Dataset Recommendation for Heterogeneous and Dynamic Scholarly Graphs via Multimodal Representation Learning*

ORNELLA IRRERA, University of Padua, Italy
MATTEO LISSANDRINI, University of Verona, Italy
DANIELE DELL'AGLIO, Aalborg University, Denmark
GIANMARIA SILVELLO, University of Padua, Italy

arXiv:2611.05071v1 [cs.IR] 3 Nov 2026

Researchers struggle to find relevant datasets amid the volume of data across repositories and the growth of scholarly publications. Manual dataset retrieval is time-consuming and often fails to surface valuable resources not easily discoverable through direct links or popularity metrics. Dataset recommendation — identifying datasets relevant to a paper, author, or query — has become essential for improving dataset discoverability and reuse. Yet this task is difficult because Scholarly Knowledge Graphs (SKGs) are large-scale, heterogeneous, dynamic, and incomplete. Although openly available SKGs organize scientific knowledge by linking scholarly outcomes, they often lack meaningful connections between datasets, leaving them isolated and underutilized. These graphs contain millions of evolving relations but still suffer from missing or inaccurate metadata, duplicate records, and weak dataset integration.

This work presents the Multimodal Scholarly Attention Network (MM-SAN), a heterogeneous multimodal graph representation learning method for recommending datasets relevant to a research publication within real-world SKGs. MM-SAN combines topology-aware and text-based embeddings, enabling effective recommendations even when textual metadata is limited or absent. It operates in transductive, semi-inductive, and fully inductive settings, adapting to evolving SKGs without retraining. Experiments show that MM-SAN surpasses state-of-the-art methods, providing a robust, scalable solution for dataset recommendation.

CCS Concepts: • **Information systems** → **Recommender systems**; *Digital libraries and archives.*

Additional Key Words and Phrases: Dataset recommendation, multimodal heterogeneous graph representation learning, scholarly knowledge graphs

1 Introduction

With the continuous growth of scientific data, finding relevant datasets to validate hypotheses, reproduce experiments, and train algorithms has become increasingly crucial. However, the absence of standardized practices for describing and citing datasets often results in poorly documented resources that are either unlinked or only loosely connected to related research outputs. This lack of descriptive metadata and proper documentation makes datasets difficult to discover and significantly limits their impact within the scientific community [5–7, 39, 50].

To improve dataset discoverability, recent methods propose relying on structured and interlinked representations of scholarly information in the form of Scholarly Knowledge Graphs (SKGs) [31]. SKGs provide a rich, structured network of scholarly items, such as papers, datasets and authors, and their relationships. As such, SKGs form a foundational infrastructure that supports a variety of scholarly tasks, including dataset discovery, information retrieval, and knowledge exploration.

SKGs like Microsoft Academic Knowledge Graph (MAKG) [14], Open Research Knowledge Graph (ORKG) [28], and OpenAIRE Graph (OAG) [32], explicitly include some dataset nodes, yet their information is quite scarce and their interconnections with scientific publications are also extremely limited. Datasets are frequently isolated or only loosely linked to related research outputs. Additionally, SKGs are *dynamic*, continuously expanding with new publications and datasets.

*This paper has been accepted as a full paper at *ACM Transactions on Recommender Systems*.

Authors' Contact Information: Ornella Irrera, ornella.irrera@unipd.it, University of Padua, Padua, Italy; Matteo Lissandrini, matteo.lissandrini@univr.it, University of Verona, Verona, Italy; Daniele Dell'Aglio, dade@cs.aau.dk, Aalborg University, Aalborg, Denmark; Gianmaria Silvello, gianmaria.silvello@unipd.it, University of Padua, Padua, Italy.

However, they are affected by *incompleteness* (i.e., publications usually contain rich metadata, whereas datasets often lack sufficient descriptions), and by *sparseness*, since links between datasets and publications are relatively weak or rare. These issues limit the ability of SKGs to represent dataset-related information, thereby making dataset discovery difficult.

In this context, *dataset recommendation*, which involves identifying relevant datasets for a paper, author, or textual query, is crucial for improving dataset discoverability and fostering data reuse. However, developing effective dataset recommendation methods requires addressing the inherent challenges of sparse, heterogeneous, dynamic, and incomplete real-world SKGs. Dataset recommendation has to deal with *data sparsity* because, unlike some settings that rely on extensive user-item interaction data, dataset recommendation involves a limited number of data points, making it difficult to identify meaningful patterns [10]. Additionally, the *lack of user feedback* eliminates a crucial source of information for refining and validating recommendation models. Moreover, dataset recommendation must handle *heterogeneous and incomplete data*. Datasets vary in descriptive metadata quality, and the available information is often inconsistent across different sources, complicating integration and comparison. The diversity in data representations and dimensions further exacerbates the challenge of establishing common ground for analysis.

Finally, dataset recommendation is hindered by a *scarcity of ground-truth labeled data*. Verified information for training and evaluation is often limited, restricting large-scale experiments and necessitating reliance on unsupervised or semi-supervised learning approaches. While these methods can be effective, they generally do not achieve the accuracy and reliability of supervised learning techniques that depend on abundant labeled data.

Addressing these challenges requires novel approaches explicitly tailored to dataset recommendation, leveraging the structure of SKGs while accounting for their limitations.

State-of-the-art dataset recommendation methods typically address these challenges in isolation, making them poorly suited for real-world applications. Most approaches leverage SKGs to rank relevant datasets [3, 9, 48, 49] but are evaluated on homogeneous, static, and specialized SKGs, failing to handle the complexity of real-world graphs. While link prediction on real-world SKGs [26] may appear as a possible approach, it is inherently limited for recommendation tasks, and no multimodal method currently exists to fully exploit heterogeneous, sparse, dynamic, incomplete SKGs.

As a result, dataset recommendation methods struggle with sparse, dynamic graphs and incomplete descriptions. Other methods rely solely on textual metadata from publications and datasets [16, 43], making them ineffective when metadata are unavailable or of low quality.

In this paper, we tackle the task of dataset recommendation, where the goal is to recommend datasets that are relevant to a given research publication. We propose Multimodal Scholarly Attention Network (MM-SAN), a multimodal heterogeneous graph representation method for dataset recommendation, accommodating heterogeneous, sparse, and incomplete real-world SKGs. MM-SAN integrates topology-based and text-based embeddings, exploiting a multimodal learning design, making it effective even when textual features are absent or of low quality. MM-SAN works in transductive and inductive scenarios and is effective whether complete or incomplete textual metadata are available. For evaluation, we rely on subsets of OAG, the backbone of the European Open Science Cloud (EOSC, https://eosc.eu/), which is open-access, frequently updated, and contains scientific datasets. We address three research questions:

RQ1 (Robustness). To what extent can MM-SAN provide effective dataset recommendations in the presence of the topological sparsity typical of real-world SKGs, characterized by limited connectivity between publication and dataset nodes?

RQ2 (Flexibility). To what extent can MM-SAN guarantee effective recommendation performance across transductive, semi-inductive, and fully inductive settings, consequently adapting to new, previously unseen nodes without requiring the model to be retrained?
RQ3 (Versatility). To what extent can MM-SAN preserve effective recommendation performance under progressively reduced textual metadata availability by leveraging graph topology to compensate for missing semantic information?

Regarding the first question, we examine how well MM-SAN performs in the presence of sparse and incomplete graphs, which are typical in real-world SKGs scenarios. The second research question focuses on *inductiveness*, defined as the model's ability to adapt to an evolving graph without requiring full retraining. In other words, the model is trained on an initial version of the graph and evaluated on an updated version containing new nodes and edges not used for training. Hence, we present for the first time a method that is able to generalize across different learning scenarios, i.e., transductive, inductive, and semi-inductive settings. Lastly, the third question focuses on whether the method can effectively exploit graph structure to mitigate the impact of missing or limited textual metadata.

Our main **contributions** are the following:

- A novel multimodal heterogeneous graph representation approach tailored for recommending scientific datasets relevant to research publications, demonstrating strong performance and robustness across transductive, inductive, and semi-inductive scenarios.
- We provide an extensive evaluation curriculum that for the first time spans across a wide range of settings, from metadata-rich to metadata-free environments. This curriculum allows us to show how MM-SAN's architecture achieves superior performance and sets a new standard for works in this area.
- We conduct a thorough comparative analysis of MM-SAN against state-of-the-art graph-based baselines and a newly introduced competitive text-based baseline, evidencing consistent improvements across all tested scenarios.

Outline. In Section 2 we describe the baselines; in Section 3 we formalize the dataset recommendation task; in Section 4 we describe the MM-SAN method; in Sections 5 and 6 we describe the experimental setup and discuss the results. In Section 7 we report an ablation study. In Section 8 we discuss the results and the real-world practical implications. Finally, in Section 9, we draw some conclusions.

2 Related work

In this work, we study methods for dataset recommendation at the intersection of two important areas: SKGs and graph representation learning. First, we present SKGs, which act as the structural backbone for modeling and connecting entities in the scholarly ecosystem, and support tasks like dataset recommendation. This overview helps contextualize the structure and characteristics of the data on which dataset recommendation methods operate. Second, we examine Graph Representation Learning (GRL) methods, which enable the extraction of informative and expressive node embeddings from SKGs. These methods play a crucial role in various downstream tasks, including link prediction, node classification, and recommendation systems, among others. Finally, we review existing approaches that have been applied to the dataset recommendation task.

Scholarly knowledge graphs. SKGs organize and structure scholarly knowledge, providing a means to access, discover and explore outcomes such as publications, datasets, software, authors and venues. Several SKGs, like AMiner [40] and DBLP [1], still focus mainly on representing connections among publications and related authors and do not include dataset nodes.

Conversely, an increasing number of SKGs explicitly model datasets as first-class entities. The most recent example is the European Marine Science (MES) graph [27] which has been released as a dataset-centric subgraph of OAG; thanks to its semi-automatic curation process, it offers a high-quality, reliable ground-truth resource that is particularly suited for evaluating methods on dataset discovery and related tasks. Other prominent examples include MAKG [14], which comprises roughly 200M nodes and more than 10B edges; ORKG, with about 5M nodes and 50M relationships; Data Set Knowledge Graph (DSKG) [15], which links around $2K$ datasets to $635K$ publications and connects them to other Linked Data sources such as MAKG, ORCID, and Wikidata; and OAG [32], the largest of these graphs, with approximately 227M nodes and 15B relationships, openly available and including over 60M datasets.

Existing evaluations of dataset recommendation have largely relied on narrow, ad-hoc collections tailored to metadata-based prediction and recommendation, which in many cases do not constitute full-fledged SKGs. Consequently, in this work we deliberately exclude the following resources:

- the LinearSVM_Dataset, a bipartite graph built from 1,691 dataset titles derived from DSKG [15] and 88K publication abstracts from MAKG, which is specifically engineered for a particular experimental setup;
- the DataFinder dataset [43], which cannot be regarded as a proper SKG because it omits key node types such as authors, venues, and organizations, and instead only provides 17K textual queries paired with 7K relevant datasets;
- the Delve [2] subsets used to evaluate the HVGAE [3] and AMENDER [9] baselines, whose graph structure is unusually dense and therefore unrepresentative of the typically sparse nature of real-world SKGs where datasets are usually connected to at most five publications [25].

Graph representation learning methods. Graph Representation Learning (GRL) provides a general paradigm in which graph-structured data are mapped into low-dimensional continuous vector spaces, enabling the learning of expressive node embeddings that capture both structural and semantic properties of the graph. These representations are then exploited by a wide range of methods to support downstream tasks such as node classification, link prediction, and recommendation.

Graph Convolutional Networks (GCNs)[29] represent one of the foundational architectures for learning node embeddings, where each layer updates a node's representation by propagating and combining information from its neighbors through graph convolutions. For homogeneous graphs, GraphSAGE [22] generalizes this idea by sampling a fixed number of neighbors per node and aggregating their features via different schemes, such as mean, LSTM-based, or pooling aggregators. GAT [42] advances this paradigm by using attention mechanisms to assign adaptive weights to neighboring nodes, leading to more expressive embeddings and strong performance, particularly in inductive learning scenarios.

When dealing with heterogeneous graphs that contain multiple node types, GRL methods must cope with additional complexity. HetGNN [51] tackles this by sampling type-specific neighbors from each node's neighborhood and combining their information through a bi-LSTM-based aggregator. RGCN [38] generalizes GCNs by explicitly modeling edge and node types, thereby enabling more expressive relational representations, while KGCN [44] aggregates neighborhood information in a knowledge graph using biased, user-specific relation weights to capture both structural and semantic relatedness.

Other techniques leverage attention mechanisms to handle heterogeneous graphs better. GATNE [8] tackles multiplex graphs by integrating edge attributes, while HGT [24] uses a transformer-based, type-specific attention mechanism to manage diverse node and edge types. KGAT [46] applies

attentive propagation over a unified collaborative knowledge graph to weigh neighbor contributions by relation-aware importance. HiNormer [33] introduces two encoders – a local structure encoder and a heterogeneous relation encoder – tailored to handle the complexity of heterogeneous graphs. Methods like HAN [47] and MAGNN [18] utilize *metapaths*, which are sequences of node and edge types that define relational patterns. HAN aggregates information from different metapaths, whereas MAGNN performs intra- and inter-metapath aggregation. These baselines have been tested on various heavily cleaned AMiner and DBLP subsets without considering datasets. SAN [26] is another GRL approach tailored to link prediction in real-world SKGs, which performs random-walk-based neighbor sampling and then aggregates the sampled information using an attention mechanism.

Dataset recommendation methods. In recent years, dataset recommendation has garnered significant attention, leading to specialized methods for this task. Most of the proposed approaches focus on recommending scientific datasets that are relevant to research publications. Some approaches focus on leveraging textual metadata from publications and datasets to generate recommendations. For instance, LinearSVM [16] recommends datasets based on a lengthy textual description, such as the abstract of a paper. Another approach introduced a neural bi-encoder model [43] to recommend datasets given a set of publication abstracts. Other approaches focus on graph exploration to identify relevant datasets. Wang et al. [48, 49] combined citation and co-author network exploration in MAKG to recommend both publications and datasets. However, this approach considered a dataset with fewer than 1K datasets, thus the produced output seldom included datasets. HVGAE [3] instead adopts a heterogeneous graph variational autoencoder. Similarly, AMENDER [9] introduces an attentive multi-layered network for recommending datasets to authors. HVGAE and AMENDER have been tested on two subsets of the Delve dataset [2], extracted and processed ad-hoc for these tasks. HetGNN [51] has been used for author-to-venue recommendations, which are more straightforward than dataset recommendations because venues are well-connected in a SKG, providing valuable information that aids recommendations. There are fewer venues than publications and datasets, simplifying the final ranking task. Most dataset recommendation methods have been evaluated only on dense, ad-hoc graphs, which limits their applicability to more complex, real-world SKGs. This is especially problematic since comprehensive textual descriptions of datasets are often unavailable. In [25], we reproduced these methods and found them less effective in real-world scenarios. Our analysis also tested the generalizability of the dataset recommendation task using three popular recommendation baselines (TopPop [12], BPR [37], and LightGCN [23]), underscoring the need for a new, robust, and versatile method that can perform well in real-world settings.

MM-SAN, introduced in this work, is designed to overcome the limitations of current dataset recommendation methods. Unlike existing approaches, MM-SAN is evaluated on real-world scholarly datasets, a setting that has so far been largely overlooked in this domain. Moreover, its ability to operate effectively in inductive scenarios distinguishes it within the broader landscape of graph representation learning. Previous work on real-world SKGs has shown that graph-based models can be effective for the related task of link prediction [26]. Motivated by this evidence, MM-SAN tackles the problem of dataset recommendation by leveraging a multimodal setting tailored to this task.

3 The dataset recommendation task

In what follows, we introduce the formal definitions of SKGs and heterogeneous GRL, and we characterize the dataset recommendation task together with the main open challenges it poses.

Definition 1. *A Scholarly Knowledge Graph (SKG) is a tuple*

$$\mathcal{G}{:}(\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{R}, \mathcal{T}_{\mathcal{V}}, \alpha, \beta, \phi_{\mathcal{V}})$$

where $\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{R}, \mathcal{T}_{\mathcal{V}}$ are the sets of nodes, edges, node types, edge types, and node attributes respectively. Then, $\alpha{:}\mathcal{V}\to\mathcal{A}$ is a node type function that maps each node $v\in\mathcal{V}$ to one node type $a\in\mathcal{A}$, $\beta{:}\mathcal{E}\to\mathcal{R}$ is an edge type function that maps each edge $e\in\mathcal{E}$ to an edge type $r\in\mathcal{R}$, and $\phi_{\mathcal{V}}{:}\mathcal{V}\to\mathcal{T}_{\mathcal{V}}$ is a function that maps each node to a textual attribute $t\in\mathcal{T}_{\mathcal{V}}$. A SKG is heterogeneous *if it has multiple node and edge types such that $|\mathcal{A}| + |\mathcal{R}| > 2$.*

Generally, an SKG incorporates both structural (edges) and semantic (nodes, their types, and their attributes) information, where the attributes are the textual metadata of a scholarly entity. This provides a rich representation for the dataset recommendation task. Its heterogeneous nature adds complexity and allows for modeling the graph's diverse relationships and attributes between various entities.

Definition 2. *Given a heterogeneous SKG $\mathcal{G}$, where $\mathcal{V} = \mathcal{V}_p \cup \mathcal{V}_d$ represents the set of publication nodes $\mathcal{V}_p$ and dataset nodes $\mathcal{V}_d$, and $q\in\mathcal{V}_p$ is the* query node*, the* ***dataset recommendation task*** *aims to learn a scoring function $\rho : \mathcal{V}_p \times \mathcal{V}_d \to \mathbb{R}$. This function assigns a relevance score for each dataset node $d_k\in\mathcal{V}_d$ to the query node q. The task output is a ranked list $\pi_q = (d_{(1)}, d_{(2)}, \ldots, d_{(n)})$, with $n = |\mathcal{V}_d|$, obtained by ordering $\mathcal{V}_d = \{d_1, \ldots, d_n\}$ according to $\rho(q, \cdot)$, such that*

$$\rho(q, d_{(1)}) \geq \rho(q, d_{(2)}) \geq \cdots \geq \rho(q, d_{(n)}).$$

In this work, the dataset recommendation task is defined as identifying and ranking datasets that are most relevant to a specific research publication, which serves as the query.

Definition 3. *Given a heterogeneous SKG $\mathcal{G}$, the objective of GRL is to learn a function $f{:}\mathcal{V}\to\mathbb{R}^d$ that maps each node $v\in\mathcal{V}$ to a d-dimensional embedding space, where $d\ll|\mathcal{V}|$. This embedding is designed such that for any two nodes $u \in \mathcal{V}_p$ and $v \in \mathcal{V}_d$, the dot-product of their corresponding vectors $\vec{u} \cdot \vec{v}$ approximates the scoring function ρ.*

GRL provides a way to map heterogeneous nodes into a shared embedding space, capturing their features and the graph's structure. This is crucial for the dataset recommendation task, where the relationship between query nodes (publications) and potential datasets must be effectively learned and represented in the embedding space.

Challenges. The dataset recommendation task faces significant challenges due to the nature of SKGs. These graphs often contain multiple sparse connected components, duplicated nodes, missing links, and heterogeneous incomplete metadata. Moreover, the extreme class imbalance in ranking exacerbates the task's complexity. In real-world SKGs, the number of datasets can reach millions, but only a small subset is relevant to a given publication [25]. As a result, the task involves accurately retrieving a small set of relevant datasets from a large pool of irrelevant ones.

The first key challenge concerns the *heterogeneity* and *incompleteness* of the metadata available in SKGs. Publications and datasets are typically deposited across different repositories, each adopting its own metadata schema, controlled vocabulary, and level of descriptive detail. As a result, even when a textual attribute is present for a node, its content can vary substantially in structure, length, and informativeness — ranging from a rich abstract to a bare title, or to a short, non-descriptive label such as a repository-generated identifier. This inconsistency is particularly pronounced for datasets: while publications are subject to relatively standardized editorial and peer-review processes that ensure a minimum level of descriptive quality, datasets are often deposited with

minimal or inconsistent documentation. This heterogeneity also complicates the integration of information across sources, as textual descriptions referring to conceptually similar resources may differ substantially in vocabulary, granularity, and completeness, making it difficult to establish a common ground for comparison. Addressing this challenge requires methods that do not depend exclusively on the availability or quality of textual metadata, but can instead complement it with alternative, structurally-derived sources of evidence.

Another challenge in dataset recommendation is the *lack of user feedback*. In conventional recommender systems, relevance signals typically arise from direct user–item interactions, such as clicks, purchases, ratings, or reviews. These interactions are continuously generated and provide abundant, relatively inexpensive supervision for both model training and evaluation. In contrast, publication–dataset relationships in scholarly graphs emerge through a much slower and less systematic process. A dataset is first produced and deposited, usually with associated metadata, and may subsequently be used, mentioned, or formally referenced by a publication. However, the presence of such a reference in a publication does not necessarily result in a publication–dataset link being recorded in the SKG. In some cases, publication–dataset relationships are deposited by researchers; in most cases, however, they are derived through automated extraction procedures applied to the textual content of scientific publications. Whether a dataset mention is successfully identified and converted into a formal graph edge depends on how the dataset is referred to in the text and on the structure and recognizability of the reference. Consequently, a publication may use or mention a relevant dataset while the corresponding relationship remains unrecorded in the graph. The observed graph should therefore be regarded as an incomplete representation of the underlying publication–dataset relationships, potentially containing only a fraction of the links that exist in reality, thereby leading to a high *sparsity* of publication and dataset nodes, and to the *scarcity of ground-truth labeled data for evaluation*.

These aspects affect both learning and evaluation. Since each publication is associated with only a small number of observed datasets, the amount of positive supervision available for learning robust publication–dataset associations is inherently limited. At the same time, the incomplete and sparse nature of the graph introduces a substantial false-negative problem: a dataset predicted by the model may in fact be relevant to a publication, while being treated as incorrect simply because the corresponding relationship has not been extracted or recorded.

This phenomenon is related to, but importantly distinct from, the long-tail problem commonly discussed in recommender systems. In conventional long-tail settings, the majority of interactions are concentrated around a relatively popular and densely connected head, while recommendation becomes progressively more difficult for less frequently interacted users or items. The head nevertheless provides substantial training and evaluation signals that models can exploit. Our setting differs fundamentally in that sparsity is pervasive throughout the SKG. Both publications and datasets are associated, on average, with at most one counterpart on average in the opposite set. Thus, the challenge is not restricted to a relatively small population of tail nodes; rather, the entire publication–dataset graph is characterized by extremely limited observed connectivity. There is consequently no reliably observed head from which models can derive abundant interaction patterns, and the incompleteness of the graph may affect both training and evaluation.

4 MM-SAN

This section presents the MM-SAN architecture illustrated in Figure 1 comprising the augmentation, neighbor sampling, and aggregation phases. Prior work has explored a graph-based model for the link prediction task [26], where the underlying augmentation, sampling, and aggregation pipeline proved effective on real-world scholarly graphs. Motivated by these results, in this work we target the dataset recommendation task: we adopt a similar augmentation and sampling strategy, and

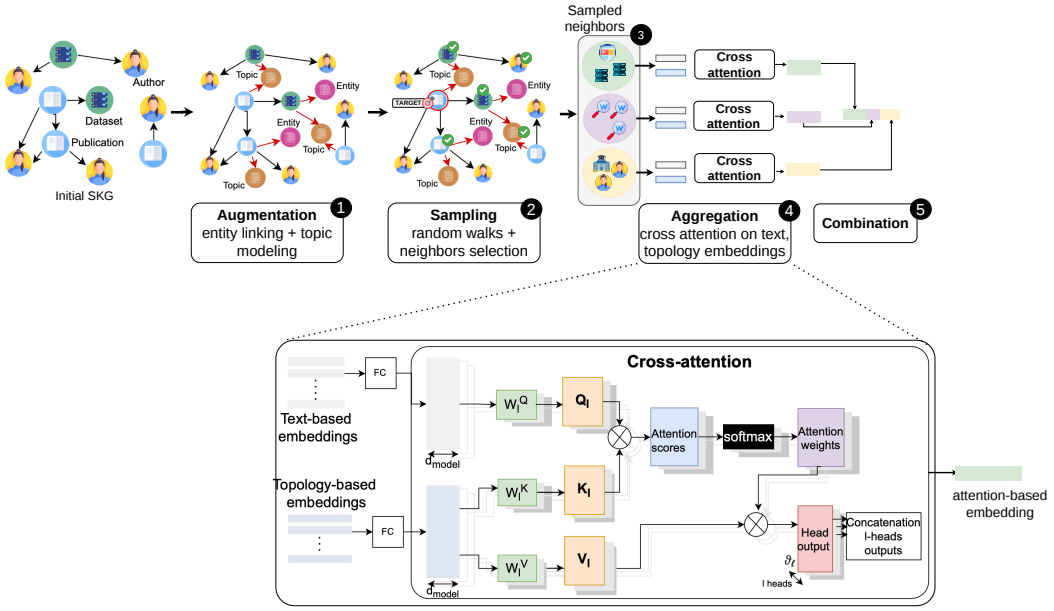


Fig. 1. The MM-SAN Architecture comprising the augmentation, neighbors sampling, and the aggregation phases. ① *Augmentation*: the input SKG is enriched offline with *topic* and *entity* nodes derived from entity linking and topic modeling, to mitigate structural sparsity. ② *Sampling*: starting from a target node, a set of random walks is generated to explore its local neighborhood. ③ *Sampled neighbors*: from the collected walks, MM-SAN selects a representative and type-balanced set of neighbors, grouped into three categories according to their similarity in metadata – (i) publications and datasets, (ii) keywords, topics, and entities, and (iii) authors, venues, and organizations. ④ *Aggregation*: for each of the three groups, a cross-attention module combines the text-based and topology-based embeddings of the sampled neighbors to produce group-specific attention-based embeddings (see the cross-attention detail at the bottom of the figure). ⑤ *Combination*: the three group embeddings are concatenated to form the final representation of the target node.

introduce a novel multimodal aggregation mechanism that explicitly accounts for the contribution of graph topology and textual information, while remaining robust to the absence of either modality.

Augmentation. The augmentation phase (❶) is executed offline to alleviate the sparsity that typically characterizes SKGs, thereby improving *robustness (RQ1)* by injecting additional nodes and edges that densify the graph. Starting from an input SKG, MM-SAN first collects the textual metadata available for *publication* and *dataset* nodes. It then applies entity linking [4], i.e., the process of detecting mentions of real-world entities in text and aligning them with entries in a knowledge base, which helps disambiguate entity mentions. In parallel, topic modeling is used to collect the main themes present in the corpus of textual descriptions. For this, we rely on BERTopic [20], which combines transformer-based embeddings with c-TF-IDF to cluster documents according to their shared topical structure. We selected BERTopic for the topic modeling step because it provides an effective combination of semantic quality, ease of use, and computational efficiency on heterogeneous corpora of varying length. These properties are particularly relevant in our setting, where the augmentation phase must remain scalable to large graphs such as PubMed.

Within this framework, topics capture high-level themes that act as hubs in the SKG, connecting fragments of the graph that would otherwise remain weakly or not at all related. Entities, instead, are fine-grained signals: when multiple nodes mention the same entity, this suggests that they focus on

closely related or overlapping content. Exploiting this complementarity, topics are primarily used to promote global connectivity, while entities emphasize local, content-specific relationships between nodes. Both are then injected into the SKG as new nodes of type *topic* and *entity*, respectively.

After enrichment, MM-SAN derives for each node two distinct representations: a structural embedding learned from the graph topology via `node2vec` [21], and a textual embedding computed from its metadata. For the latter, we employ the `all-MiniLM-L6-v2` sentence transformer[1] for longer textual descriptions and `Phrase-BERT` [45] for shorter texts such as topic labels.

For structural information, we adopt `node2vec`, a well-established and efficient approach for learning topological representations from large graphs. Its random-walk parameters (p, q) provide control over the balance between local and more distant neighborhood exploration, which is useful for our SKGs, characterized by tens of thousands of nodes and highly imbalanced node types.

For textual information, we distinguish between long and short textual attributes. Longer descriptions, including publication and dataset titles and abstracts, are encoded using `all-MiniLM-L6-v2`, a lightweight sentence-transformer that provides a favorable trade-off between semantic representation quality and computational cost. Its widespread adoption for semantic similarity tasks also facilitates reproducibility and comparison with existing approaches. In contrast, short textual attributes, such as topics, entities, and keywords, typically consist of only one to five tokens. These are therefore encoded using `Phrase-BERT`, which is specifically designed to capture the semantics of short phrases and is better suited to this setting than general-purpose sentence encoders optimized for longer textual inputs. We validated these choices by replacing `node2vec` with `DeepWalk` [36] and the transformer-based text encoders with mean-pooled GloVe embeddings [35] (Appendix A.3.1, Table A6). The simpler encoders consistently reduce effectiveness, with the performance gap widening in the inductive setting with partial metadata. This indicates that richer encoders play a key role when limited information is available about the target nodes, enabling them to be represented through more informative and representative embeddings.

Sampling. Starting from a target node v_t in the enriched SKG, phase ❷ focuses on exploring its local neighborhood and selecting a representative set of neighbors across the different node types, while keeping their counts as balanced as possible. A straightforward strategy is to sample neighbors uniformly at random, as done in methods such as GraphSAGE. Yet, in heterogeneous SKGs this often leads to poorly representative samples, since node-type distributions are typically highly skewed. For example, author nodes are frequently under-disambiguated, giving rise to many weakly connected or isolated nodes, whereas venue nodes are fewer but usually more densely connected. As a result, approaches like random sampling tend to over-sample authors and under-sample other crucial node types, like venues, producing a biased and less diverse neighborhood. To mitigate this problem, we adopt a neighbor selection scheme based on random walks.

For a given target node v_t, we generate a set of random walks of length l. We ignore edge directions because for each edge, we assume that there exists an edge in the opposite direction (e.g., for each `cites` edge, we can assume a `cited by` exists in the opposite direction). We collect the k random walks with the highest similarity scores with the target node v_t, computed as follows:

$$\text{sim_score}(v_t, \text{walk}_j) = \frac{1}{m_j} \sum_{i \in \text{walk}_j} \frac{cos(\mathbf{v}_t, \mathbf{v}_i)}{d(v_t, v_i)}$$

where walk_j is one random walk, m_j is the number of publication and dataset nodes in the walk j, $v_i \in \{\mathcal{V}_\text{p} \cup \mathcal{V}_\text{d}\}$ is either a *publication* or a *dataset* node, $\mathbf{v}_t$ and $\mathbf{v}_i$ are the embeddings encoding the textual metadata of v_t and v_i respectively, $cos(\mathbf{v}_\text{t}, \mathbf{v}_i)$ is the cosine similarity defined as $\frac{\mathbf{v}_\text{t} \cdot \mathbf{v}_i}{\|\mathbf{v}_\text{t}\| \|\mathbf{v}_i\|}$, and

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

d is the distance measured as the number of hops between the target v_t and v_i in the walk. This formula ensures that publications and datasets that are both similar and closer are given higher priority than those that are similar but more distant.

In step ❸, MM-SAN selects the top n publications and datasets with the highest cosine similarity to the target v_t from the k selected walks. For all other node types, it selects the n nodes that are closest to v_t within each type. The central concept is that entities and topics are essential for revealing previously disconnected parts of the graph, enabling a more thorough exploration of the target node's neighborhood. In contrast, publication and dataset nodes are important for deciding which walks to prioritize as they provide insights into content similarity.

The resulting set of neighbors can be divided into three groups: (i) one containing publications and datasets, (ii) one containing keywords, topics, and entities, and (iii) one containing authors, venues, and organizations. This classification is based on the similarity of the metadata of the nodes within the same group: publications and datasets share titles and descriptions; keywords, topics, and entities are sets of words that describe a research outcome; and proper nouns often identify authors, venues, and organizations. In the sampling phase, we aim to improve *flexibility (RQ2)* by focusing on a subset of representative neighbors, rather than learning a global representation for the entire neighborhood of a node. This approach enables the model to adapt to dynamic graphs of different sizes without needing global information about all the nodes.

Aggregation. The aggregation step aims to construct the final embedding representation of the target node v_t by aggregating the neighbor vectors. In particular, this step is essential to ensure that the absence or incompleteness of metadata for v_t and its selected neighbors does not affect its final embedding. In this phase, MM-SAN integrates heterogeneous multimodal data to generate robust node embeddings. This step takes inspiration from classical multimodal settings, where different data types, such as text and images, are considered together. Similarly, MM-SAN treats textual and graph topological information as two distinct modalities, each contributing to generating the final node representation.

The aggregation process (❹) ensures that even when metadata is missing or incomplete for a node v_t or its neighbors, the model can still generate a meaningful embedding for v_t. This is achieved by utilizing cross-attention [41] to effectively combine neighbor embeddings based on both textual and topological features. Cross-attention allows the model to integrate information from both modalities; for instance, if a node has sparse connections in the graph, the model can rely more on textual embeddings to enrich the representation. Conversely, if textual metadata is absent, topological information from the graph structure plays a stronger role. This approach mirrors the flexibility found in traditional multimodal settings, where the model integrates diverse data sources to make informed predictions. Hence, this phase enhances *versatility (RQ3)* by tackling the integration of diverse information sources, allowing the model to adapt to various real metadata scenarios.

MM-SAN first linearly projects the text-based and topology-based embeddings (FC blocks in Figure 1). Then, MM-SAN combines three input matrices: Query ($\mathbf{Q}$), Key ($\mathbf{K}$), and Value ($\mathbf{V}$), of sizes (n_Q, d_{model}), (n_K, d_{model}), (n_V, d_{model}) where n_Q, n_K, n_V are the numbers of queries, keys, values embeddings respectively, and d_{model} their dimension. The queries are the linear projected text-based embeddings while keys and values are the linearly projected topology-based embeddings of the selected neighbor nodes. For each attention head $\vartheta_0, \ldots, \vartheta_\ell, \ldots, \vartheta_H$, these matrices are transformed as follows: $\mathbf{Q}_\ell = \mathbf{Q}\mathbf{W}_\ell^Q$, $\mathbf{K}_\ell = \mathbf{K}\mathbf{W}_\ell^K$, and $\mathbf{V}_\ell = \mathbf{V}\mathbf{W}_\ell^V$, where $\mathbf{W}_\ell^Q$, $\mathbf{W}_\ell^K$, and $\mathbf{W}_\ell^V$ are projection matrices of size (d_{model}, d_k) and $d_k = \frac{d_{model}}{H}$ is the dimensionality of the projection space for each head. The embeddings for $\mathbf{Q}$ are based on text, while the embeddings for $\mathbf{K}$ and $\mathbf{V}$ are based on topology. For each attention head ϑ_ℓ the attention scores are computed using the scaled dot-product attention

mechanism, and:

$$\vartheta_\ell = \text{softmax}\left(\frac{\mathbf{Q}_\ell \mathbf{K}_\ell^T}{\sqrt{d_k}}\right)\mathbf{V}_\ell.$$

The outputs from all heads are then concatenated to form the final output:

$$\text{CrossAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\vartheta_0, \ldots, \vartheta_\ell, \ldots, \vartheta_H).$$

The result is a matrix of size (n_Q, d_{model}) combining information from different representation spaces. Cross-attention is applied independently to the three groups of node types: (i) publications and datasets, (ii) keywords, entities, and topics, and (iii) authors, venues, and organizations. The embeddings resulting for each group are finally concatenated (❺) to obtain the final target node embedding representation.

We experimented with various embedding combination methods, including multihead attention [41], bi-LSTM [19], and GRU [11], both as replacements for cross-attention and for the final concatenation step. These results are presented in the ablation study (cf. Sec. 7). However, cross-attention excels at capturing the intricate relationships between topological structures and textual metadata, allowing the model to understand how individual node features interact across different nodes and their descriptions. This capability is especially crucial in heterogeneous graphs, where different types of nodes may have varying levels of importance.

Training. To train the model, we employed a mini-batch gradient descent approach. The optimization was done by minimizing the cross-entropy loss, using negative sampling. The loss function is:

$$\mathcal{L} = -\left(\sum_{(u,v)\in\mathcal{E}_{p,d}} \log\sigma\left(\mathbf{h}_u^\top \mathbf{h}_v\right) + \sum_{(u',v')\notin\mathcal{E}_{p,d}} \log\sigma\left(-\mathbf{h}_{u'}^\top \mathbf{h}_{v'}\right)\right),$$

where σ is the sigmoid function, $\mathbf{h}_u$ and $\mathbf{h}_v$ are the embedding representations for publication u and dataset v, $\mathcal{E}_{\text{p,d}}$ denotes the set of positive edges between publications and datasets. For each positive edge $(u, v)\in\mathcal{E}_{p,d}$, we sample uniformly at random a negative pair $(u', v')\notin\mathcal{E}_{p,d}$, of publications and datasets not connected.

5 Experimental setup

Inspired by the evaluation conducted in [26], which systematically considers realistic conditions of increasing complexity, we evaluate MM-SAN under three learning scenarios of increasing difficulty, distinguished by their level of inductive generalization – i.e., by how much the model must handle nodes that were unseen during training in the SKG. In the **transductive** setting, all publication and dataset nodes are observed during training, the SKG is static, and the model can exploit the full graph structure and all relationships when generating recommendations. This corresponds to the standard setup adopted in prior work [18, 33, 51]. In the **semi-inductive** setting, new publications appear only at test time, while all datasets are already known during training. The task is to rank the existing datasets according to their relevance for these unseen publications. Note that this scenario is often referred to as "inductive" in the literature [51]. Finally, in the **fully inductive** setting, both new publications and new datasets are introduced during testing. This setting mirrors realistic conditions in which the SKG is continuously updated with new research outputs and datasets, and recommendations must be computed for items that were not part of the training graph.

In addition, we assess the robustness of the models to missing textual metadata by considering two metadata configurations for each learning setting. The **ideal metadata setup** represents a best-case scenario in which all datasets are fully documented, i.e., every dataset is associated with both a title and a description. The **real metadata setup** captures more realistic situations where

Table 1. Nodes and edges statistics of the used datasets [26]. $\mathcal{E}_{source,target}$ indicates the set of edges interconnecting the *source* and the *target*. Sparsity is computed over the whole graph as $1 - 2|\mathcal{E}|/(|\mathcal{V}|(|\mathcal{V}| - 1))$, where $|\mathcal{V}|$ and $|\mathcal{E}|$ denote the total number of nodes and edges, respectively; higher values indicate higher sparsity.

Dataset	Sparsity	Nodes		Edges			
MES	99.97%	Publications (p)	2.1K	$\mathcal{E}_{p,p}$	450	$\mathcal{E}_{p,a}$	9.8K
		Datasets (d)	3K	$\mathcal{E}_{d,d}$	1.2K	$\mathcal{E}_{d,a}$	10.2K
		Authors (a)	9.5K	$\mathcal{E}_{p,d}$	2.5K		
PubMed	99.99%	Publications (p)	42.6K	$\mathcal{E}_{p,p}$	18.3K	$\mathcal{E}_{p,k}$	6.6K
		Datasets (d)	33.8K	$\mathcal{E}_{d,d}$	8K	$\mathcal{E}_{d,k}$	17K
		Authors (a)	334K	$\mathcal{E}_{p,d}$	30.5K	$\mathcal{E}_{p,v}$	42.6K
		Venues (v)	6.7K	$\mathcal{E}_{p,a}$	180K	$\mathcal{E}_{p,o}$	87.5K
		Organizations (o)	14.8K	$\mathcal{E}_{d,a}$	94.6K	$\mathcal{E}_{d,o}$	515
		Keywords (k)	10.8K				

only 75%, 50%, or 25% of the datasets have complete metadata, where the datasets with complete metadata are selected at random. This setup reflects typical SKGs, in which datasets may or may not be accompanied by rich, descriptive metadata. The implementation of MM-SAN and the related datasets are available at: https://github.com/OrnellaIrrera/MM-SAN.

Datasets. We employ two SKGs extracted from the OAG [32] and used in previous works [26, 27]: MES [27] and PubMed[2]. MES is a curated scholarly graph linking publications, datasets, and authors, and offers a reliable snapshot of the scientific activity of the MES community within OpenAIRE. PubMed, instead, is a larger subgraph of OAG that connects publications and datasets, as well as authors, venues, organizations, and keywords. In contrast to MES, PubMed is not curated. The two graphs are extremely sparse, with PubMed (99.99%) being sparser than MES (99.97%).

The statistics of the SKGs are presented in Table 1 [26], where we report the sparsity of the two SKGs, along with the number of nodes and edges grouped by type.

We split the set of publication–dataset edges, $\mathcal{E}_{p,d}$, into five disjoint subsets: a training set, a validation set, and three test sets corresponding to the transductive, semi-inductive, and fully inductive scenarios.

To construct the evaluation splits, we start from the complete set of publication–dataset edges. Approximately 10% of these edges are randomly sampled to form the fully inductive test set and a further 10% to form the semi-inductive test set. The two sets are generated so that they involve disjoint collections of publications and datasets, ensuring that no node appears in both inductive test sets. Once the inductive test edges have been selected, they are removed from the original graph. The inductive constraints are then enforced by removing the corresponding nodes from the remaining graph. Specifically, for the semi-inductive setting, all publications involved in the sampled test edges are removed, whereas for the fully inductive setting both the sampled publications and datasets are removed. Node removal is applied to the entire heterogeneous graph, including all incident relations (e.g., publication–author and dataset–author edges), thereby ensuring that no information about the held-out nodes is available during training and preventing information leakage. The validation and transductive test sets are subsequently sampled from the remaining publication–dataset edges. Their sizes are chosen to match those of the semi-inductive and fully inductive test sets, respectively, while all residual publication–dataset edges are assigned to the

[2]MES and PubMed datasets are available at https://doi.org/10.6084/m9.figshare.26661382.v1

training set. Finally, we verify that every publication and dataset appearing in either the validation or transductive test set is represented by at least one edge in the training graph.

On MES, this procedure yields 2K edges for training, 155 for validation, and 155 for each of the three test sets. On PubMed, we obtain 26.8K training edges, 1.7K validation edges, and 1.7K edges for each test set.

For evaluation, we treat the publications appearing in the test-set edges $\mathcal{E}_{p,d}$ as queries and aim to retrieve their associated datasets. The same SKGs were also used in previous work on dataset recommendation [25]. We report standard recommendation metrics – Recall, mean reciprocal rank (MRR), and nDCG – computed at cut-offs $k \in \{5, 10\}$. These small values of k reflect the fact that real-world SKGs typically associate each publication with only a limited number of datasets [25, 26]. Recall measures the fraction of relevant datasets retrieved within the top-k positions, nDCG evaluates ranking quality by accounting for both relevance and position, and MRR captures how early the first relevant dataset appears in the ranked list, assigning higher scores when it is ranked near the top.

Parameters of MM-SAN. We represent each publication and dataset node with semantic vectors obtained from the pre-trained sentence-transformer `all-MiniLM-L6-v2`. Short textual labels – namely keywords, entities, and topics, which typically consist of one to five tokens – are instead encoded using `Phrase-BERT`. We derive structural representations for all nodes by learning topological embeddings with `node2vec` [21].

To enrich the graph with entity information, we run DBpedia Spotlight [34] over the textual metadata of publications and datasets, using a confidence threshold of 0.75 so as to keep only high-precision matches and thus minimize noisy links in the SKG. Topics are discovered with BERTopic [20], enforcing a minimum cluster size of two documents, and fine-tuned using KeyBERT to obtain representative textual labels.

MM-SAN is trained for up to 100 epochs with early stopping, using a learning rate of 10^{-5} and a mini-batch size of 1024. For each target node, we sample five neighbors for each group of nodes, namely: five of type publication or dataset, five of type keyword/topic/entity, and five of type venue/author/organization. The cross-attention module employs 8 attention heads and produces embeddings of dimension 128. The training hyperparameters – number of epochs, learning rate, batch size, number of attention heads, and per-type neighborhood size – are selected via grid search.

For model training, we consider four configurations corresponding to 100%, 75%, 50%, and 25% of available dataset metadata. Across these configurations, the underlying training graph is kept fixed, while the metadata associated with dataset nodes is progressively masked according to the desired availability level. A separate model is trained for each configuration, enabling the model to learn and be evaluated under consistent metadata conditions. Each model is subsequently evaluated under all three recommendation regimes: transductive, semi-inductive and fully inductive. At test time, for the semi-inductive and fully inductive settings, the held-out nodes are reintroduced into the graph together with their non-target incident edges (e.g., publication–author, dataset–venue, publication–publication relations), while the publication–dataset edges under evaluation remain hidden from the sampling procedure to prevent information leakage. For each configuration with incomplete metadata (from 75% to 25% availability), datasets lacking textual metadata are assigned a fixed all-ones vector in place of their textual embedding. This effectively suppresses the textual modality for those datasets while preserving the input structure required by the cross-attention mechanism, allowing the model to continue processing both modalities in a consistent manner. Overall, the experimental design independently controls two factors: *metadata availability* during training and *node novelty* at evaluation time. Metadata availability is varied by training separate

Table 2. Overview of the baselines used for comparison with MM-SAN. For each baseline, we report: name, the type (transductive or inductive), task (general, or dataset recommendation), and a brief description.

Model	Type	Task	Description
GraphSAGE [22]	Inductive	General	Learns node embeddings by sampling and aggregating information from a node's local neighbors, enabling inductive learning on unseen nodes.
GAT [42]	Inductive	General	Employs an attention mechanism to assign different weights to neighboring nodes during message passing.
KGAT [46]	Inductive	General	Combines the user-item bipartite graph with a knowledge graph into a unified collaborative knowledge graph, applying attentive propagation with relation-aware weights.
HAN [47]	Inductive	General	Applies hierarchical attention – at the node level and across semantic types of relations – to capture diverse relational patterns in multi-type networks.
HGT [24]	Inductive	General	Transformer-based model that incorporates type-specific parameters and relational attention to model complex interactions across node and edge types.
HetGNN (HGNN) [51]	Inductive	General	Sampling-and-aggregation architecture integrating random-walk–based sampling with bi-LSTM and self-attention aggregation.
ST-T	Transductive/Semi	Dataset recsys	Text-based baseline computed as the dot product between pre-trained sentence-transformer embeddings of a publication and a dataset ST-T$(u, v) = \mathbf{v}_u \cdot \mathbf{v}_v$ where $\mathbf{v}_u$ and $\mathbf{v}_v$ are the embeddings of the publication and dataset respectively.
KGCN [44]	Transductive	General	Extends GCNs to knowledge graphs for recommendation, aggregating neighborhood information with biased weights learned from user-specific relation scores.
MetaPath2Vec (MP2Vec) [13]	Transductive	General	Skip-gram-based representation learning method that performs metapath-guided random walks over heterogeneous networks.
RGCN [38]	Transductive	General	Relational extension of GCNs that explicitly models different edge and node types through relation-specific transformation matrices.
VGAE [30]	Transductive	General	Variational graph autoencoder that learns latent node embeddings by reconstructing the adjacency matrix, without exploiting node or edge type information.
HVGAE [3]	Transductive	Dataset recsys	Heterogeneous graph variational autoencoder tailored for dataset recommendation, extending VGAE to multi-type scholarly networks.
LinearSVM (LSVM) [16]	Transductive	Dataset recsys	Text-based baseline that recommends datasets relying on lengthy textual descriptions, such as publication abstracts, without exploiting graph structure.
DataFinder (DF) [43]	Transductive	Dataset recsys	Neural bi-encoder model that recommends datasets given a set of publication abstracts, matching textual queries to dataset descriptions.
EnsembleCN (Ens) [48]	Transductive	Dataset recsys	Ensemble method that recommends scientific datasets by exploiting author and citation networks.

models under different metadata-masking configurations while keeping the graph structure fixed. Conversely, node novelty is varied only at evaluation time by assessing the same trained model under transductive, semi-inductive, and fully inductive conditions. This design allows us to disentangle the impact of missing metadata from the challenges introduced by inductive generalization.

Baselines. We compare MM-SAN against: (i) six general-purpose inductive graph-based baselines (also used in [24, 26, 47, 51]), (ii) eight transductive baselines, and (iii) one text-based baseline. We report an overview of the baselines in Table 2, where, for each baseline, we report the learning setting it operates in (transductive or inductive), the task it was originally designed for (general-purpose graph representation learning or dataset recommendation), and a brief description of its underlying approach.

For each baseline method, we report the best-performing configuration selected via grid search. The hyperparameter ranges explored for each baseline are summarized in the appendix. It is important to note that most of these baselines were not originally conceived for the dataset recommendation problem. Instead, they are general-purpose graph representation models designed

to operate on attributed graphs. Evaluating MM-SAN against such widely used, domain-agnostic approaches allows us to more clearly position its contribution within the broader GRL landscape.

For all models, node features are initialized by concatenating textual embeddings from the pre-trained Sentence Transformer `all-MiniLM-L6-v2` with structural embeddings learned via `node2vec`. GraphSAGE [22] (SAGE) and GAT [42] are trained for 200 epochs using a learning rate of 10^{-5} and early stopping, with an embedding size of 128. Similarly, HAN [47] and HGT [24] are optimized for 200 epochs with a learning rate of 10^{-4}, also with early stopping and a 128-dimensional output space. HetGNN [51] (HGNN) is trained for 100 epochs with early stopping, a learning rate of 10^{-5}, and 10 sampled neighbors per node type, following the original setup. KGAT was trained for up to 150 epochs with early stopping, using a learning rate of 10^{-4}, 2 layers, and an embedding dimension of 128. For what concerns the transductive methods, RGCN was trained for up to 150 epochs with early stopping, using a learning rate of 3×10^{-4}, 2 layers, a hidden dimension of 128, and a dropout rate of 0.10. KGCN followed the same training budget (150 epochs, patience 20), with a learning rate of 1×10^{-4}, 2 layers, a hidden dimension of 128, dropout of 0.10. VGAE was optimized for up to 150 epochs with early stopping, a learning rate of 5×10^{-3}, 2 relational graph convolutional layers with 16 bases, a latent dimension of 128, and a KL-divergence term linearly warmed up over 20 epochs to a final weight of 0.001. HVGAE shared the same training schedule (150 epochs, patience 20), using a learning rate of 1×10^{-3}, 2 layers, a hidden dimension of 128 projected onto a 16-dimensional latent space, and an analogous KL warm-up (30 epochs, final weight of 0.001) applied independently to the publication and dataset encoders. Across all the selected models, negative sampling used 5 negatives per positive edge, and parameters were optimized with AdamW using a weight decay of 1×10^{-5}. All baselines have been implemented using PyTorch Geometric [17].

For the text-only baseline, we represent publications and datasets using embeddings derived from the pre-trained Sentence Transformer `all-MiniLM-L6-v2`. This baseline can be instantiated only when (i) all datasets are present at training time and (ii) every dataset is associated with full textual metadata. These conditions are met in the transductive and semi-inductive scenarios with complete metadata, since ranking scores can be computed only if the entire candidate set of datasets and their textual descriptions are available.

Several existing methods are not included in our comparison. GATNE [8] is omitted because it is tailored to settings with multiple edge types over a single set of nodes and relies on user-specific information, whereas our task involves recommending datasets to publications rather than to users. MAGNN [18] is excluded as it was originally developed for node classification and does not naturally extend to inductive evaluation protocols. Similarly, HiNormer [33] is not considered, since it is not designed for recommendation tasks.

6 Results

This section presents the results separately for the inductive and transductive baselines. To assess the statistical significance of our results, we conducted paired t-tests ($\alpha = 0.05$) on a per-query basis, comparing MM-SAN against each baseline. All differences were statistically significant ($p < 0.05$).

In Appendix A.3.3, we further extend our analysis by evaluating MM-SAN on dataset recommendation for authors, a more challenging variant of the task in which authors, rather than publications, serve as query nodes.

Inductive results. Tables 3, 4 report the results on MES and PubMed, respectively, at cutoffs 5 and 10, across the three evaluation settings: transductive, semi-inductive, and fully inductive. Each condition is assessed under progressively reduced proportions of datasets with available metadata: 100%, 75%, 50%, and 25%. Concerning the MES dataset, MM-SAN demonstrates strong

Table 3. Recall@k (R@k), nDCG@k (N@k) and Mean Reciprocal Rank@k (MRR@k), with $k \in \{5, 10\}$, on MES in transductive (Tran), semi-inductive (Semi), and inductive (Ind) settings, and with $\{100, 75, 50, 25\}$% of datasets with available metadata. Best results are in **bold**, second-best are underlined.

			MES $k = 5$								**MES** $k = 10$							
Metadata %	Setting	Metric	GAT	KGAT	SAGE	HGT	HAN	HGNN	ST-T	MM-SAN	GAT	KGAT	SAGE	HGT	HAN	HGNN	ST-T	MM-SAN
100%	Tran	R@k	0.625	0.136	0.489	0.023	0.010	0.581	0.693	**0.727**	0.748	0.211	0.571	0.037	0.019	0.656	0.721	**0.795**
		N@k	0.475	0.089	0.398	0.020	0.009	0.438	**0.651**	0.546	0.515	0.113	0.424	0.030	0.011	0.455	**0.661**	0.565
		MRR@k	0.462	0.073	0.448	0.020	0.009	0.378	0.470	**0.484**	0.503	0.083	0.469	0.031	0.009	0.392	0.493	**0.515**
	Semi	R@k	0.621	0.106	0.446	0.007	0.000	0.587	0.696	**0.719**	0.719	0.129	0.522	0.027	0.006	0.655	0.753	**0.799**
		N@k	0.479	0.066	0.316	0.003	0.000	0.417	**0.644**	0.524	0.484	0.074	0.329	0.010	0.004	0.438	**0.663**	0.551
		MRR@k	0.454	0.053	0.437	0.006	0.000	0.362	0.450	**0.461**	0.463	0.057	0.485	0.007	0.005	0.370	0.472	**0.499**
	Ind	R@k	0.590	0.072	0.439	0.000	0.000	0.582	-	**0.687**	0.707	0.112	0.509	0.000	0.000	0.657	-	**0.743**
		N@k	0.442	0.048	0.315	0.000	0.000	0.422	-	**0.519**	0.506	0.062	0.341	0.000	0.000	0.447	-	**0.537**
		MRR@k	0.434	0.041	0.417	0.000	0.000	0.370	-	**0.466**	0.449	0.047	0.447	0.000	0.000	0.379	-	**0.492**
75%	Tran	R@k	0.481	0.020	0.452	0.022	0.008	0.473	-	**0.623**	0.560	0.034	0.527	0.032	0.010	0.578	-	**0.759**
		N@k	0.370	0.014	0.294	0.011	0.006	0.356	-	**0.447**	0.396	0.019	0.317	0.021	0.007	0.400	-	**0.493**
		MRR@k	0.391	0.012	0.421	0.014	0.003	0.300	-	**0.495**	0.415	0.014	0.464	0.019	0.007	0.349	-	**0.508**
	Semi	R@k	0.483	0.030	0.432	0.000	0.000	0.515	-	**0.598**	0.536	0.061	0.491	0.018	0.001	0.571	-	**0.704**
		N@k	0.404	0.018	0.247	0.000	0.000	0.373	-	**0.468**	0.421	0.029	0.272	0.006	0.001	0.409	-	**0.478**
		MRR@k	0.383	0.015	0.398	0.000	0.000	0.283	-	**0.428**	0.409	0.019	0.408	0.010	0.000	0.295	-	**0.481**
	Ind	R@k	0.409	0.026	0.385	0.000	0.000	0.423	-	**0.598**	0.496	0.046	0.416	0.000	0.000	0.541	-	**0.726**
		N@k	0.324	0.015	0.235	0.000	0.000	0.383	-	**0.441**	0.354	0.022	0.245	0.000	0.000	0.408	-	**0.511**
		MRR@k	0.313	0.012	0.339	0.000	0.000	0.336	-	**0.398**	0.344	0.014	0.351	0.000	0.000	0.339	-	**0.403**
50%	Tran	R@k	0.361	0.014	0.286	0.015	0.007	0.443	-	**0.539**	0.417	0.020	0.338	0.022	0.008	0.560	-	**0.667**
		N@k	0.319	0.005	0.212	0.012	0.006	0.324	-	**0.398**	0.334	0.008	0.288	0.021	0.008	0.377	-	**0.418**
		MRR@k	0.311	0.003	0.237	0.008	0.005	0.276	-	**0.340**	0.343	0.004	0.306	0.012	0.006	0.313	-	**0.404**
	Semi	R@k	0.349	0.008	0.240	0.000	0.000	0.484	-	**0.518**	0.396	0.015	0.294	0.010	0.003	0.519	-	**0.700**
		N@k	0.277	0.005	0.209	0.000	0.000	0.352	-	**0.359**	0.295	0.007	0.226	0.004	0.001	0.367	-	**0.441**
		MRR@k	0.302	0.004	0.203	0.000	0.000	0.253	-	**0.323**	0.313	0.005	0.244	0.002	0.002	0.325	-	**0.382**
	Ind	R@k	0.321	0.007	0.196	0.000	0.000	0.434	-	**0.526**	0.371	0.013	0.234	0.000	0.000	0.482	-	**0.598**
		N@k	0.258	0.004	0.150	0.000	0.000	0.339	-	**0.391**	0.274	0.006	0.166	0.000	0.000	0.357	-	**0.415**
		MRR@k	0.273	0.003	0.181	0.000	0.000	0.304	-	**0.326**	0.298	0.004	0.221	0.000	0.000	0.323	-	**0.345**
25%	Tran	R@k	0.230	0.004	0.190	0.007	0.003	0.336	-	**0.467**	0.256	0.020	0.210	0.015	0.005	0.480	-	**0.599**
		N@k	0.203	0.002	0.165	0.007	0.002	0.242	-	**0.329**	0.212	0.012	0.171	0.009	0.004	0.265	-	**0.372**
		MRR@k	0.210	0.002	0.160	0.005	0.003	0.169	-	**0.283**	0.239	0.009	0.189	0.007	0.004	0.181	-	**0.340**
	Semi	R@k	0.206	0.003	0.145	0.000	0.000	0.367	-	**0.454**	0.213	0.005	0.172	0.005	0.002	0.443	-	**0.606**
		N@k	0.172	0.001	0.125	0.000	0.000	0.270	-	**0.306**	0.174	0.002	0.135	0.002	0.002	0.295	-	**0.354**
		MRR@k	0.177	0.001	0.081	0.000	0.000	0.247	-	**0.256**	0.180	0.002	0.107	0.000	0.001	0.259	-	**0.275**
	Ind	R@k	0.159	0.007	0.095	0.000	0.000	0.322	-	**0.453**	0.193	0.008	0.102	0.000	0.000	0.394	-	**0.539**
		N@k	0.128	0.004	0.069	0.000	0.000	0.233	-	**0.321**	0.139	0.005	0.075	0.000	0.000	0.257	-	**0.348**
		MRR@k	0.136	0.003	0.070	0.000	0.000	0.194	-	**0.241**	0.147	0.004	0.098	0.000	0.000	0.216	-	**0.255**

performance, outperforming all graph-based baselines across all scenarios. Overall, the models achieve their highest scores when full metadata is available (100% rows in Table 3). However, this scenario is the least representative of real-world academic contexts. MM-SAN preserves strong performance even in the more demanding semi-inductive and fully inductive settings, where predictions must be generated for previously unseen nodes, and remains the best graph-based model in all three settings. Moreover, when metadata availability is reduced (from 75% to 25% of the datasets), MM-SAN maintains the highest effectiveness across all settings and metrics, indicating greater robustness to limited metadata. When the amount of available metadata decreases, the shift from the transductive to the semi- and fully inductive settings has a much stronger negative impact on the performance of the homogeneous baselines, such as GAT and SAGE. This trend underscores the importance of text-based features in enhancing recommendation effectiveness, reinforcing MM-SAN's adaptability in handling sparse metadata scenarios. MM-SAN is the only model that consistently achieves Recall@5 above 0.450, Recall@10 above 0.530, and nDCG@5 and nDCG@10 above 0.300 and 0.345 respectively, even when descriptive textual metadata is available for only 25% of the nodes. In the most challenging scenario – the fully inductive setting with 25% metadata – MM-SAN improves over the strongest baseline (HGNN) by 41% in Recall@5, 38% in nDCG@5, 37% in Recall@10, and 35% in nDCG@10. This indicates that the MM-SAN multimodal approach leveraging topology-based features successfully reduces the effect of missing text-based features, a condition that significantly affects the performance of other baselines. HGNN, which is designed explicitly for heterogeneous SKGs, emerges as the second top-performing model after MM-SAN in

Table 4. Recall@k (R@k), nDCG@k (N@k) and Mean Reciprocal Rank@k (MRR@k), with $k \in \{5, 10\}$, on PubMed in transductive (Tran), semi-inductive (Semi), and inductive (Ind) settings, and with $\{100, 75, 50, 25\}$% of datasets with available metadata. Best results are in **bold**, second-best are underlined.

			PubMed $k = 5$								**PubMed** $k = 10$							
Metadata %	Setting	Metric	GAT	KGAT	SAGE	HGT	HAN	HGNN	ST-T	MM-SAN	GAT	KGAT	SAGE	HGT	HAN	HGNN	ST-T	MM-SAN
100%	Tran	R@k	0.241	0.021	0.239	0.016	0.030	0.284	**0.590**	0.380	0.304	0.063	0.308	0.026	0.043	0.341	**0.639**	0.465
		N@k	0.181	0.011	0.179	0.015	0.023	0.202	**0.523**	0.272	0.201	0.022	0.210	0.019	0.032	0.223	**0.539**	0.297
		MRR@k	0.203	0.011	0.204	0.013	0.024	0.209	**0.456**	0.234	0.205	0.021	0.235	0.015	0.031	0.287	**0.465**	0.241
	Semi	R@k	0.231	0.020	0.221	0.006	0.010	0.246	**0.596**	0.373	0.303	0.052	0.292	0.007	0.021	0.301	**0.639**	0.441
		N@k	0.171	0.010	0.169	0.005	0.008	0.177	**0.527**	0.266	0.194	0.031	0.192	0.007	0.011	0.196	**0.541**	0.289
		MRR@k	0.202	0.010	0.204	0.006	0.010	0.222	**0.450**	0.232	0.229	0.023	0.219	0.006	0.012	0.285	**0.462**	0.252
	Ind	R@k	0.196	0.013	0.207	0.000	0.000	0.271	-	**0.344**	0.265	0.018	0.252	0.003	0.002	0.334	-	**0.425**
		N@k	0.134	0.012	0.133	0.000	0.000	0.199	-	**0.246**	0.156	0.015	0.154	0.002	0.001	0.220	-	**0.272**
		MRR@k	0.182	0.011	0.193	0.000	0.000	0.187	-	**0.214**	0.184	0.012	0.215	0.002	0.001	0.193	-	**0.225**
75%	Tran	R@k	0.201	0.018	0.183	0.006	0.012	0.241	-	**0.319**	0.251	0.030	0.237	0.008	0.023	0.301	-	**0.402**
		N@k	0.149	0.013	0.139	0.005	0.009	0.171	-	**0.227**	0.166	0.020	0.157	0.007	0.012	0.191	-	**0.254**
		MRR@k	0.176	0.011	0.166	0.005	0.007	0.149	-	**0.197**	0.192	0.015	0.190	0.007	0.010	0.156	-	**0.209**
	Semi	R@k	0.186	0.012	0.174	0.000	0.007	0.237	-	**0.319**	0.250	0.022	0.227	0.004	0.009	0.293	-	**0.381**
		N@k	0.140	0.010	0.133	0.000	0.007	0.174	-	**0.222**	0.161	0.016	0.150	0.003	0.009	0.192	-	**0.244**
		MRR@k	0.138	0.010	0.168	0.000	0.006	0.150	-	**0.199**	0.187	0.011	0.174	0.003	0.007	0.157	-	**0.209**
	Ind	R@k	0.145	0.009	0.136	0.000	0.000	0.202	-	**0.289**	0.192	0.012	0.186	0.000	0.002	0.254	-	**0.360**
		N@k	0.104	0.005	0.097	0.000	0.000	0.145	-	**0.205**	0.119	0.010	0.115	0.000	0.002	0.162	-	**0.228**
		MRR@k	0.112	0.008	0.120	0.000	0.000	0.135	-	**0.183**	0.138	0.010	0.127	0.000	0.001	0.141	-	**0.192**
50%	Tran	R@k	0.156	0.016	0.145	0.005	0.007	0.203	-	**0.276**	0.196	0.022	0.174	0.006	0.013	0.253	-	**0.365**
		N@k	0.116	0.009	0.111	0.005	0.007	0.140	-	**0.194**	0.129	0.018	0.120	0.006	0.012	0.156	-	**0.223**
		MRR@k	0.115	0.010	0.123	0.004	0.006	0.131	-	**0.166**	0.163	0.016	0.138	0.005	0.011	0.150	-	**0.196**
	Semi	R@k	0.144	0.013	0.129	0.000	0.001	0.201	-	**0.259**	0.185	0.025	0.167	0.003	0.002	0.257	-	**0.342**
		N@k	0.110	0.010	0.100	0.000	0.001	0.142	-	**0.181**	0.123	0.021	0.112	0.002	0.002	0.151	-	**0.208**
		MRR@k	0.104	0.011	0.110	0.000	0.001	0.124	-	**0.164**	0.130	0.020	0.115	0.002	0.002	0.137	-	**0.174**
	Ind	R@k	0.096	0.012	0.097	0.000	0.000	0.167	-	**0.242**	0.131	0.019	0.126	0.000	0.001	0.213	-	**0.312**
		N@k	0.086	0.009	0.079	0.000	0.000	0.120	-	**0.169**	0.098	0.015	0.089	0.000	0.001	0.135	-	**0.192**
		MRR@k	0.089	0.010	0.088	0.000	0.000	0.117	-	**0.121**	0.114	0.011	0.094	0.000	0.001	0.124	-	**0.155**
25%	Tran	R@k	0.113	0.008	0.108	0.000	0.000	0.156	-	**0.230**	0.138	0.012	0.120	0.006	0.002	0.208	-	**0.306**
		N@k	0.084	0.006	0.089	0.000	0.000	0.108	-	**0.155**	0.092	0.009	0.096	0.006	0.001	0.122	-	**0.180**
		MRR@k	0.103	0.007	0.100	0.000	0.000	0.122	-	**0.162**	0.122	0.010	0.111	0.003	0.001	0.137	-	**0.195**
	Semi	R@k	0.098	0.005	0.096	0.000	0.000	0.152	-	**0.239**	0.123	0.009	0.115	0.002	0.002	0.204	-	**0.307**
		N@k	0.077	0.002	0.075	0.000	0.000	0.107	-	**0.167**	0.085	0.003	0.081	0.001	0.002	0.120	-	**0.169**
		MRR@k	0.085	0.002	0.095	0.000	0.000	0.120	-	**0.169**	0.103	0.004	0.105	0.000	0.002	0.125	-	**0.188**
	Ind	R@k	0.053	0.001	0.057	0.000	0.000	0.129	-	**0.205**	0.065	0.001	0.073	0.000	0.000	0.176	-	**0.278**
		N@k	0.038	0.001	0.043	0.000	0.000	0.090	-	**0.138**	0.041	0.001	0.048	0.000	0.000	0.106	-	**0.162**
		MRR@k	0.030	0.001	0.025	0.000	0.000	0.106	-	**0.162**	0.050	0.001	0.052	0.000	0.000	0.110	-	**0.170**

terms of Recall and nDCG in the fully inductive setting when 75%, 50%, and 25% of metadata are available. We observe that SAGE and GAT – two GRL methods designed for homogeneous graphs – when metadata is fully available, remain competitive with, and in the case of MES even surpass, HGNN, which is specifically designed for heterogeneous attributed graphs. On the other hand, they experience a substantial performance drop as the proportion of available metadata decreases. This behavior can be attributed to the ability of convolution-based message passing to fully leverage the semantic richness of node representations when all feature dimensions are available. However, once the feature space becomes sparse or incomplete, these methods face a substantial performance decline. The extremely sparse and imbalanced heterogeneous structure of the graph prevents HAN and HGT from learning stable relation-aware representations, which results in severe performance degradation – particularly in inductive settings where unseen nodes cannot rely on pre-learned type-specific parameters. The same holds for KGAT, which achieves Recall@5 higher than 0.1 only when all the metadata are available. KGAT's attentive propagation relies on relation-specific embeddings to weigh the contribution of each neighbor. Reliably estimating these relation embeddings requires a sufficient number of observed triples per relation type. In our setting, the paper–dataset relation is extremely sparse: this scarcity limits how well the corresponding relation embedding, and consequently the attention weights computed from it, can be learned. While the text-based ST-T baseline achieves higher nDCG in the transductive and semi-inductive setups, MM-SAN obtains higher Recall and MRR in both settings, showcasing its robustness across different conditions on the MES dataset.

As shown in Table 4, MM-SAN exhibits strong robustness when applied to real-world, sparse, an large SKGs like PubMed. Similar to the MES dataset, all tested methods experience a performance decline in semi-inductive and fully inductive settings as textual metadata becomes less available – a natural outcome given the increasing difficulty of the task. Compared to MES, performance on PubMed in terms of Recall, nDCG and MRR is noticeably lower. PubMed is considerably more sparse, authors are not disambiguated – further increasing sparsity – it is not manually curated, and its metadata is of lower quality. These factors collectively and directly degrade the overall model performance. MM-SAN demonstrates the highest resilience to reduced textual metadata and is the best graph-based model in all settings, with the only exception of MRR@10 with full metadata in the transductive and semi-inductive settings, where HGNN performs better. It is the only method that consistently maintains a Recall@5 above 0.200, Recall@10 above 0.275, an nDCG@5 above 0.135, and an nDCG@10 at least 0.162 across all settings. Also, in the most challenging scenario, with only 25% of datasets containing available metadata and within a full-inductive setup, MM-SAN surpasses the best baseline (HGNN) by 59% in Recall@5, 58% in Recall@10, and 53% in nDCG@5 and nDCG@10. Consistently with the results on the MES dataset, SAGE and GAT, designed for homogeneous graphs, rank among the strongest baselines after HGNN, achieving their best performance in the ideal scenario with 100% of metadata available. KGAT, HAN and HGT exhibit the lowest performance on the PubMed dataset, highlighting their limited effectiveness in real-world SKGs. The text-based baseline, ST-T, achieves the highest effectiveness in transductive and semi-inductive scenarios with full metadata availability, surpassing all graph-based methods (including MM-SAN) in Recall, nDCG and MRR. However, ST-T's effectiveness is limited to ideal conditions, making it less suited to addressing the open challenges of real-world SKGs.

Transductive results. In Table 5, we compare the performance of MM-SAN against a set of baselines specifically designed for purely transductive recommendation. We report the results on the MES and PubMed datasets at a cutoff of 5 under the transductive setup, considering progressively decreasing proportions of available metadata. MetaPath2Vec and LinearSVM are reported only at 100% metadata, as neither uses dataset textual descriptions: MetaPath2Vec relies purely on graph structure, and LinearSVM only on publication text. Their results are therefore unaffected by dataset metadata removal. RGCN, KGCN, and HVGAE all perform poorly (R@5 < 0.1 on MES, < 0.01 on PubMed). RGCN's relation-specific weights cannot be reliably learned from so few publication–dataset edges (2K on MES, 27K on PubMed). KGCN's neighbor weighting depends on a user-specific interaction history, but each publication-query is linked to fewer than one dataset on average; the model thus degenerates to near-uniform aggregation. HVGAE's dataset encoder relies solely on the (mostly empty) bipartite row of paper–dataset links, with no textual features to fall back on, so it cannot distinguish datasets under such extreme sparsity. Its GraphSAGE paper encoder finds almost no neighbors to aggregate on MES or PubMed and effectively collapses to encoding each paper in isolation. EnsembleCN (Ens) performs moderately on MES (R@5 = 0.312 at 100%, 0.010 at 25%) but fails on PubMed (R@5 ≤ 0.005). It relies on co-author and citation network exploration, which is only viable on MES, where authors are disambiguated and consequently well-connected; on PubMed, undisambiguated authors vastly outnumber papers, fragmenting the co-authorship network. VGAE follows the same pattern (R@5 = 0.414$\rightarrow$0.292 on MES; ≤ 0.029 on PubMed). Lacking type information, VGAE treats all nodes and edges as homogeneous, so its embeddings depend purely on local neighborhood density rather than node/edge semantics. This proxy is informative enough on the smaller, curated MES graph, but breaks down on the larger, sparser PubMed graph, where local density becomes too weak and noisy to distinguish nodes meaningfully. DataFinder (DF), a purely text-based bi-encoder, is the strongest transductive baseline at full metadata (R@5 = 0.420 on MES, 0.280 on PubMed), confirming the value of semantic

Table 5. Recall@k (R@k), nDCG@k (N@k), and Mean Reciprocal Rank@k (MRR@k) with $k = 5$ on MES and PubMed for the purely transductive methods and MM-SAN, with $\{100, 75, 50, 25\}$% of datasets with available metadata. Best results are in **bold**, second-best are underlined. LSVM=LinearSVM, DF=DataFinder, Ens=EnsembleCN, MP2Vec=MetaPath2Vec.

		MES $k = 5$								
Metadata %	Metric	LSVM	DF	Ens	RGCN	KGCN	HVGAE	VGAE	MP2Vec	MM-SAN
100%	R@k	0.095	0.420	0.312	0.085	0.027	0.007	0.414	0.032	**0.727**
	N@k	0.076	0.322	0.129	0.054	0.009	0.004	0.330	0.021	**0.546**
	MRR@k	0.073	0.310	0.229	0.046	0.009	0.003	0.313	0.018	**0.484**
75%	R@k	-	0.325	0.120	0.048	0.020	0.007	0.308	-	**0.623**
	N@k	-	0.312	0.100	0.023	0.012	0.003	0.249	-	**0.447**
	MRR@k	-	0.301	0.140	0.016	0.004	0.002	0.224	-	**0.495**
50%	R@k	-	0.412	0.090	0.014	0.014	0.003	0.333	-	**0.539**
	N@k	-	0.302	0.076	0.005	0.005	0.003	0.287	-	**0.398**
	MRR@k	-	0.298	0.101	0.002	0.003	0.002	0.250	-	**0.340**
25%	R@k	-	0.299	0.010	0.008	0.007	0.002	0.292	-	**0.467**
	N@k	-	0.198	0.008	0.003	0.003	0.002	0.232	-	**0.329**
	MRR@k	-	0.229	0.012	0.006	0.004	0.001	0.218	-	**0.283**
		PubMed $k = 5$								
Metadata %	Metric	LSVM	DF	Ens	RGCN	KGCN	HVGAE	VGAE	MP2Vec	MM-SAN
100%	R@k	0.074	0.280	0.005	0.001	0.006	0.001	0.024	0.061	**0.380**
	N@k	0.060	0.210	0.002	0.001	0.003	0.001	0.012	0.040	**0.272**
	MRR@k	0.057	0.121	0.005	0.001	0.002	0.001	0.008	0.033	**0.234**
75%	R@k	-	0.129	0.003	0.001	0.001	0.000	0.029	-	**0.319**
	N@k	-	0.108	0.001	0.000	0.001	0.000	0.015	-	**0.227**
	MRR@k	-	0.103	0.003	0.000	0.000	0.000	0.010	-	**0.197**
50%	R@k	-	0.109	0.001	0.001	0.001	0.000	0.018	-	**0.276**
	N@k	-	0.090	0.001	0.001	0.003	0.000	0.009	-	**0.194**
	MRR@k	-	0.080	0.002	0.001	0.000	0.000	0.007	-	**0.166**
25%	R@k	-	0.092	0.000	0.001	0.000	0.000	0.012	-	**0.230**
	N@k	-	0.072	0.000	0.000	0.000	0.000	0.007	-	**0.155**
	MRR@k	-	0.078	0.000	0.000	0.000	0.000	0.005	-	**0.162**

information when reliable. However, it cannot exploit graph structure or generalize to unseen nodes, and degrades sharply as metadata shrinks (R@5 = 0.299 on MES, 0.092 on PubMed at 25%), making it poorly suited to realistic, metadata-incomplete settings.

Overall, we can see that the dataset recommendation methods proposed so far (EnsembleCN, DataFinder, LinearSVM, HVGAE) often assume the availability of reliable textual metadata and highly interconnected networks, while largely overlooking the dynamic nature of real-world graphs. These assumptions limit their applicability to real-world SKGs, which are typically characterized by incomplete or noisy metadata, sparse and heterogeneous connectivity, and continuously evolving structures.

7 Ablation study

In this section, we present an ablation study on the MES and PubMed datasets. Tables 6 and 7 assess the contribution of each phase of the pipeline, reporting Recall, nDCG, and MRR. Table 8 reports a modality ablation, in which the textual or the topological embeddings of dataset nodes are partially or completely removed, to assess the relative importance of each modality and the ability of MM-SAN to compensate for a missing one.

We consider two experimental settings: a transductive scenario with complete metadata availability (Tran 100%) and an inductive scenario where only 50% of the metadata is available. To assess the

Table 6. Ablation study on MES. Results are reported for the transductive setting with 100% metadata (Tran[100%]) and the fully inductive setting with 50% metadata (Ind[50%]). Results are grouped by the pipeline phase being ablated (Augmentation, Sampling, Aggregation, Combination). The best result for each metric and phase is underlined; the full MM-SAN model (bottom row) is shown in **bold**.

		Tran [100%]			**Ind [50%]**		
Component	Variant	R@5	nDCG@5	MRR@5	R@5	nDCG@5	MRR@5
Augmentation	No augmentation	0.642	0.476	0.450	0.415	0.310	0.268
	No entity linking	0.676	0.501	0.469	0.463	0.325	0.289
	No topic modeling	0.708	0.512	0.471	0.482	0.349	0.301
Sampling	Random, 1-hop	0.602	0.434	0.413	0.398	0.299	0.278
	Random, 2-hop	0.637	0.471	0.419	0.403	0.311	0.291
	Random, 3-hop	0.644	0.475	0.428	0.414	0.313	0.297
Aggregation	GRU	0.642	0.492	0.442	0.519	0.367	0.318
	LSTM	0.640	0.455	0.403	0.509	0.360	0.307
	Multi-head attention	0.615	0.416	0.428	0.495	0.347	0.311
	Linear projection	0.520	0.406	0.371	0.444	0.339	0.302
Combination	GRU	0.687	0.485	0.416	0.431	0.332	0.266
	LSTM	0.659	0.425	0.414	0.415	0.310	0.242
	Multi-head attention	0.693	0.508	0.420	0.446	0.310	0.264
	Mean pooling	0.621	0.497	0.339	0.329	0.256	0.218
MM-SAN		**0.727**	**0.546**	**0.484**	**0.526**	**0.391**	**0.326**

Table 7. Ablation study on PubMed. Results are reported for the transductive setting with 100% metadata (Tran[100%]) and the fully inductive setting with 50% metadata (Ind[50%]). Results are grouped by the pipeline phase being ablated (Augmentation, Sampling, Aggregation, Combination). For each metric and phase, the best result for each metric is underlined; the full MM-SAN model (bottom row) is shown in **bold**.

		Tran [100%]			**Ind [50%]**		
Component	Variant	R@5	nDCG@5	MRR@5	R@5	nDCG@5	MRR@5
Augmentation	No augmentation	0.311	0.201	0.181	0.168	0.120	0.081
	No entity linking	0.341	0.239	0.199	0.194	0.134	0.110
	No topic modeling	0.329	0.225	0.203	0.188	0.123	0.099
Sampling	Random, 1-hop	0.314	0.243	0.220	0.190	0.101	0.099
	Random, 2-hop	0.313	0.225	0.197	0.176	0.110	0.078
	Random, 3-hop	0.295	0.212	0.184	0.172	0.120	0.069
Aggregation	GRU	0.360	0.264	0.215	0.220	0.127	0.102
	LSTM	0.363	0.266	0.222	0.200	0.154	0.119
	Multi-head attention	0.284	0.203	0.177	0.187	0.127	0.112
	Linear projection	0.257	0.191	0.169	0.162	0.121	0.100
Combination	GRU	0.315	0.216	0.180	0.188	0.139	0.099
	LSTM	0.290	0.181	0.126	0.210	0.126	0.098
	Multi-head attention	0.320	0.192	0.221	0.215	0.127	0.101
	Mean pooling	0.189	0.123	0.108	0.192	0.119	0.082
MM-SAN		**0.380**	**0.272**	**0.234**	**0.242**	**0.169**	**0.121**

statistical significance of the results, we conducted paired t-tests ($\alpha = 0.05$) on a per-query basis to compare MM-SAN against the tested ablation configurations. All comparisons yielded statistically significant differences ($p < 0.05$).

Augmentation. For the augmentation phase, we evaluated the performance of MM-SAN under three settings: without both entity and topic nodes, without entity nodes only, and without topic nodes only. This analysis was conducted to assess the importance of the augmentation phase and

to determine which type of node contributes most significantly to the model's performance. The results are summarized in Tables 6 and 7 in *Augmentation* block.

The results show that the augmentation phase is crucial for improving graph connectivity and, consequently, enhancing the dataset recommendation task by mitigating the effects of graph sparsity. Interestingly, we observe opposite trends on the MES and PubMed datasets. On MES, removing entity nodes leads to a larger performance drop than removing topic nodes (R@5: 0.676 vs. 0.708 in the Tran 100% setting). This can be explained by the fact that MES is a curated and relatively dense dataset, where entities provide highly informative links between publications and datasets. In contrast, topics mainly act as hub nodes, offering more generic semantic connections and therefore contributing less to the recommendation performance.

On PubMed, however, the opposite behavior emerges. While entities still provide useful information, their impact is less pronounced than that of topics. Due to the much larger size and higher sparsity of the PubMed graph, topic nodes play a more important role in connecting otherwise distant regions of the graph, resulting in a greater performance gain. This is reflected by the larger drop observed when topic nodes are removed compared to entity nodes (R@5: 0.329 vs. 0.341 in tran 100% setting), highlighting the importance of topic-based augmentation in sparse and large-scale graphs.

Sampling. For the sampling phase, we compared the random-walk strategy used in our original pipeline with random sampling at 1-, 2-, and 3-hop distances, i.e., by randomly sampling the neighbors of a node at increasing distances from the target. Overall, random walk proves to be a more effective strategy, as it allows us to weight candidate neighbors according to both their distance from and their similarity to the initial node. This enables the model to prioritize more relevant nodes while still exploring a broader portion of the graph. The results are provided in Tables 6, 7 in *Sampling* blocks.

The results reveal different trends across the two datasets. On MES, randomly sampling neighbors at greater hop distances leads to a slight performance improvement across all learning settings (R@5 Ind [50%]: 0.398 $\rightarrow$ 0.403 $\rightarrow$ 0.414). This is consistent with the denser and better-connected structure of MES, where exploring further hops is more likely to uncover diverse and informative nodes. In contrast, on PubMed, performance is relatively stable across different hop distances (R@5 Ind [50%]: 0.190 $\rightarrow$ 0.176 $\rightarrow$ 0.172), with 1-hop sampling providing the best performance among the random-sampling strategies. This behavior is expected given the higher sparsity and strong node-type imbalance of PubMed: exploring further hops tends to introduce a larger number of nodes of the same type that are only weakly connected to the target, providing limited additional information. Nevertheless, the random-walk strategies integrated into our pipeline consistently outperform pure random sampling. By explicitly balancing graph proximity and textual similarity, random walks can construct more diverse and informative neighborhoods around the target nodes, while avoiding the limitations imposed by the strong structural and type imbalance of the underlying SKG.

Aggregation. For the aggregation phase, we evaluated different strategies for combining the topological and textual information of the nodes. Cross-attention jointly models the two modalities by allowing the representation of each modality to selectively incorporate information from the other, thereby learning their interactions rather than simply combining them through a fixed transformation. We also evaluated alternative aggregation strategies, including GRU, LSTM, multi-head attention, and linear projection. The results are provided in Tables 6, 7 in *Aggregation* blocks.

Overall, cross-attention provides the best performance across both datasets and learning settings, confirming the importance of explicitly modeling the interactions between topology and textual information. On MES, the best alternative is GRU, which achieves an R@5 of 0.642 and 0.519 in the

transductive and inductive settings, respectively (MM-SAN: 0.727 and 0.526). On PubMed, LSTM is the strongest alternative in the transductive setting, with an R@5 of 0.363 (MM-SAN: 0.380), while GRU performs best in the inductive setting, with an R@5 of 0.220 (MM-SAN: 0.242). The same trend is reflected in the other metrics, with MM-SAN consistently achieving the highest overall performance. These results indicate that recurrent architectures can effectively capture dependencies between the modalities, but remain less effective than the explicit cross-modal interactions learned by cross-attention. The performance gap becomes more pronounced for simpler aggregation strategies. In particular, linear projection achieves the lowest R@5 among the alternatives in the transductive setting, with 0.520 on MES and 0.257 on PubMed, while multi-head attention achieves 0.615 and 0.284, respectively. In the inductive setting, linear projection similarly obtains lower R@5 values on both datasets (0.444 on MES and 0.162 on PubMed), compared to 0.526 and 0.242 for MM-SAN. Overall, the consistent advantage of cross-attention under both complete and incomplete metadata availability suggests that explicitly modeling cross-modal interactions is particularly beneficial for dataset recommendation, where effectively integrating structural and textual information is crucial.

Combination. For the combination phase, we compared the concatenation strategy adopted by MM-SAN with four alternative ways of merging the group-specific representations produced by the aggregation module: GRU, LSTM, multi-head attention, and mean pooling. Results are reported in the *Combination* block of Tables 6 and 7. Multi-head attention is the strongest alternative across both datasets and settings, reaching an R@5 of 0.693 on MES and 0.320 on PubMed in the transductive setting (MM-SAN: 0.727 and 0.380), and 0.446 and 0.215 in the inductive setting (MM-SAN: 0.526 and 0.242). GRU performs close to multi-head attention on MES in the transductive setting (R@5 = 0.687), but falls further behind in the inductive one (R@5 = 0.431 on MES and 0.188 on PubMed). Mean pooling is the weakest strategy overall: in the transductive setting its R@5 drops to 0.621 on MES and 0.189 on PubMed, and it also yields the lowest inductive score on MES (R@5 = 0.329). On PubMed in the inductive setting, instead, all four alternatives fall within a narrow range (R@5 = 0.188–0.215), well below MM-SAN. Overall, these results suggest that concatenation is more effective than applying further learned or averaging operations, likely because it preserves the complementary information carried by each group of node types rather than compressing it into a single shared representation.

Modality ablation. We further extend our ablation study to assess the relative contribution of the topological and textual modalities and to evaluate the robustness of the cross-attention mechanism when one modality is partially or completely unavailable for the target dataset nodes. Following the missing-metadata setting described in Section 5, an unavailable modality is represented by replacing its embedding with a fixed constant vector of ones, preserving the dimensionality of the original representation while removing any node-specific information from that modality.

We consider two complementary settings. First, in the *Partially missing modality* setting, one modality is made unavailable for 50% of the dataset nodes, while the other modality remains fully available. We evaluate two symmetric variants: *Topology unavailable*, where the topological embeddings of the selected dataset nodes are replaced by the constant vector while their textual embeddings are preserved, and *Text unavailable*, where the textual embeddings are replaced while topology remains available. Second, in the *Complete missing modality* setting, the same procedure is extended to all dataset nodes (0% availability). In this case, either all textual embeddings or all topological embeddings are replaced by the constant vector, allowing us to isolate the contribution of each modality to the recommendation of dataset nodes.

Importantly, we remove modalities only from dataset nodes, which represent the target nodes in our recommendation task, while preserving the multimodal information available for the remaining node types. This setting allows us to specifically investigate the impact of missing information on

Table 8. Modality ablation analysis. Each percentage indicates the proportion of dataset nodes for which the masked modality is available (e.g., [50%]: the modality is missing for half of the dataset nodes; [0%]: it is missing for all of them), while the other modality remains available for all dataset nodes. **Topology unavailable** denotes the setting in which topology is masked, i.e., available only for the indicated percentage of dataset nodes, while text is always fully available; **Text unavailable** denotes the symmetric setting (the original MM-SAN configuration), in which textual metadata is available only for the indicated percentage of dataset nodes while topology is always fully available.

		MES		**PubMed**	
Setting	Metric	Topology unavailable	Text unavailable	Topology unavailable	Text unavailable
Tran [50%]	R@5	**0.635**	0.539	**0.312**	0.276
	nDCG@5	**0.423**	0.398	**0.267**	0.194
	MRR@5	**0.369**	0.340	**0.208**	0.166
Ind [50%]	R@5	**0.603**	0.526	**0.254**	0.242
	nDCG@5	**0.401**	0.391	**0.219**	0.169
	MRR@5	**0.345**	0.326	**0.192**	0.121
Tran [0%]	R@5	**0.446**	0.395	**0.232**	0.198
	nDCG@5	**0.331**	0.301	**0.196**	0.134
	MRR@5	**0.220**	0.199	**0.154**	0.126
Ind [0%]	R@5	**0.376**	0.338	**0.180**	0.149
	nDCG@5	**0.298**	0.276	**0.127**	0.108
	MRR@5	**0.262**	0.212	**0.158**	0.107

recommendation targets without reducing the entire graph to a unimodal setting. It also provides a direct assessment of whether MM-SAN can effectively exploit either modality when the other is unavailable and whether its cross-attention mechanism remains robust when topology, rather than text, constitutes the missing modality. Table 8 reports Recall@5, nDCG@5, and MRR@5 for MES and PubMed under both transductive and inductive settings.

The results confirm the bidirectional nature of MM-SAN's cross-attention mechanism: regardless of which modality is masked, the remaining one is able to partially compensate for the missing signal, allowing the model to preserve reasonably strong performance in both settings. At the same time, the two settings are not perfectly symmetric. Across both datasets and learning settings, the *Topology unavailable* setting tends to yield slightly higher scores than the *Text unavailable* — e.g., on MES in the transductive setting at 50% availability, Recall@5 is 0.635 when topology is missing versus 0.539 when text is missing, with a comparable gap on PubMed (0.312 vs. 0.276). This suggests that, when forced to rely on a single modality, textual information tends to provide a stronger fallback signal than topology alone, likely because it more directly encodes the semantic relevance between publications and datasets, while topology offers a more indirect, relational form of evidence.

8 Discussion

In the following, we discuss the results with respect to each research question introduced in Section 1, and then outline the theoretical and practical implications of MM-SAN for dataset recommendation over real-world scholarly infrastructures.

RQ1 (Robustness). The first research question concerns robustness to the structural sparsity of SKGs. Our results show that MM-SAN is robust to sparsity thanks to two complementary components of its architecture. First, the augmentation phase connects otherwise unrelated regions

of the graph: entity and topic nodes act as hubs that link publications and datasets sharing no direct structural path. Tables 6 and 7 confirm that removing augmentation, or either of its two sources, consistently degrades performance on both MES and PubMed, showing that the improvement stems from bridging distant regions of the graph rather than from merely adding nodes. The relative contribution of the two sources depends on the graph: entity nodes are more beneficial on denser and curated graphs such as MES, whereas topic nodes provide a larger benefit on sparser graphs such as PubMed, where topic-based links recover connections that the original topology lacks. Second, the weighted random-walk sampling favors neighbors that are both structurally close and semantically similar to the target, instead of spending neighbor slots on weakly informative nodes, as uniform sampling tends to do. Accordingly, it consistently outperforms uniform random sampling at every hop distance.

Takeaway. MM-SAN copes with the structural sparsity of SKGs by enriching the graph with entity and topic nodes, which connect otherwise distant regions, and by using weighted random walks to select the most informative neighbors within sparse and unbalanced neighborhoods.

RQ2 (Flexibility). Flexibility refers to the ability of MM-SAN to recommend datasets in transductive, semi-inductive, and fully inductive settings, i.e., also for nodes not seen during training. Across both datasets and all metadata-availability levels (Tables 3–4), all methods achieve their best results in the transductive setting, but MM-SAN shows the smallest degradation when moving to the semi- and fully inductive settings, and is the only method that remains competitive in all of them, which makes it suitable for dynamic graphs where retraining is impractical or infeasible. This behavior follows from how MM-SAN builds node representations: rather than relying on node-specific parameters learned during training, it constructs the representation of a node at inference time from its sampled neighborhood. For a previously unseen publication or dataset, the weighted random walks select the most structurally close and semantically relevant nodes in its local context, so that MM-SAN can build a meaningful representation. Since no node-specific parameters need to be learned or updated, the same trained model can be directly applied to newly introduced nodes without retraining.

Takeaway. MM-SAN generalizes to unseen nodes because it builds their representations at inference time from neighbors selected through weighted random walks, rather than from node-specific parameters. The same trained model can therefore serve new publications and datasets without retraining, making MM-SAN particularly suitable for dynamic scholarly graphs.

RQ3 (Versatility). We define versatility as the ability to limit performance degradation as textual metadata becomes progressively scarcer (100% → 75% → 50% → 25%). Tables 3 and 4 show that, while several baselines perform well when textual metadata is fully available, MM-SAN degrades considerably more gracefully as metadata decreases, since its cross-attention aggregation can shift weight toward structural information. This highlights the benefit of jointly exploiting complementary modalities rather than relying exclusively on textual information. This adaptation works in both directions: the modality ablation in Table 8 shows that MM-SAN also preserves reasonable performance when topology, rather than text, is missing, by relying more heavily on the textual signal. Cross-attention therefore does not implement a fixed fallback strategy, but adjusts

the contribution of each modality according to its availability and informativeness, which makes MM-SAN well suited to real-world SKGs with incomplete or unreliable metadata.

Takeaway. MM-SAN's versatility stems from its cross-attention mechanism, which adjusts the contribution of textual and topological information according to their availability, allowing the model to rely on whichever modality remains informative.

Theoretical and practical implications of MM-SAN. From a theoretical perspective, our findings suggest that dataset recommendation over real-world SKGs should jointly account for the semantic and structural dimensions of scholarly resources. Since datasets are sparsely connected to publications and their metadata is often incomplete, the two modalities provide complementary evidence: text captures the semantic relationship between a publication and a dataset, while graph structure provides useful signals when descriptions are missing or of low quality. This is reflected in our experiments, where MM-SAN maintains comparatively strong performance as metadata availability decreases and in the more challenging inductive settings. More broadly, robustness to incomplete multimodal information and the ability to generalize to evolving graphs emerge as key properties for recommender systems operating on scholarly infrastructures.

From a practical perspective, MM-SAN could be integrated into scholarly repositories, digital libraries, or dataset discovery services in two complementary ways: to suggest candidate datasets to authors during manuscript preparation or submission, and to surface related datasets to readers exploring a published work, supporting reproducibility and data reuse. For example, when a researcher searches for a publication, the system could identify potentially relevant datasets from the underlying SKG, even when they are not directly linked to it, providing a discovery mechanism that complements keyword-based search. Such applications are particularly valuable because existing SKGs contain hundreds of millions of nodes yet only sparse publication–dataset links, as reflected in the extreme sparsity of MES and PubMed.

The choice of learning setting should be guided by the update frequency and growth rate of the target SKG. The transductive setting, which yields the best performance, suits relatively stable resources for which the model can be periodically retrained offline, e.g., on a monthly or quarterly basis. The semi-inductive setting fits repositories that continuously receive new publications while recommending from an established dataset catalog. Finally, the fully inductive setting suits large aggregators such as OAG, where both publications and datasets are added continuously and a complete retraining cycle at every update would be impractical; here, the ability to represent entirely unseen nodes, on both the publication and the dataset side, avoids the computational and operational cost of continuous retraining.

Lastly, deployment should be framed within the FAIR principles – **F**indability, **A**ccessibility, **I**nteroperability, and **R**eusability [50]. MM-SAN mainly contributes to Findability, by surfacing datasets that lack explicit links to publications, and thereby indirectly to Reusability. However, identifying a relevant dataset is only the first step towards its reuse: a dataset may require authentication or an access request, and its license may restrict redistribution, commercial use, or the creation of derivative resources. Recommendations should therefore be accompanied by up-to-date information on access conditions, licenses, and provenance, allowing researchers to distinguish immediately reusable resources from those requiring additional permissions. In future extensions, such metadata could also be incorporated as additional signals or constraints in the recommendation process. More generally, while the inductive capability reduces the need for frequent retraining, it does not remove the underlying data-management and governance requirements of real-world deployment.

9 Final remarks

Data is a cornerstone of modern research, with significant efforts dedicated to its curation for learning and analytical tasks. However, data reuse remains limited due to the challenges of finding and assessing relevant datasets for specific research outputs. Dataset recommendation is essential to addressing this issue, yet it remains highly challenging due to the inherent noise, sparsity, and incompleteness of SKGs, which serve as the primary source for linking datasets and publications.

We propose MM-SAN, a multimodal graph representation learning method for enriching SKGs and recommending relevant datasets for publications. MM-SAN's effectiveness stems from a three-phase approach: (1) enhancing graph connectivity by introducing new nodes derived from textual attributes; (2) selecting relevant neighbors for a target node using random walks; and (3) combining neighbor representations for each type with cross attention to produce the final target representation. This architecture seamlessly integrates textual information with graph topology while not relying on complete metadata coverage.

We conduct extensive experiments across three settings (transductive, semi-inductive, fully inductive) and under ideal and real metadata conditions using two benchmarks extracted from the OAG. Our evaluation emphasizes the often-overlooked importance of generating predictions for newly added items without requiring model retraining. The results demonstrate MM-SAN's effectiveness in all tested settings.

We analyzed the impact of missing textual metadata across all methods and datasets, focusing on **robustness**, **flexibility**, and **versatility**. MM-SAN maintains strong performance across both large, sparse SKGs (i.e., PubMed) and smaller, well-curated datasets (i.e., MES), demonstrating robustness to noise and sparsity. Performance declines as dataset size increases, especially when textual descriptions are absent. MM-SAN also excels in flexibility, performing well in transductive, semi-inductive, and fully inductive settings, making it suitable for dynamic real-world applications. MM-SAN emerges as the most effective model in most conditions, although text-only approaches remain stronger when high-quality textual metadata is fully available. Future work could explore mixture-of-experts architectures to better exploit such metadata. Additionally, we plan to extend MM-SAN to incorporate FAIR-related accessibility considerations into the recommendation process, investigating how access conditions, licensing restrictions, provenance, and interoperability-related metadata can be used to distinguish between relevant datasets and those that are also practically accessible and reusable. We also plan to extend this analysis to venue-based recommendation, examining how these FAIR-related characteristics may vary across different node types and connectivity patterns.

Acknowledgments

This work is supported by the HEREDITARY Project, as part of the European Union's Horizon Europe research and innovation programme under Grant Agreement No GA 101137074.

References

[1] M. R. Ackermann, H. Bast, B. M. Beckermann, J. Kalmbach, P. Neises, and S. Ollinger. 2024. The dblp Knowledge Graph and SPARQL Endpoint. *TGDK* 2, 2 (2024), 3:1–3:23. doi:10.4230/TGDK.2.2.3

[2] U. Akujuobi and X. Zhang. 2017. Delve: A Dataset-Driven Scholarly Search and Analysis System. *SIGKDD Explorations* 19, 2 (2017), 36–46. doi:10.1145/3166054.3166059

[3] B. Altaf, U. Akujuobi, L. Yu, and X. Zhang. 2019. Dataset recommendation via variational graph autoencoder. In *2019 IEEE International Conference on Data Mining, ICDM 2019, Beijing, China, November 8-11, 2019*. IEEE, 11–20. doi:10.1109/ICDM.2019.00011

[4] K. Balog. 2018. *Entity-oriented search*. The Information Retrieval Series, Vol. 39. Springer. doi:10.1007/978-3-319-93935-3

[5] D. Brickley, M. Burgess, and N. Noy. 2019. Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In *Proceedings of the ACM Web Conference 2019, WWW*. ACM, 1365–1375. doi:10.1145/3308558.3313685

[6] P. Buneman, G. Christie, J. A. Davies, R. Dimitrellou, S. D. Harding, A. J Pawson, J. L Sharman, and Y. Wu. 2020. Why data citation isn't working, and what to do about it. *Database* 2020 (2020). doi:10.1093/DATABA/BAAA022

[7] P. Buneman, D. Dosso, M. Lissandrini, and G. Silvello. 2021. Data citation and the citation graph. *Quantitative Science Studies* 2, 4 (2021), 1399–1422. doi:10.1162/qss_a_00166

[8] Y. Cen, X. Zou, J. Zhang, H. Yang, J. Zhou, and J. Tang. 2019. Representation learning for attributed multiplex heterogeneous network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD*. ACM, 1358–1368. doi:10.1145/3292500.3330964

[9] Y. Chen, Y. Wang, Y. Zhang, J. Pu, and X. Zhang. 2019. Amender: an attentive and aggregate multi-layered network for dataset recommendation. In *2019 IEEE International Conference on Data Mining, ICDM 2019, Beijing, China, November 8-11, 2019*. IEEE, 988–993. doi:10.1109/ICDM.2019.00112

[10] Z. Chen, W. Gan, J. Wu, K. Hu, and H. Lin. 2025. Data Scarcity in Recommendation Systems: A Survey. *ACM Trans. Recomm. Syst.* 3, 3, Article 27 (2025), 31 pages. doi:10.1145/3639063

[11] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR* abs/1412.3555 (2014). arXiv:1412.3555 http://arxiv.org/abs/1412.3555

[12] P. Cremonesi, M. Picozzi, and M. Matera. 2012. A comparison of recommender systems for mashup composition. In *Proceedings of the Third International Workshop on Recommendation Systems for Software Engineering, RSSE 2012, Zurich, Switzerland, June 4, 2012*. IEEE, 54–58. doi:10.1109/RSSE.2012.6233411

[13] Y. Dong, N. V. Chawla, and A. Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 135–144. doi:10.1145/3097983.3098036

[14] M. Färber. 2019. The Microsoft Academic Knowledge Graph: A Linked Data Source with 8 Billion Triples of Scholarly Data. In *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 11779)*. Springer, 113–129. doi:10.1007/978-3-030-30796-7_8

[15] M. Färber and D. Lamprecht. 2021. The data set knowledge graph: Creating a linked open data source for data sets. *Quantitative Science Studies* 2, 4 (2021), 1324–1355.

[16] M. Färber and A.-K. Leisinger. 2021. Recommending datasets for scientific problem descriptions. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*. ACM, 3014–3018. doi:10.1145/3459637.3482166

[17] M. Fey and J. E. Lenssen. 2019. Fast Graph Representation Learning with PyTorch Geometric. *CoRR* abs/1903.02428 (2019). arXiv:1903.02428 http://arxiv.org/abs/1903.02428

[18] X. Fu, J. Zhang, Z. Meng, and I. King. 2020. Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding. In *Proceedings of the ACM Web Conference 2020, WWW*. ACM, New York, NY, USA, 2331–2341. doi:10.1145/3366423.3380297

[19] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber. 2009. A novel connectionist system for unconstrained handwriting recognition. *IEEE transactions on pattern analysis and machine intelligence* 31, 5 (2009), 855–868. doi:10.1109/TPAMI.2008.137

[20] M. Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *CoRR* abs/2203.05794 (2022). arXiv:2203.05794 doi:10.48550/ARXIV.2203.05794

[21] A. Grover and J. Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 855–864. doi:10.1145/2939672.2939754

[22] W. Hamilton, Z. Ying, and J. Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*. Curran Associates, Inc., 1024–1034.

[23] X. He, K. Deng, X. Wang, Y. Li, Y.-D. Zhang, and M. Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research*

and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020. ACM, 639–648. doi:10.1145/3397271.3401063

[24] Z. Hu, Y. Dong, K. Wang, and Y. Sun. 2020. Heterogeneous graph transformer. In *Proceedings of the ACM Web Conference 2020, WWW*. ACM, 2704–2710. doi:10.1145/3366423.3380027

[25] O. Irrera, M. Lissandrini, D. Dell'Aglio, and G. Silvello. 2024. Reproducibility and Analysis of Scientific Dataset Recommendation Methods. In *Proceedings of the 18th ACM Conference on Recommender Systems, RecSys 2024, Bari, Italy, October 14-18, 2024*. ACM, 570–579. doi:10.1145/3640457.3688071

[26] O. Irrera, M. Lissandrini, D. Dell'Aglio, and G. Silvello. 2025. Heterogeneous Graph Representation for Dataset Link Prediction on Dynamic and Sparse Scholarly Graphs. In *Linking Theory and Practice of Digital Libraries - 29th International Conference on Theory and Practice of Digital Libraries, TPDL 2025, Tampere, Finland, September 23-26, 2025, Proceedings (Lecture Notes in Computer Science, Vol. 16097)*. Springer, 452–469. doi:10.1007/978-3-032-05409-8_26

[27] O. Irrera, A. Mannocci, P. Manghi, and G. Silvello. 2023. A novel curated scholarly graph connecting textual and data publications. *ACM Journal of Data and Information Quality* 15, 3 (2023), 1–24. doi:10.1145/3597310

[28] M. Y. Jaradeh, A. Oelen, K. E. Farfar, M. Prinz, J. D'Souza, G. Kismihók, M. Stocker, and S. Auer. 2019. Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge. In *Proceedings of the 10th International Conference on Knowledge Capture*. ACM, New York, NY, USA, 243–246. doi:10.1145/3360901.3364435

[29] T. N. Kipf and M. Welling. 2016. Semi-Supervised Classification with Graph Convolutional Networks. *CoRR* abs/1609.02907 (2016). arXiv:1609.02907 http://arxiv.org/abs/1609.02907

[30] T. N. Kipf and M. Welling. 2016. Variational Graph Auto-Encoders. *CoRR* abs/1611.07308 (2016). arXiv:1611.07308 http://arxiv.org/abs/1611.07308

[31] K. Lin, T. Alrashed, and N. F. Noy. 2024. Relationships Are Complicated! An Analysis of Relationships Between Datasets on the Web. In *The Semantic Web - ISWC 2024 - 23rd International Semantic Web Conference, Baltimore, MD, USA, November 11-15, 2024, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 15231)*. Springer, 47–66. doi:10.1007/978-3-031-77844-5_3

[32] P. Manghi, C. Atzori, A. Bardi, M. Baglioni, J. Schirrwagen, H. Dimitropoulos, S. La Bruzzo, I. Foufoulas, A. Mannocci, M. Horst, A. Czerniak, K. Iatropoulou, A. Kokogiannaki, M. De Bonis, M. Artini, A. Lempesis, A. Ioannidis, N. Manola, P. Principe, T. Vergoulis, S. Chatzopoulos, and D. Pierrakos. 2022. *OpenAIRE Research Graph Dump*. doi:10.5281/zenodo.7488618

[33] Q. Mao, Z. Liu, C. Liu, and J. Sun. 2023. Hinormer: Representation learning on heterogeneous information networks with graph transformer. In *Proceedings of the ACM Web Conference 2023, WWW*. ACM, 599–610. doi:10.1145/3543507.3583493

[34] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *Proceedings the 7th International Conference on Semantic Systems, I-SEMANTICS 2011, Graz, Austria, September 7-9, 2011 (ACM International Conference Proceeding Series)*. ACM, 1–8. doi:10.1145/2063518.2063519

[35] J. Pennington, R. Socher, and C. D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL, 1532–1543. doi:10.3115/V1/D14-1162

[36] B. Perozzi, R. Al-Rfou, and S. Skiena. 2014. DeepWalk: online learning of social representations. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*. ACM, 701–710. doi:10.1145/2623330.2623732

[37] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*. AUAI Press, 452–461. https://www.auai.org/uai2009/papers/UAI2009_0139_48141db02b9f0b02bc7158819ebfa2c7.pdf

[38] M. S. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling. 2018. Modeling Relational Data with Graph Convolutional Networks. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings (Lecture Notes in Computer Science, Vol. 10843)*. Springer, 593–607. doi:10.1007/978-3-319-93417-4_38

[39] G. Silvello. 2018. Theory and practice of data citation. *Journal of the Association for Information Science and Technology* 69, 1 (2018), 6–20.

[40] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. 2008. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 990–998. doi:10.1145/1401890.1402008

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention Is All You Need. *CoRR* abs/1706.03762 (2017). arXiv:1706.03762 http://arxiv.org/abs/1706.03762

[42] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. 2017. Graph Attention Networks. *CoRR* abs/1710.10903 (2017). arXiv:1710.10903 http://arxiv.org/abs/1710.10903

[43] V. Viswanathan, L. Gao, T. Wu, P. Liu, and G. Neubig. 2023. DataFinder: Scientific Dataset Recommendation from Natural Language Descriptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL, 10288–10303. doi:10.18653/v1/2023.acl-long.573

[44] H. Wang, M. Zhao, X. Xie, W. Li, and M. Guo. 2019. Knowledge Graph Convolutional Networks for Recommender Systems. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*. ACM, 3307–3313. doi:10.1145/3308558.3313417

[45] S. Wang, L. Thompson, and M. Iyyer. 2021. Phrase-BERT: Improved Phrase Embeddings from BERT with an Application to Corpus Exploration. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Association for Computational Linguistics, 10837–10851. doi:10.18653/V1/2021.EMNLP-MAIN.846

[46] X. Wang, X. He, Y. Cao, M. Liu, and T.-S. Chua. 2019. KGAT: Knowledge Graph Attention Network for Recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*. ACM, 950–958. doi:10.1145/3292500.3330989

[47] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu. 2019. Heterogeneous Graph Attention Network. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*. ACM, 2022–2032. doi:10.1145/3308558.3313562

[48] X. Wang, F. Van Harmelen, M. Cochez, and Z. Huang. 2022. Scientific Item Recommendation Using a Citation Network. In *Knowledge Science, Engineering and Management - 15th International Conference, KSEM 2022, Singapore, August 6-8, 2022, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 13369)*. Springer, 469–484. doi:10.1007/978-3-031-10986-7_38

[49] X. Wang, F. Van Harmelen, and Z. Huang. 2022. Recommending scientific datasets using author networks in ensemble methods. *Data Science* 5, 2 (2022), 167–193. doi:10.3233/DS-220056

[50] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, N. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J.G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A.C 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 1 (2016), 160018. doi:10.1038/sdata.2016.18

[51] C. Zhang, D. Song, C. Huang, A. Swami, and N. V. Chawla. 2019. Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 793–803. doi:10.1145/3292500.3330961

Appendix

In this appendix, we provide additional details regarding some key aspects of our study. First, we report the hyperparameter configurations of all baseline methods: for each baseline, we list the complete set of hyperparameter values explored during grid search, while the best-performing configurations are reported in Section 5.

Second, we present a component analysis of our pipeline. In particular, we investigate the contribution of two critical stages, the augmentation phase and the sampling phase, evaluating their individual impact on the overall performance, also when applied to the baseline methods.

Third, we report additional analyses on MM-SAN. We first evaluate MM-SAN with simpler encoders, replacing `node2vec` with `DeepWalk` and the transformer-based text encoders with mean-pooled GloVe embeddings, to assess how much the choice of encoders contributes to its effectiveness. We also analyze the sensitivity of MM-SAN to the number of neighbors sampled for each node type.

Finally, we evaluate MM-SAN on a more challenging variant of the task, in which authors rather than publications serve as query nodes, to assess whether it generalizes beyond publication-based dataset recommendation.

A.1 Baselines hyperparameters

We grouped the baseline methods into three categories based on their underlying architectural mechanism, rather than by data type alone, since this let us define a shared and meaningful hyperparameter search space for each group.

The first group consists of the variational graph autoencoders (VGAE and HVGAE), which learn probabilistic node representations and reconstruct the graph structure through an inner-product decoder. The second group brings together the attention-based models (GAT, KGAT, HGT, HAN, and HGNN), all of which rely on a multi-head attention mechanism to weigh the contribution of neighboring nodes. The third group covers non-attention message-passing models with explicit neighbor sampling (GraphSAGE, RGCN, and KGCN), which instead aggregate neighborhood information through fixed aggregation functions or relation-specific weight decompositions.

For each group, we report the hyperparameters explored via grid search, including learning rate, hidden/embedding dimension, number of layers, dropout, and weight decay, along with group-specific parameters such as the number of attention heads, the neighbor sampling size, and the aggregation strategy. The hyperparameter ranges explored for each baseline are summarized in Tables A1, A2, A3.

Table A1. Hyperparameters for variational graph autoencoder models.

Model	Learning Rate	Embedding Dim. L	Layers	Weight Decay (λ)	Optimizer
VGAE	{1e-5, 1e-4, 5e-4, 1e-3, 5e-3, 1e-2}	{16, 32, 64, 128}	{2, 3}	{0, 1e-4, 1e-3}	Adam
HVGAE	{1e-6, 1e-5, 1e-4, 5e-4, 1e-3}	{8, 16, 32, 64}	{2}	{1e-4, 1e-3, 1e-2}	Adam

Table A2. Hyperparameters for attention-based models.

Model	Learning Rate	Hidden Dim.	Layers	Heads	Neighbors	Dropout	Weight Decay
GAT	{1e-5, 1e-4, 5e-4, 1e-3, 5e-3, 1e-2}	{64, 128, 256}	{2, 3}	{1, 2, 4, 8}	-	{0.0, 0.1, 0.3, 0.5}	{0, 1e-5, 1e-4}
KGAT	{1e-5, 1e-4, 5e-4, 1e-3, 5e-3, 1e-2}	{64, 128, 256}	{2, 3}	{1, 2, 4, 8}	{5, 10, 15}	{0.0, 0.1, 0.3, 0.5}	{0, 1e-5, 1e-4}
HGT	{1e-5, 1e-4, 5e-4, 1e-3, 5e-3, 1e-2}	{64, 128, 256}	{2, 3}	{1, 2, 4, 8}	{5, 10, 15}	{0.0, 0.1, 0.3, 0.5}	{0, 1e-5, 1e-4}
HAN	{1e-5, 1e-4, 5e-4, 1e-3, 5e-3, 1e-2}	{64, 128, 256}	{2, 3}	{1, 2, 4, 8}	-	{0.0, 0.1, 0.3, 0.5}	{0, 1e-5, 1e-4}
HGNN	{1e-5, 1e-4, 5e-4, 1e-3, 5e-3, 1e-2}	{64, 128, 256}	{2, 3}	{1, 2, 4, 8}	{5, 10, 15}	{0.0, 0.1, 0.3, 0.5}	{0, 1e-5, 1e-4}

Table A3. Hyperparameters for non-attention message-passing models.

Model	Learning Rate	Hidden Dim.	Layers	Neighbors	Aggregator / Num. Bases	Dropout	Weight Decay
GraphSAGE	{1e-5, 1e-4, 5e-4, 1e-3, 5e-3, 1e-2}	{64, 128, 256}	{2, 3}	{5, 10, 15}	Aggr.: {mean, pool, LSTM}	{0.0, 0.1, 0.3, 0.5}	{0, 1e-5, 1e-4}
RGCN	{1e-5, 1e-4, 5e-4, 1e-3, 5e-3, 1e-2}	{64, 128, 256}	{2, 3}	{5, 10, 15}	Bases: {0, 4, 8, 16}	{0.0, 0.1, 0.3, 0.5}	{0, 1e-5, 1e-4}
KGCN	{1e-5, 1e-4, 5e-4, 1e-3, 5e-3, 1e-2}	{64, 128, 256}	{2, 3}	{5, 10, 15}	Aggr.: score-weighted	{0.0, 0.1, 0.3, 0.5}	{0, 1e-5, 1e-4}

A.2 Component analysis

We analyze the impact of the two main components of our pipeline: the augmentation and sampling stages. Specifically, we evaluate five baseline methods (GAT, SAGE, HAN, HGT, and HGNN) alongside MM-SAN on datasets with and without augmentation, to assess whether and to what extent augmentation affects their performance. We further investigate the impact of the sampling strategy by comparing the performance of the selected methods under the random-walk sampling configuration. Since the Recall@10, nDCG@10, and MRR@10 metrics exhibit trends that closely mirror those at cutoff 5, the latter are considered indicative of the system's overall performance and are therefore used as the primary basis for our analysis.

Table A4. Augmentation analysis. Recall@5 (R@5), nDCG@5 (N@5), and Mean Reciprocal Rank (MRR@5) scores on MES and PubMed under ideal conditions with and without augmentation. Best results are in **bold**, second-best are underlined.

Method	MES						PubMed					
	Augmentation			No Augmentation			Augmentation			No Augmentation		
	R@5	N@5	MRR@5	R@5	N@5	MRR@5	R@5	N@5	MRR@5	R@5	N@5	MRR@5
GAT	0.578	0.422	0.443	0.625	0.475	**0.462**	0.228	0.165	0.198	0.241	0.181	0.203
SAGE	0.472	0.312	0.411	0.489	0.398	0.448	0.217	0.172	0.191	0.239	0.179	0.204
HGT	0.020	0.020	0.010	0.023	0.020	0.020	0.014	0.015	0.009	0.016	0.017	0.018
HAN	0.008	0.009	0.008	0.008	0.009	0.009	0.026	0.021	0.010	0.030	0.023	0.024
HGNN	0.601	0.438	0.406	0.581	0.438	0.378	0.294	0.222	0.218	0.284	**0.202**	**0.209**
MM-SAN	**0.727**	**0.546**	**0.484**	**0.642**	**0.476**	0.450	**0.380**	**0.272**	**0.234**	**0.311**	0.201	0.181

Augmentation. We first assess how the augmentation phase ❶ in Figure 1 – specifically, the integration of entity and topic nodes to reduce overall sparsity – affects system effectiveness. Table A4 presents the performance of six systems with (`augmentation` column) and without (`no augmentation` column) this phase on the MES and PubMed datasets. Moreover, we focus on the ideal transductive setup with full metadata availability, as it is the configuration in which all baselines perform their best, and consequently provides the clearest assessment of the augmentation impact.

For GAT and GraphSAGE, which were originally developed for homogeneous graphs, the introduction of multiple node types tends to degrade their effectiveness. These architectures implicitly assume neighborhoods composed of a single node type and a shared feature space across all nodes, so when applied to heterogeneous SKGs they struggle to properly model type-specific semantics and relations, leading to suboptimal representations and weaker recommendation performance. In other words, the added nodes increase connectivity but also introduce noise that these models are not designed to handle. MM-SAN and HGNN benefit from introducing new nodes, improving performance with the enriched graph. This suggests that the augmentation phase is particularly beneficial for systems that exploit topological features and are designed explicitly for heterogeneous graphs. For HAN and HGT, removing the augmentation phase does not significantly affect their final performance. In summary, the augmentation phase proves useful for models like MM-SAN and HGNN that are tailored for heterogeneous graphs and leverage topological features. However, it may not benefit – and can even hinder – models not explicitly designed to handle heterogeneous structures.

Sampling. To test the effectiveness of our sampling phase (phase ❷ in Figure 1), we reimplemented the best-performing baselines – namely SAGE, GAT and HGNN – and adapted them to incorporate our sampling strategy. In SAGE and GAT we replaced their random sampling strategy with our random walk-based one, while in HGNN we modified the random walk–based sampling to select walks and neighbors according to their semantic similarity to the target node. The resulting performance scores for the MES and PubMed datasets at 100% of available metadata are presented in Table A5. Similarly to the previous case, the sampling strategy does not rely on metadata availability; as a consequence, we provide the results for the ideal scenario with 100% of available metadata. We find that integrating our sampling strategy into the baseline models leads to consistent performance gains when evaluated on real-world datasets like MES and PubMed. On MES, GraphSAGE benefits the most, achieving improvements of up to 35% in Recall and 28% in nDCG. On PubMed, GAT and GraphSAGE exhibit the largest boosts over their original implementations, with GraphSAGE achieving up to 51% improvement in nDCG and GAT reaching 38%. Despite these gains from the modified sampling strategy, MM-SAN remains the most effective model, consistently outperforming

Table A5. Sampling analysis. Comparison between original baselines and their variants enhanced with our sampling strategy. The table reports Recall@5 (R@5), nDCG@5 (N@5), and Mean Reciprocal Rank@5 (MRR@5) on the MES and PubMed datasets under transductive (Tran), semi-inductive (Semi), and inductive (Ind) settings. 100% indicates full metadata availability. The Gain% column shows the relative improvement achieved by our sampling method. Best results are in **bold**, second-best are underlined.

Dataset	Setting	Metric	GAT	GAT-RW	Gain %	SAGE	SAGE-RW	Gain %	HGNN	HGNN-RW	Gain %	MM-SAN
MES [100%]	Tran	R@5	0.625	0.700	12.00	0.489	0.610	24.74	0.581	0.700	20.48	**0.727**
		N@5	0.475	0.520	9.47	0.398	0.510	28.14	0.438	0.520	18.72	**0.546**
		MRR@5	0.462	0.473	2.38	0.448	0.470	4.91	0.378	0.390	3.17	**0.484**
	Semi	R@5	0.621	0.700	12.72	0.446	0.600	34.53	0.587	0.700	19.25	**0.719**
		N@5	0.479	0.490	2.30	0.316	0.400	26.58	0.417	0.510	22.30	**0.524**
		MRR@5	0.454	0.456	0.44	0.437	0.455	1.79	0.362	0.380	4.97	**0.461**
	Ind	R@5	0.590	0.622	5.42	0.439	0.512	16.63	0.582	0.634	8.94	**0.687**
		N@5	0.442	0.513	16.06	0.315	0.346	9.84	0.422	0.475	12.56	**0.519**
		MRR@5	0.434	0.437	0.69	0.417	0.438	0.23	0.370	0.377	1.89	**0.466**
PubMed [100%]	Tran	R@5	0.241	0.298	23.65	0.239	0.278	16.32	0.284	0.301	5.96	**0.380**
		N@5	0.181	0.202	11.60	0.179	0.201	12.29	0.202	0.256	26.73	**0.272**
		MRR@5	0.203	0.214	5.42	0.204	0.219	7.35	0.209	0.225	7.66	**0.234**
	Semi	R@5	0.231	0.242	4.76	0.221	0.253	14.48	0.246	0.271	10.16	**0.373**
		N@5	0.171	0.205	19.88	0.169	0.182	7.69	0.177	0.201	13.56	**0.266**
		MRR@5	0.202	0.205	1.49	0.204	0.202	-0.98	0.222	0.213	-4.05	**0.232**
	Ind	R@5	0.196	0.240	22.45	0.207	0.239	27.81	0.271	0.320	18.08	**0.344**
		N@5	0.134	0.186	38.81	0.133	0.201	51.13	0.199	**0.267**	34.17	0.246
		MRR@5	0.182	0.199	9.34	0.193	**0.214**	0.47	0.187	0.201	7.49	**0.214**

Table A6. Encoders analysis. The *Simpler* column reports the performance of MM-SAN when the original encoders are replaced with simpler alternatives: DeepWalk for graph embeddings and mean-pooled GloVe embeddings for textual representations.

		MES		**PubMed**	
Setting	Metric	Simpler	MM-SAN	Simpler	MM-SAN
Tran [100%]	R@5	0.700	**0.727**	0.317	**0.380**
	nDCG@5	0.520	**0.546**	0.219	**0.272**
	MRR@5	0.453	**0.484**	0.217	**0.234**
Ind [50%]	R@5	0.449	**0.526**	0.179	**0.242**
	nDCG@5	0.291	**0.391**	0.106	**0.169**
	MRR@5	0.231	**0.326**	0.109	**0.121**

the adapted baselines in Recall, nDCG, and MRR. These results demonstrate that our sampling strategy yields consistent and meaningful improvements when applied to real-world SKGs with high incompleteness, node imbalance, sparseness, and dynamicity.

A.3 Additional experiments

A.3.1 Encoders analysis. To assess the contribution of the encoding models employed in MM-SAN, we compared the original `node2vec` graph encoder and the Transformer-based text encoders (`Phrase-BERT` and `all-MiniLM-L6-v2`) against simpler alternatives. Specifically, we replaced `node2vec` with `DeepWalk` for graph representation learning and the Transformer-based text encoders with GloVe embeddings [35], where textual representations were obtained by mean-pooling the embeddings of the words contained in each text. The results are presented in Table A6. The results show that the original encoders consistently outperform the simpler alternatives across both datasets and learning settings. In the transductive setting, MM-SAN improves R@5 from 0.700 to 0.727 on MES and from 0.317 to 0.380 on PubMed when using the original encoders. The difference becomes more pronounced in the inductive setting, where R@5 increases from 0.449 to 0.526 on MES and from 0.179 to 0.242 on PubMed. Similar improvements are observed for

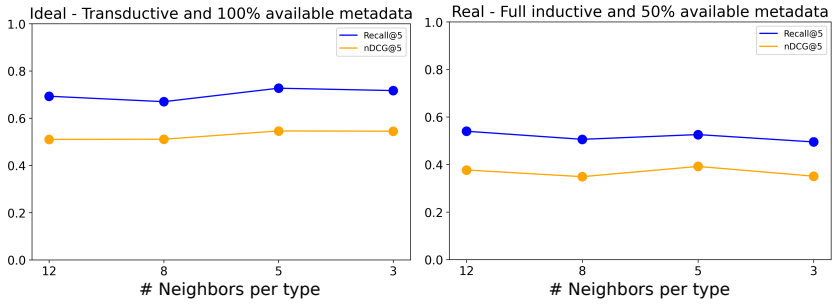


Fig. A1. Hyperparameter analysis: number of neighbors of a target node sampled for each node type.

Table A7. Dataset recommendation for authors results. Recall@k (R@k), nDCG@k (N@k) and Mean Reciprocal Rank@k (MRR@k), with $k = 5$, on MES and PubMed in transductive (Tran), semi-inductive (Semi), and inductive (Ind) settings, and with $\{100, 50\}$% of datasets with available metadata. Best results are in **bold**, second-best are underlined.

			MES $k = 5$							**PubMed** $k = 5$						
Metadata %	Setting	Metric	GAT	KGAT	SAGE	HGT	HAN	HGNN	MM-SAN	GAT	KGAT	SAGE	HGT	HAN	HGNN	MM-SAN
100%	Tran	R@k	0.120	0.098	0.105	0.010	0.008	0.567	**0.677**	0.081	0.018	0.098	0.004	0.003	0.189	**0.212**
		N@k	0.100	0.075	0.095	0.004	0.005	0.422	**0.539**	0.066	0.009	0.069	0.001	0.001	0.129	**0.180**
		MRR@k	0.118	0.078	0.091	0.006	0.005	0.375	**0.407**	0.071	0.010	0.083	0.002	0.001	0.101	**0.106**
	Semi	R@k	0.092	0.078	0.085	0.008	0.007	0.530	**0.613**	0.078	0.010	0.082	0.001	0.001	0.150	**0.200**
		N@k	0.076	0.055	0.053	0.003	0.004	0.382	**0.446**	0.065	0.008	0.071	0.000	0.000	0.100	**0.123**
		MRR@k	0.089	0.071	0.064	0.004	0.003	0.332	**0.391**	0.053	0.008	0.061	0.000	0.001	0.109	**0.113**
	Ind	R@k	0.065	0.054	0.064	0.000	0.000	0.362	**0.468**	0.068	0.010	0.062	0.000	0.000	0.099	**0.102**
		N@k	0.041	0.033	0.042	0.000	0.000	0.276	**0.363**	0.041	0.006	0.041	0.000	0.000	0.089	**0.098**
		MRR@k	0.054	0.041	0.051	0.000	0.000	0.271	**0.354**	0.041	0.004	0.056	0.000	0.000	0.070	**0.082**
50%	Tran	R@k	0.101	0.006	0.024	0.006	0.005	0.410	**0.456**	0.031	0.009	0.043	0.003	0.002	0.098	**0.106**
		N@k	0.081	0.004	0.013	0.001	0.002	0.234	**0.272**	0.010	0.003	0.023	0.000	0.002	0.090	**0.098**
		MRR@k	0.094	0.004	0.019	0.001	0.001	0.193	**0.216**	0.021	0.004	0.025	0.000	0.000	0.091	**0.103**
	Semi	R@k	0.099	0.005	0.013	0.002	0.002	0.328	**0.356**	0.019	0.004	0.020	0.001	0.002	0.098	**0.102**
		N@k	0.072	0.004	0.009	0.000	0.000	0.190	**0.212**	0.011	0.003	0.013	0.000	0.000	0.064	**0.092**
		MRR@k	0.091	0.004	0.010	0.001	0.000	0.152	**0.205**	0.012	0.002	0.015	0.000	0.000	0.084	**0.100**
	Ind	R@k	0.095	0.002	0.003	0.001	0.001	0.208	**0.307**	0.012	0.003	0.015	0.000	0.001	0.077	**0.089**
		N@k	0.054	0.001	0.002	0.000	0.000	0.139	**0.189**	0.008	0.002	0.008	0.000	0.000	0.046	**0.073**
		MRR@k	0.082	0.001	0.000	0.000	0.000	0.106	**0.130**	0.007	0.001	0.009	0.000	0.000	0.038	**0.064**

nDCG@5 and MRR@5. These results suggest that richer structural and contextual representations are particularly beneficial when only partial metadata is available. Nevertheless, the simpler encoders remain competitive, indicating that the overall MM-SAN framework remains effective even with less expressive representations.

A.3.2 Neighbors analysis. We analyzed the impact of the number of sampled neighbors on MM-SAN's overall effectiveness. Figure A1 reports the results on the MES dataset under two representative settings: an ideal scenario (transductive with 100% metadata availability) and a more realistic scenario (fully inductive with 50% metadata availability). In both cases, performance varies only slightly with the number of sampled neighbors and reaches its best overall performance when five neighbors per node type are selected. This configuration provides a good balance between representational richness and computational cost, supporting its use as the default sampling size. The limited sensitivity to the number of neighbors also suggests that MM-SAN's cross-attention mechanism can effectively identify relevant information within the sampled neighborhoods, contributing to its robustness across different graph conditions.

A.3.3 Dataset recommendation for authors. In this section, we evaluate the ability of MM-SAN to generalize to a more challenging dataset recommendation scenario in which authors, rather than publications, are used as query nodes. Compared to publications and datasets, author nodes do not directly benefit from the graph augmentation process, making the recommendation task inherently more difficult. Furthermore, author nodes are substantially sparser due to the absence of reliable author disambiguation, which often results in multiple nodes representing the same individual and, consequently, a large number of low-degree nodes. As a result, the information available from the query node itself is often limited, increasing the importance of effectively exploiting the surrounding graph structure. In this context, neighborhood exploration becomes a key component of the recommendation process. MM-SAN mitigates the effects of author sparsity by leveraging the local graph structure whenever authors are connected to multiple research artifacts or when those artifacts are embedded in a richer context of related entities, such as organizations, topics, keywords, and other scholarly objects. Through its graph augmentation and neighborhood sampling mechanisms, MM-SAN can exploit indirect relational signals that would otherwise remain inaccessible from the author node alone, enabling more effective recommendations despite the sparsity of the author layer.

Following the same procedure adopted for the publication–dataset recommendation task, we constructed a second set of five edge splits for the author–dataset recommendation task, where authors serve as query nodes and datasets as candidate items to be ranked. A similar splitting strategy was applied to the complete set of author–dataset edges $\varepsilon_{d,a}$, resulting in training, validation, transductive test, semi-inductive test, and fully inductive test sets. For the MES dataset, the validation and each of the three test sets contain approximately 400 edges, while the training set contains approximately 2,800 edges. For PubMed, the validation and each test set contain approximately 1,000 edges, with approximately 20,000 edges assigned to training.

In Table A7, we report the results of the inductive models for dataset recommendation using authors as query nodes, considering settings with 100% and 50% of datasets with available metadata. Performance is evaluated using Recall@5, nDCG@5, and MRR@5. MM-SAN consistently achieves the best performance among the evaluated inductive methods. Approaches such as GAT, SAGE, HAN, HGT and KGAT perform substantially worse than MM-SAN, with only GAT and SAGE achieving Recall higher than 0.1 in the transductive setting on MES. This can be attributed to the high sparsity of author nodes and the resulting difficulty in identifying representative neighborhoods. Similarly to the publication-based task, HGNN is the strongest baseline both on MES and PubMed, suggesting that heterogeneous message passing is effective in exploiting the limited connectivity of the graph.

As the proportion of datasets with available metadata decreases, MM-SAN's performance also declines, as its learned representations become less informative. When metadata is missing from a larger fraction of dataset neighbors, less semantic information can be propagated to the author representation, resulting in lower recommendation performance. The inductive setting exhibits the largest performance degradation. On MES, MM-SAN's R@5 decreases from 0.677 to 0.468 with 100% metadata, and from 0.456 to 0.307 with 50% metadata. On PubMed, the corresponding decreases are from 0.212 to 0.102 and from 0.106 to 0.089, respectively. Similar trends are observed for nDCG@5 and MRR@5. The larger degradation observed in the inductive setting can be attributed to the combination of node novelty and author sparsity. For unseen authors, the model must infer relevant datasets from the available graph context surrounding the author, which is often sparse and only weakly informative. In this respect, when these nodes have few connections, and particularly when the available neighboring datasets provide limited metadata, the information available for constructing an informative representation is substantially reduced. This effect is

especially pronounced in the fully inductive setting, where both the query authors and the target datasets are unseen during training.

Overall, these findings highlight the intrinsic difficulty of author-based dataset recommendation and the effectiveness of MM-SAN in making use of the limited structural and semantic information available in the SKG.