Scientometrics
https://doi.org/10.1007/s11192-025-05497-7

Interpretable XGBoost-SHAP machine learning model for identifying scientific breakthroughs

Libo Sheng[1,2] · Xuanmin Ruan[3] · Yi Wang[1,2] · Dongqing Lyu[4] · Ying Cheng[1,2]

Received: 20 June 2024 / Accepted: 17 November 2025
© Akadémiai Kiadó Zrt 2025

Abstract

Identifying scientific breakthroughs is of great significance for research evaluation and policy-making. Thus, it has been the central focus in the realm of science. This study leverages a new dataset of Nobel and Lasker prize-winning publications and employs the eXtreme Gradient Boosting (XGBoost) algorithm to establish a predictive model for scientific breakthroughs. The Input-Process-Output-Outcome (IPOO) framework serves as the fundamental perspective to deconstruct the potential factors associated with breakthroughs into four dimensions: input, process, output, and outcome. We demonstrate that XGBoost achieves the best predictive accuracy among traditional machine learning models, with F1 scores of 0.613 and 0.611 in Dataset 1 and Dataset 2, and AUC values of 0.898 and 0.880, respectively. Large language models (LLMs), used as additional baselines, exhibit higher recall scores on both datasets. In addition, we utilize the SHapley Additive exPlanations (SHAP) approach to enhance the interpretability of our model, enabling a deeper understanding of how features influence the prediction of scientific breakthroughs, which has been overlooked in previous research. This study introduces an explainable machine learning approach for tracing breakthrough research in science with bibliographic information, yielding valuable insights into future research.

Keywords Scientific breakthrough · Nobel prize · Lasker prize · Disruption · SHAP · Machine learning

Libo Sheng and Xuanmin Ruan have contributed equally to this work.

✉ Ying Cheng
chengy@nju.edu.cn

[1] Laboratory of Data Intelligence and Interdisciplinary Innovation, Nanjing University, Nanjing, China

[2] School of Information Management, Nanjing University, Nanjing, China

[3] Faculty of Humanities and Social Sciences, Nanjing Forestry University, Nanjing, China

[4] College of Information Engineering, Nanjing University of Finance & Economics, Nanjing, China

Published online: 01 December 2025

Introduction

It is generally acknowledged that breakthroughs at the cutting edge of science are linked to exceptional innovation (Häyrynen, 2007). Although such discoveries may be few in number, they have the potential to challenge the established paradigm and cause radical changes in our perception of the world. These discoveries are also recognized as crucial for further scientific progress and may pave the way for technological applications (Kuhn, 1970; Winnink et al., 2019). Thus, facilitating and fostering scientific breakthroughs has attracted increased attention in various countries (Häyrynen, 2007; Wang et al., 2021).

The identification of scientific breakthroughs is of great interest to a wide range of scholars in the realm of science. The availability of harvest databases with bibliographic data from publications makes it possible to use bibliographic information to quantitively detect such discoveries (Min et al., 2021a; Winnink & Tijssen, 2015; Winnink et al., 2019). Specifically, existing literature has used citation count-based models (Ponomarev et al., 2014a, 2014b; Schneider & Costas, 2017; Winnink et al., 2019), citation network-based models (Funk & Owen-Smith, 2017; Min et al., 2021a, 2021b; Wang et al., 2023a; Wei et al., 2023; Wu et al., 2019), and novelty indicators (Savov et al., 2020; Wang et al., 2017) to identify scientific breakthroughs. However, these methods mainly rely on the *ex-post* measure of impact and are controversial due to a concentration on a specific property of breakthroughs. For example, citation-based analysis is biased as it fails to comprehensively assess the innovativeness of scholarly publications (Xu et al., 2022c). In addition, prior studies have shown considerable inconsistency in the criteria for defining scientific breakthroughs. Some scholars have used the quantitative approach to define breakthroughs, such as operationalizing breakthroughs as the top 0.1%, 1%, or 10% of highly cited publications. Others have used peer review results to determine breakthroughs, as noted by Schneider and Costas (2017). They contended that “What eventually is considered breakthrough research is a matter to be decided by peers” (p. 711). However, previous studies mainly relied on a limited dataset consisting solely of Nobel prize-winning publications. Unlike previous research, we consider discoveries that have won significant prizes as scientific breakthroughs, including the Nobel Prize and the Lasker Prize. It could complement the larger sample of scientific breakthroughs compared to using only examining Nobel prize-winning papers.

This paper aims to predict major discoveries based on peer review using machine learning methods. This approach enhances prior research in two unique ways. First, this study represents the first attempt to adopt the Input-Process-Output-Outcome (IPOO) framework as a fundamental lens for predicting scientific breakthroughs. This framework systematically deconstructs potential factors into input, process, output, and outcome dimensions. It also enables a more structured and interpretable foundation for breakthrough forecasting. Second, while machine learning methods have advanced researchers’ capability to predict innovativeness, it introduces limitations from the inherently “black box” nature of the prediction process, which hinders interpretability. Therefore, the machine-learning-based SHapley Additive exPlanations (SHAP) approach is used in our study to decode and elucidate the influence of features related to scientific breakthroughs.

Review

Definitions and characteristics of breakthroughs

Kuhn (1970) theorized that scientific progress does not follow a cumulative unified path, but follows nonlinear laws. Specifically, two states of "normal" science and "revolutionary" science appear alternatively. Scientific advancements are not solely dependent on numerous small, incremental advances that are carried out within existing and accepted pathways. They are also driven by occasional major discoveries that alter the existing paradigm, leading to dramatic changes in science. Breakthroughs are usually aligned with these latter discoveries, with scientists using synonyms such as "revolutionary discoveries" (Kuhn, 1970), "transformative research" (Chen et al., 2009) or "disruptive research" (Wu et al., 2019). The lack of a generally accepted definition for breakthroughs is illustrated by these varied synonyms. To date, no consensus has been reached on what constitutes such research, and the definition throughout the scientific community is not specific.

The concept of scientific breakthroughs has been interpreted in various ways. For example, Winnink (2017) stated that breakthroughs are discoveries that have a major impact on science. Breakthroughs refer to advancements that are highly useful to numerous scientists in addressing scientific problems (Hollingsworth, 2008). Based on this view, an essential feature of scientific breakthroughs is a study's "major impact," which has the potential to influence subsequent studies (Wang et al., 2023a) and contribute to further progress in science. However, remarkably, breakthroughs also have an impact that goes beyond its own domain to impact other fields of science. Schneider and Costas (2017, p. 711) indicated that breakthroughs lead to "important citation spread over its own field and also other fields of science."

Existing research has also indicated that breakthroughs do not follow existing findings and must "have a genuine relevance on its own" (Schneider & Costas, 2017, p. 711). This relevance usually requires novel approaches (Uzzi et al., 2013; Wang et al., 2017), or a new way of thinking about a problem (Hollingsworth, 2008). The distinctive nature may consequently lead to "reorientations of established research streams onto new frontiers" (Wang et al., 2023a, p. 3), or "dramatically change the direction of future research" (Wei et al., 2023, p. 1). These views emphasize "originality" as the key characteristic that distinguishes breakthroughs from non-breakthroughs. Wang et al., (2023a, p. 3) also mentioned that breakthroughs are innovative discoveries that "make an original contribution to the knowledge system of science." This new knowledge plays a key role in paving the way for a new avenue of exploration.

Prediction of scientific breakthroughs

In bibliometrics and scientometrics, much attention has been paid to predicting scientific success such as predicting the impact of a paper (Hu et al., 2023; Wang et al., 2019a, 2019b), the success of a scientist (Daud et al., 2015; Kumar et al., 2023), the success of research collaborations (Hückstädt, 2023), and research grants (Tohalino & Amancio, 2022). In this study, we focus on scientific breakthroughs and review related literature using bibliometrics and machine learning methods.

Predicting scientific breakthroughs using bibliometric methods

Scientific breakthroughs are usually studied quantitatively. Using bibliometric information to identify breakthroughs has been an aim for decades. The three mainstream approaches are citation count-based, citation network-based, and novelty-based methods. One way to understand breakthrough research is to characterize it as a highly cited discovery (Mukherjee et al., 2017; Schilling & Green, 2011; Uzzi et al., 2013). It is based on the assumption that followers who are inspired by or build on previous work acknowledge its value by citing it, and the number of citations implies its impact on the scientific community (Lee et al., 2015; Mugabushaka et al., 2020). Previous research has modeled predictive citation count-based approaches to detect breakthroughs (Ponomarev et al., 2014a, 2014b; Schneider & Costas, 2017; Winnink et al., 2019). For example, Ponomarev et al. (2014b) fit linear and nonlinear models to citation data based on early citation counts (at the 6th, 12th, and 24th months) to predict later citation counts in the fifth year. The predicted values were compared with quantile thresholds to determine whether a paper could be classified as a breakthrough.

Detecting scientific breakthroughs from a knowledge structure perspective has been of interest to scholars. The hypothesis is based on the fact that breakthrough discoveries are linked to dramatic, structural changes in the existing body of knowledge in science. The potential value of a discovery can be measured by the degree of structure change it brings to the intellectual space (Chen, 2012; Xu et al., 2022c). Profound scientific discoveries usually arise from structural holes in the intellectual network, and such discoveries establish unexpected linkages between structures of knowledge (Chen et al., 2009). Based on structural-entropy methods, Xu et al. (2022c) pointed out that detecting significant knowledge-structure variations is useful for identifying the generation of breakthroughs. Funk and Owen-Smith (2017) and Wu et al. (2019) also captured the degree of disruption to the existing knowledge structure caused by a focal paper by examining the extent to which future research deviates from its intellectual forebearers. As a result, they designed the disruption indicator. However, recently, improved metrics to measure disruption have been developed to quantify and predict scientific breakthroughs (Lin et al., 2025; Wang et al., 2023a; Wei et al., 2023). Lin et al. (2025) introduced a two-dimensional metric that integrates the dimensions of disruption and scientific impact, considering both the breadth and depth of the impact. The results showed that the CIB index achieved the highest AUC score (0.79) for identifying scientific breakthroughs in the computer science field.

Another stream of research has focused on identifying breakthroughs based on the novelty of the discovery. The underlying concept is that scientific breakthroughs often require and are driven by novel approaches (Veugelers & Wang, 2019; Wang et al., 2017). From a knowledge combination perspective (Schumpeter, 1939), novelty is derived from the combination of existing bits of knowledge in an unusual or unprecedented way (Uzzi et al., 2013; Wang et al., 2017). Previous studies have concluded that novelty is an essential property of creative ideas (Lin et al., 2022; Ruan et al., 2023; Sheng et al., 2023; Uzzi et al., 2013; Wang et al., 2017). Savov et al. (2020) also identified breakthroughs by concentrating on the novelty expressed in the paper. They devised an innovation score to identify breakthroughs based on the assumption that the less similar the topic is to the past (the more similar to future papers), the more innovative it is.

Predicting scientific breakthroughs using machine learning methods

Machine learning methods that mine bibliometric information have become one of the most powerful tools to predict scientific breakthroughs. Some studies have explored and constructed citation-related features to predict scientific breakthroughs. For example, Min et al. (2021a) developed metrics based on the citing citation network and employed a logistic regression model to classify Nobel prize-winning papers and their counterexamples. The study revealed disciplinary differences, with the optimal models achieving AUC scores of 0.657 and 0.695 in the natural sciences and economics, respectively. Building on their work, Yu et al. (2024) further proposed an optimization strategy from the perspective of dynamic citation structures, capturing information from snapshots of 90 citation cascade networks. The model enhanced the prediction accuracy, improving the AUC by 7%. In addition to citation-related features, other bibliographic information has been adopted such as paper-, journal-, and author-related features (Tahamtan et al., 2016) to predict breakthroughs (Li et al., 2022, 2024; Wolcott et al., 2016). Wolcott et al. (2016) extracted bibliographic information about the papers, journals, authors, and citations, and developed a random forest predictive model to identify potential breakthroughs at earlier stages.

Other studies have focused on mining the in-depth content of articles using natural language processing (NLP) (Savov et al., 2020; Wang et al., 2021). Savov et al. (2020) applied the LDA model and support vector machine (SVM) to predict publication dates. They characterized a paper's innovation based on the degree to which the predicted date preceded or lagged behind the actual publication date, thereby identifying potentially groundbreaking research. Wang et al. (2021) proposed a breakthrough identification method that combined self-evaluation and others' evaluation of the significance of the work. They demonstrated that breakthroughs are linked to positive words about the ideas in the abstracts or citing sentences, such as "first," "new," and "novel." Using the deep learning approach, they identified breakthrough research using judgement sentences with positive words. More recently, Yu and Liang (2024) proposed a prediction framework based on graph signal processing, which integrates multi-dimensional information, including textual content and citation structures, achieving an AUC of approximately 80%.

Comparison with existing work

In summary, prior studies have predicted scientific breakthroughs from various perspectives, mainly using bibliometric and machine learning methods. Although considerable efforts have been dedicated to detecting such discoveries, several limitations remain. First, regarding feature selection to predict scientific breakthroughs, previous studies have concentrated more on *ex-post* measures based on citation-related features. However, the factors associated with breakthroughs have not been fully investigated. To identify potential factors associated with breakthroughs, we adopt the IPOO model comprising four key dimensions: input, process, output, and outcome. Specifically, we incorporate new features including combination recency, impact, novelty, and homogeneity of knowledge inputs in a paper. We also incorporate important antecedents such as researchers' knowledge base and experience. Previous studies have identified these factors as influencing breakthrough ideas. Ignoring such factors may limit the accuracy of scientific breakthrough predictions.

The second limitation is that machine learning algorithms are generally black-box models and lack adequate interpretability to elucidate how the prediction is made. Thus, previous interpretable machine learning models to predict breakthroughs are limited and have garnered little attention. This study addresses these research gaps and endeavors to interpret the black-box nature of machine learning in predicting scientific breakthroughs using the SHAP approach. Specifically, we analyze feature contributions to identify important predictors. We also reveal previously uncovered relationships between the predictors and scientific breakthroughs.

Another limitation is that using highly cited achievements to predict scientific breakthroughs may lead to a biased sample, making the model and results less reliable. Citations are subject to bias, as they are influenced by many factors that are unrelated to the content of the paper (Lyu et al., 2021b). For example, several scholars have identified the "Matthew effect" or "Nobel prize effect" indicating that eminent scientists' or laureates' publications are usually given more credit (Dong et al., 2023; Frandsen & Nicolaisen, 2013; Liao, 2021). Therefore, adopting a percentile approach that operationalizes breakthroughs as highly cited papers is not appropriate, such as identifying the 0.1%, 1%, or 10% most cited publications. Determining whether a work is a breakthrough requires the judgement of peers (Min et al., 2021a; Schneider & Costas, 2017). Our study considers prize-winning publications as the gold standard for breakthroughs. However, an expanded dataset of Nobel prize-winning and Lasker prize-winning publications can serve as a supplement to the data on scientific breakthroughs.

Data construction

Works that are recognized by the scientific community and authoritative organizations in the form of a prize or honor can generally be considered scientific breakthroughs (Mugabushaka et al., 2020). In our study, prize-winning papers, including Nobel and Lasker prize-winning papers, are regarded as ground-truth scientific breakthroughs. These prize-winning papers have not only made original contributions to the stock of knowledge, which is one of the foundations on which they are rewarded (Mugabushaka et al., 2020; Schneider & Costas, 2017), but they have also had a profound impact on science and society.

The steps for data collection are as follows. We first collected Nobel prize-winning and Lasker prize-winning papers. Specifically, Nobel prize-winning papers were drawn from the dataset constructed by Li et al. (2019a), which includes 545 Nobel laureates with 874 prize-winning papers from 1900 to 2016 in the fields of physics, chemistry, and medicine. Details of each prize-winning paper including the "Laureate name," "Title," and "Journal" are available at https://dataverse.harvard.edu/. The Lasker prize-winning papers were crawled from the official website (see https://laskerfoundation.org/all-awards-winners/), which lists selected publications for each Lasker winner from 1998 to 2022. A total of 107 Lasker winners with 661 prize-winning papers were obtained from the website. It should be acknowledged that nineteen individuals have been awarded both the Nobel and Lasker prizes, so their Nobel or Lasker prize-winning papers were merged in the collective body of the author's papers.[1]

In the second step, we linked the available Nobel and Lasker prize-winning papers to the PubMed Knowledge Graph (PKG) dataset (Xu et al., 2020) to access the bibliographic

[1] The 19 winners are: David Baltimore, Thomas C. Südhof, John Gurdon, Shinya Yamanaka, Ralph M. Steinman, Elizabeth H. Blackburn, Carol W. Greider, Jack W. Szostak, James Rothman, Randy Schekman, Mario Capecchi, Oliver Smithies, Robert Edwards, Aaron Ciechanover, Avram Hershko, Sydney Brenner, Roderick MacKinnon, Lee Hartwell, Paul Nurse.

information of each paper. For each paper, the unique PMID (a unique identifier for each paper) was obtained by manually matching and checking with the PKG dataset according to the title, publication year, journal, and author list. Our dataset included 952 award-winning publications from 305 laureates with 855 unique breakthrough papers. Figure 1 presents the distribution of winning-prize papers annually from 1910 to 2019.

In the third stage, we systematically matched each breakthrough paper with a set of non-breakthrough counterparts based on established practices in the science of science. Specifically, we adopted two well-established criteria from prior literature: (1) selecting non-award-winning papers authored by the same researcher (Capponi et al., 2022); (2) pairing breakthrough papers with non-breakthrough papers published in the same journal and year (Min et al., 2021a; Wei et al., 2023). Based on these criteria, we constructed two datasets. When determining the ratio of scientific breakthroughs to non-breakthroughs, we follow previous empirical studies (Wang et al., 2023a; Wolcott et al., 2016) and set the ratio to approximately 1:5, meaning that at least five non-breakthrough papers are matched to each breakthrough paper.

Methodology

We begin by generating a set of features to predict scientific breakthroughs within the IPOO framework: input-related, process-related, output-related, and outcome-related features. We then introduce the data splitting, data pre-processing process and oversampling, and the machine learning models used in our study. Finally, we describe SHapley Additive exPlanations (SHAP) and provide explanations for the best identification model. The overall process of the machine-learning-based SHAP approach is shown in Fig. 2.

Feature selection framework

The Input-Process-Output-Outcome (IPOO) model has been adopted as a systematic theoretical framework for performance evaluation across various fields (Cammarano et al., 2022; Choi & Choi, 2014; Ferreira et al., 2018; Hsu et al., 2020). It embodies a "systems view" that conceptualizes performance indicators into four stages: input, process, output, and outcome (Cammarano et al., 2022; Choi & Choi, 2014). Building on this conceptual

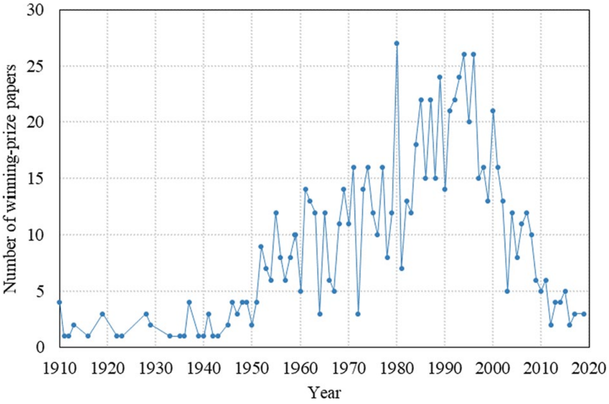


Fig. 1 Distribution of prize-winning papers

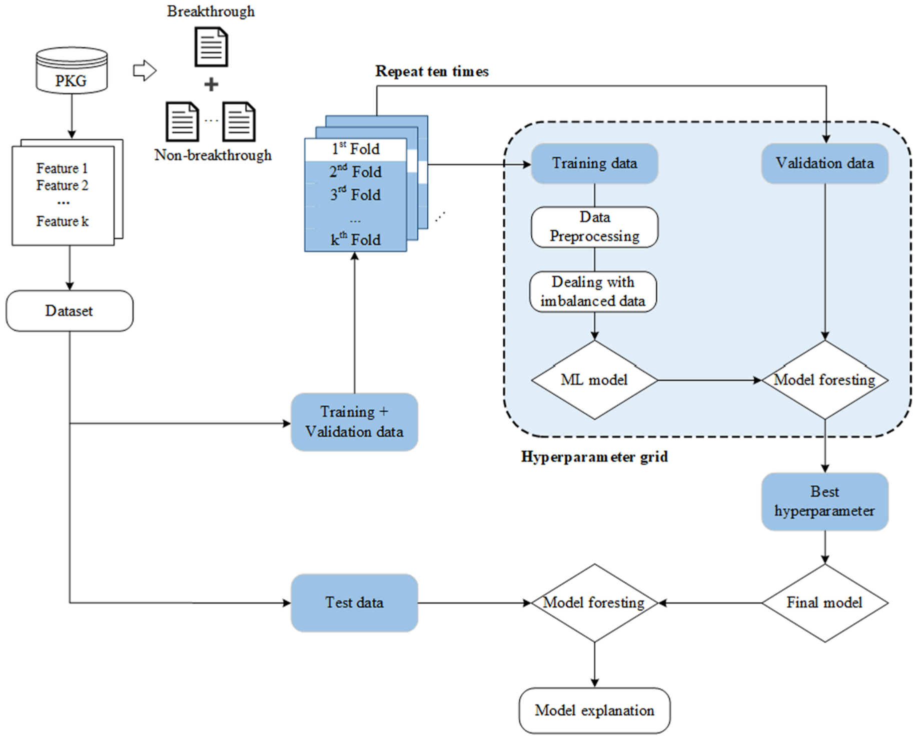


Fig. 2 The methodology framework

foundation, we posit that scientific practice can also be considered a system (Winnink et al., 2019). Therefore, we employ the IPOO model to uncover the predictive features associated with scientific breakthroughs and then structure the resulting typology of features according to these four dimensions.

Inputs are antecedent factors that are the "raw materials" or resources used in the subsequent processes (de Carvalho et al., 2017). Knowledge resources are the key inputs for future innovation (Cammarano et al., 2022; Chen et al., 2021). In scientific practice, it is widely acknowledged that new ideas rarely come from nothing but, rather, scientists recombine different streams of existing knowledge (Hur & Oh, 2021; Liang et al., 2020; Mukherjee et al., 2017). The fundamental bits of knowledge constituting an innovative idea (i.e., prior knowledge on which the new idea is built) are the "raw materials" or "ingredients" that innovators combine to form outputs (Petruzzelli et al., 2018). In other words, prior knowledge is the key input that affects the generation of new ideas (Heeley & Jacobson, 2008). In particular, the characteristics of prior knowledge have been identified as crucial determinants of innovation success (Heeley & Jacobson, 2008; Papazoglou & Nelles, 2023). Previous studies have confirmed that the amount (Schoenmakers & Duysters, 2010), recency (Katila, 2002; Liang et al., 2020; Nerkar, 2003; Papazoglou & Nelles, 2023; Petruzzelli et al., 2018), impact (Kwon & Geum, 2020; Mukherjee et al., 2017), combination novelty (Lin et al., 2022; Wang et al., 2017), and homogeneity (Hur & Oh, 2021) of prior knowledge can lead to varying levels of innovation performance. Therefore, we consider these five factors within this dimension.

Processes refer to the transformation of these inputs into meaningful outputs (Marks et al., 2001). In the context of scientific research, processes are related to the innovation activities carried out or implemented to achieve the goals of new discoveries. Processes are usually associated with creation, and innovators are seen as the heart of innovative processes (Jones, 2009; Lee et al., 2015; Wagner et al., 2011). Integrating various knowledge sources and developing logical linkages are usually accomplished by members of a team (Dahlin et al., 2005; Porter & Rafols, 2009). Therefore, team members play a vital role in driving innovative outputs, and their diverse characteristics contribute to different levels of innovation performance (Ma et al., 2023). Previous studies have investigated factors of knowledge-producing teams that are associated with innovative discoveries. These factors which include (1) team composition (e.g., gender, career age, and organizational background) (Ao et al., 2023; Li et al., 2019b; Yang et al., 2022); (2) team structure (Xu et al., 2022a, 2022b); (3) team collaboration, including collaboration size, inter-institutional or international collaboration, and collaboration freshness (Lyu et al., 2021a; Wu et al., 2019; Zeng et al., 2021); (4) the strategies for selecting the topic (e.g., diversity, popularity and novelty) (Chai & Menon, 2019; Ruan et al., 2023); (5) team members' knowledge variety and heterogeneity (Huo et al., 2019; Ma et al., 2023); (6) team members' productivity and citations (Wang et al., 2012, 2019a); (7) team members' social ties (Wang et al., 2023b), and (8) funding received (Lyu et al., 2021a). Given this evidence, we consider these factors within the dimension, as they have been shown to be supportive conditions for innovative outputs.

Outputs are the final results produced by the system (MacCuspie et al., 2014). The innovative outputs in the context of scientific research usually take the form of papers published in journals. Within this dimension, we take into consideration detailed paper-related and journal-related factors (Tahamtan et al., 2016; Wang et al., 2012).

Outcomes refer to the effects generated by these outputs. In the present study, outcome indicators are related to innovation performance, which encompasses two dimensions: output impact and disruption (Wei et al., 2023). Previous studies have demonstrated that innovative outputs exhibit varying levels of performance, not only in terms of their impact (Min et al., 2021a, 2021b; Schneider & Costas, 2017; Wang et al., 2012; Winnink & Tijssen, 2015; Winnink et al., 2019), but also in their potential to disrupt or reshape future trajectories (Funk & Owen-Smith, 2017; Wu et al., 2019). Consequently, within this dimension, we integrate these factors into our feature set. Table 1 provides an overview of the features used in our study.

Data splitting

Sample selection. We excluded papers based on the following exclusion criteria to ensure the computability of features: (1) published after 2016 to ensure that each paper has a 5-year citation time window since the PKG dataset only covers the complete citation information up until 2019; (2) have fewer than two references since our combination novelty and homogeneity measures {X4, X6} cannot be computed with fewer than two references; or (3) have fewer than two citations in the first five years after publication since the citation network-related indicators {X52-X59} cannot be constructed. We finally obtained 756 breakthrough papers and 4219 non-breakthrough papers in Dataset 1, and 765 breakthrough papers and 3791 non-breakthrough counterparts in Dataset 2.

Data splitting. In this phase, we performed a stratified split of the dataset into training-validation data (90% of the total data) and test data (10% of the total data). The former was

Table 1 Overview of features used in our study

Category	Sub-category	Feature	Abbreviation	No	Definition	Calculation	Evidence
Input	Amount	Amount	*R_numb*	X1	Amount of prior knowledge	Number of references in a paper	Schoenmakers and Duysters (2010)
	Recency	Recency	*R_mean* *R_cov*	X2 X3	Mean age and the degree of dispersion of ages of prior knowledge	See Online Appendix A	Liang et al. (2020); Mukherjee et al. (2017)
	Combination novelty	Combination novelty	*R_novelty*	X4	Atypicality of the pairwise combination of prior knowledge	See Online Appendix A	Uzzi et al. (2013); Lin et al. (2022)
	Reference impact	Reference impact	*R_impact*	X5	Impact of prior knowledge of a paper	Average number of citations of a paper's references (5-year citation window)	Sheng et al. (2023)
	Homogeneity	Homogeneity	*R_hom*	X6	Similarity with prior knowledge	See Online Appendix A	Sheng et al. (2023)

Table 1 (continued)

Category	Sub-category	Feature	Abbreviation	No	Definition	Calculation	Evidence
Process	Team composition	Gender	*Gender*	X7	Team members' gender composition	A binary variable equal to 1 if the team includes both female and male members, and 0 if all team members are the same gender[1]	Yang et al. (2022)
		Career age	*Career_age*	X8	Average amount of time team members have spent in a given field	Mean time elapsed between each member's first publication year and the observation year	Li et al. (2019b)
		Organization	*Organization*	X9	Prestige and ranking of the institutions with which team members are affiliated	A binary variable equal to 1 if any team member is affiliated with a top 100 university according to the QS ranking, and 0 otherwise[2]	Ao et al. (2023)
	Team structure	Team hierarchy	*Hierarchy*	X10	Degree of dispersion of career ages among team members	See Online Appendix A	Xu et al. (2022a); Xu et al. (2022b)

Table 1 (continued)

Category	Sub-category	Feature	Abbreviation	No	Definition	Calculation	Evidence
	Team collaboration	Author collaboration	*Co_author*	X11	Number of team members	The number of authors listed on the paper	Wu et al. (2019)
		Inter-institutional collaboration	*Co_institution*	X12	Collaboration among team members affiliated with different institutions	A binary variable equal to 1 if team members are affiliated with different institutions, and 0 otherwise	Lyu et al. (2021a)
		International collaboration	*Co_country*	X13	Collaboration among team members from different countries	A binary variable equal to 1 if team members are from different countries, and 0 otherwise	
		Team freshness	*Co_fressness*	X14	Fraction of new members within the team	Fraction of team members with no prior coauthorship relationships with any other team member before they coauthored the focal paper	Zeng et al. (2021)
	Topic select strategy	Topic novelty	*T_novelty*	X15	Atypical combination of topic combinations	See Online Appendix A	Ruan et al. (2023); Chai and Menon (2019)
		Topic popularity	*T_popularity*	X16	Popularity or hotness of the topic		
		Topic diversity	*T_diversity*	X17	Diversity of the topic	Number of major main heading terms (major MeSH terms) of a paper	

Table 1 (continued)

Category	Sub-category	Feature	Abbreviation	No	Definition	Calculation	Evidence
	Team knowledge	Knowledge variety	*K_variety*	X18	Team members' knowledge variety	See Online Appendix A	Ma et al. (2023); Huo et al. (2019)
		Knowledge heterogeneity	*K_heterogeneity*	X19	Team members' knowledge dissimilarity		
	Team experience	Team productivity	FA_{num} *MA_num*	X20-X21	Team members' productivity	Number of papers published by the first author before the year the focal paper was published; Maximum number of papers published by the authors before the year the focal paper was published	Wang et al. (2012)
		Team reputation	*FA_citation* *FA_avgcitation* *FA_H* *MA_citation* *MA_avgcitation* *MA_H*	X22-X27	Team members' academic impact in their field of study	Total number of citations, average number of citations and H index of the first author before the year the focal paper was published; Maximum number of citations, average number of citations and H index of the authors before the year focal the paper was published	Wang et al. (2012); Wang et al. (2019a)

Table 1 (continued)

Category	Sub-category	Feature	Abbreviation	No	Definition	Calculation	Evidence
	Social capital	Social capital	*Co_avgdegree* *Co_avgbetweenness* *Co_avgcloseness* *Co_avgeigenvector*	X28-X31	Set of resources available to a team through social relationships	See Online Appendix A	Wolcott et al. (2016); Wang et al. (2023b)
	Funding	Funding	*A_funding*	X32	Funding and grants received by team members	Number of funding sources the paper received	Lyu et al. (2021a)

Table 1 (continued)

Category	Sub-category	Feature	Abbreviation	No	Definition	Calculation	Evidence
Output	Paper	Discipline	*P_discipline*	X33	Discipline of a paper	Determined based on the journal category (21 categories) defined in the 2021 Journal Citation Report (JCR) of the WoS	
		Characteristics of title, abstract, and keywords	*P_title* *P_abstract* *P_keyword*	X34-X36	Informativity of the title, abstract, and keywords	Number of words in the title and abstract, and number of keywords presented in the paper	
		Length	*P_length*	X37	Length of a paper	Number of pages	Wang et al. (2012)
		Document type	*P_type*	X38	Document type of the document	A binary variable equal to 1 if the document type is a research article, and 0 otherwise	Wu et al. (2019)
		Age	*P_year*	X39	Publication year of the paper	Publication year of the paper	
		Visibility	*P_visibility*	X40	A paper is visible if it is published in an open-access journal, which means it can be accessed without any subscription restrictions	A binary variable equal to 1 if the paper is open access, and 0 otherwise	

Table 1 (continued)

Category	Sub-category	Feature	Abbreviation	No	Definition	Calculation	Evidence
	Journal	Journal impact	*JTC* *JIF* *JIFP*	X41-X43	Journal impact of the journal in which the paper is published	The journal impact is measured by three indicators obtained from the 2021 JCR: the total citations received, the Journal Impact Factor (JIF), and the average JIF Percentile	Wang et al. (2017)
		Language	*J_language*	X44	Language of the journal	Language of the journal	Wang et al. (2012)
		Scope	*J_scope*	X45	Scope of the journal where a paper is published	Number of categories assigned to the journal based on the 254 WoS classification	
Outcome	Impact	Speed	*I_first*	X46	Timeliness of a paper's dissemination within the scientific community	Number of citations in the first cited year	Wang et al. (2012); Min et al. (2021a); Min et al. (2021b); Winnink and Tijssen (2015); Winnink et al. (2019); Schneider and Costas (2017)
			I_ct	X47		Reciprocal of the paper's first-cited age	

Table 1 (continued)

Category	Sub-category	Feature	Abbreviation	No	Definition	Calculation	Evidence
		Scope	*I_total*	X48	Breadth of a paper's dissemination within the scientific community	Number of citations in the first five years after publication	
			I_author	X49		Number of new authors in citing papers who were not authors of the focal paper	
			I_field	X50		Number of distinct disciplines among citing papers	
			I_self	X51		Number of citing papers that share at least one author with the focal paper	
		Strength	*CNedge* *CNdegree* *CNdensity* *CNclusting* *CNbetweenness* *CNcloseness* *CNeigenvector* *CNcomponent*	X52-X59	Impact on the structure changes in the scientific space	See Online Appendix A	
	Disruption	Disruption	*D*	X60	Extent to which a paper consolidates or destabilizes the subsequent use of the components on which it builds	See Online Appendix A	Funk and Owen-Smith (2017); Wu et al. (2019); Wu and Yan (2019)
			Ni	X61			
			Npi	X62			

[1] We identified the gender of each author in the dataset with the help of the Genderize.io tool (e.g., see https://genderize.io/), inferring gender by taking the first name as input. It has been widely used in bibliometrics research

[2] Following Gu and Blackmore (2019), the top 100 universities in the Global Best University QS ranking were defined as the leading universities

utilized for hyperparameter optimization, while the latter was used to evaluate the final performance of the machine learning models. Specifically, within the training-validation set, all parameter combinations were exhaustively explored using a tenfold cross-validation procedure. In each loop, one fold served as the validation set and the other folds were employed as the training set. The optimal hyperparameter set was selected based on the highest mean F1 score across validation folds.

Data pre-processing and data oversampling

The following strategies were executed:

Missing values imputation. We checked for samples containing missing values. We deleted indicators with a substantial number of missing values, either in the positive class or both classes. For continuous features, the remaining indicators were dealt with K-nearest neighbor (KNN) imputation, which has been a widely used algorithm to handle missing values (Jadhav et al., 2019). The core idea of this method is to impute values calculated from the values of the k nearest neighbors. We set the default parameters using the Euclidean distance function and $k = 5$ to select the nearest neighbors and calculated the average for imputation. In addition, for categorical variables, we encoded the categorical data columns into one-hot vectors.

Data normalization. Data normalization is an essential pre-processing step to improve the data quality in machine learning studies (Singh & Singh, 2020). In this study, each continuous feature was normalized using the max–min normalization method, which was transformed into a value within the range of [0, 1].

Data oversampling. Our dataset presented an imbalanced classification problem in that the positive class had a much smaller sample than the negative class. Balance can be achieved by increasing the number of samples in the positive class (over-sampling) (Elreedy & Atiya, 2019). We adopted the synthetic minority over-sampling technique for nominal continuous features (SMOTE-NC) to balance the number of samples in each class by generating synthetic data for the minority class (Chawla et al., 2002). This technique was selected because it is an effective method that enhances performance on various imbalance ratios of data (Doan et al., 2022), and it can handle both numerical and categorical features well in our dataset.

Model selection and evaluation

Extreme gradient boosting – XGBoost

XGBoost is an optimized implementation of the ensemble method gradient boosting decision tree (GBDT), which was designed by Chen and Guestrin (2016). XGBoost is a significant improvement as it uses the second-order Taylor expansion to approximate the loss function, which makes the converge faster. Another improved feature is that it avoids the over-fitting problem by incorporating a regularization term into the objective function. The core of the algorithm is to achieve a prime solution of the objective function, which is expressed in Eq. 1.

$$Objective = \sum_{i} L\left(y_i, \hat{y}_i\right) + \sum_{k} \Omega\left(f_k\right) \tag{1}$$

The first term is the loss function that denotes the difference between the predicted and actual values, while the second term serves as the regularization term that represents the complexity of the model. XGBoost has been utilized extensively in various fields to address challenging tasks due to its high accuracy and fast processing time (Ekanayake et al., 2022; Wang et al., 2022). Consequently, we employed the XGBoost model in this study to develop a prediction model for scientific breakthroughs by leveraging its high predictive performance in classification tasks (Joung & Kim., 2023; Ma et al., 2022; Parsa et al., 2020).

Baseline models

To demonstrate the superiority of the XGBoost model (XGB), we compared it with other typical machine learning models, including logistic regression (LR), random forest (RF), support vector machine (SVM), multilayer perceptron (MLP), decision tree (DT), and AdaBoost (ADB). The first four algorithms have been employed in previous studies to identify scientific breakthroughs (Li et al., 2022; Min et al., 2021a; Wolcott et al., 2016), and the last two tree-based models have been commonly used to build classification models (Ma et al., 2022).

We further compared the results with Min et al.'s (2021a) study that used a machine learning model to predict Nobel prize-winning papers. In their study, 116 Nobel prize-winning papers and their counterpart papers were used in the logistic model. The model included nine features that quantify the structure of citation networks: X52-X59.

Large language models (LLMs)

We also conducted additional experiments using large language models (LLMs). We employed frozen local LLMs, including Llama-3.2-1B/3B and Qwen3-1.7B/4B, and trained them with a lightweight fusion head for the downstream prediction task. Regarding the data preprocessing, we took the same setting as previous machine learning models. Then, all features were projected through a multilayer perceptron to align with the hidden dimensionality of the LLM. The projected representation was then fused with the LLM's final-layer output via cross-attention, and the fused embedding was fed into a lightweight MLP classifier. During training, we optimized only the projection, fusion, and classifier parameters using AdamW, while keeping the LLM backbone frozen. This design aimed to leverage the semantic richness of the LLM embeddings without incurring the cost challenges of fine-tuning the entire model.

Model evaluation

Predicting scientific breakthroughs is a binary classification task. We adopted the F1 score and the area under the receiver operating characteristic (ROC) as the main metrics to evaluate the performance of a classification task (Ma et al., 2022; Ragini et al., 2018).

The **F1 score** is the harmonic average of precision and recall using the following formula:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2}$$

where precision represents the ratio of the number of samples in the positive class to the number of samples that are predicted as the positive class. Recall denotes the ratio of the number of samples that are correctly predicted as the positive class to the total number of samples in that category, calculated as:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

where TP indicates the number of samples with the correct classification to the positive class. FP denotes the incorrect classification to the positive class, and FN represents the incorrect classification to the negative class.

The **ROC** was created on a space where the false positive rate (FPR) is on the X coordinate, and the true positive rate (TPR) is on the Y coordinate at various threshold settings. The area under the ROC curve (AUC) was used to evaluate the performance of the model, ranging from [0, 1]. The closer the value is to 1, the better the prediction of the model. AUC = 0.5 represents the result of random guessing.

Model interpretation: SHapley Additive exPlanations (SHAP)

SHAP is an effective way to explain the output of machine learning models based on the game theoretic approach proposed by Shapley and Shubik (1954). It was first proposed by Lundberg and Lee (2017). For each sample, it assigns a SHAP value to each input variable (feature) to represent its contribution to model prediction. For an input dataset of size $N \times M$ (N denotes the number of samples, and M represents the number of features), the weighted sum of the marginal contribution is calculated when the feature is added, producing an $N \times M$ matrix with the SHAP values. It is expressed as:

$$\emptyset = \sum_{\mathrm{S} \subseteq F \backslash \mathrm{X_i}} \frac{|\mathrm{S}|!(|\mathrm{M}| - |\mathrm{S}| - 1)!}{|\mathrm{M}|!} \left(f\left(S \cup \mathrm{X_i}\right) - f(S)\right) \tag{5}$$

where F is the set of all features $\{X_1, X_2, \ldots X_m\}$. $\mathrm{F} \backslash \{\mathrm{X_i}\}$ represents the set of features of F without the feature $\{X_i\}$. S are all feature subsets of $\mathrm{F} \backslash \{\mathrm{X_i}\}$. $|M|$ is the total number of features in M while $|S|$ is the total number of features in S. The model $f\left(S \cup \{\mathrm{X_i}\}\right)$ is trained with the set of features S and $\{X_i\}$, and the model $f(S)$ is trained with the feature set S. We explore the possible relationship between the feature value and the impact on the model prediction with SHAP.

Results

Prediction results

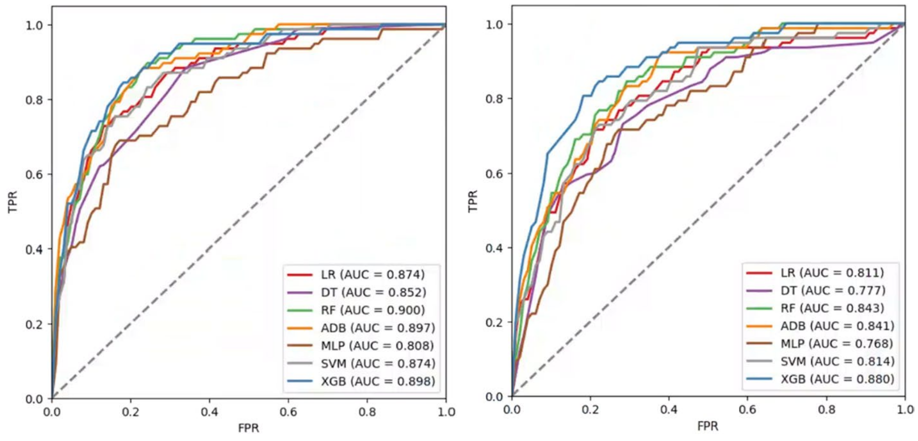
In this section, the performance of seven machine learning models is evaluated on the test set to investigate whether XGB outperforms other widely used models. The results in Table 2 show the performance measures: F1 score, precision, and recall for all models. The recall scores of XGB are 0.636 and 0.623 for Dataset 1 and Dataset 2, respectively, ranking

Table 2 The performance of machine learning models

	Dataset 1				Dataset 2			
Models	F1 score	Precision	Recall	AUC	F1 score	Precision	Recall	AUC
LR	0.584	0.667	0.519	0.874	0.497	0.433	0.584	0.811
DT	0.531	0.576	0.494	0.852	0.467	0.383	0.597	0.777
RF	0.559	0.606	0.519	0.900	0.531	0.470	0.610	0.843
ADB	0.591	0.709	0.506	0.897	0.511	0.455	0.584	0.841
MLP	0.472	0.630	0.377	0.808	0.452	0.385	0.545	0.768
SVM	0.577	0.597	0.558	0.874	0.520	0.460	0.597	0.814
XGB	0.613	0.590	0.636	0.898	0.611	0.600	0.623	0.880

1st. In terms of F1 scores, XGB achieves the best performance with scores of 0.613 and 0.611, respectively. Since the DT is a weak learner, it yields relatively weaker predictions than the tree-based algorithms, namely RF, ADB, and XGB. The ensemble learning methods, including boosting (ADB, XGB) and bagging (RF) algorithms, provide an improved version of the classic DT and are more accurate and robust than individual learning methods. In addition, as displayed in Fig. 3, XGB performs quite well in terms of the AUC score, ranking 2nd in Dataset 1 and 1st in Dataset 2 (0.898 and 0.880, respectively).

Overall, based on the evaluation scores, our results imply the potential of XGB in the breakthrough identification task across both datasets. Additionally, we found differences in the evaluation scores between the two datasets, which suggests that the construction of the non-breakthrough class affects prediction accuracy, with Dataset 1 demonstrating superior performance. Our study indicates that the model's performance depends not only on the selection of algorithm but also on the construction method of negative samples. Compared with Dataset 1, Dataset 2 was designed to include non-breakthrough papers that are more comparable to breakthroughs in terms of journal-related features (e.g., journal impact),



(a) (b)

Fig. 3 ROC curves of the seven algorithms (Dataset 1 and Dataset 2)

which may reduce the distinctions between the breakthrough and non-breakthrough groups. The relatively smaller differences between the two classes in these features may have led to relatively lower predictive accuracy for models based on Dataset 2. Future studies could further examine the robustness of our findings under different sampling strategy.

Comparison with related work

The results are compared with the baseline model (with nine features) constructed by Min et al. (2021a). Figure 4 presents the changes in performance measures of our model relative to the baseline model in the two datasets. In Dataset 1 (orange bars), while MLP yields a lower F1 score compared to the baseline, all other algorithms demonstrate superior performance. Notably, XGB achieves remarkable improvements with F1 score increases of 12.5%, and recall score increases of 13%. In Dataset 2 (green bars), the F1 and AUC scores of all models demonstrate significant improvement over the baseline, with the exception of MLP. Notably, XGB achieves the greatest enhancement with an F1 score increase of 14.8% and an AUC score increase of 9.6%. In summary, our study achieves predictions

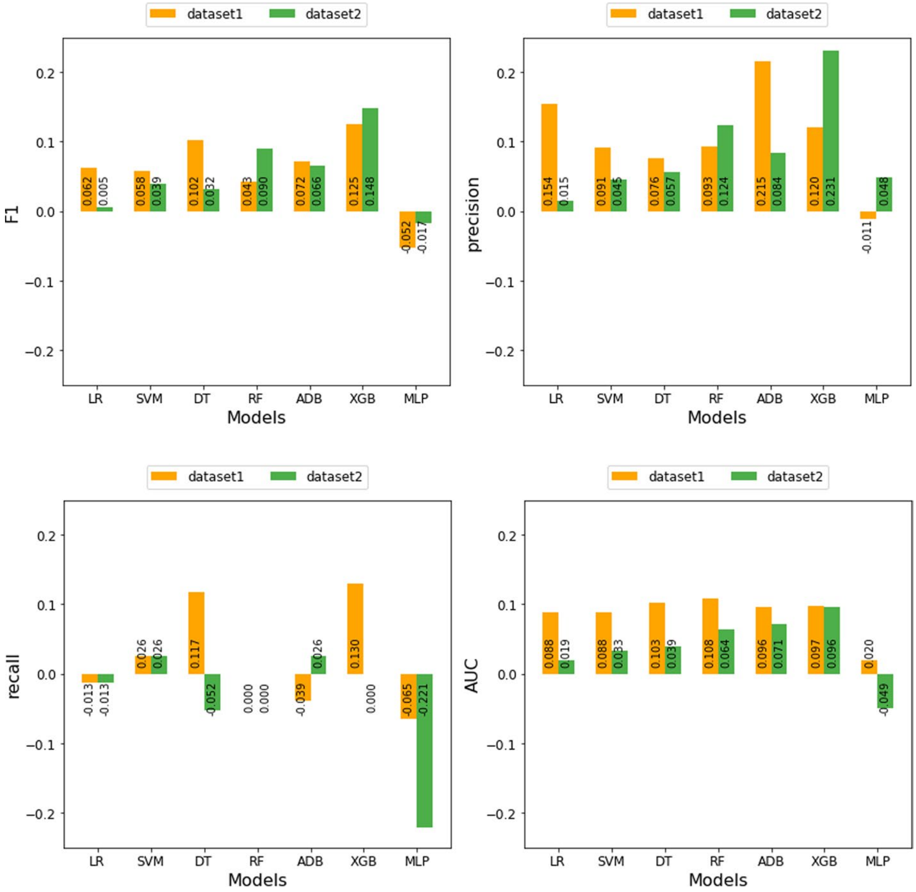


Fig. 4 Comparison of F1, precision, recall, and AUC of predictions with related work (Dataset 1 and Dataset 2)

with acceptable accuracy compared to Min et al. (2021a), with XGB achieving superior improvements across evaluation metrics in both datasets.

It should be noted that when using the same baseline model, our results are higher than those of Min et al. (2021a) with an AUC of 0.619. The differences are likely attributed to the following reason. The dataset of non-breakthroughs in Min et al.'s (2021a) study was constructed from papers that received approximately equivalent citation counts as breakthroughs, which means that one of the features, citation counts, showed limited predictive power in their model.

Comparison with LLMs

Table 3 reports the results of four LLMs, including Llama-1B/3B and Qwen3-1.7B/4B. As shown in Table 3, QWen3-1.7B achieves the best recall and F1 values in Dataset 1, with a recall score of 0.61 and an F1 score of 0.584, respectively. The AUC results further confirm that QWen3-1.7B performs well, ranking first among all models. In Dataset 2, Llama-3B demonstrates the best performance in terms of F1 and AUC values, while Llama-1B achieves the highest recall (0.805). Comparing the classification results in Tables 2 and 3, we find that in Dataset 1, the LLMs generally achieve lower evaluation scores than traditional machine learning models in terms of F1 and AUC scores, except for DT and MLP. Notably, the best-performing XGB surpasses QWen3-1.7B. However, both Llama-3B and QWen3-1.7B achieve higher recall scores than all other traditional machine learning models, except XGB. In Dataset 2, we find that SVM, RF, ADB and XGB perform better than LLMs in terms of F1 and AUC scores. Both Llama-1B/3B achieve higher recall scores than all seven traditional machine learning models.

In summary, there are several main findings: (1) In both datasets, XGB performs best among traditional machine learning models, and also surpasses LLMs; (2) LLMs (Qwen3-1.7B in Dataset 1 and Llama-1B in Dataset 2) have higher recall scores than the majority of traditional machine learning methods; and (3) Similar to the results of traditional machine learning models, LLMs demonstrate superior performance in Dataset 1 compared to Dataset 2, depending on the construction of the non-breakthrough class.

Model interpretability

For a global explanation, we employed the XGBoost model's built-in feature importance analysis to obtain the feature importance ranking. The results revealed the extent to which each feature impacts the model prediction. Table 4 lists the top ten features, with *Ni* identified as the most important feature. These top features fall into four categories: paper impact

Table 3 The performance of LLMs

	Dataset 1				Dataset 2			
Models	F1 score	Precision	Recall	AUC	F1 score	Precision	Recall	AUC
Llama-1B	0.544	0.571	0.520	0.865	0.5	0.363	0.805	0.820
Llama-3B	0.548	0.506	0.597	0.867	0.505	0.413	0.649	0.821
QWen3-1.7B	0.584	0.560	0.610	0.879	0.455	0.404	0.520	0.810
QWen3-4B	0.557	0.619	0.507	0.869	0.503	0.457	0.558	0.817

Table 4 The top ten important features

Rank	Feature	Importance
1	*Ni*	0.159
2	*MA_H*	0.039
3	*CNedge*	0.036
4	*MA_citation*	0.036
5	*MA_num*	0.031
6	*I_total*	0.031
7	*JIF*	0.021
8	*CNdegree*	0.020
9	*I_self*	0.019
10	*FA_H*	0.018

(*CNedge*, *I_total*, *CNdegree*, and *I_self*), paper disruption (*Ni*), team experience (*MA_H*, *MA_citation*, *MA_num* and *FA_H*), and journal impact (*JIF*). However, XGBoost model's built-in feature importance analysis cannot interpret the relationship between the features and model prediction. Therefore, we further used the SHAP approach to interpret the model. We conducted both SHAP importance analysis and SHAP dependency analysis to explore the possible relationship between the feature value and the impact on the model prediction. The corresponding results are provided in Online Appendix B, which supplements the XGBoost built-in feature importance analysis.

In addition to the global explanations mentioned above, we adopted SHAP methods to provide local explanations for each individual sample. Figure 5 illustrates the explanations for the instance obtained from the SHAP waterfall plot. SHAP values decompose the prediction of the model into the sum of the contributions of each input variable. All variables (features) collectively contribute to the deviation of prediction from the base value, ultimately determining whether the output is breakthrough or non-breakthrough. Red ones denote variables that push the prediction toward breakthrough, while blue ones represent

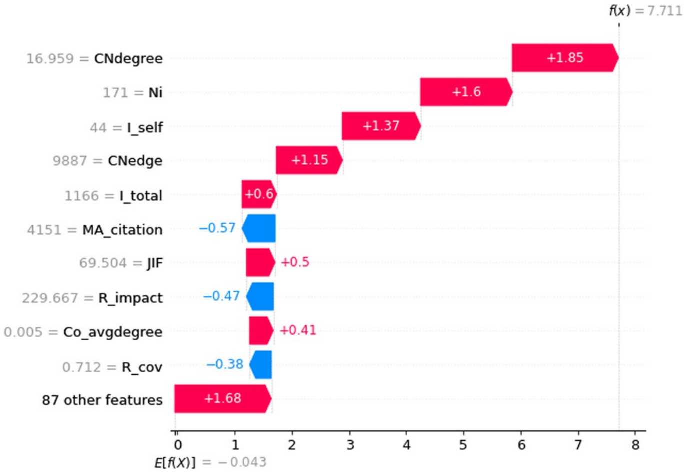


Fig. 5 The most important SHAP local explanation

variables that influence the prediction toward non-breakthrough. The length of the bar reflects the magnitude of the contribution to the prediction.

Following Saarela and Kaerkkaeinen (2020), we show the most important local explanation for the sample. This explanation has the highest predicted probability to be a breakthrough and was actually a breakthrough (true positive). The sample is from Okita et al. (2007), in which Shinya Yamanaka was honored with the 2009 Lasker Basic Medical Research Award for nuclear reprogramming discoveries. First, our analysis shows that the relatively high *CNdegree* plays the most significant predictive role in this case, exhibiting positive effects on the prediction of a breakthrough (SHAP = 1.85). In addition, the relatively high values of *Ni* (171), *I_self* (44), *CNedge* (9,887), and *I_total* (1,166) collectively lead to further divergence from the base value, as these factors exhibit positive impacts on the model. These top-ranked features indicate that outcome-related features dominate the prediction in this individual case. In addition, a publication in *Nature* emerged as another significant feature with a SHAP value of 0.50. For other features, *MA_citation*, *R_impact* and *R_cov* have opposite effects (SHAP = −0.57, −0.47 and −0.38, respectively), which pushes the model away from the positive class. In summary, when taking all feature contributions into consideration during the prediction process, the model accurately predicts the breakthrough classification.

Discussion and conclusion

Identifying breakthrough research is a significant and challenging issue not only for scientists in the scientific community, but also for R&D management and policymakers. This paper presents an interpretable machine learning model to predict scientific breakthroughs utilizing a new dataset of Nobel and Lasker prize-winning publications. Specifically, we designed an upgraded framework that integrates possible factors that are associated with breakthroughs with the IPOO perspective. Traditional machine learning models and large language models are adopted to evaluate prediction performance. We also applied the XGBoost model's built-in importance method and SHAP to identify critical factors and quantify their influence on the model. This approach improves the transparency and interpretability of the prediction and provides new insights.

Main conclusion

Research has begun to highlight the importance of identifying scientific breakthroughs, regarded as major innovations in the advancement of science. The key findings of this study are as follows:

(1) XGBoost exhibits the best predictive performance among traditional machine learning methods in two datasets.
(2) Compared with the results of the baseline model constructed by Min et al. (2021a), our study achieved better accuracy in identifying scientific breakthroughs. We found that XGBoost demonstrates remarkable improvements of 12.5% and 14.8% in the F1 score in the two datasets, respectively.
(3) Qwen3-1.7B and Llama-3B are the best models among the LLMs in Dataset 1 and Dataset 2, respectively. In terms of overall performance measured by the F1 score, traditional machine learning models perform better than LLMs, except for DT and MLP.

LLMs (Qwen3-1.7B in Dataset1 and Llama-1B in Dataset 2) have higher recall scores than most traditional machine learning methods.

(4) The XGBoost model's built-in feature importance analysis suggests that *Ni*, *MA_H*, *CNedge*, *MA_citation*, *MA_num*, *I_total*, *JIF*, *CNdegree*, *I_self* and *FA_H* are the most influential features.

(5) We compared the consistency of the top contributing features using SHAP analysis, logistic regression analysis, and the model's built-in feature importance analysis. First, the results show that *Ni* plays the most important role and exhibits a positive correlation with the prediction of breakthroughs. This finding serves as the foundation for exploring the capabilities of the *Ni* indicator in subsequent studies. Second, *Ni*, *I_self*, *CNedge*, *MA_H*, and *I_total* consistently appear and are ranked among the top features across the three methods. In addition, the logistic regression analysis shows that approximately 85% of the features that are statistically significant exhibit directional consistency with SHAP values (for details, please refer to Online Appendix B-D).

Our study built upon the work of Min et al. (2021a) by extending the set of observable features and offering a more in-depth interpretation of their interrelationships with scientific breakthroughs. The findings demonstrate that XGBoost achieved the best performance, with F1 scores of 0.613 and 0.611 for Dataset 1 and Dataset 2, respectively. Although the prediction results are better than random guessing and that of Min et al. (2021a), there is still room for improvement in the prediction of scientific breakthroughs. Although opportunities remain for further improvement, our results demonstrate that the features we constructed are capable of capturing important aspects of the underlying patterns associated with scientific breakthroughs. Importantly, the features we employed align with those that have been widely adopted in citation prediction studies, including predicting citation counts (Bai et al., 2019; Ruan et al., 2020; Zhang et al., 2021), technological impact (Gao et al., 2024), clinical citations (Liu et al., 2024), and the identification of highly cited papers (Hu et al., 2023; Wang et al., 2019a, 2019b). While citation prediction tasks often achieve AUC scores exceeding 0.86 or F1 scores above 0.69 (Akella et al., 2021; Fu & Aliferis, 2010; Hu et al., 2023), breakthrough prediction presents a fundamentally different and more challenging problem. Citation outcomes are heavily influenced by extrinsic factors, such as author reputation, journal prestige, or network visibility (Fu & Aliferis, 2010), which are relatively easier to quantify. In contrast, breakthroughs are likely driven more by intrinsic characteristics, including the originality, quality, and transformative potential of the research itself, which are more difficult to observe and operationalize. Accordingly, although our study focuses on observable, bibliometric-based features, we recognize that enhancing the prediction of breakthroughs may require incorporating richer representations of research content. Future work could integrate semantically interpretable patterns derived from the full texts of articles, going beyond traditional metadata approaches (Beranová et al., 2022), thereby further advancing predictive performance.

Theoretical implications

This study provides a novel lens for predicting scientific innovation grounded in the theoretical foundations of the IPOO model. While previous studies have demonstrated the effectiveness of the IPOO model in developing performance evaluation frameworks, its potential for scientific breakthrough prediction remains underexplored. Our study makes a primary theoretical contribution by establishing a pathway that demonstrates how the

IPOO model can be operationalized for breakthrough prediction. Through this lens, we decompose the relevant factors into knowledge input, team process, innovative output, and innovative outcome dimensions, and further identify potential factors within each dimension. By empirically validating our framework on two datasets, we advance our understanding of the mechanisms underlying scientific breakthroughs and reveal the most significant predictors influencing breakthroughs.

Another implication is that, unlike existing studies, we employ the XGBoost-SHAP machine learning approach to identify scientific breakthroughs. First, the findings demonstrate the effectiveness of the XGBoost model and show that it is capable of higher predictive accuracy. In addition, the findings related to feature contributions indicate that outcome-related factors (*Ni*, *CNedge*, *I_total*, *CNdegree*, and *I_self*), team experience (*MA_H*, *MA_citation*, *MA_num* and *FA_H*), and journal impact (*JIF*) play crucial roles in identifying breakthroughs. In addition, the findings reveal several new relationships. For example, *Ni* is positively correlated with the prediction of breakthroughs.

Practical implications

This study provides several practical implications for science policymakers and academic evaluators. First, our results show that machine learning methods are more effective in detecting scientific breakthroughs. Given the exponential growth of scholarly publications each year, an automated identification method may be a preferable way for databases to identify potential excellent papers. For example, the "Clarivate Citation Laureates" are selected based on the paper and citations in the Web of Science platform. They could be considered candidates to win a Nobel prize. Similarly, researchers can leverage well-established research infrastructures with open datasets containing available bibliographic data of publications, such as the PKG dataset (Xu et al., 2020) and the SciSciNet (Lin et al., 2023), which initially offers commonly used metrics. The method proposed in our study may help researchers identify potential outstanding papers in the database and effectively assist in the evaluation of research papers.

Our findings also provide potential quantitative guidance for identifying ground-breaking discoveries using an alternative perspective. A wealth of research has been dedicated to measuring significant innovations in science using *ex-post* measures, including citation count-based and citation network-based approaches. Our study suggests that *Ni* is the strongest predictor in *ex-post* measures. This finding suggests the potential utility of the measurs in future applications. However, future work should verify our results.

Limitations and future directions

This study is not without limitations. First, the calculation of features is dependent on the bibliographic information in the PKG database. Consequently, the results are influenced by the quality of the database. For example, citation-based indicators are intricately related to the coverage and accuracy of the citation data in the database. Second, we acknowledge that many unobservable factors, such as the quality or originality of the discovery, influence the prediction of breakthroughs. However, they were not incorporated into our study. Examining these factors is challenging, as it requires an in-depth analysis of the content of articles. Consequently, our study is limited to observable factors that have been quantified and examined in previous studies. Therefore, the scope of predictive features considered in our model is constrained. Future endeavors should combine both literature-related features and content-related features to enhance the detection of breakthrough discoveries. In addition, for feasibility, we adopted a 1:5 ratio of scientific breakthroughs to non-breakthroughs, which is far less imbalanced than the real-world distribution. This difference may affect the generalizability of the results under more realistic distributions. Future work could evaluate the method under more realistic ratios on larger datasets.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11192-025-05497-7.

Acknowledgements This research was supported by the National Natural Science Foundation of China (NSFC No. 72304133). We sincerely thank Professor Jiang Li at Nanjing University for his constructive suggestions and invaluable guidance in revising the manuscript. We are also grateful to Dr. Yuanyang Zhu at Nanjing University for his assistance in performing LLMs, as well as to Tianxing Pan at Nanjing University for his assistance.

Funding Funding was provided by National Natural Science Foundation of China,72304133,Dongqing Lyu

Data availability The data and code used in this study are available at GitHub: https://github.com/Libo-S/XGBoost_SHAP.

References

Akella, A. P., Alhoori, H., Kondamudi, P. R., Freeman, C., & Zhou, H. (2021). Early indicators of scientific impact: Predicting citations with altmetrics. *Journal of Informetrics, 15*(2), 101128.

Ao, W., Lyu, D., Ruan, X., Li, J., & Cheng, Y. (2023). Scientific creativity patterns in scholars' academic careers: Evidence from PubMed. *Journal of Informetrics, 17*(4), 101463.

Bai, X., Zhang, F., & Lee, I. (2019). Predicting the citations of scholarly paper. *Journal of Informetrics, 13*(1), 407–418.

Beranová, L., Joachimiak, M. P., Kliegr, T., Rabby, G., & Sklenák, V. (2022). Why was this cited? Explainable machine learning applied to COVID-19 research literature. *Scientometrics, 127*(5), 2313–2349.

Cammarano, A., Michelino, F., & Caputo, M. (2022). Extracting firms' R&d processes from patent data to study inbound and coupled open innovation. *Creativity and Innovation Management, 31*(2), 322–339.

Capponi, G., Martinelli, A., & Nuvolari, A. (2022). Breakthrough innovations and where to find them. *Research Policy, 51*(1), 104376.

Chai, S., & Menon, A. (2019). Breakthrough recognition: Bias against novelty and competition for attention. *Research Policy, 48*(3), 733–747.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Chen, C. (2012). Predictive effects of structural variation on citation counts. *Journal of the American Society for Information Science and Technology, 63*(3), 431–449.

Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., & Pellegrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics, 3*(3), 191–209.

Chen, K. Y., Altinay, L., Chen, P. Y., & Dai, Y. D. (2021). Market knowledge impacts on product and process innovation: Evidence from travel agencies. *Tourism Review, 77*(1), 271–286.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785–794).

Choi, S., & Choi, J. S. (2014). Dynamics of innovation in nonprofit organizations: The pathways from innovativeness to innovation outcome. *Human Service Organizations: Management, Leadership & Governance, 38*(4), 360–373.

Dahlin, K. B., Weingart, L. R., & Hinds, P. J. (2005). Team diversity and information use. *Academy of Management Journal, 48*(6), 1107–1123.

Daud, A., Ahmad, M., Malik, M. S. I., & Che, D. (2015). Using machine learning techniques for rising star prediction in co-author network. *Scientometrics, 102*(2), 1687–1711.

de Carvalho, G. D. G., Cruz, J. A. W., de Carvalho, H. G., Duclós, L. C., & de Fátima Stankowitz, R. (2017). Innovativeness measures: A bibliometric review and a classification proposal. *International Journal of Innovation Science, 9*(1), 81–101.

Doan, Q. H., Mai, S. H., Do, Q. T., & Thai, D. K. (2022). A cluster-based data splitting method for small sample and class imbalance problems in impact damage classification. *Applied Soft Computing, 120*, 108628.

Dong, X., Lin, K., Gao, Y., & Hu, B. (2023). Nobel citation effects on scientific publications: A case study in physics. *Information Processing & Management, 60*(4), 103410.

Ekanayake, I. U., Meddage, D. P. P., & Rathnayake, U. (2022). A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). *Case Studies in Construction Materials, 16*, e01059.

Elreedy, D., & Atiya, A. F. (2019). A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Information Sciences, 505*, 32–64.

Ferreira, M. A., Johnson, D., da Silva, C. P., & Ramos, T. B. (2018). Developing a performance evaluation mechanism for Portuguese marine spatial planning using a participatory approach. *Journal of Cleaner Production, 180*, 913–923.

Frandsen, T. F., & Nicolaisen, J. (2013). The ripple effect: Citation chain reactions of a nobel prize. *Journal of the American Society for Information Science and Technology, 64*(3), 437–447.

Fu, L., & Aliferis, C. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics, 85*(1), 257–270.

Funk, R. J., & Owen-Smith, J. (2017). A dynamic network measure of technological change. *Management Science, 63*(3), 791–817.

Gao, X., Wu, Q., Liu, Y., & Wang, Y. (2024). Predicting the technological impact of papers: Exploring optimal models and most important features. *Journal of Information Science*. https://doi.org/10.1177/01655515241261056

Gu, X., & Blackmore, K. L. (2019). Developing a scholar classification scheme from publication patterns in academic science: A cluster analysis approach. *Journal of the Association for Information Science and Technology, 70*(11), 1262–1276.

Häyrynen, M. (2007). *Breakthrough research: Funding for high-risk research at the Academy of Finland*. Academy of Finland.

Heeley, M. B., & Jacobson, R. (2008). The recency of technological inputs and financial performance. *Strategic Management Journal, 29*(7), 723–744.

Hollingsworth, J. R. (2008). Scientific discoveries: An institutionalist and path-dependent perspective. Biomedical and Health ResearchIn C. Hannaway (Ed.), *Biomedicine in the Twentieth Century: Practices, Policies, and Politics* (Vol. 72, pp. 317–353). National Institutes of Health.

Hsu, Y. C., Wang, C. W., & Lan, J. B. (2020). Evaluating the performance of employee assistance programs (EAP): A checklist developed from a large sample of public agencies. *Asia Pacific Journal of Management, 37*, 935–955.

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

Hückstädt, M. (2023). Ten reasons why research collaborations succeed—a random forest approach. *Scientometrics, 128*(3), 1923–1950.

Huo, D., Motohashi, K., & Gong, H. (2019). Team diversity as dissimilarity and variety in organizational innovation. *Research Policy, 48*(6), 1564–1572.

Hur, W., & Oh, J. (2021). A man is known by the company he keeps?: A structural relationship between backward citation and forward citation of patents. *Research Policy, 50*(1), 104117.

Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence, 33*(10), 913–933.

Jones, B. F. (2009). The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *The Review of Economic Studies, 76*(1), 283–317.

Joung, J., & Kim, H. (2023). Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews. *International Journal of Information Management, 70*, 102641.

Katila, R. (2002). New product search over time: Past ideas in their prime? *Academy of Management Journal, 45*(5), 995–1010.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (Vol. 111). University of Chicago Press.

Kumar, D., Bhowmick, P. K., & Paik, J. H. (2023). Researcher influence prediction (ResIP) using academic genealogy network. *Journal of Informetrics, 17*(2), 101392.

Kwon, U., & Geum, Y. (2020). Identification of promising inventions considering the quality of knowledge accumulation: A machine learning approach. *Scientometrics, 125*(3), 1877–1897.

Lee, Y. N., Walsh, J. P., & Wang, J. (2015). Creativity in scientific teams: Unpacking novelty and impact. *Research Policy, 44*(3), 684–697.

Li, X., Ma, X., & Feng, Y. (2024). Early identification of breakthrough research from sleeping beauties using machine learning. *Journal of Informetrics, 18*(2), 101517.

Li, J., Yin, Y., Fortunato, S., & Wang, D. (2019a). A dataset of publication records for Nobel laureates. *Scientific Data, 6*(1), 33.

Li, J., Yin, Y., Fortunato, S., & Wang, D. (2019b). Nobel laureates are almost the same as us. *Nature Reviews Physics, 1*(5), 301–303.

Li, X., Wen, Y., Jiang, J., Daim, T., & Huang, L. (2022). Identifying potential breakthrough research: A machine learning method using scientific papers and Twitter data. *Technological Forecasting and Social Change, 184*, 122042.

Liang, G., Hou, H., Ding, Y., & Hu, Z. (2020). Knowledge recency to the birth of Nobel Prize-winning articles: Gender, career stage, and country. *Journal of Informetrics, 14*(3), 101053.

Liao, C. H. (2021). The Matthew effect and the halo effect in research funding. *Journal of Informetrics, 15*(1), 101108.

Lin, R., Li, Y., Ji, Z., Xie, Q., & Chen, X. (2025). Quantifying the degree of scientific innovation breakthrough: Considering knowledge trajectory change and impact. *Information Processing & Management, 62*(1), 103933.

Lin, Y., Evans, J. A., & Wu, L. (2022). New directions in science emerge from disconnection and discord. *Journal of Informetrics, 16*(1), 101234.

Lin, Z., Yin, Y., Liu, L., & Wang, D. (2023). SciSciNet: A large-scale open data lake for the science of science research. *Scientific Data, 10*(1), 315.

Liu, Z., Wang, C., & Wang, R. (2024). From bench to bedside: Determining what drives academic citations in clinical trials. *Scientometrics, 129*(11), 6813–6837.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Lyu, D., Gong, K., Ruan, X., Cheng, Y., & Li, J. (2021a). Does research collaboration influence the "disruption" of articles? Evidence from neurosciences. *Scientometrics, 126*, 287–303.

Lyu, D., Ruan, X., Xie, J., & Cheng, Y. (2021b). The classification of citing motivations: A meta-synthesis. *Scientometrics, 126*, 3243–3264.

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.

Ma, Y., Li, T., Mao, J., Ba, Z., & Li, G. (2022). Identifying widely disseminated scientific papers on social media. *Information Processing & Management, 59*(3), 102945.

MacCuspie, R. I., Hyman, H., Yakymyshyn, C., Srinivasan, S. S., Dhau, J., & Drake, C. (2014). A framework for identifying performance targets for sustainable nanomaterials. *Sustainable Materials and Technologies, 1*, 17–25.

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review, 26*(3), 356–376.

Min, C., Bu, Y., & Sun, J. (2021a). Predicting scientific breakthroughs based on knowledge structure variations. *Technological Forecasting and Social Change, 164*, 120502.

Min, C., Bu, Y., Wu, D., Ding, Y., & Zhang, Y. (2021b). Identifying citation patterns of scientific breakthroughs: A perspective of dynamic citation process. *Information Processing & Management, 58*(1), 102428.

Mugabushaka, A. M., Sadat, J., & Faria, J. C. D. (2020). In Search of Outstanding Research Advances: Prototyping the creation of an open dataset of "editorial highlights". arXiv preprint arXiv:2011.07910.

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), e1601315.

Nerkar, A. (2003). Old is gold? The value of temporal exploration in the creation of new knowledge. *Management Science, 49*(2), 211–229.

Okita, K., Ichisaka, T., & Yamanaka, S. (2007). Generation of germline-competent induced pluripotent stem cells. *Nature, 448*(7151), 313–317.

Papazoglou, M. E., & Nelles, J. (2023). Keeping pace with technological change: Insights into the recency of internal knowledge inputs. *Journal of the Knowledge Economy, 14*(4), 3724–3740.

Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S., & Mohammadian, A. K. (2020). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accident Analysis & Prevention, 136*, 105405.

Petruzzelli, A. M., Ardito, L., & Savino, T. (2018). Maturity of knowledge inputs and innovation value: The moderating effect of firm age and size. *Journal of Business Research, 86*, 190–201.

Ponomarev, I. V., Lawton, B. K., Williams, D. E., & Schnell, J. D. (2014a). Breakthrough paper indicator 2.0: Can geographical diversity and interdisciplinarity improve the accuracy of outstanding papers prediction? *Scientometrics, 100*, 755–765.

Ponomarev, I. V., Williams, D. E., Hackett, C. J., Schnell, J. D., & Haak, L. L. (2014b). Predicting highly cited papers: A method for early detection of candidate breakthroughs. *Technological Forecasting and Social Change, 81*, 49–55.

Porter, A., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics, 81*(3), 719–745.

Ragini, J. R., Anand, P. R., & Bhaskar, V. (2018). Big data analytics for disaster response and recovery through sentiment analysis. *International Journal of Information Management, 42*, 13–24.

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., 51*(5), 1033–1046.

Ruan, X., Zhu, Y., Li, J., & Cheng, Y. (2020). Predicting the citation counts of individual papers via a BP neural network. *Journal of Informetrics, 14*(3), 101039.

Saarela, M., & Kaerkkaeinen, T. (2020). Can we automate expert-based journal rankings? Analysis of the Finnish publication indicator. *Journal of Informetrics, 14*(2), 101008.

Savov, P., Jatowt, A., & Nielek, R. (2020). Identifying breakthrough scientific papers. *Information Processing & Management, 57*(2), 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.

Schneider, J. W., & Costas, R. (2017). Identifying potential "breakthrough" publications using refined citation analyses: Three related explorative approaches. *Journal of the Association for Information Science and Technology, 68*(3), 709–723.

Schoenmakers, W., & Duysters, G. (2010). The technological origins of radical inventions. *Research Policy, 39*(8), 1051–1059.

Schumpeter, J. A. (1939). *Business Cycles; A Theoretical, Historical, and Statistical Analysis of the Capitalist Process* (1st ed). McGraw-Hill Book Company, inc.

Shapley, L. S., & Shubik, M. (1954). A method for evaluating the distribution of power in a committee system. *American Political Science Review, 48*(3), 787–792.

Sheng, L., Lyu, D., Ruan, X., Shen, H., & Cheng, Y. (2023). The association between prior knowledge and the disruption of an article. *Scientometrics, 128*(8), 4731–4751.

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing, 97*, 105524.

Tahamtan, I., Safipour Afshar, A., & Ahamdzadeh, K. (2016). Factors affecting number of citations: A comprehensive review of the literature. *Scientometrics, 107*, 1195–1225.

Tohalino, J. A., & Amancio, D. R. (2022). On predicting research grants productivity via machine learning. *Journal of Informetrics, 16*(2), 101260.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science, 342*(6157), 468–472.

Veugelers, R., & Wang, J. (2019). Scientific novelty and technological impact. *Research Policy, 48*(6), 1362–1372.

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., Rafols, I., & Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics, 5*(1), 14–26.

Wang, D., Thunéll, S., Lindberg, U., Jiang, L., Trygg, J., & Tysklind, M. (2022). Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based machine learning methods. *Journal of Environmental Management, 301*, 113941.

Wang, F., Fan, Y., Zeng, A., & Di, Z. (2019a). Can we predict ESI highly cited publications? *Scientometrics, 118*, 109–125.

Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy, 46*(8), 1416–1436.

Wang, M., Wang, Z., & Chen, G. (2019b). Which can better predict the future success of articles? Bibliometric indices or alternative metrics. *Scientometrics, 119*, 1575–1595.

Wang, M., Yu, G., Xu, J., He, H., Yu, D., & An, S. (2012). Development a case-based classifier for predicting highly cited papers. *Journal of Informetrics, 6*(4), 586–599.

Wang, S., Ma, Y., Mao, J., Bai, Y., Liang, Z., & Li, G. (2023a). 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.

Wang, X., Yang, X., Du, J., Wang, X., Li, J., & Tang, X. (2021). A deep learning approach for identifying biomedical breakthrough discoveries using context analysis. *Scientometrics, 126*, 5531–5549.

Wang, Y., Li, N., Zhang, B., Huang, Q., Wu, J., & Wang, Y. (2023b). The effect of structural holes on producing novel and disruptive research in physics. *Scientometrics, 128*(3), 1801–1823.

Wei, C., Li, J., & Shi, D. (2023). Quantifying revolutionary discoveries: Evidence from Nobel prize-winning papers. *Information Processing & Management, 60*(3), 103252.

Winnink, J. J. (2017). Early-stage detection of breakthrough-class scientific research: using micro-level citation dynamics. Social and Behavioural Sciences, Leiden University PhD Thesis.

Winnink, J. J., & Tijssen, R. J. (2015). Early stage identification of breakthroughs at the interface of science and technology: Lessons drawn from a landmark publication. *Scientometrics, 102*, 113–134.

Winnink, J. J., Tijssen, R. J., & Van Raan, A. F. J. (2019). Searching for new breakthroughs in science: How effective are computerised detection algorithms? *Technological Forecasting and Social Change, 146*, 673–686.

Wolcott, H. N., Fouch, M. J., Hsu, E. R., DiJoseph, L. G., Bernaciak, C. A., Corrigan, J. G., & Williams, D. E. (2016). Modeling time-dependent and-independent indicators to facilitate identification of breakthrough research papers. *Scientometrics, 107*, 807–817.

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature, 566*(7744), 378–382.

Wu, Q., & Yan, Z. (2019). Solo citations, duet citations, and prelude citations: New measures of the disruption of academic papers. arxiv preprint arXiv:1905.03461.

Xu, F., Wu, L., & Evans, J. (2022a). Flat teams drive scientific innovation. *Proceedings of the National Academy of Sciences, 119*(23), e2200927119.

Xu, H., Bu, Y., Liu, M., Zhang, C., Sun, M., Zhang, Y., & Ding, Y. (2022b). Team power dynamics and team impact: New perspectives on scientific collaboration using career age as a proxy for team power. *Journal of the Association for Information Science and Technology, 73*(10), 1489–1505.

Xu, H., Luo, R., Winnink, J., Wang, C., & Elahi, E. (2022c). A methodology for identifying breakthrough topics using structural entropy. *Information Processing & Management, 59*(2), 102862.

Xu, J., Kim, S., Song, M., Jeong, M., Kim, D., Kang, J., & Ding, Y. (2020). Building a PubMed knowledge graph. *Scientific Data, 7*(1), 205.

Yang, Y., Tian, T. Y., Woodruff, T. K., Jones, B. F., & Uzzi, B. (2022). Gender-diverse teams produce more novel and higher-impact scientific ideas. *Proceedings of the National Academy of Sciences, 119*(36), e2200841119.

Yu, H., & Liang, Y. (2024). A framework for predicting scientific disruption based on graph signal processing. *Information Processing & Management, 61*(6), 103863.

Yu, H., Liang, Y., & Xie, Y. (2024). Predicting scientific breakthroughs based on structural dynamic of citation cascades. *Mathematics, 12*(11), 1741.

Zeng, A., Fan, Y., Di, Z., Wang, Y., & Havlin, S. (2021). Fresh teams are associated with original and multidisciplinary research. *Nature Human Behaviour, 5*(10), 1314–1322.

Zhang, X., Xie, Q., & Song, M. (2021). Measuring the impact of novelty, bibliometric, and academic-network factors on citation count using a neural network. *Journal of Informetrics, 15*(2), 101140.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.