HeCiX: Integrating Knowledge Graphs and Large Language Models for Biomedical Research

Prerana Sanjay Kulkarni[1*], Muskaan Jain[2*], Disha Sheshanarayana[3*], and Srinivasan Parthiban[4]

[1] Dept. of Computer Science and Engineering, PES University, Bengaluru, India
prer.kulk@gmail.com
[2] Dept. of Mathematics, IIT Madras, Chennai, India
muskaan40389@gmail.com
[3] Dept. of Computer Science and Engineering, Manipal University, Jaipur, India
disha.229301161@muj.manipal.edu
[4] Dept. of Data Science and Engineering, Indian Institute of Science Education and Research, Bhopal, India
parthi@iiserb.ac.in

Abstract. Despite advancements in drug development strategies, 90% of clinical trials fail. This suggests overlooked aspects in target validation and drug optimization. In order to address this, we introduce HeCiX-KG, Hetionet-Clinicaltrials neXus Knowledge Graph, a novel fusion of data from ClinicalTrials.gov and Hetionet in a single knowledge graph. HeCiX-KG combines data on previously conducted clinical trials from ClinicalTrials.gov, and domain expertise on diseases and genes from Hetionet. This offers a thorough resource for clinical researchers. Further, we introduce HeCiX, a system that uses LangChain to integrate HeCiX-KG with GPT-4, and increase its usability. HeCiX shows high performance during evaluation against a range of clinically relevant issues, proving this model to be promising for enhancing the effectiveness of clinical research. Thus, this approach provides a more holistic view of clinical trials and existing biological data.

Keywords: Knowledge Graph · Large Language Model · LangChain · Clinical Trials.

1 Introduction

The pharmaceutical industry faces significant challenges in drug discovery, with alarming clinical trial failure rates of almost 90% [1]. The rise in attrition rates reflects not only huge financial losses but also delayed implementation of life-saving treatments for patients.

One of the major reasons underlying this is the fragmented nature of the available data. Hetionet [2] contains vast domain knowledge about diseases, genes,

* These authors contributed equally to this work.

and anatomy, yet it lacks sufficient information about previously conducted clinical trials and experiments. Conversely, ClinicalTrials.gov [3] houses extensive information about clinical trials and experiments conducted worldwide, including details about Principal Investigators of studies which can be useful in identifying Key Opinion Leaders (KOLs). However, it offers limited insights into the diseases themselves. This disparity between our understanding of fundamental biology and clinical trial outcomes hinders effective drug development.

To address this, we propose a novel knowledge graph, HeCiX-KG (Hetionet-Clinicaltrials neXus Knowledge Graph), which integrates information from clinicaltrials.gov [3] and Hetionet [2]. HeCiX-KG is a single knowledge graph, connecting biological knowledge with clinical trial data. This integration can provide better understanding, revealing linkages and patterns that were previously missed, but are vital for the effective repurposing and discovery of new drugs.

Building upon HeCix-KG, we introduce HeCiX, a system that utilizes OpenAI's GPT-4 [4] and LangChain [5] to enable seamless interaction with the knowledge graph. HeCiX translates natural language queries into CQL (Cypher Query Language) queries, which makes it possible to efficiently retrieve relevant context from the knowledge graph. Subsequently, the system displays the result in human-understandable format, rendering the information available to clinical and biomedical researchers.

We evaluate HeCiX's performance against a wide array of question-answering tasks relevant to the domain. The results show notable advancements in the scope and depth of data obtained, thus providing a helpful tool to enhance the efficiency of clinical research, and improve the success rates of drug repurposing and development. HeCiX overcomes significant shortcomings in the current resources by offering a holistic view of disease biology, clinical trial history, and expert knowledge at the user's fingertips.

The structure of the paper is as follows. We describe the background of the work and the knowledge graph construction in Section 2 and Section 3 respectively. Section 4 talks about the detailed methodology. Experimentation is described in Section 5 which is followed by results and discussions in Section 6. Finally, Section 7 discusses the conclusion.

2 Background and Related Work

2.1 Knowledge Graphs in Biomedical Research

Knowledge graphs have emerged as powerful tools for representing and integrating complex biomedical information, facilitating efficient data integration and knowledge discovery. Notable examples include Bio2RDF [6], CTKG by Chen et al. [7], Hetionet [2], and others. They have been used to enhance drug discovery, understand disease mechanisms, and identify biomedical relationships.

2.2 Large Language Models in Healthcare

Large Language Models (LLMs) have significantly improved medical literature analysis, clinical note interpretation, and diagnosis support. Important models

in this field include DeepMind's MedIC, Microsoft's BioGPT [8], TrialGPT [9], BioBart [10], and BioMistral [11], among others. They have impacted domains such as drug discovery, patient data analysis, and clinical decision support.

2.3 Hetionet: A Comprehensive Biomedical Knowledge Graph

Hetionet is a heterogeneous network of biomedical knowledge, integrating data such as genes, compounds, diseases, and their interrelationships [2]. It has been used in predicting drug-target interactions, identifying disease mechanisms, and supporting drug repurposing efforts, making it ideal for integration with other data sources to enhance clinical research capabilities.

2.4 ClinicalTrials.gov: A Repository of Clinical Trial Data

ClinicalTrials.gov is an extensive data source providing information about clinical trials, studies conducted on various diseases, principal investigators, study locations, and trial outcomes [3]. It thereby supports clinical research and drug development.

3 Knowledge Graph Construction

HeCiX-KG is constructed from two primary sources of data, Hetionet [2] and ClinicalTrials.gov [3]. It combines their data into a single knowledge source and includes data related to six specific diseases, namely Vitiligo, Atopic Dermatitis, Alopecia Areata, melanoma, Epilepsy, and Hypothyroidism.

3.1 Hetionet

Hetionet is a highly interconnected knowledge base, which combines data from 29 distinct databases. It comprises a total of 47,031 nodes across 11 types: Disease, Compound, Gene, Symptom, Side Effect, Biological Process, Molecular Function, Anatomy, Cellular Component, Pathway, and Pharmacologic Class [2]. For the purpose of constructing HeCiX-KG, we have extracted a subgraph of Hetionet, consisting of data related to the six chosen diseases. This comprises a total of 1071 nodes and 1125 relations.

3.2 ClinicalTrials.gov

ClinicalTrials.gov provides massive amounts of information about clinical trials and studies on various diseases and conditions [3]. While the total number of records in ClinicalTrials.gov exceeds 500,000, our research focuses on a selected subset of 1,200 records, spanning the six selected diseases. This subset when constructed as a knowledge graph, contains 5,454 nodes and 11,466 edges. The nodes in this subset are classified into 9 types: Disease, Principal Investigators (PI), Study, Conditions, Phases, Locations, Interventions, Age, and Sex. There are 10 types of relationships connecting these nodes.

3.3 Schema

By taking inspiration from the schemas of the knowledge graphs constructed in Hetionet [2] and the work of Devarakonda et al. [12], we have constructed a comprehensive schema for HeCiX-KG. The 'Disease' node serves as the main connecting point between the two integrated databases. This schema has been illustrated in Figure 1. By populating the schema with our data, we have obtained a knowledge graph consisting of 6,509 nodes and 14,377 edges.

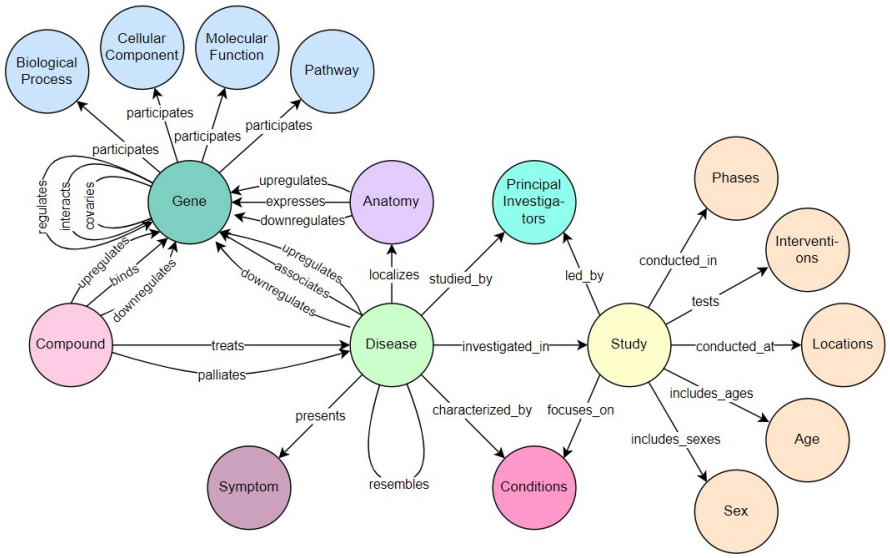

Fig. 1. HECIX-Schema

4 Methodology

Our approach consists of two major stages, construction of HeCiX-KG, and its integration with GPT-4 using LangChain.

4.1 Knowledge Graph Construction

HeCiX-KG is constructed by extracting and integrating relevant data from Hetionet [2] and ClinicalTrials.gov [3] for six specific diseases. The resulting knowledge graph has 6,509 nodes and 14,377 edges. The construction process involves data extraction, schema design, entity-relationship mapping, and graph population.

4.2 LLM Integration using LangChain

To enhance the usability of HeCiX-KG, we developed HeCiX, a system that integrates our knowledge graph with GPT-4 using LangChain. Specifically, we utilized the GraphCypherQAChain component from the LangChain ecosystem for this integration. As indicated in Figure 2, our query processing pipeline is as follows:

1. **User Query Input**: A user submits a natural language prompt to LangChain.
2. **Query and Prompt Processing**: The user's question is combined with a set prompt template, and then sent to GPT-4
3. **Cypher Query Generation**: GPT-4 generates a Cypher query based on the user's input and sends it back to LangChain.
4. **Database Querying**: LangChain executes the generated Cypher query on HeCiX-KG.
5. **Raw Results Retrieval**: HeCiX-KG returns the raw query results (the 'Full Context') to LangChain.
6. **Context Forwarding**: LangChain forwards the full context to GPT-4 for interpretation and conversion into a human-readable format.
7. **Human-Readable Response Generation**: GPT-4 generates a human-readable response based on the full context sent to it, and sends it to LangChain.
8. **User Response**: Finally, LangChain returns the human-readable response to the user, thereby providing the user with the answer to their query.

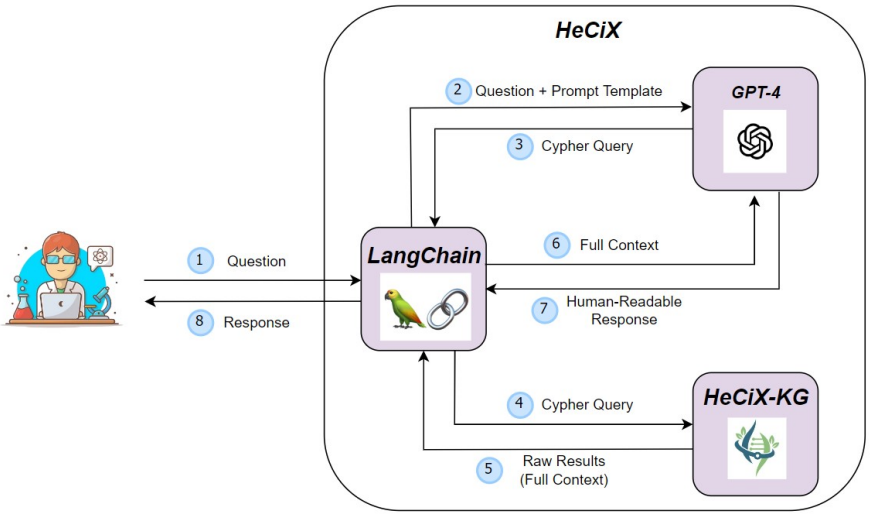

Fig. 2. Query processing pipeline

This integration allows users to interact with the complex HeCiX-KG using natural language queries, significantly enhancing its accessibility and usability for clinical researchers.

5 Experimentation

5.1 Experimental Setup

Our experimental setup consisted of the following major steps:

1. We set up our AuraDB [13] server to host HeCiX-KG.
2. We constructed the schema of our knowledge graph, based on the individual structures of Hetionet and clinical trials data.
3. We populated the schema with data from both Hetionet and ClinicalTrials.gov to create HeCiX-KG.
4. We integrated HeCiX-KG with GPT-4 using LangChain's GraphQueryQAChain component.

5.2 Evaluation Methodology

To properly assess and analyse the question-answering abilities of our system, we designed a set of question-answering tasks for HeCiX to answer. These tasks were carefully crafted to cover various aspects of clinical research, including drug discovery, identification of KOLs, and analysis of biomedical data, among others.

6 Results and Discussion

6.1 Experimentation Results

We utilized the RAGAS [14] framework to evaluate our model's performance. The framework calculates several key metrics: faithfulness, answer relevance, context precision, and context recall. The results of our experimentation have been displayed in Table 1.

Table 1. Performance metrics of HeCiX using the RAGAS framework

Metric	**Score**
Faithfulness	0.8572
Answer Relevance	0.9340
Context Precision	0.9202
Context Recall	0.6654

6.2 Comparative Analysis

In addition to evaluating our model on the evaluation metrics seen earlier, we also carried out a comparative analysis with OpenAI's GPT-4 and Anthropic's Claude 3 Sonnet [15] in a zero-shot scenario. We tested these models on the same set of questions used to evaluate HeCiX. We discovered that while these models excel at answering questions based on domain-specific knowledge, they struggle with queries related to clinical trials.

Our experiments with HeCiX demonstrated promising results in enhancing clinical research capabilities. Figure 3 shows a sample interaction with HeCiX, where HeCiX is answering questions which test both domain expertise and knowledge about clinical trials.

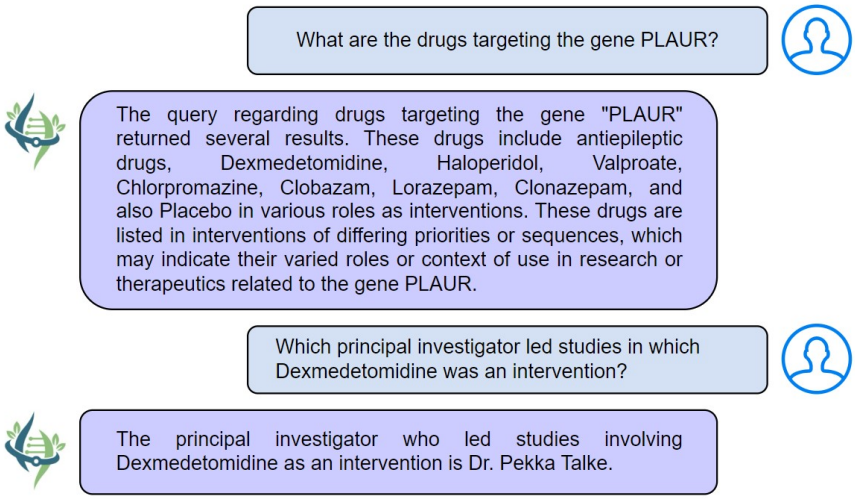

Fig. 3. Sample interaction with HeCiX demonstrating QA capabilities

6.3 Limitations

While HeCiX has shown promising results, it is important to acknowledge its limitations.

- Uncertainty of the model's performance as the knowledge graph expands.
- Additional testing on a wider range of diseases to ensure robustness of the system.

7 Conclusion

This paper introduces HeCiX, an innovative system that connects knowledge graphs from clinical trials data and Hetionet with large language models to

address major challenges in clinical research. Our experiments showcase that HeCiX enhances drug discovery processes and uses existing scattered biomedical data effectively. HeCiX uncovers all possible relationships between diseases, genes, and treatments, potentially accelerating drug development which can lead to unexpected discoveries.

HeCiX is a major advancement in using AI for clinical research. We believe that HeCiX will play a crucial role in shaping the future of biomedical research, providing a connected and innovative ecosystem in the field.

The scope of future enhancements for HeCiX could include adding SNOMED CT to the knowledge graph for better clinical term standardization and employing meta-path matching techniques to identify complex relationships more effectively.

References

1. Sun D, Gao W, Hu H, Zhou S. Why 90% of clinical drug development fails and how to improve it? Acta Pharm Sin B. 2022 Jul;12(7):3049-3062. doi: 10.1016/j.apsb.2022.02.002. Epub 2022 Feb 11. PMID: 35865092; PMCID: PMC9293739.
2. Himmelstein, Daniel Scott, et al. "Systematic integration of biomedical knowledge prioritizes drugs for repurposing." *Elife* **6** (2017): e26726.
3. U.S. National Library of Medicine, "ClinicalTrials.gov," https://www.clinicaltrials.gov/, accessed July 14, 2024.
4. OpenAI, "GPT-4 Model Documentation," https://openai.com/research/gpt-4/, accessed July 14, 2024.
5. Harrison Chase, "LangChain," https://github.com/langchain-ai/langchain, accessed July 14, 2024.
6. Belleau, François, et al. "Bio2RDF: towards a mashup to build bioinformatics knowledge systems." *Journal of biomedical informatics* 41.**5** (2008): 706-716.
7. Chen, Ziqi, et al. "Ctkg: A knowledge graph for clinical trials." *medRxiv* (2021): 2021-11.
8. Luo, Renqian, et al. "BioGPT: generative pre-trained transformer for biomedical text generation and mining." *Briefings in bioinformatics* 23.**6** (2022): bbac409.
9. Jin, Qiao, et al. "Matching patients to clinical trials with large language models." *ArXiv* (2023).
10. Yuan, Hongyi, et al. "BioBART: Pretraining and evaluation of a biomedical generative language model." *arXiv preprint arXiv:2204.03905* (2022).
11. Labrak, Yanis, et al. "Biomistral: A collection of open-source pretrained large language models for medical domains." *arXiv preprint arXiv:2402.10373* (2024).
12. Devarakonda, Murthy V., et al. "Clinical trial recommendations using Semantics-Based inductive inference and knowledge graph embeddings." *Journal of biomedical informatics* 154 (2024): 104627.
13. Neo Technology, "Neo4j AuraDB Documentation," https://neo4j.com/docs/aura/current/, accessed July 14, 2024.
14. Es, Shahul, et al. "Ragas: Automated evaluation of retrieval augmented generation." *arXiv preprint arXiv:2309.15217* (2023).
15. Anthropic, "Claude 3 Sonnet," https://www.anthropic.com/news/claude-3-5-sonnet, accessed July 14, 2024.