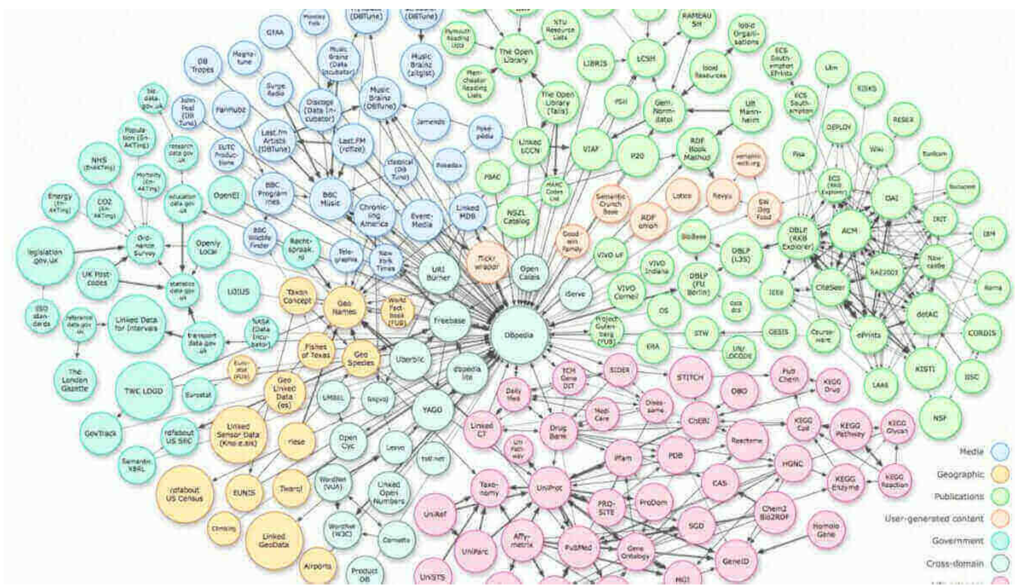
Enhancing Multimodal Knowledge Graph Representation Learning through Triple Contrastive Learning

Yuxing Lu, Weichen Zhao, Nan Sun and Jinzhuo Wang

Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)

Background

- Knowledge Graphs (KGs) model entities and relationships across various domains (e.g., healthcare, recommendation systems)
- Traditional KGs rely on symbolic data, which doesn't fully align with human perception

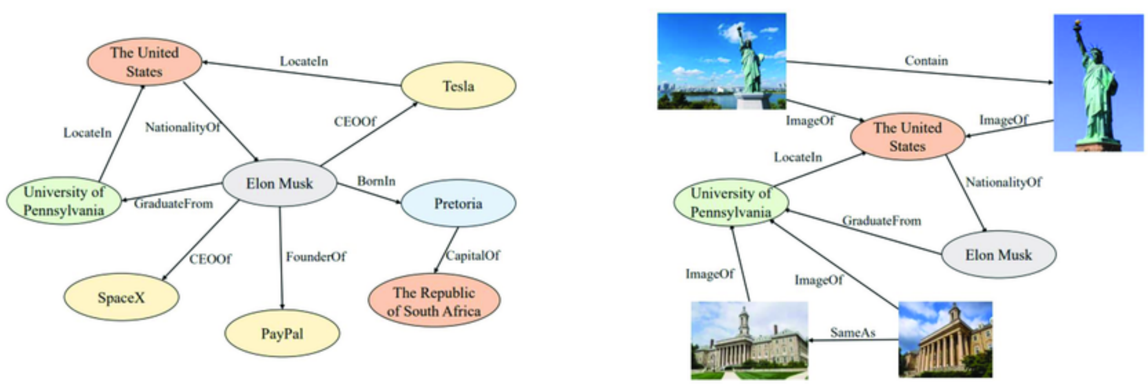


Challenges in Multimodal Knowledge Graph Embedding

- Fusing different modalities (like text and images) into a unified KG representation
- Most existing methods only handle one or two modalities and struggle to combine information effectively

Motivation for Multimodal Knowledge Graphs

- Multimodal integration is key for holistic understanding (e.g., observing an apple).
- Diverse modalities can better capture human-like cognition in representation learning.
- In healthcare, multimodal integration improves outcomes (e.g., diagnosis, recommendations).
- Advances in foundation models help represent different modalities effectively.



(a) Knowledge Graph (b) Multimodal Knowledge Graph

KG-MRI

- Multimodal representation integration model for knowledge graph representation learning
- Incorporate different modalities' representations from foundation models with knowledge graph embeddings
- Used triple contrastive learning (TCL) and apply a dual-phase training strategy to optimize alignment among varied representations

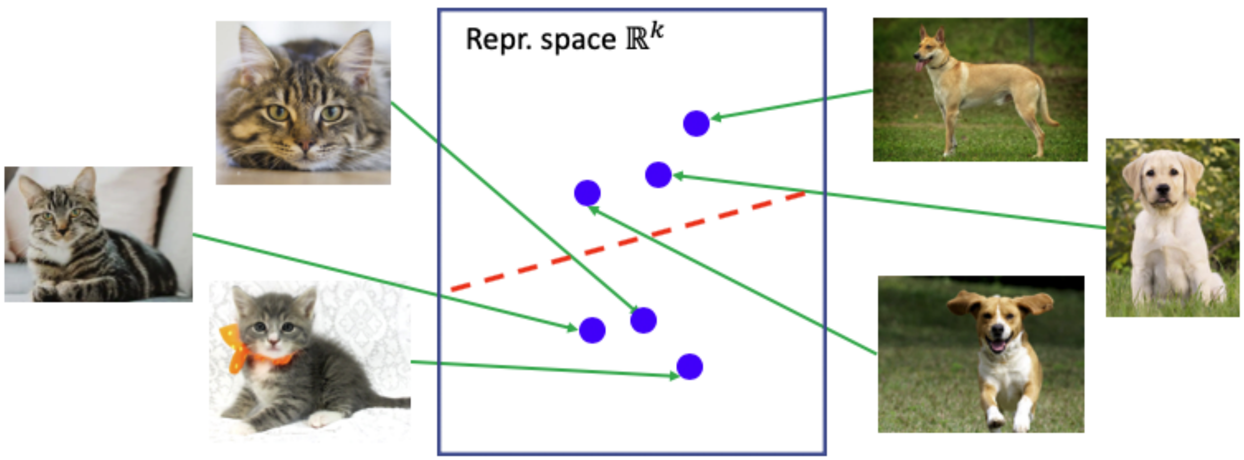
Knowledge Graph Embedding[1]

- A technique for embedding these graphs into low-dimensional vector spaces (numerical representation)
- Facilitate efficient storage and enable downstream tasks (E.g., knowledge graph completion and inference)

[1] Tan, J., Wang, D., Sun, J., Liu, Z., Li, X., & Feng, Y. (Year). Towards assessing the quality of knowledge graphs via differential testing.

Contrastive Learning

- Contrastive learning is a self-supervised learning technique used to teach a model to distinguish between similar and dissimilar examples
- It works by comparing pairs of inputs and trying to bring similar things closer together in vector space while pushing dissimilar things further apart



Ref: https://www.linkedin.com/pulse/what-contrastive-learning-aionlinecourse-5diwc/

Methods

- Consists of following modules: multimodal representation acquisition, triple contrastive learning, and dual-phase training

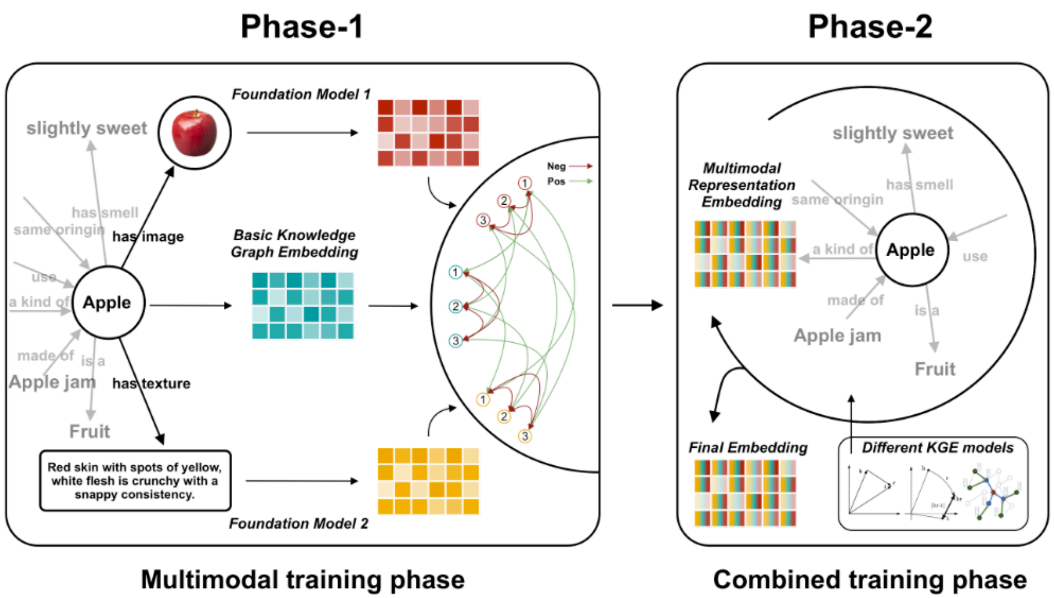


Figure 1: The overall framework of the multimodal representation learning (MRI) algorithm. Two different modalities of an entity are retrieved from the knowledge graph and are represented to vector representations through foundation models respectively. These two distinct representations, along with 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.

Multimodal Representation Acquisition

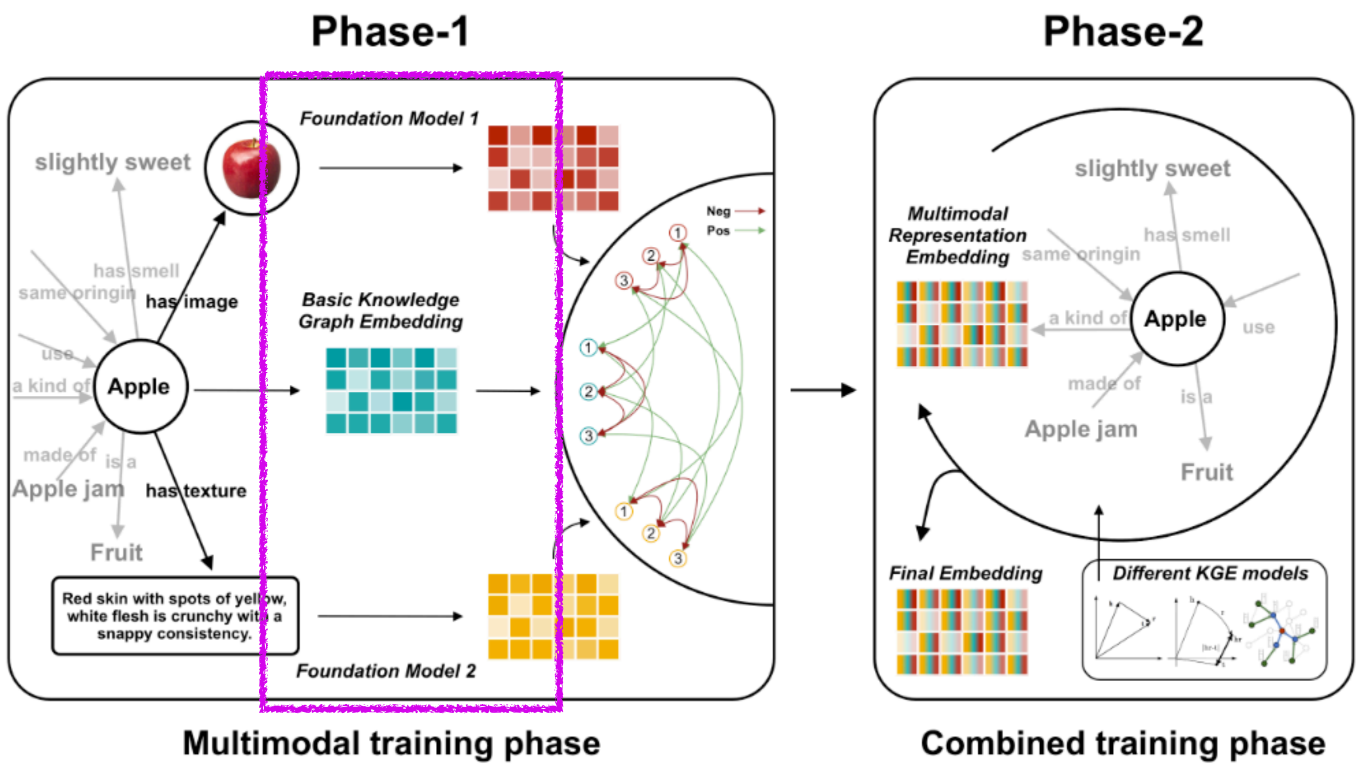


Figure 1: The overall framework of the multimodal representation learning (MRI) algorithm. Two different modalities of an entity are retrieved from the knowledge graph and are represented to vector representations through foundation models respectively. These two distinct representations, along with 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.

Multimodal Representation Acquisition

Two foundation models

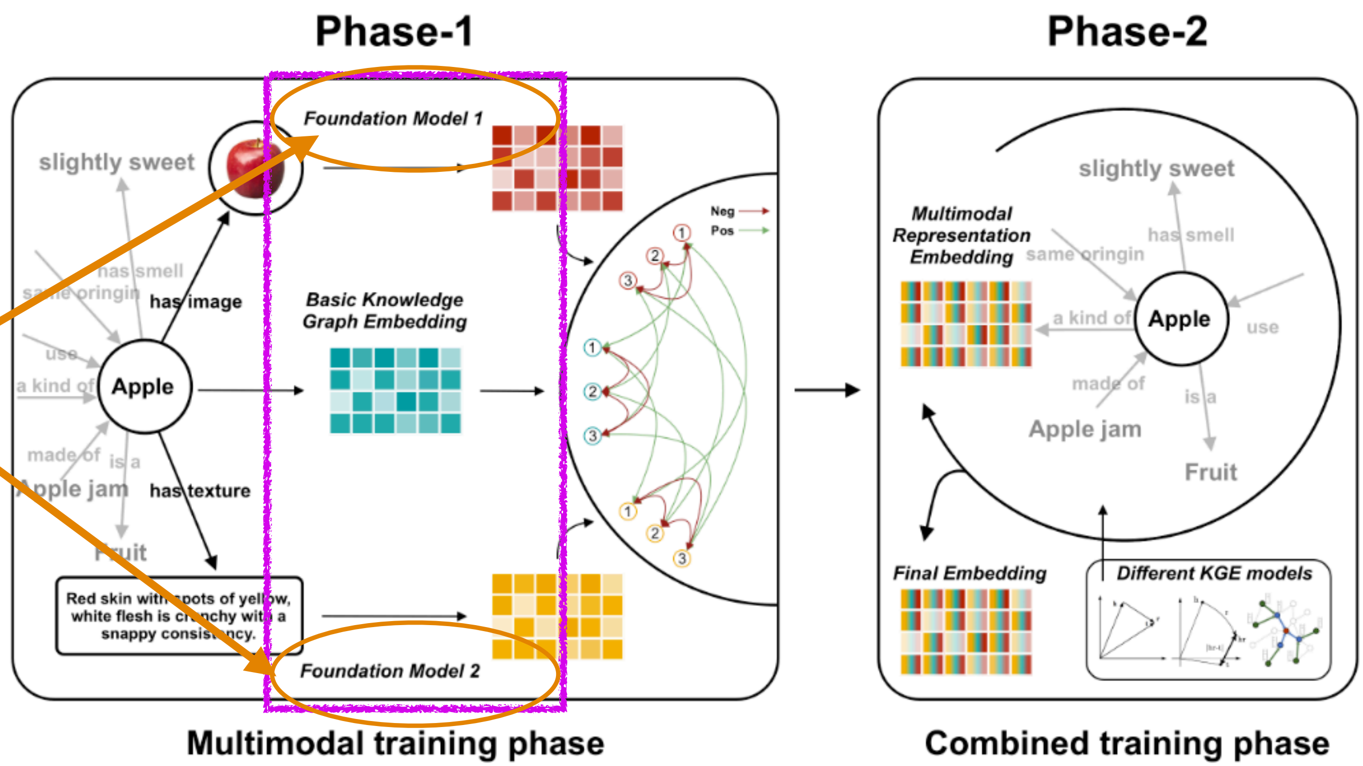


Figure 1: The overall framework of the multimodal representation learning (MRI) algorithm. Two different modalities of an entity are retrieved from the knowledge graph and are represented to vector representations through foundation models respectively. These two distinct representations, along with 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.

Multimodal Representation Acquisition

Knowledge Graph Embedding Model

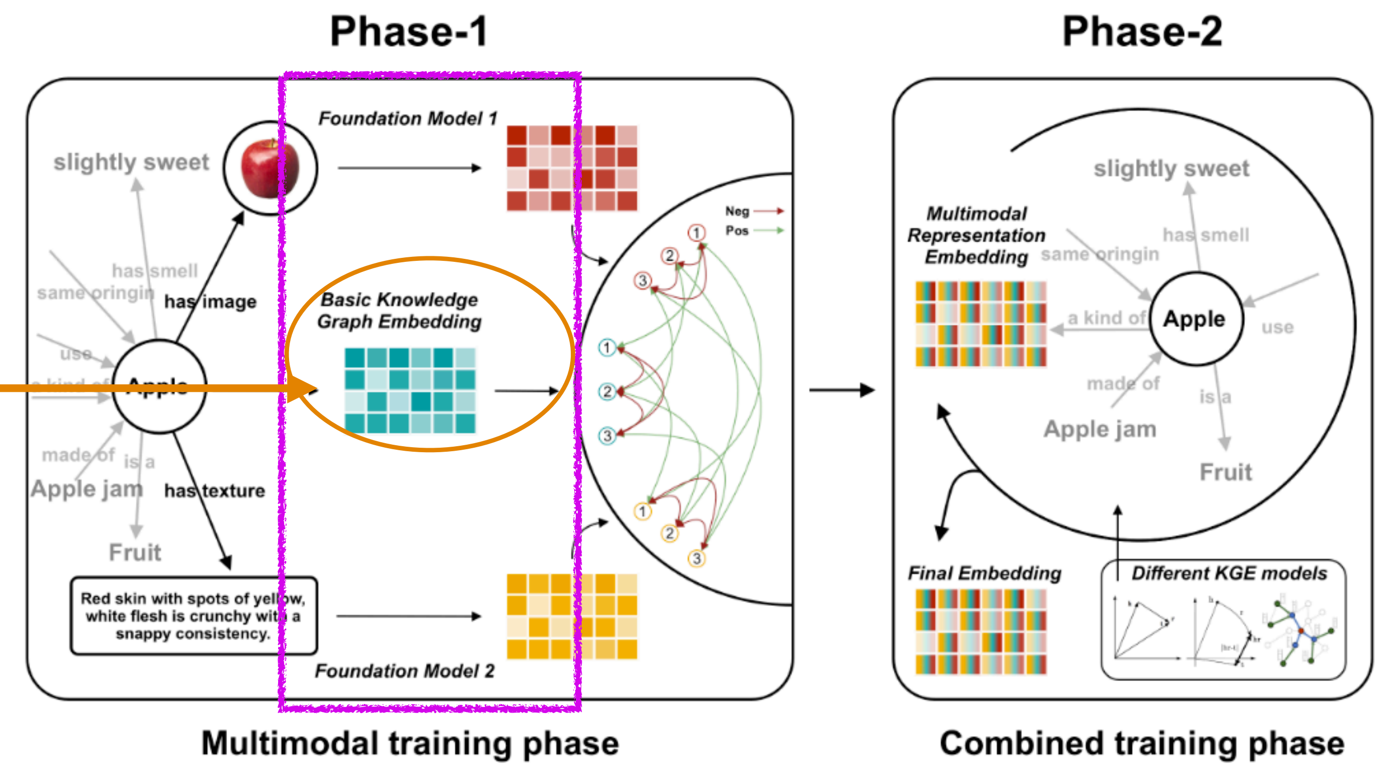


Figure 1: The overall framework of the multimodal representation learning (MRI) algorithm. Two different modalities of an entity are retrieved from the knowledge graph and are represented to vector representations through foundation models respectively. These two distinct representations, along with 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.

Triple Contrastive Learning

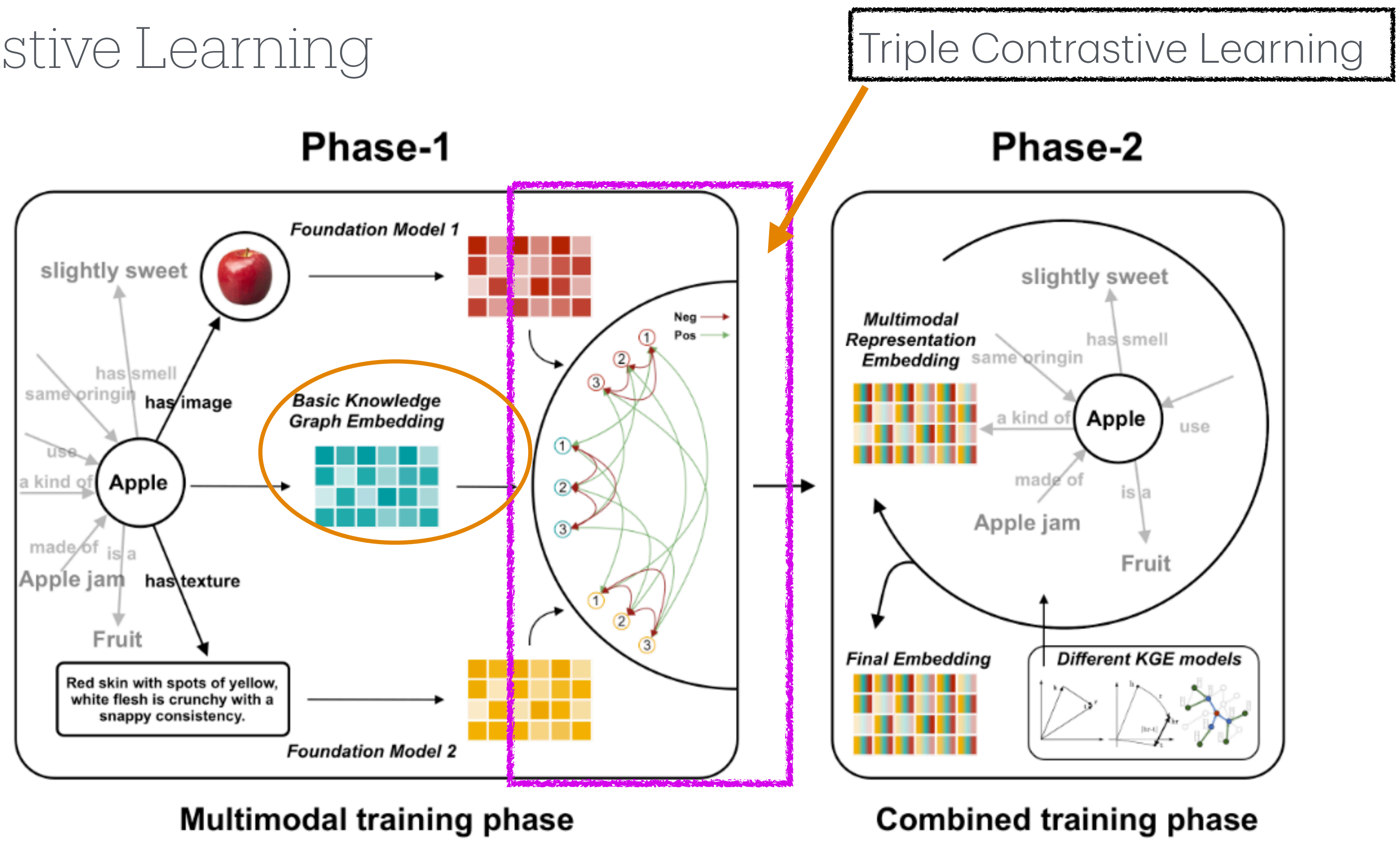


Figure 1: The overall framework of the multimodal representation learning (MRI) algorithm. Two different modalities of an entity are retrieved from the knowledge graph and are represented to vector representations through foundation models respectively. These two distinct representations, along with 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.

Triple Contrastive Learning

- Main Idea: aligns the three different embeddings into a unified vector space

$$e_s[i] = \frac{1}{k}\sum_{j=1}^{k} s_{\text{emb}}[i \times k + j], \quad m_s \in \mathbb{R}^d \longrightarrow$$ i-th element after average-pooling (first representation)

$$e_t[i] = \frac{1}{k}\sum_{j=1}^{k} t_{\text{emb}}[i \times k + j], \quad m_t \in \mathbb{R}^d \longrightarrow$$ i-th element after average-pooling (second representation)

- To get the same dimension by average pooling network

Triple Contrastive Learning

- Main Idea: brings together the distances between representations e_s, e_t, e_{emb}, while pulling away other samples

$$S_{\text{pos}} = \sum_{p=1}^{3} \exp\left(e_i^T p e_{i(p \mod 3)+1} / \tau\right) \longrightarrow$$ Score for positive pairs

$$S_{\text{neg1}} = \sum_{p=1}^{3} \exp\left(e_i^T p e_{i(p \mod 3)+1} / \tau\right) \longrightarrow$$ Score for negative pairs

$$S_{\text{neg2}} = \sum_{q=1}^{K} \sum_{p=1}^{3} \exp\left(e_i^T e_j (q \mod 3) + 1 / \tau\right) \longrightarrow$$ Score for negative pairs

$$L_{\text{TCL}} = -\frac{1}{n} \sum_{i=1}^{n} \log \frac{S_{\text{pos}}}{S_{\text{neg1}} + S_{\text{neg2}}}$$

$$e'_{\text{emb}} = \frac{e_{\text{emb}} + e_s + e_t}{3}$$

Triple Contrastive Learning

- Main Idea: brings together the distances between representations e_s, e_t, e_{emb}, while pulling away other samples

$$S_{\text{pos}} = \sum_{p=1}^{3} \exp\left(e_i^T p e_{i(p \mod 3)+1} / \tau\right) \longrightarrow \text{Score for positive pairs}$$

$$S_{\text{neg1}} = \sum_{p=1}^{3} \exp\left(e_i^T p e_{i(p \mod 3)+1} / \tau\right) \longrightarrow \text{Score for negative pairs}$$

$$S_{\text{neg2}} = \sum_{q=1}^{K} \sum_{p=1}^{3} \exp\left(e_i^T e_j(q \mod 3) + 1 / \tau\right) \longrightarrow \text{Score for negative pairs}$$

$$L_{\text{TCL}} = -\frac{1}{n} \sum_{i=1}^{n} \log \frac{S_{\text{pos}}}{S_{\text{neg1}} + S_{\text{neg2}}}$$

$$e'_{\text{emb}} = \frac{e_{\text{emb}} + e_s + e_t}{3}$$

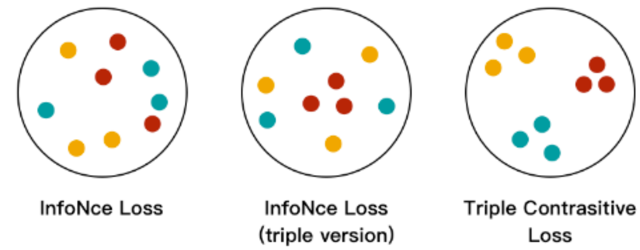
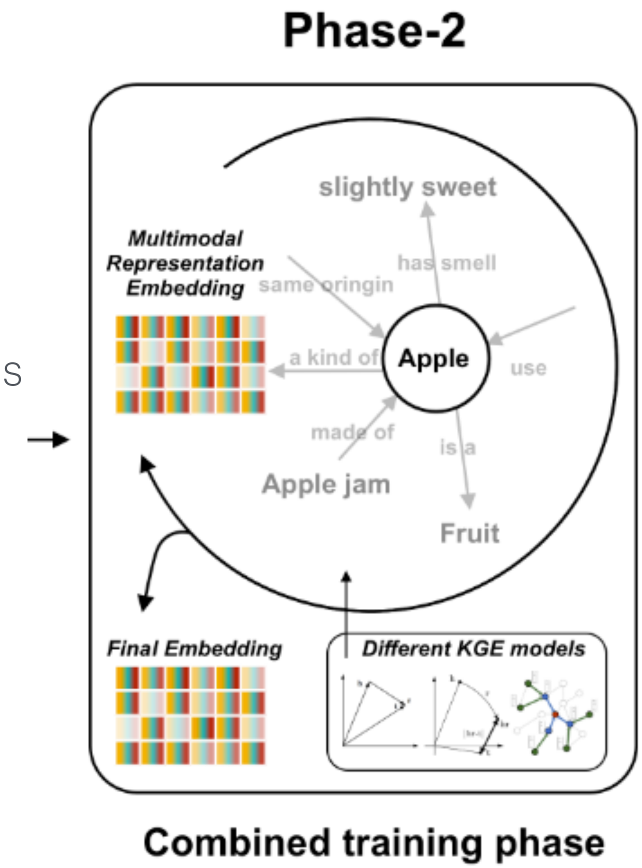


Figure 2: An illustration comparing triple contrastive learning with other contrastive learning techniques. While methods like InfoNCE loss focus on contrasting pairs, our triple contrastive learning iexcels in refining sample representations. It not only aligns identical samples more closely but also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

Dual-phase Training

- Phase 1: Multimodal Training Phase
  - The model learns embeddings for each modality
  - Triple contrastive learning is used to align these embeddings into a unified vector space
- Phase 2: Combined Training Phase
  - In the second phase, the aligned embeddings from Phase 1 are fine-tuned as a single, combined embedding
  - KGE method is used to refine the entity representations



Experiments - Dataset

- Constructed a biomedical knowledge graph named HMKG from HMDB (Human Metabolome Database)

- Offers details on small molecule compounds present in human body

Diabetes mellitus type 1 (222100)

Show 10 entries

Search:

Metabolite	Biospecimen	Concentration	Patient Status	Age	Sex	Details
alpha-Carotene (HMDB0003993)	Blood	0.0800 +/- 0.0700 uM	Abnormal	Adult (>18 years old)	Both	details
beta-Carotene (HMDB0000561)	Blood	0.390 +/- 0.260 uM	Abnormal	Adult (>18 years old)	Both	details
beta-Cryptoxanthin (HMDB0033844)	Blood	0.420 +/- 0.330 uM	Abnormal	Adult (>18 years old)	Both	details
L-Lactic acid (HMDB0000190)	Blood	500.0 +/- 130.0 uM	Abnormal	Adult (>18 years old)	Both	details
Lutein (HMDB0003233)	Blood	0.210 +/- 0.0900 uM	Abnormal	Adult (>18 years old)	Both	details
Lycopene (HMDB0003000)	Blood	0.560 +/- 0.290 uM	Abnormal	Adult (>18 years old)	Both	details
Zeaxanthin (HMDB0002789)	Blood	0.0700 +/- 0.0300 uM	Abnormal	Adult (>18 years old)	Both	details
2-(2-Phenylacetoxy)propionylglycine (HMDB0059732)	Urine	Not Quantified	Abnormal	Adult (>18 years old)	Male	details
2-(2-Phenylacetoxy)propionylglycine (HMDB0059732)	Urine	Not Quantified	Abnormal	Adult (>18 years old)	Male	details
3-Hydroxyisovaleric acid (HMDB0000754)	Urine	10.78 (4.94 – 21.07) umol/mmol creatinine	Abnormal	Adult (>18 years old)	Both	details

Showing 1 to 10 of 21 entries

Previous 1 2 3 Next

Diabetes mellitus type 2 (125853)

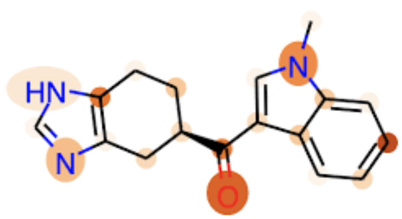
Show 10 entries

Search:

Metabolite	Biospecimen	Concentration	Patient Status	Age	Sex	Details
(R)-3-Hydroxyisobutyric acid (HMDB0000336)	Blood	38.0 +/- 5.0 uM	Abnormal	Adult (>18 years old)	Both	details
(S)-3-Hydroxyisobutyric acid (HMDB0000023)	Blood	38.0 +/- 5.0 uM	Abnormal	Adult (>18 years old)	Both	details
1,5-Anhydrosorbitol (HMDB0002712)	Blood	50.0 (18.5 - 95.0) uM	Abnormal	Adult (>18 years old)	Not Specified	details
1,5-Anhydrosorbitol (HMDB0002712)	Blood	62.0 +/- 38.0 uM	Abnormal	Adult (>18 years old)	Both	details
1,5-Anhydrosorbitol (HMDB0002712)	Blood	30.0 (9.7 - 66.4) uM	Abnormal	Adolescent (13-18 years old)	Both	details
1-Butanol (HMDB0004327)	Blood	0.11 (0 - 2.43) uM	Abnormal	Adult (>18 years old)	Both	details
3-Hydroxybutyric acid (HMDB0000011)	Blood	7700.0 +/- 300.0 uM	Abnormal	Children (1-13 years old)	Both	details

Experiments - Multimodal Information Representation

- Extracted all SMILES sequences of each compound
- Chemical structure (SMILES) encoded with ChemBERTa-2
- Text descriptions encoded with SciBERT



SMILES sequences → ChemBERTa-2 → Embedding for SMILES

Text descriptions → SciBERT → Embeddings for text descriptions

Results-Comparison with other KGE methods

	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]	✓			<u>0.538</u>	<u>0.664</u>	<u>0.699</u>	<u>0.742</u>	656	<u>0.614</u>
MRI-RotatE(Ours)	✓		✓	**0.572**	**0.698**	**0.731**	**0.770**	<u>550</u>	**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.

Application: NAFLD Diagnosis Pipeline

- Studied a non-alcoholic fatty liver disease (NAFLD) cohort
- Integrated compound-level data with knowledge graph embeddings to improve NAFLD prediction accuracy

	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).

NAFLD Analytical Pipeline

- Sampling: to establish reference ranges for each compound
- Classify compounds into three categories
- Retrieve vector embeddings from HMKG for each selected compound based on regulation category
- Patient-Level Matrix Creation
- Classification with MLP Model

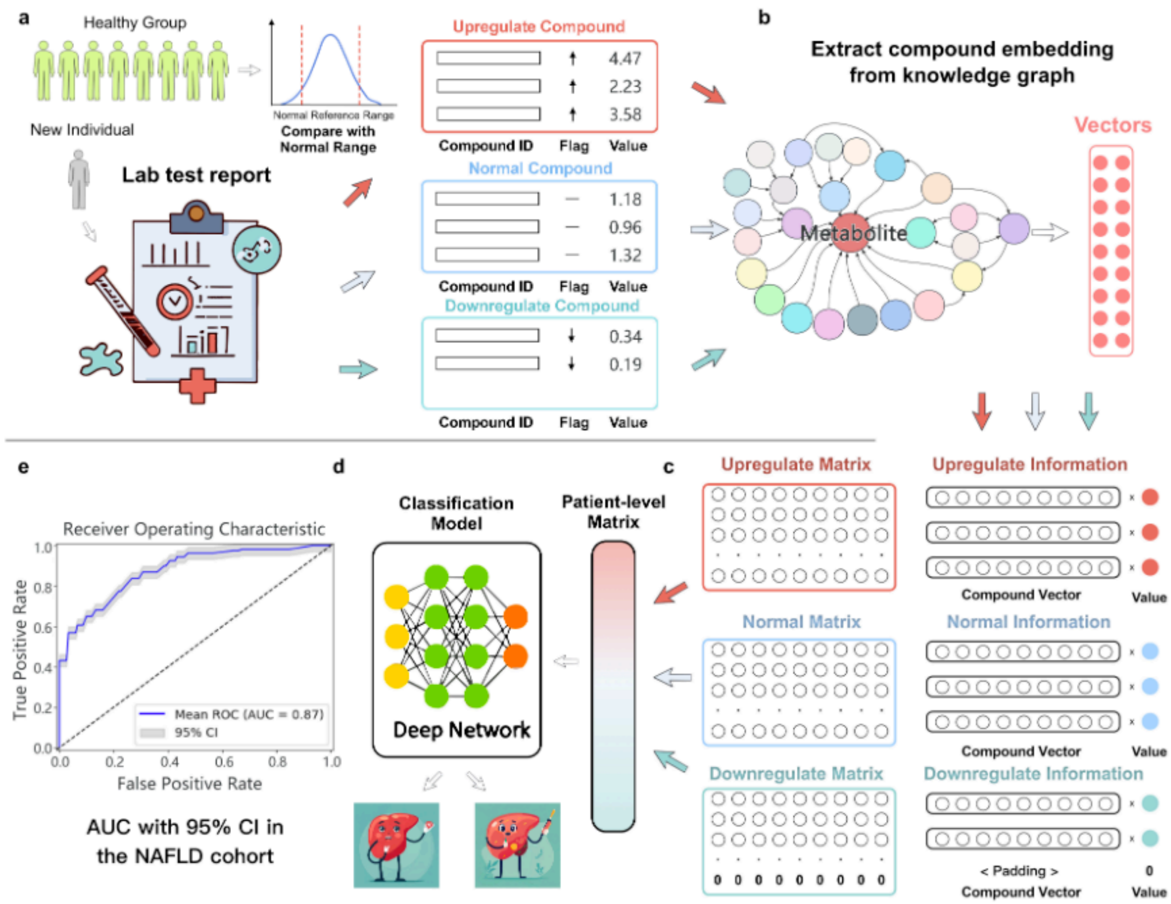


Figure 3: NAFLD diagnosis pipeline using HMKG. The normal range compound expression are calculated as thresholds for distinguishing upregulating, normal, and downregulating compounds. Representation retrieved from HMKG are fed into a NAFLD classification model.

NAFLD Diagnosis Results

- Compared the performance with directly applying traditional ML models to classify patients' different compounds data

	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.

Conclusion

- Proposed a multimodal integration method for knowledge graph representation learning
- Introduced triple contrastive learning and a dual-phase training strategy for aligning multimodal representations
- Demonstrated KG-MRI's effectiveness compared to other KGE methods on a real-world dataset