Journal of Informetrics 19 (2025) 101606

Contents lists available at ScienceDirect

Journal of Informetrics

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

Early identification of breakthrough technologies: Insights from science-driven innovations

Dan Wang [a], Xiao Zhou [a,*], Pengwei Zhao [a], Juan Pang [a], Qiaoyang Ren [a]

[a] School of Economics and Management, Xidian University, Xian, 710126, PR China

ARTICLE INFO

Keywords:
Technological breakthrough
science-driven innovation
Science-technology linkage
Knowledge network
Link prediction

ABSTRACT

Identifying breakthrough technologies is crucial for advancing technological innovation and, in this sense, the innovation patterns driven by science are considered to be key pathways for forming breakthrough technologies. Building on this premise, this paper presents a framework for identifying breakthrough technologies that starts with these signals of scientific innovation. The first step in the method is to construct a science-technology knowledge network based on papers and patents. Then a two-stage selection method funnels the scientific innovation signals, filtering out those with the potential to trigger technological breakthroughs. Next, a machine learning-based link prediction model, integrating three types of features, identifies new links between science-driven signals and existing technologies. A community detection algorithm then identifies sub-networks of technologies formed around these new links. Finally, a structural entropy index is used to evaluate these sub-networks to determine potential breakthrough technologies. By systematically characterizing the content and core features of scientific innovation signals, this study reveals the driving sources of technological breakthroughs and sheds light on the absorption and diffusion processes of scientific innovation. We validated the method through a use case in the field of artificial intelligence. Those who manage technological innovation should find the insights of this research particularly valuable.

1. Introduction

According to Schumpeter (Schumpeter, 1934), who was the first to introduce the concept of innovation into the field of economic growth, technological breakthroughs serve as a significant driver of both social progress and economic development. Subsequently, Ettlie, Bridges, and O'keefe (1984) categorized innovation into two paradigms: breakthroughs and incremental innovations. Incremental innovations include improvements to existing technologies through regular and continual evolution. They tend to be frequent occurrences that robustly drive industry growth. Moreover, markets and industries are generally adept at capturing the signals of these advancement. By contrast, breakthroughs are more intermittent and frequently have the potential to induce a paradigm shift in the status quo (Anderson & Tushman, 1990). In other words, breakthroughs catalyze industrial change. Yet identifying a breakthrough technology is incredibly difficult. Not only are they sporadic, but their sudden emergence tends to surprise both the industry and the market. Additionally, they often involve fresh approaches and require novel resource allocations so as to adjust to the evolving technological and market environment (Capponi, Martinelli, & Nuvolari, 2022). Take, for example, OpenAI's GPT-3, a third-generation text generation model introduced in 2020, was recognized as one of the notable innovations in *MIT Technology*

* Corresponding author.
E-mail address: belinda1214@126.com (X. Zhou).

https://doi.org/10.1016/j.joi.2024.101606
Received 18 March 2024; Received in revised form 24 September 2024; Accepted 7 November 2024
Available online 19 November 2024
1751-1577/

Review's 2021 list of '10 Breakthrough Technologies (Zhang & Li, 2021). The ChatGPT natural language question-and-answer model garnered over one million users within five days of its public launch and now boasts more than 50 million active users per month. This rapid adoption sparked a surge in research interest in generative artificial intelligence (AI) (Lund & Wang, 2023). This clearly illustrates that recognizing and forecasting breakthrough technologies is critical for enhancing competitiveness between firms and driving industrial transformation. Consequently, the primary focus of this study is to develop a method for identifying breakthrough technologies. Our solution is based on science-driven technological innovation patterns (S-T patterns), which have emerged as a critical path for pinpointing breakthrough technologies.

S-T patterns refer to the notion that scientific discoveries not only drive technological development but are also considered to be primary patterns of technological innovation. From this perspective, the ability to integrate scientific knowledge into technological development is crucial to a nation's capacity for innovation and its competitiveness (Malva, Kelchtermans, Leten, & Veugelers, 2015). Empirical research has demonstrated that technological breakthroughs often originate from scientific discoveries at three different levels (Veugelers & Wang, 2019). First, at the industry level, groundbreaking scientific discoveries are a major driving force behind technological breakthroughs and industrial advancements (Ke, 2020). For example, in the late 1980s, Nobel Laureates Albert Fert and Peter Grunberg discovered the principle of giant magneto-resistance – a discovery that not only led to theoretical breakthroughs in physics but also significantly accelerated the development and commercialization of disk drive technology across the globe (Dedrick & Kraemer, 2015). Second, at the company level, innovative science provides theoretical, data-driven problem-solving capabilities that offer reliable evidence for the novelty and creativity of corporate research and development (Malva et al., 2015). For instance, the Lord Manufacturing Company aimed to develop controllable fluids. After a decade of exploration in the field of electricity with no notable progress, Chief Research Scientist Dave Carlson's insight into the potential power of magnetic fields – being two orders of magnitude higher than the potential power of electric fields – led to a breakthrough in controllable fluid technology, highlighting the significant role of scientific innovation discoveries in solving practical problems (Fleming & Sorenson, 2004). Third, at the invention level, inventions are seen as a combinatorial search process, where innovative science provides several key prerequisites to generating breakthroughs (Narayanamurti & Tsao, 2024; Chung, Ko, Kim, & Yoon, 2021). It broadens the scope of the search for knowledge. It alters the search process of inventors, guiding them to find useful combinations in a more direct manner. And it closes off irrelevant research paths, saving time and resources.

Many studies identify breakthrough technologies by integrating scientific publications and patents to reveal the interactions between science and technology (S&T). Most of these approaches identify breakthrough technologies by analyzing non-patent citations (Min & Ke, 2021; Poege, Harhof, Gaessler, & Barufaldi, 2019). However, this method has some substantial limitations. For example, citation relationships are often incomplete, the time delay between an invention and publication can be lengthy, and these methods do not reveal the semantic relationships between S&T at a micro level (Block, Wustmans, Laibach, & Bröring, 2021). In recent years, text-based research techniques have developed the ability to identify technologies that might potentially be breakthroughs from the perspective of topic relevance (Xu, Luo, Winnink, Wang, & Elahi, 2022b). However, more research is needed to confirm that this method works in the early stages of a breakthrough, i.e., when the signals are weak. Weak signals refer to uncertain early indicators that have significant potential impact and often serve as early warnings of future opportunities (Ansoff, 1980). Yoon (2012) views weak signals as emerging topics connected to keywords that have not been widely interpreted by people. In the field of technology identification, weak signal analysis provides additional solutions for recognizing breakthrough technologies (Ma, Mao, & Li, 2024). Building on this concept, we introduce the term "scientific innovation signals," hereafter referred to as scientific signals, which denote potential but widely unexplored scientific discoveries within the scientific domain that have the potential to trigger technological breakthroughs. Systematically identifying and analyzing these signals and exploring the dynamic evolutionary pathways of scientific signals in inducing technological breakthroughs could help to identify breakthrough technologies much earlier.

Existing methods for identifying breakthrough technologies primarily rely on two different approaches to represent such technologies. The first approach uses coarse-grained IPC classification codes or individual patents or papers to characterize technological knowledge (Li, Ma, & Feng, 2024; Capponi et al., 2022). This method is limited in its ability to monitor detailed changes in technology at a micro level. With advancements in natural language processing (NLP), fine-grained text-based technology, text mining, and measurement have increasingly become significant research methods in technology identification and forecasting. The second approach uses keywords and keyword clusters as the most basic units for representing knowledge elements (Sun, Kolesnikov, Goldstein, & Chan, 2021). Through the innovative reorganization of these knowledge elements, this method provides semantic features that more precisely reveal technological content and details. This study relies on a fine-grained approach to representation.

We therefore developed a framework for identifying breakthrough technologies based on S-T patterns. The core idea of the research is to use "scientific signals" as a starting point to explore the mechanisms by which these signals lead to technological breakthroughs, and, in turn, identify the technologies with the potential to be breakthroughs based on this exploration. Our analytical framework is unique in three respects. First, we construct a science-technology network using high-quality research papers and patents. Building on this, we identify scientific signals using the weakness and novelty of the signals to effectively distinguish signals with the potential to lead to technological transformation. This is the starting point of science-driven technological breakthroughs. Second, we perform machine learning-based link prediction using a multi-feature fusion method to uncover potential new links between scientific signals and existing technologies. This approach captures the propagation paths and influence ranges of these signals within existing technological domains. Finally, a community detection algorithm identifies sub-networks containing new links. The influence of these sub-networks is then evaluated using structural entropy to ensure that breakthrough technologies are identified early. The effectiveness of this framework is validated through a use case in the field of AI.

The remainder of this paper is organized as follows. Section 2 reviews the related research. Section 3 proposes a framework for identifying breakthrough technologies. Section 4 presents an empirical analysis in the field of AI. Section 5 discusses the conclusions of

the study.

2. Literature review

2.1. Breakthrough technology identification methods

Most research on identifying breakthrough technologies early in their development focuses on qualitative approaches such as the Delphi method, expert opinions, and technology roadmaps (Sainio & Puumalainen, 2007; Dixon, Eames, Britnell, Watson, & Hunt, 2014). However, identifying breakthroughs using these methods not only requires significant time and effort, it also heavily depends on the expertise and knowledge of experts, which means they are susceptible to bias. Fortunately, today we have abundant resources for data processing, machine analysis, and algorithms that provide a less labor-intensive and more objective alternative to expert judgment. Such methods have evolved from the traditional expert-centric qualitative approaches into a hybrid blend of qualitative and quantitative approaches, such as bibliometrics, text analysis, and machine learning.

Bibliometric approaches identify breakthrough technologies by analyzing metrics like citation counts and structures like citation networks (Min et al., 2021; Xu, Winnink, Pang, Wen, & Chen, 2023). For example, the "highly cited" metric is commonly used to identify breakthrough advancements in current research (Mukherjee, Romero, Jones, & Uzzi, 2017). In fact, the greatest value of citation counts is their simplicity. However, while counting citations essentially reflects the impact of a technology, relying solely on this statistic may not accurately pinpoint groundbreaking technologies (Wuestman, Hoekman, & Frenken, 2020). For this and other reasons, recent research has shifted from prioritizing citation counts to favoring citation networks. For example, Funk and Owen-Smith (2017) introduced the consolidation-or-destabilization (CD) index, a dynamic citation-based network metric that evaluates how a particular patent challenges or reinforces a priori knowledge. Building on this, Wu, Wang, and Evans (2019) developed the disruption index (DI) as a simplified version of the CD index to gauge scientific and technological innovations.

The identification approaches based on text analysis primarily rely on technologies such as NLP to extract keywords, topic terms, semantic structures, and other strings from texts to identify breakthrough technologies. These approaches mainly include co-word analysis, topic modeling, subject-action-object (SAO) semantic structures, and burst word detection. Xu et al. (2022a) constructed a topic co-occurrence network based on data from scientific articles and implemented forward-looking prediction of breakthrough topics using link prediction and structural entropy methods. To cope with large-scale dataset processing and to uncover the hidden relationships in big data, machine learning has gradually become an important research approach. For example, Li et al. (2024) used a machine learning model to establish a relationship between historical sleeping beauties and their citation trends to predict potential sleeping beauties. They then further constructed breakthrough indicators to identify breakthrough research.

2.2. How S-T linkage analysis is used to identify breakthrough technologies

2.2.1. S-T linkage analysis

Science and technology are recognized as crucial elements in fostering innovation, with their dynamic interplay leading to numerous technological breakthroughs and advancements in science (Ba, Meng, Ma, & Xia, 2024). Following the linear model of "science drives technology, and technology drives the economy", innovation ideas from basic scientific research inspire applied research, leading to the creation of highly feasible technologies with significant economic returns (Balconi, Brusoni, & Orsenigo, 2010). However, as time has passed, studies have shown that technology can also reciprocally promote scientific development. This feedback loop has been added to the linear model, illustrating that S&T engage in a synergistic and mutually beneficial nonlinear relationship (Petrescu, 2009). This relationship can be broadly classified into three patterns of interaction: science-driven technology, technology-pulled science, and S&T synergy (Ba et al., 2024). This paper primarily focuses on S-T patterns. Particularly, in fields such as life sciences and AI, which are rooted in science and knowledge-intensive pursuits, scientific research serves as a crucial driver for achieving technological breakthroughs and innovations (Rosenberg & Birdzell 1990; Balconi et al., 2010).

2.2.2. Application of S-T in breakthrough technology identification

Existing research has extensively explored the role S-T connections play in helping to both identify key technologies and evaluate technological opportunities. Overall, these connections have been found to be good informers of breakthrough technologies. These studies primarily focus on two main approaches. The first approach involves citation network analysis. Numerous studies employ patent citations to non-patent literature to investigate the contribution of science to technological development (Xu et al., 2022b; Choi & Yoon, 2022). In terms of citation quantity, citing a greater number of recent innovative papers increases the impact of the patents and the likelihood of the technology being a breakthrough (Min & Ke, 2021). In terms of citation quality, a high-quality scientific foundation helps to generate a high-quality patent (Wang & Li, 2021). Poege et al. (2019) further find that the higher the quality of the scientific publications cited by a patent, the greater the value of the patent.

The second approach involves text mining methods that analyze the relevance of technological topics. Existing research mainly employs two categories of methods to analyze the innovation paths between science and technology to identify breakthrough technologies. The first category is link prediction, which is based on analyzing complex networks. Link prediction estimates the likelihood that a link between two nodes in a network will exist in the future based on the network nodes and structures known at a given point in time. In this sense, link prediction reveals future development trends in a technological field (Lü & Zhou, 2011). Han, Jeon, and Geum (2022) identified potential innovation opportunities by extracting the signals of innovation from scientific papers and converting them into service networks. They then used cosine similarity to predict where future links might form in these service networks. The second

category involves machine learning (Roh & Yoon, 2023). Machine learning effectively extracts information from complex, unstructured data (e.g., graphics and text) and has been applied in recent years to answer an array of different questions relating to ST&I management (Chen et al., 2023). For instance, Roh and Yoon (2023) proposed a sentence generation algorithm to discover technology and science innovation opportunities, with the generated sentences reflecting the science and technology relationship and specific directions of innovation. Chen et al. (2023) introduced a novel deep learning method for exploring science-technology linkages. Yu and Yan (2022) proposed using main path analysis in combination with several machine learning methods to study science-technology associations and identify potential knowledge discoveries.

Hence, the academic community has made substantial efforts to identify breakthrough technologies. However, three primary limitations have hampered these efforts. First, there has not been enough research into how to identify breakthrough technologies in their early stages of development. We know the scientific domain contains signals that indicate potential scientific innovation, but techniques to systematically identify and then capture these signals have not yet been widely explored. Second, although S-T patterns are a crucial source of breakthrough technologies, core issues regarding the dynamic process of how these patterns evolve and how they are transmitted, i.e., the specific ways in which science drives technological breakthroughs, have not been fully mapped. Third, existing methods largely quantify the links between S&T based on citations or similarity, but what we do not have is a multi-feature perspective that reveals the deeper fusions between S&T.

To address these gaps, we propose a framework for identifying breakthrough technologies based on S-T patterns. First, we identify scientific signals, which indicate a breakthrough in its early stages, quantified in terms of its weak signals and novelty. Second, by integrating three types of features – similarity metrics, centrality metrics, and text semantic metrics – we developed a machine learning-based link prediction method to identify potential new links between the scientific signals and the existing technologies. Subsequently, a community detection algorithm identifies sub-networks that contain new links. These sub-networks are assessed using structural entropy to discern breakthrough technologies.

3. Research framework and methodology

Scientific research not only provides theoretical support and guidance for technological development but also offers significant insights into the direction and pathways for technological innovation. Based on this concept, we constructed a research framework for identifying breakthrough technologies driven by science. The analytical framework is shown in Fig. 1.

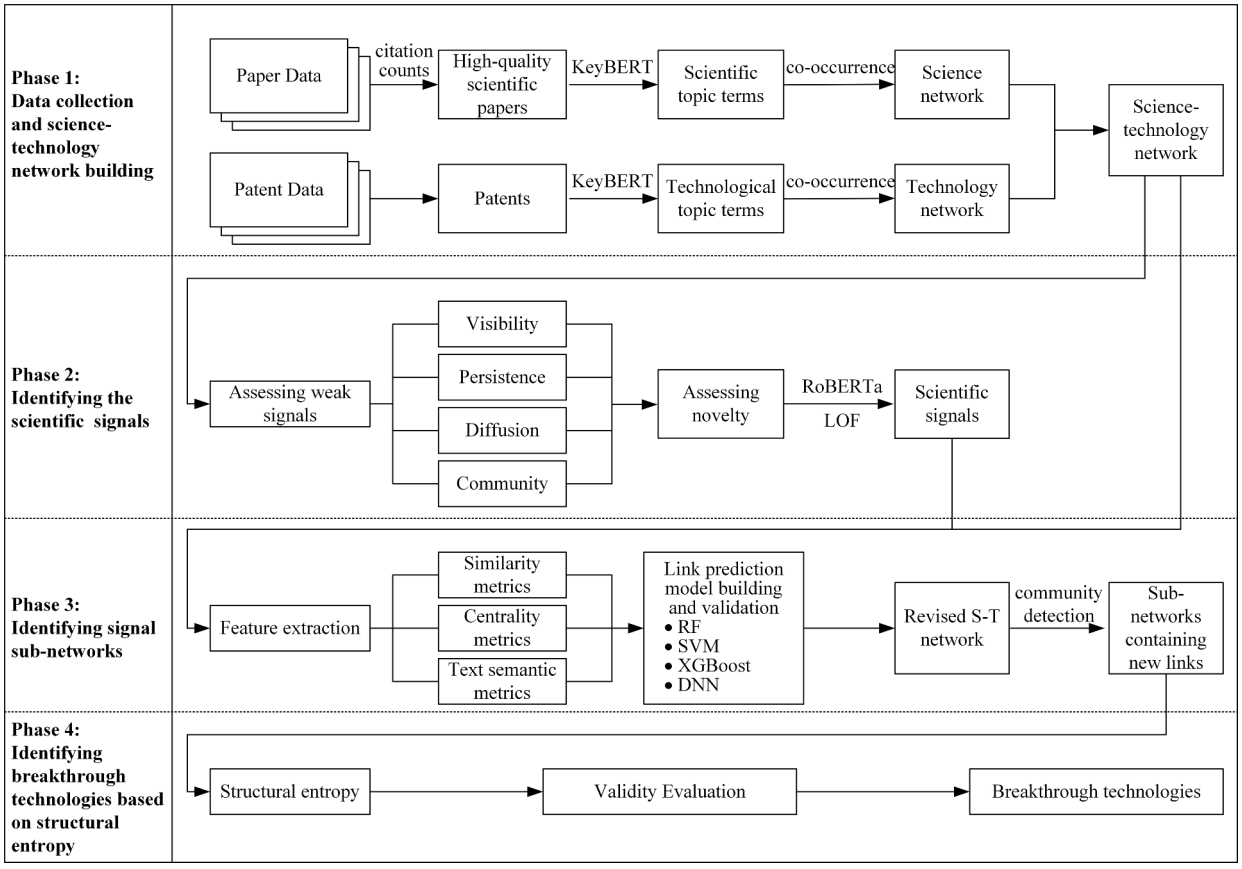

Fig. 1. The research framework.

3.1. Data collection and science-technology network building

This stage primarily involves collecting data and constructing a science-technology knowledge network. This procedure involves three steps: data collection and preprocessing, knowledge extraction, and network construction.

3.1.1. Data collection and preprocessing

The first two steps in data collection are to select a target domain and then formulate a retrieval strategy. The dataset was obtained from the Web of Science (WoS), where we retrieved the title, abstract, publication year, and total citation count of the papers. We combined the titles and abstracts of the papers as the text data for the papers. Subsequently, we collected patent data from the Incopat database, merging the titles and abstracts as the patent texts.

Existing studies indicate that high-quality scientific papers are more likely to lead to technological breakthroughs than lower quality papers (Poege et al., 2019). Therefore, in addition to collecting papers from the field, we further filtered out the high-quality papers. The literature widely recognizes a positive correlation between high-quality papers and higher citation counts (Jang, Woo, & Lee, 2017), with Bornmann et al. (2014)'s research suggesting that highly cited papers have greater scientific value and make more significant contributions to scientific progress. Hence, in this paper, we used citation counts as a measure of paper quality. Note that, over time, older papers will accumulate more citations than more recently published papers. To address this issue, we introduced a time-slicing method, selecting papers with citation counts exceeding a certain threshold year by year as high-quality papers.

3.1.2. Extracting knowledge

To extract keywords from both the papers and patent texts, we chose an unsupervised and domain-independent keyword extraction algorithm called KeyBERT (Khan et al., 2022). KeyBERT uses document embeddings to represent texts at the document level, then extracts N-gram word/phrase embeddings and uses cosine similarity to find the words/phrases most similar to the document, effectively assigning keywords representing the knowledge concepts for each text. The extracted keywords undergo manual semantic cleaning to merge synonyms and remove irrelevant words. We filtered scientific and technological terms by setting frequency thresholds, identifying scientific topic terms and technological topic terms. It is crucial to carefully consider the characteristics of scientific signals when setting the threshold for the frequency of scientific topic terms.

3.1.3. Constructing the science-technology network (S-T network)

Based on the knowledge elements and their co-occurrence relationships, we retained the largest connected graph and used that to construct one science network (S-Net) and one technology network (T-Net). We then used semantic similarity to construct a combined S-T network. Given the significant differences in writing style and expression between patent texts and academic papers, we bridged these two domains using semantic similarity method. Specifically, we first used RoBERTa (Liu et al., 2019) to convert scientific and technological topic terms into fixed vector representations. Then, we calculated the pairwise cosine similarities between the topic terms in the science network and those in the technology network. Finally, we retained pairs of scientific and technical topic terms with high similarity, and using expert opinions, merged synonyms from among the selected topic terms to integrate terms with similar meanings in both fields. Through these shared terms, we established mappings between the science and technology networks, thereby constructing an S-T network that maps the connectivity between scientific research and existing technologies.

3.2. Identifying the scientific signals

This stage focuses on identifying the scientific signals in the S-T network. This paper represents these signals by extracting keywords from academic papers. Essentially, scientific signals are a type of weak signal. Hiltunen (2008) proposed a three-dimensional framework for weak signals based on the dimensions of signal, issue, and interpretability. Building on this theoretical framework, Yoon (2012) quantified these dimensions, using visibility, diffusion, and influence as new analytical frameworks, and identified these weak signals through the frequency and growth rate of keywords. However, this research did not consider the aspect of novelty. Therefore, we adopted a two-stage method of funneling the terms: the first stage employs bibliometric analysis to evaluate and identify the characteristics of weak signals, while the second stage applies machine learning methods to assess novelty.

3.2.1. Assessing weak signals

Drawing upon the studies of Yoon (2012) and Porter, Garner, Carley, and Newman (2019), four characteristics are measured in the first stage of assessing weak signals: visibility, persistence, diffusion, and community. Visibility is assessed through term frequency. Scientific signals typically emerge in the early stages of research, characterized by a limited number of related publications and underdeveloped technical applications. Then, for a scientific signal to be significant, it must exhibit persistence and diffusion; otherwise, it is just a "one-hit wonder". Finally, a scientific signal should be able to attract interest within the academic community. Our evaluation of weak signals is therefore based on four thresholds:

(1) Visibility: Scientific terms that exceed a certain threshold frequency (related to the threshold set for the frequency of scientific terms in Section 3.1.2) and appear exclusively within the scientific domain, but not within the technological domain.
(2) Persistence: The scientific topic term must appear in at least two out of the five years.
(3) Diffusion: The frequency of the term in the subsequent three years should be at least twice that of the frequency in the preceding two years.

(4) Community: The scientific topic term must be present in the papers of at least two different organizations.

3.2.2. Assessing novelty

The second stage involves assessing novelty. Novelty measures the degree to which an idea is related to current prevailing paradigms. An idea is considered more novel if it significantly differs from existing paradigms. At this stage, we adopt RoBERTa and local outlier factor (LOF) (Breunig, Kriegel, Ng, & Sander, 2000) to measure the novelty of scientific topic terms. At the core of this approach, RoBERTa represents scientific and technological topic terms as vectors of textual information, while LOF measures the novelty of topic terms on a numerical scale. LOF is a density-based outlier detection method that quantifies the novelty of an object by analyzing the local density of the target object compared to its neighboring objects in the embedding space. Thus, LOF is calculated as follows:

$$LOF_k(p) = \frac{1}{k} \sum_{q \in N_K(p)} \frac{lrd(q)}{lrd(p)} \tag{1}$$

where p is an object, q represents its k*th* nearest neighbor, with k being the number of neighbors. *lrd* is the local reachability density, and $N_K(P)$ is defined as the set of objects within k-distance(p) from the object p. The k-distance(p) is computed as the Euclidean distance between p and its k-th nearest neighbor.

After calculating the LOF scores for all topic terms, the scientific topic terms are filtered from the previous step in descending order based on their LOF values. The top 20% of terms are selected and, with expert input, terms with little similarity to mainstream technological knowledge are identified. Terms with high LOF scores are more novel. These highly novel terms represent the final scientific signals, indicating scientific discoveries that are underdeveloped and underutilized, potentially serving as starting points for technological breakthroughs.

3.3. Identifying signal sub-networks

In this stage, a link prediction algorithm is used on the S-T network to uncover potential connections between scientific signals and technological topic terms. Then, a community detection algorithm is used to identify sub-networks within the main S-T network that include these science-technology signal associations. Using this approach, we can identify and examine clusters of closely connected scientific signals and technological topic terms within the network, providing a deeper understanding of how these signals of innovation have propagated or impacted the field.

3.3.1. Performing link prediction over the S-T network

This step aims to identify potential links between scientific signals and existing technological topic terms in the S-T network using a machine learning-based link prediction approach. Following the research of Park and Geum (2022) and Wang and Lee (2023), we consider three types of feature metrics: (1) technological similarity metrics; (2) technological centrality metrics; and (3) technological text semantic metrics.

First, the similarity between two nodes can provide clues for future links. Six representative similarity-based link prediction indices are used for this purpose, including the Common Neighbor index (CN), Salton index (Salton), Jaccard index (Jaccard), Adamic–Adar index (AA), Resource Allocation index (RA), and Katz index (Katz) (Clauset, Moore, & Newman, 2008; Martínez, Berzal, & Cubero, 2016). Second, technological centrality represents the importance of each node, which is based on the number of connections it has or the importance of its adjacent nodes. Three types of node centrality are considered: degree centrality, closeness centrality, and betweenness centrality (Wang & Lee, 2023). Lastly, for textual semantic features, we use RoBERTa to vectorize the topic terms of all nodes in the network and calculate the semantic similarity between pairwise nodes using the cosine similarity algorithm (Park & Geum, 2022).

These three types of feature metrics constitute the input for the machine learning model, as shown in the Table 1.

In terms of model selection, we chose four classic machine learning models that have been successfully applied in various fields: random forest (RF), support vector machine (SVM), extreme gradient boosting (XGBoost), and a deep neural network (DNN) (Park & Geum, 2022; Wang & Lee, 2023). When implementing the link prediction, we treated the existing links between any two nodes as a positive sample and defined the absence of a link as a negative sample. Due to the sparsity of the network, the number of negative samples greatly exceeded the number of positive samples, which led to an imbalanced data issue. To address this problem, we use undersampling to match the number of positive and negative samples. Next, we split the samples into training and validation sets according to a certain ratio and evaluated the model's performance using the validation data. We assessed the classification results of the algorithms using accuracy, precision, recall, and F1 scores (Forman, 2003).

Table 1
Link prediction measures.

Categories	Link prediction measures	References
Similarity metrics	CN, Salton, Jaccard, AA, RA, and Katz	Martínez et al. (2016))
Centrality metrics	Degree centrality, Closeness centrality, and Betweenness centrality	Wang and Lee (2023)
Text semantic metrics	Text similarity	Park and Geum (2022)

$$Accuracy = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{2}$$

$$Precision = \frac{TP_i}{TP_i + FP_i} \tag{3}$$

$$Recall = \frac{TP_i}{TP_i + FN_i} \tag{4}$$

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5}$$

Moreover, we continuously adjusted the model parameters to improve classification performance, and then selected the best-performing model to predict the new links in the S-T network. Our goal was to investigate how scientific signals diffuse within existing technological systems and ultimately lead to evolutionary trajectories of technological breakthroughs. Among the newly predicted links, we focused specifically on the links between scientific signals and technological topic terms, excluding other types of links. This approach not only simplified the complex network structure but also enabled us to more precisely capture the diffusion paths of scientific signals within technological systems. These were the only new links we incorporated into the original network to form a revised S-T network.

3.3.2. Performing community detection on the revised S-T network

The goal of this step is to use community detection to identify sub-networks that contain new links between scientific signals and technological topic terms in the revised S-T network. The community detection algorithm divides the complex revised S-T network into relatively independent sub-networks. Specifically, we use the Fluid C method (Parés et al., 2017), which considers the dynamic changes of nodes within the network, resulting in greater cohesion within the same community and relatively weaker links between different communities. This allows us to effectively delineate different communities, with each community focusing on a specific set of topics. Since we are particularly interested in technological changes driven by innovative science, we consider breakthrough technologies to be composed of closely related scientific signals and technological topic terms. Therefore, the algorithm identifies sub-networks containing new links as candidates for breakthrough technologies.

3.4. Identifying breakthrough technologies based on structural entropy

The purpose of this stage is to identify breakthrough technologies by evaluating the impact of sub-networks containing new links. From the perspective of network dynamics, the addition of new nodes and new links can impact the existing technological network, and this impact can be measured by the extent of changes in the network structure (Chen, 2012). Therefore, we regard significant changes in the structure of the revised S-T network as important signals of technological transformation.

To better capture the changes in the structure of the revised S-T network, we incorporate structural entropy to measure structural changes to the network. Structural entropy theory, which is derived from the concept of entropy in thermodynamics, explores the characteristics of complex network structures from a topological perspective and can reveal the overall information of the network (Xu et al., 2022a). Thus, it can provide a methodological basis for numerically characterizing a network's structural features. Adding scientific signals and their links to existing technologies so as to create a revised S-T network impacts the knowledge network. This impact is measured using a structural entropy index, which identifies sub-networks that alter the state of the knowledge network.

More specifically, we chose the structural entropy calculation method proposed by Xu et al. (2022a). This method focuses on the differences in structural entropy reflected by different sub-networks. Specifically, the method compares the entropy values of the overall network with and without particular sub-networks. The first step is to take a measurement of the overall structural entropy of the originally revised S-T network, denoted as *original_structuralentropy*. Then, for each sub-network containing new links, we remove one sub-network at a time from the revised S-T network and re-measure the network's structural entropy, denoted as new_*structuralentropy*. This provides an assessment of the impact of each sub-network on the overall network structure. The structural entropy impact of each sub-network *Subnetwork_Impact* is then calculated as follows:

$$Subnetwork\_Impact = |original\_structuralentropy - new\_structuralentropy| \tag{6}$$

The focus here is on the magnitude of the sub-network's impact on the network structure, regardless of the direction of the impact. Therefore, only the absolute value of the results is important. The greater the impact of the sub-network, the greater the technology's breakthrough potential. The median structural entropy value becomes the threshold for determining whether a sub-network has high impact. Hence, if the impact value is above the median, the sub-network is selected as a potential breakthrough technology.

4. Empirical analysis

To demonstrate the effectiveness of the proposed method, we choose the field of AI as a case study. AI is a typical field where scientific development precedes technological innovation. As a core component of the most recent scientific and technological revolution, AI is having a profound impact on the world's socioeconomic development and competitive international landscape. From the emergence of neural network-based intelligent decision-making algorithms in 2017 to the improvement in natural language

understanding capabilities of large-scale models like GPT by 2022, waves of AI advancements are emanating from technological breakthroughs and laying the groundwork for further intelligent applications in future scenarios. Therefore, the ability to predict breakthrough technologies in the field of AI would be highly beneficial for governments and enterprises involved in strategic research and planning.

4.1. Data acquisition and construction of the S-T network

4.1.1. Data collection and selection of high-quality papers

Publications in the field of AI and granted patents were obtained from the WoS and the Incopat patent database, respectively. The search strategy was based on the method proposed by Tsay and Liu (2020). For scientific publications, further restrictions were applied at the disciplinary level, retaining papers within disciplines including 'Engineering', 'Computer Science', 'Electrical & Electronic', 'Artificial Intelligence', 'Theory & Methods', 'Multidisciplinary', 'Interdisciplinary Applications', 'Automation & Control Systems', 'Robotics', 'Hardware & Architecture', and 'Acoustics'. After removing duplicates, a total of 699,930 publications related to AI published from 2010 to 2022 and 99,712 patents were obtained. It is worth noting that the annual growth rate of scientific publications has increased by more than around seven times that of patents in the past decade, as shown in Fig. 2. This indicates that scientific achievements in this field are more abundant than patents. However, one can also see an undeniable time delay between basic research and a technological outcome, known as the experience lag period, which, here, is typically around four years. Patent value significantly depreciates five years after issuance (Zhang, You, Tang, & Wen, 2023). Therefore, we focused on the period from 2014 to 2018, which netted 234,285 publications and 29,468 patents.

Existing studies typically set the threshold for high-quality papers between 1% and 10% (Hu, Cui, & Lin, 2023). If the threshold is set at the top 1%, the sample size may be insufficient. Conversely, if the threshold is set at the top 10%, the increase in sample size could obscure key characteristics of scientific signals, reducing the accuracy of identification. We use the time slice approach and set our threshold for quality to the top 5% of papers by annual citations each year as high-quality papers, retaining 11,711 high-quality papers from this exercise.

4.1.2. Knowledge extraction and construction of the S-T network

First, we performed preliminary data cleaning on the paper and patent texts, including removing special symbols and lemmatizing words. Next, we used the KeyBERT algorithm to extract keywords from both the paper and patent texts. The extraction process was configured using the all-MiniLM-L6-v2 model (Shooshtarian, Gurmu, & Sadick, 2023), with the "top_n" parameter set to 15 to retrieve the 15 most relevant keywords and the "keyphrase_ngram_range" set to (2, 3). Irrelevant words were removed, and synonyms were merged. In setting the frequency threshold for technological terms, we found that too low a threshold introduced a great deal of noise, obscuring core technological terms, while too high a threshold risked missing important terms, affecting accuracy. Therefore, referencing existing research and experimental results (Xu et al., 2022a), we set the term frequency threshold to 10, retaining technological terms exceeding this threshold. Ultimately, 2334 technological terms were extracted. Based on the co-occurrence relationships among these terms, we retained the largest connected component to construct the T-Net. For scientific terms, considering the visibility of scientific signals and avoiding short-term random phenomena (Porter et al., 2019), we set the threshold to 5, retaining scientific terms exceeding this threshold. This resulted in 2055 scientific terms. Based on the co-occurrence relationships among these terms, we retained the largest connected component to construct the S-Net.

Using RoBERTa, we vectorized the scientific and technological terms into vectors of 768 dimensions. We calculated the similarity

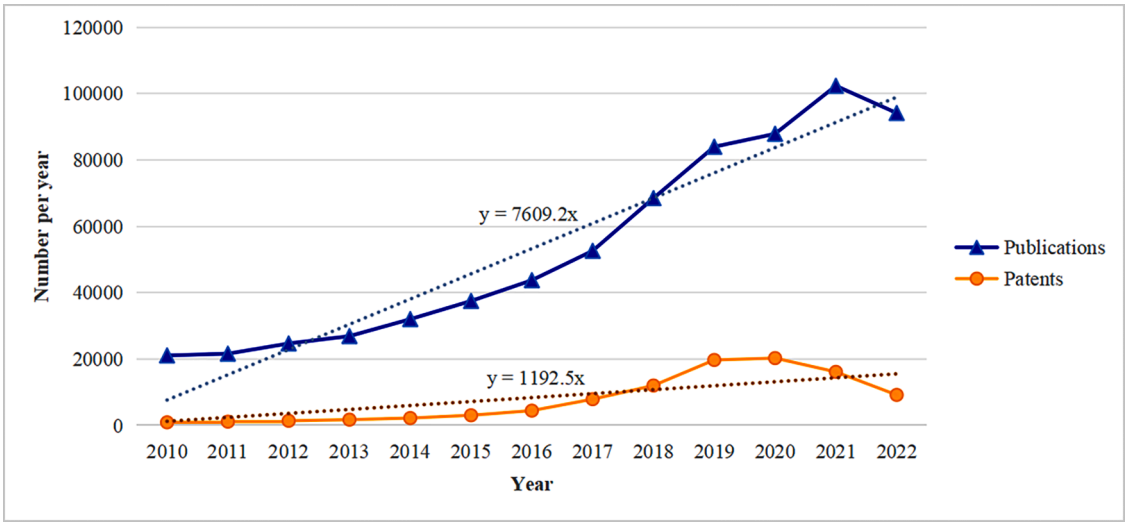

Fig. 2. The number of scientific publications and patents published annually in the AI field.

between nodes in the S-Net and T-Net and merged highly similar scientific-technological term pairs based on expert opinions, resulting in 993 shared terms. Then, using the mapping relationships between these term pairs, we constructed the S-T network. Descriptive statistics of the three networks are shown in Table 2.

4.2. Identification of scientific signals

Following the framework design outlined in Section 3.2, we quantitatively identified scientific signals in terms of their weakness and novelty. In the first stage, which involves identifying the weak signals, we found that 1062 scientific terms did not appear in technological terms by comparing the scientific terms with the technological terms. After applying the three additional criteria, we screened out 561 scientific terms. The second stage concerns assessing novelty. Here, we combined technological and scientific terms, resulting in 3396 terms. Using RoBERTa, we vectorized the semantic text of these terms into 768-dimensional vectors, which served as the input for LOF to calculate the novelty score of each term. We set the parameter k in the LOF algorithm to 20. Focusing on the novelty scores of the scientific terms screened in the first stage, we then ranked these scores in descending order. After prioritizing the top 20% of candidate terms and supplementing the results with expert opinions, we identified 96 scientific terms as signals of scientific innovation. Table 3 shows the scores of the top 10 scientific signals.

4.3. Identification of signal sub-networks using link prediction and community detection

This step involved calculating the feature indicators within the S-T network as defined in Section 3.3.1. From the three types of feature indicators, we obtained a 13-dimensional vector as the input data. There were 3396 nodes in the original S-T network, including 316,924 positive links and 5447,786 negative links. Hence, we divided the dataset into training and validation sets according to a ratio of 8:2. To balance the data, we randomly selected an equal number of negative links in the training and validation sets. We trained and evaluated the model using four machine learning techniques. Comparing the performance of each model across different evaluation metrics, as shown in Table 4, it is evident that RF outperformed the others on most metrics. Therefore, we selected the RF model to predict unlinked edges. Among the predicted new links, 630 new links between scientific signals and technological terms were identified and added to the original S-T network, resulting in a revised S-T network.

Following link prediction analysis, we applied the Fluid C community detection method to the revised S-T network, clustering the network into 20 sub-networks. Among these, 8 sub-networks did not contain scientific signals and were excluded, leaving 12 sub-networks for subsequent research.

4.4. Identification of breakthrough technologies based on structural entropy

Since calculating structural entropy involves edge weights, we computed the correlation between the feature metrics and the links, finding that the Adamic-Adar metric had a correlation coefficient of 0.72 with the links, indicating a significant correlation. Therefore, this metric was used as the edge weight for calculating the structural entropy. The structural entropy impact of each sub-network was calculated according to the method described in Section 3.4, as shown in Fig. 3.

We identified sub-networks above the median as candidate breakthrough technologies. The final results were decided upon by consulting experts. Ultimately, the study identified six breakthrough technologies: multi-modal natural language processing (multi-modal NLP), brain-computer interface, cloud computing and edge computing, visual tracking,video image processing technology, and intelligent diagnosis, as shown in the Fig. 4. Among them, the impact of multi-modal NLP, as the 16th technology, was particularly significant. We therefore conducted a detailed analysis of this breakthrough technology. It is evident that scientific signals such as multi-modal, encoder-decoder, CNN-RNN combined with natural language processing systems, emotion analysis, text semantics, and language translation, have had a profound impact. Specifically, multi-modal NLP leverages information from text, speech, and visual sources, encompassing a richer information set than single-modal approaches. The integration of different modalities allows for complementary content, aiding in the resolution of ambiguities and enhancing semantic understanding, thereby significantly improving the performance of various NLP tasks, such as sentiment analysis and machine translation. Multi-modal sentiment analysis represents an emerging and prominent area within the field of NLP. By integrating vocal intonation, facial expressions, and textual content, it is possible to more accurately capture and understand emotional changes, leading to a notable increase in the accuracy of speech emotion recognition (Singh & Kapoor, 2023). In machine translation, compared to conventional text-only Neural Machine Translation (NMT), Multi-modal NMT can utilize visual information closely related to text semantics (Huang, Zhang, & Xu, 2023). This integration helps the system better understand the specific context of the text, resulting in translations that are more contextually appropriate. The methods based on the encoder-decoder model not only effectively integrate and process multi-modal information but also offer new possibilities for future research and applications in multi-modal technologies. Additionally, Chen et al. (2023)'s research

Table 2
Descriptive statistics of the S-Net, T-Net and S-T network.

Network	Number of nodes	Number of edges	Network density
S-Net	2055	146,637	0.069
T-Net	2334	202,348	0.074
S-T network	3396	316,924	0.055

Table 3
Scientific signals and LOF Score (Top 10).

Rank	Scientific signals	LOF Score
1	google street view	1.85
2	self drive car	1.76
3	neighborhood rough set	1.67
4	aperture radar	1.66
5	high dimensional data	1.60
6	multi modal	1.59
7	remote sense imagery	1.55
8	medical synthetic image	1.54
9	inference network	1.51
10	visual question answer	1.50

Table 4
Performance matrix.

Model	Accuracy%	Precision%	Recall%	F1 score%
DNN	89.87	88.09	92.22	90.10
RF	**89.90**	87.92	**92.50**	**90.15**
SVM	89.70	**89.79**	89.70	89.69
XGBoost	89.71	**89.79**	89.71	89.70

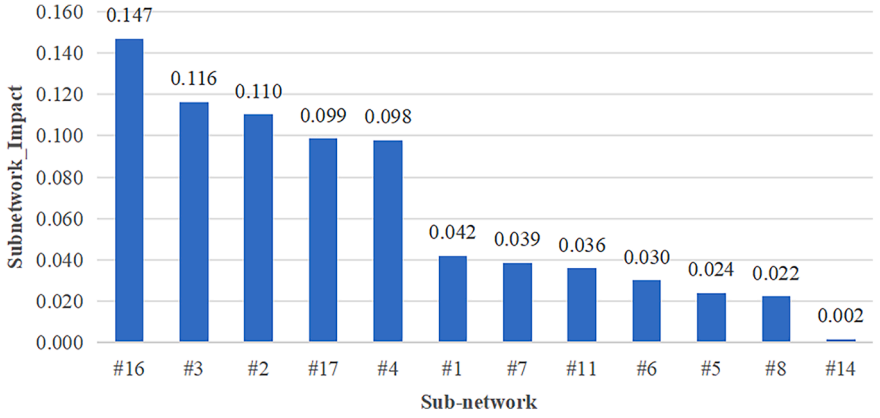

Fig. 3. Structural entropy influence results of sub-networks.

indicates that from 2018 - 2021, the topic pair "CS90 (Multi-modal NLP) - CT3 (Speech Emotion Analysis)" has the highest coupling strength. The finding aligns with our results, suggesting that multi-modal NLP technology will be a significant direction for future research in the field.

4.5. Verification

To verify the effectiveness of our proposed method, we conducted a comprehensive evaluation, including keyword validation, novelty verification, and a quantitative analysis of the identified breakthrough technologies.

4.5.1. Keyword validation

We selected three common algorithms for validation: KeyBERT, RAKE, and TF-IDF, adopting the method proposed by Li and Yan (2019) to evaluate the performance of these three algorithms. Specifically, we randomly selected 10 high-quality papers and 10 patent texts from 2014 to 2018, making a total of 100 passages of text. Two coders independently rated the keywords extracted from these 100 texts. The rating criteria included three levels: Level 1 includes meaningful keywords that reflect the theme of the original text; Level 2 includes meaningful keywords that do not reflect the theme of the original text; and Level 3 includes meaningless keywords unrelated to the theme of the original text. To ensure the reliability and validity of the evaluation results, each coder independently rated the keywords, and Cohen's Kappa coefficient was used to measure the consistency between the coders' ratings. The final Cohen's Kappa coefficient was 0.78, indicating a high level of agreement among the coders. According to the coding results presented in Table 5, KeyBERT demonstrated the best performance in keyword extraction, while the traditional TF-IDF algorithm performed the worst.

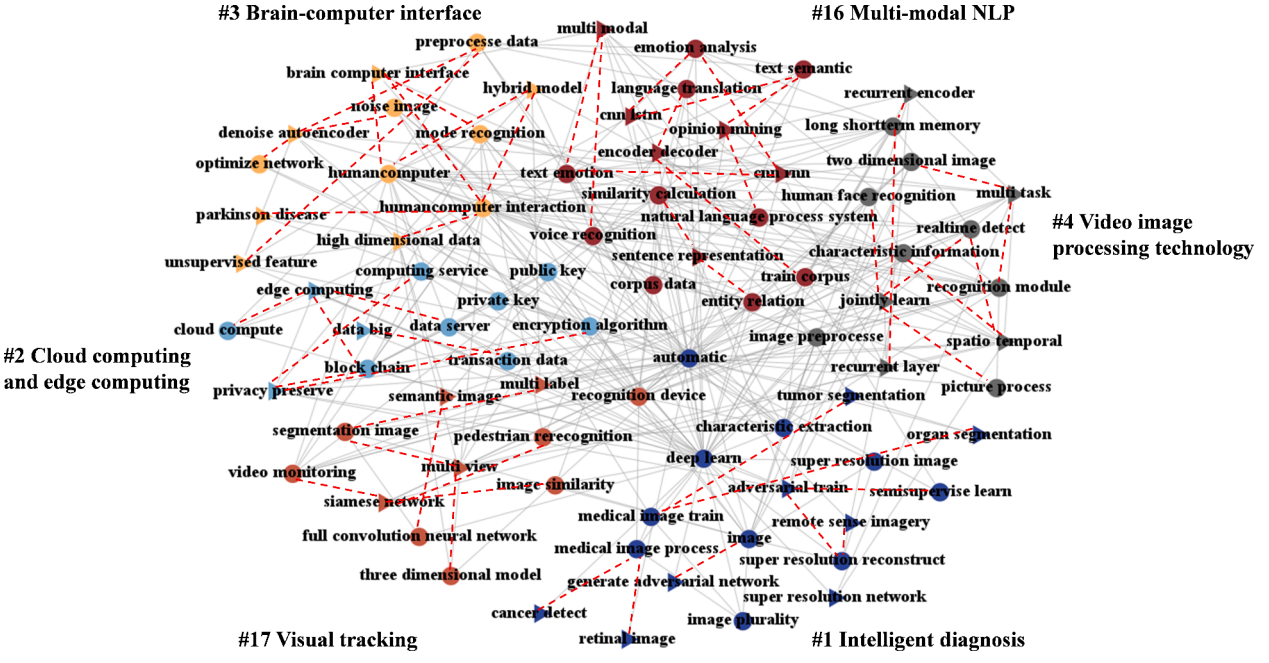

Fig. 4. Breakthrough technologies (triangle node represents scientific signals, dot node represents technology terms. A solid line represents the direct connection between terms. While a red dotted line represents new links from linkage prediction).

4.5.2. Validation of novelty

Building on the method proposed by Jeon, Lee, Ahn, and Lee (2023) to evaluate the novelty of scientific terms, we applied four different word embedding algorithms—RoBERTa, BERT, SciBERT, and Word2Vec—and combined them with two outlier detection algorithms: LOF and isolation forest (IF) (Liu, Ting, & Zhou, 2008). We constructed eight different algorithm combinations and computed the novelty scores of the topic terms for each. The top 50 scientific terms based on each algorithm's scores were selected as candidate novelty terms for evaluation. To effectively validate the novelty of the terms, we constructed an emerging terms list based on key industry reports, major reviews, government planning documents, and expert opinions. This list was used as "scientific innovation" labels to test the effectiveness of our algorithms. We evaluated the algorithms by the proportion of terms identified as labels. The comparative results of the eight algorithm combinations are shown in Table 6. These results demonstrate that, in the field of AI, the combination of RoBERTa + LOF has certain advantages.

4.5.3. Quantitative validation of breakthrough technologies

In this section, we conducted quantitative experimental comparisons to verify the accuracy of the breakthrough technology results. We retrieved 10 patents most similar to the 20 sub-network technology combinations from AI patents filed between 2019 and 2023, ultimately obtaining 200 patents. We classified these patents into three categories: 1) breakthrough patents, similar to the six identified breakthrough technologies, which included 60 patents; 2) non-breakthrough patents, similar to the six non-breakthrough technologies, which included 60 patents; and 3) ordinary patents, similar to the eight technology combinations that did not include new links. These numbered 80 patents. We then compared these three types of patents using four indicators from the Incopat database: number of forward citations, number of transfers, technology advancement and patent value. Citations reflect the patent's impact, transfers reflect the patent's commercial value, technology advancement reflects the degree of innovation, and patent value reflects the quality of the patent. The performance of the three types of patents on these four indicators is shown in Table 7. The validation results show that breakthrough patents score higher against all four indicators. This is consistent with existing research, indicating that breakthrough patents have higher technological impact and a higher commercial value (Arora, Belenzon, & Suh, 2022).

5. Conclusion

5.1. Main conclusion

The main purpose of this study was to develop a framework for identifying breakthrough technologies based on S-T patterns. First, we constructed an S-T network based on the semantic similarity between high-quality papers and patents to reveal the links between science and technology. Second, we identified scientific signals against two metrics: the weakness of the signals and novelty, effectively filtering out those signals that could trigger a technological revolution. This marks the starting point of identifying a science-driven technological breakthrough. Then, we integrated three types of features: similarity, centrality, and text semantic, and employed machine learning-based link prediction to predict potential new links between scientific signals and existing technologies. This reveals the diffusion path of scientific signals through technological domains. Next community detection isolates sub-networks that contain newly predicted links, while structural entropy is used to assess the relative impact of these sub-networks, identifying potential breakthrough technologies. Using the field of AI as an example, the results identify six breakthrough technologies, with both quantitative and qualitative assessments confirming the effectiveness of the proposed method. The promotion and application of these technologies is expected to have a significant impact on future technological and industrial development.

5.2. Major contributions

The main innovations and contributions of this study are reflected in three respects:

EnclosedCircleChinese1 Based on S-T patterns, this study proposes a systematic method for identifying breakthrough technologies. This method uses a fine-grained approach based on terms to dynamically track and measure how scientific signals trigger technological breakthroughs. It significantly improves upon traditional citation-based methods in providing insights into the relationship between S&T. Additionally, it addresses some of the limitations of static perspectives in identifying breakthrough technologies. The proposed method can be integrated into interactive systems for identifying scientific signals and predicting breakthrough technologies, offering a valuable tool with high adaptability and flexibility. This system can modularly integrate various components of the analytical framework, facilitating real-time monitoring and early warning of breakthrough technologies in target areas. The developed software and quantitative results can provide

Table 5
Performance results of algorithms for extracting keywords.

Algorithms	Max	Min	Mean	Average standard error
TF-IDF	1	0	0.52	0.16
RAKE	1	0.3	0.67	0.11
KeyBERT	1	0.5	0.82	0.10

Table 6
A comparison of novelty keywords measured by alternative methods.

Combination measure	Proportion of novelty keywords\%	Combination measure	Proportion of novelty keywords\%
BERT+LOF	62	BERT+IF	70
RoBERTa+LOF	88	RoBERTa+IF	70
SciBERT+LOF	80	SciBERT+IF	72
Word2vec+LOF	82	Word2vec+IF	76

Table 7
Comparative analysis of core value indicators of breakthrough patents, non-breakthrough patents, and ordinary patents.

	Average number of forward citations	Average number of transfers	Average technology advancement	Average patent value
Breakthrough patents	3.46	1.83	8.78	8.95
Non-breakthrough patents	2.81	1.38	7.29	8.33
Ordinary patents	2.73	1	6.91	8.29

unique competitive advantages and informed decision support for decision-makers across government, industry, and academia, playing a crucial role in policy-making, technological planning, and innovation management.

EnclosedCircleChinese2 We designed a two-stage funnel selection method for identifying scientific signals. This method characterizes and measures the core features of scientific signals from two perspectives: weak signal assessment and novelty assessment. In the first stage, bibliometric techniques are employed to evaluate weak signals, while in the second stage, machine learning methods are used to assess novelty. By using this approach, we can effectively identify scientific signals with the potential to trigger technological breakthroughs, thereby addressing the core question of which scientific signals drive technological breakthroughs.

EnclosedCircleChinese3 We developed a machine learning-based link prediction method that integrates three types of features to track the diffusion path of scientific signals. This method combines network topology, node centrality, and textual semantics, revealing the deep links between scientific signals and technological terms. This method not only effectively promotes the knowledge flow and diffusion from basic scientific research to applied research but also provides a powerful tool for advancing the integration of S&T, accelerating the knowledge transfer of scientific innovation results. This cross-disciplinary knowledge integration enables inventors to break through original knowledge boundaries, opening new research directions and innovation paths. It is of great significance for promoting cross-field collaboration and enhancing the overall efficiency of scientific research.

5.3. Limitations

Despite its contributions, this study has some limitations. First, this research mainly considers the driving role of science in technological breakthroughs. However, future research could further explore the identification of breakthrough technologies under different interaction modes between science and technology. Second, beyond the driving role of science, commercial drivers are also important for technological development. Future research could integrate more types of data sources to explore the influence mechanisms of various factors on technological breakthroughs. Lastly, in evaluating breakthrough technologies, this study primarily focuses on the technical aspects of the inventions. Future evaluations could consider incorporating market dimensions, social impact, policy support, and other characteristics for a more comprehensive assessment of breakthrough technologies.

CRediT authorship contribution statement

Dan Wang: Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Xiao Zhou:** Writing – review & editing, Writing – original draft, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Pengwei Zhao:** Writing – review & editing, Formal analysis, Conceptualization. **Juan Pang:** Writing – review & editing, Data curation. **Qiaoyang Ren:** Writing – review & editing, Visualization.

Acknowledgment

This work was supported by the General Program of National Natural Science Foundation of China [grant numbers 72374165].

References

Anderson, P., & Tushman, M. L. (1990). Technological discontinuities and dominant designs: A cyclical model of technological change. *Administrative Science Quarterly, 35*(4), 604–633. https://doi.org/10.1007/s00191-020-00661-z

Ansoff, H. I. (1980). Strategic issue management. *Strategic Management Journal, 1*(2), 131–148. https://doi.org/10.1002/smj.4250010204

Arora, A., Belenzon, S., & Suh, J. (2022). Science and the market for technology. *Management Science, 68*(10), 7176–7201. https://doi.org/10.1287/mnsc.2021.4268

Ba, Z., Meng, K., Ma, Y., & Xia, Y. (2024). Discovering technological opportunities by identifying dynamic structure-coupling patterns and lead-lag distance between science and technology. *Technological Forecasting and Social Change, 200*, Article 123147. https://doi.org/10.1016/j.techfore.2023.123147

Balconi, M., Brusoni, S., & Orsenigo, L. (2010). In defence of the linear model: An essay. *Research Policy, 39*(1), 1–13. https://doi.org/10.1016/j.respol.2009.09.013

Block, C., Wustmans, M., Laibach, N., & Bröring, S. (2021). Semantic bridging of patents and scientific publications–The case of an emerging sustainability-oriented technology. *Technological Forecasting and Social Change, 167*, Article 120689. https://doi.org/10.1016/j.techfore.2021.120689

Bornmann, L. (2014). How are excellent (highly cited) papers defined in bibliometrics? A quantitative analysis of the literature. *Research Evaluation, 23*(2), 166–173. https://doi.org/10.1093/reseval/rvu002

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on management of data* (pp. 93–104). https://doi.org/10.1145/342009.335388

Capponi, G., Martinelli, A., & Nuvolari, A. (2022). Breakthrough innovations and where to find them. *Research Policy, 51*(1), Article 104376. https://doi.org/10.1016/j.respol.2021.104376

Chen, C. (2012). *Turning points: The nature of creativity*. Springer Science & Business Media.

Chen, X., Ye, P., Huang, L., Wang, C., Cai, Y., Deng, L., & Ren, H. (2023). Exploring science-technology linkages: A deep learning-empowered solution. *Information Processing & Management, 60*(2), Article 103255. https://doi.org/10.1016/j.ipm.2022.103255

Choi, J., & Yoon, J. (2022). Measuring knowledge exploration distance at the patent level: Application of network embedding and citation analysis. *Journal of Informetrics, 16*(2), Article 101286. https://doi.org/10.1016/j.joi.2022.101286

Chung, J., Ko, N., Kim, H., & Yoon, J. (2021). Inventor profile mining approach for prospective human resource scouting. *Journal of Informetrics, 15*(1), Article 101103. https://doi.org/10.1016/j.joi.2020.101103

Clauset, A., Moore, C., & Newman, M. E. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature, 453*(7191), 98–101. https://doi.org/10.1038/nature06830

Dedrick, J., & Kraemer, K. L. (2015). Who captures value from science-based innovation? The distribution of benefits from GMR in the hard disk drive industry. *Research Policy, 44*(8), 1615–1628. https://doi.org/10.1016/j.respol.2015.06.011

Dixon, T., Eames, M., Britnell, J., Watson, G. B., & Hunt, M. (2014). Urban retrofitting: Identifying disruptive and sustaining technologies using performative and foresight techniques. *Technological Forecasting and Social Change, 89*, 131–144. https://doi.org/10.1016/j.techfore.2013.08.027

Ettlie, J. E., Bridges, W. P., & O'keefe, R. D. (1984). Organization strategy and structural differences for radical versus incremental innovation. *Management Science, 30*(6), 682–695. https://doi.org/10.1287/mnsc.30.6.6822

Fleming, L., & Sorenson, O. (2004). Science as a map in technological search. *Strategic Management Journal, 25*(8-9), 909–928. https://doi.org/10.1002/smj.384

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research, 3*, 1289–1305.

Funk, R. J., & Owen-Smith, J. (2017). A dynamic network measure of technological change. *Management Science, 63*(3), 791–817. https://doi.org/10.1287/mnsc.2015.2366

Han, J., Jeon, B., & Geum, Y. (2022). Link prediction revisited: New approach for anticipating new innovation chances using technology convergence. *IEEE Transactions on Engineering Management*, 1–17. https://doi.org/10.1109/TEM.2022.3213867

Hiltunen, E. (2008). The future sign and its three dimensions. *Futures, 40*(3), 247–260. https://doi.org/10.1016/j.futures.2007.08.021

Hu, Z., Cui, J., & Lin, A. (2023). Identifying potentially excellent publications using a citation-based machine learning approach. *Information Processing & Management, 60*(3), Article 103323. https://doi.org/10.1016/j.ipm.2023.103323

Huang, Y., Zhang, T., & Xu, C. (2023). Learning to decode to future success for multi-modal neural machine translation. *Journal of Engineering Research, 11*(2), Article 100084. https://doi.org/10.1016/j.jer.2023.100084

Jang, H. J., Woo, H. G., & Lee, C. (2017). Hawkes process-based technology impact analysis. *Journal of Informetrics, 11*(2), 511–529. https://doi.org/10.1016/j.joi.2017.03.007

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

Ke, Q. (2020). Technological impact of biomedical research: The role of basicness and novelty. *Research Policy, 49*(7), Article 104071. https://doi.org/10.1016/j.respol.2020.104071

Khan, M. Q., Shahid, A., Uddin, M. I., Roman, M., Alharbi, A., Alosaimi, W., et al. (2022). Impact analysis of keyword extraction using contextual word embedding. *PeerJ Computer Science, 8*, e967. https://doi.org/10.7717/peerj-cs.967

Li, K., & Yan, E. (2019). Are NIH-funded publications fulfilling the proposed research? An examination of concept-matchedness between NIH research grants and their supported publications. *Journal of Informetrics, 13*(1), 226–237. https://doi.org/10.7717/10.1016/j.joi.2019.01.001

Li, X., Ma, X., & Feng, Y. (2024). Early identification of breakthrough research from sleeping beauties using machine learning. *Journal of Informetrics, 18*(2), Article 101517. https://doi.org/10.1016/j.joi.2024.101517

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Luke Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized Bert pretraining approach. arXiv preprint arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692.

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. In *Paper presented at the 2008 eighth IEEE international conference on data mining* (pp. 413–422).

Lü, L., & Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and Its Applications, 390*(6), 1150–1170. https://doi.org/10.1016/j.physa.2010.11.027

Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: How may AI and GPT impact academia and libraries? *Library Hi Tech News, 40*(3), 26–29. https://doi.org/10.1108/LHTN-01-2023-0009

Ma, M., Mao, J., & Li, G. (2024). Discovering weak signals of emerging topics with a triple-dimensional framework. *Information Processing & Management, 61*(5), Article 103793. https://doi.org/10.1016/j.ipm.2024.103793

Malva, A. D., Kelchtermans, S., Leten, B., & Veugelers, R. (2015). Basic science as a prescription for breakthrough inventions in the pharmaceutical industry. *The Journal of Technology Transfer, 40*, 670–695. https://doi.org/10.1007/s10961-014-9362-y

Martínez, V., Berzal, F., & Cubero, J. C. (2016). A survey of link prediction in complex networks. *ACM Computing Surveys (CSUR), 49*(4), 1–33. https://doi.org/10.1145/3012704

Min, C.& Ke, Q. (2021). Temporal search in the scientific space predicts breakthrough inventions. arXiv preprint arXiv:2107.09176. https://doi.org/10.48550/arXiv.2107.09176.

Mukherjee, S., Romero, D. M., Jones, B., & Uzzi, B. (2017). The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: The hotspot. *Science Advances, 3*(4), Article E1601315. https://doi.org/10.1126/sciadv.1601315

Narayanamurti, V., & Tsao, J. Y. (2024). How technoscientific knowledge advances: A Bell-Labs-inspired architecture. *Research Policy, 53*(4), Article 104983. https://doi.org/10.1016/j.respol.2024.104983

Parés, F., Gasulla, D. G., Vilalta, A., Moreno, J., Ayguadé, E., Labarta, J., et al. (2017). Fluid communities: A competitive, scalable and diverse community detection algorithm. In *Paper presented at the sixth international conference on complex networks and their applications* (pp. 229–240). https://doi.org/10.1007/978-3-319-72150-7_19

Park, M., & Geum, Y. (2022). Two-stage technology opportunity discovery for firm-level decision making: GCN-based link-prediction approach. *Technological Forecasting and Social Change, 183*, Article 121934. https://doi.org/10.1016/j.respol.2024.104983

Petrescu, A. S. (2009). Science and technology for economic growth. New insights from when the data contradicts desktop models 1. *Review of Policy Research, 26*(6), 839–880. https://doi.org/10.1111/j.1541-1338.2009.00420.x

Poege, F., Harhof, D., Gaessler, F., & Barufaldi, S. (2019). Science quality and the value of inventions. *Science Advances, 5*(12), eaay7323. https://doi.org/10.1126/sciadv.aay7323

Porter, A. L., Garner, J., Carley, S. F., & Newman, N. C. (2019). Emergence scoring to identify frontier R&D topics and key players. *Technological Forecasting and Social Change, 146*(March), 628–643. https://doi.org/10.1016/j.techfore.2018.04.016

Roh, T., & Yoon, B. (2023). Discovering technology and science innovation opportunity based on sentence generation algorithm. *Journal of Informetrics, 17*(2), Article 101403. https://doi.org/10.1016/j.joi.2023.101403

Rosenberg, N., & Birdzell, L. E. (1990). Science, technology and the Western miracle. *Scientific American, 263*(5), 42–55. https://www.jstor.org/stable/24996974.

Sainio, L. M., & Puumalainen, K. (2007). Evaluating technology disruptiveness in a strategic corporate context: A case study. *Technological Forecasting and Social Change, 74*(8), 1315–1333. https://doi.org/10.1016/j.techfore.2006.12.004

Schumpeter, J. A. (1934). *The theory of economic development*. Harvard University Press.

Shooshtarian, S., Gurmu, A. T., & Sadick, A. M. (2023). Application of natural language processing in residential building defects analysis: Australian stakeholders' perceptions, causes and types. *Engineering Applications of Artificial Intelligence, 126*, Article 107178. https://doi.org/10.1016/j.engappai.2023.107178

Singh, N., & Kapoor, R. (2023). Multi-modal Expression Detection (MED): A cutting-edge review of current trends, challenges and solutions. *Engineering Applications of Artificial Intelligence, 125*, Article 106661. https://doi.org/10.1016/j.engappai.2023.106661

Sun, B., Kolesnikov, S., Goldstein, A., & Chan, G. (2021). A dynamic approach for identifying technological breakthroughs with an application in solar photovoltaics. *Technological Forecasting and Social Change, 165*, Article 120534. https://doi.org/10.1016/j.techfore.2020.120534

Tsay, M. Y., & Liu, Z. W. (2020). Analysis of the patent cooperation network in global artificial intelligence technologies based on the assignees. *World Patent Information, 63*, Article 102000. https://doi.org/10.1016/j.wpi.2020.102000

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., & Lee, J. J. (2023). Predicting and analyzing technology convergence for exploring technological opportunities in the smart health industry. *Computers & Industrial Engineering, 182*, Article 109352. https://doi.org/10.1016/j.cie.2023.109352

Wang, L., & Li, Z. (2021). Knowledge flows from public science to industrial technologies. *The Journal of Technology Transfer, 46*, 1232–1255. https://doi.org/10.1007/s10961-019-09738-9

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

Wuestman, M., Hoekman, J., & Frenken, K. (2020). A typology of scientific breakthroughs. *Quantitative Science Studies, 1*(3), 1203–1222. https://doi.org/10.1162/qss_a_00079

Xu, H., Luo, R., Winnink, J., Wang, C., & Elahi, E. (2022a). A methodology for identifying breakthrough topics using structural entropy. *Information Processing & Management, 59*(2), Article 102862. https://doi.org/10.1016/j.ipm.2021.102862

Xu, H., Winnink, J., Pang, H., Wen, S., & Chen, L. (2023). Breakthrough potential of emerging research topics based on citation diffusion features. *Journal of Information Science, 49*(5), 1390–1416. https://doi.org/10.1177/01655515211061219

Xu, H., Yue, Z., Pang, H., Elahi, E., Li, J., & Wang, L. (2022b). Integrative model for discovering linked topics in science and technology. *Journal of Informetrics, 16*(2), Article 101265. https://doi.org/10.1016/j.joi.2022.101265

Yoon, J. (2012). Detecting weak signals for long-term business opportunities using text mining of Web news. *Expert Systems with Applications, 39*(16), 12543–12550. https://doi.org/10.1016/j.eswa.2012.04.059

Yu, D., & Yan, Z. (2022). Combining machine learning and main path analysis to identify research front: From the perspective of science-technology linkage. *Scientometrics, 127*(7), 4251–4274. https://doi.org/10.1007/s11192-022-04443-1

Zhang, M., & Li, J. (2021). A commentary of GPT-3 in MIT technology review 2021. *Fundamental Research, 1*(6), 831–833. https://doi.org/10.1016/j.fmre.2021.11.011

Zhang, N., You, D., Tang, L., & Wen, K. (2023). Knowledge path dependence, external connection, and radical inventions: Evidence from Chinese Academy of Sciences. *Research Policy, 52*(4), Article 104738. https://doi.org/10.1016/j.respol.2023.104738