Citation Prediction and Recommendation with Semantic and Graph-Based Learning

Shenghao Luo
University of Illinois - Urbana-Champaign, USA

Abstract: Citation prediction can be formulated as both a link prediction and recommendation task. Given a source paper A and a candidate reference paper B, the objective is to estimate the probability that A cites B. Unlike undirected link prediction, citation prediction is inherently directed because the relationship from A to B is not equivalent to that from B to A, and it is temporally constrained because a valid reference must be published before the source paper. This study begins with a text-based baseline using semantic document embeddings and logistic regression, then examines probability calibration, Top-*K* recommendation, embedding robustness, leakage-free graph message passing, directed message passing, candidate-reference graph features, and transfer experiments on the larger OGBL-citation2 benchmark. The results show that semantic text representations provide a strong foundation, while graph information offers complementary signals when incorporated without temporal or edge leakage. End-to-end graph representation learning is not consistently superior to a comparable neural text-only model, especially in cold-start settings where the source paper has limited citation history. In contrast, simple historical graph features derived only from the candidate reference paper consistently improve citation identification and ranking. Overall, combining semantic text representations with carefully controlled candidate-side graph features improved Cit-*HepT*h F1 by about 0.54% and OGBL-citation2 subset MRR by about 22.6% over the corresponding text-only baselines.

Keywords: Citation Prediction; Citation Recommendation; Graph Neural Networks; Semantic Embedding; Link Prediction; Graph Leakage

1. Introduction

Citation prediction is important because citations are one of the main mechanisms through which scientific knowledge is connected, organized, and discovered. When researchers write a new paper, they need to identify prior studies that provide background, support their arguments, introduce related methods, or serve as comparison points. However, the rapid growth of scientific publications makes it increasingly difficult to manually find all relevant references. Citation recommendation can reduce this burden by automatically suggesting candidate papers that are likely to be cited by a given source paper. Beyond recommendation, citation prediction is also useful for analyzing scientific influence, identifying missing citations, modeling knowledge flow, and detecting research communities or emerging topics. Early citation recommendation work, such as He et al. (2010), showed that citation recommendation can be formulated as a context-aware retrieval problem, while later neural approaches such as Ebesu and Fang (2017) demonstrated the value of learned representations for citation recommendation. Citation-informed representation learning, such as SPECTER proposed by Cohan et al. (2020), further shows that citation links can provide powerful supervision for scientific paper embeddings [2].

Traditional citation prediction methods include text-similarity models, supervised machine learning classifiers, and classical graph-based algorithms. Text-based methods represent papers using titles, abstracts, or local citation contexts, and then recommend papers based on semantic relevance [1,6](Bhagavatula et al., 2018; He et al., 2010). These methods are effective because related papers often discuss similar topics, but they may miss structural citation signals such as paper influence, shared-reference patterns, and research-community membership. Traditional classifiers such as logistic regression or multilayer perceptrons can combine manually designed features, including cosine similarity, publication year, citation count, and embedding

Copyright @ STEMM Institute Press

difference, but they still depend on predefined features. Classical graph methods such as PageRank and link-prediction heuristics use citation structure, but they are usually fixed and indirect: they measure specific graph patterns rather than learning flexible representations from both text and graph structure [10,13](Liben-Nowell & Kleinberg, 2007; Page et al., 1999). More recent hybrid methods, such as BERT-GCN, combine language models with graph convolutional networks, suggesting that citation recommendation can benefit from both semantic text representations and graph structure [8](Jeong et al., 2020).

Graph Neural Networks provide a natural framework for citation prediction because a citation network is itself a graph: papers are nodes, and citations are directed edges [16]. The idea of neural networks on graphs can be traced back to early work by Gori et al. (2005), who proposed a neural model capable of directly processing graph domains, and Scarselli et al. (2009), who formalized the Graph Neural Network model for different types of graphs, including directed, undirected, cyclic, and acyclic graphs. Later models such as Graph Convolutional Networks, GraphSAGE, and Graph Attention Networks made GNNs more scalable and effective for representation learning on large graphs [5,9,17](Hamilton et al., 2017; Kipf & Welling, 2017; Veličković et al., 2018). Through message passing, a GNN can learn node representations that combine paper content with neighborhood information, allowing it to capture research communities, co-citation relationships, shared references, and citation influence. However, directly applying a standard GNN to citation prediction can introduce edge leakage if the same citation links are used both for message passing and as supervised prediction labels. To address this, this project uses a leakage-free design with three main modifications: first, message-passing edges are separated from supervised prediction edges; second, temporal validation and testing splits are used so future citation information is not available during training; third, directed message passing and target-only historical graph features are explored to better represent citation direction and candidate-paper influence.

2. Related work

2.1 Definition and Development of Graph Neural Networks

Graph Neural Networks (GNNs) are a class of neural network models designed to learn from graph-structured data. A graph is usually represented as $G=(V,E)$, where V is the set of nodes and E is the set of edges. Different from traditional neural networks that operate on independent data samples, GNNs explicitly model the relationships between objects. In a GNN, each node updates its representation by aggregating information from its neighboring nodes. Through this message-passing process, the final node embedding contains both the node's own features and information from its local graph structure. This makes GNNs especially suitable for citation networks, where papers are nodes and citation links are directed edges between papers.

The development of GNNs has made it possible to apply deep learning directly to relational data. Early studies introduced neural models for graph domains and formalized the general GNN framework[4,14] (Gori et al., 2005; Scarselli et al., 2008). Later models such as Graph Convolutional Networks, GraphSAGE, and Graph Attention Networks improved the scalability and flexibility of GNNs by introducing neighborhood aggregation, inductive learning, and attention-based message weighting [5,9,17](Hamilton et al., 2017; Kipf & Welling, 2016; Veličković et al., 2017). In practical applications, these models are used to generate node embeddings that combine original node features with graph neighborhood information. For citation prediction, this means a paper representation can include not only its title and abstract, but also information from related papers in the citation network. Therefore, GNNs are useful because they can learn research communities, co-citation patterns, shared-reference structure, and paper influence directly from the citation graph.

2.2 Related Work on Citation Prediction and Citation Recommendation

Early work on citation prediction and citation recommendation mainly focused on textual relevance, citation context, and information retrieval. These methods treated citation recommendation as the task of matching a citation context or a manuscript with relevant candidate papers. The general idea was that if the local writing context, title, abstract, or topic of a source paper is semantically related to a candidate paper, then the candidate paper is more likely to be cited. For example, He et al. (2010) formulated citation recommendation as a context-aware retrieval

 Copyright @ STEMM Institute Press

problem, where references are recommended based on the local text around the citation position. Tang et al. (2014) extended this direction to cross-language citation recommendation, showing that citation prediction may also require language alignment [16]. Later, Ebesu and Fang (2017) proposed the Neural Citation Network, which used neural representations, attention, and author information for context-aware citation recommendation [3]. Bhagavatula et al. (2018) focused on content-based citation recommendation, using paper text such as titles and abstracts to recommend relevant references. These studies show that textual information is highly useful for citation recommendation, but text-only methods may miss graph-structural signals such as paper influence, citation communities, and shared-reference patterns.

Recent work has moved toward stronger neural representation learning and hybrid text-graph modeling. Instead of relying only on handcrafted text features or local citation contexts, recent methods use deep language models, citation-informed embeddings, and graph neural networks to represent scientific papers and citation relationships. Cohan et al. (2020) proposed SPECTER, a citation-informed Transformer model that learns scientific paper embeddings from titles and abstracts using citation links as supervision. This is closely related to this project because the baseline uses SPECTER2 embeddings as the main semantic representation. Jeong et al. (2020) further combined BERT with Graph Convolutional Networks, using BERT for citation-context representation and GCN for graph information. This hybrid direction supports the motivation of this project: citation recommendation should use both semantic similarity and citation-network structure. However, this work focuses more specifically on leakage-free graph evaluation, directed message passing, and target-only historical graph features, because directly using citation edges in a GNN can accidentally expose the links that the model is supposed to predict.

3. Methodology

This project follows a complete pipeline from problem definition and text-based modeling to graph-based improvement and external dataset transfer, as shown in Figure 1. The overall goal is to predict whether a source paper A will cite a candidate paper B, and then use this prediction score to recommend possible references. The main dataset is the Cit-HepTh citation network, which contains high-energy theory papers and directed citation links. To further test generalization, the proposed method is also evaluated on ogbl-citation2, a large-scale citation recommendation benchmark from the Open Graph Benchmark [7](Hu et al., 2020). The methodology can be divided into three main parts: text-based citation prediction, leakage-free graph-based modeling, and transfer evaluation.

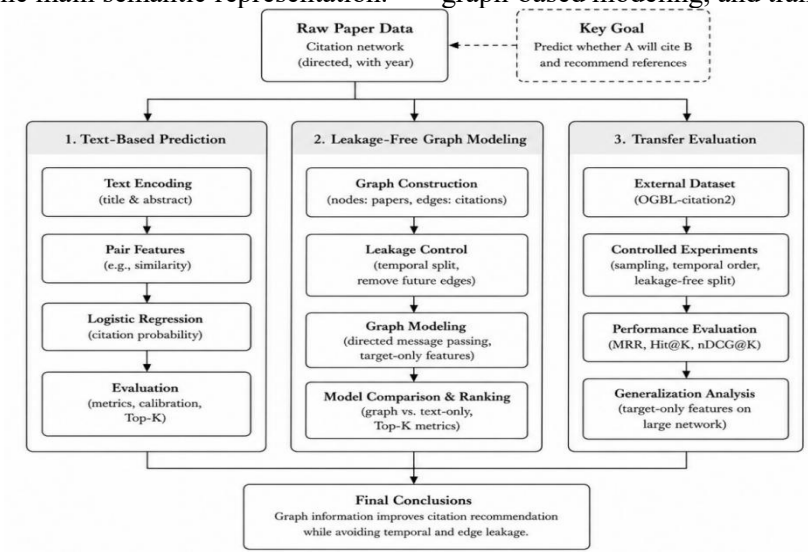


Figure 1. Overview of methodology

3.1 Text-Based Citation Prediction and Evaluation

The first part establishes the basic citation prediction pipeline. Citation prediction is formulated as a directed link-prediction problem, where each paper is treated as a node and each citation is treated as a directed edge from the citing paper to the cited paper. Because citation is temporal, the candidate reference paper must be

published no later than the source paper. This constraint prevents future information leakage and makes the task closer to a real recommendation scenario.

Each paper's title and abstract are encoded into a 768-dimensional SPECTER2 embedding. For each paper pair (A,B), three feature groups are constructed: the elementwise product of the two embeddings, the absolute difference between them, and their cosine similarity. These features are used to train a logistic regression baseline. This baseline is important because it measures how much citation prediction can be solved using semantic similarity alone.

The model is evaluated in two ways. First, balanced pair classification is used, where the model predicts whether a given paper pair has a citation link. Second, Top- K citation recommendation is used, where the model ranks many candidate papers for one source paper. The second setting is harder and more realistic because only a very small percentage of candidate papers are true references.

Score calibration is also applied so that the predicted probabilities better reflect real citation likelihood. In a balanced training set, positive and negative citation pairs are sampled in similar numbers, but in a real recommendation setting, true citations are much rarer than non-citations. Therefore, the raw model scores may be useful for ranking but may not represent reliable probabilities. Score calibration adjusts these raw scores on a validation set so that, for example, pairs predicted with probability 0.8 are actually cited about 80% of the time. In addition, embedding robustness is tested by duplicating a sentence, adding an irrelevant sentence, comparing with an unrelated paper, and adding Gaussian noise to embeddings. These experiments evaluate whether small text or embedding changes strongly affect the prediction and ranking outputs.

3.2 Leakage-Free Graph-Based Modeling

The second part introduces graph-based models to use citation-network structure. GraphSAGE is used to learn paper representations from both text embeddings and citation neighborhoods. In this setting, papers exchange information with neighboring papers through message passing. This allows the model to capture research communities, shared-reference patterns, co-citation relationships, and paper influence.

However, directly applying a GNN to citation prediction can create an edge leakage problem. If the citation edges used as supervised prediction labels are also present in the message-passing graph, the model may indirectly observe the answers during training. To avoid this problem, the graph experiment is redesigned with stricter controls. Message-passing edges are separated from supervised prediction edges, and temporal validation and test splits are used. In this way, the GNN can use historical graph structure without seeing the exact citation links it is asked to predict.

Two graph-based improvements are then explored. First, directed message passing is tested because incoming citations and outgoing references have different meanings. Incoming citations can reflect influence or authority, while outgoing citations describe the references used by a paper. Second, target-only historical graph features are added, including historical log in-degree, historical log out-degree, in contrast, and out contrast. These features are attached only to the candidate paper B, not the source paper A, because in real citation recommendation the source paper may be new and may not have stable citation history. This design allows graph information to be used in a realistic and leakage-free way.

3.3 Transfer Evaluation on OGBL-Citation2

The third part evaluates whether the proposed graph-feature idea transfers beyond the original Cit-HepTh corpus. For this purpose, the method is tested on ogbl-citation2, a large, directed citation network from the Open Graph Benchmark [7](Hu et al., 2020). In this dataset, each node is a paper with a 128-dimensional word2vec feature vector, and each directed edge represents a citation. The official task is to rank a true missing reference above negative candidate references, which closely matches the citation recommendation goal of this project.

Because the full ogbl-citation2 graph is very large, controlled subset experiments are used first. Text-only MLP, GraphSAGE, and text plus target graph feature models are compared under the same subset setting. These experiments test whether historical candidate-paper graph features can still improve ranking performance on a much larger citation network. Overall, the methodology demonstrates a progression from semantic text modeling to carefully controlled graph-based citation prediction, showing how graph information can improve recommendation while avoiding temporal and edge leakage.

 Copyright @ STEMM Institute Press

4. Experiment

The experiments are designed to evaluate whether citation prediction can be improved by combining semantic text representations with graph-based information. The experimental process first tests a text-based baseline, then evaluates calibration, Top- K recommendation, and embedding robustness. After that, graph-based models are introduced under a leakage-free protocol to test whether citation-network structure provides additional predictive value. Finally, external experiments on ogbl-citation2 and a RefSeer-style citation context dataset are used to examine whether the proposed graph and influence features generalize beyond the main Cit-HepTh dataset.

4.1 Dataset Introduction

This project uses three citation-related datasets: Cit-HepTh, ogbl-citation2, and a RefSeer-style citation context dataset. Cit-HepTh is the main experimental dataset. It is a high-energy theory citation network containing 29,555 papers published between 1992 and 2003. Each paper is treated as a node, and each citation is treated as a directed edge from the citing paper to the cited paper. The dataset contains paper text, publication year, and citation relationships, so it supports both text-based citation prediction and graph-based citation recommendation.

Cit-HepTh is split temporally to simulate a realistic recommendation setting. Papers from 1992 to 2000 are used for training, papers from 2001 are used for validation and score calibration, and papers from 2002 to 2003 are used for final testing. This temporal split is important because citation prediction should not use future citation information to predict earlier citation links.

The project also uses ogbl-citation2 as a large-scale external benchmark. It is a directed citation graph from the Open Graph Benchmark, containing nearly three million papers and over thirty million citation edges [11]. Each node has a 128-dimensional word2vec feature vector and publication year information. Its official task is to rank a true missing reference above negative candidate references, which closely matches the citation recommendation goal of this project. Therefore, ogbl-citation2 is used to test whether the graph-feature method developed on Cit-HepTh can transfer to a larger citation network.

Finally, a RefSeer-style citation context dataset, stored locally as arxiv_data.csv, is used for the comparison experiment in Section 4.3.2. This dataset contains citation contexts, citing authors, cited paper titles, and cited authors. Unlike Cit-HepTh and ogbl-citation2, it is mainly used for context-aware citation recommendation: given a local citation context, the model ranks the true cited title against sampled negative candidate titles. This dataset helps compare the project with previous citation recommendation work that used RefSeer-style citation context data.

4.2 Experimental Setup

All experiments were implemented in Python. The main machine learning libraries include scikit-learn for the logistic regression baseline and score calibration, PyTorch for neural network training, and PyTorch Geometric for GraphSAGE-based graph neural network models. The paper embeddings were generated with SPECTER2, where each paper title and abstract were encoded into a 768-dimensional vector. For the external benchmark, ogbl-citation2 was loaded with the Open Graph Benchmark package [12]. The experiments were run locally on a personal computer environment, using a Python virtual environment to manage dependencies. Since ogbl-citation2 is much larger than the original Cit-HepTh dataset, the OGB experiments were conducted on sampled subsets rather than the full benchmark.

For the baseline model, each paper pair was represented using three feature groups: the elementwise product of the source and target embeddings, the absolute difference between the two embeddings, and their cosine similarity. These produced a 1537-dimensional pair feature vector. A Standard scaler was used to normalize features, followed by L2-regularized logistic regression trained with SGD Classifier. The model was trained for five epochs with a batch size of 8,192. The training set used citation links from 1992–2000, the validation set used 2001, and the final test set used 2002–2003.

For the graph-based experiments, each paper was treated as a node and each citation was treated as a directed edge. The GraphSAGE encoder used two layers with dimensions $(768 \rightarrow 128 \rightarrow 64)$. A residual projection from the original text embedding to 64 dimensions was added to preserve textual information. The pair decoder concatenated the source representation, target representation, elementwise product, and absolute difference, then predicted one citation score with an MLP. To avoid leakage, message-

Copyright @ STEMM Institute Press

passing edges and supervised prediction edges were separated, and validation/test evaluation used only historical graph information available before the prediction period. Additional graph variants included directed GraphSAGE and target-only historical graph features.

Model performance was evaluated from both classification and recommendation perspectives. For pair classification, Accuracy, Precision, Recall, F1, ROC-AUC, and Average Precision were used. Accuracy measures overall correctness, Precision measures the proportion of predicted citation links that are correct, Recall measures the proportion of true citation links that are found, and F1 is computed as

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (1)$$

ROC-AUC and Average Precision evaluate how well the model ranks positive citation pairs above negative pairs across different thresholds. For citation recommendation, Top-K ranking metrics were used, including Precision@K, Recall@K, HitRate@K, MRR, and nDCG@K. Precision@K is calculated as

$$Precision@K = \frac{\#\ relevant\ papers\ in\ Top{-}K}{K} \quad (2)$$

while Recall@K is calculated as

$$Recall@K = \frac{\#\ relevant\ papers\ in\ Top{-}K}{\#\ all\ relevant\ papers} \quad (3)$$

MRR is the average reciprocal rank of the first correct citation,

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i} \quad (4)$$

where $rank_i$ is the position of the first relevant citation for query i. nDCG@K gives higher reward when relevant citations appear earlier in the ranked list. These ranking metrics are important because real citation recommendation requires selecting a few useful papers from a large candidate pool.

Score calibration was also evaluated because models trained on balanced positive and negative pairs may produce probabilities that are too high for real recommendation settings, where true citations are rare. Platt scaling was applied on the validation year to recalibrate prediction scores. Calibration quality was measured using Expected Calibration Error and Brier score, where lower values indicate more reliable probability estimates.

4.3.1 Experiment results

In this project, I compare four representative models: the text-based logistic regression baseline, leakage-free GraphSAGE, directed GraphSAGE, and the final text + target-only graph feature model. These four models show the development of the project from a simple non-GNN baseline to graph-based methods. Since results from other citation prediction papers are often reported on different datasets and different splits, it is not fair to directly compare their exact scores with mine. Therefore, this section mainly compares my models under the same Cit-HepTh temporal split, while discussing how they relate to previous citation prediction methods, as shown in Table 1.

Table1. Pair-classification Performance of Baseline and Graph-based Models on the Cit-HepTh Test Set

Model	Main idea	Graph information	Test ROC-AUC	Test AP	Test F1
Logistic regression baseline	SPECTER2 pair features	No	0.8947	0.8977	0.8125
Leakage-free GraphSAGE	Text embedding + citation neighbors	Yes	0.9124	0.9084	0.8378
Directed GraphSAGE	Separate incoming/outgoing citation messages	Yes	0.9226	0.9189	0.8477
Text + target-only graph features	Text + candidate paper history	Yes	**0.9395**	**0.9358**	**0.8701**

The results show a clear improvement trend in Figure 2. The logistic regression baseline already performs well because SPECTER2 embeddings contain strong semantic information about papers. This is similar to content-based citation recommendation methods in previous work, where title and abstract representations are used to measure whether two papers are related. However, the baseline does not directly use the citation network, so it cannot capture structural signals such as paper influence, citation communities, or shared-reference patterns.

The leakage-free GraphSAGE model improves over logistic regression, which shows that citation graph structure is useful when it is used correctly. The key point is “leakage-free”: message-passing edges are separated from supervised prediction edges, so the model cannot see the exact citation links it is asked to predict. This makes the result more reliable than a direct GNN setup.

Directed GraphSAGE performs better than ordinary leakage-free GraphSAGE. This result is

 Copyright @ STEMM Institute Press

important because citation graphs are directed. Incoming citations and outgoing references have different meanings. Incoming citations can represent a paper's influence, while outgoing references describe the background papers it uses. By separating these two directions, directed GraphSAGE better matches the nature of citation networks.

The best result comes from the text + target-only graph feature model. Instead of using full GNN message passing, this model adds historical graph features only to the candidate paper, including historical log in-degree, historical log out-degree, and in/out contrast. This design works well because in real citation recommendation, the source paper may be new and have no citation history, while candidate reference papers are usually older and already have historical citation information. Compared with the matched text-only neural model, this model improves ROC-AUC, AP, and F1, showing that graph information adds value beyond text similarity.

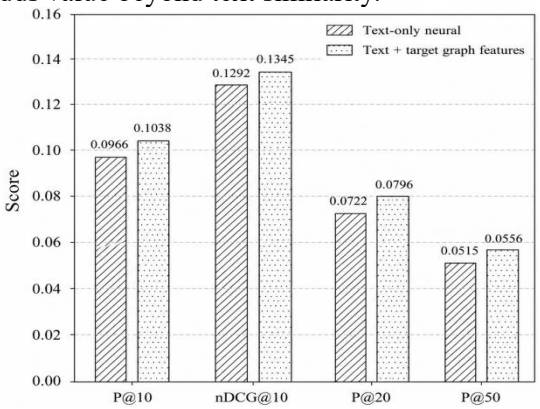


Figure 2. Top-K ranking Performance on Cit-HepTh Test Set

For ranking evaluation, the result is more nuanced, as shown in Table 2:

Table 2. Full-candidate Top-*K* Ranking Performance on the Cit-HepTh Test Set

Ranking model	MRR	Precision@10	Recall@10	nDCG@10	Precision@20	Precision@50
Text-only neural	**0.2992**	0.0966	0.0772	0.1292	0.0722	0.0515
Text + target graph features	0.2912	**0.1038**	**0.0772**	**0.1345**	**0.0796**	**0.0556**

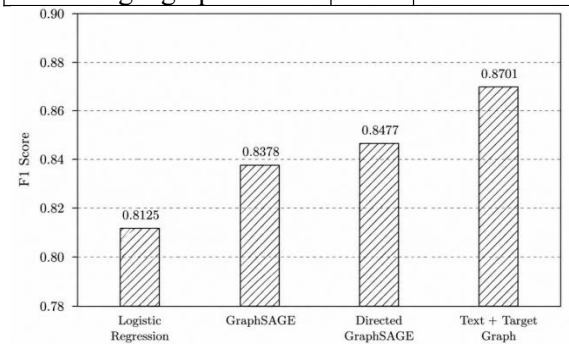


Figure 3. F1 Score Comparison of Different Models on the Cit-HepTh Test Set.

This ranking result means that target graph features help place more true citations into the recommendation list, especially at Top-20 and Top-50(see Figure 3). However, MRR does not improve, which means the model does not always move the first correct citation earlier. Therefore, the graph feature model is useful for increasing the overall quality of the recommendation list, but future work should use ranking-specific training to improve the very top result.

Overall, compared with previous non-GNN citation prediction methods, my work shows that text embeddings are a strong foundation, but graph information provides extra value. Compared with direct GNN methods, my work also shows that graph models must be designed carefully to avoid leakage. The strongest conclusion is that candidate-side historical graph features are simple, interpretable, and effective for citation recommendation.

4.3.2 Comparison with related work and same-dataset experiments

This section compares the experimental results with representative citation recommendation and graph-based citation prediction studies(see Table 3). Since different papers use different datasets, candidate pools, node features, and evaluation protocols, their scores cannot be treated as a direct leaderboard comparison. Therefore, the comparison is divided into two levels. First, related work is used to position this project within the broader citation prediction literature. Second, same-dataset experiments are used to show whether the proposed influence-based improvement is effective under a controlled experimental setting.

Previous citation recommendation studies have used different strategies. He et al. introduced context-aware citation recommendation, where citations are recommended based on the local citation context around the citation position. Ebesu and Fang proposed the Neural Citation Network, using a neural encoder-decoder model with attention and author information on the RefSeer citation context dataset. In addition, the Open Graph Benchmark provides large-scale citation link prediction results on ogbl-citation2; the official GraphSAGE benchmark reports test MRR 0.8260. These works show that citation prediction can benefit from context modeling,

Copyright @ STEMM Institute Press

neural representation learning, and graph structure. However, their reported scores are not directly comparable to this project because they are evaluated under different settings. For this reason, the main comparison in this section focuses on experiments conducted in the same local environment.

Table 3. Comparison with Related Work and Local Experiments

Work / model	Dataset	Main method	Reported result
He et al.	Citation context data	Context-aware probabilistic recommendation	Evaluated with citation-context retrieval metrics
Ebesu & Fang	RefSeer	Neural Citation Network with attention and author information	Improved Recall, MAP, MRR, and nDCG over baselines
OGB official GraphSAGE	ogbl-citation2	Full-batch GraphSAGE	Test MRR 0.8260
My TF-IDF text baseline	RefSeer-style arxiv_data.csv	Citation context and cited-title text similarity	MRR 0.5340, Hits@1 0.4080, Hits@10 0.7703
My influence-only model	RefSeer-style arxiv_data.csv	Historical cited-title frequency	MRR 0.3114, Hits@1 0.2060, Hits@10 0.5057
My text + influence model	RefSeer-style arxiv_data.csv	Text similarity + historical target influence	MRR 0.5681, Hits@1 0.4697, Hits@10 0.7527
My text-only MLP	OGB subset	Text/node features only	MRR 0.243
My GraphSAGE	OGB subset	Graph neighborhood aggregation	MRR 0.270
My text + target graph features	OGB subset	Text/node features + target graph history	MRR 0.298

The RefSeer-style experiment provides a same-dataset comparison closer to context-aware citation recommendation. In this experiment, the task is to rank the true cited title above sampled negative cited titles for a given citation context. The first 70% of rows are used as pseudo-history for fitting the TF-IDF representation and computing historical influence, while the last 15% are used for testing. Each test query contains one true cited title and 100 sampled negative candidate titles, and the experiment is repeated over three random seeds.

Models are evaluated using MRR and Hits@1. The text + influence model achieves the best performance on both metrics, showing that historical target influence helps move the correct citation closer to the top of the recommendation list.

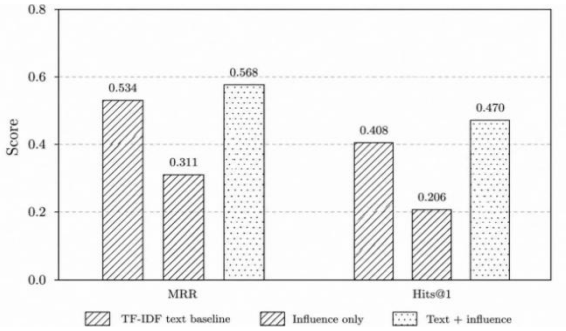


Figure 4. Citation Recommendation Performance on the RefSeer-Style Dataset under the Same-Dataset Setting

The results show that citation context text is already a strong signal, as shown in Figure 4. The TF-IDF text baseline achieves MRR 0.5340 and Hits@1 0.4080, which means that local citation context can often identify the correct cited paper title. However, adding historical target influence further improves the most important ranking metrics. The text + influence model improves MRR from 0.5340 to 0.5681, an absolute increase of 0.0341 and a relative improvement of about 6.4%. Hits@1 improves from 0.4080 to 0.4697, a relative improvement of about 15.1%. This indicates that influence information helps the model place the correct citation at the first position more often.

The influence-only model performs worse than the text-based models, with MRR 0.3114 and Hits@1 0.2060. This result is important because it shows that historical citation frequency alone is not enough. A frequently cited paper is not necessarily the correct citation for a specific context. However, when influence is combined with text similarity, it becomes useful. Text provides semantic relevance, while influence helps distinguish among semantically plausible candidates by favoring historically important cited papers.

The OGB subset experiment provides additional evidence from a graph-based citation recommendation setting. On the same sampled ogbl-citation2 subset, the text-only MLP achieves MRR 0.243. GraphSAGE improves MRR to

 Copyright @ STEMM Institute Press

0.270, showing that graph neighborhood information adds value beyond node features alone. The best result is achieved by the text + target graph feature model, which reaches MRR 0.298. Compared with the text-only MLP, this is an absolute improvement of 0.055 and a relative improvement of about 22.6%. Hits@10 also improves from 0.447 to 0.533, a relative improvement of about 19.2%.

Overall, the comparison supports the main conclusion of the project. Related work shows that citation recommendation benefits from neural models and graph structure, while the same-dataset experiments show that target-side influence features provide clear improvements over text-only baselines. On the RefSeer-style citation context dataset, influence features improve MRR and Hits@1. On the OGB subset, target graph features improve MRR and Hits@10. These results suggest that graph or influence information is most useful when it is introduced in a controlled, interpretable, and leakage-aware way.

4.4 Effect of Adding Influence Features to the GNN-Based Model

In the original graph-based model, citation prediction mainly depended on paper text embeddings and neighborhood aggregation from the citation graph. This allows the model to learn whether two papers are semantically and structurally related. However, citation behavior is not determined only by topic similarity. Some papers are more likely to be cited because they are influential, widely recognized, or frequently used as background references. Therefore, I added influence-related graph features to the candidate paper, including historical log in-degree, historical log out-degree, and in/out contrast.

The most important influence feature is historical log in-degree, which measures how many citations a candidate paper has already received before the prediction time. A higher in-degree usually means the paper has stronger historical influence or authority. Historical log out-degree measures how many references the candidate paper made in the past. This can reflect the paper's citation behavior or how broadly it connects to previous work. In/out contrast measures the difference between citation received and references made, helping the model distinguish papers that are mainly influential from papers that are mainly citation heavy.

To test whether these influence features are useful, I compared the model without target graph features and the model with target-only graph features under the same experimental setting. The comparison is shown in Table 4.

Table 4. Effect of Adding Target-side Influence Features on Cit-HepTh Pair Classification

Model	Added influence features	Test ROC-AUC	Test AP	Test F1
Matched text-only neural model	No	0.9348	0.9296	0.8654
Text + target-only graph features	Yes	**0.9395**	**0.9358**	**0.8701**

The results show that adding influence features improves all three pair-classification metrics. ROC-AUC increases from 0.9348 to 0.9395, AP increases from 0.9296 to 0.9358, and F1 increases from 0.8654 to 0.8701. This means the model becomes better at ranking true citation links above non-citation links and also improves its final binary prediction quality.

I also tested which influence feature contributes most by running an ablation experiment, as shown in Table 5.

Table 5. Ablation Study of Target-Side Graph Influence Features on Cit-HepTh

Ablation model	Target graph feature(s)	Test ROC-AUC	Test AP	Test F1
Text-only neural	none	0.9348	0.9296	0.8654
+ target in-degree	historical log citations received	0.9347	0.9296	0.8647
+ target out-degree	historical log references made	0.9320	0.9260	0.8618
+ target in/out contrast	log in-degree − log out-degree	0.9348	0.9298	0.8651
+ all target graph features	all three features	**0.9395**	**0.9358**	**0.8700**

This ablation shows that the improvement is not caused by one single feature alone. In-degree alone is almost the same as the text-only model, out-degree alone is weaker, and in/out contrast alone is also close to neutral. The best result appears when all three features are used together. This suggests that the model benefits from the full historical citation profile of the candidate paper, not just simple popularity.

The reason these features help is that citation recommendation has both semantic and social-structural components. Text embeddings tell the

model whether two papers are about similar topics, but influence features tell the model whether the candidate paper has a citation history that makes it likely to be cited. For example, two candidate papers may be semantically similar to the source paper, but the one with stronger historical citation authority may be more likely to appear in the reference list. Similarly, out-degree and in/out contrast help the model understand whether a paper is a broad survey-like paper, a method paper, or an influential canonical work.

However, the ranking-level result is more mixed. In full-candidate recommendation, target graph features improve Precision@10, nDCG@10, Precision@20, and Precision@50, but MRR slightly decreases, as shown in Table 6.

Table 6. Effect of Target-side Graph Features on Full-candidate Citation Recommendation

Ranking model	MRR	Precision@10	Recall@10	nDCG@10	Precision@20	Precision@50
Text-only neural	**0.2992**	0.0966	0.0772	0.1292	0.0722	0.0515
Text + target graph features	0.2912	**0.1038**	**0.0772**	**0.1345**	**0.0796**	**0.0556**

This means the influence features help add more relevant papers into the recommendation list, especially at Top-20 and Top-50, but they do not always place the first correct citation earlier. Therefore, the result is positive but nuanced. The influence features improve the overall quality and density of the recommendation list, but future work may need ranking-specific training to optimize the very top positions.

Overall, adding target-side influence features improves the model because citation is not only a semantic relationship but also a graph-structural relationship. Papers are cited not only because they are textually similar, but also because they are influential, authoritative, or commonly used by a research community. The experiment shows that these features are useful, especially when combined together and computed only from historical graph information. This makes them effective, interpretable, and safe from temporal leakage.

5. Conclusion

This project studies citation prediction as both directed link prediction and citation recommendation. A strong text-based baseline is first built with SPECTER2 embeddings and logistic regression, then extended with ranking evaluation, score calibration, embedding robustness tests, GraphSAGE-based models, directed message passing, target-only influence features, and transfer experiments on ogbl-citation2. The results show that semantic text embeddings are already effective, but graph information can further improve citation prediction when used carefully. Separating message-passing edges from supervised prediction edges avoids edge leakage, while temporal validation and testing prevent future information leakage. The strongest improvement comes from target-only historical influence features, including historical log in-degree, historical log out-degree, and in/out contrast. This suggests that citation behavior depends not only on semantic similarity, but also on the candidate paper's historical influence and citation profile.

Future work can improve the system in several directions. Ranking-specific loss functions should be explored because the final goal is recommendation rather than only binary classification. Larger-scale neighbor sampling can support experiments on the full OGB benchmark. Stronger directed GNN architectures may better distinguish incoming citations from outgoing references. More advanced calibration methods can also make predicted citation probabilities more reliable in real settings, where true citation links are rare. Overall, the project shows that citation prediction benefits from combining semantic text representations with leakage-free and interpretable graph information.

References

[1] Bhagavatula, C., Feldman, S., Power, R., & Ammar, W. (2018). Content-based citation recommendation. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 238–251. https://doi.org/10.18653/v1/N18-1022

[2] Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. (2020). SPECTER: Document-level representation learning using citation-informed transformers. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2270–2282. https://doi.org/10.18653/v1/2020.acl-main.207

[3] Ebesu, T., & Fang, Y. (2017). Neural citation network for context-aware citation recommendation. *Proceedings of the 40th International ACM SIGIR Conference on*

 Copyright @ STEMM Institute Press

Research and Development in Information Retrieval, 1093–1096. https://doi.org/10.1145/3077136.3080730

[4] Gori, M., Monfardini, G., & Scarselli, F. (2005). A new model for learning in graph domains. *Proceedings of the IEEE International Joint Conference on Neural Networks*, 729–734. https://doi.org/10.1109/IJCNN.2005.1555942

[5] Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems, 30*, 1024–1034. https://papers.nips.cc/paper_files/paper/2017/hash/5dd9db5e033da9c6fb5ba837a7ebea9-Abstract.html

[6] He, Q., Pei, J., Kifer, D., Mitra, P., & Giles, C. L. (2010). Context-aware citation recommendation. *Proceedings of the 19th International Conference on World Wide Web*, 421–430. https://doi.org/10.1145/1772690.1772734

[7] Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., & Leskovec, J. (2020). Open Graph Benchmark: Datasets for machine learning on graphs. *Advances in Neural Information Processing Systems, 33*, 22118–22133.

[8] Jeong, C., Jang, S., Park, E., & Choi, S. (2020). A context-aware citation recommendation model with BERT and graph convolutional networks. *Scientometrics, 124*(3), 1907–1922. https://doi.org/10.1007/s11192-020-03561-y

[9] Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*. https://arxiv.org/abs/1609.02907

[10] Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology, 58*(7), 1019–1031. https://doi.org/10.1002/asi.20591

[11] Open Graph Benchmark. (n.d.). *Leaderboards for link property prediction: ogbl-citation2*. Retrieved August 12, 2026, from https://ogb.stanford.edu/docs/leader_linkprop/

[12] Open Graph Benchmark. (n.d.). *Link property prediction: ogbl-citation2*. Retrieved August 12, 2026, from https://ogb.stanford.edu/docs/linkprop/#ogbl-citation2

[13] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab. http://ilpubs.stanford.edu:8090/422/

[14] Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). The graph neural network model. *IEEE Transactions on Neural Networks, 20*(1), 61–80. https://doi.org/10.1109/TNN.2008.2005605

[15] Stanford Network Analysis Project. (n.d.). *High-energy physics theory citation network*. SNAP Datasets. Retrieved August 12, 2026, from https://snap.stanford.edu/data/cit-HepTh.html

[16] Tang, X., Wan, X., & Zhang, X. (2014). Cross-language context-aware citation recommendation in scientific articles. *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, 817–826.

[17] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. *International Conference on Learning Representations*. https://arxiv.org/abs/1710.10903

Copyright @ STEMM Institute Press