Team LLMKG (SIMP-LLM)

Overall impact score: 3
Overall impact summary: The investigators proposed a cutting-edge machine learning technique to tackle an important problem. However, the investigators' focus is unclear within the proposal: although the significance section is focused on drug repurposing, the innovation section discusses predicting binary drug-disease interactions, and the approach section discusses evaluating the models' performance on standardized medical exams compared to general biomedical large-language models.

Significance score: 1
Significance summary: The need for identifying drugs to tackle rare diseases was articulated well. The fact that 10% of Americans are affected by rare diseases but that fewer than 10% of these diseases have an approved treatment is a clear problem.

Innovation score: 3
Innovation summary: The investigators plan to use a cutting-edge approach: augmenting an existing drug-repurposing knowledge graph, DRKG, with biomedical literature in order to build a graph neural network for drug repurposing. The proposal addresses two weaknesses in knowledge graphs: their inability to be updated with the latest information, as well as their sensitivity to contradictions and uncertainties. The investigators plan to address the former by embedding the knowledge graph entities with a PubMed-based large language model, BioLinkBert; they plan to address the latter using a large language model. This will be used to create a graph neural network that will be used to predict drug-disease interactions.

This plan seems reasonable. However, I am confused by the investigators' end goal: the prediction of a drug-disease interaction. This binary classification may not be very useful for the stated use case of drug repurposing. For drug repurposing, a ranked list will likely be far more useful than an unranked set of candidates, as the candidates will need to be further validated manually. I would recommend that the investigators look into a relevant paper: Hsieh et. al., 2021. Hsieh built a COVID-19 knowledge graph using curated literature, performed transfer learning with DRKG, and built a neural network ranking model to select the most potent drugs.

Approach score: 3
Approach summary: The team's approach for the first two aims is strong, although I am concerned that the root of this proposal, DRKG, was created for the purpose of identifying drug repurposing candidates for COVID-19. In fact, their Github repository states that the data are collected from recent publications related to COVID-19. Will this be an issue, given that this proposal is focused on rare diseases, or will the addition of BioLinkBERT address this?

I am confused about the investigators' end goal. The investigators plan to evaluate the model's performance on a dataset containing medical licensing exam-type questions. However, the model is built with drug repurposing in mind. In addition, comparing its performance to general-purpose large-language biomedical models does not appear to be a fair comparison. I believe that there may be more relevant benchmark models and datasets. Could the model be evaluated against its own sub-components, DRKG and BioLinkBERT, to examine how these methods are complementary? Also, could the model be compared to parallel ones built exclusively to predict drugs for repurposing such as PREDICT (Gottlieb 2011), SAEROF (Jiang 2020), or HINGRL (Zhao 2022)? A relevant dataset for this aim is the comparative toxicogenomics database (CTD), which includes 228,901 curated chemical-disease interactions. This is significantly larger than the list available on the FDA's orphan drug designations and approvals list, which contains 6,479 drugs.

Investigators score: 1
Investigators summary: The investigators are well equipped to tackle this problem. They are all PhD or Masters candidates with the Department of Biomedical Data Science with relevant backgrounds in natural language processing, deep learning, and statistical methods.

Environment score: 1
Environment summary: The investigators are at Stanford with good access to computational resources, as well as experts within the field of drug discovery.