

Aligning Biomedical Texts and Knowledge Graphs: A Systematic Comparison of Lightweight Alignment Strategies

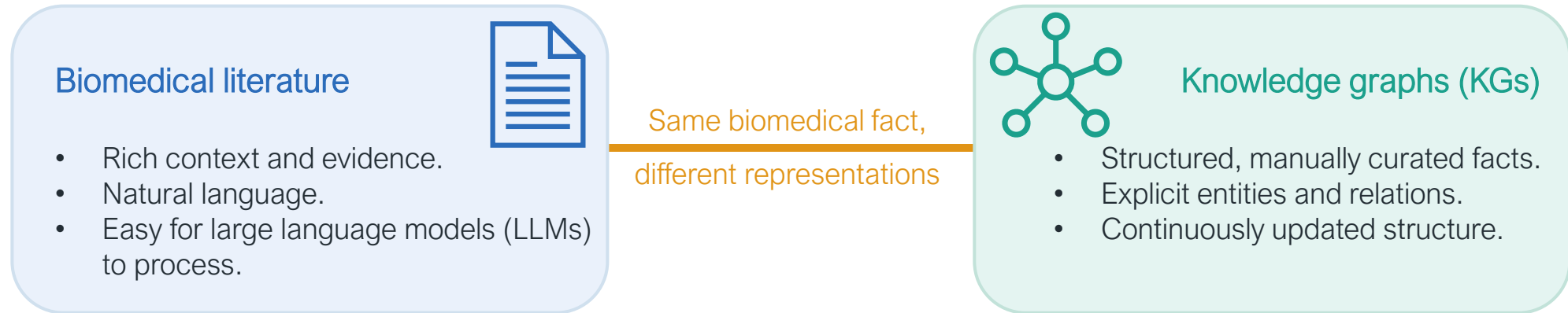
Artem Bisliouk*, Elizaveta Nosova*, Heiko Paulheim, Andreea Iana*, and Rita T. Sousa*



3rd Workshop on Knowledge Graphs and Neurosymbolic AI (KG-NeSy 2026)

October 25, 2026

Biomedical knowledge is available in two complementary forms



Explicit text ↔ triple correspondences

- **Ground LLM responses** in KG information to mitigate hallucinations.
- **Retrieve textual evidence** to support/contradict a given KG fact.
- Extract relations from the literature to **enrich/ complete** KGs.

Which chemicals increase the expression of Parkinson's disease-associated genes, and what experimental evidence supports this?

1 Initial text retrieval

[Curcumin study]

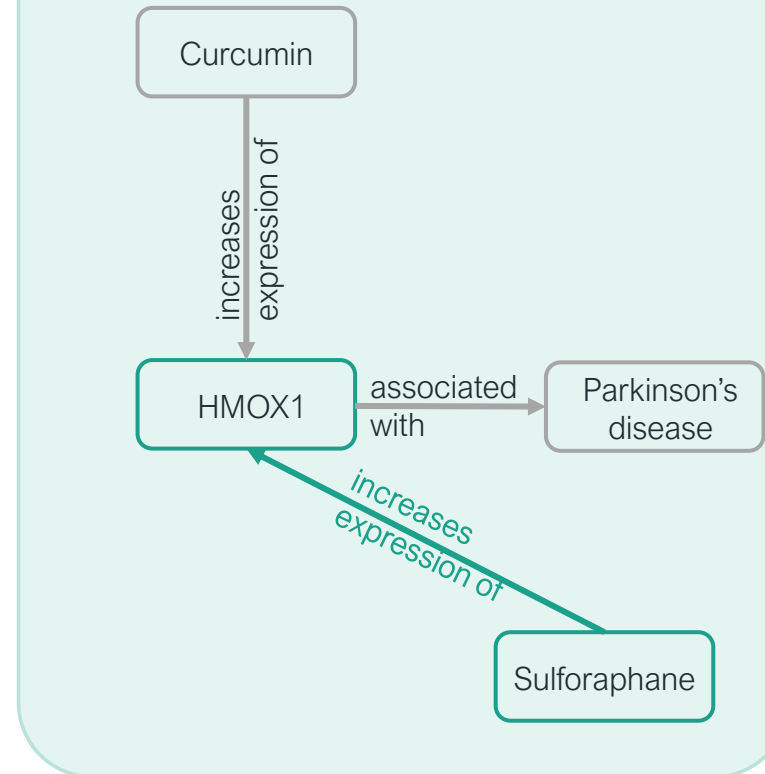
Curcumin increased HO-1 mRNA and protein in human renal proximal tubule cells.

PMID: 11592943

alignment

Match the KG triple

2 Graph expansion



alignment

Find text evidence for the KG triple

3 Retrieve evidence

[Sulforaphane study]

Sulforaphane increased HMOX1 mRNA in human Het-1A epithelial cells (10 μ M, 8 h).

PMID: 31661135

Existing methods do not explicitly align biomedical free-text evidence with KG triples

1. Most methods are optimized for downstream tasks, rather than for learning an explicit mapping between free-text evidence and structured triples.
2. Text-graph alignment methods typically align individual nodes instead of complete facts.

ERNIE: Enhanced Language Representation with Informative Entities

Zhengyan Zhang^{1,2,3*}, Xu Han^{1,2,3*}, Zhiyuan Liu^{1,2,3†}, Xin Jiang⁴, Maosong Sun^{1,2,3}, Qun Liu⁴

¹Department of Computer Science and Technology, Tsinghua University, Beijing, China

²Institute for Artificial Intelligence, Tsinghua University, Beijing, China

³State Key Lab on Intelligent Technology and Systems, Tsinghua University, Beijing, China

⁴Huawei Noah's Ark Lab

{zhangzhengyan14, hanxu17}@mails.tsinghua.edu.cn

K-BERT: Enabling Language Representation with Knowledge Graph

Weijie Liu,¹ Peng Zhou,² Zhe Zhao,² Zhiruo Wang,³ Qi Ju,^{2*}

Haotang Deng,² Ping Wang^{1*}

¹Peking University, Beijing, China

²Tencent Research, Beijing, China

³Beijing Normal University, Beijing, China

{dataliu, pwang}@pku.edu.cn, SherronWang@gmail.com,
{rickzhou, nlpzhe, zhiruowang, damonju, haotangdeng}@tencent.com

Representing Text for Joint Embedding of Text and Knowledge Bases

Kristina Toutanova
Microsoft Research
Redmond, WA, USA

Danqi Chen*
Computer Science Department
Stanford University

Patrick Pantel
Microsoft Research
Redmond, WA, USA

Hoifung Poon
Microsoft Research
Redmond, WA, USA

Pallavi Choudhury
Microsoft Research
Redmond, WA, USA

Michael Gamon
Microsoft Research
Redmond, WA, USA

KG-BERT: BERT for Knowledge Graph Completion

Liang Yao, Chengsheng Mao, Yuan Luo*

Northwestern University

Chicago IL 60611

{liang.yao, chengsheng.mao, yuan.luo}@northwestern.edu

Knowledge Enhanced Contextual Word Representations

Matthew E. Peters¹, Mark Neumann¹, Robert L. Logan IV², Roy Schwartz^{1,3},
Vidur Joshi¹, Sameer Singh², and Noah A. Smith^{1,3}

¹Allen Institute for Artificial Intelligence, Seattle, WA, USA

²University of California, Irvine, CA, USA

³Paul G. Allen School of Computer Science & Engineering, University of Washington

{matthewp, markn, roys, noah}@allenai.org

{rlogan, sameer}@uci.edu

KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation

Xiaozi Wang¹, Tianyu Gao³, Zhaocheng Zhu^{4,5}, Zhengyan Zhang¹
Zhiyuan Liu^{1,2*}, Juanzi Li^{1,2}, and Jian Tang^{4,6,7*}

¹Department of CST, BNRist; ²KIRC, Institute for AI, Tsinghua University, Beijing, China

{wangxz20, zy-z19}@mails.tsinghua.edu.cn

{liuzy, lijuanzi}@tsinghua.edu.cn

³Department of Computer Science, Princeton University, Princeton, NJ, USA

tianyug@princeton.edu

⁴Mila - Québec AI Institute; ⁵Université de Montréal; ⁶HEC, Montréal, Canada

zhaocheng.zhu@umontreal.ca, jian.tang@hec.ca

⁷CIFAR AI Research Chair

SimKGC: Simple Contrastive Knowledge Graph Completion with Pre-trained Language Models

Liang Wang^{1*} and Wei Zhao² and Zhuoyu Wei² and Jingming Liu²

¹Microsoft Research Asia

²Yuanfudao AI Lab, Beijing, China

wangliang@microsoft.com

{zhaowei01, weizhuoyu, liujm}@yuanfudao.com

Research question and contributions

Which design choices drive the alignment quality between structured and unstructured representations of biomedical facts?

1 Modular framework

Plug in any pre-trained text-encoder and KG embedding model, then train a lightweight projection using InfoNCE.

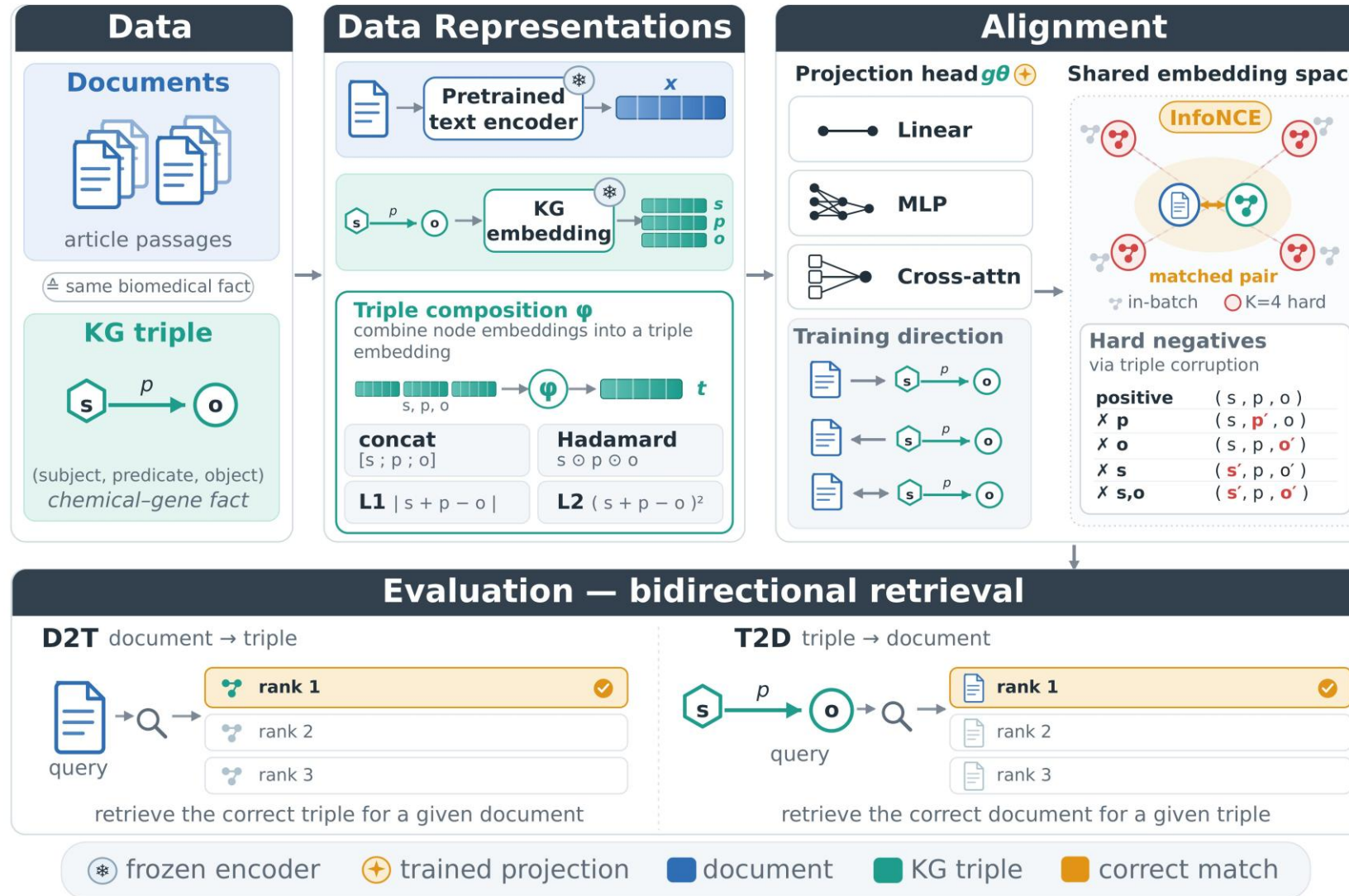
2 CTD-Align dataset

A reusable corpus of 22 293 one-to-one chemical-gene interactions triples linked to source documents with PubMed provenance.

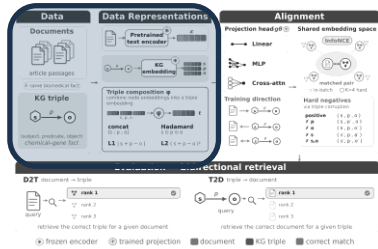
3 Systematic ablation

Compare six design dimensions (text encoder, KG embedding model, projection head, triple composition, training direction, hard-negative sampling strategy).

Methodology

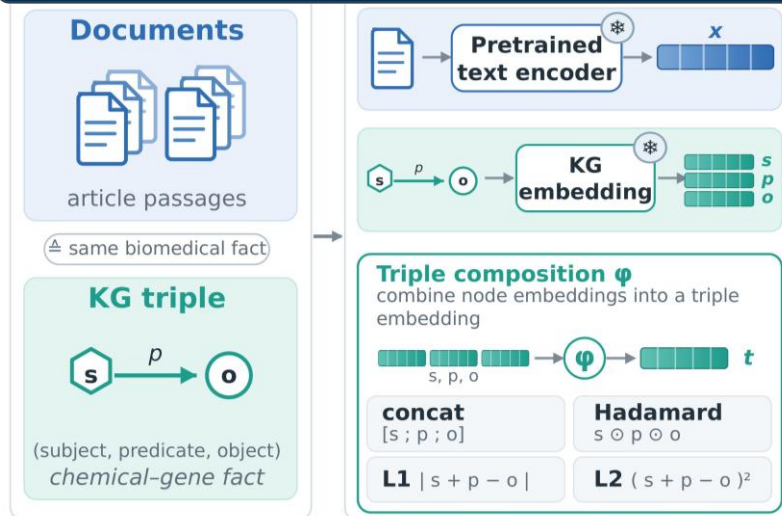


Methodology



Data and data representations

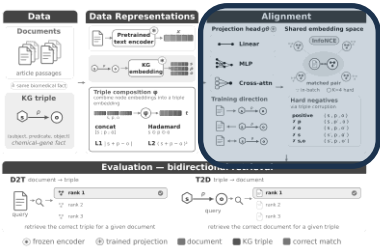
Given corresponding biomedical documents and KG triples, each modality is encoded with a text encoder and a KG embedding model.



The input data consists of biomedical facts, each expressed in two complementary modalities:

- **Document passage embedding**
Sentence or paragraph expressing the relation between the subject s and the object o.
- **KG triple (s,p,o) embedding**
Subject, predicate, object embeddings are combined into one triple representation using four triple-composition operators (concatenation, Hadamard, L1, L2).





Methodology

Projection head

Increasing model complexity.

Linear · MLP · Cross-attention

Training direction

Varying the retrieval target.

text→KG · KG→text · bidirectional

text→KG projects document embeddings into the KG triple space;
KG→text projects triple embeddings into the document space;
bidirectional randomly assigns each batch to one of the two directions.

Contrastive training with the InfoNCE loss

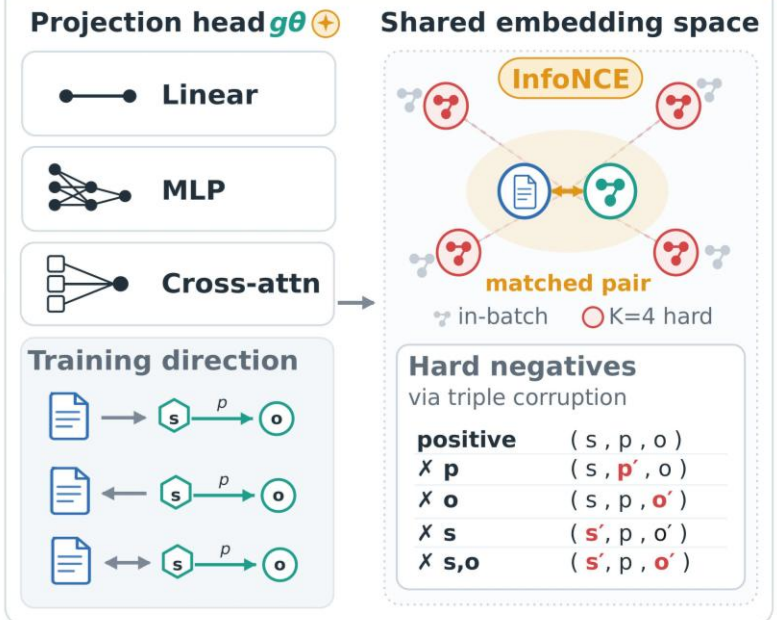
For each anchor (document or triple), in-batch negatives are generated.

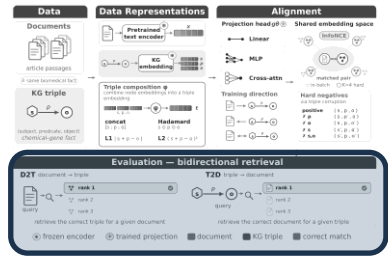
+ optional hard negatives from corrupted triples using 4 strategies

None · perturb s · perturb o · perturb p · perturb s,o · All

Alignment

It learns a lightweight projection that maps one modality into the other modality's embedding space.



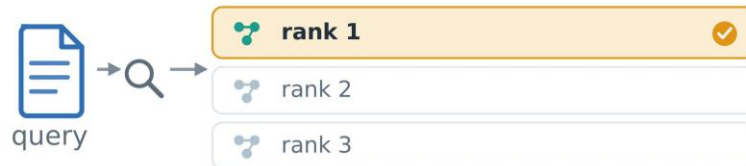


Methodology

Evaluation – bidirectional retrieval

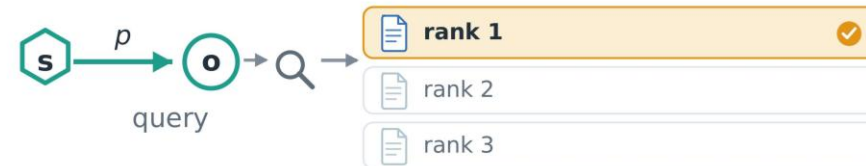
Given a document or a triple, we project it through the learned mapping and rank all candidates in the target modality by cosine similarity.

D2T document → triple



retrieve the correct triple for a given document

T2D triple → document



retrieve the correct document for a given triple

Two complementary tasks:

- D2T (document-to-triple): retrieve the correct KG triple for a given evidence document.
- T2D (triple-to-document): retrieve the correct document for a given triple.

CTD-Align: the triple-document dataset



Comparative Toxicogenomics Database (CTD) KG

- Chemical-gene interactions form the core KG.
- The core KG is enriched by adding related chemical-disease, gene-disease, and gene-pathway associations.
- **1M** edges in the enriched CTD KG.



PubMed Corpus

- Retrieve abstracts for CTD PubMed IDs.
- PubTator provides entity-normalized annotations.
- **63 239** PubMed abstracts with PubTator entity annotations.

22 293 one-to-one triple↔document pairs

- Filter compound-action interactions.
- Candidate spans for each triple are ranked by verbalizing the triple and scoring each candidate with MedCPT.
- When the same span is selected for several triples, the triple with the rarer predicate is selected.

Experimental setup

Text encoders

Embedding the PubMed corpus using pretrained LMs.

BioBERT · PubMedBERT

KG embedding models

Training on the enriched CTD KG to capture different aspects of the structure.

RotatE · TuckER · RDF2Vec

Baseline

Verbalize each triple as text, embed it with the same text encoder, and rank by cosine similarity.

Evaluation protocol

5-fold CV grouped by PubMed paper

MRR · Hits@k · Rank

Full candidate pool (derived from all 22 293 dataset pairs)

Results

Overall performance against baselines

Retrieval performance for the best alignment configuration under each training with the verbalization baseline, in both retrieval settings.

		D2T (document→triple)				T2D (triple→document)			
Model	Config	MRR	Hits@1	Hits@10	Rank	MRR	Hits@1	Hits@10	Rank
Baseline Verbalized	BioBERT / –	8.41 ±0.3	5.47 ±0.3	13.69 ±0.4	927 ±45	7.41 ±0.1	4.48 ±0.1	12.98 ±0.2	845 ±51
Baseline Verbalized	PubMedBERT / –	4.43 ±0.3	2.81 ±0.2	7.45 ±0.4	4826 ±96	2.83 ±0.2	1.68 ±0.2	4.87 ±0.3	5066 ±136
Aligned (text→KG)	PubMedBERT / RDF2Vec	15.99 ±0.6	8.74 ±0.6	30.17 ±0.7	38 ±2	12.13 ±0.3	5.92 ±0.2	24.30 ±0.8	55 ±2
Aligned (KG→text)	BioBERT / RotatE	3.75 ±0.3	1.34 ±0.3	7.84 ±0.6	349 ±18	5.89 ±0.2	2.21 ±0.3	12.41 ±0.3	147 ±7
Aligned (bidirectional)	PubMedBERT / RotatE	15.15 ±0.6	8.51 ±0.6	28.27 ±0.6	48 ±3	5.74 ±0.3	2.29 ±0.3	11.97 ±0.5	159 ±5

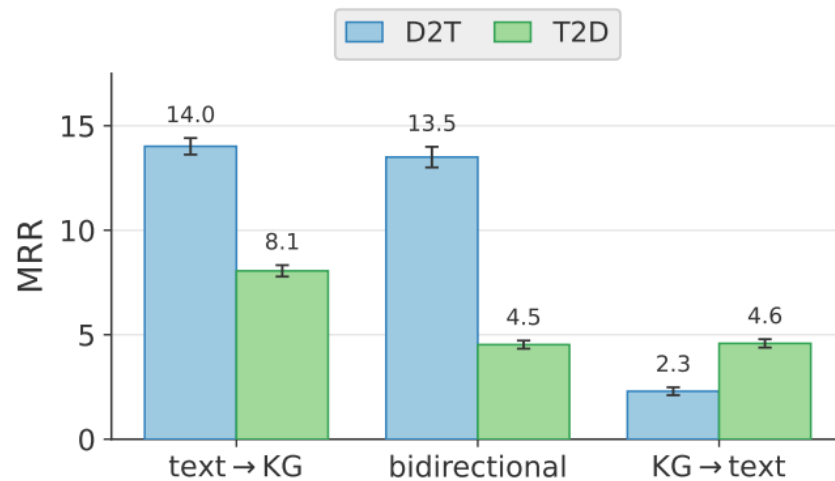
+90% MRR Strongest D2T vs. strongest baseline

+64% MRR Strongest T2T vs. strongest baseline

Learning a projection between the two frozen text and KG embedding spaces substantially **outperforms the baseline** in both retrieval settings.

Results

Training direction



Impact of training direction

Mean MRR across text models, KG embeddings, and architectures (concatenation, in-batch negatives only).

Projecting **text** into the **KG** space **outperforms** the other training directions in both evaluation settings.

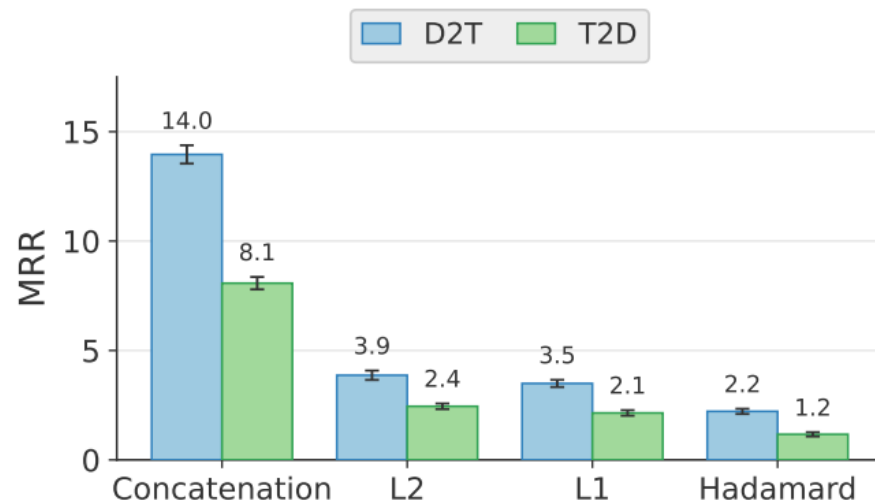
Document → **triple** is a reduction:
discard context that is not needed
once the relation is identified.

Triple → **document** is an addition:
reconstruct missing context and
surface form.

Results

Triple composition is the dominant design choice

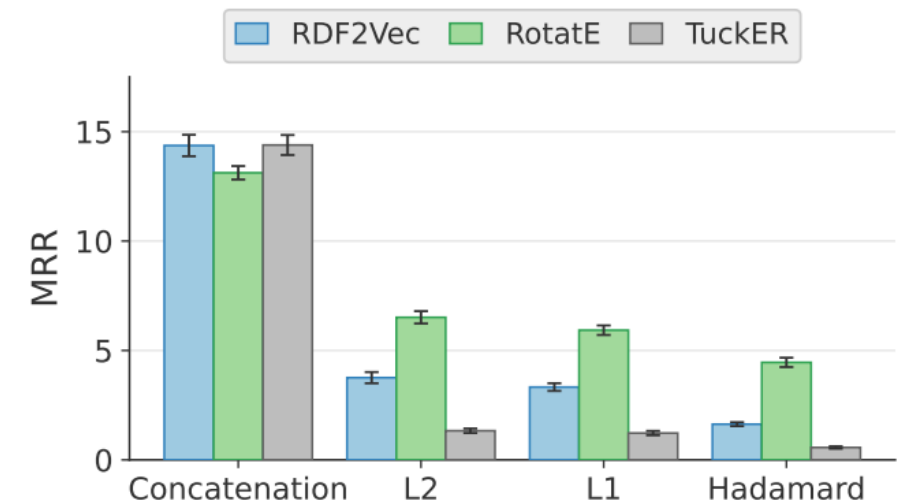
(explains 87% of D2T and 62% of T2D MRR variance)



Impact of triple composition overall

Mean MRR across text models, KG embeddings, architectures, and negative strategies (text→KG training).

Concatenation outperforms the other operators and preserves the separable subject-predicate-object structure in a $3d$ vector, the alternatives collapse it into a d dimensions.



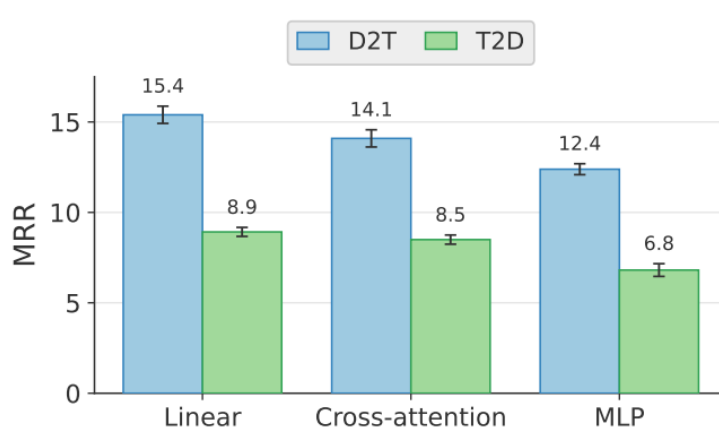
Impact of triple composition by KG embedding

Mean MRR across text models, architectures, and negative strategies (text→KG training, D2T).

Under concatenation, RDF2Vec and TuckER outperform RotatE. **RDF2Vec** may be easier to align because its graph-walk geometry is more distributional.

Results

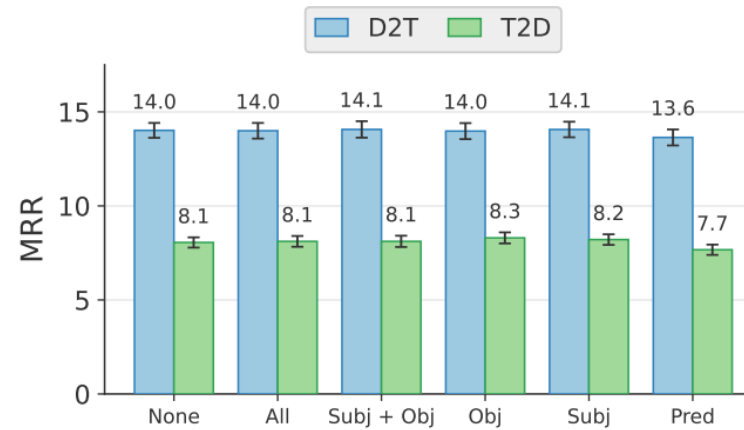
Secondary choice: simpler is usually enough



Projection head

Mean MRR across text models, KG embeddings, and negative strategies (text→KG training, concatenation).

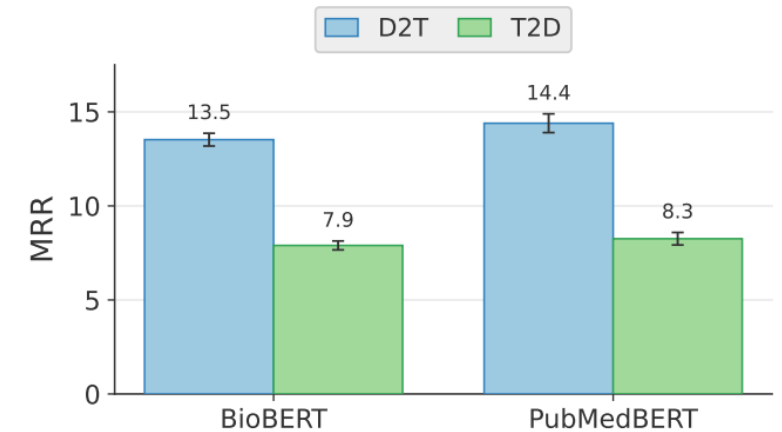
Linear outperforms MLP and cross-attention.



Hard-negatives selection

Mean MRR across text models, KG embeddings, and architectures (text→KG training, concatenation).

Hard negative strategies based on corrupting triples add no benefit.



Text encoder

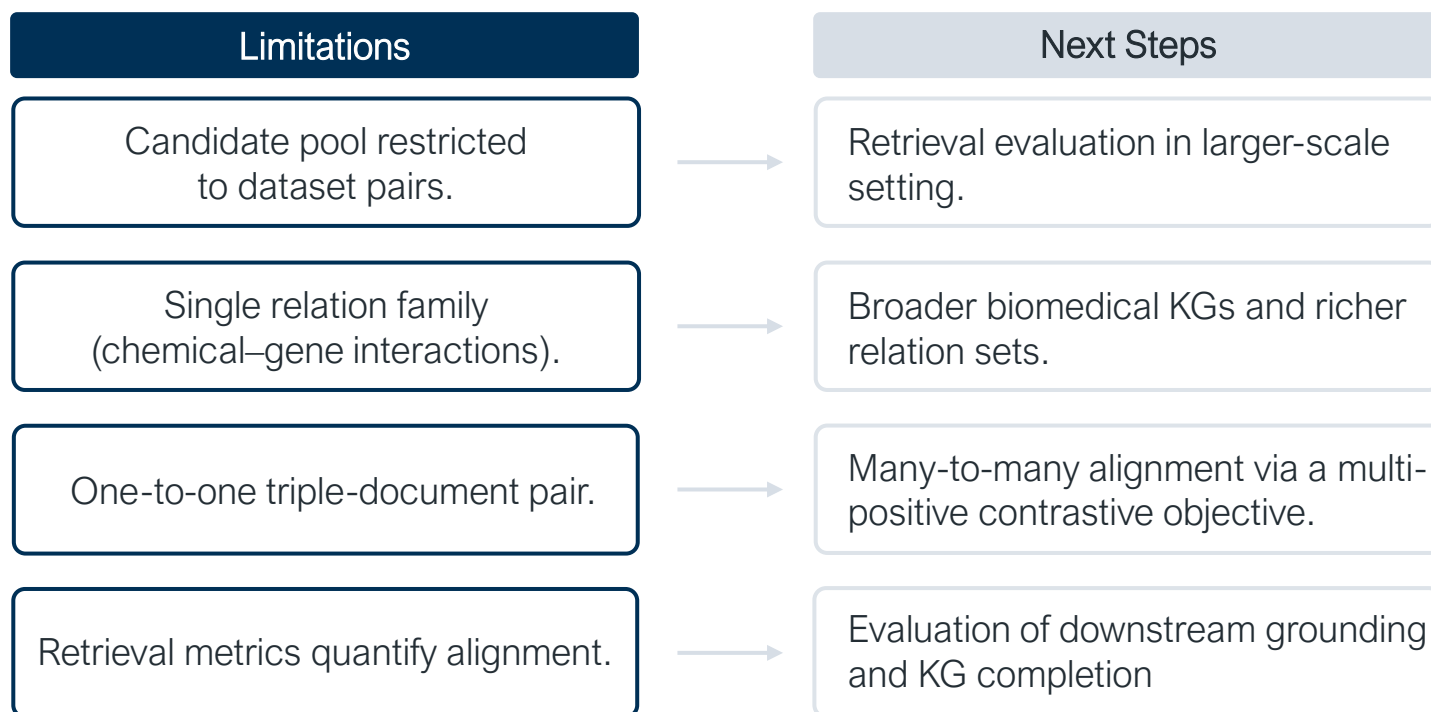
Mean MRR across KG embeddings, architectures, and negative strategies (text→KG training, concatenation).

BioBERT vs. PubMedBERT matters little once alignment is learned.

Takeaways, limitations, and next steps



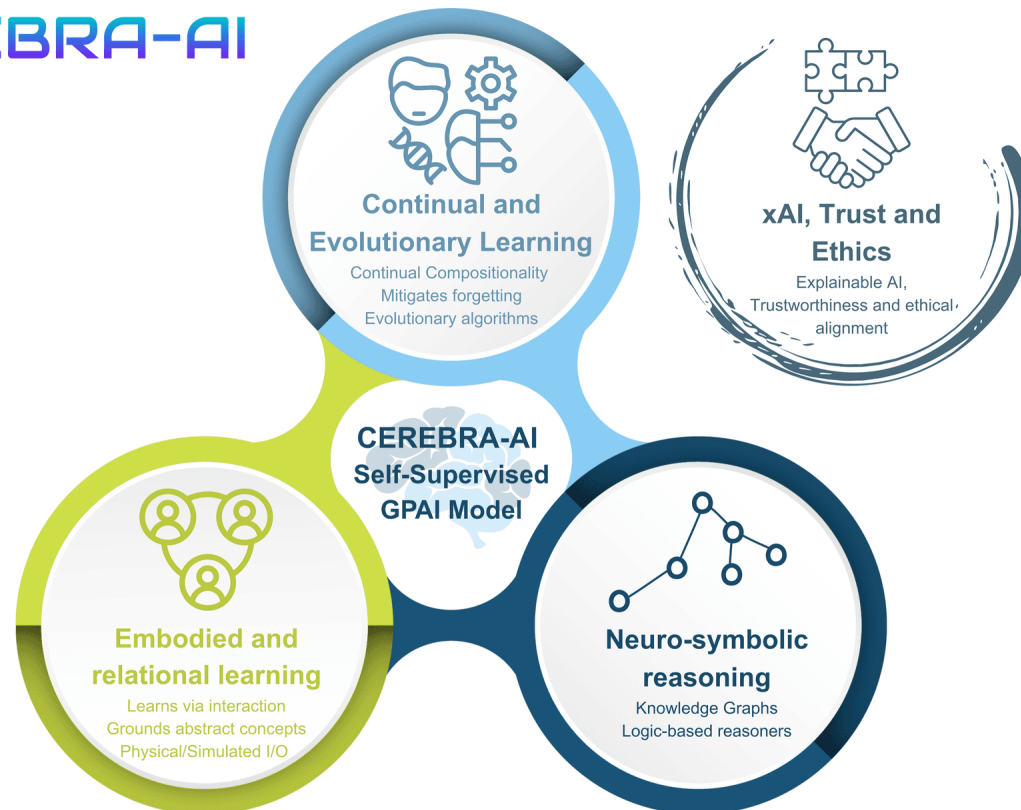
Our findings can inspire more transparent evaluation, including systematic ablations that uncover the components driving performance.



We are Hiring @ Uni Mannheim!

Research position in the EU Horizon project

CEREBRA-AI



Interested?
Contact us to learn more
about the position.

Sven Hertling
sven.hertling@uni-mannheim.de

Rita T. Sousa
rita.sousa@uni-mannheim.de

Heiko Paulheim
heiko.paulheim@uni-mannheim.de

Thank you for your attention!

✉ rita.sousa@uni-mannheim.de

🌐 <https://ritatsousa.github.io/>



3rd Workshop on Knowledge Graphs and Neurosymbolic AI (KG-NeSy 2026)

October 25, 2026