

From Newborn to Impact: Bias-Aware Citation Prediction

Mingfei Lu

University of Technology Sydney
Sydney, Australia
mingfei.lu@student.uts.edu.au

Mengjia Wu

University of Technology Sydney
Sydney, Australia
mengjia.wu@uts.edu.au

Jiawei Xu

University of Texas at Austin
Austin, United States
jiaweixu@utexas.edu

Weikai Li

University of California, Los Angeles
Los Angeles, United States
weikaili@cs.ucla.edu

Feng Liu

The University of Melbourne
Melbourne, Australia
feng.liu1@unimelb.edu.au

Ying Ding

University of Texas at Austin
Austin, United States
ying.ding@austin.utexas.edu

Yizhou Sun

University of California, Los Angeles
Los Angeles, United States
yzsun@cs.ucla.edu

Jie Lu

University of Technology Sydney
Sydney, Australia
jie.lu@uts.edu.au

Yi Zhang*

University of Technology Sydney
Sydney, Australia
yi.zhang@uts.edu.au

Abstract

As a key to accessing research impact, citation dynamics underpins research evaluation, scholarly recommendation, and the study of knowledge diffusion. Citation prediction is particularly critical for newborn papers, where early assessment must be performed without citation signals and under highly long-tailed distributions. We identify two key research gaps: (i) insufficient modeling of implicit factors of scientific impact, leading to reliance on coarse proxies; and (ii) a lack of bias-aware learning that can deliver stable predictions on lowly cited papers. We address these gaps by proposing a Bias-Aware Citation Prediction Framework, which combines multi-agent feature extraction with robust graph representation learning. First, a multi-agent \times graph co-learning module derives fine-grained, interpretable signals, such as reproducibility, collaboration network, and text quality, from metadata and external resources, and fuses them with heterogeneous-network embeddings to provide rich supervision even in the absence of early citation signals. Second, we incorporate a set of robust mechanisms: a two-stage forward process that routes explicit factors through an intermediate exposure estimate, GroupDRO to optimize worst-case group risk across environments, and a regularization head that performs what-if analyses on controllable factors under monotonicity and smoothness constraints. Comprehensive experiments on two real-world datasets demonstrate the effectiveness of our proposed model. Specifically, our model achieves around a 13% reduction in error metrics (MAE and RMSLE) and a notable 5.5% improvement in the ranking metric (NDCG) over the baseline methods. The code can be found at <https://github.com/Maekfei/BA-Cite>.

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'17, Washington, DC, USA

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

CCS Concepts

• Information systems → World Wide Web.

Keywords

Citation prediction; Multi-agent systems; Graph neural networks

1 Introduction

Citation dynamics, as a key to accessing research impact, is crucial for research evaluation, scholarly recommendation, and the study of knowledge diffusion. Citation prediction has thus become a particularly significant task for identifying scientific innovation from newborn papers. However, citation data are highly biased. Implicit factors such as reproducibility are not well reflected in modeling, while overly correlated explicit factors lead to shortcuts, e.g., top venue = high citations [14, 26]. Thus, accurately and robustly predicting citations in such a cold-start scenario remains a pressing challenge. As citation behaviors unfold across large-scale web-based scholarly platforms, addressing this challenge also contributes to robust and generalizable web mining of academic content and information diffusion.

Previous work on citation prediction falls largely into two categories. (1) Early-cascade models use initial citation dynamics as predictors. DeepCas [15] treats the citation cascade of a paper as a sequence, generating diffusion pathways through random walks and modeling them with BiGRU and attention to capture early spread signals. SI-HDGNN [35] further embeds these processes in heterogeneous dynamic academic graphs, combining multi-relational structures with early citation sequences to forecast long-term scientific impact. However, these models require years of waiting for early citations to accumulate, making them ineffective in the cold-start stage when timely prediction is most needed. (2) Metadata-driven models leverage the author, institution, venue, and related descriptors. For example, HINTS [12] models dynamic heterogeneous information networks, via graph neural networks (GNNs) to capture temporal evolution in citation time series. Cluster-Aware Text-Enhanced HGNN [36] integrates signals with cluster-level and textual features to improve prediction. However, citation distributions are highly long-tailed [12, 20], and these models perform

poorly on lowly cited papers. Moreover, they tend to overfit correlations with explicit factors while overlooking deeper implicit factors that may have comparably higher potential to influence citation behaviors in real-world scientific activities, leading to significant performance degradation under distribution shifts.

Despite some promising solutions, existing methods cannot achieve accurate and robust citation prediction for particularly lowly cited papers due to the following two research gaps: **Gap 1: Insufficient attention to implicit factors, resulting in reliance on strong correlations with explicit factors and poor generalization across domains.** Current models rely heavily on factors such as author reputation and venue prestige, which correlate with citations but cannot comprehensively cover all decisive determinants of citation behaviors. Lacking fine-grained representations of implicit factors such as topic hotness, reproducibility, and collaboration structure, models would default to superficial correlations, which ultimately degrade performance when the data distribution shifts. **Gap 2: A lack of bias-aware models that remain robust on lowly cited papers.** In this work, bias refers to the group-level prediction disparity induced jointly by the long-tailed citation distribution, feature sparsity in cold-start settings, and the empirical risk minimization (ERM) objective that minimizes average risk [7, 17]. This structural bias causes models to systematically underperform on low-citation subgroups. Most existing methods overlook the issue, since minimizing overall error inherently biases training toward highly cited papers that dominate the loss. As a result, lowly cited papers remain underrepresented and their citation dynamics is poorly predicted, undermining the early evaluation of underrepresented elements in the research community, e.g., early career researchers and emerging research directions. Based on these gaps, we pose our core question:

How can we design bias-aware models that deliver stable predictions on low-citation papers while revealing the underlying drivers of scientific impact?

To bridge these two significant gaps, we propose a Bias-Aware Citation Prediction Framework, termed *BA-Cite*, which combines fine-grained feature extraction with robust GNN learning. Specifically, **to tackle Gap 1**, we design a multi-source informed graph learning framework that jointly models agent-derived implicit factors and heterogeneous graph structures. Six agents automatically extract fine-grained features, including reproducibility, text quality, collaboration network, topical hotness, venue prestige and role-aware author reputation, from metadata and external resources. Instead of serving as isolated inputs, these signals are fused with graph-based paper embeddings and propagated through a two-stage predictor, allowing the model to integrate implicit factors with graph context. In this way, the framework reduces reliance on explicit proxies and achieves stronger generalization. **To overcome Gap 2**, we center learning on a two-stage predictor: The model initially estimates an intermediate exposure variable from graph embeddings and agent-derived features, and then predicts citations based on both the exposure estimate and the remaining features, while excluding superficial correlates from direct inputs. Robust learning objectives are attached to the second stage’s outputs: (i) Group Distributionally Robust Optimization (GroupDRO)

minimizes the worst-group risk on the prediction loss, countering head-dominated bias; and (ii) a regularization module performs what-if interventions on controllable factors, recomputing exposure and predictions while enforcing monotonicity and smoothness constraints. These objectives are jointly optimized and back-propagated through both stages and the graph encoder, shaping the pipeline toward bias-resistant and consistent behavior. We conduct systematic evaluations on two large-scale academic datasets, AMiner and OpenAlex. Experimental results demonstrate that, compared with the state-of-the-art baselines, our framework achieves around a 13% reduction in error metrics (MAE and RMSLE) and a notable 5.5% improvement in the ranking metric (NDCG).

The main contributions of this paper are highlighted as follows:

- (1) Empirical finding. We identify that prior citation predictors overfit explicit signals, leading to degradation on long-tail papers and under distribution shifts.
- (2) Multi-source Informed Graph Learning Framework. We propose a collaborative framework that combines agent-based fine-grained implicit feature extraction with graph representation learning, enabling robust prediction even without early citation information.
- (3) Bias-Aware GNN Learning. We design a robust mechanism that integrates Stage-A/Stage-B modeling, GroupDRO, and a regularization module, thereby suppressing superficial correlations, highlighting true correlations, and enhancing both accuracy and robustness.

2 Related Work

Research on scientific impact prediction and citation-network modeling falls into three directions: early-cascade modeling, metadata-driven graph learning, and dynamic or contrastive graph approaches.

Early-cascade models frame citation accumulation as a diffusion process. DeepCas [15] captures early propagation via random-walk citation paths and BiGRU attention, while SI-HDGNN [35] embeds such cascades into heterogeneous academic graphs. Despite their effectiveness with sufficient citations, these models perform poorly in cold-start settings where early prediction is crucial.

Metadata-driven graph models leverage structural and textual metadata such as authors, venues, and topics. HINTS [12] encodes dynamic heterogeneous networks with R-GCN and GRU, while CATE-HGNN [36] and HLM-Cite [6] enrich semantics via clustering and pretrained language models. However, they remain sensitive to long-tail imbalance and often overfit superficial correlations.

Dynamic and contrastive frameworks model evolving citation contexts more explicitly. H2CGL [8] builds hierarchical heterogeneous graphs with citation-aware GIN, relation-aware GAT, and contrastive learning to integrate structural and temporal cues. Related work such as TGN-TRec [27], NETEVOLVE [19], and ResearchTown [39] explores dynamic or agent-based paradigms emphasizing interpretability and network evolution. More recently, From Words to Worth [41] proposes newborn article impact prediction, showing that fine-tuned LLMs can infer normalized impact (TNCSISP) from titles and abstracts alone, achieving competitive performance without citation history or external metadata.

LLM-based semantic feature extraction. Recent studies leverage large language models (LLMs) to extract high-level semantic

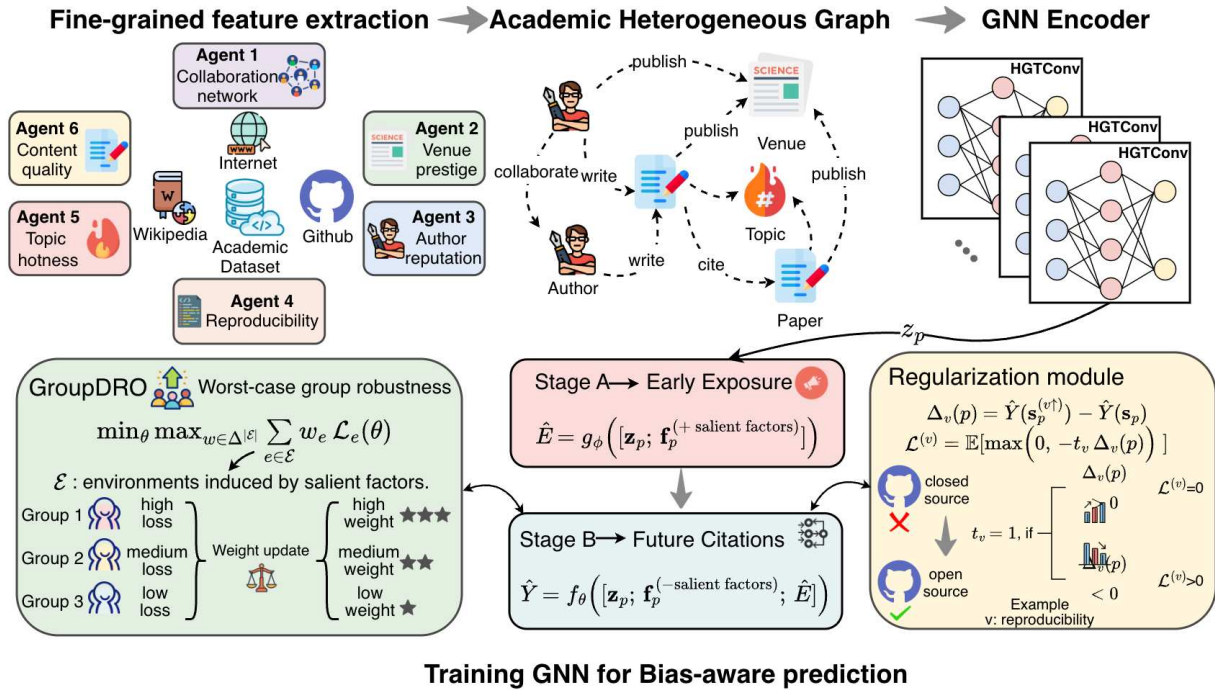


Figure 1: Overall architecture of the proposed BA-Cite framework.

features[3, 16, 33, 34, 38] from titles, abstracts, and metadata for scholarly impact prediction. LLMs provide contextualized representations and latent signals such as novelty or interdisciplinarity, which are incorporated into downstream graph models as auxiliary embeddings[2, 11, 37], especially under cold-start settings. However, these features are typically treated as static enhancements, without explicitly modeling their interaction with dynamic graph structures or citation bias.

Despite these advances, few studies examine the mechanisms and bias dynamics driving citation disparities. We address this gap with a unified multi-source graph learning framework, enabling robust citation prediction under cold-start and distribution shifts.

3 Methodology

In this section, we present our framework, BA-Cite, illustrated in Fig. 1. It consists of two parts: (i) agent-based fine-grained feature extraction, where multiple agents derive implicit factors from metadata and external resources; and (ii) graph learning on dynamic heterogeneous networks, where a GNN encoder integrates these signals through a two-stage forward process with bias- and robustness-oriented objectives. In the following parts, we first outline the motivation, then describe the functions of individual agents, and finally detail the three GNN modules that incorporate agent-derived features into bias-aware representation learning.

3.1 Motivation

3.1.1 Addressing Bias in Long-Tailed Citation Prediction. In citation prediction, the inherently long-tailed distribution creates a persistent imbalance: a few highly cited papers dominate the learning process, while the majority receive limited attention. This skew causes models to produce inaccurate predictions—often overestimating lowly cited papers and failing to generalize under distribution shifts, such as when predicting citations for evaluating new venues or identifying emerging research topics. Such instability weakens the predictive reliability and limits the model’s ability to capture the real scientific impact across domains and time, for example, leading to biased predictive results that underestimate the contributions of early-career researchers and non-mainstream research. Addressing bias under the long-tailed and shifting distributions is therefore essential not only for improving prediction robustness but also for promoting fairness and inclusiveness in scientific assessment.

3.1.2 Capturing Core Factors Driving Citation Dynamics with Agents and GNN(s). Existing citation prediction models often rely on hand-crafted metadata or early citation signals, which are unavailable for newborn papers in cold-start settings and difficult to model manually. Agents, by contrast, can autonomously mine and reason over multi-source information — such as reproducibility, topic hotness, and collaboration network — to derive fine-grained, high-level semantic factors that are otherwise implicit or sparsely encoded in metadata. However, semantic cues alone cannot capture the structural and temporal dependencies that shape citation dynamics. While agents excel at extracting implicit knowledge, GNNs effectively model heterogeneous academic networks. Combining them

enables semantic knowledge to be structurally grounded, ensuring both predictive accuracy and robustness.

3.2 Fine-grained feature extraction

To enrich paper representations, we design six domain-specific agents that automatically derive implicit features from metadata and external resources. Each agent focuses on a distinct dimension of citation dynamics, leveraging domain heuristics and multi-source knowledge to uncover fine-grained semantic cues that are difficult to encode manually. Together, these agents capture complementary aspects such as author reputation, venue prestige, collaboration patterns, reproducibility, topic hotness, and text quality, thereby expanding the feature space and enabling more balanced and generalizable representations. Given a paper p , the outputs are concatenated into a unified feature vector $\mathbf{f}_p = [A, V, R, C, H, Q]$, where each component corresponds to one agent’s extracted factor. Below we describe the reason for choosing these factors and how each agent extracts them from metadata and external sources.

Role-aware Author Reputation (A). Readers tend to cite works by reputable scholars or rising researchers, yet an author’s position within the byline also matters—first authors indicate primary contribution, last authors reflect senior leadership, while middle authors exert weaker influence [21].

Extraction process. We assess author reputation by partitioning the author list into three roles: first author, last author, and other co-authors. For each role, we retrieve metadata such as institutional affiliation, publication count, and total citations; institutional prestige is further enriched via external sources (e.g., Wikipedia). These signals are aggregated into a continuous score on a 1–5 scale. During training, the scores of different roles are assigned different weights to reflect their varying influence on citation outcomes.

Venue Prestige (V). Prestigious venues act as credibility signals, making their publications more visible and trusted, hence more likely to be cited [1].

Extraction process. We assess venue prestige by matching the venue name against external rankings (China Computer Federation Recommended Rankings (CCF) and Computing Research and Education Association of Australasia (CORE)) using both exact and fuzzy matching. The agent outputs a score on a 1–5 scale.

Reproducibility (R). Open-source code or data enhances transparency, and reuse, driving credibility and long-term impact [23].

Extraction process. We assess reproducibility by scanning the content for open-source indicators (e.g., GitHub/GitLab links). If links are detected, we verify whether the repository contains code or datasets. The agent outputs a binary score (0/1).

Collaboration Network (C). Broad and diverse collaborations, especially across institutions or countries, increase attention and citation potential through higher credibility and dissemination [32].

Extraction process. We assess collaboration characteristics, including team size, institutional diversity, and cross-country collaboration. Institutional metadata are complemented with external lookups (e.g., Wikipedia) to estimate prestige and geographic dispersion. The agent outputs a composite score on a 1–5 scale.

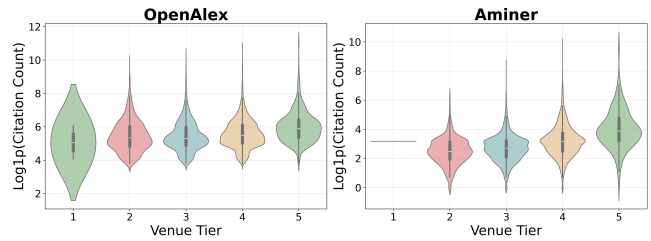


Figure 2: Distribution of citation counts across venues of different prestige levels on the Aminer and OpenAlex datasets.

Topic Hotness (H). Work in trending or growing areas gains citations faster by aligning with the community’s interests [31]. *Extraction process.* We assess topical hotness using the paper’s keywords. For each keyword, we count the number of papers in the previous year; the mean count across keywords is used as the hotness score. The agent outputs a continuous value.

Text Quality (Q). Clear, well-structured titles and abstracts improve readability, directly influencing citation outcomes [13].

Extraction process. We assess text quality by prompting an LLM with the paper’s title and abstract, together with best-paper exemplars as references. The LLM evaluates structural clarity and professional expression and produces a score on a 1–5 scale.

Output. The six features are concatenated into \mathbf{f}_p and injected as attributes of the paper node p in the heterogeneous academic graph. These enriched node features are then propagated through the GNN encoder, enabling the model to jointly capture structural patterns and implicit semantic signals.

3.3 Bias-Aware GNN Learning Modules

With agent-derived features injected as paper-node attributes in the heterogeneous graph, the GNN encoder propagates them together with structural and relational information to form enriched paper representations. A key challenge is that conventional metadata-based models often overfit explicit correlations such as venue, which reflect distributional skew rather than true importance. This shortcut undermines prediction on high-impact papers outside top venues and widens gaps on lowly cited cases. To address these issues, we introduce three complementary modules: (i) a two-stage predictor that first estimates an intermediate exposure variable from graph embeddings and agent-derived features, and then predicts future citations based on both the exposure estimate and remaining signals; (ii) a GroupDRO objective applied to the prediction loss, which minimizes worst-group risk and alleviates head-dominated bias; and (iii) a counterfactual intervention head that perturbs controllable factors and regularizes predictions through monotonicity and smoothness constraints. Together, these modules reduce over-reliance on explicit venue cues, leverage implicit drivers of scientific impact, and yield more robust citation predictions.

3.3.1 Two-Stage Forward Strategy for Fair Representation. Venue serves as the most salient shortcut feature in citation prediction: it is strongly correlated with citation counts but does not constitute a true determinant [5, 12, 29]. As shown in Fig. 2, higher-ranked

venues typically exhibit higher average citation counts, but individual papers still show wide variation in citations. In computer science, this distributional pattern often arises because top venues cluster popular topics, attract well-known researchers, and also a larger proportion of papers release open-source code. A robust predictor should therefore capture the joint influence of multiple factors on future citations rather than being misled by a single explicit signal. To this end, we isolate venue effects by enforcing the influence pathway $V \rightarrow E \rightarrow Y$, where E denotes *early exposure* and Y denotes future citations. A heterogeneous graph encoder first produces paper embeddings; Stage A estimates E using venue (among other features), while Stage B predicts Y without direct venue input, letting V influence Y only through \hat{E} .

Stage A: Exposure Estimation ($V \rightarrow E$). Stage A estimates the latent early exposure variable \hat{E} from both the graph encoder and metadata features. The encoder operates on the full heterogeneous graph that includes venue nodes, so the paper embedding \mathbf{z}_p already incorporates venue effects. Together with the venue-inclusive feature vector $\mathbf{f}_p^{(+V)} = [V, R, C, H, Q, Y, A_1, A_2, A_3]$, which includes venue prestige, reproducibility (R), collaboration network (C), topic hotness (H), text quality (Q), publication year (Y), and differentiated author reputations for first (A_1), last (A_2), and other co-authors (A_3), these signals are passed to a feed-forward head g_ϕ to produce \hat{E} .

Stage B: Venue-Excluded Prediction ($E \rightarrow Y$). Stage B predicts the final citation count using a simplified graph where venue nodes and edges are removed, yielding a venue-excluded paper embedding. The input to the predictor f_θ is the concatenation of this embedding, the venue-excluded feature vector $\mathbf{f}_p^{(-V)}$, and the Stage A estimate \hat{E} . In this way, venue influences the outcome only indirectly via exposure, enforcing the influence path $V \rightarrow E \rightarrow Y$.

Log-MSE Output and Prediction Loss. Instead of a negative-binomial likelihood, we adopt a mean squared error (MSE) objective after applying a logarithmic transformation to citation counts in order to mitigate the heavy-tailed distribution. Specifically, the predictor outputs $\hat{Y} = f_\theta([\mathbf{z}_p; \mathbf{f}_p^{(-V)}; \hat{E}])$, and the loss is defined as

$$\mathcal{L}_{\text{pred}} = \left(\log(1 + y_p) - \log(1 + \hat{Y}) \right)^2, \quad (1)$$

where y_p denotes the observed citation count of paper p .

Module Discussion. In this module, we restructure the prediction pipeline into two stages:

$$\text{Stage A: } \hat{E} = g_\phi([\mathbf{z}_p; \mathbf{f}_p^{(+V)}]), \quad \text{Stage B: } \hat{Y} = f_\theta([\mathbf{z}_p; \mathbf{f}_p^{(-V)}; \hat{E}]),$$

Stage B does not directly observe V , so the only pathway is $V \rightarrow E \rightarrow Y$. From an information-theoretic view, $I(S; Y | \hat{E}) < I(S; Y)$, meaning the shortcut influence of S on Y is strictly reduced and the model is forced to rely more on other implicit drivers.

3.3.2 Environment-Aware Optimization for Robust Generalization. We apply GroupDRO [25] to the Stage-B prediction loss to prioritize the worst environment and improve cross-environment generalization. To mitigate the venue-dominated shortcut, we partition the training data into two environments by venue tier, $\mathcal{E} = \{\text{low}, \text{high}\}$. For environment $e \in \mathcal{E}$ with index set \mathcal{D}_e , the group risk is defined

as the mean loss on samples from that environment:

$$\mathcal{L}_e(\theta) = \frac{1}{|\mathcal{D}_e|} \sum_{i \in \mathcal{D}_e} \ell(f_\theta(x_i), y_i), \quad (2)$$

where $f_\theta(x_i)$ denotes the model prediction for paper i , y_i is the observed citation count, and $\ell(\cdot)$ is the *Stage-B* prediction loss defined in Eq. 1 (Log-MSE on $\log(1 + y)$).

GroupDRO objective. We optimize a worst-group risk via adversarial reweighting:

$$\min_{\theta} \max_{w \in \Delta^2} \sum_{e \in \{\text{low}, \text{high}\}} w_e \mathcal{L}_e(\theta), \quad \Delta^2 = \left\{ \begin{array}{l} w_{\text{low}}, w_{\text{high}} \geq 0, \\ w_{\text{low}} + w_{\text{high}} = 1 \end{array} \right\}, \quad (3)$$

where $w = (w_{\text{low}}, w_{\text{high}})$ denotes nonnegative environment weights lying in the probability simplex Δ^2 . The inner maximization allocates more weight to the environment with larger risk, forcing the model to improve performance on the worst-performing group.

Weight update. Let $\bar{\mathcal{L}} = \frac{1}{2} \sum_e \mathcal{L}_e$ and $\sigma_{\mathcal{L}} = \sqrt{\frac{1}{2} \sum_e (\mathcal{L}_e - \bar{\mathcal{L}})^2}$ (the population standard deviation of group risks). Here \mathcal{L}_e is the average loss of environment e , $\bar{\mathcal{L}}$ is the mean loss across environments, and $\sigma_{\mathcal{L}}$ is their standard deviation. With step size $\alpha > 0$ and small constant ε ,

$$\tilde{w}_e \propto w_e^{(t)} \exp\left(\alpha \frac{\mathcal{L}_e - \bar{\mathcal{L}}}{\sigma_{\mathcal{L}} + \varepsilon}\right), \quad \sum_e \tilde{w}_e = 1, \quad (4)$$

where $w_e^{(t)}$ is the current weight of environment e . Higher-than-average loss increases the weight, while lower loss decreases it. Finally, weights are clamped and renormalized:

$$\hat{w}_e = \text{clip}(\tilde{w}_e, w_{\min}, w_{\max}), \quad w_e^{(t+1)} = \frac{\hat{w}_e}{\sum_{e'} \hat{w}_{e'}}. \quad (5)$$

Here $[w_{\min}, w_{\max}]$ bounds prevent degenerate values, and $w_e^{(t+1)}$ denotes the updated weight for environment e .

Module Discussion. This environment-aware optimization prevents the model from collapsing onto dominant groups shaped by highly cited or high-prestige papers. By forcing improvements on underrepresented environments, GroupDRO enhances fairness, promoting robust generalization across diverse citation contexts.

3.3.3 Regularization for Reasonable and Stable Prediction. To provide actionable “what-if” estimates, we augment Stage B with a regularization head. This module turns abstract features into interpretable regularization effects: it quantifies the predicted citation change if *only* a controllable factor v were improved (e.g., toggling $R: 0 \rightarrow 1$), while all other attributes remain fixed and the induced change in early exposure is propagated consistently. This yields predictions that are both actionable and constrained to be directionally reasonable and stable.

For a controllable factor v (e.g., reproducibility R), let $\mathbf{s}_p = [\mathbf{z}_p; \mathbf{f}_p^{(-V)}; \hat{E}]$ be the Stage B input constructed from the observed features, and let $\mathbf{s}_p^{(v\uparrow)}$ be the same input after setting v to a high value (keeping all other features fixed) and recomputing \hat{E} under this change. The per-factor counterfactual effect is defined as

$$\Delta_v(p) = \hat{Y}(\mathbf{s}_p^{(v\uparrow)}) - \hat{Y}(\mathbf{s}_p), \quad (6)$$

where $\hat{Y}(\cdot) = f_{\theta}(\cdot)$ and the early-exposure estimate in $\mathbf{s}_p^{(\sigma\uparrow)}$ is $\hat{E}^{(\sigma\uparrow)} = g_{\phi}([\mathbf{z}_p; \mathbf{f}_p^{(+V)} \text{ with } v \leftarrow \text{high}])$.

Monotonicity and smoothness regularization. Let $t_v \in \{+1, -1\}$ denote the expected direction of improvement (typically $t_v = +1$), and let τ_v be a threshold that marks the “low” region of v (e.g., $R=0$). We enforce that raising v should not hurt citations for low-value cases, and keep effects calibrated via a smoothness penalty:

$$\mathcal{L}_{\text{mono}}^{(v)} = \mathbb{E}[\max(0, -t_v \Delta_v(p)) \mathbf{1}\{v(p) < \tau_v\}], \quad (7)$$

$$\mathcal{L}_{\text{smooth}}^{(v)} = \mathbb{E}[\Delta_v(p)^2]. \quad (8)$$

Aggregating over controllable factors \mathcal{V} , the total regularizer is

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{mono}} \sum_{v \in \mathcal{V}} \mathcal{L}_{\text{mono}}^{(v)} + \lambda_{\text{smooth}} \sum_{v \in \mathcal{V}} \mathcal{L}_{\text{smooth}}^{(v)}. \quad (9)$$

Module Discussion. One major source of bias in citation prediction arises from the limited and coarse metadata used in traditional models, which makes explicit factors dominate the learning process. Within our regularization module, the model leverages fine-grained, implicitly derived features to regularize representation learning, thereby mitigating shortcut reliance. By distributing explanatory power across multiple implicit drivers—such as collaboration patterns, topic dynamics, and text quality—the model becomes less dependent on any single explicit factor and achieves more balanced generalization across varying citation environments.

3.3.4 Objective. Our training objective focuses on environment-aware risk (GroupDRO) and the proposed sensitivity regularization:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{main}} \mathcal{L}_{\text{groupdro}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}. \quad (10)$$

We also employ two lightweight auxiliaries on Stage B—an exposure calibration loss on \hat{E} and an adversarial venue-invariance loss—which are reported in ablations and described in Appendix B, but omitted here for brevity.

4 Experiments

In this section, we conduct extensive experiments to answer the following four research questions:

RQ1: How does BA-Cite perform compared with other models in terms of predictive accuracy, and ranking quality?

RQ2: How does each component contribute to the overall performance of BA-Cite?

RQ3: Does BA-Cite achieve robust prediction across different data distributions after bias mitigation?

RQ4: How do different hyper-parameters affect BA-Cite ?

4.1 Experimental Setup

4.1.1 Datasets. We select two widely used publicly available academic datasets, **Aminer** [28] and **OpenAlex** [22], to verify the effectiveness of our proposed framework. Both datasets focus on the *computer science* domain and contain heterogeneous nodes, as well as temporal relations such as *cites*, *writes*, and *has_topic*. The temporal coverage ranges from **2010 to 2025**.

For each paper, the prediction target is its future citation number within the next five years (starting from the second year) after publication. We split the data by year, using papers published during 2010–2018 for training, 2019 for validation, and 2020 for testing.

Within each split, we randomly sample 10,000 papers for training, 1,000 for validation, and 1,000 for testing. To reduce randomness, we repeat the sampling process three times and conduct five runs with different random seeds for each sample. The reported results include the mean and standard deviation across all runs.

4.1.2 Baselines. We compare our framework with representative methods from four categories, covering classical GNNs, sequential models, large language models, and metadata-based neural models. **Graph Neural Network-based methods.**

- **GAT (ICLR’18)** [30]: models citation relations using multi-head graph attention.
- **HINTS (WWW’21)** [12]: encodes temporal heterogeneous information networks for citation time-series prediction.
- **DyGFormer (NeurIPS’23)** [40]: applies transformer-style temporal encoding for dynamic graphs.

Sequence-based methods.

- **BiLSTM-Meta (Scientometrics’21)** [18]: captures citation sequences via bidirectional recurrent modeling.
- **DeepCas (WWW’17)** [15]: learns citation cascade representations through random walks and BiGRU-based attention.
- **SI-HDGNN (KBS’22)** [35]: builds heterogeneous dynamic academic networks for impact propagation.

Large Language Model-based methods.

- **GPT-4o (OpenAI’24)** [9]: leverages LLM reasoning and knowledge for citation impact estimation.
- **Llama-3.1-405B (Meta’24)** [4]: employs open-source LLM embeddings for academic impact inference.
- **NAIP (AAAI’25)** [41]: formulates newborn article impact prediction by fine-tuning large language models on title–abstract pairs with the TNCSISP metric, enabling content-only impact estimation without external metadata.

Metadata-based Neural Methods.

- **BP-NN (J. Informetrics’20)** [24]: A four-layer feed-forward neural network that predicts five-year citation counts.

These baselines represent diverse paradigms in scientific impact prediction, from early cascade modeling to metadata-driven, dynamic, and LLM-enhanced frameworks, providing a comprehensive comparison foundation. For all feature-dependent baselines, we supply the fine-grained semantic features extracted in Section 3.2 to ensure consistent and enriched input representations.

4.1.3 Implementation Details. We implement the counterfactual two-stage HGT in PyTorch. The heterogeneous encoder uses two HGT layers (hidden size 128, 4 attention heads, dropout 0.4). Node features follow our schema: paper nodes have an 8-dimensional vector $[R, Q, C, H, Y, A_1, A_2, A_3]$, while author/venue/topic nodes are initialized with 1-dimensional metadata features. For counterfactual learning, we enable *reproducibility* (R) and *content quality* (Q) as actionable variables and apply monotonicity regularization so that larger (R, Q) should not decrease predicted citations. We also employ adversarial training and an auxiliary loss. We adopt GroupDRO over 2 environments with step size $\alpha = 0.1$ and group-weight clipping to $[0.1, 0.9]$. Environments are constructed by venue prestige threshold $\tau = 0.8$. Optimization uses AdamW (lr = 10^{-3} , weight decay = 10^{-4}) with a 10-epoch warm-up followed by cosine decay

Table 1: Comparison on Aminer and OpenAlex: MALE, RMSLE, NDCG@10, and NDCG@20 (mean \pm std). Lower is better for MALE/RMSLE; higher is better for NDCG. Best results are in bold, second best are underlined. “ \dagger ” indicates statistically significant improvement over all baselines under a paired t-test with $p < 0.05$.

Baselines	Aminer				OpenAlex			
	MALE \downarrow	RMSLE \downarrow	NDCG@10 \uparrow	NDCG@20 \uparrow	MALE \downarrow	RMSLE \downarrow	NDCG@10 \uparrow	NDCG@20 \uparrow
<i>Graph Neural Network-based methods</i>								
GAT	0.94 \pm 0.02	1.09 \pm 0.02	0.09 \pm 0.11	0.10 \pm 0.11	<u>0.87 \pm 0.04</u>	1.14 \pm 0.03	0.23 \pm 0.17	0.25 \pm 0.16
HINTS	0.99 \pm 0.01	1.14 \pm 0.01	0.03 \pm 0.01	0.04 \pm 0.01	0.98 \pm 0.00	1.14 \pm 0.00	0.04 \pm 0.02	0.05 \pm 0.03
DyGFormer	1.36 \pm 0.36	1.55 \pm 0.38	0.27 \pm 0.09	0.26 \pm 0.07	1.09 \pm 0.23	1.36 \pm 0.24	<u>0.39 \pm 0.06</u>	<u>0.43 \pm 0.05</u>
<i>Sequence-based methods</i>								
BiLSTM	0.79 \pm 0.02	0.98 \pm 0.03	<u>0.32 \pm 0.08</u>	<u>0.33 \pm 0.06</u>	2.21 \pm 0.11	2.35 \pm 0.11	0.17 \pm 0.06	0.13 \pm 0.08
DeepCas	0.90 \pm 0.02	1.30 \pm 0.03	0.04 \pm 0.02	0.05 \pm 0.02	1.09 \pm 0.01	1.33 \pm 0.01	0.06 \pm 0.02	0.08 \pm 0.03
SI-HDGNN	<u>0.73 \pm 0.02</u>	1.01 \pm 0.02	0.16 \pm 0.10	0.17 \pm 0.09	1.05 \pm 0.04	1.31 \pm 0.04	0.11 \pm 0.07	0.05 \pm 0.05
<i>Large Language Model-based methods</i>								
GPT-4o	1.73 \pm 0.03	1.91 \pm 0.04	0.13 \pm 0.02	0.23 \pm 0.07	0.90 \pm 0.01	<u>1.12 \pm 0.00</u>	0.38 \pm 0.00	0.36 \pm 0.04
Llama_3.1_405b	1.62 \pm 0.01	1.80 \pm 0.01	0.12 \pm 0.02	0.18 \pm 0.01	1.02 \pm 0.08	1.23 \pm 0.10	0.31 \pm 0.09	0.34 \pm 0.03
NAIP_Llama	0.77 \pm 0.02	<u>0.97 \pm 0.03</u>	0.05 \pm 0.01	0.06 \pm 0.01	2.53 \pm 0.01	2.73 \pm 0.01	0.01 \pm 0.01	0.01 \pm 0.02
NAIP_Qwen	0.83 \pm 0.02	1.04 \pm 0.03	0.05 \pm 0.02	0.06 \pm 0.01	2.46 \pm 0.01	2.66 \pm 0.01	0.08 \pm 0.03	0.10 \pm 0.04
<i>Metadata-based Neural methods</i>								
BP-NN	0.84 \pm 0.05	1.05 \pm 0.06	0.27 \pm 0.12	0.31 \pm 0.12	2.25 \pm 0.25	2.38 \pm 0.24	0.23 \pm 0.07	0.20 \pm 0.04
<i>Our Method</i>								
BA-Cite	0.71 \pm 0.02\dagger	0.88 \pm 0.03\dagger	0.34 \pm 0.12\dagger	0.37 \pm 0.13\dagger	0.67 \pm 0.01\dagger	0.92 \pm 0.04\dagger	0.51 \pm 0.04\dagger	0.47 \pm 0.01\dagger

Table 2: Ablation study of BA-Cite on Aminer and OpenAlex: MALE, RMSLE, NDCG@10, and NDCG@20 (mean \pm std).

Ablation Type	Aminer				OpenAlex			
	MALE \downarrow	RMSLE \downarrow	NDCG@10 \uparrow	NDCG@20 \uparrow	MALE \downarrow	RMSLE \downarrow	NDCG@10 \uparrow	NDCG@20 \uparrow
<i>w/o Feature</i>	0.86 \pm 0.02	1.00 \pm 0.03	0.10 \pm 0.07	0.11 \pm 0.06	1.88 \pm 0.00	2.04 \pm 0.02	0.07 \pm 0.00	0.09 \pm 0.01
<i>w/o Two-Stage</i>	0.70 \pm 0.02	0.87 \pm 0.02	<u>0.26 \pm 0.18</u>	<u>0.29 \pm 0.16</u>	<u>0.78 \pm 0.03</u>	<u>0.99 \pm 0.05</u>	0.16 \pm 0.08	0.19 \pm 0.07
<i>w/o Reg</i>	0.78 \pm 0.02	0.97 \pm 0.03	0.14 \pm 0.08	0.14 \pm 0.08	1.21 \pm 0.20	1.30 \pm 0.18	<u>0.18 \pm 0.09</u>	<u>0.21 \pm 0.07</u>
<i>w/o GroupDRO</i>	0.79 \pm 0.02	0.89 \pm 0.03	0.01 \pm 0.00	0.03 \pm 0.01	0.99 \pm 0.02	1.09 \pm 0.03	0.05 \pm 0.02	0.07 \pm 0.03
BA-Cite	<u>0.71 \pm 0.02</u>	<u>0.88 \pm 0.03</u>	0.34 \pm 0.12	0.37 \pm 0.13	0.67 \pm 0.01	0.92 \pm 0.04	0.51 \pm 0.04	0.47 \pm 0.01

to 10^{-5} . We train up to 200 epochs with batch size 128 and early stopping on validation loss.

4.1.4 Evaluation Metrics. We evaluate model performance from two perspectives: (i) point accuracy using Mean Absolute Log Error (MALE) and Root Mean Squared Log Error (RMSLE); and (ii) ranking quality using Normalized Discounted Cumulative Gain (NDCG@K, $K=10, 20$) [10]. All log-based metrics adopt $\log(1+y)$ to mitigate heavy-tailed citation counts. Lower is better for MALE/RMSLE, while higher is better for NDCG. The detailed calculation formulas are provided in the Appendix A.

4.2 Overall Performance (RQ1)

Table 1 summarizes the overall results on the Aminer and OpenAlex datasets. In general, BA-Cite achieves the best performance across all metrics. On Aminer, it outperforms the best baselines by reducing MALE by 2.7% and RMSLE by 9.3%, while improving NDCG@10

and NDCG@20 by 2% and 4%. On OpenAlex, BA-Cite further reduces MALE by 20.3% and RMSLE by 20.1%, and boosts NDCG@10 and NDCG@20 by 12% and 4%. These consistent gains demonstrate that integrating agent-derived fine-grained features with dynamic heterogeneous graph learning effectively improves both accuracy and ranking quality. Compared to other methods, GNN-based models generally perform better than sequence- or metadata-based baselines, highlighting the importance of structural and temporal modeling. LLM-based methods achieve competitive ranking performance, which may benefit from pretrained knowledge rather than structural understanding. Overall, BA-Cite provides balanced and stable improvements, confirming its advantage in handling cold-start and long-tailed citation prediction scenarios.

4.3 Ablation Study (RQ2)

We examine the effect of each module in BA-Cite on Aminer and OpenAlex (Table 2). Removing fine-grained feature extraction (*w/o*

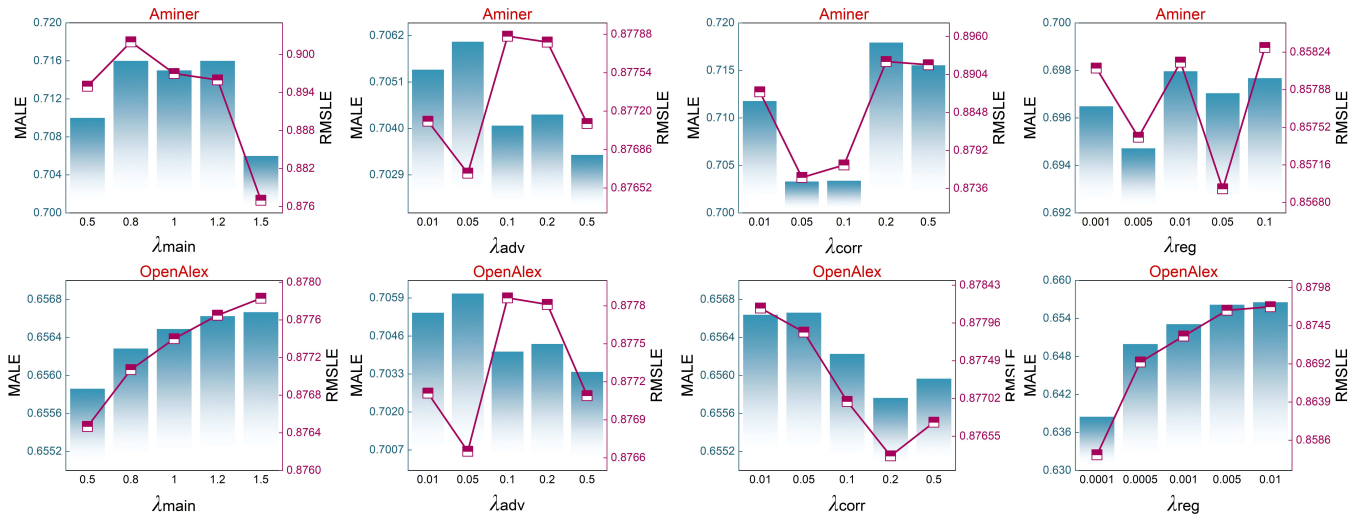


Figure 3: Parameter sensitivity results of BA-Cite on the Aminer and OpenAlex datasets.

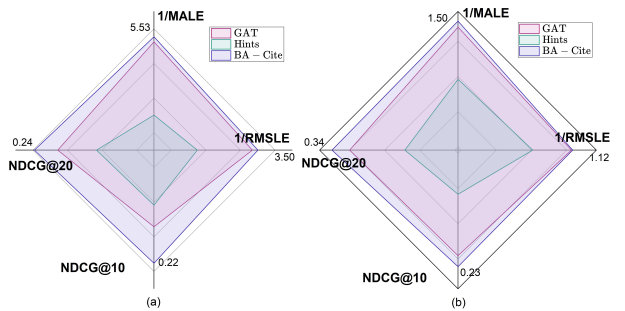


Figure 4: Performance comparison of three models on Aminer dataset with different citation levels. (a) Results on lowly cited papers. (b) Results on highly cited papers.

Feature) leads to the largest drop, confirming the importance of agent-derived semantic cues. Eliminating the two-stage process (w/o Two-Stage) also reduces ranking quality, showing that staged optimization enhances representation robustness. However, on the relatively unbiased Aminer dataset (as evidenced in Fig. 2), the two-stage mechanism introduces slight structural redundancy and optimization noise, which marginally reduces fitting efficiency despite its advantage in bias mitigation. Without bias regularization (w/o Reg), both MALE and RMSLE slightly increase, indicating reduced robustness. The absence of environment-aware training (w/o GroupDRO) causes severe degradation in NDCG, proving its necessity for balanced performance across domains. Overall, BA-Cite achieves robust and balanced performance, demonstrating that both agent-based feature extraction and bias-aware two-stage optimization are crucial for accurate and fair citation prediction.

4.4 Analysis of Robustness (RQ3)

As shown in Fig. 4, we further evaluate whether BA-Cite maintains stable performance after bias mitigation under varying data distributions. Following the partitioning strategy in [12], we divide

papers into lowly cited and highly cited subsets by citation counts. Across all three compared models, BA-Cite achieves the best performance, showing strong results on both low- and high-citation papers, which indicates its robustness to distributional variation.

4.5 Analysis of Parameter Sensitivity (RQ4)

As shown in Fig. 3, we analyze the sensitivity of four loss weights. (1) λ_{main} . On Aminer, errors decrease with larger λ_{main} , peaking at 1.2–1.5, indicating that a stronger main objective improves fit under mild bias. On OpenAlex, MALE and RMSLE increase with λ_{main} , so smaller values (0.5–0.8) are preferable to preserve capacity for debiasing. (2) λ_{adv} . A small-to-moderate adversarial strength works best: Aminer achieves its lowest errors around 0.05, while OpenAlex shows a trade-off—MALE near 0.05 and RMSLE near 0.2—suggesting λ_{adv} in 0.05–0.2. Larger values introduce instability without gains. (3) λ_{corr} . Both datasets exhibit a U-shaped trend: mild regularization helps (optimum around 0.1–0.2), whereas overly weak or strong settings (e.g., 0.01 or 0.5) degrade performance by allowing redundancy or over-constraining embeddings. (4) λ_{reg} . Fairness regularization should be conservative. Aminer reaches minimum errors around 0.05, while OpenAlex favors very small values. Strong regularization degrades accuracy on both datasets.

5 Conclusion

We present BA-Cite, a bias-aware citation prediction framework combining multi-agent semantic extraction with dynamic heterogeneous graph learning. By modeling author, venue, and topic dynamics, BA-Cite provides a strong semantic basis for citation reasoning. Its two-stage optimization with GroupDRO enhances robustness and mitigates overfitting to high-prestige environments. Experiments on Aminer and OpenAlex show consistent improvements over strong baselines, with stability confirmed by ablation and sensitivity analyses. BA-Cite generalizes well under distribution shifts, supporting real-world scholarly impact evaluation. Future work

includes impact explanation, cross-domain transfer, and reinforcement learning-based adaptive bias mitigation.

References

- [1] Michael Callahan, Robert L Wears, and Ellen Weber. 2002. Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *Jama* 287, 21 (2002), 2847–2850.
- [2] Junyi Chen, Mengjia Wu, Qian Liu, and Yi Zhang. 2026. Explainable prediction of knowledge recombination: A synergized method with heterogeneous hypergraph learning and large language models. *Information Processing & Management* 63, 1 (2026), 104336. doi:10.1016/j.ipm.2025.104336
- [3] Siming Deng, Runsong Jia, Chunjuan Luan, Mengjia Wu, and Yi Zhang. 2026. AI-enhanced multi-dimensional measurement of technological convergence through heterogeneous graph and semantic learning. *Scientometrics* (2026), 1–28.
- [4] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints* (2024), arXiv:2407.
- [5] Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. 2018. Science of science. *Science* 359, 6379 (2018), eaao0185.
- [6] Qianyue Hao, Jingyang Fan, Fengli Xu, Jian Yuan, and Yong Li. 2024. Hlm-cite: Hybrid language model workflow for text-based scientific citation prediction. *Advances in Neural Information Processing Systems* 37 (2024), 48189–48223.
- [7] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*. PMLR, 1929–1938.
- [8] Guoxiu He, Zhikai Xue, Zhuoren Jiang, Yangyang Kang, Star Zhao, and Wei Lu. 2023. H2CGL: Modeling dynamics of citation network for impact prediction. *Information Processing & Management* 60, 6 (2023), 103512.
- [9] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).
- [10] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.
- [11] Runsong Jia, Mengjia Wu, Ying Ding, Jie Lu, and Yi Zhang. 2025. HetGCOT: Heterogeneous Graph-Enhanced Chain-of-Thought LLM Reasoning for Academic Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 15950–15963. doi:10.18653/v1/2025.findings-emnlp.864
- [12] Song Jiang, Bernard Koch, and Yizhou Sun. 2021. HINTS: Citation Time Series Prediction for New Publications via Dynamic Heterogeneous Information Network Embedding. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) (WWW '21). Association for Computing Machinery, New York, NY, USA, 3158–3167. doi:10.1145/3442381.3450107
- [13] Tan Jin, Huiqiong Duan, Xiaofei Lu, Jing Ni, and Kai Guo. 2021. Do research articles with more readable abstracts receive higher online attention? Evidence from Science. *Scientometrics* 126, 10 (2021), 8471–8490.
- [14] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased learning-to-rank with biased feedback. In *Proceedings of the tenth ACM international conference on web search and data mining*. 781–789.
- [15] Cheng Li, Jiaqi Ma, Xiaoxiao Guo, and Qiaozhu Mei. 2017. Deepcas: An end-to-end predictor of information cascades. In *Proceedings of the 26th international conference on World Wide Web*. 577–586.
- [16] Xin Li, Zhuo Cai, Shoujin Wang, Kun Yu, and Fang Chen. 2025. A survey on enhancing causal reasoning ability of large language models. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 399–416.
- [17] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. 2021. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*. PMLR, 6781–6792.
- [18] Anqi Ma, Yu Liu, Xiujuan Xu, and Tao Dong. 2021. A deep-learning based citation count prediction model with paper metadata semantic features. *Scientometrics* 126, 8 (2021), 6803–6823.
- [19] Kentaro Miyake, Hiroyoshi Ito, Christos Faloutsos, Hirotomo Matsumoto, and Atsuyuki Morishima. 2024. Netevolve: Social network forecasting using multi-agent reinforcement learning with interpretable features. In *Proceedings of the ACM Web Conference 2024*. 2542–2551.
- [20] Mark EJ Newman. 2005. Power laws, Pareto distributions and Zipf’s law. *Contemporary physics* 46, 5 (2005), 323–351.
- [21] Alexander Michael Petersen, Santo Fortunato, Raj K Pan, Kimmo Kaski, Orion Penner, Armando Rungi, Massimo Riccaboni, H Eugene Stanley, and Fabio Pamolli. 2014. Reputation and impact in academic careers. *Proceedings of the National Academy of Sciences* 111, 43 (2014), 15316–15321.
- [22] Jason Priem, Heather Piwowar, and Richard Orr. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833* (2022).
- [23] Edward Raff. 2023. Does the market of citations reward reproducible work?. In *Proceedings of the 2023 ACM Conference on Reproducibility and Replicability*. 89–96.
- [24] Xuanmin Ruan, Yuanyang Zhu, Jiang Li, and Ying Cheng. 2020. Predicting the citation counts of individual papers via a BP neural network. *Journal of Informetrics* 14, 3 (2020), 101039.
- [25] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731* (2019).
- [26] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*. PMLR, 1670–1679.
- [27] Junhao Shen, Mohammad Ausaf Ali Haqqani, Beichen Hu, Cheng Huang, Xihao Xie, Tsengdar Lee, and Jia Zhang. 2024. Temporal Graph Neural Network-Powered Paper Recommendation on Dynamic Citation Networks. *arXiv preprint arXiv:2408.15371* (2024).
- [28] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 990–998.
- [29] Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. 2013. Atypical combinations and scientific impact. *Science* 342, 6157 (2013), 468–472.
- [30] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).
- [31] Tian Wei, Menghui Li, Chensheng Wu, Xiao-Yong Yan, Ying Fan, Zengru Di, and Jinshan Wu. 2013. Do scientists trace hot topics? *Scientific reports* 3, 1 (2013), 2207.
- [32] Lingfei Wu, Dashun Wang, and James A Evans. 2019. Large teams develop and small teams disrupt science and technology. *Nature* 566, 7744 (2019), 378–382.
- [33] Mengjia Wu, Gunnar Sivertsen, Lin Zhang, Fan Qi, and Yi Zhang. 2025. Scaling research aim identification: Language models for classifying scientific and societal-oriented studies. *Journal of the Association for Information Science and Technology* 76, 11 (2025), 1470–1487.
- [34] MJ Wu, Yi Zhang, Robin Haunschild, and Lutz Bornmann. 2025. Leveraging large language models for post-publication peer review: Potential and limitations. In *Proceedings of the 20th International Conference on Scientometrics & Informetrics (ISSI 2025)*. 1207–1226.
- [35] Xovee Xu, Ting Zhong, Ce Li, Goce Trajcevski, and Fan Zhou. 2022. Heterogeneous dynamical academic network for learning scientific impact propagation. *Knowledge-Based Systems* 238 (2022), 107839.
- [36] Carl Yang and Jiawei Han. 2023. Revisiting citation prediction with cluster-aware text-enhanced heterogeneous graph neural networks. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. IEEE, 682–695.
- [37] Xiaoyu Yang, Yufei Chen, Xiaodong Yue, Shaoxun Xu, and Chao Ma. 2023. T-distributed spherical feature representation for imbalanced classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 10825–10833.
- [38] Xiaoyu Yang, Jie Lu, and En Yu. 2025. Adapting Multi-modal Large Language Model to Concept Drift From Pre-training Onwards. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=b20VK2GnSs>
- [39] Haofei Yu, Zhaochen Hong, Zirui Cheng, Kunlun Zhu, Keyang Xuan, Jinwei Yao, Tao Feng, and Jiaxuan You. 2024. Researchtown: Simulator of human research community. *arXiv preprint arXiv:2412.17767* (2024).
- [40] Le Yu, Leilei Sun, Bowen Du, and Weifeng Lv. 2023. Towards better dynamic graph learning: New architecture and unified library. *Advances in Neural Information Processing Systems* 36 (2023), 67686–67700.
- [41] Penghai Zhao, Qinghua Xing, Kairan Dou, Jinyu Tian, Ying Tai, Jian Yang, Ming-Ming Cheng, and Xiang Li. 2025. From Words to Worth: Newborn Article Impact Prediction with LLM. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 1183–1191.

A Experiments Details

Point accuracy. Given n papers with ground-truth five-year citations y_i and predictions \hat{y}_i , we compute:

$$\text{MALE} = \frac{1}{n} \sum_{i=1}^n |\log(1 + \hat{y}_i) - \log(1 + y_i)|, \quad (11)$$

$$\text{RMSLE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\log(1 + \hat{y}_i) - \log(1 + y_i))^2}, \quad (12)$$

Ranking quality. For papers published in the same year, let π denote the predicted ranking and π^* the ground-truth ranking:

$$\text{NDCG@K} = \frac{\sum_{i=1}^K \frac{y_{\pi(i)}}{\log_2(i+1)}}{\sum_{i=1}^K \frac{y_{\pi^*(i)}}{\log_2(i+1)}}. \quad (13)$$

B Auxiliary Losses for Stage B

To further stabilize optimization and mitigate residual bias, we incorporate two lightweight auxiliary objectives at Stage B: **(i) exposure calibration loss** and **(ii) adversarial venue-invariance loss**. Both are designed to regularize the learned citation representations without introducing additional parameters or inference overhead.

B.1 Exposure Calibration Loss

Empirical studies have shown that citation counts are strongly correlated with *exposure factors*—such as publication venue, collaboration size, or open-source visibility—which may distort predictive learning. To prevent the model from over-amplifying these factors, we impose an auxiliary calibration constraint on the predicted exposure score \hat{E} :

$$\mathcal{L}_{\text{exp}} = \text{KL}(p(\hat{E}) \| p(E^*)), \quad (14)$$

where E^* denotes the empirical exposure distribution estimated from the training data. This term penalizes deviations between the predicted and empirical exposure distributions, ensuring that the model’s intermediate exposure estimation remains statistically consistent and well-calibrated.

B.2 Adversarial Venue-Invariance Loss

Venue prestige is one of the most dominant shortcut features in citation prediction. To enhance robustness against venue bias, we introduce an adversarial objective that enforces venue-invariant latent representations. A discriminator D_v is trained to predict the venue label from the Stage B feature embedding s_p , while the main encoder f_{θ} attempts to fool it:

$$\min_{\theta} \max_{\phi} \mathbb{E}_{p \sim \mathcal{D}} [\log D_{v_{\phi}}(v_p | f_{\theta}(s_p))]. \quad (15)$$

This min-max game discourages the encoder from encoding venue-specific artifacts, resulting in a fairer and more transferable citation representation.

B.3 Implementation Note

Both auxiliaries are assigned small weights ($\lambda_{\text{exp}} = 0.1$, $\lambda_{\text{adv}} = 0.05$) relative to the primary Stage B objective. They are only active during training and disabled during inference. Ablation results in Table 2 confirm that incorporating these auxiliaries improves fairness and stability, particularly under long-tailed or domain-shifted scenarios.

C Dataset Diagnostics and Descriptive Statistics

This appendix reports basic diagnostics of the evaluation splits used in our experiments, including overall scale and central tendency of citation counts, temporal coverage, authorship statistics, and

venue/document-type compositions. These summaries help contextualize the long-tailed nature of citations and the salience of venue as a shortcut factor discussed in the main text.

C.1 Global Characteristics

Table 3 summarizes dataset-level statistics. We report the number of papers, citation central tendencies (mean/median/max), year range, average number of authors, and the highest venue tier present in each split.

C.2 Venue Tier Distribution

Table 4 reports the venue-tier composition for each split. Percentages are computed over all papers in the split.

Table 3: Dataset characteristics across splits.

Dataset	Avg Cit.	Max Cit.	Avg Authors	Top Venue Tier
Aminer1	49.4	18,816	3.5	Tier 4
Aminer2	47.9	9,517	3.5	Tier 4
Aminer3	47.7	33,132	3.6	Tier 4
OpenAlex1	442.2	71,274	4.3	Tier 4
OpenAlex2	450.0	194,890	4.3	Tier 4
OpenAlex3	436.5	72,981	4.3	Tier 4

Table 4: Venue tier distribution by split.

Dataset	Tier 1	Tier 2	Tier 3	Tier 4	Tier 5
Aminer1	2 (0.0%)	813 (6.8%)	3190 (26.6%)	7698 (64.1%)	297 (2.5%)
Aminer2	8 (0.1%)	782 (6.5%)	3215 (26.8%)	7681 (64.0%)	311 (2.6%)
Aminer3	4 (0.0%)	823 (6.9%)	3152 (26.3%)	7736 (64.5%)	285 (2.4%)
OpenAlex1	2 (0.0%)	1094 (9.1%)	2528 (21.1%)	7602 (63.3%)	774 (6.5%)
OpenAlex2	1 (0.0%)	1060 (8.8%)	2541 (21.2%)	7652 (63.8%)	745 (6.2%)
OpenAlex3	2 (0.0%)	1080 (9.0%)	2524 (21.0%)	7605 (63.4%)	789 (6.6%)