Response to reviewers

We sincerely appreciate the time and effort the reviewers and editors have dedicated to evaluating our manuscript. Their insightful comments have been extremely valuable in improving the quality of our work. Below, we provide point-by-point responses to each comment. Changes to the document have been highlighted in yellow.

Reviewer 1

Comment 1

*On The Methodology ** The method, as I understand it, is at least partially incremental. The authors put together three pre-existing components (edge2vec, a GNN, and GNNExplainer) and only use a different dataset. I guess that the major contribution here could be the domain and the task, which I agree is very valuable and important. I have a few questions regarding the setup, but I do apologize if I missed something. It seems to me that edge2vec is used before training the GNN, but it also seems that edge2vec is used on the entire graph before training the GNN. Isn't this process leaking some information in the second step? Like, when the authors remove the links to create the testing data, nodes embedding will still encode that kind of information? My assumption here is that the authors are actually using edge2vec as pre-initialization for the GNN.*

Response 1

Thank you for your insightful comment. Indeed, edge2vec is used before training the GNN, and it is trained on the entire graph. However, we believe this does not necessarily introduce data leakage in our setup. As you mention, the edge2vec embeddings serve as pretrained representations of the graph structure, similar to how word embeddings are used in LLMs (ie. in BERT [https://arxiv.org/abs/1810.04805]). Since edge2vec is an unsupervised method, it does not utilize task-specific labels, meaning the GNN still needs to learn how to map these embeddings to the target task. Nonetheless, we have added a paragraph to the main text to clarify it.

Comment 2

There are many things that would require more details. For example, how is edge2vec used for link prediction in Table 5? How was the optimal value for edge2vec selected (see Table S5)? Was optimization with raytune also run for the baselines (Table5)? In addition to this, the authors' model was tuned with raytune, but is this run even at cross validation time? I could not find the model architecture so I looked at the code and seem complicated enough (few graph layers + batch norms) that it might be worth adding to the paper for reproducibility purposes.

Response 2

Thank you for the feedback. To clarify the usage of edge2vec for link prediction in Table 5, we employed the dot product of the edge embeddings to compute the likelihood of a link between nodes. The optimal values for the parameters of edge2vec were fine tuned using RayTune, with the parameters being reported in Table S5. Additionally, the optimization process using RayTune was only applied to our proposed model, not to the baseline models, as they were directly taken from the respective papers with their default hyperparameters. Regarding the model architecture, we have now included a detailed description in the paper for better reproducibility.

Comment 3

*** On Writing ** In general, writing would benefit from more work as there are a few sentences that are unclear: "Next, it is the link prediction for each drug-symptom node embeddings pair by using the dot product as scoring function" (page 4). >> I think this sentence is missing a verb or something similar (there are a few typos in the paper). "in the training dataset the supervision edges and the message passing edges are the same; in the validation dataset the message passing edges are the training edges (message and supervision) ... that are different from the training and validation supervision edges" (page 6). >> This is very long and would benefit from being split and rephrased. I'd recommend the authors to improve the general presentation and add more details. In particular, related work never introduces the methods in details (e.g., GraphSAGE). Table 1 is a bit too "high-level" as a summary of an entire field (which I understand you don't need to cite all of, but I'd still argue that some polishing here could improve the paper). I often found myself scrolling up and down to retrieve information from previous parts of the paper. There are several minor writing problems (e.g., typos) that, when taken together, become a bit distracting. The paper sometimes uses British and sometimes American English. "The code is freely accessible with an open license at https://github.com/PPerdomoQ/rare-disease-explainer" appears at least 3 times. Versions of the Python packages are reported in multiple sections where they could probably be appendix material. First I read "GraphSAGE and GNNExplainer were implemented using PyTorch Geometric version 2.0.4" (page 4) but then "The GraphSAGE model was created using the DeepSNAP library" (page 6). >> I guess this is because DeepSNAP uses PyTorch Geometric, but again, this is something that could have been polished. Another issue comes from reference [3]: "and only in Europe about 36 million people suffer from rare diseases." When I click on the link, the actual figure is "between 27 and 36 million people live with a rare disease," which I understand might look like a minor thing, but I think this should be reported correctly. The bibliography contains many typos/papers with missing journal names, etc. Also, there are >150 references but the paper seems to have less than 70? Quality of the images seem low, I'd try to export them using higher resolution.*

Response 3

Thank you very much for all your comments. We have reviewed the spelling and grammar of the paper and hope to make it much more readable. We have removed unnecessary parts (versions of the Python, repeated links…) and clarified sentences that could be confusing. We have also added an extra column to Table 1 which includes possible applications of the

methods hoping to offer a more complete view of the XAI field. We have corrected the errors involving the references and tried improving the quality of the images.

Reviewer 2

Comment 1

One question I would have for the drug repurposing part is the extent to which the system returns false positives specifically. These types of repurposing recommendations could be imagined to have a relatively high rate of false positives.

Response 1

Thank you very much for your question. There are different strategies that have been used to reduce the false positive rate. First, we performed negative sampling with a 1:1 ratio, ensuring a balanced training set that helps the model distinguish between true and false associations. Nonetheless, this ratio can be increased to further reduce the false positive rate. Additionally, since the model demonstrates a high AUROC, we can adjust the decision threshold to be more stringent, reducing false positives at the cost of recall. Finally, our explainable AI module can help the user to interpret whether a certain prediction is a false positive (ie. if the provided explanation makes no sense). These combined strategies help mitigate the risk of false positives while maintaining predictive performance.

Comment 2

Of course clinical validation of a given recommendation for drug repurposing would be out of scope for the current study but perhaps clinical validation could be approximated by determining which drugs had failed in investigations for particular indications and using that as a quasi validation dataset.

Response 2

Thank you very much for your feedback. Using a clinical dataset is indeed a great idea. In this paper we tried to capture this information by performing a manual bibliographical search when evaluating the predictions. Nonetheless, making use of a clinical trial dataset is a great idea for a future version of the model.

Comment 3

I didn't fully understand if the chosen interpretability method was reproducible and if not what steps are taken to mitigate stochastic variability in performance?

Response 3

Thank you very much for your question. The method follows an approach similar to training a neural network: two people can have the same neural network and the same dataset but still train two different models. This is due to the random processes in the training (ie. random split of the datasets, dropout affecting different neurons, etc.). Our approach is similar, as training GNNExplainer implies different random processes. However, we have tried an implement different methods to reduce this randomness (ie. multiple runs of the model). Nonetheless, one could always use a fixed seed to ensure reproducibility).

Comment 4

Why use the 2021 versions of the source databases Monarch/DrugCentral/TTD without updates to more recent versions?

Response 4

Thank you very much for your comment. At the time the work was conducted, the 2021 versions of Monarch, DrugCentral, and TTD were the most up-to-date resources available, and our primary objective was to showcase the effectiveness of our novel method. We focused on demonstrating the methodology rather than conducting an exhaustive update of the source databases. In future work, we plan to incorporate the latest versions of these databases to further enhance and validate our findings.

Comment 5

Would there be expected to be significant changes in the underlying databases since then that might affect the outcome of the analysis?

Response 5

This is indeed a very interesting comment. While there have been updates to the underlying databases since 2021, we do not expect these changes to significantly impact the overall outcome of our analysis. The core relationships and patterns our method captures remain largely stable over time. If the structure or format of the databases were to change, our approach would need minor adaptations, but these would be easy to implement. Importantly, the number of records alone does not affect our methodology, as it is designed to be robust to dataset size variations.

Comment 6

"this explanation is classified into complete..." <-- maybe "classified as complete" would work better?

Response 6

Thank you very much for pointing this out. We have reviewed and corrected the spelling and grammar of the manuscript.

Reviewer 3

Comment 1

In Section 3.2: The overview of the explainable AI in graphML is a bit empty. Many more methods have been proposed for this task (e.g., GRETEL, PGExplainer) , and it would be useful to have a table for them in the style of Table 1.

Response 1

We really appreciate the suggestion to expand the discussion on XAI in graphs. In response, we have added references to additional methods, including PGExplainer and GRETEL. Furthermore, we have included a summary table (Table 2) that categorizes these methods based on their explanation type, applicable tasks, and main drawbacks. We believe this addition provides a clearer and more complete view of explainability methods in KGs.

Comment 2

Then an explaination of why GNNExplainer was picked out of all the others should be provided. Once all the pros of this model have been identified, then - for the sake of transparency - the disadvatages should also be given. For example, it should be highlightedt he fact that GNNExplainer is a post-hoc explainer, hence the explanations might not always be faithful. For example, if the GNN is trained with noisy data, GNNExplainer may highlight irrelevant edges or nodes simply because they correlate with predictions. Or again, GNNExplainer often focuses on the local neighborhood of a node, while some GNNs (e.g., Graph Attention Networks) might base predictions on long-range dependencies.

Response 2

We appreciate your comment. We would like to highlight that the requested discussion regarding the strengths and weaknesses of GNNExplainer was already present in the manuscript (page 4). However, we have now further clarified these aspects by explicitly mentioning that as a post-hoc method, GNNExplainer’s explanations may not always be faithful. We believe these clarifications address the reviewer's concerns while maintaining the original structure of the section.

Comment 3

In Section 4.1: the authors write that they used the dot product as a scoring function for each drug-symptom node embeddings pair> Why was this operator used an not some other (e.g., cosine similarity)? -

Response 3

This raises a really good point. We chose the dot product as the scoring function for drug-symptom node embeddings because it directly reflects the strength of the relationship between entities in our Knowledge Graph. Unlike cosine similarity, which only measures the directional alignment of vectors, the dot product captures both directional similarity and magnitude, allowing us to better model the intensity and confidence of potential links. Furthermore, many Knowledge Graph embedding methods (e.g., TransE, DistMult) rely on bilinear scoring functions that inherently use dot product-based similarity. This makes it a natural choice for ranking predicted links, as higher dot product values indicate stronger associations without requiring additional normalization. Given that our task involves link prediction, using dot product ensures consistency with widely adopted KG embedding frameworks while preserving meaningful variations in relationship strength.

Comment 4

In Section 4.2: how is the information gathered from these three databases? were they compatible? which features does a note have and which features does a edge have? The paper in this sense might benefit from some restructuring, as some of this information is introduced in this section but then more can be found in Section 5.1. In general, it would be nice to have a figure where max 5 rows are considered for each database, and then it is shown how the graph for each of the mini-databases is built and then how they are merged.

Response 4

Thank you very much for your points. Our idea was that Section 4.2 describes how data was gathered and structured, while Section 5.1 presents the results of its aggregation. The databases had differences in format, but we standardized entity representations to ensure compatibility. While a figure could help illustrate this, we believe our textual explanation sufficiently clarifies the integration process and adding might not provide substantial additional clarity.

Comment 5

In order to train a GNN for link prediction, negative examples are needed and how the negative sampling is performed has a great impact on the final results.
Could you please elaborate on this aspect?

Response 5

This is indeed a very good point. In our case we performed a random 1:1 negative sampling. This is, for each true supervision edge, a false edge is created. To create it the tail of a relationship is removed (considering that an edge consists of a head node and a tail node connected by a relationship) and a new tail node is randomly selected from the set of all possible nodes, ensuring that the newly formed edge does not exist in the original dataset. We also tested different negative sampling ratios which can be seen in Table 5.

Comment 6

Section 4.3.5: The authors write: "Regarding the parameters of GNNExplainer, because the graphs are highly connected, explanations were generated by using the 1-hop neighborhood around the graph." This is a very tight neighbourhood, was any ablation study conducted to support this choice? -

Response 6

Thank you very much for your feedback. This is indeed a very good point. There were two main reasons to take this decision. On the one hand, as mentioned in the paper, we want to obtain simple and understandable explanations. Looking at broader neighborhoods increased the chances of hitting a hub node and leading to over complex explanations. And the second reason is that biological explanations often happen locally in disease modules, as indicated by Paci P. et. al. (2022) (https://www.nature.com/articles/s41540-022-00221-0)

Comment 7

Minor comments: 1. the figures occupy much more space then needed, with some being vertically aligned when there was enough space for a horizontal alignment and with a suboptimal allocation of the parts in the figure (e.g., figure 5). I strongly encourage the authors to move figures and table around to minimise the waste of space. 2. sometimes abbreviations are used (e.g., can't) Best Regards,

Response 7

Thank you very much for your suggestions. With the incorporation of new tables and paragraphs throughout the paper we have reviewed the display of the paper taking into account your considerations and hope to have provided a much more comprehensive and structured presentation of our findings.