Aspect-Aware Content-Based Recommendations for Mathematical Research Papers

Ankit Satpute
Ankit.Satpute@fiz-karlsruhe.de
FIZ Karlsruhe
Berlin, Germany

André Greiner-Petter
Greiner-Petter@gipplab.org
University of Göttingen
Göttingen, Germany

Noah Gießing
Noah.Giessing@fiz-karlsruhe.de
FIZ Karlsruhe
Berlin, Germany

Olaf Teschke
Olaf.Teschke@fiz-karlsruhe.de
FIZ Karlsruhe
Berlin, Germany

Moritz Schubotz
Moritz.Schubotz@fiz-karlsruhe.de
FIZ Karlsruhe
Berlin, Germany

Akiko Aizawa
aizawa@nii.ac.jp
National Institute of Informatics
Tokyo, Japan

Bela Gipp
gipp@uni-goettingen.de
University of Göttingen
Göttingen, Germany

Abstract

Content-based research paper recommendation (CbRPR) has seen advances in domains like computer science and biomedicine, but remains unexplored for mathematics, a field where paper relatedness is more conceptual than explicit textual or citation-based similarity. In mathematics, papers may be connected through shared proof techniques, logical implications, or natural generalizations, yet exhibit minimal textual or citation overlap, rendering standard embedding or citation-based CbRPR ineffective. To address this gap, we first conduct an expert-driven study characterizing mathematical recommendations, revealing that relevance is inherently *aspect*-driven. Grounded in this insight, we introduce GoldRiM (small and expert-annotated) and SilverRiM (large and automatically derived), the first datasets for *aspect*-aware CbRPR in mathematics. Recognizing that LLM embeddings of mathematical content alone yields suboptimal representation, we propose AchGNN, an *aspect*-conditioned heterogeneous GNN that jointly models textual semantics, citation structure, and author identity. Across GoldRiM and SilverRiM, AchGNN consistently outperforms prior *aspect*-based CbRPR methods, achieving substantial gains across all evaluated *aspects*. We conduct ablation studies to analyze the contribution of individual *aspect* supervisions, authorship lineage, and graph structural signals to AchGNN's performance. To assess domain generality, we further evaluate AchGNN on the *Papers with Code* dataset of machine learning publications, demonstrating that our *aspect*-aware approach effectively transfers beyond mathematics. We deploy our system on the MaRDI platform to help mathematicians with recommendations and release datasets and code publicly for reproducibility: github.com/gipplab/MathAspectRecSys.

CCS Concepts

• **Information systems** → **Recommender systems**; Network data models; • **Computing methodologies** → *Neural networks*.

Keywords

Math similarity, Recommender System, Content Similarity

ACM Reference Format:
Ankit Satpute, André Greiner-Petter, Noah Gießing, Olaf Teschke, Moritz Schubotz, Akiko Aizawa, and Bela Gipp. 2026. Aspect-Aware Content-Based Recommendations for Mathematical Research Papers. In *Proceedings of (SIGIR'26).* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Content-based Research Paper Recommendation (CbRPR) systems suggests scholarly work that are similar in content. While such systems have seen progress in computer science (CS) [2, 25] and biomedicine (BM) [14, 20], they remain unexplored for mathematics. The gap is non-trivial as mathematical research differs fundamentally from CS and BM. First, unlike experimental sciences, where overlap in terminologies, datasets, or methods is common, mathematical papers often connect through abstraction, symbolic compression, and deductive reasoning [22]. Even citations in mathematics reference foundational concepts or proof techniques, rather than closely related prior work in the empirical sense [17]. Second, candidates with high textual similarity (basis effectively used in existing CbRPR systems [25, 36]) could be insufficient for capturing mathematical domain-traits. This is because two papers may be strongly related through, for instance, a dual formulation of a theorem or an extension of a proof strategy, while exhibiting minimal

SIGIR'26, Melbourne | Naarm, Australia

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

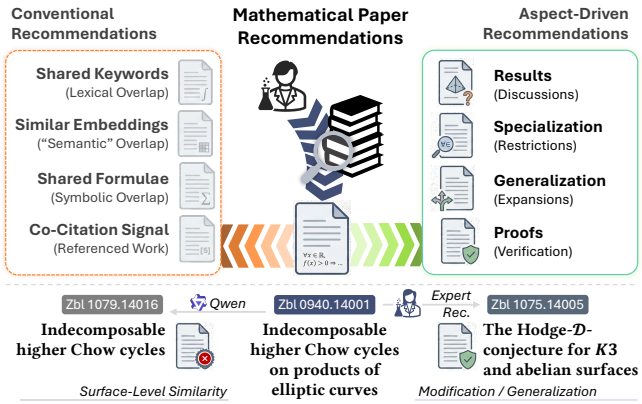

Figure 1: For a seed paper (center: Zbl 0940.14001), a CbRPR system (Qwen3-7B [53]) retrieves a surface-level similar paper (left: Zbl 1079.14016), whereas an expert mathematician recommends a paper (right: Zbl 1075.14005) that constitutes an important modification/generalization of the seed, despite minimal textual, semantic, or citation overlap.

textual and citation overlap. Figure 1 demonstrates broader patterns: in mathematics, relevance is often determined by the domain-specific roles (e.g., generalization, restriction, or proof reuse), rather than by observable similarity signals. We later show in Section 3.1 that there is in fact no clear correlation between relevance and similarity (lexical or semantical via embeddings) in mathematics. Such domain-specific challenges, the scarcity of available datasets, and the lack of empirical evidence characterizing CbRPR in mathematics motivate our first research question (RQ1): **What constitutes content-based relevance in mathematical research papers?**

To address RQ1, we conducted the first expert-driven study of mathematical recommendations. Our analysis in this work reveals that mathematical relevance is inherently *aspect*-driven, encompassing conceptual, methodological, or structural connections (e.g., generalization of theorem, dual formulation, shared proof technique). Crucially, these math-specific *aspect* categories are absent from existing CbRPR datasets[1]. To empirically validate and operationalize this insight, we introduce two novel resources: (1) **GoldRiM**, a small high-quality test-dataset curated with expert recommendations; and (2) **SilverRiM**, a large-scale dataset derived from implicit recommendation signals in zbMATH Open[2] abstracts. GoldRiM uncovers the structure of mathematical relevance, while SilverRiM provides the first large-scale benchmark for domain-specific CbRPR in mathematics, with *aspects* indicating *why* papers are related in both datasets.

While leading CbRPR approaches rely on Large Language Model (LLM) embedding similarities [29, 36, 54], mathematical texts' can cause LLM embeddings to underperform [13, 41], suggesting that embeddings alone may be insufficient for mathematical CbRPR. In this work, we find through experiments that LLM embeddings indeed underperform on mathematical CbRPR datasets. A mathematical CbRPR should, therefore, prioritize domain-salient signals, such as keywords, venues, or classification codes. However, these signals are typically coarse-grained and often unavailable across databases, limiting their applicability beyond specifically curated resources. In contrast, authorship information and citations are typically universally available, providing a robust foundation. Authorships also serve as a proxy for intellectual lineage, as mathematicians tend to work within narrowly defined subfields, reuse proof styles and foundational results are popular with author names [17, 31]. Hence, mathematical CbRPR should jointly model semantic similarity through citations and authorship to allow a system to prioritize works that are both semantically related and embedded within the same mathematical lineage. This motivates our second research question (RQ2): **How does the joint modeling of textual and authorship similarity affect CbRPR in mathematics?**

To address RQ2, we require a modeling framework that can jointly captures semantic relatedness and intellectual lineage in mathematical research, which naturally calls for relational learning. We leverage prior evidence that jointly modeling author–paper citation graphs through a Graph Neural Network (GNN) is effective for CbRPR [11, 21, 50]. Since we find via RQ1 that mathematical recommendations are inherently *aspect*-driven, our model must further preserve *aspect* specificity rather than collapsing all relations into a single notion of similarity. To this end, we formalize paper–paper recommendation using an Aspect-conditioned Heterogeneous GNN (AchGNN). The model operates on a heterogeneous graph with paper and author nodes and integrates semantic textual similarity, *aspect* information, and authorship relations within a unified framework.

We evaluate AchGNN on GoldRiM and SilverRiM against several baselines, including fine-tuned LLM CbRPR [34], a heterogeneous GNN [51], and state-of-the-art LLM-embeddings [1, 53]. AchGNN establishes a new state of the art on both GoldRiM and SilverRiM, consistently outperforming all competing baselines on GoldRiM and SilverRiM. We further evaluate AchGNN on the Papers with Code (PwC) dataset [19] to assess its applicability beyond heavily mathematics-oriented benchmarks, where it demonstrates competitive performance. The recommendations generated by AchGNN have already been fully integrated into the Mathematical Research Data Initiative platform (MaRDI)[3] and are planned for incorporation into zbMATH Open in the near future. Lastly, all our annotated datasets, the source code, and additional materials are publicly available[1].

2 Related Work

Aspect-based CbRPR: Aspect-based CbRPR has been approached mainly through supervised classification and embedding-based retrieval [36, 54]. Supervised classification-based approaches [9, 23, 35] classify pairs of papers with respect to a predefined sets of *aspects* to produce recommendations. However, their quadratic computational complexity makes them impractical for large-scale

[1]We provide additional materials, including an overview table of existing CbRPR datasets, alternative representations of results, aspect definitions, and more plots also in our repository: github.com/gipplab/MathAspectRecSys
[2]https://zbmath.org/

[3]An example document with recommendations: https://portal.mardi4nfdi.de/wiki/Publication:3361952

datasets, and they have shown limited ability to distinguish *aspects* compared to embedding-based retrieval [34, 35]. Early TF-IDF embedding-based retrieval methods [6, 7] were outperformed [29, 34, 42] by specialized, fine-tuned models such as SciBERT [3] and SPECTER [10]. Although fine-tuned embeddings consistently outperform general-purpose models, prior work has not addressed mathematical CbRPR. The empirical success of existing *aspect*-based CbRPR methods has largely been demonstrated on datasets[1] drawn from CS and BM, which dominate current benchmarks and typically follow the Introduction–Method–Results– Discussion (IMRD) structure [20, 37]. As a result, these methods often base on LLM-derived document embeddings and are optimized for structurally regular documents. Their effectiveness in domains with substantially different writing conventions remains underexplored [42]. Mathematical text, in particular, is semantically dense, highly symbolic, and frequently lacks explicit structural markers [15, 26]. As a result of these domain-specific characteristics, recent LLM embeddings from Massive Text Embedding Benchmark (MTEB) [32] exhibit markedly lower performance gains in mathematics compared to other domains [41].

Heterogeneous GNNs in CbRPR: Heterogeneous GNNs have shown consistent improvements in CbRPR over collaborative filtering and standalone embedding approaches by jointly modeling multiple scholarly entities, such as papers, authors, and venues [12, 25, 46]. These methods construct heterogeneous graphs and propagate information over citation, authorship, and venue relations to learn paper representations for recommendation via embedding similarity. While effective at leveraging structural signals, these models treat research papers as monolithic entities and do not produce *aspects*-based recommendations. Recent surveys confirm that although heterogeneous GNNs are widely adopted for CbRPR, *aspect*-based CbRPR using GNNs have not been explored [2, 28, 36]. In contrast, *aspect*-aware GNNs have been extensively studied in e-commerce recommendations, where user–item interaction graphs are enriched with *aspect*-level information. In recent works, FigGNN [45] and MA-GNN [52] incorporate *aspect*-conditioned user–item interactions to learn *aspect*-specific embeddings that improve recommendation rankings.

3 Methodology

Motivated by the success in e-commerce settings, we address the lack of an analogous formulation for mathematical CbRPR by integrating *aspect*-aware modeling with authorship, where shared authorship is treated as a signal of conceptual lineage. Figure 2 provides an overview of our proposed approach for a CbRPR in mathematics. Our pipeline consists of two main stages: (1) grounding mathematical CbRPR, and (2) AchGNN: generating *aspect*-based recommendations using a Heterogeneous GNN. Because no dataset exists for mathematical CbRPR, the first stage focuses on analyzing how relevance is expressed in mathematical research and constructing datasets that operationalize this notion. In the second stage, we leverage the insight gained from high-quality recommendations to design AchGNN, which integrates citation graph with authorship information to model *aspect*-based mathematical recommendations.

3.1 Grounding Mathematical CbRPR

Existing *aspect*-based CbRPR datasets do not specifically contain mathematical research papers [2, 25, 36]. We therefore curate our own annotated dataset specifically for mathematical CbRPR. We choose zbMATH Open as the source database for our dataset as it generally excels in terms of coverage, metadata quality, availability of research content, and accessibility through public APIs, compared to arXiv[4], MathSciNet[5], Google Scholar, or JSTOR[6]. However, most importantly, zbMATH Open's reviewing approach provides a unique opportunity for curating expert-level recommendations at scale. zbMATH Open provides summaries (typically and hereafter referred to as *abstracts*) written by human expert reviewers. These abstracts offer extensive historical coverage, spanning from 1763 to the present and include many foundational works that are frequently cited. Those reviewers are domain experts, making their judgments a reliable source of relevance signals. This is critical because relevance in mathematical research is often implicit and conceptually complex, rendering expert judgment indispensable for establishing ground truth. Utilizing zbMATH Open's unique infrastructure, we study and collect recommendations through two complementary datasets. GoldRiM: emphasizes annotation quality and supports qualitative analysis of how mathematical recommendations are formed. SilverRiM: emphasizes scale, enabling large-scale modeling and realistic quantitative evaluation.

3.1.1 GoldRiM. The goal of GoldRiM is to explicitly characterize mathematical relevance as perceived by domain experts, providing a qualitative foundation for mathematical CbRPR. To this end, we collaborated with zbMATH Open to collect expert-provided recommendations, a common approach for constructing high-quality human-annotated datasets in CbRPR [2, 25, 36]. Due to time and budget constraints, annotations were obtained from a single senior expert and active reviewer. The reviewer had over 30 years of experience in curating mathematical literature, ensuring high-quality and internally consistent judgments. We selected 80 seed documents, all drawn from the expert's primary domain of expertise, Algebraic Geometry, to minimize domain mismatch and maximize annotation reliability. For each seed document, the expert provided multiple recommendations judged to be most relevant, yielding 420 seed–recommendation pairs (between three to 11 per seed). The expert was deliberately given only vague instructions to provide recommendations based solely on the content of the seed documents, enabling subsequent analysis of the expert's decision-making process. This collection constitutes a small but high-quality **Gold**-standard dataset of **R**ecommendations **i**n **M**athematics (**GoldRiM**). We acknowledge likely bias and limited scale of GoldRiM, but it is designed for qualitative and diagnostic analysis only. Prior work has shown that even small, expert-curated datasets can yield meaningful and durable insights when the goal is to uncover underlying relevance structures rather than estimate population-level distributions [18, 33]. In the following, we statistically analyze GoldRiM recommendations.

Lexical overlap: Naively, relevance has been often measured via surface-level lexical overlap between seeds and recommendations

[4] https://arxiv.org/
[5] https://mathscinet.ams.org/mathscinet
[6] https://www.jstor.org/

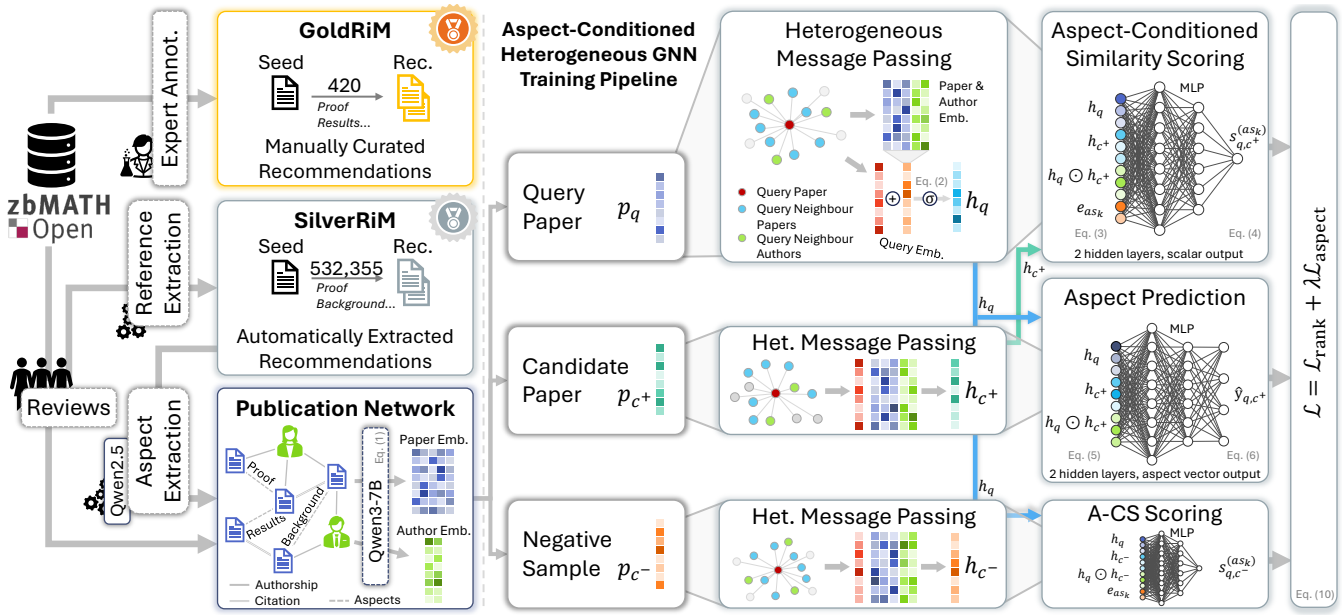

Figure 2: Architecture and learning pipeline of AchGNN.

(see TF-IDF approaches in Section 2). In mathematics, however, relevance is often rooted in deep connections between mathematical concepts that may or may not be easily visible by lexical overlaps. Figure 3 shows the Jaccard similarity based on 2-gram[7] overlaps in GoldRiM and PwC. Across GoldRiM sets, similarity values are uniformly low compared to PwC sets. Specifically, GoldRiM recommendations exhibit a mean similarity of just 0.031, with a tightly concentrated distribution near zero, demonstrating that expert recommendations are not driven by lexical overlap. Interestingly, citation-based pairs show an even lower overlap, indicating that citation adjacency does not correspond to lexical similarity in mathematical literature. In PwC, the overlap and similarity score varies significantly more depending on the selected seeds with some seeds showing similar distributions as in zbMATH Open. However, the other end of the spectrum, shown in Figure 3, exhibits significantly more lexical overlap.

Embeddings analysis: Semantic similarity analysis via embeddings has typically yielded promising results in many search or recommendation tasks in the past [29, 34]. Figure 4 illustrates the semantic distances between seeds and recommendations based on Qwen3-7B [53] embeddings in GoldRiM and PwC. Again, GoldRiM citations and recommendations exhibit the lowest similarity compared to the other datasets. However, the difference to recommendations in PwC is not as pronounced as for the lexical comparison above. More crucially, however, is the clear overlap between similarity scores of PwC recommendations and the most similar articles in the PwC dataset. This strongly suggests that highly similar embeddings correlate with relevance in PwC. In contrast, this correlation is not evident in GoldRiM. In fact, recommendations

[7]We further analyzed 1- and 3-gram overlap. However, 1-grams produced higher scores but exhibits the same patterns as 2-grams, and 3-gram overlap yielded near-zero overlap. We, therefore, only discuss 2-gram analysis here.

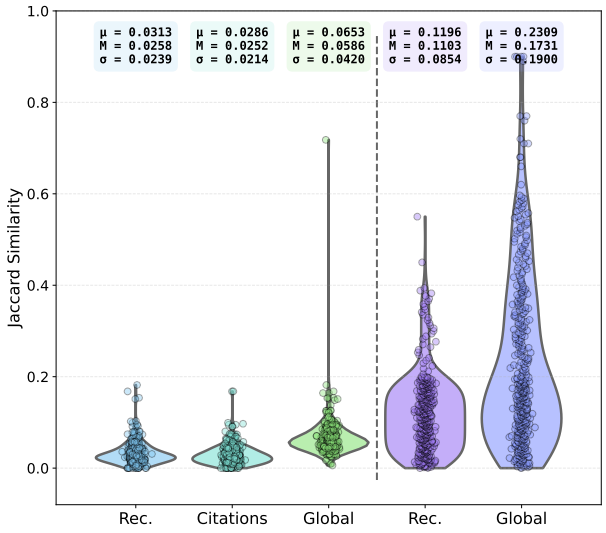

Figure 3: Distribution of 2-gram overlap between the GoldRiM seeds (left) and PwC seeds (right) and their respective recommendations (Rec.), citations (Citations), and all documents in the respective datasets (Global). Citations and global entries are capped to 420 top scoring articles. μ: mean, M: median, and σ: standard deviation.

and citations have significantly lower similarity scores compared to what is available within zbMATH Open (Global), further underlying that embeddings alone are also insufficient for capturing expert-level recommendations in mathematics.

Citation-proximity analysis: Both analysis above indicate towards a correlation between citations and recommendations within GoldRiM. However, we find that 23% of GoldRiM recommendations

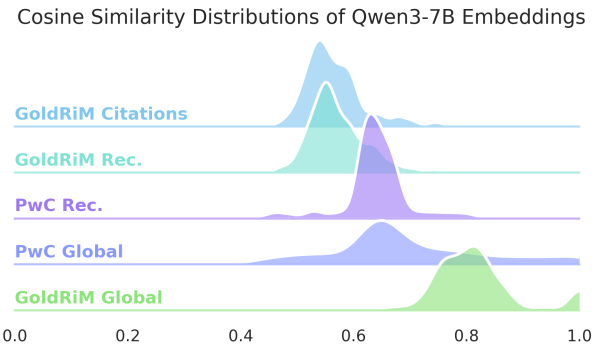

Figure 4: Distribution of 420 Qwen3-7B embedding similarities in GoldRiM and PwC sorted by their respective mean values between the seeds and their recommendations (Rec.), citations (Citations) and all documents from the representative datasets (Global).

have no citation path between the seed and the recommended paper. Among the remaining pairs, 43% are connected by a single citation edge and only 3% by two edges. Such short paths are not discriminative, as many non-recommended papers are also reachable within one or two jumps. For the remaining 31% of recommendations, the shortest citation path exceeds three edges, requiring traversal through multiple intermediary documents. Bibliographic coupling achieves a recall of 34%, while co-citation achieves 35%. These results indicate that citation structures are inherently important for recommendations, but capture only a rather limited subset of expert judgments. As such, solely relying on citation networks alone likely result in an insufficient pool of recommendation candidates. After consulting with the expert from zbMATH Open, taking authorship lineage into account can be considered a natural extension of the citation network. Mathematicians tend to reuse proof styles and foundational results or methodologies are often referred to by the author's names rather than specialized descriptive terms [17, 31].

Aspect-based similarity: The inability of relying solely on text similarity or citation-based signals to account for expert recommendations indicates that mathematical relevance is governed by implicit relationships that are not directly observable. This motivated a qualitative analysis of how expert recommendations are conceptually related. Together with the expert, we manually analyzed all 420 recommendation pairs and assigned labels describing the underlying relationship between each seed and its recommendation. This analysis revealed that recommendations are connected through distinct *aspects*, each capturing a specific mode of mathematical reasoning. The initial labeling yielded 66 *aspects*. Many labels, however, reflected surface-level linguistic variation rather than fundamentally different relationships. For example, labels such as *restriction*, *reduction*, and *limiting case* differ lexically but express the same conceptual relationship. Furthermore, *aspect*-based retrieval requires *aspect* spaces small enough to allow statistically meaningful evaluation and tractable supervision. In order to enable comparability with community standards, we consolidated the original labels into four high-level *aspects*, *Specialization / Restriction* (size: 32.14%), *Modification / Generalization* (40%), *Results* (17.86%), and *Prove / Cases* (10%)—through iterative discussion with a domain expert (detailed descriptions of each category are available in our repository[1]). This abstraction preserves conceptual distinctions while enabling generalization and interpretability.

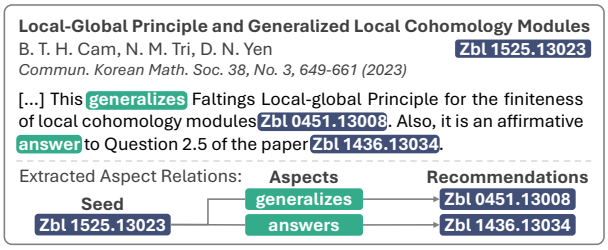

Figure 5: Example of *aspect* extractions.

3.1.2 SilverRiM. In essence, the analysis of GoldRiM answers RQ1. GoldRiM shows that relevance does not seem to correlate with lexical or embedding similarities but can be partially derived from citations and *aspect* connections. GoldRiM's inherent flaw is limited size and the bias from a single annotator. It results in the necessity of a scalable dataset that covers a larger area of zbMATH Open and more variety in generating *aspects* and relevance. During the construction of GoldRiM, we observed that many of the manually identified *aspects* were explicitly articulated in zbMATH Open reviewer-written abstracts, often through references to prior work (see Figure 5). Considering zbMATH Open's reviewers are often highly decorated field experts themselves, we can consider such in-abstract references as recommendations. In combination, this allows us to generate comparable data to GoldRiM at scale.

As of October 2025, zbMATH Open contains approximately 312K abstracts with 786K in-abstract references by over 13k different reviewers, providing a natural basis for large-scale extraction of mathematical recommendations and *aspect*-labels. Prior work shows that few-shot LLM-based phrase extraction can approximate human annotations with competitive quality in citation-related tasks [38, 44]. Therefore, we use Qwen2.5-14B [39], which performs strongly in citation-intent prediction and extraction [24], in order to extract *aspect*-labels. Given an abstract and its in-abstract references, the model is prompted to extract short descriptive phrases characterizing the relationship between the seed paper and each referenced work. After filtering malformed outputs and missing contexts, we obtain 212K seed papers, 532K recommendation pairs, and 7,326 distinct relationship labels. This collection constitutes a large-scale **Silver**-standard dataset of **R**ecommendations **i**n **M**athematics (**SilverRiM**).

The extracted relationship labels capture fine-grained linguistic expressions used by reviewers, but they are inherently noisy and highly sparse. Treating them as independent ground-truth *aspects* hinders generalization, evaluation, and interpretability. In *aspect*-based CbRPR, aspects encode more semantically informative relationships than their exact textual wording, a pattern consistently observed in prior domain-specific CbRPR work [29, 34, 36]. Hence, we align the SilverRiM aspect-labels with the established ontology of four categories from GoldRiM utilizing k-means clustering. Manual inspection of representative labels from each cluster confirms semantic alignment with the corresponding expert-defined *aspects*. Figure 6 shows the most frequently *aspect*-labels of SilverRiM in their associated cluster. Although automatically derived, the dataset resembles a very similar distribution as in GoldRiM, with

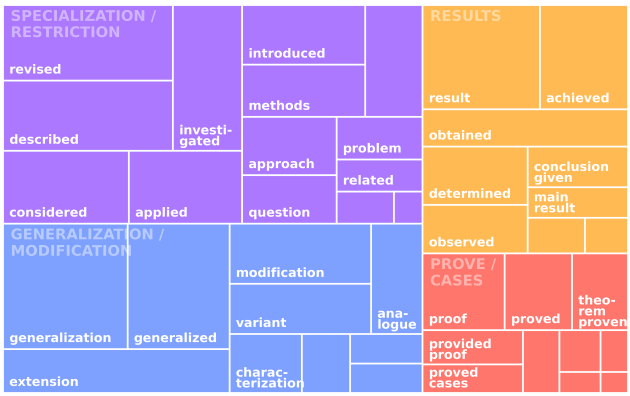

Figure 6: Fine-grained *aspect* labels on SilverRiM.

Table 1: Statistics of the evaluation datasets. *Aspect*-labels refers to the number of fine- grained generated labels.

Characteristic	GoldRiM	SilverRiM	PwC
Total Seeds	80	212,382	157,606
Rec. Pairs	420	532,355	1,227,058
Unique *Aspects*	4	4	3
Aspect-labels	66	7,326	3,952
Avg. Rec. per Seed	5.25	2.50	7.78

Specialization/Restriction (37.83%) and *Generalization/Modification* (29.30%), followed by *Results* (21.04%) and *Prove/Cases* (11.83%). Table 1 provides an overview of the curated datasets in comparison with PwC.

A comparison with existing *aspect*-based CbRPR datasets, which largely focus on CS and BM literature, reveals limited overlap and substantial domain-specific divergence. While the *Results aspect* is shared, other *aspects* in our dataset are mathematics-specific and capture relations not represented in prior datasets. This highlights mathematics as a distinct and previously underrepresented domain for CbRPR, reflecting fundamentally different retrieval and recommendation needs. Our dataset therefore fills an important gap and enables the study of *aspect*-aware recommendations in a new and challenging domain.

3.2 Aspect-Conditioned Heterogeneous GNN

Prior work [25, 49] has shown that heterogeneous GNNs effectively exploit citation and authorship structure for CbRPR; however, existing formulations learn a single, *aspect*-agnostic representation per paper (see Section 2). Motivated by *aspect*-conditioned graph training in e-commerce recommendations [45, 52], we adapt *aspect*-aware modeling to mathematical CbRPR by integrating *aspect* conditioning into a heterogeneous citation–authorship graph. Authorship serves as a domain-salient signal of conceptual lineage, while *aspect* conditioning enables fine-grained distinctions among relevance relations during training. This results in an **A**spect-**c**onditioned **h**eterogeneous **GNN** (AchGNN), which combines heterogeneous message passing with *aspect*-conditioned scoring to support *aspect*-based CbRPR in mathematics.

3.2.1 Graph Initialization. We define a heterogeneous undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with two node types: papers $\mathcal{P}$ and authors $\mathcal{A}$, such that $\mathcal{V} = \mathcal{P} \cup \mathcal{A}$. The edge set captures three complementary relations: **Aspect-labeled edges** $(p_i, p_j, as_k) \in \mathcal{E}_{as}$, indicating that papers p_i and p_j are similar with *aspect* $as_k \in \mathcal{AS}$, **Citations** $(p_i, p_j) \in \mathcal{E}_c$, and **Authorship edges** $(a_m, p_i) \in \mathcal{E}_a$, connecting authors to their papers and encoding intellectual lineage with $\mathcal{E} = \mathcal{E}_{as} \cup \mathcal{E}_a \cup \mathcal{E}_c$. This construction enables information flow not only between textually linked papers but also across papers connected through shared authors, to compliment textual similarity. Paper nodes are initialized using titles and abstract embeddings (with in-abstract citations removed). Author nodes are initialized using Author-name embeddings.

$$h_i^{(0)} = f_{\text{text}}(p_i) \in \mathbb{R}^d, h_l^{(0)} = f_{\text{text}}(a_m) \in \mathbb{R}^d \tag{1}$$

where f_{text} is an embedding obtained using Qwen3-7B, as it yields the strongest performance among embedding-based base model baselines on GoldRiM and SilverRiM (see Section 4.1). The same model is used to generate author name embeddings to ensure consistent embedding dimensionality across node types.

3.2.2 Heterogeneous Message Passing. We employ a GraphSAGE-inspired heterogeneous GNN Message passing [16] given its scalability and empirical robustness in scholarly graphs [55]. We stack L GNN layers and use fixed neighborhood sampling to balance coverage and noise that has been observed optimal in heterogeneous scholarly graphs [16, 51].

Let $h_v^{(l)}$ denote the aggregated embedding of node v at layer l. Node updates are computed as:

$$h_v^{(l+1)} = \sigma\left(W_0^{(l)} h_v^{(l)} + \sum_{r \in \mathcal{R}} \sum_{u \in \mathcal{N}_r(v)} \frac{1}{|\mathcal{N}_r(v)|} W_r^{(l)} h_u^{(l)}\right), \tag{2}$$

where $\mathcal{R}$ denotes the relation types $\{c, a\}$ (citation or authorship), $\mathcal{N}_r(v)$ the neighbors of v under relation r, and $W_r^{(l)}$ relation-specific transformation matrices. The resulting embeddings encode both connected papers via citations and authorship lineage.

3.2.3 Aspect-Conditioned Similarity Scoring. *Aspect* conditioning is applied at the scoring level rather than during message passing to avoid entangling multiple *aspects* within node embeddings. We combine the paper-to-node transformation proposed in GraphCL [51] with *aspect*-conditioning used in FigGNN [45] and MA-GNN [52]. Given a query paper p_q, a candidate paper p_c, and an *aspect* $as_k \in \mathcal{AS}$, we construct an *aspect*-conditioned interaction vector:

$$z_{q,c}^{(as_k)} = \left[\, h_q \parallel h_c \parallel h_q \odot h_c \parallel e_{as_k} \,\right], \tag{3}$$

where $h_q, h_c \in \mathbb{R}^d$ are the final GNN embeddings of the query and candidate papers, $\odot$ denotes element-wise multiplication, and $e_{as_k} \in \mathbb{R}^{d_{as}}$ is a learnable embedding associated with *aspect* as_k. This design explicitly exposes both symmetric interactions ($h_q \odot h_c$) and *aspect* identity, allowing the scoring function to learn *aspect*-specific similarity patterns. The interaction vector is mapped to a scalar relevance score by a standard Multi-Layer Perceptron (MLP) f_{score}:

$$s_{q,c}^{(as_k)} = f_{\text{score}}\left(z_{q,c}^{(as_k)}\right). \tag{4}$$

3.2.4 Aspect Prediction Objective. We introduce an auxiliary *aspect*-prediction task to encourage *aspect*-discriminative representations. For a paper pair (p_q, p_c), we construct an aspect-prediction vector:

$$z_{q,c} = \left[\, h_q \parallel h_c \parallel h_q \odot h_c \,\right]. \tag{5}$$

The interaction vector is passed through an aspect-classification network:

$$\hat{\mathbf{y}}_{q,c} = f_{\text{aspect}}\left(z_{q,c}\right), \tag{6}$$

where $\hat{\mathbf{y}}_{q,c} \in \mathbb{R}^{|\mathcal{AS}|}$ denotes the predicted aspect logits. The aspect classifier is implemented as a MLP with output dimension equal to the number of aspects:

$$f_{\text{aspect}}(z) = W_a\, \sigma(W_z z + b_z) + b_a. \tag{7}$$

The auxiliary loss is the standard cross-entropy loss:

$$\mathcal{L}_{\text{aspect}} = \text{CE}\left(\text{softmax}\left(\hat{\mathbf{y}}_{q,c}\right), as_k\right). \tag{8}$$

Aspect prediction is formulated as a multi-class classification problem over observed positive paper pairs. Cross-entropy loss provides implicit negative supervision through competition among aspect classes, and negative paper pairs are not used, as their aspects are undefined.

3.2.5 Learning Objectives. We jointly optimize AchGNN for (i) aspect-conditioned separation and (ii) aspect prediction. For aspect-conditioned separation, we adopt a pairwise Bayesian Personalized Ranking (BPR) loss [40]:

$$\mathcal{L}_{\text{rank}} = -\log \sigma\left(s_{q,c^+}^{(e_{as_k})} - s_{q,c^-}^{(e_{as_k})}\right). \tag{9}$$

encouraging higher scores for aspect-consistent positive edges (c^+) and lower scores for negative edges (c^-). The classification loss is standard cross-entropy over aspects. The final objective is:

$$\mathcal{L} = \mathcal{L}_{\text{rank}} + \lambda\, \mathcal{L}_{\text{aspect}}, \tag{10}$$

where λ controls the contribution of aspect supervision. This joint formulation regularizes ranking by enforcing aspect-discriminative representations.

3.2.6 Inference. At inference time, AchGNN computes contextual embeddings for all papers, evaluates aspect-conditioned scores $s_{q,c}^{(as)}$, and ranks candidates accordingly:

$$\text{TopN}^{(as)}(p_q) = \text{argsort}_c\; s_{q,c}^{(as)}. \tag{11}$$

This yields aspect-specific recommendation lists that reflect both relational proximity and mathematical lineage.

4 Experiments

We conduct experiments on GoldRiM, SilverRiM, and PwC [19] to address RQ2. PwC consists of machine learning papers annotated with structured aspects: *task*, *method*, and *dataset* and enables evaluation beyond the mathematics domain. Furthermore, PwC has been widely used as a benchmark for aspect-based CbRPR in the past [29, 34, 36].

4.1 Evaluated models

We compare AchGNN against a set of diverse state-of-the-art approaches consisting of embedding-based and graph-based baselines. **SPECTER** [10] and **SciBERT** [3] are both BERT-based language models that have been effectively used in *aspect*-based CbRPR and often serve for robust baseline evaluations [29, 35]. The **SciBERT-FT** [34] variant employs a fine-tuned Siamese architecture based on SciBERT and currently represents the best-performing method for *aspect*-based CbRPR on the PwC dataset [19]. Since LLM embeddings are predominant in CbRPR, we select state-of-the-art LLM embedding models from MTEB [32]. The MTEB leaderboard ranking[8] lists **Qwen2-7B** [39] as the model with the highest overall performance as of October 2025. The MTEB benchmark further includes domain-specific tasks. Notably, **Qwen3-7B** [53] attains state-of-the-art performance on the arXiv clustering task, with a majority of mathematics and physics research papers. Building on this, **Qwen3-7B-FT** denotes a fine-tuned variant trained using instruction-tuned samples and contrastive-loss [27, 43]. Another MTEB subtask involves clustering of question–answer pairs from Math Stack Exchange. On this task, **Memtron** [1] demonstrates the best performance among all evaluated models. Lastly, Graph Contrastive Learning (**GraphCL**) [51] is a top-performing approach [5, 48] leveraging a joint graph for recommendation, making it a strong graph-based baseline for our evaluation. All models contain fewer than 8B parameters, reflecting realistic deployment constraints in large-scale digital libraries.

4.2 Evaluation & Training Setup

We evaluate all models under an *aspect*-specific retrieval paradigm, where recommendations are generated and assessed independently for each *aspect*. Let $\mathcal{AS} = \{as_1, \ldots, as_n\}$ denote the set of n *aspects* defined in the dataset. For each *aspect* $as_j \in \mathcal{AS}$, we construct independent training and test splits, resulting in n *aspect*-specific retrieval tasks. During evaluation, each document in the test split corresponding to *aspect* a_j is treated as a query seed. Given a seed document d_s, the model computes an *aspect*-specific representation $\vec{d}_s(a_j)$, which is used to retrieve the top-k candidate documents via k-nearest neighbor search. Document similarity is measured using cosine similarity, following standard practice in *aspect*-based CbRPR [36, 54]. In addition to *aspect*-wise evaluation, we report results under a *general* setting. Here, all test documents from all *aspects* are pooled into a single test set, and retrieval is performed without specifying any *aspect*. This evaluation measures overall recommendation effectiveness and allows us to assess whether performance gains observed in *aspect*-specific scenarios generalize to an *aspect*-agnostic retrieval setting.

Trained models (SciBERT-FT, Qwen3-7B-FT, GraphCL, and AchGNN) are provided with *aspect* information, whereas the remaining base models produce *aspect*-agnostic representations. To prevent models from exploiting abstract reference cues (such as in Figure 5), we remove all explicit mentions of other zbMATH Open documents from abstracts prior to training and evaluation. Due to its expert-curated nature and limited size, GoldRiM is used exclusively for testing. SilverRiM and PwC are each split into 75% training and 25% test sets, with 5-fold cross-validation employed to ensure robustness.

[8] https://huggingface.co/spaces/mteb/leaderboard

Table 2: Overall results for $k = 10$ retrieved documents in GoldRiM and SilverRiM. Precision (P), recall (R), and mean reciprocal rank (MRR) are reported.

Aspects ⟶	General			Specialization / Restriction			Results			Prove / Cases			Modification / Generalization		
Methods ↓	P	R	MRR	P	R	MRR	P	R	MRR	P	R	MRR	P	R	MRR
Dataset: GoldRiM															
SPECTER	0.023	0.050	0.111	0.017	0.038	0.083	0.012	0.029	0.061	0.008	0.021	0.044	0.006	0.016	0.031
SciBERT	0.007	0.016	0.035	0.005	0.012	0.026	0.004	0.009	0.019	0.003	0.007	0.014	0.002	0.005	0.010
SciBERT-FT	0.075	0.133	0.261	0.061	0.112	0.228	0.047	0.091	0.193	0.035	0.072	0.161	0.026	0.054	0.129
Qwen2-7B	0.077	0.156	0.391	0.062	0.131	0.332	0.048	0.102	0.274	0.035	0.078	0.221	0.026	0.059	0.176
Qwen3-7B	0.106	0.213	0.413	0.089	0.184	0.361	0.071	0.148	0.302	0.054	0.116	0.248	0.041	0.089	0.201
Qwen3-7B-FT	0.122	0.245	0.475	0.102	0.212	0.415	0.082	0.170	0.347	0.062	0.133	0.285	0.047	0.102	0.231
Memtron	0.091	0.187	0.343	0.075	0.159	0.297	0.059	0.128	0.243	0.044	0.101	0.196	0.033	0.077	0.156
GraphCL	0.271	0.352	0.471	0.238	0.319	0.431	0.203	0.284	0.392	0.172	0.251	0.351	0.143	0.217	0.309
AchGNN	**0.341**	**0.425**	**0.593**	**0.312**	**0.398**	**0.556**	**0.279**	**0.361**	**0.512**	**0.243**	**0.326**	**0.468**	**0.211**	**0.289**	**0.423**
Dataset: SilverRiM															
SPECTER	0.043	0.345	0.252	0.046	0.363	0.270	0.039	0.317	0.221	0.039	0.313	0.219	0.039	0.317	0.228
SciBERT	0.005	0.044	0.036	0.007	0.054	0.045	0.004	0.030	0.024	0.002	0.018	0.015	0.004	0.033	0.026
SciBERT-FT	0.055	0.350	0.262	0.056	0.362	0.274	0.051	0.331	0.243	0.050	0.327	0.241	0.050	0.329	0.242
Qwen2-7B	0.057	0.442	0.310	0.059	0.457	0.327	0.052	0.415	0.275	0.053	0.420	0.287	0.053	0.419	0.287
Qwen3-7B	0.072	0.566	0.399	0.074	0.576	0.413	0.068	0.542	0.366	0.072	0.569	0.385	0.069	0.547	0.371
Qwen3-7B-FT	0.075	0.583	0.409	0.082	0.593	0.425	0.070	0.558	0.377	0.074	0.586	0.387	0.071	0.563	0.373
Memtron	0.064	0.502	0.362	0.066	0.516	0.379	0.060	0.474	0.335	0.062	0.488	0.342	0.061	0.479	0.333
GraphCL	0.083	0.615	0.396	0.079	0.620	0.410	0.076	0.598	0.372	0.079	0.620	0.388	0.081	0.596	0.372
AchGNN	**0.086**	**0.634**	**0.410**	**0.084**	**0.640**	**0.447**	**0.082**	**0.614**	**0.381**	**0.081**	**0.636**	**0.388**	**0.083**	**0.613**	**0.374**

Table 3: Overall results for the $k = 10$ retrieved documents in PwC. SciBERT-FT results as reported by Ostendorff et al. [34].

Aspects ⟶	General			Task		
Methods ↓	P	R	MRR	P	R	MRR
SciBERT-FT	-	-	-	**0.569**	**0.242**	**0.708**
AchGNN	0.486	0.217	0.529	0.478	0.214	0.524
	Method			Dataset		
SciBERT-FT	0.407	0.168	**0.588**	**0.270**	0.374	**0.533**
AchGNN	**0.512**	**0.236**	0.563	0.201	**0.469**	0.501

For training, SilverRiM recommendation pairs are used as positive samples. Hard negatives are constructed by retrieving an equal number of top-ranked candidates using Qwen3-7B embeddings with cosine similarity, excluding documents that appear in the training data.

We report Precision@k (P), Recall@k (R), and Mean Reciprocal Rank (MRR). While the average number of GoldRiM and SilverRiM recommendations per query is 5.25 and 2.50, respectively, we set $k = 10$ to evaluate ranking quality beyond minimal recall and to remain comparable with prior work on PwC. The retrieval corpus consists of 3.8 million English-language zbMATH Open documents. For AchGNN, the number of paper neighbors was set to 15, author neighbors to five, the number of layers $L = 2$, and $\lambda = 0.2$.

4.3 Experimental results

Tables 2 summarizes the results on GoldRiM and SilverRiM. All BERT-based models underperform on these datasets. Although SPECTER outperforms SciBERT among base models, its gains do not translate into improved fine-tuning outcomes for mathematical CbRPR; this behavior has also been observed in prior state-of-the-art work, where fine-tuned variants of SPECTER underperform relative to SciBERT [34]. SciBERT-FT, fine-tuned on SilverRiM, performs significantly better, but still falls short in comparison to the other approaches. This is particularly noteworthy since SciBERT-FT is the best performing model on the PwC task [29, 34]. Table 3 compares AchGNN on the PwC dataset and its three *aspects* with the reported results of SciBERT-FT [34] (here fine-tuned on PwC). The relative competitive performance of AchGNN on PwC indicates a more robust approach across multiple datasets, this is promising, since a general lack of generalizability across datasets of recommender systems is rather common [8, 30] and also evident from our evaluation of the baselines on our math-specific datasets.

Among base LLM embedding models, Qwen3-7B achieves the strongest performance. This result is consistent with its strong performance on math-intensive clustering tasks in MTEB, suggesting improved representations of formal and abstract text. The fine-tuned variant, Qwen3-7B-FT, further improves performance over the base model. However, its performance still falls short of graph-based approaches. This suggests that simple fine-tuning alone is insufficient to capture implicit CbRPR in mathematics, and that performance bottlenecks in this domain are not solely due to limited supervision [47].

Among graph-based methods, GraphCL sufficiently outperforms embedding-only baselines across most *aspects* and both datasets, highlighting the importance of structural signals obtained via a

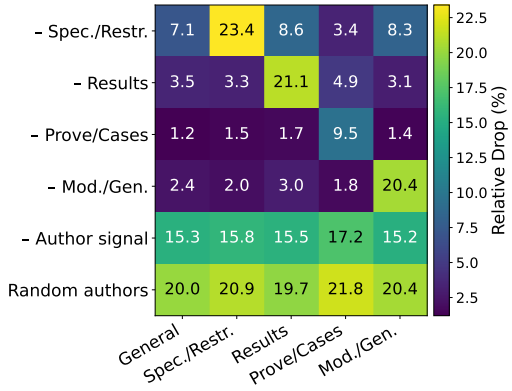

Figure 7: Relative R@10 performance drops for removing and modifying *aspects* (y-axis) on GoldRiM.

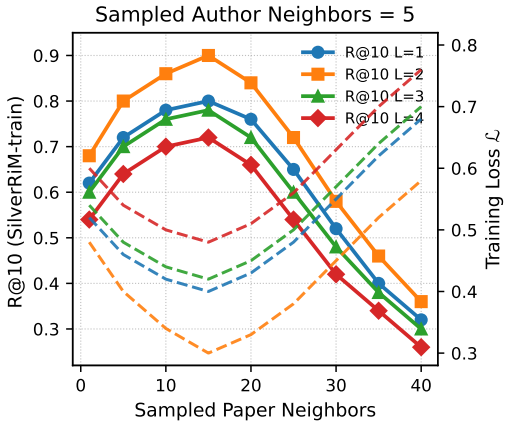

Figure 8: Effects of sampled neighborhood sizes and GNN depths (L) on SilverRiM training, where sampled papers refers to the number of neighbors sampled at each layer.

joint paper and author graph. AchGNN achieves the best performance across all *aspects*, metrics and datasets (GoldRiM and SilverRiM), with a *General* MRR of 0.593 (on GoldRiM). This corresponds to a 26% relative improvement over GraphCL, despite operating on the same graph structure, and more than a five-fold improvement over SciBERT-FT. Addressing RQ2, AchGNN's modeling of *aspects* during representation learning captures representations for mathematical documents more effectively than the *aspect*-agnostic contrastive learning employed by GraphCL. Lastly, the performance change between evaluations on SilverRiM and GoldRiM across all models relative to each other are remarkably consistent, mostly preserving the same order of models (see the bar chart representations of Table 2 in our repository). This indicates that GoldRiM, despite the inherent bias from just one annotator and limited scope, can still be used as a meaningful evaluation dataset.

4.4 Ablation Study

In addition to the general evaluation, we conduct an ablation study to quantify the contribution of individual *aspects* and graph signals in AchGNN. Each variant is trained by removing one supervision source while keeping the architecture, optimization, and evaluation protocol fixed. We further analyze authorship contribution added to the graph structure and evaluate performances with varying parameters for GNN layers and the number of neighborhood nodes.

4.4.1 Aspect Removal. Figure 7 summarizes the relative performance drops on Recall@10. Removing *Specialization/Restriction* supervision leads to the largest performance degradation. In contrast, removing *Prove / Cases* results in the smallest performance drop, reflecting its limited training size (see Figure 6) and highly localized relevance patterns. *Aspect* ablations primarily degrade performance on the removed *aspect*, with limited spillover to others. This is further underlined by the relatively small performance drops on *General*. This indicates that AchGNN does not collapse multiple relevance dimensions into a single embedding. Instead, *aspect*-conditioned scoring successfully disentangles relevance criteria [52]. We further normalized the performance degradation against the number of removed edges during training but found the same relative performance changes with respect to each *aspect*.

4.4.2 Validation of Authorship Lineage. A potential concern is that the performance gains of AchGNN may stem merely from adding additional edges to the graph, rather than from meaningful authorship information. To validate the role of authorship lineage, we perform authorship ablation and study the affect of purposefully adding noise to the data by evaluating the effects of randomized authorship. Both variants cause substantial degradation, often even greatly exceeding that of any single-*aspect* removal. This is further apparent on the normalized performance drop, since authorship signal only attributes to about 12% of the connections in the graph, which is on par with the smallest *aspect* set *Prove / Cases*. Randomized authorship performs worst and significantly degrade performance even compared to removing authorship information entirely in absolute terms. This strongly indicates that authorship information contribute to semantic lineage rather than mere added graph connectivity [17]. Furthermore, it highlights that AchGNN leverages meaningful intellectual structure beyond textual similarity.

4.4.3 Effect of Neighborhood Sampling Size and GNN Depth. Figure 8 evaluates the sensitivity of AchGNN to paper-node neighborhood sampling and GNN depth on SilverRiM, with the author-node neighborhood size fixed to five. AchGNN performs best with moderate neighborhood sizes, achieving its lowest training loss and highest Recall@10 under the configuration of 15 paper neighbors, 5 author neighbors, and $L = 2$ GNN layers. Increasing the paper neighborhood size initially improves performance. This confirms the benefit of incorporating citation and authorship context beyond single node representation through embedding [16, 49, 55]. Further increases lead to degraded Recall@10 and higher loss due to over-smoothing [4, 51]. For author nodes, sampling a single neighbor reduces performance (over 5%). In contrast, increasing the cap from five to ten yields only marginal changes (below 1%). This behavior aligns with the SilverRiM average author-degree distribution of approximately 2.5 authors, causing author neighborhoods to saturate once moderate sampling thresholds are reached. Overall, AchGNN benefits most from moderate neighborhood sizes, balancing structural context aggregation with over-smoothing avoidance.

5 Conclusion and Future Work

This work addresses a fundamental gap in content-based research paper recommendation (CbRPR) in mathematics, a domain in which relatedness is rarely captured by surface-level similarity. Through an expert-driven analysis, we demonstrate that mathematical CbRPR are inherently *aspect*-driven, grounded in conceptual relations such as theorem generalization, dual formulations, and proof reuse relations that remain invisible to state-of-the-art CbRPR approaches. To support a systematic study of this problem, we introduced GoldRiM, a high-quality test-dataset, designed to study and uncover mathematical relevance, and SilverRiM, a scaled version which allows for training and evaluation of *aspect*-aware CbRPR in mathematics. Motivated by the limitations of standalone LLM embeddings for mathematical content, we proposed AchGNN, an *aspect*-conditioned heterogeneous graph neural network that jointly models textual semantics, citation structure, and author identity. Experimental results show that AchGNN consistently outperforms state-of-the-art *aspect*-based CbRPR methods. Beyond quantitative improvements, our evaluation demonstrated that effective mathematical CbRPR requires structural and relational modeling than purely semantic. Ablation analyses confirm that AchGNN's improvements stem from *aspect* supervision, meaningful authorship lineage, and selected neighborhood and depth configurations. We publicly released our datasets and code: https://github.com/gipplab/MathAspectRecSys.

Hosting all data on the MaRDI platform further allows us to evaluate user-based interactions in the future, enabling us to explore state-of-the-art approaches on user-item interaction graphs. Furthermore, we plan to validate identified math-specific *aspects* with additional reviewers across larger subfields. In addition, the current graph size is limited to zbMATH Open. zbMATH Open articles, however, often refer to outside articles, most notably arXiv. Expanding the network of paper relationships and authorship lineage across multiple datasets potentially improves generalizability and manifests AchGNN as a standard for mathematical research recommender systems.

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 437179652; 567156310. The authors gratefully acknowledge the computing time granted by the KISSKI project. The calculations for this research were conducted with computing resources under the project *MathRecSys*. We utilized AI models to proofread, enhance grammar, and improve sentence clarity. Every sentence in the resulting text is checked by the authors, and citations are provided wherever available. We take full responsibility for the text in this manuscript.

References

[1] Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. 2025. Llama-Embed-Nemotron-8B: A Universal Text Embedding Model for Multilingual and Cross-Lingual Tasks. arXiv:2511.07025 [cs.CL] https://arxiv.org/abs/2511.07025

[2] Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. 2016. Paper recommender systems: a literature survey. *International Journal on Digital Libraries* 17, 4 (2016), 305–338.

[3] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: Pretrained Language Model for Scientific Text. In *EMNLP*. arXiv:arXiv:1903.10676

[4] Chen Cai and Yusu Wang. 2020. A note on over-smoothing for graph neural networks. *arXiv preprint arXiv:2006.13318* (2020).

[5] Xuheng Cai, Chao Huang, Lianghao Xia, and Xubin Ren. 2023. LightGCL: Simple Yet Effective Graph Contrastive Learning for Recommendation. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=FKXVK9dyMM

[6] Tanmoy Chakraborty, Amrith Krishna, Mayank Singh, Niloy Ganguly, Pawan Goyal, and Animesh Mukherjee. 2016. Ferosa: A faceted recommendation system for scientific articles. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 528–541.

[7] Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. Solvent: A mixed initiative system for finding analogies between research papers. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–21.

[8] Jin Yao Chin, Yile Chen, and Gao Cong. 2022. The Datasets Dilemma: How Much Do We Really Know About Recommendation Datasets?. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*. ACM, 141–149. doi:10.1145/3488560.3498519

[9] Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural Scaffolds for Citation Intent Classification in Scientific Publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 3586 3596.

[10] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 2270–2282. doi:10.18653/v1/2020.acl-main.207

[11] Daniel Cummings and Marcel Nassar. 2020. Structured Citation Trend Prediction Using Graph Neural Networks. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 3897–3901. doi:10.1109/ICASSP40776.2020.9054769

[12] Ziheng Duan, Yueyang Wang, Weihao Ye, Qilin Fan, and Xiuhua Li. 2022. Connecting latent relationships over heterogeneous attributed network for recommendation. *Applied Intelligence* 52, 14 (Nov. 2022), 16214–16232. doi:10.1007/s10489-022-03340-7

[13] Maryam Fatima. 2025. FIRMA: Bidirectional Formal-Informal Mathematical Language Alignment with Proof-Theoretic Grounding. In *Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP 2025)*, Marco Valentino, Deborah Ferreira, Mokanarangan Thayaparan, Leonardo Ranaldi, and Andre Freitas (Eds.). Association for Computational Linguistics, Suzhou, China, 62–76. doi:10.18653/v1/2025.mathnlp-main.5

[14] Xiaoyue Feng, Hao Zhang, Yijie Ren, Penghui Shang, Yi Zhu, Yanchun Liang, Renchu Guan, and Dong Xu. 2019. The Deep Learning–Based Recommender System "Pubmender" for Choosing a Biomedical Publication Venue: Development and Validation Study. *J Med Internet Res* 21, 5 (24 May 2019), e12957. doi:10.2196/12957

[15] Heather Graves, Shahin Moghaddasi, and Azirah Hashim. 2013. Mathematics is the method: Exploring the macro-organizational structure of research articles in mathematics. *Discourse Studies* 15, 4 (2013), 421–438. arXiv:https://doi.org/10.1177/1461445613482430 doi:10.1177/1461445613482430

[16] William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 1025–1035.

[17] Klaus Hulek and Olaf Teschke. 2023. How do mathematicians publish?–Some trends. *European Mathematical Society Magazine* 129 (2023), 36–41.

[18] David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the Evolution of a Scientific Field through Citation Frames. *Transactions of the Association for Computational Linguistics* 6 (07 2018), 391–406.

[19] Marcin Kardas, Piotr Czapla, Pontus Stenetorp, Sebastian Ruder, Sebastian Riedel, Ross Taylor, and Robert Stojnic. 2020. AxCell: Automatic Extraction of Results from Machine Learning Papers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 8580–8594. doi:10.18653/v1/2020.emnlp-main.692

[20] Özge Kart, Alexandre Mestiashvili, Kurt Lachmann, Richard Kwasnicki, and Michael Schroeder. 2022. Emati: a recommender system for biomedical literature based on supervised learning. *Database* 2022 (12 2022), baac104. arXiv:https://academic.oup.com/database/article-pdf/doi/10.1093/database/baac104/47779573/baac104.pdf doi:10.1093/database/baac104

[21] Abdalsamad Keramatfar, Mohadeseh Rafiee, and Hossein Amirkhani. 2022. Graph Neural Networks: A bibliometrics overview. *Machine Learning with Applications* 10 (2022), 100401. doi:10.1016/j.mlwa.2022.100401

[22] Vitaly Kiryushchenko. 2023. *Diagrams, Visual Imagination, and Continuity in Peirce's Philosophy of Mathematics*. Springer.

[23] Yuta Kobayashi, Masashi Shimbo, and Yuji Matsumoto. 2018. Citation Recommendation Using Distributed Representation of Discourse Facets in Scientific Articles.

In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries* (Fort Worth, Texas, USA) *(JCDL '18)*. Association for Computing Machinery, New York, NY, USA, 243–251. doi:10.1145/3197026.3197059

[24] Paris Koloveas, Serafeim Chatzopoulos, Thanasis Vergoulis, and Christos Tryfonopoulos. 2025. Can llms predict citation intent? an experimental analysis of in-context learning and fine-tuning on open llms. In *International Conference on Theory and Practice of Digital Libraries*. Springer, 207–224.

[25] Christin Katharina Kreutz and Ralf Schenkel. 2022. Scientific paper recommendation systems: a literature review of recent publications. *International journal on digital libraries* 23, 4 (2022), 335–369.

[26] Maria Kuteeva and Lisa McGrath. 2015. The theoretical research article as a reflection of disciplinary practices: The case of pure mathematics. *Applied Linguistics* 36, 2 (2015), 215–235.

[27] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models. arXiv:2405.17428 [cs.CL] https://arxiv.org/abs/2405.17428

[28] Xiao Li, Li Sun, Mengjie Ling, and Yan Peng. 2023. A survey of graph neural network based recommendation in social networks. *Neurocomputing* 549 (2023), 126441. doi:10.1016/j.neucom.2023.126441

[29] Kehan Long, Shasha Li, Jintao Tang, and Ting Wang. 2025. Leveraging multiple control codes for aspect-controllable related paper recommendation. *Inf. Process. Manage.* 62, 1 (Jan. 2025), 19 pages. doi:10.1016/j.ipm.2024.103879

[30] Duncan C. McElfresh, Sujay Khandagale, Jonathan Valverde, John Dickerson, and Colin White. 2022. On the Generalizability and Predictability of Recommender Systems. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*. http://papers.nips.cc/paper_files/paper/2022/hash/1c446a652e50b1ea5618b66c07bfc0c5-Abstract-Conference.html

[31] H Mihaljević-Brandt and O Teschke. 2014. Journal Profiles and Beyond: What Makes a Mathematics Journal "General"? *Newsletter of the European Mathematical Society* 91 (2014), 55–56.

[32] Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 2014–2037. doi:10.18653/v1/2023.eacl-main.148

[33] Sheshera Mysore, Tim O'Gorman, Andrew McCallum, and Hamed Zamani. 2021. CSFCube - A Test Collection of Computer Science Research Articles for Faceted Query by Example. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. https://openreview.net/forum?id=8Y50dBbmGU

[34] Malte Ostendorff, Till Blume, Terry Ruas, Bela Gipp, and Georg Rehm. 2022. Specialized Document Embeddings for Aspect-based Similarity of Research Papers. In *2022 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (Cologne, Germany). doi:10.1145/3529372.3530912

[35] Malte Ostendorff, Terry Ruas, Moritz Schubotz, Georg Rehm, and Bela Gipp. 2020. Pairwise Multi-Class Document Classification for Semantic Relations between Wikipedia Articles. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (Virtual Event, China) *(JCDL '20)*. Association for Computing Machinery, New York, NY, USA, 127–136. doi:10.1145/3383583.3398525

[36] Iratxe Pinedo, Mikel Larrañaga, and Ana Arruarte. 2025. Recent Advances and Trends in Research Paper Recommender Systems: A Comprehensive Survey. *arXiv preprint arXiv:2508.08828* (2025).

[37] Santiago Posteguillo. 1999. The Schematic Structure of Computer Science Research Articles. *English for Specific Purposes* 18, 2 (1999), 139–160. doi:10.1016/S0889-4906(98)00001-5

[38] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a General-Purpose Natural Language Processing Task Solver?. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 1339–1384. doi:10.18653/v1/2023.emnlp-main.85

[39] Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL] https://arxiv.org/abs/2412.15115

[40] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (Montreal, Quebec, Canada) *(UAI '09)*. AUAI Press, Arlington, Virginia, USA, 452–461.

[41] Ankit Satpute, Noah Gießing, André Greiner-Petter, Moritz Schubotz, Olaf Teschke, Akiko Aizawa, and Bela Gipp. 2024. Can LLMs Master Math? Investigating Large Language Models on Math Stack Exchange. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2316–2320. doi:10.1145/3626772.3657945

[42] Tim Schopf, Emanuel Gerber, Malte Ostendorff, and Florian Matthes. 2023. AspectCSE: Sentence Embeddings for Aspect-Based Semantic Textual Similarity Using Contrastive Learning and Structured Knowledge. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, Ruslan Mitkov and Galia Angelova (Eds.). INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 1054–1065. https://aclanthology.org/2023.ranlp-1.113/

[43] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (2024).

[44] Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large Language Models for Data Annotation and Synthesis: A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 930–957. doi:10.18653/v1/2024.emnlp-main.54

[45] Gang Wang, Hanru Wang, Jing Liu, and Ying Yang. 2022. Leveraging the fine-grained user preferences with graph neural networks for recommendation. *World Wide Web* 26, 4 (Sept. 2022), 1371–1393. doi:10.1007/s11280-022-01099-y

[46] Gang Wang, Li Zhou, Junqiao Gong, and Xuan Zhang. 2024. Heterogeneous graph neural network with hierarchical attention for group-aware paper recommendation in scientific social networks. *Applied Soft Computing* 167 (2024), 112448. doi:10.1016/j.asoc.2024.112448

[47] Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. 2025. On the theoretical limitations of embedding-based retrieval. *arXiv preprint arXiv:2508.21038* (2025).

[48] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021. Self-supervised Graph Learning for Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 726–735. doi:10.1145/3404835.3462862

[49] Qianqian Xie, Yutao Zhu, Jimin Huang, Pan Du, and Jian-Yun Nie. 2021. Graph Neural Collaborative Topic Model for Citation Recommendation. *ACM Trans. Inf. Syst.* 40, 3, Article 48 (Nov. 2021), 30 pages. doi:10.1145/3473973

[50] 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)*. 682–695. doi:10.1109/ICDE55515.2023.00058

[51] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. *Advances in neural information processing systems* 33 (2020), 5812–5823.

[52] Chenyan Zhang, Shan Xue, Jing Li, Jia Wu, Bo Du, Donghua Liu, and Jun Chang. 2023. Multi-Aspect enhanced Graph Neural Networks for recommendation. *Neural Networks* 157 (2023), 90–102. doi:10.1016/j.neunet.2022.10.001

[53] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176* (2025).

[54] Yang Zhang, Yufei Wang, Quan Z. Sheng, Lina Yao, Haihua Chen, Kai Wang, Adnan Mahmood, Wei Emma Zhang, Munazza Zaib, Subhash Sagar, and Rongying Zhao. 2025. Deep learning meets bibliometrics: A survey of citation function classification. *Journal of Informetrics* 19, 1 (2025), 101608. doi:10.1016/j.joi.2024.101608

[55] Zhiqiang Zhong, Cheng-Te Li, and Jun Pang. 2023. Hierarchical message-passing graph neural networks. *Data Mining and Knowledge Discovery* 37, 1 (2023), 381–408.