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

Anonymous Author(s)

## Abstract

Researchers increasingly struggle to find relevant datasets due to the overwhelming volume of data across repositories and the rapid growth of scholarly publications. Manual dataset retrieval is time-consuming and often fails to surface valuable datasets that are not easily discoverable through direct links or popularity metrics.

Dataset recommendation – identifying relevant datasets for a paper, author, or query – has become essential for improving dataset discoverability and reuse. However, this task is challenging due to the vast scale, heterogeneous structure, dynamic nature, and inherent incompleteness of Scholarly Knowledge Graphs (SKGs). While openly available SKGs organize scientific knowledge by linking papers, authors, venues, and datasets, they often fail to establish meaningful connections between datasets, leaving them isolated and underutilized. These graphs contain millions of evolving connections, yet suffer from missing or inaccurate metadata, duplicate entries, and weak dataset integration.

This work introduces MultiModal Scholarly Attention Network (MM-SAN), a novel multimodal heterogeneous graph representation learning method designed to recommend datasets relevant to a given research publication within real-world SKGs. MM-SAN integrates topology-based and text-based embeddings, enabling effective recommendations even when textual metadata is sparse or absent. Furthermore, it is designed to operate in transductive, semi-inductive, and entirely inductive settings, allowing it to adapt to evolving SKGs without retraining. Experimental results demonstrate that MM-SAN outperforms state-of-the-art approaches, offering a robust and scalable solution to dataset recommendation.

## CCS Concepts

• **Information systems → Recommender systems**.

## Keywords

Dataset recommendation, multimodal recommendation, scholarly knowledge grpahs

## 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 [4–6, 32, 42].

To improve dataset discoverability it is crucial to rely on structured and interlinked representations of scholarly information [27]. Scholarly Knowledge Graphs (SKGs) support this process by providing a rich, structured network of entities and their relationships, such as papers, datasets, authors, and organizations. SKGs serve as a foundational infrastructure for various scholarly tasks, including dataset discovery, information retrieval, and knowledge exploration.

Despite their importance, SKGs – including Microsoft Academic Knowledge Graph (MAKG) [37], Open Research Knowledge Graph (ORKG) [24], and OpenAIRE Graph (OAG) [29], which explicitly include datasets, often lack sufficient representation of datasets and their interconnections with scientific publications. Datasets are frequently isolated or only loosely linked to related research outputs. Additionally, SKGs are *dynamic*, continuously expanding with new publications and datasets. Yet, they suffer from *incompleteness* (where publications typically have detailed metadata, while datasets often lack adequate descriptions) and *sparseness* (where datasets and publications are weakly connected). These limitations hinder the ability of SKGs to capture dataset-related information, making dataset discovery challenging.

In this context, *dataset recommendation*, which involves identifying relevant datasets for a paper, author, or textual query, is crucial in improving dataset discoverability and fostering data reuse. However, developing effective dataset recommendation methods requires addressing the inherent challenges of sparse, heterogeneous, 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 constrained by a *scarcity of 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 [2, 9, 40, 41] but are evaluated on homogeneous, static, and specialized SKGs, failing to handle the complexity of real-world graphs. Other methods rely solely on textual metadata from publications and datasets [14, 36], 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 and 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 comprises scientific datasets. We address three research questions:

**RQ1 (Robustness).** To what extent can dataset recommendation algorithms effectively handle data sparsity?
**RQ2 (Flexibility).** Can the models work effectively both in transductive and inductive settings?
**RQ3 (Versatility).** Can we enhance the influence of graph topology in scenarios where textual metadata is incomplete or absent?

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 SKG scenarios. For the second, we evaluate the model's ability to generalize across various learning settings, including transductive, inductive, and semi-inductive configurations. Lastly, the third question focuses on whether the method can effectively exploit graph structure to mitigate the impact of missing or limited textual metadata.

Central to this study is MM-SAN, and we outline three primary **contributions**:

- We propose MM-SAN, 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 of MM-SAN across a wide range of settings, from metadata-rich to metadata-free environments, consistently showing MM-SAN's superior performance.
- 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. Finally, in Section 8, we draw some conclusions.

## 2 Related work

This section presents two core areas of related work that represent the basis of our approach. 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 explore graph representation learning methods, which allow for the extraction of informative and expressive node embeddings from SKGs, and we review existing approaches that have been applied to the dataset recommendation task.

*Scholarly Knowledge Graphs.* SKGs are structured representations of the scientific outputs interconnecting entities such as publications, authors, and datasets. Several SKGs, like AMiner [33] and DBLP [28], still focus mainly on representing connections among publications. On the other hand, more SKGs started including datasets, namely: MAKG [37], with a total of 200M nodes and over 10B relationships; ORKG, containing approximately 5M nodes and 50M relationships; Data Set Knowledge Graph (DSKG) [13], with $2K$ datasets linked to $635K$ publications (also linked to other Linked Data sources such as MAKG, ORCID, and Wikidata); and OAG [29], the largest with 227M nodes and 15B relationships; it is also open-access and includes more than 60M datasets. Recently, the European Marine Science (MES) graph [23] was published as a subgraph of OAG; it has been semi-automatically curated and provides a reliable ground truth dataset.

On the other hand, we found that several state-of-the-art methods have been tested on homogeneous and ad-hoc datasets applicable only for metadata-based prediction and recommendation; in frequent cases, they are not proper SKGs. For this reason, in this work, we do not consider: (i) The LinearSVM_Dataset, which is a bipartite graph containing $1,691$ dataset titles, extracted from the DSKG [13], and 88K abstracts from MAKG; (ii) The DataFinder Dataset [36], which is not a proper SKG as it lacks information (nodes) about authors, venues, and organizations, consisting only of textual queries (17K) associated with relevant datasets (7K); (iii) The Delve [1] subsets, where the baselines HVGAE [2] and AMENDER [9] were initially tested, which is quite dense and thus far from a typically sparse real-world SKGs.

*Graph-based baselines.* Graph Representation Learning (GRL) is adopted by various methods to learn node embeddings by transforming graph structures into continuous vector spaces for node

classification, link prediction, and recommendation tasks. Graph Convolutional Networks (GCNs)[25] is adopted as a foundational approach that learns node embeddings by aggregating and transforming information from neighboring nodes through graph convolutions. For homogeneous graphs, GraphSAGE[20] extends this idea by sampling a fixed set of representative neighbors for each node and aggregating their representations using techniques like mean aggregation, LSTM, or pooling. GAT [35] further refines this method by applying attention mechanisms to learn node representations, proving particularly effective in inductive settings.

In the case of heterogeneous graphs involving multiple node types, GRLs methods face additional challenges. HetGNN [43] addresses these challenges by sampling neighbors of different types from a node's neighborhood and aggregating them with a bi-LSTM. RGCN [8] extends GCNs by incorporating edge and node types, allowing for more flexible graph representation.

Other techniques leverage attention mechanisms to handle heterogeneous graphs better. GATNE [7] tackles multiplex graphs by integrating edge attributes, while HGT [21] uses a transformer-based, type-specific attention mechanism to manage diverse node and edge types. HiNormer [30] introduces two encoders – a local structure encoder and a heterogeneous relation encoder – tailored to handle the complexity of heterogeneous graphs. Methods like HAN [39] and MAGNN [16] 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.

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 [14] recommends datasets based on a lengthy textual description, such as the abstract of a paper. Another approach introduced a neural bi-encoder model [36] to recommend datasets given a set of publication abstracts. Other approaches focus on graph exploration to identify relevant datasets. Wang et al. [40, 41] 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 [2] 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 [1], extracted and processed ad-hoc for these tasks. HetGNN [43] 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. Irrera et al. [22] reproduced these methods and found them less effective in real-world scenarios. Their analysis also tested the generalizability of the dataset recommendation task using three popular recommendation baselines (TopPop, BPR, and LightGCN), underscoring the need for a new, robust, and versatile method that can perform well in real-world settings.

## 3 The Dataset Recommendation Task

This section provides a formal definition of SKG, heterogeneous GRL, and the dataset recommendation task with its open challenges.

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

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

*where* $\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{R}$ *are the sets of nodes, edges, node types, edge types, 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* $a \in \mathcal{A}$*, while* $\beta:\mathcal{E} \to \mathcal{R}$ *is an edge type function that maps each edge* $e \in \mathcal{E}$ *to an edge type* $r \in \mathcal{R}$*. An SKG is* heterogeneous *if it has multiple node and edge types such that* $|\mathcal{A}| + |\mathcal{R}| > 2$.

An SKG is a Knowledge Graph (KG) tailored to the scholarly domain, representing academic entities and their relationships. Generally, an SKG incorporates both structural (edges) and semantic (nodes and their types) information, providing 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 output is a datasets ranking such that* $\forall \{d_i, d_j\} \in \mathcal{V}_d$*, if* $\rho(q, d_i) \geq \rho(q, d_j)$*, then* $d_i$ *is ranked above* $d_j$.

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. We study the dataset recommendation task as a node-representation learning problem.

DEFINITION 3. *Given an 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* $\rho$.

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.

The dataset recommendation task faces significant challenges due to the sparse and incomplete 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 [22]. As a result, the task involves accurately retrieving a small set of relevant datasets from a large pool of irrelevant ones.

## 4 MM-SAN

This section presents the MM-SAN architecture illustrated in Figure 1 comprizing augmentation, sampling, and aggregation.

*Augmentation.* The augmentation phase (❶) is performed offline to mitigate the sparsity issue common in SKGs, enhancing *robustness (RQ1)* by introducing new nodes that increase the graph's connectivity. Given an input SKG, MM-SAN extracts the available textual metadata from the *publication* and *dataset* nodes. Entity linking [3] is applied to identify and resolve ambiguities of entities within the text. Topic modeling is performed with BERTopic [18], which leverages transformer-based embeddings and c-TF-IDF to group similar documents based on common topics. Topics represent broader themes that serve as hubs within the SKG, linking multiple disconnected components. In contrast, entities are more specific and, when shared among nodes, indicate that those nodes likely cover similar or closely-related content. This distinction enriches the graph's structure by using topics to increase overall connectivity and entities to highlight precise content overlaps between nodes. Entities and topics are then incorporated into the SKG as new nodes of type *entity* and *topic*.

For each node of the enriched SKG, MM-SAN extracts two vectors: one derived from the graph topology using node2vec [19], and the other from textual metadata using the all-MiniLM-L6-v2 sentence transformer [1] for longer textual metadata and phrase-BERT [38] for shorter texts, such as topics.

*Sampling.* Given a target node $v_t$ in the enriched SKG, phase ❷ aims to explore the neighborhood of $v_t$ and collect a representative set of neighbors of different types, ensuring that the number of nodes of each kind is balanced. Random sampling is a standard method for sampling neighbors, and it is used in approaches like GraphSAGE. However, this method may yield an unrepresentative sample because SKGs are heterogeneous, with a highly unbalanced distribution of node types. In practical scenarios, author nodes often lack disambiguation, resulting in many isolated nodes. In contrast, there are generally fewer venue nodes, which tend to be more interconnected. To address this issue, 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 $\text{walk}_j$ is one random walk, $m_j$ is the number of publication and dataset nodes in the walk $j$, $v_i \in \{\mathcal{V}_p \cup \mathcal{V}_d\}$ 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, $cos(\mathbf{v}_t, \mathbf{v}_i)$ is the cosine similarity, and $d$ 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 $k$ chosen 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 modes, 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 [34] to effectively combine neighbor embeddings based on both textual and topological features. Cross-attention allows the model to integrate information from both modes; 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 linear projects the text-based and topology based embeddings (FC blocks in Figure 1). Then, MM-SAN combines three input matrices: Query (**Q**), Key (**K**), and Value (**V**), of sizes
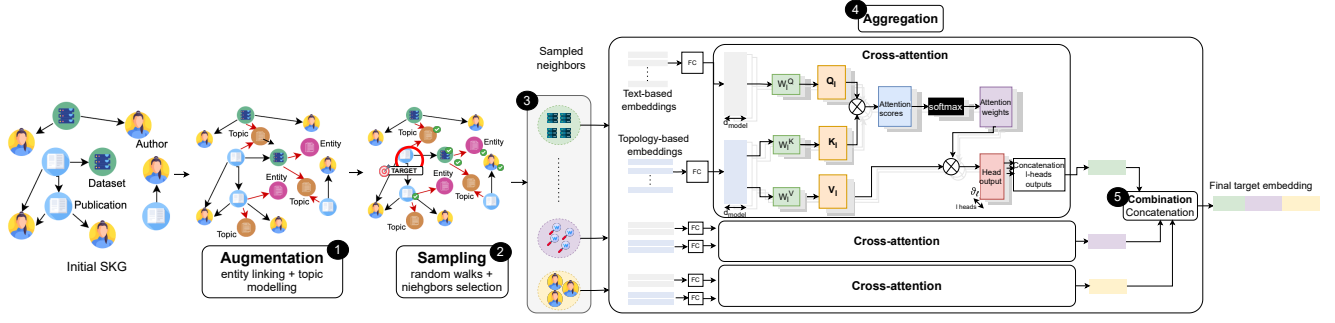
**Figure 1: The MM-SAN Architecture comprizing the collection and augmentation, neighbors sampling, and the aggregation phases.**

$(n_Q, d_{model})$, $(n_K, d_{model})$, $(n_V, d_{model})$ 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: $\mathbf{Q}_\ell = \mathbf{QW}_\ell^Q$, $\mathbf{K}_\ell = \mathbf{KW}_\ell^K$, and $\mathbf{V}_\ell = \mathbf{VW}_\ell^V$, 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 $\vartheta_\ell$ 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 has been 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 (**5**) to obtain the final target node embedding representation.

We experimented with various embedding combination methods, including multihead attention [34], bi-LSTM [17], and GRU [11], both as replacements for cross-attention and for the final concatenation step. These results are presented in the ablation study (cfr. 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.

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 $\sigma$ is the sigmoid function, $\mathbf{h}_u$ and $\mathbf{h}_v$ are the embedding representations for publication $u$ and dataset $v$, $\mathcal{E}_{p,d}$ 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 directly connected.

## 5 Experimental setup

We evaluate MM-SAN across three increasingly complex settings. In the **transductive** setting, all publications and datasets are available during training, the SKG is static, and the model leverages the complete structure and relationships to make the recommendations. This is the classical setting adopted in previous studies [16, 30, 43]. In the **semi-inductive** setting, new publications are introduced at test time, while all datasets have been seen during training. We rank the existing datasets based on their relevance to the new publications. It is worth noting that our semi-inductive setting is sometimes called "inductive" in the literature [43]. Lastly, the **fully inductive** setting is where new publications and new datasets are added in the test phase. The fully inductive scenario reflects real-world conditions, where the SKG evolves with a continuous introduction of new outcomes, requiring recommendations to consider new publications and datasets.

**Table 1: Nodes and edges statistics of the used datasets.**

| Dataset | Node | | Edges | |
|---|---|---|---|---|
| MES | Publications (P): 2, 1$K$<br>Datasets (D): 3$K$<br>Authors (A): 9, 5$K$ | | P–P: 450<br>D–D: 1, 2$K$<br>P–D: 2, 5$K$ | P–A: 9, 8$K$<br>D–A: 10, 2$K$ |
| PubMed | Publications (P): 42, 6$K$<br>Datasets (D): 33, 8$K$<br>Authors (A): 334$K$<br>Venues (V):6, 7$K$<br>Organizations (O): 14, 8$K$<br>Keywords (K): 10, 8$K$ | | P–P: 18, 3$K$<br>D–D: 8$K$<br>P–D: 29, 5$K$<br>P–A: 180$K$<br>D–A: 94, 6$K$ | P–K: 6, 6$K$<br>D–K: 17$K$<br>P–V: 42, 6<br>P–O: 87, 5$K$<br>D–O: 515 |

Further, we test the models' robustness without textual metadata under two different metadata conditions for each setting. The **ideal metadata setup** represents the ideal scenario where all datasets are fully described by complete metadata – i.e., all the datasets have a title and a description. The **real metadata setup** represents intermediate scenarios where only 75%, 50%, or 25% of the datasets have complete metadata, with the datasets containing metadata randomly selected. This setting reflects real-world conditions, as in real SKGs, datasets may or may not have comprehensive and descriptive metadata.

The implementation of MM-SAN and the related datasets are available at: https://anonymous.4open.science/r/MM-SAN.

We employ two SKGs extracted from the OAG [29]: MES [23] and PubMed[2]. MES is a curated scientific graph interconnecting publications, datasets, and authors; it represents a reliable snapshot of the scientific activity of the European Marine Science (MES) community of OpenAIRE. PubMed is a larger subgraph of the OAG that also interconnects publications and datasets, authors, venues, organizations, and keywords. Unlike MES, PubMed is not curated.

The statistics of the SKGs are presented in Table 1. We partitioned the $P - D$ (publication-dataset) edges into five subsets: training, validation, and three test sets – transductive, semi-inductive, and fully inductive. The $P - D$ edges for the three test sets were randomly selected from the original MES and PubMed datasets. Publications and datasets assigned to the semi-inductive and fully inductive test sets were removed from the original MES and PubMed datasets. After filtering, we extracted a set of $P - D$ edges from MES and PubMed for validation, while the remaining data was used for training. In MES, we allocated $2K$ edges for training, 155 for validation, and 155 for each of the three test sets. In PubMed, we used $26.8K$ edges for training, $1.8K$ for validation, and $1.8K$ for each test set.

To evaluate the model, we treated the publications in the $P - D$ edges of each test set as queries. These SKGs were also utilized in a previous study [22] that focused on the dataset recommendation task. We use evaluation metrics suited to recommendation tasks, specifically Recall and nDCG with cut-off at $k = 5$. Recall quantifies the proportion of relevant items among the top-5 recommendations, while nDCG assesses their quality by considering both relevance and ranking. The online repository also includes results for nDCG, precision, recall, F1, hits, and mean reciprocal rank, evaluated at cut-offs of 5, 10, 20.

*Parameters of MM-SAN.* We generated embeddings for publication and dataset nodes using the `all-MiniLM-L6-v2` pre-trained sentence transformer model. For keywords, entities, and topics – each ranging from one to five words – we assigned embeddings produced by `phrase-BERT`. Additionally, we computed topological embeddings for all nodes using `node2vec` [19]. For entity extraction, we annotated the textual metadata of publications and datasets using DBPedia Spotlight [31], setting a confidence threshold of 0.75. Topics were identified with BERTopic [18], using a minimum cluster size of two documents, and further refined with KeyBERT. MM-SAN was trained for 100 epochs with early stopping to minimize training time, with learning rate $1 \times 10^{-5}$ and a 1024 mini-batch size. For each target node, we collected five neighboring nodes of the

publication or dataset type, five of the keyword, topic, or entity type, and five of the venue, author, or organization type. In the cross-attention attention mechanism, we employed 8 heads and an output dimension of 128. The number of epochs, learning rate, mini-batch size, number of attention heads, and number of neighbors sampled for each type were optimized through a grid search.

*Baselines.* We compare MM-SAN against five general-purpose inductive graph-based baselines (also used in [21, 39, 43]) and one text-based baseline, reporting the best results obtained through grid search. It is important to note that none of these baselines were specifically designed for the dataset recommendation task. All graph-based methods are general-purpose models that operate on attributed graphs. This approach enables a clearer contextualization of MM-SAN's contribution by benchmarking it against broadly adopted, domain-agnostic models. The node embeddings were initialized by concatenating the textual embeddings using the `all-MiniLM-L6-v2` pre-trained Sentence Transformer with those from node2vec. GraphSAGE [20] (SAGE for brevity) and GAT [35] were trained for 200 epochs with a learning rate of $10^{-5}$, and early stopping. The output dimension was set to 128. Similarly, HAN [39] and HGT [21] were trained for 200 epochs with a learning rate of $10^{-4}$, also utilizing early stopping, with an output dimension of 128. HetGNN [43] (HGNN for brevity) was trained for 100 epochs with early stopping, a learning rate of $10^{-5}$, and 10 neighbors per node type, as specified in the original paper. The baselines used in this study have been widely employed in previous research on GRL, and they are applicable in both transductive and inductive settings. To implement the baselines, we used PyTorch Geometric [15]. All the graph-based baselines have been optimized through a grid search.

The text-based baseline is a Sentence-Transformer (ST-T for brevity) and consists of the dot product between the pre-trained embedding of the publication and the one of the dataset: $\text{ST-T}(u, v) = \mathbf{v}_u \cdot \mathbf{v}_v$ where $\mathbf{v}_u$ and $\mathbf{v}_v$ are the embeddings of the publication and dataset respectively obtained with `all-MiniLM-L6-v2` pretrained model. This baseline is available when all the datasets have complete metadata in transductive and semi-inductive settings and when all the datasets are available during training. This design choice is because the ranking can only be generated when all the datasets are available, which occurs in transductive and semi-inductive scenarios, and when the textual content is available. Some baselines were excluded from this study. GATNE [7] was not considered because it requires multiple relation types within the same set of nodes and user metadata to provide effective recommendations, which our datasets do not provide since we recommend datasets to papers, not to users. MAGNN [16] was excluded since it was designed for the node classification task and was not intended for the inductive settings. HiNormer [30] was also not included, as it was not designed for link prediction and recommendation tasks.

We also compared MM-SAN with RGCN [8], HVGAE [2], VGAE [26], and Metapath2Vec [12]. These models performed significantly worse than most of the baselines reported in this work and are limited to transductive settings, making them unsuitable for inductive scenarios. Therefore, their results are not included in this study.

---

[2]MES and PubMed datasets are available at https://figshare.com/s/1e11a6f03fbf97d61936?file=48494335

**Table 2: Recall@5 (R@5) and nDG@5 (N@5) score over the MES dataset in transductive (Tran), semi-inductive (Semi), and inductive (Ind) settings. 100% indicates that all the datasets have textual metadata. [75%, 50%, 25%] indicates the percentage of datasets having textual metadata.**

| MES | | | GAT | SAGE | HGT | HAN | HGNN | ST-T | MM-SAN |
|---|---|---|---|---|---|---|---|---|---|
| 100% | Tran | R@5 | 0.625 | 0.489 | 0.023 | 0.008 | 0.581 | 0.693 | **0.727** |
| | | N@5 | 0.475 | 0.398 | 0.020 | 0.009 | 0.438 | **0.651** | 0.546 |
| | Semi | R@5 | 0.621 | 0.446 | 0.007 | 0.000 | 0.587 | 0.696 | **0.719** |
| | | N@5 | 0.479 | 0.316 | 0.003 | 0.000 | 0.417 | **0.644** | 0.524 |
| | Ind | R@5 | 0.590 | 0.439 | 0.000 | 0.000 | 0.582 | - | **0.687** |
| | | N@5 | 0.442 | 0.315 | 0.000 | 0.000 | 0.422 | - | **0.519** |
| 75% | Tran | R@5 | 0.481 | 0.352 | 0.022 | 0.008 | 0.473 | - | **0.623** |
| | | N@5 | 0.370 | 0.294 | 0.011 | 0.006 | 0.356 | - | **0.447** |
| | Semi | R@5 | 0.483 | 0.332 | 0.000 | 0.000 | 0.515 | - | **0.598** |
| | | N@5 | 0.404 | 0.247 | 0.000 | 0.000 | 0.373 | - | **0.442** |
| | Ind | R@5 | 0.409 | 0.325 | 0.000 | 0.000 | 0.523 | - | **0.598** |
| | | N@5 | 0.324 | 0.235 | 0.000 | 0.000 | 0.383 | - | **0.468** |
| 50% | Tran | R@5 | 0.361 | 0.286 | 0.015 | 0.007 | 0.443 | - | **0.539** |
| | | N@5 | 0.319 | 0.212 | 0.012 | 0.006 | 0.324 | - | **0.398** |
| | Semi | R@5 | 0.349 | 0.240 | 0.000 | 0.000 | 0.484 | - | **0.518** |
| | | N@5 | 0.277 | 0.209 | 0.000 | 0.000 | 0.352 | - | **0.359** |
| | Ind | R@5 | 0.321 | 0.196 | 0.000 | 0.000 | 0.434 | - | **0.526** |
| | | N@5 | 0.258 | 0.150 | 0.000 | 0.000 | 0.339 | - | **0.391** |
| 25% | Tran | R@5 | 0.230 | 0.190 | 0.007 | 0.003 | 0.336 | - | **0.467** |
| | | N@5 | 0.203 | 0.165 | 0.007 | 0.002 | 0.242 | - | **0.329** |
| | Semi | R@5 | 0.206 | 0.145 | 0.000 | 0.000 | 0.367 | - | **0.454** |
| | | N@5 | 0.172 | 0.125 | 0.000 | 0.000 | 0.270 | - | **0.306** |
| | Ind | R@5 | 0.159 | 0.075 | 0.000 | 0.000 | 0.322 | - | **0.453** |
| | | N@5 | 0.128 | 0.069 | 0.000 | 0.000 | 0.233 | - | **0.321** |

**Table 3: Recall@5 (R@5) and nDG@5 (N@5) score over the PubMed dataset in transductive (Tran), semi-inductive (Semi), and inductive (Ind) settings. 100% indicates that all the datasets have textual metadata. [75%, 50%, 25%] indicates the percentage of datasets having textual metadata.**

| PubMed | | | GAT | SAGE | HGT | HAN | HGNN | ST-T | MM-SAN |
|---|---|---|---|---|---|---|---|---|---|
| 100% | Tran | R@5 | 0.241 | 0.239 | 0.016 | 0.030 | 0.284 | **0.590** | 0.380 |
| | | N@5 | 0.181 | 0.179 | 0.017 | 0.023 | 0.202 | **0.523** | 0.272 |
| | Semi | R@5 | 0.231 | 0.221 | 0.006 | 0.010 | 0.246 | **0.596** | 0.373 |
| | | N@5 | 0.171 | 0.169 | 0.005 | 0.008 | 0.177 | **0.527** | 0.266 |
| | Ind | R@5 | 0.196 | 0.187 | 0.000 | 0.000 | 0.271 | - | **0.344** |
| | | N@5 | 0.134 | 0.133 | 0.000 | 0.000 | 0.199 | - | **0.246** |
| 75% | Tran | R@5 | 0.201 | 0.183 | 0.006 | 0.012 | 0.241 | - | **0.319** |
| | | N@5 | 0.149 | 0.139 | 0.005 | 0.009 | 0.171 | - | **0.227** |
| | Semi | R@5 | 0.186 | 0.174 | 0.000 | 0.007 | 0.237 | - | **0.319** |
| | | N@5 | 0.140 | 0.133 | 0.000 | 0.007 | 0.174 | - | **0.222** |
| | Ind | R@5 | 0.145 | 0.136 | 0.000 | 0.000 | 0.202 | - | **0.289** |
| | | N@5 | 0.104 | 0.097 | 0.000 | 0.000 | 0.145 | - | **0.205** |
| 50% | Tran | R@5 | 0.156 | 0.145 | 0.005 | 0.007 | 0.203 | - | **0.276** |
| | | N@5 | 0.116 | 0.111 | 0.005 | 0.007 | 0.140 | - | **0.194** |
| | Semi | R@5 | 0.144 | 0.129 | 0.000 | 0.001 | 0.201 | - | **0.259** |
| | | N@5 | 0.110 | 0.100 | 0.000 | 0.002 | 0.142 | - | **0.181** |
| | Ind | R@5 | 0.096 | 0.097 | 0.000 | 0.000 | 0.167 | - | **0.242** |
| | | N@5 | 0.086 | 0.079 | 0.000 | 0.000 | 0.120 | - | **0.169** |
| 25% | Tran | R@5 | 0.113 | 0.108 | 0.000 | 0.000 | 0.156 | - | **0.230** |
| | | N@5 | 0.084 | 0.089 | 0.000 | 0.000 | 0.108 | - | **0.155** |
| | Semi | R@5 | 0.098 | 0.096 | 0.000 | 0.000 | 0.152 | - | **0.239** |
| | | N@5 | 0.077 | 0.075 | 0.000 | 0.000 | 0.107 | - | **0.167** |
| | Ind | R@5 | 0.053 | 0.057 | 0.000 | 0.000 | 0.129 | - | **0.205** |
| | | N@5 | 0.038 | 0.043 | 0.000 | 0.000 | 0.090 | - | **0.138** |

## 6 Results

Tables 2 and 3 report the results for the MES and PubMed datasets under three conditions: transductive, semi-inductive, and full-inductive.

Each setting was evaluated with progressively lower proportions of datasets containing metadata: 100%, 75%, 50%, and 25%.

Table 2 shows that MM-SAN reports strong performance on the MES dataset, consistently outperforming all graph-based baselines in recall@5 and nDCG@5 across all scenarios. While the ST-T baseline achieves slightly higher nDCG@5 scores in transductive and semi-inductive setups, MM-SAN remains competitive, showcasing its robustness across different conditions. All models, including MM-SAN, perform highest in the transductive setting with full metadata availability (100% row in Table 2). However, this scenario is the least representative of real-world academic contexts. MM-SAN maintains stable performance even in more challenging semi-inductive and full-inductive settings, where predictions must be made for unseen nodes. When tested under more realistic conditions – where metadata availability is limited – MM-SAN, like all models, experiences a slight performance decline as the proportion of datasets with metadata decreases (from 75% to 25%). 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 method that consistently maintains a recall@5 above 0.450 and an nDCG@5 score of at least 0.320, even when only 25% of the dataset nodes have descriptive textual metadata. In this respect, in the most challenging condition (inductive settings and 25% of available metadata), we see that MM-SAN improves the recall of the best baseline (HGNN) of the 41% and the nDCG of the 38%. This indicates that the MM-SAN multimodal approach leveraging topology-based features successfully reduces the effect of missing text-based features, significantly affecting the performance of other baselines. HAN and HGT achieve recall and nDCG scores below 0.1 in the transductive setting and 0.0 in the inductive setting. Notably, SAGE and GAT, which are designed for homogeneous graphs, demonstrate greater robustness compared to HAN and HGT, which are tailored for heterogeneous graphs. Interestingly, GAT achieves the highest performance in the ideal scenario when all the metadata have descriptive metadata and are in a transductive setting. As expected, HGNN, which is designed explicitly for heterogeneous SKGs, emerges as the second top-performing model after MM-SAN in a real scenario when 75%, 50%, and 25% of metadata are available in inductive settings.

As shown in Table 3, MM-SAN exhibits strong robustness when applied to real-world, sparse, large, and uncurated SKGs like PubMed. Similar to the MES dataset, all tested methods experience a performance decline in semi-inductive and full-inductive settings as textual metadata becomes less available – a natural outcome given the increasing difficulty of the task.

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 and nDCG. This underscores the significant role of text-based features in generating effective recommendations when metadata is abundant. However, ST-T's effectiveness is limited to ideal conditions, making it less suited to addressing the open challenges of real-world SKGs.

Conversely, MM-SAN demonstrates the highest resilience without textual metadata and in the inductive setting. It is the only method that consistently maintains a recall@5 above 0.200 and an

**Table 4: Recall@5 and nDG@5 scores over the MES dataset in transductive and 100% metadata available setting. The column "no augmentation" refers to the models run excluding topics and entities; "augmentation" reports the results on the enriched graph.**

| Method | Augmentation | | No Augmentation | |
|--------|------|------|------|------|
| | R@5 | N@5 | R@5 | N@5 |
| GAT | 0.578 | 0.422 | 0.625 | 0.475 |
| SAGE | 0.472 | 0.312 | 0.489 | 0.398 |
| HGT | 0.020 | 0.020 | 0.023 | 0.020 |
| HAN | 0.008 | 0.009 | 0.008 | 0.009 |
| HGNN | 0.601 | 0.488 | 0.581 | 0.438 |
| MM-SAN | **0.727** | **0.546** | **0.642** | **0.476** |

nDCG@5 above 0.135 across all settings, highlighting its adaptability and robustness in more challenging and realistic scenarios.

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 58% in recall@5 and 53% in nDCG@5. Like the results observed on the MES dataset, HAN and HGT exhibit the lowest performance on the PubMed dataset, highlighting their limited effectiveness in real-world SKGs.

To further validate the consistency and robustness of our experimental findings, we re-evaluated the strongest baseline methods — namely SAGE, GAT, and HGNN — by replacing their original sampling strategies with the one proposed in MM-SAN. This replacement led to performance improvements from 0.1% to 25% in recall and from 0.01% to 28% in nDCG, underscoring the significant influence of the sampling phase on the overall effectiveness of graph-based baselines. However, despite the performance increase, none of the baselines outperformed MM-SAN. For completeness, we report the detailed results of this additional evaluation in the online appendix [3].

## 7 Ablation study

*Components analysis.* We first assess how the augmentation phase ❶ – specifically, the integration of entity and topic nodes to reduce overall sparsity – affects system effectiveness. Table 4 presents the performance of six systems with (augmentation column) and without (no augmentation column) this phase on the MES dataset. For GAT and GraphSAGE, incorporating diverse node types diminishes the effectiveness of these models. In SKGs, node embeddings vary by type; thus, aggregating embeddings with different features can negatively impact overall performance. 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. The results on the PubMed dataset, confirm the same trends observed on MES. These results are available in the online repository.

In summary, the augmentation phase proves useful for models like MM-SAN and HGNN that are tailored for heterogeneous graphs

---

[3] https://anonymous.4open.science/r/MM-SAN/rebuttal/RecSys_2025_annex.pdf

**Table 5: Analysis on sampling phase. Recall@5 and nDG@5 scores over the MES and PubMed datasets in transductive and 100% metadata available setting. With RS-$i$ we denote Random Sampling (RS) and $i$ refers to the number of hops considered.**

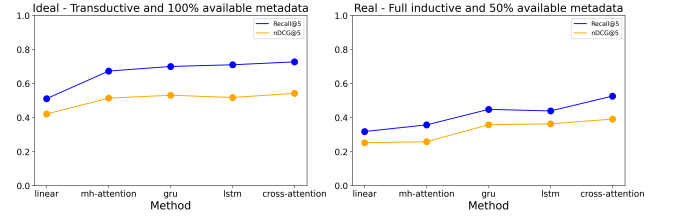| Method | MES | | PubMed | |
|--------|------|------|------|------|
| | R@5 | N@5 | R@5 | N@5 |
| RS-1 | 0.644 | 0.534 | 0.314 | 0.243 |
| RS-2 | 0.637 | 0.471 | 0.313 | 0.225 |
| RS-3 | 0.602 | 0.465 | 0.295 | 0.212 |
| MM-SAN | **0.727** | **0.546** | **0.380** | **0.270** |



**Figure 2: Component-based analysis on the aggregation phase of the pipeline.**

and leverage topological features. However, it may not benefit – and can even hinder – models not explicitly designed to handle heterogeneous structures.

We then investigate the performance of MM-SAN when replacing the original random walk-based sampling method with random sampling ❷ in Figure 1. We explored the effect of random sampling at different hop distances (k-hops) to evaluate how sampling neighbors from progressively deeper levels influences the results. The results for MES and PubMed datasets in an ideal setup – transductive and 100% metadata available setting – are shown in Table 5. We see that on the MES dataset, sampling neighbors at greater depths improves performance. On the other hand, on PubMed there is the opposite tendency, and the performances remains similar independently of the hops. This is reasonable: PubMed is extremly sparse and imbalanced which may lead to the sampling of the nodes of the same type; MES instead, is more connected, dense and curated. Nonetheless, the methods we proposed in our pipeline, based on random walks, yields superior results.

To evaluate the effectiveness of various aggregation methods within MM-SAN, we conducted experiments comparing cross-attention, bi-LSTM, GRU, linear transformation, and multihead attention. Performance was measured using nDCG@5 and recall@5 in both an ideal scenario (100% metadata availability in a transductive setting) and a more realistic scenario (50% metadata availability in a full-inductive setting). Figure 2 indicates that cross-attention achieves the highest robustness, followed by bi-LSTM and GRU, while linear transformation and multihead attention yield the lowest performance in the dataset recommendation task. These findings suggest that cross-attention is particularly well-suited for multimodal learning tasks, such as dataset recommendation as handled by MM-SAN, where integrating diverse data sources is crucial. This aligns with
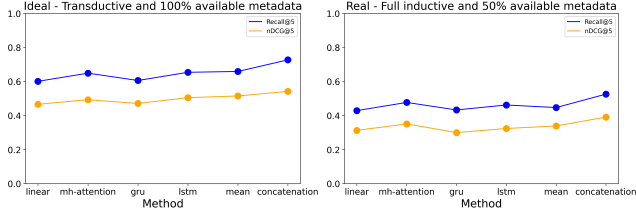
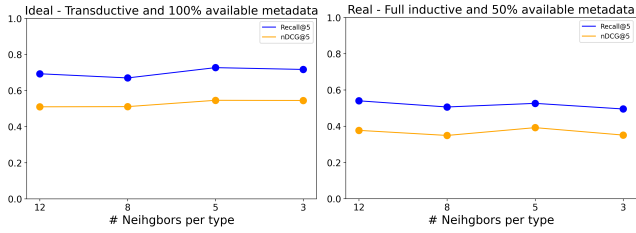**Figure 3: Component-based analysis on the combination phase of the pipeline.**



**Figure 4: Hyperparameter analysis: number of neighbors of a target node sampled for each node type.**

research indicating that cross-attention mechanisms can improve performance in multimodal applications by facilitating dynamic interactions between modalities. Finally, we evaluated the effectiveness of MM-SAN testing different combination methods in place of concatenation (❺ in Figure 1). In MM-SAN, we use concatenation to combine embeddings. We tested the following combination approaches: bi-LSTM, GRU, multihead-attention, linear projection, and mean pooling. In Figure 3, we illustrate the results of our experiments in two ideal and real scenarios on the MES dataset. Concatenation is highly effective, while GRU and linear projection provide the lowest results.

*Hyperparameter analysis.* One key efficiency factor, closely tied to SKGs, is the number of neighbors sampled for each node type when targeting a node $v_t$. SKGs exhibit an imbalanced distribution of node types, making selecting the most representative nodes essential to ensure a well-balanced and meaningful set. Sampling too many nodes can introduce noise by adding irrelevant or redundant information, while choosing too few may result in an incomplete and unrepresentative set of neighbors. Figure 4 presents our analysis on the MES dataset. As in previous experiments, we compare two settings: the ideal and the real-world scenario. In both cases, performance varies slightly with the number of neighbors considered, peaking when selecting five neighbors per type. These minor variations suggest that cross-attention effectively mitigates the impact of additional or missing nodes, demonstrating its adaptability and robustness in real-world SKGs.

## 8 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 addressed this challenge by proposing MM-SAN, a model based on multimodal representation learning for enriching SKGs and recommending datasets to publications. Our approach addresses key dataset recommendation challenges, aiming to enhance data discovery and reuse within the scientific research process.

We conduct extensive experiments across three settings (transductive, semi-inductive, 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 (e.g., PubMed) and smaller, well-curated datasets (e.g., 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 full-inductive settings, making it suitable for dynamic real-world applications. MM-SAN proves to be the most effective and versatile model, capable of handling missing textual metadata by leveraging topological features. In contrast, existing methods struggle to fully exploit high-quality textual metadata. Future work could explore mixture-of-experts architectures to address this. Additionally, we plan to extend MM-SAN to author and venue recommendation, examining its performance on nodes with different roles and connectivity patterns.

# References

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

[2] 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. https://doi.org/10.1109/ICDM.2019.00011

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

[4] 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. https://doi.org/10.1145/3308558.3313685

[5] 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). https://doi.org/10.1093/DATABA/BAAA022

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

[7] 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. https://doi.org/10.1145/3292500.3330964

[8] J. Chen, H. Hou, J. Gao, Y. Ji, and T. Bai. 2019. RGCN: recurrent graph convolutional networks for target-dependent sentiment analysis. In *Knowledge Science, Engineering and Management - 12th International Conference, KSEM 2019, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 11775)*. Springer, 667–675. https://doi.org/10.1007/978-3-030-29551-6_59

[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. https://doi.org/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. https://doi.org/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] 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. https://doi.org/10.1145/3097983.3098036

[13] 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.

[14] 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. https://doi.org/10.1145/3459637.3482166

[15] 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

[16] 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. https://doi.org/10.1145/3366423.3380297

[17] 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. https://doi.org/10.1109/TPAMI.2008.137

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

[19] 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. https://doi.org/10.1145/2939672.2939754

[20] 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.

[21] 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. https://doi.org/10.1145/3366423.3380027

[22] 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. https://doi.org/10.1145/3640457.3688071

[23] 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. https://doi.org/10.1145/3597310

[24] 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. https://doi.org/10.1145/3360901.3364435

[25] 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

[26] 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

[27] 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. https://doi.org/10.1007/978-3-031-77844-5_3

[28] Q. Lv, M. Ding, Q. Liu, Y. Chen, W. Feng, S. He, C. Zhou, J. Jiang, Y. Dong, and J. Tang. 2021. Are we really making much progress?: Revisiting, benchmarking and refining heterogeneous graph neural networks. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, Feida Zhu, Beng Chin Ooi, and Chunyan Miao (Eds.). ACM, 1150–1160. https://doi.org/10.1145/3447548.3467350

[29] 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*. https://doi.org/10.5281/zenodo.7488618

[30] 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. https://doi.org/10.1145/3543507.3583493

[31] 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. https://doi.org/10.1145/2063518.2063519

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

[33] 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. https://doi.org/10.1145/1401890.1402008

[34] 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

[35] 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

[36] 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. https://doi.org/10.18653/v1/2023.acl-long.573

[37] K. Wang, Z. Shen, C. Huang, C.H. Wu, Y. Dong, and A. Kanakia. 2020. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies* 1, 1 (2020), 396–413. https://doi.org/10.1162/qss_a_00021

[38] 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. https://doi.org/10.18653/V1/2021.EMNLP-MAIN.846

[39] 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. https://doi.org/10.1145/3308558.3313562

[40] 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. https://doi.org/10.1007/978-3-031-10986-7_38

[41] 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. https://doi.org/10.3233/DS-220056

[42] 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. https://doi.org/10.1038/sdata.2016.18

[43] 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. https://doi.org/10.1145/3292500.3330961