Supplementary Material

Susana Nunes*[1,2], Samy Badreddine[2,3,4], and Catia Pesquita[1]

[1]LASIGE, Faculty of Sciences, University of Lisbon, Lisbon, Portugal
[2]Sony AI, Barcelona, Spain
[3]University of Trento, Trento, Italy
[4]Bruno Kessler Institute, Trento, Italy

1 Theorems and Proofs

Theorem 1. *Considering each triple of the graph as independently and identically distributed, we have*

$$\mathrm{IC}(v) = -\log \frac{\deg(v)}{|\mathcal{G}|}. \tag{1}$$

where $\deg(v)$ *the total degree of an entity node in the knowledge graph,*

Proof.

$$\begin{aligned}
\mathrm{IC}(v) &= I((S = v) \cup (O = v)) &(2)\\
&= -\log[p(S = v) + p(O = v)] &(3)\\
&= -\log\left[\frac{|\{(s,r,o) \in \mathcal{G}; s = v\}|}{|\mathcal{G}|} + \frac{|\{(s,r,o) \in \mathcal{G}; o = v\}|}{|\mathcal{G}|}\right] &(4)\\
&= -\log\left[\frac{\deg^{+}(v)}{|\mathcal{G}|} + \frac{\deg^{-}(v)}{|\mathcal{G}|}\right] &(5)\\
&= -\log\frac{\deg(v)}{|\mathcal{G}|} &(6)
\end{aligned}$$

where:

- $I(E)$ is the information content of event E,
- $\deg^{-}(v)$ counts the in-degree of the node v, $\deg^{+}(v)$ its out-degree, and $\deg(v)$ its total degree.
- The third equality comes from an i.i.d. assumption over the KG data. Intuitively, it counts the frequency of the triples that contain v as a subject and as an object.

*scnunes@ciencias.ulisboa.pt

$\square$

Theorem 2. *Considering each triple of the clustered graph as independently and identically distributed, we have*

$$\mathrm{IC}_c(v) = -\log \frac{\deg(\kappa(v))}{|\mathcal{G}_c|} \tag{7}$$

where the degree of $\kappa(v)$ is calculated in the clustered graph.

Proof.

$$\begin{aligned}
\mathrm{IC}_c(v) &= \log[p(S_c = \kappa(v)) + p(O_c = \kappa(v))] && (8)\\
&= -\log \left[\frac{|\{(s,r,o) \in \mathcal{G}_c; s = \kappa(v)\}|}{|\mathcal{G}_c|} + \frac{|\{(s,r,o) \in \mathcal{G}_c; o = \kappa(v)\}|}{|\mathcal{G}_c|} \right] && (9)\\
&= -\log \frac{\deg(\kappa(v))}{|\mathcal{G}_c|} && (10)
\end{aligned}$$

$\square$

2 Experiments

2.1 Evaluation

We evaluated REx against several baseline methods, including rule-based (AnyBURL [5]), embedding-based (TransE [1], DistMult [11], ComplEx [9], ConvE [3], RESCAL [6]), graph convolutional (R-GCN [8], CompGCN [10]), neuro-symbolic (pLogicNet [7]), and RL-based (MINERVA [2], PoLo [4]) approaches. All hyperparameters respected the default settings.

Table 1: Datasets Statistics.

	Hetionet	PrimeKG	OREGANO
Triples	4499850	8096649	1571899
Entities	45159	129313	98603
Relations	51	35	41
Train	483	7510	117
Valid	121	939	29
Test	151	939	63

The datasets are available at:

- Hetionet - `https://github.com/hetio/hetionet`
- PrimeKG - `https://github.com/mims-harvard/PrimeKG`
- OREGANO - `https://gitub.u-bordeaux.fr/erias/oregano`

Table 2: Mappings across datasets.

Dataset	CHEBI	NCIT
Hetionet	2,333	4,800
PrimeKG	5278	13,210
Oregano	10,451	15,862

3 Results

Table 3: Novel explanatory paths identified by REx for Hetionet Dataset.

	Paths
1	$\textit{Compound} \xrightarrow{\text{causes}} \textit{Side Effect} \xleftarrow{\text{causes}} \textit{Compound} \xrightarrow{\text{palliates}} \textit{Disease}$
2	$\textit{Compound} \xrightarrow{\text{treats}} \textit{Disease} \xrightarrow{\text{associates}} \textit{Gene} \xleftarrow{\text{associates}} \textit{Disease}$
3	$\textit{Compound} \xrightarrow{\text{treats}} \textit{Disease} \xleftarrow{\text{palliates}} \textit{Compound} \xrightarrow{\text{palliates}} \textit{Disease}$
4	$\textit{Compound} \xrightarrow{\text{treats}} \textit{Disease} \xleftarrow{\text{palliates}} \textit{Compound} \xrightarrow{\text{treats}} \textit{Disease}$
5	$\textit{Compound} \xrightarrow{\text{treats}} \textit{Disease} \xleftarrow{\text{treats}} \textit{Compound} \xrightarrow{\text{treats}} \textit{Disease}$

3.1 Domain Expert Evaluation

For the domain expert evaluation ten explanations were randomly selected from both REx and MINERVA. Figure 1 presents the explanations generated from REx for Hetionet.

3.2 Cluster Sensitivity

Both CIC and CIC by relation produce consistent cluster sizes (avg. 10 entities), but CIC by relation yields more fine-grained, relation-specific groupings (≈800 clusters/edge). This additional granularity correlates with improved performance, indicating that CIC by relation provides a more effective clustering strategy.

Dataset	Metapath	Freq.
Hetionet	*Compound* $\xrightarrow{\text{causes}}$ *Side Effect* $\xleftarrow{\text{causes}}$ *Compound* $\xrightarrow{\text{treats}}$ *Disease*	1061
	Compound $\xleftarrow{\text{includes}}$ *Pharmacologic Class* $\xrightarrow{\text{includes}}$ *Compound* $\xrightarrow{\text{treats}}$ *Disease*	145
	Compound $\xrightarrow{\text{resembles}}$ *Compound* $\xrightarrow{\text{resembles}}$ *Compound* $\xrightarrow{\text{treats}}$ *Disease*	123
	Compound $\xrightarrow{\text{treats}}$ *Disease* $\xleftarrow{\text{treats}}$ *Compound* $\xrightarrow{\text{treats}}$ *Disease*	36
	Compound $\xrightarrow{\text{treats}}$ *Disease* $\xrightarrow{\text{localizes}}$ *Anatomy* $\xleftarrow{\text{localizes}}$ *Disease*	15
	Compound $\xrightarrow{\text{treats}}$ *Disease* $\xrightarrow{\text{associates}}$ *Gene* $\xleftarrow{\text{associates}}$ *Disease*	15
	Compound $\xrightarrow{\text{resembles}}$ *Compound* $\xrightarrow{\text{binds}}$ *Gene* $\xleftarrow{\text{associates}}$ *Disease*	14
	Compound $\xrightarrow{\text{treats}}$ *Disease* $\xrightarrow{\text{presents}}$ *Symptom* $\xleftarrow{\text{presents}}$ *Disease*	13
	Compound $\xrightarrow{\text{resembles}}$ *Compound* $\xrightarrow{\text{treats}}$ *Disease*	7
	Compound $\xrightarrow{\text{binds}}$ *Gene* $\xleftarrow{\text{associates}}$ *Disease*	1
	Compound $\xrightarrow{\text{treats}}$ *Disease* $\xleftarrow{\text{palliates}}$ *Compound* $\xrightarrow{\text{treats}}$ *Disease*	1
	Compound $\xrightarrow{\text{causes}}$ *Side Effect* $\xleftarrow{\text{causes}}$ *Compound* $\xrightarrow{\text{palliates}}$ *Disease*	1
PrimeKG	*Drug — indication — Disease — indication — Drug — indication — Disease*	5173
	Drug — indication — Disease — associated with — Gene/Protein — associated with — Disease	86
	Drug — off-label use — Disease — indication — Drug — indication — Disease	12
	Drug — synergistic interaction — Drug — synergistic interaction — Drug — indication — Disease	11
	Drug — contraindication — Disease — contraindication — Drug — indication — Disease	8
	Drug — indication — Disease — indication — Drug — off-label use — Disease	8
	Drug — off-label use — Disease — off-label use — Drug — indication — Disease	7
	Drug — indication — Disease — parent-child — Disease — parent-child — Disease	2
	Drug — side effect — Effect/Phenotype — side effect — Drug — indication — Disease	1
Oregano	*Compound* $\xrightarrow{\text{is\_affecting}}$ *Gene* $\xrightarrow{\text{causes\_condition}}$ *Disease* $\xleftarrow{\text{causes\_condition}}$ *Gene* $\xrightarrow{\text{causes\_condition}}$ *Disease*	432
	Compound $\xrightarrow{\text{has\_side\_effect}}$ *Side Effect* $\xleftarrow{\text{has\_side\_effect}}$ *Compound* $\xrightarrow{\text{is\_affecting}}$ *Gene* $\xrightarrow{\text{causes\_condition}}$ *Disease*	83
	Compound $\xrightarrow{\text{has\_indication}}$ *Indication* $\xleftarrow{\text{has\_indication}}$ *Compound* $\xrightarrow{\text{is\_affecting}}$ *Gene* $\xrightarrow{\text{causes\_condition}}$ *Disease*	49
	Compound $\xrightarrow{\text{is\_affecting}}$ *Gene* $\xleftarrow{\text{is\_affecting}}$ *Compound* $\xrightarrow{\text{is\_affecting}}$ *Gene* $\xrightarrow{\text{causes\_condition}}$ *Disease*	4
	Compound $\xrightarrow{\text{is\_affecting}}$ *Gene* $\xrightarrow{\text{acts\_within}}$ *Pathway* $\xleftarrow{\text{acts\_within}}$ *Gene* $\xrightarrow{\text{causes\_condition}}$ *Disease*	3
	Compound $\xrightarrow{\text{has\_target}}$ *Protein* $\xleftarrow{\text{has\_target}}$ *Compound* $\xrightarrow{\text{is\_affecting}}$ *Gene* $\xrightarrow{\text{causes\_condition}}$ *Disease*	3
	Compound $\xrightarrow{\text{is\_affecting}}$ *Gene* $\xleftarrow{\text{gene\_product\_of}}$ *Protein* $\xrightarrow{\text{gene\_product\_of}}$ *Gene* $\xrightarrow{\text{causes\_condition}}$ *Disease*	2
	Compound $\xrightarrow{\text{has\_code}}$ *ATC* $\xleftarrow{\text{has\_code}}$ *Compound* $\xrightarrow{\text{is\_affecting}}$ *Gene* $\xrightarrow{\text{causes\_condition}}$ *Disease*	2
	Compound $\xleftarrow{\text{increase\_efficacy}}$ *Compound* $\xleftarrow{\text{increase\_efficacy}}$ *Compound* $\xrightarrow{\text{is\_affecting}}$ *Gene* $\xrightarrow{\text{causes\_condition}}$ *Disease*	1

Table 4: Frequency of different metapaths in Hetionet, Oregano and PrimeKG, generated by REx.

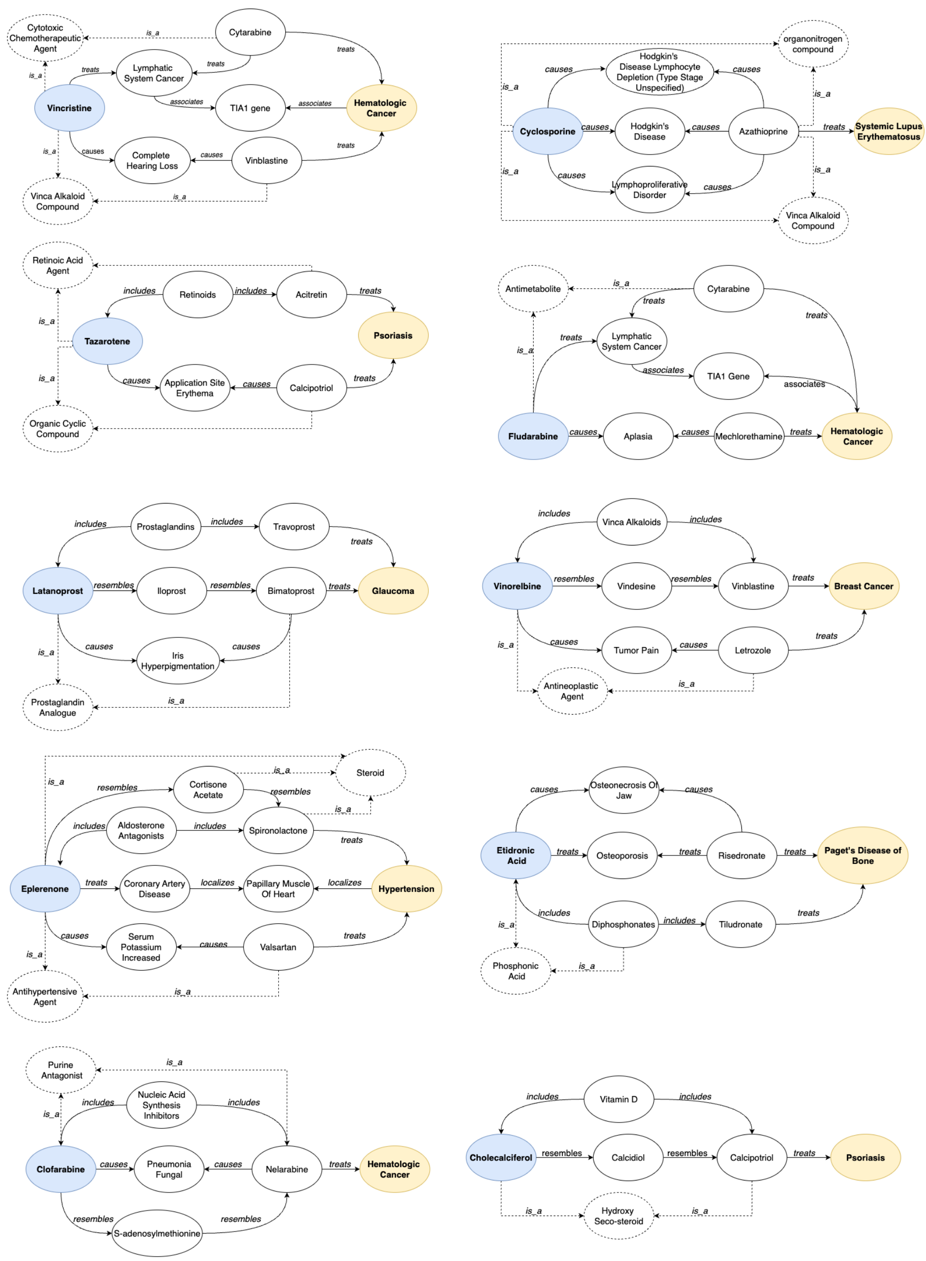

Figure 1: Explanations generated with REx for the expert evaluation, applied to Hetionet.

References

[1] 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.

[2] Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, Luke Vilnis, Ishan Durugkar, Akshay Krishnamurthy, Alex Smola, and Andrew McCallum. Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning. *arXiv preprint arXiv:1711.05851*, 2017.

[3] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.

[4] Yushan Liu, Marcel Hildebrandt, Mitchell Joblin, Martin Ringsquandl, Rime Raissouni, and Volker Tresp. Neural multi-hop reasoning with logical rules on biomedical knowledge graphs. In *The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18*, pages 375–391. Springer, 2021.

[5] Christian Meilicke, Melisachew Wudage Chekol, Manuel Fink, and Heiner Stuckenschmidt. Reinforced anytime bottom up rule learning for knowledge graph completion. *arXiv preprint arXiv:2004.04412*, 2020.

[6] Maximilian Nickel, Volker Tresp, Hans-Peter Kriegel, et al. A three-way model for collective learning on multi-relational data. In *Icml*, volume 11, pages 3104482–3104584, 2011.

[7] Meng Qu and Jian Tang. Probabilistic logic neural networks for reasoning. *Advances in neural information processing systems*, 32, 2019.

[8] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 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.

[9] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *International conference on machine learning*, pages 2071–2080. PMLR, 2016.

[10] Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. Composition-based multi-relational graph convolutional networks. *arXiv preprint arXiv:1911.03082*, 2019.

[11] 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.