ranked the papers based on novelty and median of similarity and selected the top 20 non-award papers, analyzing their statistical characteristics in terms of knowledge combination similarity, citation impact, and novelty. The results indicate that these papers exhibit consistently lower similarity(median), and similarity(average) compared to the typical non-award papers, suggesting a greater degree of novelty in their research content and methodology. Despite not receiving awards, many of these papers still garnered substantial total citation counts and influential citations, demonstrating their academic impact. For instance, the paper indexed as 532 received 27 highly influential citations and a total of 440 citations, indicating that although it did not win an award, it has been widely recognized within the research community over time. The case of paper 532 is not an isolated one. This compelling paradox, where papers demonstrate high novelty and achieve significant long-term impact without receiving formal awards, is not a random occurrence. It points to specific, systematic mechanisms within the academic evaluation process. To reveal these underlying mechanisms, we conducted a deeper analysis of the representative cases from Table 6, interpreting them through the lens of established well-known theories and research findings. Our analysis suggests that a combination of cognitive and temporal mechanisms contributes to this discrepancy.

1. A paper's success in the award selection process is shaped not only by its inherent novelty, but also by the alignment between its contribution type and the prevailing evaluation focus of the venue. A lack of such alignment can lead to undervaluation, particularly for papers whose primary contribution lies in foundational theory or in directions not yet recognized as central to the field. A representative example is Paper 532, "Robust Principal Component Analysis for Computer Vision," published at the International Conference on Computer Vision (ICCV), a conference that historically emphasizes empirical advances in vision-related tasks. The paper proposed Robust PCA (RPCA) to address critical limitations of traditional PCA in the presence of outliers. Although its contribution is a fundamental advancement in statistics and machine learning, its research direction may have been perceived by a specialized computer vision award committee as less compelling than papers presenting state-of-the-art results on a canonical visual task.
2. Another key mechanism concerns the temporal disconnect between point-in-time evaluations and the unfolding of long-term scholarly influence. Award decisions are, by nature, a snapshot prediction of future importance made by a small committee under tight deadlines. This process is inherently better at recognizing work with immediately apparent utility or applications. In contrast, foundational or pioneering ideas may require time to diffuse before their significance is broadly recognized–a dynamic captured by the "Sleeping Beauty" phenomenon (van Raan, 2004). This is exemplified by Paper 554, “Recognizing Action at a Distance,” which received no formal recognition at the time of publication. However, as interest in action understanding and long-range video modeling grew, the paper's

Table 7
Novelty and knowledge combination of sample papers.

Awarded	Title	Knowledge Combination(Top 10 by novelty)
Yes	Densely Connected Convolutional Networks	0.72(benchmark tasks, vanishing-gradient problem); 0.68(convolutional networks, feature propagation); 0.67(benchmark tasks, feature-maps); 0.65(DenseNet, benchmark tasks); 0.64(DenseNet, connections); 0.64(Dense Convolutional Network, convolutional networks); 0.64(connections, convolutional networks); 0.63(convolutional networks, feature-maps); 0.63(benchmark tasks, feature reuse)
Yes	Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting	0.56(Long sequence time-series forecasting, memory usage); 0.53(Informer, Transformer); 0.52(Long sequence time-series forecasting, prediction capacity); 0.52(Long sequence time-series forecasting, time complexity); 0.52 (Transformer, self-attention distilling); 0.51(Transformer, inference speed); 0.51(Transformer, memory usage); 0.51(Transformer, memory usage); 0.50(Transformer, prediction capacity); 0.50(Transformer, generative style decoder); 0.49(ProbSparse self-attention, Transformer)
No	Anytime Approximate Modal Reasoning	0.53(anytime proof procedure, approximation method); 0.47(approximation method, credulous approximations); 0.46(approximation method, multi-modal logics); 0.44(approximation method, classical modal tableaux); 0.41(approximation method, unbounded logical introspection); 0.30(approximation method, unbounded logical omniscience); 0.11(unbounded logical introspection, unbounded logical omniscience); 0.09(credulous approximations, unbounded logical omniscience); 0.07(multi-modal logics, unbounded logical omniscience); 0.06(classical modal tableaux, quality guarantees);
No	Anonymization for Skeleton Action Recognition	0.64(machine learning, security/privacy); 0.30(hyperparameter relaxations, protection mechanisms); 0.28(differentially private training, hyperparameter relaxations); 0.24(model and data ownership verification, robustness against model evasion); 0.22(conflicting interactions, model and data ownership verification); 0.17(model and data ownership verification, protection mechanisms); 0.13(differentially private training, robustness against model evasion); 0.10(conflicting interactions, differentially private training); 0.06(differentially private training, model and data ownership verification); 0.06(differentially private training, protection mechanisms)

relevance became increasingly apparent, leading to a substantial surge in citations several years later–ultimately exceeding 1,300. This case illustrates how long-term community validation can diverge markedly from initial judgments, highlighting a structural blind spot in the temporal scope of novelty evaluation.

5.5. Case study

To explore whether the proposed combinatorial-novelty score is interpretable to humans and consistent with expert judgements of originality, we performed a small-scale qualitative assessment. From the experimental dataset, we randomly drew four sample papers, stratified by prize status: two award-winning papers and two non-award papers. For each paper, we extracted the top 10 knowledge-unit pairs based on their novelty scores, displaying only the highest-ranking combinations according to novelty, as shown in Table 7. The first element of every tuple represents the novelty score, and the bracketed pair lists the two knowledge whose co-occurrence is being evaluated.

Densely Connected Convolutional Networks is widely recognized for three key innovations: (1) introducing a dense connectivity pattern where each layer connects to all previous layers, (2) alleviating the vanishing-gradient problem through improved gradient flow, and (3) enabling systematic feature reuse across the network. Our novelty assessment reflects these contributions through several high-scoring knowledge combinations. The pair (benchmark tasks, vanishing-gradient problem, 0.72) captures the paper's explicit link between training stability and evaluation performance. (Convolutional networks, feature propagation, 0.68) corresponds to DenseNet's emphasis on

Table 8
Conferences and the number of award-winning and non-award papers.

Conference Abbreviation	Conference Full Name	Award-Winning Papers Count	Non-Award Papers Count
MICCAI	Medical Image Computing and Computer-Assisted Intervention	20	1798
ISBI	International Symposium on Biomedical Imaging	18	3411
MLMI	Machine Learning in Medical Imaging	14	166
BHI	IEEE-EMBS International Conference on Biomedical and Health Informatics	9	662
AIME	International Conference on Artificial Intelligence in Medicine	13	293
SMC	IEEE International Conference on Systems, Man and Cybernetics	13	6718
AMIA	American Medical Informatics Association	81	8270
IPMI	Information Processing in Medical Imaging	12	683
CBMS	Symposium on Computer-Based Medical Systems	8	377
BIOSTEC	International Joint Conference on Biomedical Engineering Systems and Technologies	74	748
MeMeA	IEEE International Symposium on Medical Measurements and Applications	5	699
CHASE	IEEE/ACM International Conference on Connected Health: Cooperative and Human Aspects of Software Engineering	3	70
RECOMB	Research in Computational Molecular Biology	15	504
MIE	Medical Informatics Europe	5	1323
ICT4AWE	International Conference on Information and Communication Technologies for Ageing Well and e-Health	10	298
VCBM	Eurographics Workshop on Visual Computing for Biomedicine	8	125

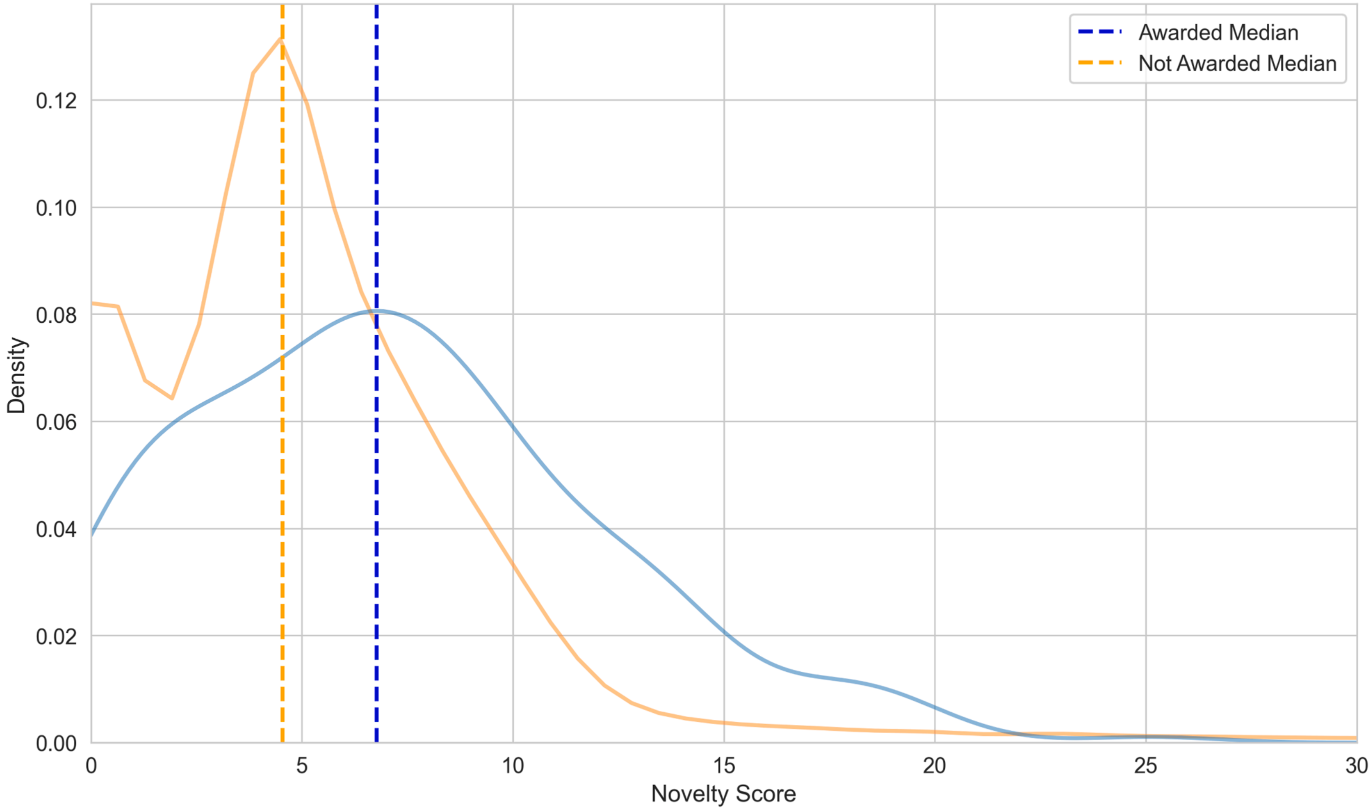


Fig. 12. Distribution of novelty score by award status of BME papers.

improving inter-layer information flow–an underexplored concern in prior CNNs. High scores for (Dense Convolutional Network, convolutional networks) and (connections, convolutional networks) further signal the conceptual shift introduced by dense connectivity. Finally, (benchmark tasks, feature reuse, 0.63) reflects the model's novel reuse mechanism tied to task-level gains. Informer introduces several innovations aimed at making Transformer-based architectures efficient and scalable for long sequence time-series forecasting. These include ProbSparse self-attention for reducing time and memory complexity, self-attention distilling for compressing intermediate representations, and a generative-style decoder for stable multi-step prediction without autoregressive roll-out. Our novelty assessment reflects these contributions through high-scoring concept pairs such as (long sequence time-series forecasting, memory usage, 0.56), (long sequence time-series forecasting, time complexity, 0.52), and (Transformer, self-attention distilling, 0.52). These combinations capture the paper's effort to integrate resource-aware optimization with architectural mechanisms–an alignment that is relatively rare in prior Transformer literature. Additionally, (Transformer, generative style decoder, 0.50) and (ProbSparse self-attention, Transformer, 0.49) reflect Informer's structural extensions to the standard Transformer framework. Together, these results suggest that our combinatorial novelty measure is able to surface meaningful conceptual linkages that correspond closely to the paper's expert-recognized innovations in both architectural design and task efficiency.

For the non-award papers, Anytime Approximate Modal Reasoning and Anonymization for Skeleton Action Recognition show several novel concepts but with relatively low novelty scores overall, indicating their contributions are more incremental in nature. For example, the pair (anytime proof procedure, approximation method, 0.53) and (machine learning, security/privacy, 0.64) suggest that these works explore important ideas, but the novelty of their combinations is not as pronounced as in the award-winning papers. The lower-scoring combinations (unbounded logical introspection, unbounded logical omniscience, 0.11) and (differentially private training, hyperparameter relaxations, 0.28) highlight that while these papers introduce new knowledge, the conceptual novelty is limited compared to the groundbreaking advances seen in DenseNet and Informer.

These findings validate that our combinatorial novelty measure is effective not only in capturing the significant innovations of high-impact papers like DenseNet and Informer but also in identifying papers with lower novelty, indicating incremental rather than transformative contributions.

5.6. Cross-field validation on biomedical engineering (BME) conferences

To evaluate the generalizability of the proposed method beyond the AI domain, we conducted an additional cross-field validation using data from the Biomedical and Medical Engineering domain,which differs from AI in terms of knowledge organization structures. A list of representative conferences was obtained from Research.com[4], which ranks venues based on Impact Score metrics. Award-winning papers were manually collected from publicly available sources, including official conference websites, academic blogs, and institutional announcements. Conferences or years without accessible award records were excluded. For each included conference-year pair, non-award papers were retrieved from DBLP[5] to construct the baseline dataset. Table 8 summarizes the composition of the final dataset.

The proposed method was initially evaluated on some conferences in the AI domain, which served as the primary experimental setting. These conferences are widely recognized as top-tier venues in the field, indicating a relatively uniform standard of paper quality across events. This homogeneity justified the use of a threshold-based evaluation strategy. However, conferences in the biomedical and medical engineering

[4] https://research.com/conference-rankings/computer-science/biomedical-bioinformatics

[5] https://dblp.org/db/conf/index.HTML

domain exhibit substantial disparities in their Impact Scores. reflecting greater variability in paper quality. Consequently, fixed-threshold comparisons are less appropriate in this context.

To address this issue, we adopted a distributional comparison approach, consistent with the evaluation framework introduced by Wang et al. (2024b). Their study demonstrated that high-novelty scientific articles tend to exhibit right-skewed distributions of novelty scores compared to those with lower novelty. Following this rationale, we hypothesize that award-winning papers in the biomedical and medical engineering domain should show a distribution of novelty scores shifted toward higher values relative to non-award papers. When visualized as probability density functions, this distinction is expected to manifest as a rightward shift in the score distribution for award-winning papers relative to non-award papers.

The empirical results confirm this expectation. As shown in Fig. 12, the probability density function of novelty scores for award-winning papers (blue curve) displays a clear rightward shift compared to non-award papers (orange curve). The distribution for non-award papers is more concentrated around lower novelty scores, whereas award-winning papers are associated with a wider and more right-skewed distribution, suggesting a tendency toward higher novelty. This difference is also reflected in the median scores, as indicated by the vertical dashed lines: the median novelty score of award-winning papers significantly exceeds that of non-award ones. These results provide empirical support for the effectiveness of our method in the Biomedical Engineering domain. Despite differences from the AI domain in research focus and content structure, the novelty metric remains capable of distinguishing between award and non-award papers in the Biomedical Engineering field, demonstrating its applicability across scientific fields.

6. Conclusion

This article proposes a method for measuring the novelty in research papers. We introduce an approach based on knowledge combinations and knowledge propagation, which consists of four sequential steps: knowledge extraction, reference knowledge co-occurrence network construction, knowledge propagation on reference knowledge co-occurrence network, and focal paper novelty computation. Experimental results based on a computer science conference paper dataset demonstrate the effectiveness of our method in quantifying the level of paper novelty.

We further conduct a multi-dimensional analysis and comparison of the characteristics of award-winning and non-award papers, including the distribution of knowledge quantity, the number of knowledge combinations, and the distribution patterns of knowledge combination similarity and novelty. Several key conclusions are drawn: (a) Award-winning papers generally incorporate more knowledge, while non-award papers contain fewer knowledge. A higher knowledge count may provide stronger theoretical support, a richer research background, and greater interdisciplinary integration, potentially enhancing novelty. (b) The combination of knowledge may influence a paper's novelty. Papers with more knowledge pairs tend to form interconnected knowledge networks, fostering novel insights through recombination. In contrast, papers with fewer knowledge pairs may adhere to established paradigms, limiting novelty. (c) The distribution of knowledge combination novelty in award-winning papers is uneven, primarily concentrated in the lower range, with considerable variation between the combined knowledge. In contrast, non-award-winning papers exhibit a more uniform distribution of knowledge combination novelty, without a distinct high-frequency range. While some combinations demonstrate high novelty, a substantial proportion of them also display lower novelty.

This study has several limitations. First, it relies on co-occurrence relationships as the basis for modeling associations between knowledge units. While this approach is partially effective, it falls short of capturing the more nuanced and fine-grained semantic relationships that often exist in scientific discourse. As a result, some conceptual connections may be oversimplified or overlooked. Second, the study evaluates novelty using a limited number of award-winning papers, which are selected annually by conferences or journals. This inherently narrow scope may exclude many highly novel but unrecognized contributions at the time of publication. This reflects a broader challenge in novelty detection: the lack of comprehensive and direct ground-truth data. The absence of universally accepted labels for novelty makes it difficult to evaluate models at scale with high fidelity.

To address these limitations, future research can explore more sophisticated techniques for modeling semantic relationships between knowledge units–moving beyond co-occurrence to incorporate contextual or causal links using advanced natural language processing or knowledge graph techniques. Additionally, we aim to investigate the mechanisms by which scientific novelty emerges, with the goal of constructing a larger-scale, theoretically grounded dataset. Rather than relying solely on indirect proxies such as awards or citation counts, this dataset would incorporate novelty indicators informed by conceptual and methodological criteria. Finally, while our current framework aggregates novelty scores through summation, future work could explore more expressive modeling approaches–such as neural architectures capable of learning the relative importance and interactions among knowledge components in a dynamic and context-aware manner. These directions have the potential to significantly enhance the granularity, accuracy, and interpretability of novelty detection models in scientific domains.

CRediT authorship contribution statement

Zhongyi Wang: Conceptualization, Methodology, Software, Writing - Original draft preparation; **Zeren Wang:** Conceptualization, Data curation, Methodology, Software, Data visualization, Writing - Original draft preparation; **Guangzhao Zhang:** Software, Writing - Original draft preparation; **Jiangping Chen:** Methodology, Writing - Review; **Markus Luczak-Roesch:** Methodology, Writing - Review; **Haihua Chen:** Methodology, Writing - Original draft preparation and Review, Project Management.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve the language and grammar of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Data Availability

Data is available at https://github.com/haihua0913/graphLLM4ScientificNovelty.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Zhongyi Wang reports article publishing charges was provided by Fundamental Research Funds for the Central China Normal University. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Prompts used in the experiments

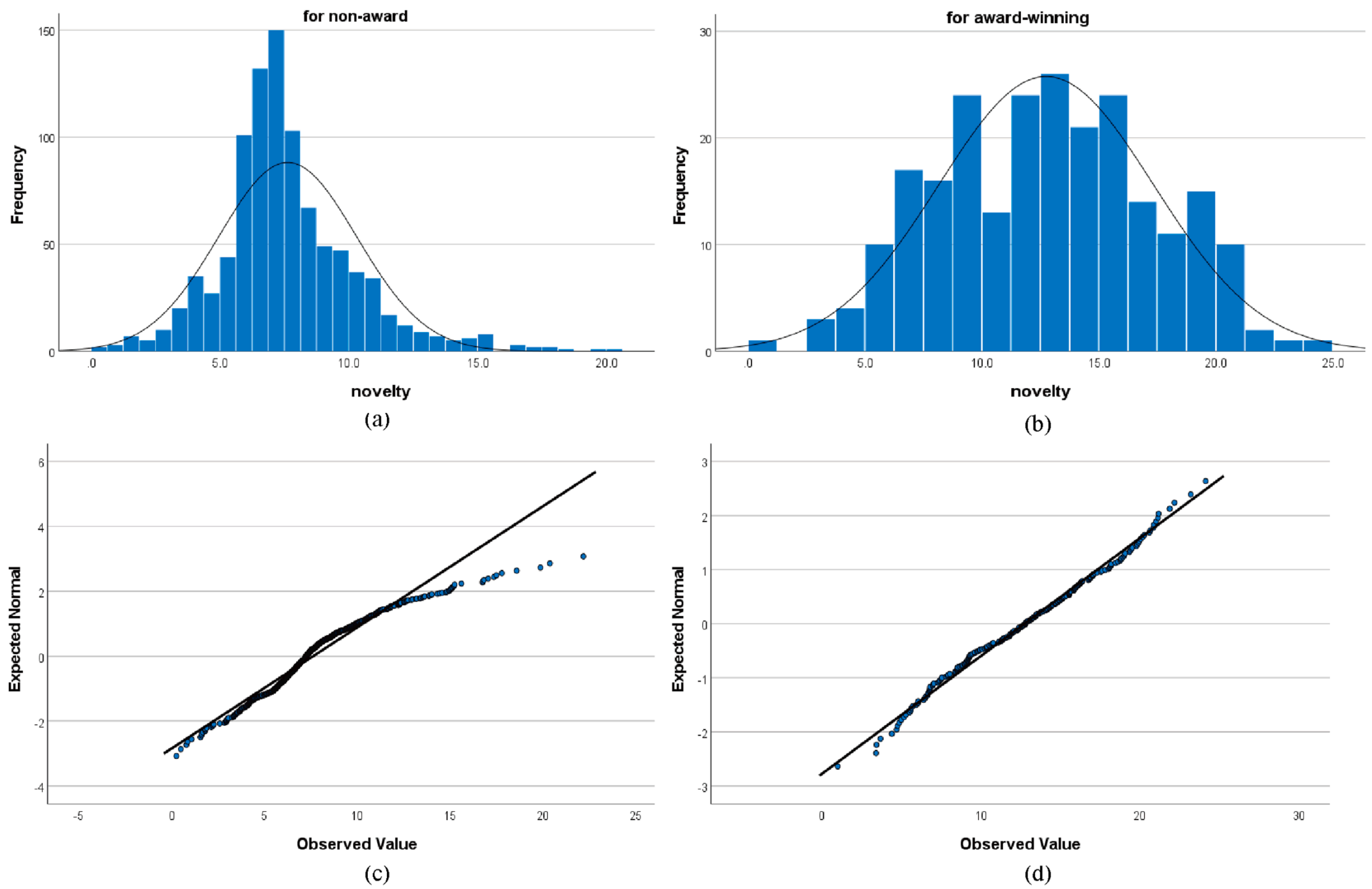
Prompts used in the experiments are shown in Table A.9, where prompt-1 includes additional instructions specifically constraining knowledge extraction.

Table A.9
Prompts used in the experiments.

Prompt-1	Task
	You are an expert assistant designed to extract critical knowledge from an academic abstract, with a special focus on innovation, methodology, and contributions. **Instructions:** - You are given an academic abstract. - Your task is to extract the most important pieces of knowledge or concepts that: - Describe the core topic of the paper. - Represent the key knowledge central to the study. - Are directly used in the paper's research method, framework, or experimental design. - Reflect the new theories, methods, or contributions introduced by the paper. - Include key theories, frameworks, or technologies used. - Represent the scientific or practical contribution to the field. - All extracted items must appear verbatim in the abstract. - Do not paraphrase, summarize, or add inferred content. - Try to extract no more than ten distinct items. - Return the final result as a single line, separated by English commas. **Now process the following abstract:** I have the following abstract: [DOCUMENT] Based on the abstract above, extract the knowledge that best describes the topic of the abstract. Prioritize any concepts that relate to the paper's innovation, methodology, and research contribution. Make sure all extracted knowledge or concepts appear verbatim in the text and try not to exceed ten pieces. Use the following format separated by commas: <Knowledge>
Prompt-2	Task
	You are an expert assistant designed to extract critical knowledge from an academic abstract. **Now process the following abstract:** I have the following abstract: [DOCUMENT] Based on the abstract above, extract the knowledge that best describes the topic of the abstract. Make sure all extracted knowledge or concepts appear verbatim in the text. Use the following format separated by commas: <Knowledge>

Appendix B. Histograms and normal Q-Q plots of predicted novelty scores by award status

This appendix presents the SPSS-generated histograms and Q-Q plots of predicted novelty scores for the award-winning and non-award groups. These visualizations supplement the normality assessments described in the main text, providing additional evidence for the approximate normality of the data distributions. Fig. B.1 displays the four subplots: panels (a) and (c) present the histogram and Q–Q plot for the non-award group, while panels (b) and (d) show the corresponding plots for the award-winning group. The figures support the conclusion that the award group exhibits approximate normality, while the non-award group shows modest deviations in the upper tail.



(a)

(b)

(c)

(d)

Fig. B.1. Normality assessment of predicted novelty scores by award status: histograms and Q-Q plots.

References

Ahuja, G., & Morris Lampert, C. (2001). Entrepreneurship in the large corporation: A longitudinal study of how established firms create breakthrough inventions. *Strategic Management Journal*, *22*(6–7), 521–543. https://doi.org/10.1002/smj.176

Al-Zaidy, R. A., & Giles, C. L. (2017). Automatic knowledge base construction from scholarly documents. In *Proceedings of the 2017 ACM symposium on document engineering* (p. 149–152). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3103010.3121043

Azoulay, P., Graff Zivin, J. S., & Manso, G. (2011). Incentives and creativity: Evidence from the academic life sciences. *The RAND Journal of Economics*, *42*(3), 527–554. https://doi.org/10.1111/j.1756-2171.2011.00140.x

Bai, X., Wang, M., Lee, I., Yang, Z., Kong, X., & Xia, F. (2019). Scientific paper recommendation: A survey. *IEEE Access*, *7*, 9324–9339. https://doi.org/10.1109/ACCESS.2018.2890388

Bao, T., Zhang, H., & Zhang, C. (2025). Enhancing abstractive summarization of scientific papers using structure information. *Expert Systems with Applications*, *261*, 125529. https://doi.org/10.1016/j.eswa.2024.125529

Berlyne, D. E. (1960). Conflict, arousal, and curiosity. New York: McGraw-Hill. https://doi.org/10.1037/11164-000

Brockman, B. K., & Morgan, R. M. (2003). The role of existing knowledge in new product innovativeness and performance. *Decision Sciences*, *34*(2), 385–419. https://doi.org/10.1111/1540-5915.02326

Brown, P. F., Della Pietra, V. J., deSouza, P. V., Lai, J. C., & Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, *18*(4), 467—479. https://aclanthology.org/J92-4003/.

Casadevall, A., & Fang, F. C. (2016). Revolutionary science. *mBio*, *7*(2), e00158–16. https://doi.org/10.1128/mbio.00158-16

Chen, L., & Fang, H. (2019). An automatic method for extracting innovative ideas based on the scopus® database. *Knowledge Organization*, *46*(3), 171–186. https://doi.org/10.5771/0943-7444-2019-3-171

Eshima, S., Imai, K., & Sasaki, T. (2024). Keyword-assisted topic models. *American Journal of Political Science*, *68*(2), 730–750. https://doi.org/10.1111/ajps.12779

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science*, *47*(1), 117–132. https://doi.org/10.1287/mnsc.47.1.117.10671

Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists' research strategies. *American Sociological Review*, *80*(5), 875–908. https://doi.org/10.1177/0003122415601618

Hernández-Castañeda, Á., García-Hernández, R.A., Ledeneva, Y., & Millán-Hernández, C.E. (2022). Language-independent extractive automatic text summarization based on automatic keyword extraction. *Computer Speech & Language*, *71*, 101267. https://doi.org/10.1016/j.csl.2021.101267

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Holland, J. H. (1992). Genetic algorithms. *Scientific American*, *267*(1), 66. https://doi.org/10.1038/scientificamerican0792-66

Hood, L., & Galas, D. (2003). The digital code of DNA. *Nature*, *421*(6921), 444–448. https://doi.org/10.1038/nature01410

Hou, J., Wang, D., & Li, J. (2022). A new method for measuring the originality of academic articles based on knowledge units in semantic networks. *Journal of Informetrics*, *16*(3), 101306. https://doi.org/10.1016/j.joi.2022.101306

Jang, H., Kim, S., & Yoon, B. (2023). An explainable AI (XAI) model for text-based patent novelty analysis. *Expert Systems with Applications*, *231*, 120839. https://doi.org/10.1016/j.eswa.2023.120839

Jeon, D., Lee, J., Ahn, J. M., & Lee, C. (2023). Measuring the novelty of scientific publications: A fasttext and local outlier factor approach. *Journal of Informetrics*, *17*(4), 101450. https://doi.org/10.1016/j.joi.2023.101450

Jeong, Y., & Kim, E. (2022). SciDeBERTa: Learning DeBERTa for science technology documents and fine-tuning information extraction tasks. *IEEE Access*, *10*, 60805–60813. https://doi.org/10.1109/ACCESS.2022.3180830

Kim, P. (2017). Convolutional neural network. In *MATLAB deep learning: With machine learning, neural networks and artificial intelligence* , pp. 121–147). Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4842-2845-6_6

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *International conference on learning representations (ICLR)*. Toulon, France: OpenReview.net. https://openreview.net/forum?id=SJU4ayYgl.

Kuhn, T. S. (1970). The structure of scientific revolutions. (2nd ed.). Chicago: University of Chicago Press.

Lee, Y.-N., Walsh, J. P., & Wang, J. (2015). Creativity in scientific teams: Unpacking novelty and impact. *Research Policy*, *44*(3), 684–697. https://doi.org/10.1016/j.respol.2014.10.007

Liu, M., Bu, Y., Chen, C., Xu, J., Li, D., Leng, Y., Freeman, R. B., Meyer, E. T., Yoon, W., Sung, M., Jeong, M., Lee, J., Kang, J., Min, C., Song, M., Zhai, Y., & Ding, Y. (2022). Pandemics are catalysts of scientific novelty: evidence from COVID-19. *Journal of the Association for Information Science and Technology*, *73*(8), 1065–1078. https://doi.org/10.1002/asi.24612

Lu, W., Huang, S., Yang, J., Bu, Y., Cheng, Q., & Huang, Y. (2021). Detecting research topic trends by author-defined keyword frequency. *Information Processing & Management*, *58*(4), 102594. https://doi.org/10.1016/j.ipm.2021.102594

Lubis, F., Nasution, D. A., Girsang, D. C., Sembiring, E., Rumahorbo, H. F. S., Naibaho, H. T., Simbolon, R. F. D., Siregar, S. M. A., & Naibaho, T. R. (2023). Analysis the role of references in scientific articles: Influence on research credibility and impact. *Formosa Journal of Science and Technology*, *2*(11), 3065—3074. https://doi.org/10.55927/fjst.v2i11.6822

Luo, Y., Wang, Y., Wang, Y., Wang, Y., Yan, N., Shiferaw, B. D., Mackay, L. E., Zhang, Z., Zhang, C., & Wang, W. (2024). Development and validation of a nomogram for predicting suicidal ideation among rural adolescents in china. *Psychology Research and Behavior Management*, *17*, 4413–4429. https://doi.org/10.2147/PRBM.S498396

Ma, Y., Ba, Z., Zhao, H., & Sun, J. (2023). How to configure intellectual capital of research teams for triggering scientific breakthroughs: Exploratory study in the field of gene editing. *Journal of Informetrics*, *17*(4), 101459. https://doi.org/10.1016/j.joi.2023.101459

Macgregor, R. B., & Poon, G. M. K. (2003). The DNA double helix fifty years on. *Computational Biology and Chemistry*, *27*(4), 461–467. https://doi.org/10.1016/j.compbiolchem.2003.08.001

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316. https://doi.org/10.1097/JTO.0b013e3181ec173d

Min, C., Bu, Y., Wu, D., Ding, Y., & Zhang, Y. (2021). Identifying citation patterns of scientific breakthroughs: A perspective of dynamic citation process. *Information Processing & Management*, *58*(1), 102428. https://doi.org/10.1016/j.ipm.2020.102428

Mishra, S., & Torvik, V. I. (2016). Quantifying conceptual novelty in the biomedical literature. *D-Lib Magazine*, *22*(9–10). https://doi.org/10.1045/september2016-mishra

Mormina, M. (2019). Science, technology and innovation as social goods for development: Rethinking research capacity building from sen's capabilities approach. *Science and Engineering Ethics*, *25*(3), 671–692. https://doi.org/10.1007/s11948-018-0037-1

Mukherjee, S., Uzzi, B., Jones, B. F., & Stringer, M. (2017). How atypical combinations of scientific ideas are related to impact: The general case and the case of the field of geography. In *Knowledge and networks* , pp. 243–267). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-45023-0_12

Nasar, Z., Jaffry, S. W., & Malik, M. K. (2018). Information extraction from scientific articles: A survey. *Scientometrics*, *117*(3), 1931–1990. https://doi.org/10.1007/s11192-018-2921-5

Popova, S., & Danilova, V. (2014). Keyphrase extraction abstracts instead of full papers. In *2014 25th international workshop on database and expert systems applications* (pp. 241–245). Munich, Germany: IEEE. https://doi.org/10.1109/DEXA.2014.57

Ruan, X., Ao, W., Lyu, D., Cheng, Y., & Li, J. (2023). Effect of the topic-combination novelty on the disruption and impact of scientific articles: Evidence from pubmed. *Journal of Information Science*, *0*(0), 01655515231161133. https://doi.org/10.1177/01655515231161133

Runhui, L., Yalin, L., Ze, J., Qiqi, X., & Xiaoyu, C. (2025). Quantifying the degree of scientific innovation breakthrough: Considering knowledge trajectory change and impact. *Information Processing & Management*, *62*(1), 103933. https://doi.org/10.1016/j.ipm.2024.103933

śauperl, A., Klasinc, J., & Luźar, S. (2008). Components of abstracts: Logical structure of scholarly abstracts in pharmacology, sociology, and linguistics and literature. *Journal of the American Society for Information Science and Technology*, *59*(9), 1420–1432. https://doi.org/10.1002/asi.20858

Savov, P., Jatowt, A., & Nielek, R. (2020). Identifying breakthrough scientific papers. *Information Processing & Management*, *57*(2), 102168. https://doi.org/10.1016/j.ipm.2019.102168

Schilling, M. A., & Green, E. (2011). Recombinant search and breakthrough idea generation: an analysis of high impact papers in the social sciences. *Research Policy*, *40*(10), 1321–1331. https://doi.org/10.1016/j.respol.2011.06.009

Schilling, M. A., & Phelps, C. C. (2007). Interfirm collaboration networks: The impact of large-scale network structure on firm innovation. *Management Science*, *53*(7), 1113–1126. https://doi.org/10.1287/mnsc.1060.0624

Shibayama, S., Yin, D., & Matsumoto, K., (2021). Measuring novelty in science with word embedding. *PLOS ONE*, *16*(7), 1–16. https://doi.org/10.1371/journal.pone.0254034

Shrivastava, S. R., & Shrivastava, P. S. (2022). Referencing in scientific writing and research. *Assam Journal of Internal Medicine*, *12*(2), 91–92. https://doi.org/10.4103/ajoim.ajoim_6_22

Sun, X., Chen, N., & Ding, K. (2022). Measuring latent combinational novelty of technology. *Expert Systems with Applications*, *210*, 118564. https://doi.org/10.1016/j.eswa.2022.118564

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, *342*(6157), 468–472. https://doi.org/10.1126/science.1240474

van Raan, A. F. J. (2004). Sleeping beauties in science. *Scientometrics*, *59*(3), 467–472. https://doi.org/10.1023/B:SCIE.0000018543.82441.f1

Verhoeven, D., Bakker, J., & Veugelers, R. (2016). Measuring technological novelty with patent-based indicators. *Research Policy*, *45*(3), 707–723. https://doi.org/10.1016/j.respol.2015.11.010

Veugelers, R., & Wang, J. (2019). Scientific novelty and technological impact. *Research Policy*, *48*(6), 1362–1372. https://doi.org/10.1016/j.respol.2019.01.019

Wang, J., Veugelers, R., & Stephan, P. (2017a). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, *46*(8), 1416–1436. https://doi.org/10.1016/j.respol.2017.06.006

Wang, J., Veugelers, R., & Stephan, P. (2017b). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, *46*(8), 1416–1436. https://doi.org/10.1016/j.respol.2017.06.006

Wang, S., Ma, Y., Mao, J., Bai, Y., Liang, Z., & Li, G. (2023). Quantifying scientific breakthroughs by a novel disruption indicator based on knowledge entities. *Journal of the Association for Information Science and Technology*, *74*(2), 150–167. https://doi.org/10.1002/asi.24719

Wang, Z., Qiao, X., Chen, J., Li, L., Zhang, H., Ding, J., & Chen, H. (2024a). Exploring and evaluating the index for interdisciplinary breakthrough innovation detection. *The Electronic Library*, *42*(4), 536–552. https://doi.org/10.1108/EL-06-2023-0141

Wang, Z., Wang, N., Zhang, H., Wang, Z., Wang, Z., Ding, J., & Chen, H. (2025). Ibid-cct: A novel model for interdisciplinary breakthrough innovation detection based on the cusp catastrophe theory. *Information Processing & Management*, *62*(4), 104121. https://doi.org/10.1016/j.ipm.2025.104121

Wang, Z., Zhang, H., Chen, J., & Chen, H. (2024b). An effective framework for measuring the novelty of scientific articles through integrated topic modeling and cloud model. *Journal of Informetrics*, *18*(4), 101587. https://doi.org/10.1016/j.joi.2024.101587

Weil, B. H. (1970). Standards for writing abstracts. *Journal of the American Society for Information Science*, *21*(5), 351–357. https://doi.org/10.1002/asi.4630210507

Weitzman, M. (1998). Recombinant growth. *The Quarterly Journal of Economics*, *113*(2), 331–360. https://doi.org/10.1162/003355398555595

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*(7744), 378–382. 10.1038/s41586-019-0941-9. https://doi.org/10.1038/s41586-019-0941-9

Wu, W., Zhang, C., Bao, T., & Zhao, Y. (2025). Sc4anm: Identifying optimal section combinations for automated novelty prediction in academic papers. *Expert Systems with Applications*, *273*, 126778. https://doi.org/10.1016/j.eswa.2025.126778

Xiao, T., Makhija, M., & Karim, S. (2022). A knowledge recombination perspective of innovation: review and new research directions. *Journal of management*, *48*(6), 1724–1777. https://doi.org/10.1177/01492063211055982

Yan, Y., Tian, S., & Zhang, J. (2020). The impact of a paper's new combinations and new components on its citation. *Scientometrics*, *122*(2), 895–913. https://doi.org/10.1007/s11192-019-03314-6

Zhang, J., & Zhu, L. (2022). Citation recommendation using semantic representation of cited papers' relations and content. *Expert Systems with Applications*, *187*, 115826. https://doi.org/10.1016/j.eswa.2021.115826

Zhang, L., Li, Y., & Li, Q. (2024). A graph-based keyword extraction method for academic literature knowledge graph construction. *Mathematics*, *12*(9), 1349. https://doi.org/10.3390/math12091349

Zhao, Y., & Zhang, C. (2025). A review on the novelty measurements of academic papers. *Scientometrics*, *130*(2), 727–753. https://doi.org/10.1007/s11192-025-05234-0

Zhao, Y., Zhang, M., Chen, X., & Zhang, Z. (2024). Early identification of scientific breakthroughs through outlier analysis based on research entities. *Journal of Data and Information Science*, *9*(4), 90–109. https://doi.org/10.2478/jdis-2024-0027

Zhu, L., Liu, X., He, S., Shi, J., & Pang, M. (2015). Keywords co-occurrence mapping knowledge domain research base on the theory of big data in oil and gas industry. *Scientometrics*, *105*(1), 249–260. https://doi.org/10.1007/s11192-015-1658-7

Zong, Q.-J., Shen, H.-Z., Yuan, Q.-J., Hu, X.-W., Hou, Z.-P., & Deng, S.-G. (2013). Doctoral dissertations of library and information science in china: A co-word analysis. *Scientometrics*, *94*(2), 781–799. https://doi.org/10.1007/s11192-012-0799-1