# Enhancing Aging Biomarker Research through Large Language Models and Knowledge Graphs (Student Abstract)

**Srikar Reddy Gadusu[1], Yiğit Küçük[1], Vania Santillana[1], Aaron King[2] and Hande Küçük McGinty[1]***

[1]Koncordant Lab, Department of Computer Science, Kansas State University, Manhattan, KS 66506
[2]Aeon Biomarkers LLC, Kihei, Hawaii, United States
srikarre@ksu.edu, yigitkucuk92@gmail.com, aaron.k.biology@gmail.com, hande@ksu.edu

## Abstract

Aging biomarkers play a crucial role in uncovering the biological mechanisms behind aging and in developing strategies to support healthy aging. However, the search for reliable aging biomarkers is particularly challenging due to the intricate and multifactorial nature of the aging process. Furthermore, biomarker names and categories are not well-standardized in the current literature. While, a formal definition of a biomarker is nonexistent in the current literature, formally defining biomarkers and standardizing the vocabulary for biomarkers can help accelerate AI research around this concept which can lead to better, faster and more accurate analyses of the existing data and literature. Thus, in this work, we generated Knowledge Graphs that can help us define and standardize biomarkers. We present our Knowledge Graphs (KGs) generated using both an LLM and expert-curated datasets. We compare both KGs to understand why systematic integration between these two models is needed. The integration of Knowledge Graphs (KGs) and Large Language Models (LLMs) presents a promising approach to advancing aging biomarker research through the inherent structured and standardized nature of ontology schemas in knowledge graphs. We showcase that the accuracy of LLM-generated KGs remains questionable but systematic methods such as KNARM can help us with the accuracy of these efforts. In future work, we will propose a synergistic framework where KGs and LLMs interact iteratively to improve both the comprehensiveness and accuracy of aging biomarker information.

## Introduction

In the age of artificial intelligence and data science, ontologies play a key role in organizing and standardizing data, making it easier to analyze and identify patterns. Ontologies not only support advanced data management but also enhance machine learning and AI techniques by enabling machines to interpret and reason with structured information (Hitzler et al. 2012). Additionally, ontologies and knowledge graphs are critical for the semantic web, allowing intelligent searches and improving how information is retrieved and understood. As data becomes more complex, the importance of ontologies will only grow, helping machines process and understand data more effectively, as well as connecting disconnected pieces of data such as diseases, biomarkers, and drug targets (McGinty 2018).

However, manually creating ontologies comes with challenges. The process of defining concepts, relationships, and properties can be slow, prone to errors, and difficult to standardize. Furthermore, transferring knowledge between domain experts and computer scientists during the development process can lead to misunderstandings and inconsistencies. This creates an additional layer of complexity, particularly as data continues to grow. While previous methods like Knowledge Acquisition and Representation Methodology (KNARM) (McGinty 2018) have addressed these challenges, newer approaches (Toro et al. 2024) now incorporate Large Language Models (LLMs) to enhance ontology building (Caufield et al. 2024).

Recognizing both the challenges and opportunities in this area, our study uses our previously established Ontology Learning with Integrated Vector Embeddings (OLIVE)(Zhang Y., Dalal A. S., Martin C., Gadusu S.R., McGinty H.K. 2024) Methodology, an evolved version of KNARM that integrates LLMs into the semi-automated process of building ontologies. This new approach builds on previous research and addresses key concerns in knowledge acquisition and ontology validation, bringing LLMs into the process to assist with automation.

## Methodology

In this paper, we aim to showcase the differences in knowledge graphs when they are built using our KNARM(McGinty 2018) and OLIVE (Zhang Y., Dalal A. S., Martin C., Gadusu S.R., McGinty H.K. 2024) methodologies. KNARM is a methodology that produces modular, semi-automated ontologies and knowledge graphs while our OLIVE methodology enhances this approach using Large Language Models (LLMs).

### Expert-Curated Knowledge Graph Using KNARM

In this approach, we collaborated with **Aeon Biomarkers**, who provided us with biomarker data derived based on data from the NHANES (National Health and Nutrition Examination Survey) datasets. This dataset includes biomarker values for patients of various ages and both sexes. To analyze this data, we followed these steps:

*Corresponding Author.

- **Data Processing:** Using Python and the pandas library, we processed the NHANES dataset to calculate key statistical values (such as the mean and median) for each biomarker across different age groups. This provided a clearer understanding of how biomarker levels vary with age.
- **Biomarker Filtering:** We implemented a filtering strategy to isolate and analyze specific biomarkers, with **Body Mass Index (BMI)** serving as a primary stratification variable. This approach enabled the examination of potential associations between BMI and other biomarker levels. Through this methodology, we aimed to elucidate whether BMI or other biomarkers significantly influence the levels of specific biomarkers under investigation.
- **Significance Testing:** Due to the large number of biomarkers and potential comparisons, manually visualizing each possible graph to identify relationships was inefficient. To address this challenge, we developed an automated approach to detect significant changes in biomarker values. This method involved analyzing values using their statistical significance to assess whether variations in one biomarker were associated with substantial changes in others. By employing this automated analysis, we effectively narrowed down the biomarkers that demonstrated meaningful interactions, thereby streamlining the identification of relevant biomarker associations. One should note that, this automated methods sometimes resulted in false positives, which was corrected by our domain experts at Aeon Biomarkers.We also utilized ML methods to explore data rapidly to identify potential new connections among biomarkers. We used **Correlation Matrix Calculation and Correlation Threshold Filtering** to reveal simple correlations among biomarkers and understand highcorrelation pairs using different cut-off thresholds.
- **Knowledge Graph Creation:** Once we identified significant relationships between biomarkers, we used the Neo4j database backend as indicated in the KNARM methodology to store these connections. **Neo4j** allowed us to represent these relationships visually and structurally as a knowledge graph, making it easier to understand the interactions between different biomarkers. Using a database backend also allowed us to employ our semi-automated ontology building approach when generating our knowledge graph.

## LLM-Based Methodology (OLIVE)

The second part of this study, we focused on using our LLM-powered workflow, OLIVE, to enhance and automate the process of knowledge graph creation. This workflow allowed us to integrate knowledge from existing research in the literature and helped us create knowledge graphs that can utilize the data extracted from the abstracts of the existing literature. The steps we followed in this methodology are: **Abstract Retrieval:** OLIVE uses LLMs to search scientific literature for abstracts related to specific biomarker keywords. Based on the biomarkers vocabulary we derived from the biomarkers present in the NHANES dataset, we

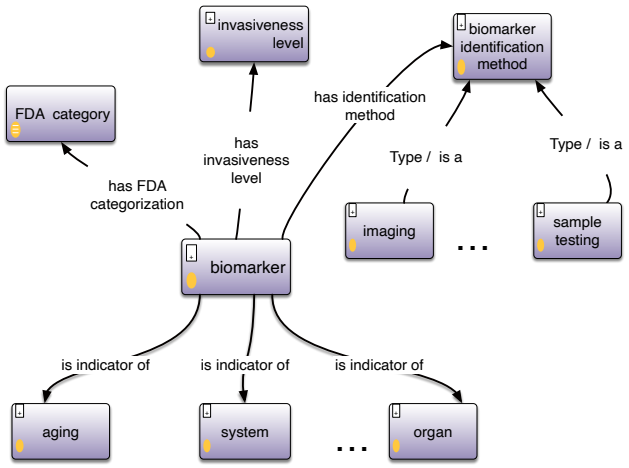


Figure 1: As part of our study we generated an expert curated Biomarkers of Aging Ontology Schema with the different aspects of biomarkers found in the NHANES dataset.

extracted existing knowledge from the literature and used it as the data source for our automated knowledge graph creation. and **Text Processing and Knowledge Graph Generation:** The LLM analyzes the abstracts to extract relevant information and automatically converts it into a knowledge graph based on this extraction and the generated knowledge graph can be visualized using a simple user interface we prepared that allows visualization using WebVOWL(Lohmann et al. 2015). In this way, we are able to generate a knowledge graph that captures the relationships between various biomarkers based on the information extracted from the literature, complementing the expert-curated knowledge graph we developed manually.

## Conclusions

In this paper, we describe our efforts towards building an Aging Biomarkers Ontology (ABO)( McGinty H.K., Gadusu S.R., Küçük, Y., King, A. 2024) and we compare Knowledge Graphs (KGs) generated through LLMs with those created by experts, demonstrating the need for systematic integration between these models. This standardization effort and formal definitions for aging biomarkers are a novel approach. Our exploration using LLMs to enhance our KG resulted in a knowledge graph that needed further attention which is addressed using our OLIVE workflow. The difference in the results mainly stems from the amount of expert curation applied on the knowledge graph building which helps with the accuracy and the quality of the modeling in the knowledge graph. However, while expert-curated KGs provide reliable and structured information, LLMs excel in processing vast amounts of unstructured data from scientific literature. Hence, using systematic methodologies like OLIVE may become a crucial part of the process when we are building such systems for Trustworthy and Explainable AI applications that are built to explore biomarkers.

## References


McGinty H.K., Gadusu S.R., Küçük, Y., King, A. 2024. Aging Biomarkers Ontology (ABO). Forthcoming.

Caufield, J. H.; Hegde, H.; Emonet, V.; Harris, N. L.; Joachimiak, M. P.; Matentzoglu, N.; Kim, H.; Moxon, S.; Reese, J. T.; Haendel, M. A.; et al. 2024. Structured prompt interrogation and recursive extraction of semantics (SPIRES): A method for populating knowledge bases using zero-shot learning. *Bioinformatics*, 40(3): btae104.

Hitzler, P.; Krötzsch, M.; Parsia, B.; Patel-Schneider, P. F.; and Rudolph, S., eds. 2012. *OWL 2 Web Ontology Language: Primer (Second Edition)*. W3C Recommendation 11 December 2012. Available from http://www.w3.org/TR/owl2-primer/.

Lohmann, S.; Link, V.; Marbach, E.; and Negru, S. 2015. WebVOWL: Web-based visualization of ontologies. In *Knowledge Engineering and Knowledge Management: EKAW 2014 Satellite Events, VISUAL, EKM1, and ARCOE-Logic, Linköping, Sweden, November 24-28, 2014. Revised Selected Papers. 19*, 154–158. Springer.

McGinty, H. K. 2018. *KNowledge Acquisition and Representation Methodology (KNARM) and Its Applications*. Ph.D. diss., Dept. of Computer Science, University of Miami.

Toro, S.; Anagnostopoulos, A. V.; Bello, S. M.; Blumberg, K.; Cameron, R.; Carmody, L.; Diehl, A. D.; Dooley, D. M.; Duncan, W. D.; Fey, P.; et al. 2024. Dynamic retrieval augmented generation of ontologies using artificial intelligence (DRAGON-AI). *Journal of Biomedical Semantics*, 15(1): 19.

Zhang Y., Dalal A. S., Martin C., Gadusu S.R., McGinty H.K. 2024. OLIVE: Ontology Learning with Integrated Vector Embeddings. Forthcoming.