Knowledge Graph and Large Language Model for Metabolomics

Yuxing Lu[1,2,3]

[1]Department of Big Data and Biomedical AI, College of Future Technology, Peking University
[2]Wallace H Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University
[3]Tencent AI Lab
yxlu0613@gmail.com

Abstract

The advancements in Knowledge Graphs (KGs) and Large Language Models (LLMs) are driving transformative changes across various research fields, including metabolomics. These tools present exceptional opportunities to elucidate complex metabolic pathways and identify biomarkers essential to biological systems. My research focuses on harnessing the potential of KGs and LLMs within metabolomics, specifically making interactions between them and with biological researches. KGs, with their structured representation of metabolic entities and relationships, provide a robust foundation for managing extensive multimodal metabolomic knowledge. Recently, I developed a metabolite-centric knowledge graph and explored innovative methodologies to leverage KGs and LLMs for enhancing predictive modeling in clinical settings. My future research aims to fully exploit the capabilities of KGs and LLMs in metabolomics, advancing our understanding and applications in this field.

Introduction

Metabolomics is the comprehensive study of small molecules known as metabolites within the human body, stands at the forefront of systems biology (Clish 2015). This field has rapidly evolved, with advancements in analytical techniques and bioinformatics. However, there remains a significant gap in tools capable of integrating, analyzing, and inferring diverse metabolomics knowledge. Knowledge Graphs (KGs) and Large Language Models (LLMs) present promising solutions to address these challenges.

Over the past decade, KGs have emerged as transformative tools for managing and analyzing complex datasets. In the context of metabolomics, KGs can encapsulate heterogeneous data, including metabolic pathways, enzymes, metabolite-disease associations, and clinical observations, thereby providing a holistic view of metabolic interactions and functions.

Additionally, KGs have proven the ability to enhance the generation of LLMs by using Retrieval-Augmented Generation (RAG) techniques (Hu and Lu 2024). This approach enables the models to not only access a broad repository of structured information stored within KGs but also to enhance their responses with up-to-date and relevant data fetched in real-time.

I am currently a third-year PhD student in a joint program among Peking University, Georgia Institute of Technology and Emory University. My past research primarily focused on constructing a metabolite-centric knowledge graph and investigating the interactions between KGs and LLMs. Moving forward, I aim to address the gaps in the application of LLMs within metabolomics and explore innovative research and clinical applications in this field.

Current Works and Contributions

My current research focuses on multimodal integration methods, metabolomic knowledge graph, and the interaction between KGs and LLMs.

1. Multimodal Clinical Data Integration. Medical data is inherently multimodal, and integrating these data types in a rational and efficient manner is crucial. One research project I was involved in predicts a patient's biological age by integrating images of the face, tongue coating, and fundus. The difference between biological age and actual age, termed AgeDiff, serves as a biomarker for predicting various diseases (Wang et al. 2024). Additionally, my past research explored multimodal integration at the genomics level. I utilized DNA methylation, mRNA expression, and miRNA expression data from cancer patients to predict corresponding cancer types and subtypes. In these two studies, I introduced methods that leverage dynamic learning and guided cross-attention to integrate multimodal data (Lu et al. 2023b,c). The results demonstrated that my approach achieved an average AUC of 87.41% across 12 cancer subtype prediction tasks, significantly surpassing the previous state of the art (SOTA) of 84.85%.

2. Metabolite-centric Knowledge Graph. I have constructed a metabolite-centric knowledge graph (MetaKG, [1]) (Lu et al. 2024) using all the metabolomics information from the most commonly used databases like HMDB (Wishart et al. 2022). Through knowledge graph representation learning and multimodal metabolite representation integration with a triple contrastive learning module, my method demonstrates the capability to reduce batch effects in metabolomics data and achieves state-of-the-art results in

[1]http://www.metakg.xyz

12 metabolite attribute prediction tasks, as well as in a real-world Non-Alcoholic Fatty Liver Disease (NAFLD) clinical diagnosis task. To the best of my knowledge, MetaKG is the first metabolomics knowledge graph to date, and this work has been accepted by IJCAI 2024.

3. KG-enhanced LLM Application. Leveraging external knowledge to enhance the generation or prediction results of large language models (LLMs) has always been a focus of my interest. In real-world applications, external knowledge often exists in various heterogeneous forms. Therefore, I proposed a method that enhances large language models by utilizing a variety of heterogeneous knowledge, both explicitly and implicitly (Lu et al. 2023a; Lu, Zhao, and Wang 2023). Specifically, I convert data from various sources (such as knowledge graphs, databases, and search engine results) into a unified natural language format and process this external knowledge using prompt learning techniques to design templates or employ LLM agents. This approach guides the model to make better predictions and generate more accurate outputs. I tested my method on an electronic health record (EHR) diagnostic classification dataset, achieving the best results in both disease diagnosis and cross-departmental transfer tasks.

Future Directions

In the rest time of my PhD study, I plan to focus on enhancing the integration of Knowledge Graphs (KGs) and Large Language Models (LLMs) within metabolomics, which includes:

1. Continuous refinement of MetaKG. My primary objective is to enhance the current MetaKG by integrating additional metabolomic databases and leveraging the latest advancements in knowledge graph algorithms. This will involve systematic updates and expansions of the datasets within MetaKG to ensure its comprehensiveness and relevance. I will investigate the application of cutting-edge technologies, such as graph neural networks and contrastive learning, to improve MetaKG's data analysis and pattern recognition capabilities.

2. Expansion of MetaKG applications. MetaKG can be utilized to facilitate a series of metabolomics research. This includes utilizing MetaKG embeddings for metabolite property classification and correcting inaccuracies in existing database records. Additionally, MetaKG can be employed to generate high-quality scientific hypotheses, which will guide subsequent experimental designs and research directions.

3. Development of a domain LLM for metabolomics. Given the extensive textual content available in existing metabolic databases and PubMed, there is a significant opportunity to train a domain-specific language model tailored for metabolomics. This specialized LLM would be proficient in understanding and generating text relevant to metabolites, thereby assisting researchers in efficiently navigating and analyzing vast amounts of scientific literature, ultimately enhancing research productivity.

4. Exploration of synergies between MetaKG and LLMs. Metabolomics is a knowledge-intensive discipline. By leveraging prompt learning and Retrieval-Augmented Generation (RAG) techniques, I can optimize the utilization of knowledge from MetaKG to enhance the accuracy and relevance of generated text (Lu, Zhao, and Wang 2024). This approach aims to improve the applicability of the generated content, significantly advancing the role of LLMs in the field by automating the development of experimental protocols and predicting outcomes, thereby providing innovative methodologies and insights in metabolic research.

Acknowledgments

I would like thank my supervisor, Jinzhuo Wang for his continued support and guidance. Additionally, I am thankful to all my co-authors for their support and contributions.

References

Clish, C. B. 2015. Metabolomics: an emerging but powerful tool for precision medicine. *Molecular Case Studies*, 1(1): a000588.

Hu, Y.; and Lu, Y. 2024. Rag and rau: A survey on retrieval-augmented language model in natural language processing. *arXiv preprint arXiv:2404.19543*.

Lu, Y.; Liu, X.; Du, Z.; Gao, Y.; and Wang, G. 2023a. Medkpl: a heterogeneous knowledge enhanced prompt learning framework for transferable diagnosis. *Journal of Biomedical Informatics*, 143: 104417.

Lu, Y.; Peng, R.; Dong, L.; Xia, K.; Wu, R.; Xu, S.; and Wang, J. 2023b. Multiomics dynamic learning enables personalized diagnosis and prognosis for pancancer and cancer subtypes. *Briefings in Bioinformatics*, 24(6): bbad378.

Lu, Y.; Peng, R.; Wang, J.; and Jiang, B. 2023c. MoTIF: a Method for Trustworthy Dynamic Multimodal Learning on Omics. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2851–2858. IEEE.

Lu, Y.; Zhao, W.; Sun, N.; and Wang, J. 2024. Enhancing Multimodal Knowledge Graph Representation Learning through Triple Contrastive Learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 5963–5971.

Lu, Y.; Zhao, X.; and Wang, J. 2023. Medical knowledge-enhanced prompt learning for diagnosis classification from clinical text. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Lu, Y.; Zhao, X.; and Wang, J. 2024. ClinicalRAG: Enhancing Clinical Decision Support through Heterogeneous Knowledge Retrieval. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, 64–68.

Wang, J.; Gao, Y.; Wang, F.; Zeng, S.; Li, J.; Miao, H.; Wang, T.; Zeng, J.; Baptista-Hon, D.; Monteiro, O.; et al. 2024. Accurate estimation of biological age and its application in disease prediction using a multimodal image Transformer system. *Proceedings of the National Academy of Sciences*, 121(3): e2308812120.

Wishart, D. S.; Guo, A.; Oler, E.; Wang, F.; Anjum, A.; Peters, H.; Dizon, R.; Sayeeda, Z.; Tian, S.; Lee, B. L.; et al. 2022. HMDB 5.0: the human metabolome database for 2022. *Nucleic acids research*, 50(D1): D622–D631.