Response to Reviewers

Reviewer 1

We thank the reviewer for their comments. We have strengthened the mentioned sections; comments are addressed in order (as extracted from the review text), below.

1. **Remark:** While the evaluation considers many knowledge graph embedding methods such as TransE, ComplEx, DistMult, RESCAL, RotatE, I think these methods cannot leverage the new information added in the extended knowledge graphs. For example, authors acknowledge that 'relations P31 and P279 are not symmetric; therefore DistMult, which can encode symmetric relations, cannot leverage P31 and P279 edges'. Following this line of reasoning, P279 (subclassof) is a hierarchical relation, which should be evaluated with methods that can encode hierarchies such as [refs]. Also in the case of P31 relation, it is well known that it is a N-1 relation, therefore a method encoding N-1 relations (such as TransD) should be chosen for this property. In summary, I think the evaluation does not match the hypothesis the authors intend to propose.
   Response: We agree that relations such as P31 (instance-of) and P279 (subclass-of) are inherently asymmetric and hierarchical, which classical models like DistMult are not well-suited to encode, we address this in 5.2. We felt that this was easy to include, due to the nature of the library that we were using (i.e., DGLKE), so we could do a comparison across all implementations it contained. As such, we left out TransD as an oversight. We have run the same experiments with TransD as we did for the other methods, and we have added them to the tables found in the paper. For reference, we also include them here, in Tables 1 & 2.

		Evaluation across KGs			Evaluation with T_{237}	
Model	Metrics	FB15k-237	FB15k-238	FB15k-239	FB15k-238	FB15k-239
TransD	MRR	0.210	0.228	**0.259**	0.0032	0.0030
	MR	330.50	287.87	**155.05**	7168.6280	6828.5341
	HITS@1	0.116	0.126	**0.146**	0.0025	0.0023
	HITS@3	0.249	0.277	**0.302**	0.0027	0.0025
	HITS@10	0.390	0.419	**0.485**	0.0033	0.0030

Table 1: TransD model evaluation on FB15k variants. Metrics shown for FB15k-237 are: Mean Rank (MR), Mean Reciprocal Rank (MRR), and Hits@K (K=1,3,10). Other values are to be filled after evaluation on FB15k-238 and FB15k-239.

2. **Remark:** "Fig. 1 is very useful, it is not self-explainable. Maybe adding edge labels such as 'type_of' or 'subclass_of' and adding a richer caption can help to improve this figure."
   Response: Unfortunately, we could not devise a way to add edge labels, without it becoming too cluttered. However, we expanded the figure caption to more clearly identify the "type_of" and "subclass_of" relationships, arrow color and presentation meaning and various node types such as color and what the letters represent, to improve interpretability without needing to refer back to the introduction.
3. **Remark:** "Section 3.5 point (d): t-sne and umap should be t-SNE and UMAP maybe."
   Response: We have corrected these consistently throughout the paper.
4. **Remark:** "There are several references pointing to ArXiv while published versions of those

Model	Metrics	$M_{238} \leftarrow$ $T_{238-237}$	$M_{239} \leftarrow$ $T_{238-237}$	$T_{239-238}$	$T_{239-237}$
Both	MRR	**0.3784**	0.0143	0.0006	0.2615
	MR	**100.8378**	8476.8943	8628.7851	128.5453
	HITS@1	**0.2485**	0.0132	0.0000	0.1510
	HITS@3	**0.4569**	0.0143	0.0002	0.3061
	HITS@10	**0.6049**	0.0151	0.0007	0.4746

Table 2: This table reports the results of our ablation-like study with TransD, where we change which component of the data against which we evaluate. M_x denotes a model being trained with FB15k-x. T_{x-y} denotes test data, where $x - y$ refers to the set difference resulting in data that can only be found in FB15k-x.

papers exist."
Response: We have corrected these consistently throughout the paper.

Reviewer 2

We thank Reviewer 2 for their remarks and and suggestions on improving the state-of-the-art and flow of the manuscript. Where appropriate, we have made corresponding changes, explained below.

5. **Remark:** "The introduction of the synthetic datasets in the Introduction and Section 3 needs better contextualization. I guess the synthetic data somehow tries to replicate real-world data patterns, to be insightful. It would be interesting to add examples for each of the SKG graphs introduced for the reader to understand what they could represent in reality."
   Response: We have added real world examples in the A.1 section of the Appendix and also enhanced Figure 2's caption in order to be more explicit.

6. **Remark:** "I didn't understand the motivation for SKG-237. Other than having a similar number of nodes and degrees with FB15k-237, does it replicate any important structural patterns? If the patterns are just random, then isn't it naturally expected that performance is bad on it? In that case, I find that it doesn't add much to the insights and the story would be clearer without it. If it does replicate important patterns, please clarify and elaborate Section 3.3 (too short at the moment)."
   Response: The purpose of this investigation was not just to replicate degree distributions and node counts, but also to evaluate what KGE models actually expect in terms of knowledge (or data) to produce effective embeddings. Specifically in this case, we study a rather aggressive change: the naive removal of any implicit semantics, and thus produce a graphs that *superficially* globally mirrors a established benchmark dataset. This is a step that distinguishes between the impact of structure and semantics in link prediction tasks. Its poor performance is informative, indicating that structure is likely more important than we realise and that real-world knowledge graphs depend on more than just degree or volume. Our experiments with SKG-237 provide evidence for this observation and initiate a conversation about the types of structural assumptions that KGE models implicitly expect. We have updated the description of SKG-237 in Section 3.3 to give a more thorough and precise explanation of its structural features. We make clear which characteristics, like degree clustering were purposefully kept and which were not.

7. **Remark:** “I find the discussions of KGE performance over SKG isotopes (in Section 5.1), which seems to be the main story of the paper, too brief. The last paragraph cuts in the middle, so I wonder if there may have been a problem with the submission."
   Response: Thank you for pointing out the omission. We are not exactly sure what happened to the original discussion. However, the discussion in Section 5.1 is now completed and to clarify the observations and emphasize our main outcomes. The relationship between richness and learning persists. Caution must be taken when introducing structural complexity and semantic distance, as these may enhance representational richness while also making the link prediction task more challenging for traditional embedding models.

8. **Remark:** “Another concern is that the second contribution, on the visual analysis of the embeddings, is impossible to review because it is present only in appendix and not attached to the submission. If the authors submit a revision, please include the appendix, or at least include the most important Figure in the main paper (since it is even part of the abstract)."
   Response: Appendix is now correctly attached in our submission. We also prompt you to explore our GitHub repository as well were all of those vizualizations reside: `https://github.com/kastle-lab/kge-impact`

9. **Remark:** “I don’t understand how it’s possible to create dataset splits for KGEs on let’s say SKG-4, if you explain in Section 3 that they all have disconnected components. For SKG-4, every single triple has an object with a node degree of 1. If you put any triple in valid/test, then it means that its object will never be seen in training. But KGE methods are designed for transductive link prediction; their results will be random for nodes that are never seen in training. How did you manage the splits?"
   Response: It is true that SKG-4 has a lot of disconnected components and that many triples involve nodes with degree 1, which means that objects only appear once. However, our choice to create dataset splits by randomly shifting and splitting triples is in line with standard knowledge graph embedding (KGE) benchmarking methods. The splits purposefully maintain the realistic challenge of learning in sparse and uneven graph structures, which is a crucial scenario for real-world applications where unseen nodes or rare entities frequently occur (Wang et al., 2017) [5]. This is despite the fact that KGE methods are naturally transductive and depend on entities being seen during training (Bordes et al., 2013; Trouillon et al., 2016) [1, 4]. Maintaining this random split also allows us to thoroughly assess how resilient embedding techniques are to this sparsity and how well they adapt over a small context, as seen in (Sun et al., 2019) [3]. Moreover, the SKG datasets are fabricated benchmarks created to examine the basic behaviors of embedding algorithms under controlled variations in structure, including extreme scenarios like nodes with degree one, in addition to being used for standard transductive link prediction (Dettmers et al., 2018; Wang et al., 2018) [2, 6]. Therefore, rather of working against the expectations of transductive learning, our method encourages the evaluation to include edge situations that highlight drawbacks and direct future model advancements. To be clearer, we have adapted Section 3 to explicitly specify the dataset splits. This clarification removes previous ambiguity and strengthens the justification of our experimental setup.

10. **Remark:** “Table 6: How are the degree centrality values computed? Maybe I’m using the wrong concept, but I thought it would be the number of edges (in or out) connecting to a node. Then doesn’t the graph in SKG4 have at least 1 edge connecting to each node? Same question for the other graphs. I expected the average degree to then be greater than 1.0, I don’t understand how the values 0.003 etc. are obtained."

Response: The degree centrality values in Table 6 were computed using NetworkX's **normalized** degree centrality function. This metric is defined as the node's degree divided by the maximum possible degree (i.e., number of nodes minus one), resulting in values normalized between 0 and 1. Therefore, a value like 0.003 indicates that the node connects to about 0.3% of all other nodes, consistent with the sparse and large graphs in SKG-4 and others. Additionally, we have adapted Section 3.7 to be more precise regarding these computations to avoid ambiguity.

11. **Remark:** "Table 6: What does the difference column represent?"
Response: The difference column showcases the change in values between graphs (up for larger down for smaller). The caption of tables 6 and 7 are expanded now to avoid any ambiguities.

12. **Remark:** "I think Figure 1 could be more clear by adding types on the edges that are different. For example, right now, for SKG5 it is stated that 'the lavender property is always attached to the top node'. But it's not explicit what defines a 'top node' (graphs don't have orientations); I assume what defines it is the type of the edge."
Response: We have enhanced Figure 1 expanding the figure caption, explaining the "type_of" and "subclass_of" relationships, arrow color and presentation meaning and various node types such as color and what do the letters represent in order to improve interpretability without needing to refer back to the introduction.

13. **Remark:** "Figure 2 caption cuts in the middle."
Response: Thank you for pointing out the omission; we have completed the caption.

14. **Remark:** "Section 5.2, line 50: In fact, DistMult can only represent symmetric relationships, since its score function cannot model asymmetry."
Response: Thank you for the correction; we have made that explicit in the paper. We also prompt you to take a look at remark **1.** where we also evaluate our graphs using another appropriate model.

References

[1] Bordes, A., Usunier, N., Garcia-Duran, A., andOksana Yakhnenko, J.W.: Translating embeddings for modeling multi-relational data. In: Proceedings of the 27th Annual Conference on Neural Information Processing Systems. p. 2787– 2795 (2013)

[2] Dettmers, T., Minervini, P., Stenetorp, P., Riedel, S.: Convolutional 2d knowledge graph embeddings (2018)

[3] Sun, Z., Deng, Z.H., Nie, J.Y., Tang, J.: Rotate: Knowledge graph embedding by relational rotation in complex space. In: Proceedings of the 7th International Conference on Learning Representations (ICLR) (2019), https://openreview.net/forum?id=HkgEQnRqYQ

[4] Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., Bouchard, G.: Complex embeddings for simple link prediction. In: Balcan, M., Weinberger, K.Q. (eds.) Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016. JMLR Workshop and Conference Proceedings, vol. 48, pp. 2071–2080. JMLR.org (2016)

[5] Wang, Q., Mao, Z., Wang, B., Guo, L.: Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering **29**(12), 2724–2743 (2017). https://doi.org/10.1109/TKDE.2017.2754499

[6] Wang, X., He, X., Cao, Y., Liu, M., Chua, T.S.: Kgat: Knowledge graph attention network for recommendation. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. p. 950–958. KDD '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3292500.3330989, https://doi.org/10.1145/3292500.3330989