Enhancing Multimodal Knowledge Graph Representation Learning through Triple Contrastive Learning

Yuxing Lu[1,2], **Weichen Zhao**[3], **Nan Sun**[4] and **Jinzhuo Wang**[1,*]
[1]Department of Big Data and Biomedical AI, College of Future Technology, Peking University
[2]Tencent AI Lab
[3]School of Computer Science and Technology, Soochow University
[4]School of Computer Science and Technology, Huazhong University of Science & Technology
wangjinzhuo@pku.edu.cn

Abstract

Multimodal knowledge graphs incorporate multimodal information rather than pure symbols, which significantly enhance the representation of knowledge graphs and their capacity to understand the world. Despite these advances, the existing multimodal fusion technique still faces significant challenges in representing modalities and fully integrating the diverse attributes of entities, particularly when dealing with more than one modality. To address this issue, the article proposes a Knowledge Graph Multimodal Representation Learning (KG-MRI) method. This method utilizes foundation models to represent different modalities and incorporates a triple contrastive learning model and a dual-phase training strategy to effectively fuse the different modalities with knowledge graph embeddings. We conducted comprehensive comparisons with several knowledge graph embedding methods to validate the effectiveness of our KG-MRI model. Furthermore, validation on a real-world Non-Alcoholic Fatty Liver Disease (NAFLD) cohort demonstrated that the vector representations learned through our methodology have enhanced representational capabilities and can remove batch effects, showing promise for broader applications in complex multimodal environments.

1 Introduction

A knowledge graph (KG) is a complex semantic network that includes various entities, each with unique attributes, linked by edges representing different semantic relationships [Ji *et al.*, 2021]. These graphs play a vital role in many contemporary applications, especially in recommendation systems [Guo *et al.*, 2020] and natural language-based question answering [Chen *et al.*, 2020b]. With the rapid development of graph management tools and analysis methods, knowledge graphs are increasingly used in scientific research, with notable applications in fields such as genomics, proteomics, and systems biology.

However, most existing knowledge graphs are represented with pure symbols in the form of natural text or identifiers, which does not align with how humans perceive and process knowledge in the real world [Zhu *et al.*, 2022]. Human cognition stores knowledge in a variety of non-symbolic modalities. For example, when a person observes an apple, they do not retrieve a symbolic identifier but rather access a rich array of sensory and contextual information, including the apple's taste, smell, color, and texture. This multimodal integration is crucial for holistic understanding. Grounding entities in a knowledge graph with these diverse modalities can enable it to mirror human-like cognition more accurately during representation learning.

Current knowledge graph embedding (KGE) models tries to represent all entities and relations with low-rank continuous vectors that preserve the graph's inherent structure and capture diverse contextual information [Wang *et al.*, 2017]. This makes them easier to work with and apply to various machine learning and deep learning tasks [Mohamed *et al.*, 2021]. Common KGE methods includes translation models like TransE [Bordes *et al.*, 2013], RotatE [Sun *et al.*, 2019], etc., semantic match models like DisMult [Yang *et al.*, 2014], SimplE [Kazemi and Poole, 2018], etc., and neural network models like ConvE [Dettmers *et al.*, 2018], ER-MLP [Dong *et al.*, 2014], etc. Recently, researchers have extended KGE methods to multimodal knowledge graphs. They proposed a series of multimodal KGE methods, ranging from conceptual frameworks [Lu *et al.*, 2022; Cao *et al.*, 2022; Wang *et al.*, 2019] to practical applications [Yao *et al.*, 2023; Li *et al.*, 2023]. Despite these advancements, the integration of more than two modalities' information remains a challenge.

Recent advances in foundation models offer an ideal solution for representing different modalities. For example, for text or natural language modalities, semantic vector representations can be obtained through large language models (LLMs) such as GPT-4 [Achiam *et al.*, 2023], LLama [Touvron *et al.*, 2023], BERT [Devlin *et al.*, 2018], and others. For image modalities, vector representations containing image information can be acquired through Vision Foundation Models (VFMs) like MAE [He *et al.*, 2022], ImageBind [Girdhar *et al.*, 2023], and similar models. Additionally, there has been a surge in domain-specific foundation models, such as CNBERT [Lu *et al.*, 2023a], ChemBERTa-2 [Ahmad *et al.*, 2022] and scGPT [Cui *et al.*, 2024]. Additionally, contrastive

learning has been proven to be an effective method for aligning information from different modalities. In the face of representations from foundation models in various domains, the CLIP model [Radford *et al.*, 2021] was the first to use contrastive learning to bring closer the representations of different modalities that convey the same semantics, and to distance the representations of different semantics. For multimodal information in knowledge graphs, using contrastive learning is also an effective approach [Zhang *et al.*, 2023; Yang *et al.*, 2022; Liang *et al.*, 2022].

In this paper, we propose a multimodal representation integration model for knowledge graph representation learning (KG-MRI). This approach aims to help incorporate different modalities' representations from foundation models with knowledge graph embeddings. To further optimize alignment among these varied representations, we employ a triple contrastive learning (TCL) module and apply a dual-phase training strategy. Comprehensive comparison experiments with various KGE models demonstrate the superiority of the KG-MRI algorithm. In addition, we conducted an empirical analysis on a non-alcoholic fatty liver disease (NAFLD) real-world clinical cohort, illustrating the practical application of KG-MRI in clinical decision-making and removing batch effects. Altogether, our contributions are three-fold:

- We proposed a multimodal representation integration method in knowledge graph representation learning to fully integrate different information of an entity.
- We introduce a triple contrastive learning module and a dual-phase training strategy in aligning multimodal representation. We demonstrate the efficacy of KG-MRI in comparison with other KGE methods.
- We test the KG-MRI's practical utility using a private NAFLD cohort from Jidong Hospital, China. Experimental results indicate that the multimodal embedding enhances diagnostic accuracy and mitigates batch effects.

2 Related Work

2.1 Multimodal Knowledge Graph Embedding

Contrary to earlier unimodal KGE methods, multimodal KGE takes advantage of the extensive knowledge derived from multiple modalities [Zhu *et al.*, 2022]. Recent advancements in multimodal KGE have introduced various models, with the objective of combine diverse data modalities for enhanced representation learning. Models like MMKRL [Lu *et al.*, 2022] and CapEnrich [Yao *et al.*, 2023] leverage multi-source knowledge to improve the semantic understanding of entities and relations in knowledge graphs, focusing on enriching traditional KGE methods that primarily rely on triplet facts, he2019integrating presents MK-BERT, a model that integrates BERT with multimodal data to enrich entity and relation embeddings in knowledge graphs. Likewise, the TransAE model [Wang *et al.*, 2019] combines a multimodal autoencoder with TransE model, addressing challenges in unifying multimodal data for better representation learning. These models underscore a shared goal to harness multimodal information to provide more robust representations of real-world entities and relations. They have demonstrated promising results in link prediction tasks and knowledge graph completion benchmarks.

In our work, we propose a multimodal representation integration algorithm, aiming to better align the different representation of entities through contrastive learning and KGE methods, thus improving the representation capability of the whole knowledge graph.

2.2 Contrastive Learning in Knowledge Graph

Contrastive learning, as a self-supervised learning method, has proven highly effective in leveraging the inherent properties and structures of knowledge graphs (KGs) to enhance the quality and utility of graph embeddings. Liang2022Relational utilizes the relational symmetrical structures in KGs to construct positive pairs for contrastive learning, significantly boosting the discriminative capacity of the embeddings. Similarly, the Cross-Scale Contrastive Graph Knowledge Synergy approach by Zhang2023Contrastive leverages hierarchical graph views to promote knowledge sharing and enhance the generalization ability of graph embeddings. Additionally, yang2022knowledge addresses the challenges of noise and sparsity in KG-enhanced recommendation systems through a novel contrastive learning framework KGCL, which uses a knowledge graph augmentation schema to mitigate noise and employs a cross-view contrastive learning paradigm to leverage unbiased user-item interactions effectively. These methods share a unified objective to harness structural and relational data inherently present in KGs, fostering more accurate and robust embeddings.

However, current contrastive learning-based KG embedding methods are primarily limited to dual modalities and tend to focus predominantly on graph structures, often overlooking the potential to integrate richer multimodal information on entities.

3 Preliminaries

In this section, we first define the concepts of a Knowledge Graph (KG) and Multimodal Representation Integration (MRI) as used in our study.

3.1 Definition of Knowledge Graph

A Knowledge Graph, denoted as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, consists of entities and relationships forming a network.

Specifically, $\mathcal{E}$ represents the set of entities in the graph $\mathcal{G}$, $\mathcal{R}$ denotes the set of relationships connecting pairs of entities, and $\mathcal{T} = \{(h, r, t) \mid h, t \in \mathcal{E}, r \in \mathcal{R}\}$ is the set of triplets. Each triplet (h, r, t) indicates that a relationship r exists between the head entity h and the tail entity t. In this work, we aim to enhance Knowledge Graph Embedding (KGE) methods by using multimodal representations of entities. To this end, we retrieve two different modalities for each entity e using the queries $(e, \text{has\_modality\_1}, ?)$ and $(e, \text{has\_modality\_2}, ?)$.

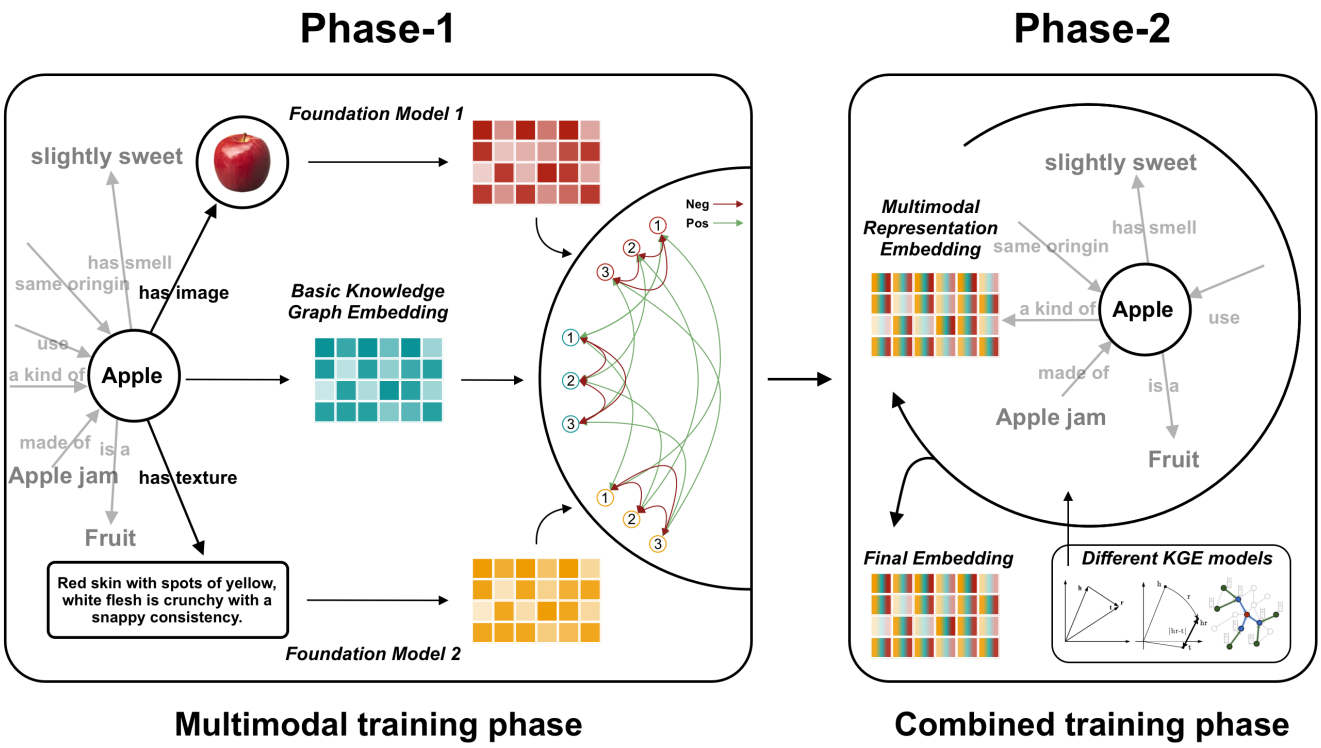

Figure 1: The overall framework of the multimodal representation learning (MRI) algorithm. Two different modalities of an entity are retrieved from the knowledge graph and are represented to vector representations through foundation models respectively. These two distinct representations, along with outputs from the basic KGE method, are integrated using a triple contrastive learning module to enhance alignment. Two separate training phases are employed to optimize integration performance. The outputs of this training process serves as the new KG embeddings for the knowledge graph.

3.2 Definition of MRI

The Multimodal Representation Integration (MRI) algorithm is designed to enhance the representation of entities by integrating their representations from different modalities.

Specifically, for an entity $e \in \mathcal{E}$ with an initial representation $e_{\text{emb}} \in \mathbb{R}^d$, and two other modalities' representations $e_{m_1} \in \mathbb{R}^{d_1}$ and $e_{m_2} \in \mathbb{R}^{d_2}$, the MRI algorithm integrates these representations to produce a refined multimodal embedding $e'_{\text{emb}} \in \mathbb{R}^d$ through a triple contrastive learning module and a dual-phase training strategy. This approach aims to improve the representational capability of the existing knowledge graph embedding methods.

4 Methods

The overall framework of MRI is shown in Figure 1. In general, the MRI algorithm consists of the following modules: multimodal representation acquisition, triple contrastive learning, and dual-phase training. First, the multimodal representation acquisition module utilizes two distinct large language models to produce the representations of two different representations of entities. Then, the triple contrastive learning models aligns different representations of the same entity into a fixed-dimension vector space while maintaining the richness of information from each modality. Finally, the dual-phase training approach guarantees a holistic and in-depth learning experience by concentrating on refining the KG representations.

4.1 Multimodal Representation Acquisition

A multimodal knowledge graph primarily stores different modalities related to a entity e. Yet, a comprehensive understanding of each entity necessitates the integration of different modalities' information. With the advancements in visual foundation models (VFMs) and large language models (LLMs), it's now feasible to obtain different representations through specially tailored LLMs.

We initially extract one modality s from a entity through knowledge graph query $(e, \text{has\_modality\_1}, ?)$. We use a foundation model FM_1 with a specified cut-off length of 128 to get the corresponding representation s_{emb}.

$$s_{emb} = f_{FM_1}(s_{1:128})[0], \quad s_{emb} \in \mathbb{R}^{768} \tag{1}$$

where f_{FM_1} is the encoder of the first foundation model, $s_{1:128}$ is the truncated or padded information, and the first element of the encoder's output is the global vector of the entity's representation of modality s.

Similarly, we retrieve the second modality t of the same entity e via a knowledge graph query $(e, \text{has\_modality\_2}, ?)$. We use a corresponding foundation model FM_2 with a specified

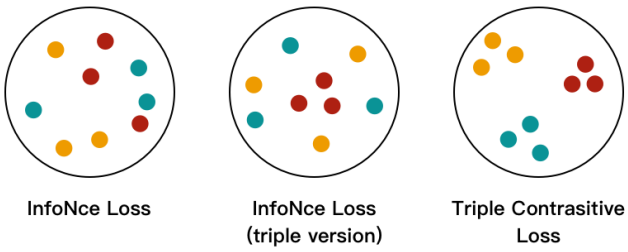

Figure 2: An illustration comparing triple contrastive learning with other contrastive learning techniques. While methods like InfoNCE loss focus on contrasting pairs, our triple contrastive learning iexcels in refining sample representations. It not only aligns identical samples more closely but also distinctly separates dissimilar samples.

cut-off length of 256 to get the corresponding representation t_{emb}.

$$t_{emb} = f_{FM_2}(t_{1:256})[0], \quad t_{emb} \in \mathbb{R}^{768} \tag{2}$$

where f_{FM_2} is the encoder of the second foundation model, $t_{1:256}$ is the truncated or padded, and the first element of the encoder's output is the global vector of the entity's representation of modality t.

4.2 Triple Contrastive Learning

With two 768-dimensional representation vectors s_{emb} and t_{emb}, we consider to integrate them into the entity's representation $e_{emb} \in \mathbb{R}^d$ calculated by the base KGE model. We first project s_{emb} and t_{emb} to the same dimension of e_{emb} by an average-pooling network.

$$e_s[i] = \frac{1}{k}\sum_{j=1}^{k} s_{\text{emb}}[i \times k + j], \quad m_s \in \mathbb{R}^d \tag{3}$$

$$e_t[i] = \frac{1}{k}\sum_{j=1}^{k} t_{\text{emb}}[i \times k + j], \quad m_t \in \mathbb{R}^d \tag{4}$$

where $e_s[i]$ and $e_t[i]$ is the i-th element after pooling and $k = \frac{768}{d}$ is the pooling stride.

With 3 same-dimension representation vectors e_{emb},e_s and e_t, we propose a triple contrastive learning module to learn an integrate vector for multiple entity representations. This module brings together the distances between the representations e_{emb}, e_s, e_t of all modalities (3 in our experiment) of the same entity while pulling away the representations e'_{emb}, e'_s, e'_t of other samples:

$$S_{pos} = \sum_{p=1}^{3} \exp\left(\mathbf{e}_{ip}^{\top}\mathbf{e}_{i(p \bmod 3)+1}/\tau\right) \tag{5}$$

$$S_{neg_1} = \sum_{p=1}^{3} \exp\left(\mathbf{e}_{ip}^{\top}\mathbf{e}_{i(p \bmod 3)+1}/\tau\right) \tag{6}$$

$$S_{neg_2} = \sum_{j=1}^{K}\sum_{q=1}^{3} \exp\left(\mathbf{e}_{iq}^{\top}\mathbf{e}_{j(q \bmod 3)+1}/\tau\right) \tag{7}$$

$$L_{\text{TCL}} = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{S_{pos}}{S_{neg_1} + S_{neg_2}} \tag{8}$$

where n is the number of samples, τ is a temperature parameter, e_1, e_2, e_3 represents e_{emb}, e_s and e_t respectively. K is the number of negative samples. We then calculate the average vector of e_{emb}, e_s and e_t

$$e'_{emb} = \frac{e_{emb} + e_s + e_t}{3}, \tag{9}$$

as the integrated multimodal representation for entity e.

Different from the traditional contrastive learning loss function calculated between two elements, such as InfoNCE loss in SimCLR [Chen *et al.*, 2020a], triple contrastive learning [Lu *et al.*, 2023b] seeks to concentrate the representations from three modalities of the same entity while ensuring their representations remain distinct from the other entities. An intuitive illustration of this concept is available in Figure 2. Both InfoNCE loss and its triple version cannot well distinguish the three-modality scenario.

4.3 Dual-phase Training

To better integrate the representations of multiple modalities, we designed a two-stage training process, namely the multimodal training phase and the combined training phase. Here, we use the RotatE [Sun *et al.*, 2019] KGE method as an example for explanation.

We first initialize all the entity and relation embeddings $E \in \mathbb{R}^{num\_entites \times d}$ and $R \in \mathbb{R}^{num\_relations \times d}$. Then, given a triplet (h, r, t), the head entity embedding is rotated by the relation embedding using element-wise multiplication:

$$h_{emb\_rot} = h_{emb} \odot r_{emb}, \quad h_{emb} \in E, \quad r_{emb} \in R, \tag{10}$$

where $\odot$ denotes element-wise multiplication. We then compute the distance score between the rotated head entity embedding and the tail entity embedding as

$$f_{score} = ||h_{emb\_rot} - t_{emb}||_1, \quad t_{emb} \in E, \tag{11}$$

and we applied margin ranking loss L_{RL} as the total loss as the supervisory signal for KGE methods.

$$L_{RL} = \sum_{i=1}^{N} \max\left(0, 1 + f_{score}(h, r, t) - f_{score}(h'_i, r'_i, t'_i)\right) \tag{12}$$

During the multimodal training phase, the gradient is updated using sum of L_{RL} and L_{TCL}. However, during the combined training phase, only L_{RL} is utilized to ensure a more consistent update of the integrated representation. In our experiments, each training phase is set to run 500 epochs. Upon the completion of training, we extract the embeddings of E and R as the representations of all entities and relations.

5 Experiments

5.1 Dataset

We constructed a biomedical knowledge graph named HMKG from the Human Metabolome Database (HMDB, https://hmdb.ca/, [Wishart *et al.*, 2022]), which is a comprehensive electronic online database offering intricate details on small molecule compounds present in the human body. HMDB presents chemical, clinical, and molecular biological

	Translation	Semantic	Neural Network	Hit@1	Hit@3	Hit@5	Hit@10	MR	MRR
TransE [Bordes *et al.*, 2013]	✓			0.059	0.168	0.213	0.276	2561	0.135
TransD [Ji *et al.*, 2015]	✓			0.147	0.389	0.44	0.512	**462**	0.294
TransH [Wang *et al.*, 2014]	✓			0.458	0.542	0.579	0.607	2730	0.512
TransR [Lin *et al.*, 2015]	✓			0.181	0.262	0.303	0.369	1740	0.244
DisMult [Yang *et al.*, 2014]		✓		0.479	0.577	0.621	0.675	783	0.551
ER-MLP [Dong *et al.*, 2014]			✓	0.096	0.220	0.299	0.427	644	0.199
SimplE [Kazemi and Poole, 2018]		✓		0.012	0.055	0.089	0.140	6223	0.054
NodePiece [Galkin *et al.*, 2021]			✓	0.185	0.194	0.201	0.218	17622	0.198
PairRE [Chao *et al.*, 2020]	✓			0.227	0.311	0.35	0.405	1703	0.289
QuatE [Zhang *et al.*, 2019]	✓			0.075	0.118	0.14	0.173	8394	0.111
RotatE [Sun *et al.*, 2019]	✓			0.538	0.664	0.699	0.742	656	0.614
MRI-RotatE(Ours)	✓		✓	**0.572**	**0.698**	**0.731**	**0.770**	550	**0.631**

Table 1: We first compared some popular knowledge graph embedding methods, including translation models, semantic match models and neural network models. Then we selected the best performing knowledge graph embedding methods and applied it as the base model for our MRI algorithm. Results are presented in terms of Hit@n, median rank (MR), and MRR (Mean Reciprocal Rank). The best results are bolded, and the second-best results are underlined.

specifics of over 200,000 compounds. Each compound is cataloged under a distinct MetaboCard, encompassing 130 data fields; roughly two-thirds pertain to chemical and clinical details, while the remaining one-third focuses on enzymatic or biological data. Many data fields are hyperlinked to other databases, which provide a high extensibility in the future. All the information in HMDB are storaged in the XML format. Based on this, we designed a specially-tailed parser to extract all the biological information in HMDB in the form of triplets.

5.2 Data Preprocessing

Regarding the data preprocessing process of HMKG, we incorporated data de-duplication, data alignment, and data disambiguation methods to enhance data quality. In the process of entity standardization, we eliminated duplicates caused by variations in letter case and special symbols. Entities representing the same concept are linked using the "the_same_as" relationship. Additionally, we preprocessed all numerical information in the knowledge graph, removing missing values, outliers, and erroneous data to ensure data accuracy and reliability. Furthermore, we perform standardization of the normal concentration values of compounds in the human body. By comparing these values with concentration values in patients with specific diseases, we establish triplets indicating upregulation or downregulation relationships between compounds and diseases. This enables us to identify compound expression changes related to diseases accurately.

5.3 Multimodal Information Representation

In our experiment, we first extract all the SMILES sequences of each compound. SMILES is a notation employed to depict chemical structures in textual form, with an average sequence length of 64. We utilize a recently developed language model named ChemBERTa-2 as the encoder for SMILES. ChemBERTa-2 is pretrained on a vast dataset comprising the SMILES sequences of 77 million compounds sourced from PubChem.

Then we extract all the text descriptions stored in HMDB of each compound and opt for SciBERT to encode these text descriptions. SciBERT is pretrained on a large multi-domain scientific publications corpus of 1.14M papers from Semantic Scholar and is widely used in biomedical natural language processing tasks.

5.4 Experiment Settings

In our experiment, we meticulously partitioned the dataset into training, validation, and testing sets following an 8:1:1 ratio. This division was aimed at ensuring a robust evaluation framework. The hyperparameters were chosen to strike a balance between the model's efficiency and reliability. Both the entity and relation embeddings were initialized with a dimensionality of 128. The learning rate was established at 1.0×10^{-3}. The training was conducted over 1000 epochs on a single Tesla A100 GPU. The total computation time varied between 3 and 60 hours, contingent on the base KGE model employed.

To uphold the reproducibility of our experimental outcomes, we utilized a fixed random seed of 42. During training, we applied AdamW [Loshchilov and Hutter, 2018] and CosineAnnealingLR [Loshchilov and Hutter, 2016] as our default optimizer and scheduler. This meticulous setup was designed to balance computational efficiency with model fidelity, thereby ensuring robust and reproducible results.

For reporting the experimental results of HMKG, common metrics including Hits@1, Hits@3, Hits@5, Hits@10, MR (Mean Rank), and MRR (Mean Reciprocal Rank) were utilized. The higher values of Hits@n and MRR and the lower values of MR explain the better performance of the method. These metrics are extensively used in evaluating the performance of models in knowledge graph embedding and information retrieval tasks.

6 Results

6.1 Comparison with Other KGE Methods

In HMKG, different patterns exist among various relationships. Some relationships are symmetric, like "the_same_as" relationship, while others are asymmetric, like the "is_a" relationship. Additionally, some relationships are inversions of each other. For instance, if Entity A has a "has_sub_class" relationship with Entity B, it typically implies that Entity B has a "has_father_class" relationship with Entity A. Moreover, some relationships exhibit properties of composition and transitivity. For example, if Compound A "has_disease" Disease B, and Disease B is "related_to" Gene C, then Compound A and Gene C may also have a "related_to" relationship.

Since different knowledge graph embedding (KGE) methods may have varying effects in knowledge graphs composed of different patterns, we selected a group of KGE methods to compare their vector representation learning capabilities within HMKG. We broadly categorized our selected KGE methods into three groups. Firstly, we considered translation model-based KGE models such as TransE [Bordes *et al.*, 2013], TransD [Ji *et al.*, 2015], TransH [Lin *et al.*, 2015], and RotatE [Sun *et al.*, 2019], etc. Secondly, we explored semantic matching models like Dismult [Yang *et al.*, 2014] and SimplE [Kazemi and Poole, 2018]. Lastly, we delved into neural network-based models including ER-MLP [Dong *et al.*, 2014] and NodePiece [Galkin *et al.*, 2021]. The comparison results are shown in Table 1.

In Table 1, we bolded the best results, and underlined the second-best results. Among all KGE methods, RotatE [Sun *et al.*, 2019] achieved the best performance in Hits@n and MRR, and significantly outperformed other methods, such as (Hit@1 was 5% higher than the second place DisMult model [Yang *et al.*, 2014], and MRR was 0.08 higher than the second place DisMult model as well). However, TransD [Ji *et al.*, 2015] and ER-MLP [Dong *et al.*, 2014]models performed somewhat better on the MR evaluation metric. We have also tested the results of other KGE methods like ConvE [Dettmers *et al.*, 2018], KG2E [He *et al.*, 2015], ConvKB [Nguyen *et al.*, 2017] and RGCN [Schlichtkrull *et al.*, 2018], etc., but their performance is not as good as the methods listed in Table 1, so we choose not to display these results.

Based on the above results, we chose RotatE [Sun *et al.*, 2019] as the base KGE model in our MRI algorithm, and we integrate the chemical embedding of compounds' SMILES sequences and semantic embedding of compounds' text descriptions to enhance the representation capability of HMKG. We report the performance of our MRI-RotatE KGE methods in the last row of Table 1. Not surprisingly, our method notably outperformed on various metrics, encompassing all Hit@n and MRR values. Furthermore, it achieved an MR of 550, marking a substantial advancement over RotatE [Sun *et al.*, 2019] and surpassing ER-MLP [Dong *et al.*, 2014] (MR=644), while nearing the performance of TransD [Ji *et al.*, 2015] (MR=462). The comparison results strongly demonstrate the effectiveness of the MRI algorithm.

	Overall	**NAFLD**	**NC**
Sex, n (%)			
Male	194 (62.6%)	109 (35.2%)	85 (27.4%)
Female	116 (37.4%)	51 (16.5%)	65 (21.0%)
Average age, years	40.3 ± 9.0	40.8 ± 9.0	39.7 ± 8.9
Age group, n (%)			
<30	16 (5.2%)	7 (2.3%)	9 (2.9%)
30-39	164 (52.9%)	82 (26.5%)	82 (26.5%)
40-49	69 (22.3%)	39 (12.6%)	30 (9.7%)
50-59	55 (17.7%)	29 (9.4%)	26 (8.4%)
≥60	6 (1.9%)	3 (1.0%)	3 (1.0%)

Table 2: Demographic statistics of the NAFLD cohort (n=310), where NAFLD stands for non-alcoholic fatty liver disease (n=160), NC stands for normal control (n=150).

6.2 An Empirical Study in NAFLD Diagnosis

In demonstrating the practical application of KG-MRI for disease diagnosis and health assessment, we carried out a study on a non-alcoholic fatty liver disease (NAFLD) cohort (160 NAFLD patients and 150 Normal Control) from Jidong Hospital, Hebei, China. Blood sample of each patient was collected after fasting and followed by medical follow-up. Detailed statistics of this cohort can be found in Table 2.

In clinical scenario, using compound information for health assessments is a crucial concern. Traditional machine learning methods have been successful in diagnosing certain single diseases. However, the challenge of integrating data from different batches, due to variations in devices and conditions, hinders the reuse of pre-trained models. In order to overcome this problem, it's essential to decouple the fixed number of compounds in patient clinical test data from predictive models. Our compound embedding through KG-MRI model allows the handling of compound data regardless of the number of compounds. Moreover, using information about individual compounds from knowledge graphs for disease prediction can provide a more detailed and thorough description of compounds. This enhances the model's generality and interpretability.

As illustrated in Figure 3, our analytical pipeline starts by sampling all normal population samples to establish reference ranges for each compound. Next, for each sample under study, the model classifies the compounds into three categories, that is, upper-regulation, normal, and lower-regulation, based on these reference ranges. Within each category, we select 50 most relevant compounds and look up into HMKG to obtain their vector representations. These vectors are then multiplied by their respective expression levels to create a comprehensive patient-level matrix. If the number of compounds in a given category is fewer than 50, zero-padding vectors are used to fill up the shortfall.

We then used the MultiLayer Preceptron (MLP) neural network to classify this NAFLD cohort using the aforementioned patient-level matrix under the cross-validation settings, achieving an average AUC of 0.87, an average F1 of 0.84, and an average accuracy of 0.83, which surpasses the clas-

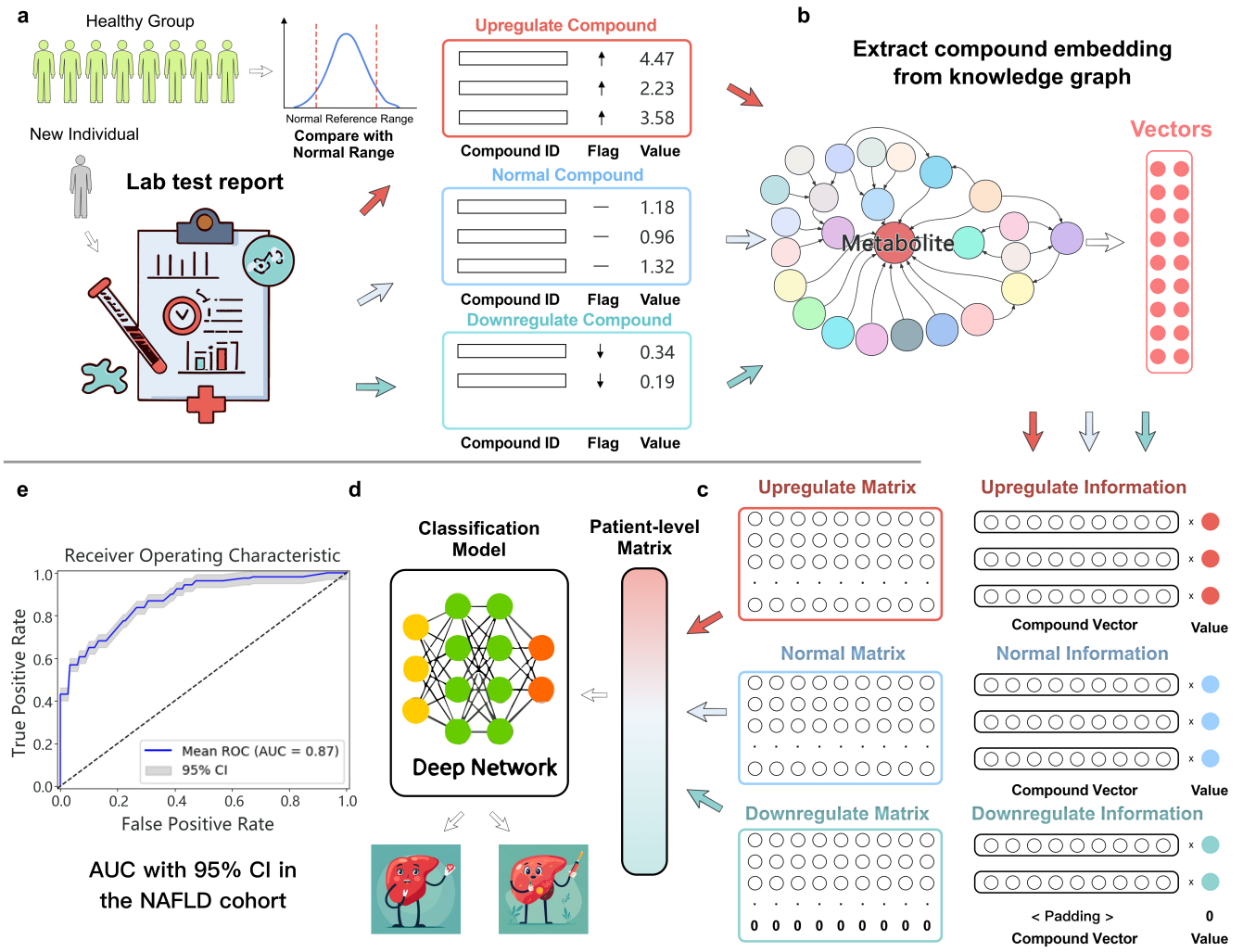

Figure 3: NAFLD diagnosis pipeline using HMKG. The normal range compound expression are calculated as thresholds for distinguishing upregulating, normal, and downregulating compounds. Representation retrieved from HMKG are fed into a NAFLD classification model.

	Acc.	F1	AUC		Acc.	F1	AUC
LR	0.72	0.71	0.65	NB	0.70	0.72	0.68
SVM	0.80	0.83	0.83	KNN	0.76	0.74	0.77
RF	0.72	0.58	0.61	KG-MRI	**0.83**	**0.84**	**0.87**

Table 3: Performance metrics for different classification models. Each model has gone through a five-fold cross-validation. The highest metric value is highlighted in bold.

sification performance attained when directly applying traditional machine learning models to classify patients' different compounds expression data (a fixed-length vector that cannot generalize to a varying lengths due to factors like batch effects of samples and disparities among different clinical test machines). The full comparison results with other machine learning models [Su *et al.*, 2012; Rish and others, 2001; Hearst *et al.*, 1998; Breiman, 2001] are shown in Table 3.

7 Conclusion

In this study, we proposed a knowledge graph multimodal representation (KG-MRI) learning model using triple contrastive learning and a dual-phase training strategy. This model uniquely integrates various modalities associated with a single entity in a knowledge graph, enhancing the multimodal coherence and utility of the derived embeddings. Our exhaustive comparative analysis reveals that KG-MRI outperforms traditional KGE models across multiple metrics. Further in-depth evaluations were conducted using a biomedical knowledge graph specifically tailored to analyze Non-Alcoholic Fatty Liver Disease (NAFLD). By applying our KG-MRI model to this domain, we were able to significantly reduce batch effects, which are common issues in traditional classification models used in biomedical contexts. This reduction in batch effects, coupled with our model's enhanced performance metrics, underscores the practical and theoretical benefits of integrating multimodal data in knowledge graph representations.

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Ahmad *et al.*, 2022] Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*, 2022.

[Bordes *et al.*, 2013] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.

[Breiman, 2001] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[Cao *et al.*, 2022] Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. Otkge: Multi-modal knowledge graph embeddings via optimal transport. *Advances in Neural Information Processing Systems*, 35:39090–39102, 2022.

[Chao *et al.*, 2020] Linlin Chao, Jianshan He, Taifeng Wang, and Wei Chu. Pairre: Knowledge graph embeddings via paired relation vectors. *arXiv preprint arXiv:2011.03798*, 2020.

[Chen *et al.*, 2020a] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[Chen *et al.*, 2020b] Xiaojun Chen, Shengbin Jia, and Yang Xiang. A review: Knowledge reasoning over knowledge graph. *Expert systems with applications*, 141:112948, 2020.

[Cui *et al.*, 2024] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, pages 1–11, 2024.

[Dettmers *et al.*, 2018] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Dong *et al.*, 2014] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 601–610, 2014.

[Galkin *et al.*, 2021] Mikhail Galkin, Etienne Denis, Jiapeng Wu, and William L Hamilton. Nodepiece: Compositional and parameter-efficient representations of large knowledge graphs. *arXiv preprint arXiv:2106.12144*, 2021.

[Girdhar *et al.*, 2023] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Armand Alwala, Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.

[Guo *et al.*, 2020] Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, 34(8):3549–3568, 2020.

[He *et al.*, 2015] Shizhu He, Kang Liu, Guoliang Ji, and Jun Zhao. Learning to represent knowledge graphs with gaussian embedding. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 623–632, 2015.

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[Hearst *et al.*, 1998] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[Ji *et al.*, 2015] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, pages 687–696, 2015.

[Ji *et al.*, 2021] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2):494–514, 2021.

[Kazemi and Poole, 2018] Seyed Mehran Kazemi and David Poole. Simple embedding for link prediction in knowledge graphs. *Advances in neural information processing systems*, 31, 2018.

[Li *et al.*, 2023] Qian Li, Shu Guo, Yangyifei Luo, Cheng Ji, Lihong Wang, Jiawei Sheng, and Jianxin Li. Attribute-consistent knowledge graph representation learning for multi-modal entity alignment. In *Proceedings of the ACM Web Conference 2023*, pages 2499–2508, 2023.

[Liang *et al.*, 2022] Ke Liang, Yue Liu, Sihang Zhou, Xinwang Liu, and Wenxuan Tu. Relational symmetry based knowledge graph contrastive learning. *ArXiv*, abs/2211.10738, 2022.

[Lin *et al.*, 2015] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[Loshchilov and Hutter, 2016] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2016.

[Loshchilov and Hutter, 2018] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[Lu *et al.*, 2022] Xinyu Lu, Lifang Wang, Zejun Jiang, Shichang He, and Shizhong Liu. Mmkrl: A robust embedding approach for multi-modal knowledge graph representation learning. *Applied Intelligence*, pages 1–18, 2022.

[Lu *et al.*, 2023a] Yuxing Lu, Xiaohong Liu, Zongxin Du, Yuanxu Gao, and Guangyu Wang. Medkpl: a heterogeneous knowledge enhanced prompt learning framework for transferable diagnosis. *Journal of Biomedical Informatics*, 143:104417, 2023.

[Lu *et al.*, 2023b] Yuxing Lu, Rui Peng, Lingkai Dong, Kun Xia, Renjie Wu, Shuai Xu, and Jinzhuo Wang. Multiomics dynamic learning enables personalized diagnosis and prognosis for pancancer and cancer subtypes. *Briefings in Bioinformatics*, 24(6):bbad378, 2023.

[Mohamed *et al.*, 2021] Sameh K Mohamed, Aayah Nounu, and Vít Nováček. Biological applications of knowledge graph embedding models. *Briefings in bioinformatics*, 22(2):1679–1693, 2021.

[Nguyen *et al.*, 2017] Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. A novel embedding model for knowledge base completion based on convolutional neural network. *arXiv preprint arXiv:1712.02121*, 2017.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Rish and others, 2001] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[Schlichtkrull *et al.*, 2018] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, et al. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 593–607. Springer, 2018.

[Su *et al.*, 2012] Xiaogang Su, Xin Yan, and Chih-Ling Tsai. Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(3):275–294, 2012.

[Sun *et al.*, 2019] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*, 2019.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Wang *et al.*, 2014] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28, 2014.

[Wang *et al.*, 2017] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE transactions on knowledge and data engineering*, 29(12):2724–2743, 2017.

[Wang *et al.*, 2019] Zikang Wang, Linjing Li, Qiudan Li, and Daniel Zeng. Multimodal data enhanced representation learning for knowledge graphs. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[Wishart *et al.*, 2022] David S Wishart, AnChi Guo, Eponine Oler, Fei Wang, Afia Anjum, Harrison Peters, Raynard Dizon, Zinat Sayeeda, Siyang Tian, Brian L Lee, et al. Hmdb 5.0: the human metabolome database for 2022. *Nucleic acids research*, 50(D1):D622–D631, 2022.

[Yang *et al.*, 2014] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.

[Yang *et al.*, 2022] Yuhao Yang, Chao Huang, Lianghao Xia, and Chenliang Li. Knowledge graph contrastive learning for recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 1434–1443, 2022.

[Yao *et al.*, 2023] Linli Yao, Weijing Chen, and Qin Jin. Capenrich: Enriching caption semantics for web images via cross-modal pre-trained knowledge. In *TheWebConf*, 2023.

[Zhang *et al.*, 2019] Shuai Zhang, Yi Tay, Lina Yao, and Qi Liu. Quaternion knowledge graph embeddings. *Advances in neural information processing systems*, 32, 2019.

[Zhang *et al.*, 2023] Yifei Zhang, Yankai Chen, Zixing Song, and Irwin King. Contrastive cross-scale graph knowledge synergy. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023.

[Zhu *et al.*, 2022] Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. Multi-modal knowledge graph construction and application: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2022.