Longevi graph: knowledge graphs with statistical approaches for aging

by

Srikar Reddy Gadusu

B.Tech., Jawaharlal Nehru Technological University Hyderabad, 2020

A THESIS

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Computer Science
Carl. R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2025

Approved by:

Major Professor
Dr.Hande Küçük McGinty

Copyright

© Srikar Reddy Gadusu 2025.

Abstract

Aging is a complex biological process influenced by various biomarkers, such as cholesterol and blood sugar levels, which serve as measurable indicators of health and disease. However, the vast amount of biomarker data presents challenges in identifying meaningful relationships. To address this, we developed Aging Biomarkers Ontology (ABO), a structured network that formally defines biomarkers from an aging perspective, organizes and connects different biomarkers, making it easier to analyze their relationships. We first applied statistical methods to detect significant differences between biomarkers across various groups. Then, we built a biomarker knowledge graph using Neo4j, a graph database that helps to represent and explore these complex relationships. To further explore aging biomarkers and provide a more complete knowledge graph, we used two approaches to uncover hidden connections: Depth-Limited Search (DLS) and Machine Learning-Based Embedding Search. The DLS approach searches the knowledge graph by traversing the nodes connected to a given node within a specified depth, whereas the embeddings-based method converts biomarker relationships into numerical representations and applies cosine similarity to predict potential associations. To compare the effectiveness of these two approaches, we analyzed how well DLS-based reasoning and embeddings-based search detected known and new relationships. By combining statistical analysis, graph-based knowledge graph completion approaches, and machine learning, this study provides a structured way to explore patterns between biomarkers. By integrating statistical analysis, knowledge graph completion approaches, and machine learning, this study provides a structured framework for aging biomarker analysis, improving data exploration, and aid biomedical research in longevity and aging.

Table of Contents

List of Figures vi

List of Tables vii

List of Nomenclature viii

Acknowledgements ix

1 Introduction 1

2 Literature Review 4

2.1 Biomarker Based Analyses in Large Scale Studies 5

2.2 Statistical Methods for Identifying Significant Differences 9

2.3 Knowledge Graphs in Biomedical Informatics 11

2.4 Algorithmic Exploration of Knowledge Graphs 13

2.5 Graph Embeddings 15

3 Methodology 20

3.1 Data Collection and Preparation 21

3.2 Statistical Analysis 23

3.2.1 Significant Differences 24

3.2.2 Automation of Analysis 27

3.3 Knowledge Graph Construction 28

3.3.1 Design and Structure of the Knowledge Graph 28

3.3.2 Database Creation 30

3.3.3 Graph-Based Representation of Biomarkers 32
3.3.4 Applications of the Knowledge Graph 32
3.4 Algorithmic Exploration of the Knowledge Graph 33
3.4.1 Graph Construction 33
3.4.2 Testing Reasoning Capabilities with a Rule-Based Graph 34
3.4.3 Depth-Limited Search Algorithm 35
3.4.4 Calculating Weights Between Relationships 37
3.4.5 Weighted Depth-Limited Search 40
3.5 Graph-Based Analysis Using Embeddings 42
3.5.1 Building the Biomarker Knowledge Graph 42
3.5.2 Building the Adjacency Matrix 43
3.5.3 Reducing Complexity with Singular Value Decomposition (SVD) . . . 44
3.5.4 Using Cosine Similarity to Predict New Associations 45
3.5.5 Validating Predictions 46

4 Conclusion and Future Work 47

Bibliography 53

List of Figures

2.1 The steps of KNowledge Acquisition and Representation Methodology 13

2.2 Graph Embeddings 15

3.1 Methodology 21

3.2 Resulted combined data of the biomarkers datasets 22

3.3 Streamlit application to analayze the data 23

3.4 DXAVisceralFatMass plot for the filtered BMI range of 0 to 25 24

3.5 DXAVisceralFatMass plot for the filtered BMI range of 26 to 100 25

3.6 Application to Find the Significant Differences for the selected biomarkers and filter biomarker ranges 28

3.7 Biomarker model 30

3.8 Biomarkers Categorization in Neo4j 33

3.9 Correlation Matrix 40

3.10 Neo4j graph with Weighted Relations 41

3.11 Result of Graph Embeddings 44

3.12 Cosine Similarity 45

4.1 Biomarkers which shows Significant Differences for BMI 49

4.2 Results obtained when traversing through BMI 50

List of Tables

3.1 Adjacency matrix representation of a simple graph 43

4.1 SVD-Predicted Associations for Body Mass Index 51

Nomenclature

- **ABO** – Aging Biomarkers Ontology
- **BMI** – Body Mass Index
- **CRP** – C-Reactive Protein
- **DLS** – Depth-Limited Search
- **FDA** – Food and Drug Administration
- **IQR** – Interquartile Range
- **KG** – Knowledge Graph
- **KNARM** – Knowledge Acquisition and Representation Methodology
- **NHANES** – National Health and Nutrition Examination Survey
- **OBO** – Open Biomedical Ontologies
- **OWL** – Web Ontology Language
- **PFA** – Perfluoroalkyl substances
- **SVD** – Singular Value Decomposition

Acknowledgments

First, I express my deepest gratitude to my advisor, Dr. Hande Küçük McGinty, for her exceptional mentorship, support and belief in my research abilities. Working under her guidance as a Research Assistant has been a transformative experience that significantly improved both my academic and professional development. Her thoughtful feedback, insightful perspectives, and encouragement were helpful throughout the course of this thesis. The courses I took with her "Concordant Systems" and "Food Informatics" have provided me with the foundation in Knowledge Graphs and interdisciplinary collaboration.

I am also sincerely grateful to my committee members, Dr. Mitchell Neilsen and Dr. Romulo Lollato, for their valuable time, insights, and support. I extend my thanks to our collaborator and domain expert, Aaron King, whose valuable input enriched this research.

A special thank you to my dear friends Sindhu Palla, Aryan Dalal, Janvi Aggarwal, Nirmal Gelal, Pavan Kumar Pativada and Atalay Ileri. Your friendship and unwavering support kept me grounded throughout this journey. Your encouragement, humor, and belief in me were vital sources of motivation during challenging times.

Most importantly, I extend my heartfelt gratitude to my mother and my sister whose love, patience, and unwavering support have been my greatest strength.

To all who have been a part of this incredible voyage, thank you.

Chapter 1

Introduction

A biomarker[1] is defined as a specific feature that researchers measure to show normal biological processes, harmful processes, or the response of the body to various factors. These factors can include environmental exposures or medical treatments. Biomarkers are different from clinical outcome assessments (COAs), which measure how a patient feels, functions, or survives in daily life.

Several key categories of biomarkers exist[1]. Diagnostic biomarkers help detect or confirm if a disease is present. They can also identify different types (subtypes) of a disease. Monitoring biomarkers are used to make repeated measurements over time so that we can see how a disease changes or how well a treatment is working. Safety biomarkers show whether someone is likely to experience toxic or harmful effects from a treatment or if they are already experiencing toxicity. Susceptibility or risk biomarkers tell us if a person is likely to develop a disease in the future. Prognostic biomarkers help predict how a disease might progress in people who already have a condition, including the chances of future clinical problems or events. It is important to note that a biomarker can belong to more than one category, but evidence must be collected to prove its usefulness in each category.

Biomarkers are important tools in both research and healthcare. They help us measure and track normal biological processes or detect diseases in their early stages. The use of biomarkers as early warning systems[2] in the evaluation of disease risk has increased markedly

in the last decade. Biomarkers are indicators of typical biological processes, pathogenic processes, or pharmacological reactions to therapy. The application and identification of biomarkers in the medical and clinical fields have a huge impact on society. In addition, biomarkers can be targeted to improve diagnosis, prognosis, and therapeutics. The identification of ideal biomarkers holds the key potential for personalized medicine and overall improved clinical outcomes. This is further supported by the fact that protein biomarkers can be used as indicators of the evolution of inflammation, immunity, and stress or related diseases, such as cancer, diabetes, cardiovascular, neurological, and other syndromes.

However, one major problem is that these biomarkers are not well organized or classified. This means that even though there is a lot of information about biomarkers in the literature, it is still difficult to organize it in a clear way and understand the full picture.[3]. Different fields of study may focus on different types of biomarkers without a common system or language to link them. As a result, researchers and healthcare professionals may miss connections that could be very important for understanding diseases, evaluating treatments, or preventing health problems before they begin.

This problem becomes even clearer in the area of aging research. Aging is a very complex process that occurs at many levels, from our genes and cells to our entire bodies. Scientists use different "aging clocks", such as epigenetic clocks[4] (focusing on DNA methylation), transcriptomic clocks[5] (focusing on RNA), and proteomic clocks[6] (focusing on proteins) to measure the biological age of a person and predict how their body might change over time. These clocks are helpful, but they typically look at only one type of data and can miss important information.

At the same time, research[4] shows that different parts of the body can age at different speeds and that even individual cells in the same tissue can vary a lot. For example, as we get older, our blood cell composition changes, our immune system changes, and certain cells can gain a competitive advantage. Due to these differences, relying on only one type of measurement may not capture the whole picture of aging. To truly understand why some people age faster or develop diseases more quickly, we need to study many types of data together and pay attention to the wide variation among cells.

For example, molecular data from our genes or proteins[7] (often called OMICS data) might show that someone is at risk for certain age-related conditions, such as heart disease or dementia. Meanwhile, functional biomarkers, such as walking speed or memory tests, could confirm whether this risk is present in real world abilities. If we do not have a clear system to combine OMICS data with functional biomarker data, we lose the opportunity to see the full picture of how a person is aging and what might be done to slow or improve that process.

One major reason for this gap is that there is no clear and organized way to classify and connect different biomarkers in a computer-operable format. Without a consistent system to name and categorize them, machine learning struggles to find patterns or connections. This brings us to an important point, If we could organize all the data in a way that is easier for computers to process, it might help machine learning identify useful patterns and relationships more effectively.

In this research, we aim to showcase that by building a more structured and systematic approach to organizing biomarker data, we can reveal hidden connections or new biomarkers that would otherwise remain unnoticed. Therefore, we present our novel approach to classifying and formalizing biomarker data which aims to enhance the discovery of these markers and allow the accelerated and accurate discovery of their relationships.

With the research question and hypothesis mentioned above, we present here our proposed solution which focuses on three main steps: developing a comprehensive biomarker ontology (a structured, computer readable classification system), using statistical methods to filter out noise and highlight the most meaningful biomarkers, and building a knowledge graph to visualize and explore how these biomarkers connect to one another. By unifying different kinds of data in this way, we aim to speed up discovery, improve understanding of these biomarkers.

Chapter 2

Literature Review

Biomarkers are characteristics objectively measured[8] that reflect biological processes or responses to treatments. They are becoming increasingly important in drug discovery and personalized medicine, but only a few have been confirmed to reliably predict how patients will respond. Proper study design and rigorous statistical methods are crucial, especially because small studies often exaggerate the results. Multiple large studies are usually needed for reliable validation, and issues such as confounding factors and multiple statistical tests must be carefully addressed. Biomarkers can be grouped into categories such as prognostic, predictive, pharmacodynamic, and surrogate endpoints. Prognostic biomarkers indicate likely outcomes regardless of treatment, while predictive biomarkers help identify which treatment will work best. Both standard randomized trials and biomarker adaptive designs are useful for studying these markers, but they are prone to the same types of bias found in therapeutic research. Strong statistical approaches are essential, and confirmation across several studies is typically required to ensure that a biomarker is valid.

In recent years, large collections of biomedical data have greatly increased our understanding of human health and disease. One major resource is the National Health and Nutrition Examination Survey (NHANES)[9], which has become essential to study population health and chemical exposures in the United States. Through its extensive collection of biological samples and health data from thousands of volunteers, NHANES gives researchers

valuable ways to explore how different biomarkers are related to human health, identify risk factors, and uncover potential treatments. During recent years[9], more and more researchers have used these data for exposure assessments and health association studies, investigating how various biomarkers relate to health issues such as heart disease, obesity, diabetes, and immune dysfunction. This wealth of information helps us better understand population-level health risks, supports the discovery of promising treatment strategies, and advances our general knowledge of how the environment affects human health.

By combining thorough data preparation, statistical analysis, structured knowledge representation, and specialized algorithms, researchers aim to gain deeper insights into how these biomarkers interact[10]. Data preparation is a critical stage that typically takes around 80 percent of total data engineering effort[11]. This is because real-world data is often incomplete, noisy, or inconsistent, and high-performance analytical systems need clean, high-quality data to function effectively. During data preparation, missing values are handled, errors are removed, conflicts are resolved, and the data are reduced to only the most relevant information. This process prevents hidden patterns from going unnoticed, ensures better system performance, and leads to more reliable output.

This chapter reviews the research supporting each step of the method described above, placing it in the wider field of biomedical informatics and data science.

2.1 Biomarker Based Analyses in Large Scale Studies

Large data sets now allow us to study health differences and clinical outcomes on a broad scale. Previously, one of the biggest challenges in health research was having too few participants, which limited the reliability of the findings. Now, thanks to large-scale surveys such as NHANES, researchers can explore various definitions of normal health with more precision. For example, the 2013–2014 NHANES survey found that 62 percent of participants had no diseases, 16 percent considered their health excellent, and 35 percent were defined by specific age categories[12]. In addition to NHANES, other large data sets, including electronic health records and insurance claims, help determine healthy reference ranges and link them

to clinical outcomes. As more countries develop national research biobanks, these growing data sources become essential to define what is considered healthy in different populations. They also allow for more rigorous testing of whether increasing precision in defining health actually leads to better healthcare.

Long-term large cohort studies are particularly valuable because they provide insight into how lifestyle, environment, and genetics together influence serious diseases[13]. One clear example is the UK Biobank, which recruited 500,000 participants and collected diverse information, including surveys, physical measurements, laboratory test results, and health outcomes. This massive scale ensures that even relatively rare diseases, those affecting a few thousand people, are captured in sufficient numbers to uncover strong connections between risk factors and health outcomes. By merging participant data with national health databases over time, researchers can follow the development of various health conditions. This approach gives a clearer picture of disease progression and the factors that could contribute to it, helping scientists and healthcare professionals make more accurate predictions and targeted recommendations.

Biomarkers play a critical role in health research by providing measurable indicators of normal body processes, diseases, and the effectiveness of treatments[14]. These markers are often used to study conditions related to heart health, infections, immune function, and metabolism. However, it is important to remember that biomarker levels can be influenced by individual characteristics such as diet habits. For instance, dietary fat can affect how well certain fat-soluble vitamins are measured. Researchers must also think about the stability of biomarkers, because some, like certain vitamins, are light-sensitive and can degrade over time if not stored or handled properly.

Careful preparation of complex biomedical data is essential before starting any in-depth analysis[15]. The first step is to identify and record missing data without trying to guess or remove those values, while also documenting why they are missing. This transparency helps maintain data integrity. Next, using a tidy data structure, where each variable is placed in its own column and each observation appears in a separate row, makes analysis and sharing simpler. Any data preprocessing steps should be explained with a clear pseudocode,

detailing every transformation applied to the raw data. In addition, noting specific software versions, system configurations, and providing assurances that the same process produces consistent results ensures reproducibility. A comprehensive code book that outlines the variables, measurement units, and any summaries or transformations is invaluable, especially for collaborations or future re-analyses. By following these practices, researchers reduce errors and strengthen the validity of their findings.

Together, these paragraphs highlight the powerful role that large data sets and well-managed studies play in understanding health outcomes. They highlight how studies such as NHANES and the UK Biobank have successfully gathered extensive data to clarify the influence of genetics, lifestyle, and the environment on disease. These resources also demonstrate the central importance of biomarkers in interpreting various health conditions and their responses to treatment. Finally, they emphasize the need for meticulous data management, from identifying missing values to using tidy structures, to maintain data quality and ensure reproducible results. Recognizing and applying these insights in our research underscores why the work is timely and critical. By adopting these approaches, we can make a more robust contribution to our understanding of disease processes, allow more nuanced interpretations of biomarker data, and ultimately help improve healthcare outcomes.

Although large-scale studies help to understand disease processes, translating these insights into clinical practice remains a complex challenge. Currently, it takes approximately 17 years to translate new biomarker knowledge into clinical practice, highlighting the need for more efficient systems. Recent research demonstrates that composite scores that use multiple layers of information provide deeper insight into the mechanism of aging. The paper[16] shows how computational analysis, including LASSO regression and cluster algorithms, has already successfully revealed new relationships between biomarkers and aging profiles. Furthermore, the complexity of biomarker validation requires addressing accuracy, specificity, and reproducibility, which could be significantly improved through better data organization and computational approaches.

Current research[17] addresses the growing use of the term "biological age" in scientific discussions and calls for a more precise language when describing aging biomarkers. The au-

thors propose three main recommendations. First, aging clocks should be named according to the type of data they use, such as epigenetic age for clocks based on epigenetic markers. Second, the authors argue that aging cannot be represented by a single metric because it occurs in multiple dimensions and differs between tissues. Third, they suggest that aging biomarkers are referred to as 'indicators of biological age' rather than direct measurements of biological age. According to the authors, most scientists (86 percent) agree that aging cannot be quantified by one measurement alone, and this consensus emphasizes the need for careful terminology. By distinguishing validated aging biomarkers from the complex process of aging of the entire body, researchers can better assess how different factors influence health and longevity. The paper also highlights the variety of aging clocks developed using biological markers such as proteins, DNA methylation, microRNA, metabolites, and microbiota. Despite these advancements, no single method can fully capture the multifaceted nature of biological aging. The authors illustrate this using grip strength: although it can predict mortality and age-related outcomes, it does not ensure overall biological youth. An individual may have strong grip strength, but still participate in unhealthy behaviors such as smoking or drinking, underscoring the importance of multiple biomarkers. This perspective is supported by a survey of longevity researchers, who generally agree that no single biomarker can fully measure biological age.

Building on these insights, reorganizing biomarker data in a more computer-operable format could enhance the speed and accuracy of discovering biomarkers and understanding their interrelationships. This approach involves creating a systematic way of classifying and formalizing biomarker data, potentially through the use of ontologies, which can help standardize terminology, improve data sharing, and enable complex analyses between different tissues and biological processes. By embracing such a structured framework, scientists can more effectively reveal patterns in aging processes, test the multidimensional aspects of biological age, and ultimately refine both clinical and research applications.

Researchers measure aging[18] using certain markers, especially in blood. Although these markers are promising for showing how our bodies age, there are still a few hurdles before they can be used in clinics. One major problem is that studies often get different results,

in part because researchers use different methods and analyze data in different ways. The authors say that we need standard rules and better methods that work for different groups of people. They also suggested more teamwork, clear reporting guidelines, and a thorough testing of various markers in multiple large groups. The paper[18] stresses the importance of making all computer code and data available so that others can repeat and confirm the findings.

2.2 Statistical Methods for Identifying Significant Differences

For this research, we performed statistical analysis to identify patterns and significant differences between biomarkers. For this we have looked into the different review of the methods and selected the ones which are appropriate for our work flow from the feedback given by the domain expert.

The Jacobson-Truax(JT)[19] method is the most commonly used way to measure meaningful clinical improvement, appearing in 35 percent of studies on clinical significance. It defines improvement as a change from a dysfunctional score before treatment to a functional score after treatment, along with a statistically reliable change. The JT method is popular because it is easy to calculate, has established cut-off points for many commonly used tests, and provides a balanced approach compared to more extreme methods. It closely matches the findings of the Gulliksen-Lord-Novick(GLN) method, both reporting similar moderate rates of clinically significant change (18.9 percent). However, for the JT method to work correctly, data from the patient and non-patient data should come from comparable groups. The GLN (Gulliksen-Lord-Novick) method was introduced by Hsu (1989) as an alternative to Jacobson's approach. Its main idea is to handle the regression to the mean, the tendency for very high or very low scores to move closer to average over time. The GLN method uses a hypothetical average (mean) and standard deviation of a suitable comparison group to adjust for this effect. Studies show that GLN agrees with the JT method more than 90

percent of the time, and both methods find a moderate level of clinically significant change (18.9 percent). The Expected Norms (EN) method gives the highest estimates of clinically significant change. It decides if a person has improved by looking at their post-treatment score compared to a confidence range around their estimated pre-treatment true score. In studies, it labels the fewest people as unchanged (42 percent) and the most as improved (46 percent). The High Accuracy (HA) method, on the other hand, is the strictest. It rarely classifies people as improved, but is more likely to say that they have worsened. It also marks the highest number of clients as unchanged (67 percent). Lastly, the Hierarchical Linear Modeling (HLM) method uses multiple data points for each person and calculates empirical Bayes estimates. It is the only method that does not depend on reliability scores, but is often impractical in regular clinical settings because it is complex and requires frequent testing.

A clustered Mann-Whitney U-test[20] can analyze clustered or grouped data, such as paired eyes in a clinical trial comparing nutritional supplements for retinitis pigmentosa. It addresses the correlation between measurements within the same cluster while preserving the non-parametric benefits for data that do not follow a normal distribution. This approach often provides greater statistical power than a standard Mann-Whitney U-test, especially when the intraclass correlation is low. However, there are some important limitations. The test is not valid for unbalanced designs, and it assumes measurements taken at different time points are exchangeable - a condition that may not hold in longitudinal studies. In highly skewed distributions, outliers can distort the average values within a cluster. Despite these drawbacks, the method is particularly useful for continuous data or categorical data with many levels.

Researchers frequently use correlation coefficients, such as Pearson's correlation, to measure how strongly two continuous variables move together[21]. Although Pearson's correlation only captures straight-line (linear) relationships, it is widely used because it is straightforward to understand and easy to calculate. Scientists often pick a cutoff value to focus on only those correlations that are strong and meaningful, filtering out weaker, less reliable ones[22]. Additionally, methods to correct for multiple comparisons, like controlling the false

discovery rate, are commonly employed to reduce the chance of mistakenly identifying links that are not really there.

Once the data are properly prepared, they can be used to find links and noticeable differences between groups of individuals, defined by their traits, environments, or medical conditions[23]. Basic statistics such as medians and interquartile ranges are helpful in showing typical values and the spread of these values, especially since biological data often do not follow a normal distribution[24]. These measures also reduce the impact of very unusual values (outliers), which are common in health data.

2.3 Knowledge Graphs in Biomedical Informatics

Knowledge graphs organize information[25] so that different systems can work together, reason about data, and answer questions. They include entities (things), relations (connections between these things), and attributes (additional details). They can be created from many sources, such as databases, text, or images, using methods such as manual annotation, rule-based extraction, or machine learning.

Knowledge graphs (KGs) are particularly valuable in public health research, allowing scientists to structure and formalize extensive information from diverse sources for better analysis and decision making[26]. In KGs, important elements like biomarkers, diseases, genes, and metabolites are shown as 'nodes', while their connections can be shown as 'edges'. This approach makes it possible to combine different types of data into one unified structure that can be easily searched. Due to this, KGs have been successfully used in areas such as finding new uses for existing drugs[27], describing diseases in detail, and discovering new biomarkers[28].

Knowledge Acquisition and Representation Methodology (KNARM)[29] is a structured way to build knowledge graphs by creating detailed and consistent ontologies. Ontologies define the main ideas and relationships in a specific area, acting as the framework for a knowledge graph. KNARM has nine steps: sub-language analysis, unstructured interview, sub-language recycling, creating metadata and modeling knowledge, structured interview, validation of knowledge acquisition, building the database, semi-automated ontology cre-

ation, and ontology validation. Each step offers guidelines for involving domain experts and knowledge engineers, as well as tools for simplifying ontology development. KNARM has been applied to several biomedical ontologies, such as the Drug Target Ontology (DTO), which focuses on drug targets; the LINCS Information Framework (LIFE) ontology, which examines different types of cellular responses; and the BioAssay Ontology (BAO) 2.0, which describes screening assays and their data. This method not only explains how to build knowledge graphs collaboratively and systematically but also how to maintain them by updating and validating with new data and expert feedback. Moreover, KNARM shows how knowledge graphs can be utilized for data integration, analysis, visualization, and discovery, and the methodology can be adapted to other fields where complex, ever-changing data sources must be integrated and analyzed.

The Open Biomedical Ontologies (OBO) Foundry[30] is an initiative to organize and standardize biomedical ontologies for more effective data integration. It includes several mature ontologies, such as Cell Ontology, Gene Ontology, and the Foundational Model of Anatomy. However, based on existing tables and descriptions, there does not appear to be a specific biomarker ontology among them. The Foundry focuses on ensuring that ontologies are open, well-structured, and do not overlap. Participating ontologies must follow certain principles, including collaborative development, using shared relationships, and clearly defining their scope. We decided to create our own biomarker ontology using the KNARM methodology.

Graph structures are an intuitive way to represent data[31] because all information about an entity can be stored in a single node, with links showing the relationships. This model is especially useful when the connections and layout of the data are just as important as the data itself. It has been applied in various specialized areas, such as spatial data, web systems, and document management. Graph-oriented operations allow data to be manipulated directly in this form, and integrity constraints can be set to maintain consistency (5). Furthermore, graph views can be customized for different management and querying needs without affecting the underlying implementation.

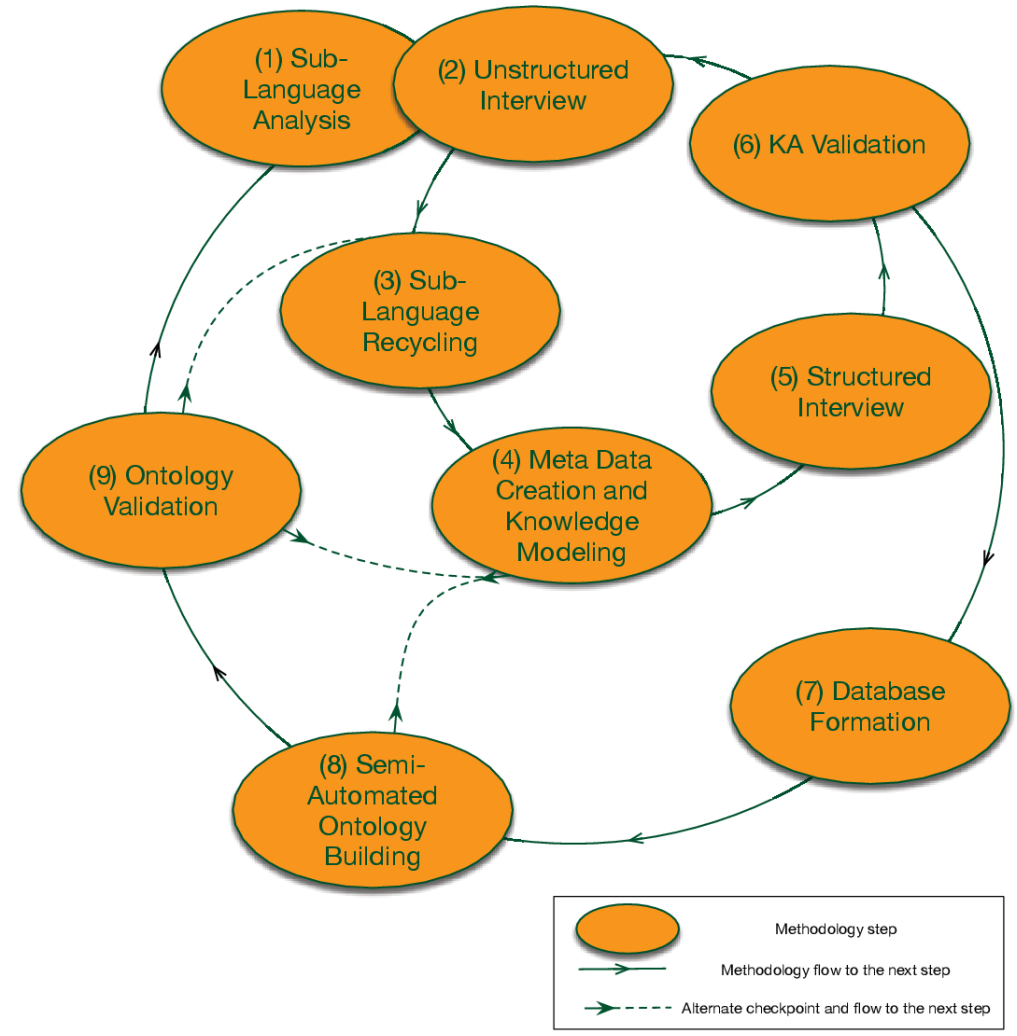

Figure 2.1: The steps of KNowledge Acquisition and Representation Methodology

2.4 Algorithmic Exploration of Knowledge Graphs

Drug2ways[32] is a new computational method that looks at causal paths in biological networks to find potential drug candidates. It uses an efficient algorithm to handle networks containing drug-protein, protein-protein, and protein-disease connections. The system can identify single-target drugs, multi-target drugs, and combination therapies. Tests using clinical trial data showed that it could reliably find known drug-disease pairs. However, Drug2ways depends on well-structured biological networks that already have clear causal relationships, which makes it better suited for drug discovery rather than for building and studying knowledge graphs.

BFS and DFS algorithms are effectively used to analyze protein-protein interactions (PPI) within biological networks[33]. They are applied to protein knowledge graphs, where they help navigate complex relationships between proteins and their functional annotations. These graphs capture intricate interactions involving the roles of proteins in various biological processes and signal transduction pathways. Both the BFS and DFS methods showed strong performance on the SHS27K and SHS148K datasets, highlighting their value in the study of protein-protein interactions. By revealing complex biological contexts, these algorithms play a key role in protein structure prediction and drug development.

However, both BFS and DFS have limitations. BFS can become inefficient if the graph is very large, because it explores all possible neighbors at each level. DFS can get lost exploring deep paths, particularly if the graph is highly connected, and it might not quickly find all the relevant nodes you care about. This is where depth-limited search (DLS) can be helpful, because it prevents the algorithm from going too deep. DLS sets a specific depth beyond which the search will not continue, making it easier to control the scope of exploration and focus on the depth of interest.

Beyond just searching, KGs can be used to infer new relationships by applying specific rules that reflect expert knowledge of the field. Combined with searching, these rules can lead to new discoveries almost automatically[34]. If the connections (edges) in the graph have weights reflecting their importance, probabilistic approaches can highlight the strongest or most likely paths[35]. In biomedical contexts, these methods can uncover hidden links between biomarkers, potentially revealing indirect associations that could shed light on how diseases work or suggest possible points for intervention.

A probabilistic approach to knowledge graphs has several important benefits[36]. First, it helps measure how certain or uncertain we are about the predictions being made, which is key to understanding how reliable the connections in the graph are. Instead of treating connections as simply true or false, this method uses probabilities, calculated through logistic regression, to represent the likelihood of a connection.

This approach is especially useful for exploring hidden patterns in the data and adding new information to the graph more effectively. It also allows us to make better decisions

about which connections to focus on, by considering both what we already know (exploitation) and what we still need to discover (exploration). Overall, this makes building and expanding knowledge graphs smarter and more flexible compared to traditional methods that treat all data as definite.

2.5 Graph Embeddings

Graph embeddings refer to methods that transform nodes, edges, or entire subgraphs into a continuous vector space while preserving the structural properties of the original graph[37]. These embeddings aim to create low-dimensional representations that maintain important graph relationships, making them useful for machine learning applications such as link prediction, community detection, and node classification.

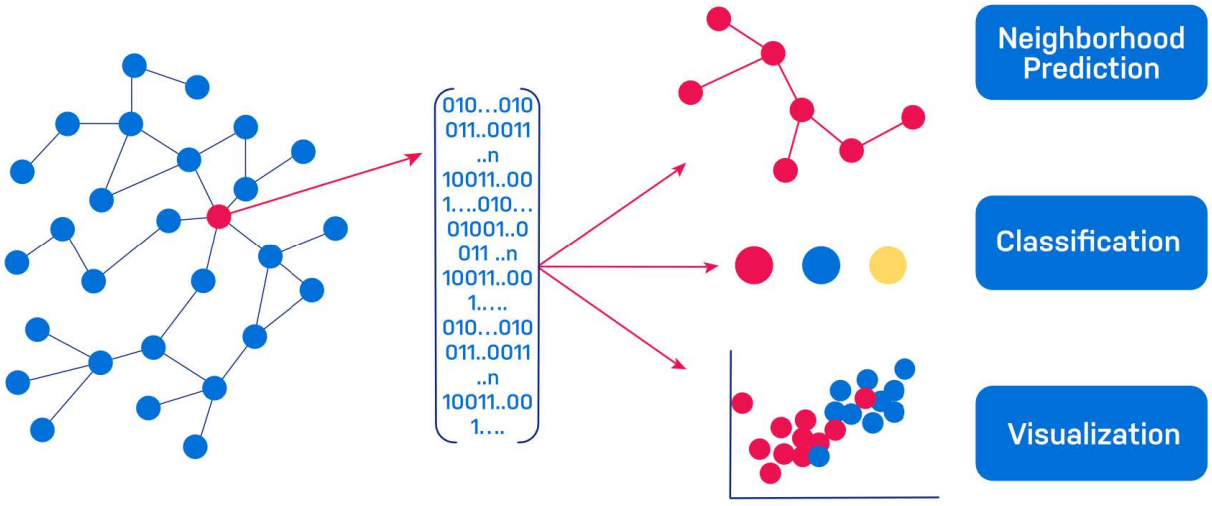

Figure 2.2: Graph Embeddings

Singular Value Decomposition (SVD) is a matrix factorization technique widely used in data compression, dimensionality reduction, and feature extraction. In graph embedding, SVD transforms the adjacency matrix into a lower-dimensional representation while preserving key structural information. This approach has been extensively applied in various domains, including biomedical networks, where it helps uncover hidden patterns in biomarker interactions and other biological relationships[37].

SVD has been used in multiple graph embedding techniques because of its ability to efficiently extract latent features from network structures. Traditional matrix factorization methods, such as Laplacian Eigenmaps and Graph Factorization, often rely on SVD to reduce high-dimensional representations while maintaining important connectivity patterns. Compared to more recent approaches, such as DeepWalk and Node2Vec, SVD-based embeddings offer a more direct and computationally efficient way to capture node relationships without requiring extensive random walk sampling[37]. Studies have shown that SVD-based embeddings can achieve competitive or even superior performance in link prediction tasks compared to deep learning-based methods[37].

A key advantage of SVD in graph embedding is its ability to reduce the dimensionality of large adjacency matrices while preserving essential structural information. By projecting high-dimensional data into a lower-dimensional space, SVD facilitates visualization and downstream machine learning tasks. Various studies have demonstrated that embedding dimensions can be adjusted based on specific use cases, with lower dimensions improving interpretability and higher dimensions enhancing predictive performance in link prediction and classification tasks[37]. The ability to capture important patterns while filtering out noise makes SVD an effective tool for embedding biomedical networks.

In biomedical graph analysis, link prediction plays a crucial role in identifying potential relationships between biomolecules, such as drug-disease associations and protein-protein interactions. SVD-based embeddings have been found to improve link prediction performance by capturing latent features that may not be immediately observable in the raw adjacency matrix. Since SVD effectively groups nodes with similar connectivity patterns, it enhances the ability to predict missing or unknown links in biological networks. Research has indicated that SVD-based methods outperform traditional similarity-based techniques in biomedical link prediction tasks, highlighting their utility in uncovering novel biomarker relationships[37].

SVD is often preferred over other graph embedding methods due to its computational efficiency. Unlike approaches such as Node2Vec and DeepWalk, which rely on random walks and stochastic optimization techniques, SVD provides a direct matrix factorization, reducing the need for hyperparameter tuning and extensive sampling. This efficiency makes it partic-

ularly suitable for smaller to medium-sized graphs, where computational resources may be limited. Additionally, studies have shown that SVD-based embeddings require less time for training while maintaining competitive performance in predictive tasks[38].

Truncated SVD is a specialized variant of SVD that focuses on retaining the most significant singular values while discarding lower-rank components. This method is particularly effective for graph embedding as it allows for efficient compression of adjacency matrices without significant loss of structural information. Compared to other dimensionality reduction techniques, Truncated SVD performs well in applications where adjacency matrices are sparse, such as in biomedical graphs. Additionally, its ability to preserve global structural properties while reducing computational overhead makes it a suitable choice for link prediction and clustering tasks in biological networks[37].

Many biological associations exhibit linear dependencies, making SVD an effective technique for graph embedding in biomedical applications. Unlike deep learning-based methods that are optimized for capturing complex nonlinear interactions, SVD is well-suited for scenarios where relationships follow a more structured pattern. Prior research has demonstrated that SVD-based embeddings perform well in biomedical network analysis, particularly when the primary goal is to identify associations based on well-defined mathematical relationships rather than complex hierarchical structures. The ability to efficiently capture linear dependencies further reinforces the relevance of SVD in biomarker discovery and disease association studies[37].

This literature review highlights the critical components that form the basis for our research and how they connect to the methodology presented in the next chapter. By exploring a wide range of studies and techniques, we identified key challenges and opportunities in biomarker research, statistical analysis, and knowledge graph construction.

The first component focuses on biomarker-based analyses in large-scale studies. Resources such as NHANES and UK Biobank have shown the value of extensive datasets in uncovering patterns in health and disease. These studies highlight how factors such as genetics, lifestyle, and environment interact to influence outcomes, emphasizing the importance of biomarkers in understanding and predicting health conditions. However, a recurring challenge in these

studies is the lack of a unified framework to organize and analyze biomarkers effectively.

Next, we review statistical methods for identifying meaningful differences and relationships between biomarkers. Techniques such as the Jacobson-Truax and Mann-Whitney U-tests have been commonly applied in clinical and biological research, but their effectiveness depends on well-prepared datasets. These methods, while powerful, often struggle when the data is incomplete or irregular, which is a frequent issue in real-world biomarker datasets.

We then examine the role of knowledge graphs in biomedical informatics. Knowledge graphs provide an intuitive way to organize and link diverse types of data, allowing researchers to visualize relationships and uncover hidden patterns. Methods such as the Knowledge Acquisition and Representation Methodology (KNARM) have demonstrated how structured ontologies can enhance data integration and enable reasoning in complex domains. Despite these advancements, there is a gap in the application of such frameworks specifically to biomarkers.

Finally, we explore the use of algorithmic approaches for graph-based reasoning and discovery. Depth-limited search has shown promise in navigating large and complex knowledge graphs, allowing for efficient identification of meaningful connections. These techniques also help balance the trade-off between exploration (discovering new insights) and exploitation (focusing on known relationships), making them particularly valuable for biomarker research. After that, we explored the applications of embeddings on Knowledge Graphs. Graph embeddings have been widely used in knowledge graphs to enhance representation learning, making it easier to perform reasoning tasks, infer missing links.

Together, these components lay the foundation for a methodology that integrates best practices from multiple disciplines. In the next chapter, we describe how our research brings these elements together through a systematic approach. First, we collect and prepare the data to address issues such as missing values and irregular distributions. Next, we apply statistical tools to identify significant relationships between biomarkers, guided by domain expertise. Using these insights, we constructed a detailed and flexible knowledge graph to represent biomarkers and their connections. Finally, we used algorithmic approaches to enable us to uncover hidden patterns and prioritize meaningful relationships.

By combining these components into a cohesive framework, our research aims to overcome the existing challenges in biomarker analysis. The methodology presented in the next chapter builds on the lessons from this review, demonstrating how these methods can work together to advance biomarker research.

Chapter 3

Methodology

In this chapter, we outline the methods used to explore and analyze the biomarker data. The methodology was carefully designed to combine statistical analysis, knowledge graph construction, and algorithmic approaches to uncover meaningful patterns and relationships in the data. The process involved the following broad steps, which are detailed in the sections of this chapter:

- **Data Collection and Preparation:** We started by collecting biomarker data and preparing them for analysis. This included cleaning the data, organizing them into meaningful categories, and handling missing values to ensure that the dataset was ready for further exploration.
- **Statistical Analysis:** Using statistical methods, we analyzed the data to identify significant relationships between biomarkers based on domain expert feedback and guidance. Pearson's correlation coefficient was used to quantify the strength of these relationships, helping us uncover patterns that could guide subsequent steps.
- **Knowledge Graph Construction:** A knowledge graph was built to represent the biomarkers and their relationships. The biomarkers were classified according to the expert's input and their relationships were weighted by the strength of their correlations, creating a powerful tool for reasoning and exploration.

- **Algorithmic Exploration of the Knowledge Graph:** To further analyze the graph, we use a depth-limited search (DLS) algorithm. This allowed us to identify direct and inferred relationships between biomarkers, prioritizing stronger connections while maintaining computational efficiency.

- **Graph-Based Analysis Using Embeddings:** To make biomarker analysis more efficient, we used embeddings to group related biomarkers based on the relations.

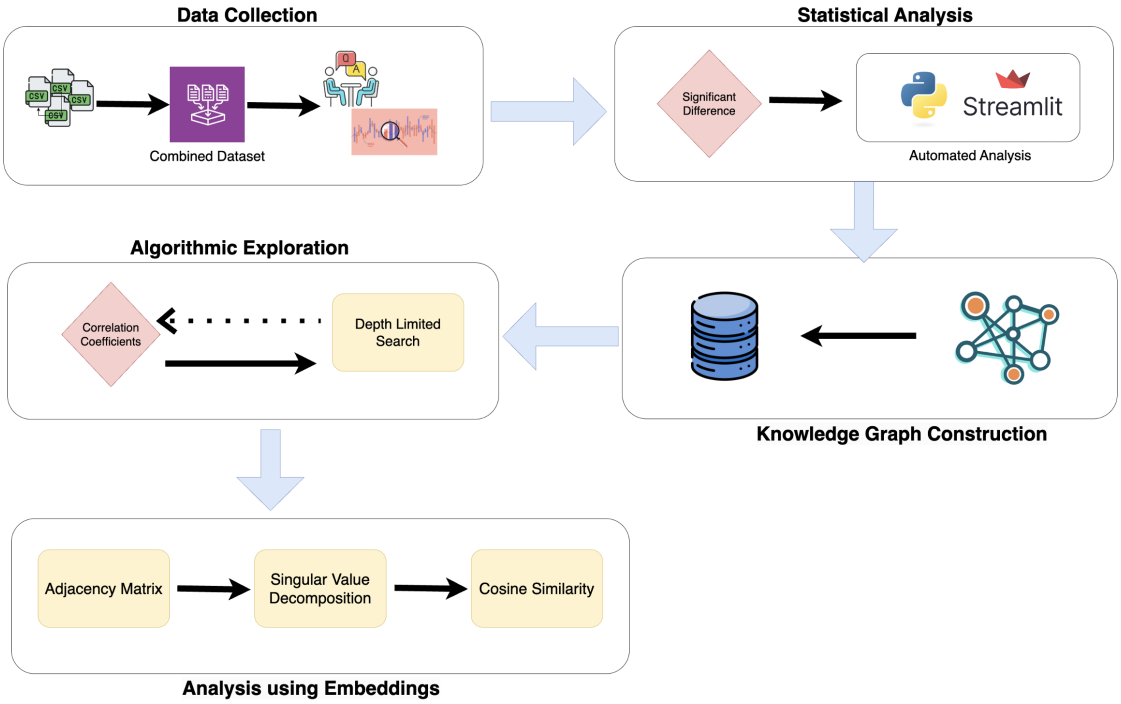

Figure 3.1: Methodology

3.1 Data Collection and Preparation

The first step was to organize the NHANES dataset, which includes biomarker data collected from a large number of individuals. Each file in the original dataset contained values for a specific biomarker, with rows corresponding to individual patients. Each patient was identified by a unique `patient_id`, along with their age and the associated biomarker value. All

biomarker measurements were collected in the same laboratory, which helps ensure consistent and comparable results.

SEQN	Age	acarotene	acratio	acryptoxanthin	adlg	albumin	alkalinephosphatase	alpha1acid	alt	amh	apob	ast	b2m	basophilper
9972	44	0.49				4.3	68.0		65.0			49.0	2.64	0.6
9978	16	0.49				4.4	112.0		16.0			16.0	1.52	0.8
9999	28	0.49				4.3	67.0		20.0			22.0	1.65	1.7
10013	58	0.49				4.2	62.0		58.0			104.0	1.77	0.1
10021	44	0.49				4.0	136.0		31.0			25.0	1.8	0.3
10023	13	0.49			10.737	4.4	309.0		17.0	5.121		26.0	1.89	0.2
10030	14	0.49				4.1	132.0		12.0			15.0	1.44	0.2
10031	24	0.49				3.7	57.0		18.0			21.0	1.21	0.7
10051	74	0.49				4.0	89.0		13.0			28.0	1.98	0.3
10090	57	0.49				3.9	59.0		15.0			19.0	2.49	0.4
10110	17	0.49				4.1	118.0		21.0			22.0	1.98	1.1
10121	39	0.49				4.7	66.0		41.0			21.0	1.84	0.4
10125	19	0.49				4.1	52.0		10.0			16.0	1.55	1.2
10168	70	0.49				4.0	86.0		11.0			49.0	2.19	1.2
10177	13	0.49				4.4	150.0		11.0			21.0	1.71	0.1
10202	26	0.49				3.6	89.0		12.0			15.0	1.76	0.8
10208	12	0.49			5.869	4.1	317.0		36.0			31.0	2.52	0.2
10226	20	0.49				4.8	81.0		39.0			30.0	2.07	0.2
10230	17	0.49			3.298	4.7	100.0		17.0	5.752		32.0	1.8	0.2
10239	41	0.49				3.8	48.0		26.0			33.0	1.37	0.8

Figure 3.2: Resulted combined data of the biomarkers datasets

To prepare the data for analysis, we merged all biomarker files into a single dataset using `patient_id` as the key. The resulting dataset includes 113,438 patients and 184 biomarkers. Patient ages mostly range from 0 to 120 years. However, a small subset (273 entries) contained implausible age values above 120, likely due to data entry errors or placeholder values. These entries were excluded from age-based analysis. Additionally, none of the patient records contain complete data for all biomarkers, which highlights the challenge of working with missing values in large-scale biomedical datasets. The dataset was sorted by age to support the analysis of biomarker trends across different age groups.

We created an application to visualize the data that made it easy to recognize the patterns in the data. In that application, users were able to interact with the dataset by selecting specific biomarkers of interest for further analysis. In addition, we introduced a "filter biomarker" as a criterion to divide individuals into two distinct groups. For example, if BMI was chosen as the filter biomarker, Group 1 included individuals with Body Mass

Index (BMI) levels within a specified range (range 1). Group 2 consisted of individuals with BMI levels within another predefined range (range 2). This grouping allowed us to perform comparative analyses between the two sets of individuals. By this application, we can compare the graph patterns of the filtered biomarker dataset which we want to compare. It was easier to find the initial differences between the biomarkers selected.

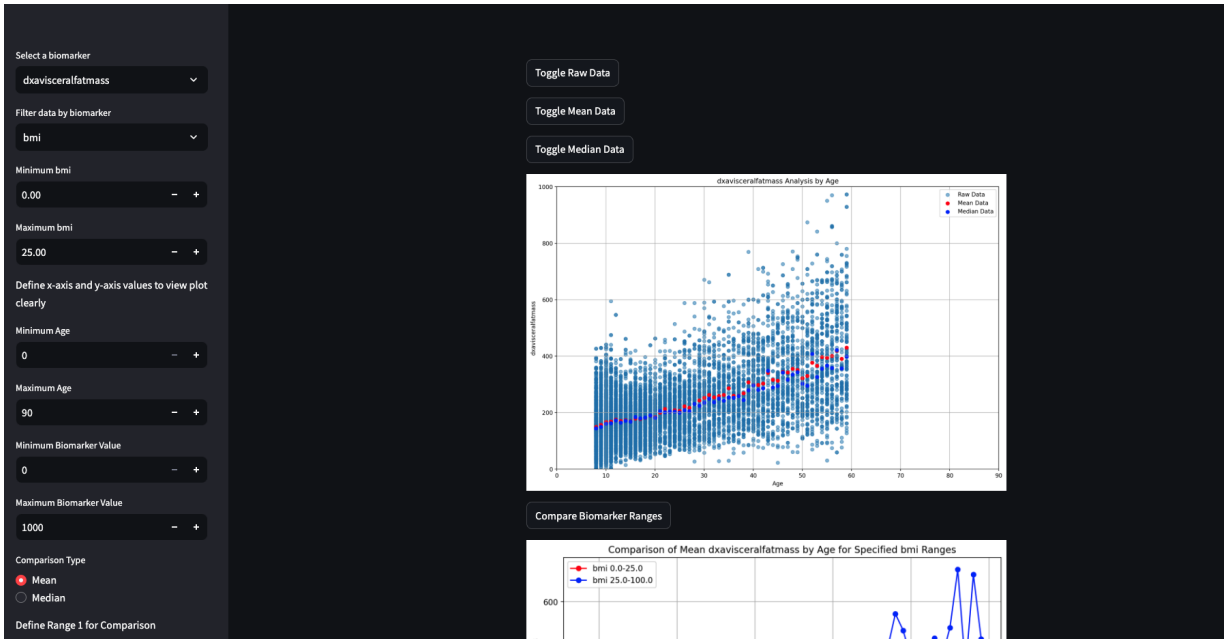

Figure 3.3: Streamlit application to analayze the data

3.2 Statistical Analysis

After visualizing various biomarker patterns in the application, we focused on comparing certain biomarkers such as, "dxasubcutfatmass" against a filtered "BMI" dataset. We chose these specific biomarkers for our initial analysis based on guidance from our domain expert. By manually reviewing the plots, we noticed that some biomarkers appear to be strongly dependent on each other.

To further analyze these relationships, we developed a statistical approach that looks at how the values of one biomarker change when another biomarker's values shift.

3.2.1 Significant Differences

Based on the feedback of the domain expert on the filtered dataset, we define two biomarkers as having a significant difference if the first biomarker's plot (range 1) shows a significant increase or decrease compared to the second biomarker's plot (range 2).

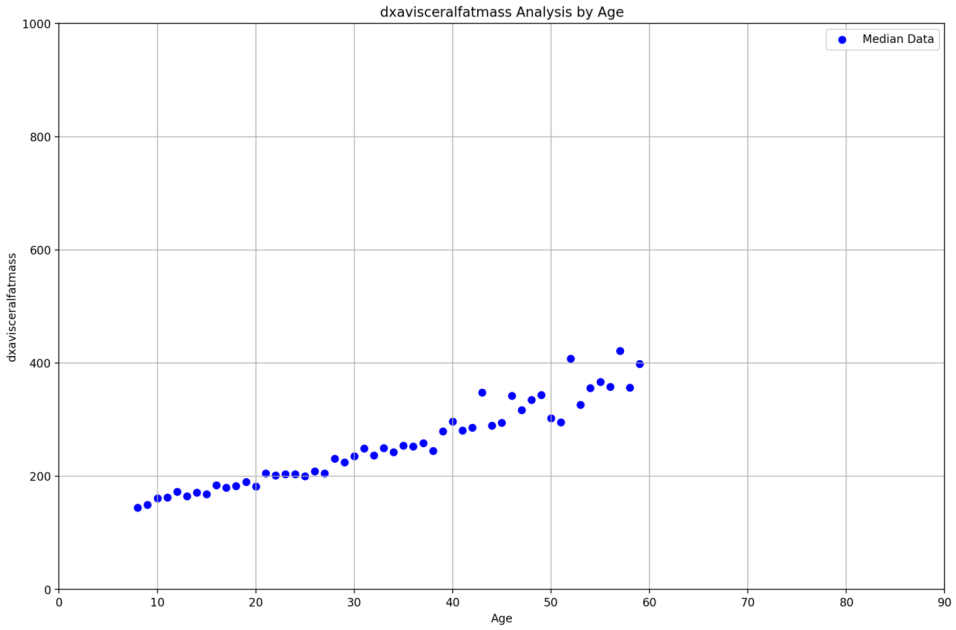

Figure 3.4: DXAVisceralFatMass plot for the filtered BMI range of 0 to 25

We want to compare different filtered biomarker datasets in a way that mathematically identifies when the values of a particular biomarker differ significantly from the filtered range, based on the plotted data. We performed statistical tests, including the T-test and the Mann-Whitney U test, to identify significant differences between groups.

A t-test is used to compare the average values of two groups to see if they are significantly different from each other. It works by calculating a t-statistic, which measures the difference between the group means relative to the variation in the data. The result is compared to a critical value or turned into a p-value, which tells us the likelihood of seeing the observed difference if there is no real difference (the null hypothesis). If the p-value is small (for example, less than 0.05), suggests that the groups are significantly different.

Whereas, Mann-Whitney U test is used to compare two groups to determine if their values are significantly different, without assuming that the data follow a normal distribution. It

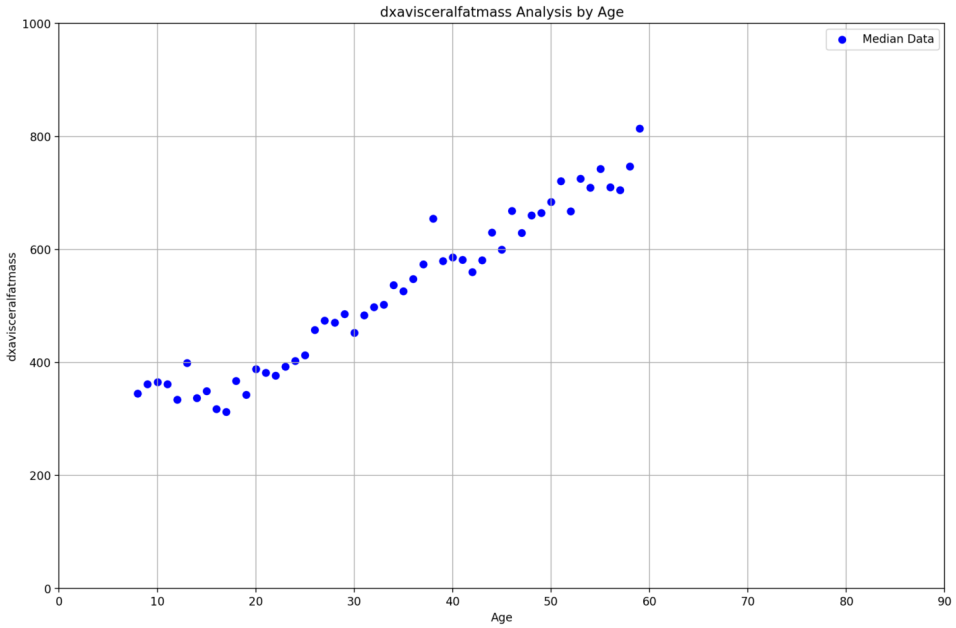

Figure 3.5: DXAVisceralFatMass plot for the filtered BMI range of 26 to 100

works by ranking all the values from both groups together and then comparing the sum of the ranks between the groups. The test calculates a U-statistic, which is used to determine a p-value, showing whether the difference between the groups is statistically significant.

However, both tests produced results that appeared unreliable, with several false positives observed in the significance levels. This was likely due to the high proportion of missing values in the dataset and the lack of normal distribution. Filtering the data without properly handling the missing values may have introduced bias and led to unrepresentative groupings, ultimately affecting the validity of the statistical outcomes. Additionally, the Mann-Whitney U test, though more flexible, assumes that the samples accurately reflect the bigger groups they belong to. If the group sizes were too small, uneven or had very different distributions, the results could also be unreliable.

Hence, we want to come up with a different approach to find these significant differences mathematically. As previously discussed, two biomarkers are considered to have a significant difference if the plot of the first biomarker (range 1) shows a notable increase or decrease compared to the plot of the second biomarker (range 2). For example, as shown in the figures 3.4 and 3.5, the two plots demonstrate significant differences when filtered by BMI (range

1: 0–25, range 2: 26–100).

We calculated the Interquartile Range (IQR) for each biomarker, which represents the range between the 25th and 75th percentiles of the data. This helped us assess how the middle 50% of values varied between groups. Specifically, we observed that in several cases, the IQR of one group did not fall within the IQR of the other indicating a notable shift in distribution. Compared to traditional methods like the T-test and Mann-Whitney U test, this approach proved to be more conservative. The statistical tests often reported a large number of significant differences. However, upon visual inspection of the plots, many of those biomarkers did not appear meaningfully different between the two groups. This suggested that the tests have produced false positives likely due to the presence of missing data. By contrast, the IQR-based method aligned more closely with the visual trends, helping us filter out potentially misleading results and focus on more differences. This mathematical observation provided a clear indication of the significant differences between the two groups. In contrast, when the IQR of one group was within the IQR of the other, no significant differences were identified.

To further evaluate central tendencies, we calculated the median values for each biomarker within both groups. The median, being less sensitive to outliers, provided a robust measure of central tendency. A key part of our comparison was determining whether the median of one group fell within the IQR of the other group. If this condition was met, we concluded there was no significant difference between the groups, based on domain expert interpretations.

Although this approach effectively identified patterns, it occasionally produced false positives. To address this, we introduced a threshold value to define the minimum difference required for classification as significantly different. This threshold was determined in collaboration with domain experts to ensure that the results were statistically robust and clinically meaningful.

3.2.2 Automation of Analysis

To efficiently analyze a large number of biomarkers and potential comparisons, we developed an automated system to detect significant relationships between biomarkers. Using Python as the primary programming language, we used libraries such as pandas for data manipulation, numpy for numerical computations, and scipy for statistical testing. These tools allowed us to process and evaluate the datasets with ease.

The script is implemented to calculate key statistical metrics, including medians and interquartile ranges (IQRs), for selected biomarkers across defined groups. It identifies significant differences based on a comparison of these metrics, using automated workflows to ensure consistency and accuracy. The integration of Streamlit enabled us to create an interactive interface for selecting biomarkers, filtering data, and dynamically visualizing results.

The interactive interface built using Streamlit, starts by loading a dataset containing various biomarkers. This allows users to easily select biomarkers for analysis and apply filters to the data. Users can choose a specific biomarker to filter by and define custom value ranges for two groups (Range 1 and Range 2). This flexibility enables customized comparisons based on specific criteria. The code calculates key statistical metrics such as, median and interquartile range for each selected biomarker in both ranges. For each biomarker, the script checks whether the median of one group falls within the IQR of the other group. If it does, it concludes that there is "no significant difference" between the groups. If not, it identifies a "significant difference." This mathematical approach ensures consistent and objective analysis. The results are displayed in a clean and interactive table, showing the median, IQR, and comparison results for each biomarker. In addition, users can export the results as a downloadable CSV file, making it easy to share or further analyze the findings.

This automation reduced the need for manual analysis and ensured a comprehensive examination of the dataset with a significant time difference, the runtime of the program is 5 to 10 seconds for the given dataset, while it took the domain expert to identify this for one biomarker, a couple of hours, with inconsistent results at times.

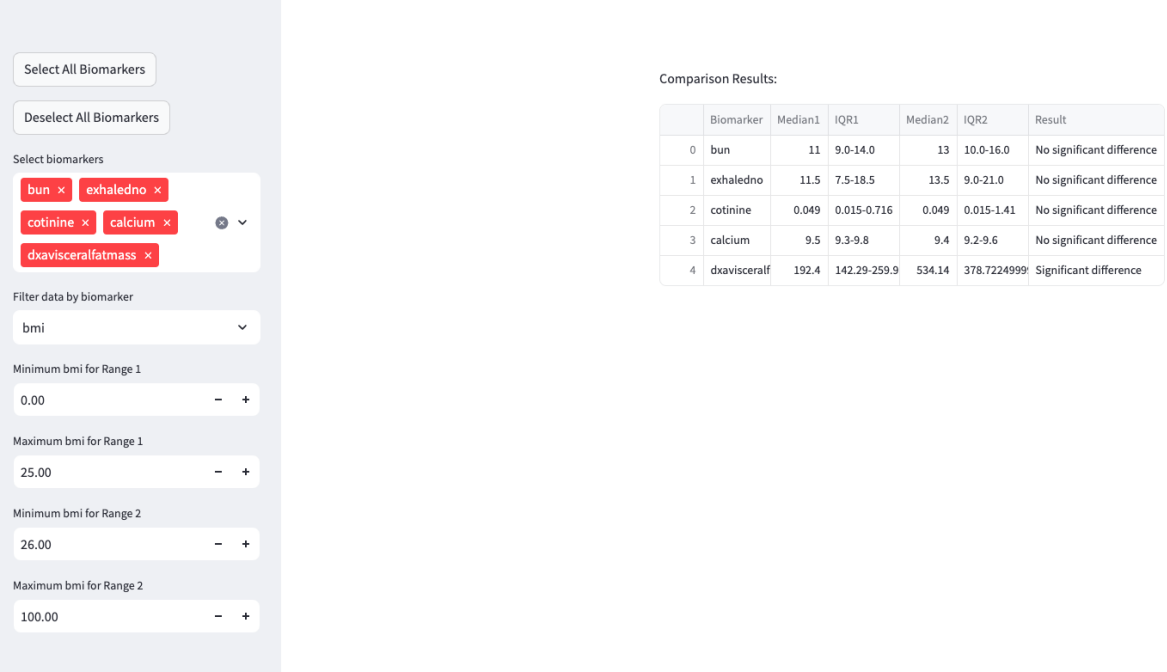

Figure 3.6: Application to Find the Significant Differences for the selected biomarkers and filter biomarker ranges

3.3 Knowledge Graph Construction

After identifying significant differences between biomarkers, the next step involved structuring these insights into a knowledge graph. This graph provided a robust and intuitive way to represent relationships between biomarkers, allowing efficient querying and visualization. The first step of constructing a knowledge graph was to create an ontology which acts as a skeletal structure for the knowledge graph. We followed the KNARM methodology to design the knowledge graph. KNARM consists of nine steps: sub-language analysis, unstructured interview, sub-language recycling, meta data creation and knowledge modeling, structured interview, KA validation, database creation, semi-automated ontology building, and ontology validation.

3.3.1 Design and Structure of the Knowledge Graph

The first steps involved analyzing the language used in the field and conducting unstructured interviews to identify key pieces of information and how they are connected within existing

knowledge about biomarkers. We consulted with a domain expert to understand the type of data they work with and their initial goals. This included learning what they were trying to analyze in the data and any early findings they had, which helped guide the direction of our analysis.

As discussed in section 2 we checked for the existing ontologies to integrate the data or modeling. However, we did not find anything related to our biomarker dataset. The Biomarker Ontology Model was created with input from a domain expert to structure and organize knowledge about biomarkers in a clear and systematic way. This model helps classify biomarkers and establish relationships between their characteristics, making it easier to understand their roles and applications. At the core of the model is the concept of a biomarker, which is linked to several key attributes. These include the FDA categorization that classifies biomarkers based on their regulatory status, the invasiveness level indicating whether the biomarker identification method requires invasive procedures, and the identification method, which details the technology or process used to detect the biomarker. The model also categorizes biomarkers based on their sample sources. Biomarkers can be derived from blood, with specific components like blood serum and blood plasma, or from urine, cerebrospinal fluid, and other bodily fluids. Furthermore, the model links biomarkers to their functions or roles. For example, some biomarkers are used to study aging, others serve as disease indicators, assess risk, or aid in diagnostics. To provide additional structure, the model organizes biomarkers into subcategories based on their attributes. For instance, biomarkers are classified as invasive or non-invasive depending on the complexity of their identification methods. Similarly, sample types like blood serum or urine are presented as subcategories under broader sample sources, creating a clear hierarchy. Relationships such as "is a" or "has part" help map connections between biomarkers and their attributes. This structured ontology provides a flexible and detailed framework for exploring biomarkers, making it useful for research and clinical applications. It can also be extended as new biomarkers or attributes are discovered. The figure 3.7 is the created model.

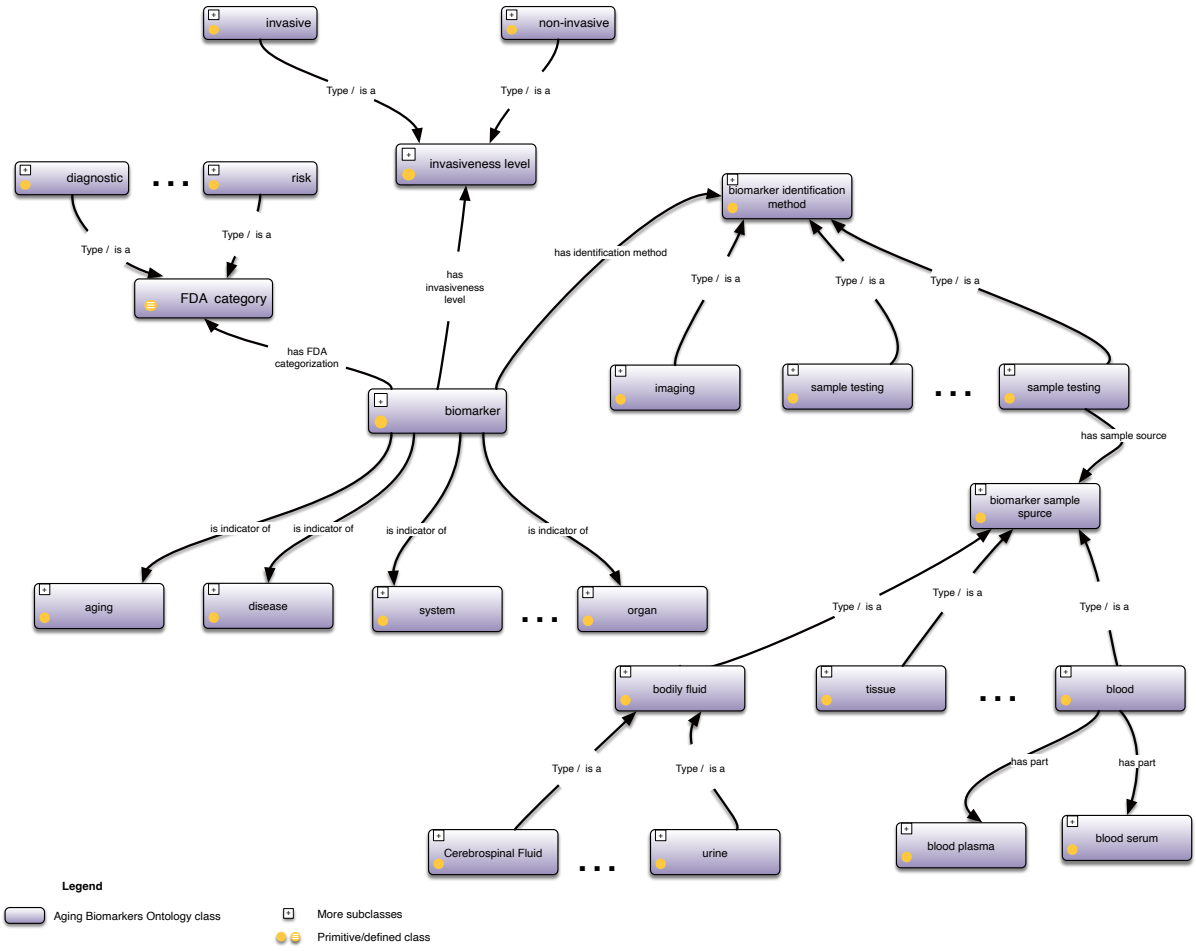

Figure 3.7: Biomarker model

3.3.2 Database Creation

As part of the knowledge graph construction, we need to create a database backend for the model to help store the data. For our database backend we used Neo4j. Neo4j is a database designed to work with graphs. Instead of storing data in tables like traditional databases, it uses nodes (representing entities) and relationships (showing how those entities are connected). This structure is useful for handling connected data, such as social networks, recommendation systems, or organizational hierarchies.

For our model, we used the following nodes to store our data:

- **Central Node:** A central node labeled Biomarkers served as the root of the graph, representing the overarching concept.

- **Category Nodes:** Each biomarker category, such as Metabolism, Fatty Acids, Iron, and Antioxidants, was represented as a node connected to the central node via the HAS_CATEGORY relationship.
- **Biomarker Nodes:** Individual biomarkers were represented as nodes within their respective categories and linked to category nodes using the HAS_BIOMARKER relationship.

Data Loading into Neo4j

The data was loaded into the Neo4j database using Cypher queries. Below is an outline of the process:

- **Creation of the Central Node:** The central Biomarkers node was created as the root of the graph.
- **Category Nodes and Relationships:** Nodes for each biomarker category were created and linked to the central node. For example:

```
CREATE (:Central {name: 'Biomarkers'});
CREATE (m:Metabolism {name: 'Metabolism'}),
    (f:FattyAcids {name: 'Fatty Acids'}), ...;
MATCH (b:Central {name: 'Biomarkers'})
    CREATE (b)-[:HAS_CATEGORY]->(m), (b)-[:HAS_CATEGORY]->(f), ...;
```

- **Biomarker Nodes and Relationships:** Nodes for individual biomarkers were created within their respective categories. These were linked to category nodes using queries like:

```
MATCH (c:Metabolism {name: 'Metabolism'})
    WITH c UNWIND ['Plasma Fasting Glucose', 'LDL Cholesterol', ...]
```

```
AS biomarker CREATE (b:Metabolism {name: biomarker})
CREATE (c)-[:HAS_BIOMARKER]->(b);
```

To maintain data integrity, unique constraints were applied to nodes to prevent duplicate entries. For example:

```
CREATE CONSTRAINT FOR (b:Metabolism) REQUIRE b.name IS UNIQUE;
CREATE CONSTRAINT FOR (b:FattyAcids) REQUIRE b.name IS UNIQUE;
```

3.3.3 Graph-Based Representation of Biomarkers

The final knowledge graph with a database provided a hierarchical structure in which the Biomarkers node was linked to various categories, and each category contained its associated biomarkers. For example:

- (Biomarkers)-[:HAS_CATEGORY]→(Metabolism)
- (Metabolism)-[:HAS_BIOMARKER]→(Plasma Fasting Glucose)

3.3.4 Applications of the Knowledge Graph

This knowledge graph served multiple purposes:

- **Querying and Reasoning:** Users could efficiently query relationships between biomarkers, categories, and their significance using Cypher.
- **Visualization and Insights:** The graph structure made it easy to visualize the biomarker relationships, making it easier to identify patterns and associations.
- **Integration with Machine Learning:** The structured data in Neo4j could be integrated with machine learning models to infer new relationships and provide reasoning for existing associations.

Figure 3.8: Biomarkers Categorization in Neo4j

3.4 Algorithmic Exploration of the Knowledge Graph

After loading the biomarker data into Neo4j, the relationships were transferred into a graph-based structure using NetworkX, a Python library for graph analysis. This allowed us to perform algorithmic approaches on the graph to infer new relationships and query the data efficiently.

3.4.1 Graph Construction

To represent the biomarker relationships as a graph, we used a directed graph structure DiGraph where:

- Each **node** in the graph represented a biomarker or an entity (e.g., BMI, DXA visceral

fat, hsCRP).

- Each **directed edge** represented a relationship between two nodes. For example, the edge (BMI, ASSOCIATED_WITH, DXA visceral fat) indicated that BMI is associated with DXA visceral fat.

The relationships were encoded using labeled edges, where the edge label ASSOCIATED_WITH described the nature of the connection between two nodes. This structure was implemented effieciently by iterating over a list of relationships and adding them to the graph using NetworkX.

For example:

- Relationships such as BMI being associated with CRP and hsCRP were represented as edges.
- These edges were added to the graph with attributes, making it possible to store and query relationship specific metadata.

3.4.2 Testing Reasoning Capabilities with a Rule-Based Graph

In addition to the biomarker graph, a separate graph was constructed to test the reasoning capabilities of the system. This graph incorporated relationships between various biomarkers, conditions, and treatments, such as ASSOCIATED_WITH, INDICATES, and RESPONDS_TO, and was designed to infer relationships that were not explicitly present in the graph but could be derived using predefined rules.

The test graph was represented as a directed graph, where: Each node represented a biomarker, condition, or treatment (e.g., CRP, Diabetes, Metformin, etc.). Each directed edge represented a relationship between two nodes, labeled with the type of relationship (e.g., ASSOCIATED_WITH, INDICATES).

For example, the following relationships were encoded:

- CRP (C-Reactive Protein) is ASSOCIATED_WITH Inflammation.

- Inflammation RESPONDS_TO Corticosteroids.
- HbA1c (Hemoglobin A1c) INDICATES Diabetes.
- HbA1c is ASSOCIATED_WITH Cardiovascular Risk.
- Metformin RESPONDS_TO Diabetes.

These edges were added using NetworkX and encoded with their respective relationship labels.

Reasoning Rules for Relationship Inference

We also used a set of rules was defined to infer relationships that were not explicitly present in the graph. For example: A condition could be inferred to RESPONDS_TO a treatment if a biomarker ASSOCIATED_WITH that condition is INDICATED by the treatment. Additional rules could infer complex relationships, such as a treatment's effectiveness on multiple conditions if shared biomarkers are present. For instance, Corticosteroids may indirectly affect conditions linked to Inflammation, such as Arthritis or Asthma, through its relationship with CRP.

The reasoning mechanism applied these rules dynamically during graph traversal, allowing the system to infer indirect relationships.

3.4.3 Depth-Limited Search Algorithm

The depth-limited search algorithm was implemented to query the graph and infer relationships within a specified depth limit. The depth limit controlled how far the algorithm could explore from the starting node, making it computationally efficient and focused.

Overview of Depth-Limited Search

Depth-limited search is a variation of the depth-first search (DFS) algorithm, where the exploration of nodes is restricted to a specific maximum depth. In our case:

- The **start node** is the entity from which the search begins (e.g., BMI).
- The **target relationship** is the type of edge we want to find or infer (e.g., ASSOCIATED_WITH).
- The **depth limit** specifies how many layers (hops) of relationships can be explored from the start node.

Key Steps in the Depth-Limited Search Algorithm

- **Initialization:** The algorithm initializes a set of visited nodes to keep track of explored nodes and avoid revisiting them, thereby preventing infinite loops. A results dictionary is also initialized to store nodes and paths that satisfy the target relationship.
- **Graph Traversal:** Starting from the specified node, the algorithm iteratively explores its neighbors (nodes directly connected by outgoing edges). For each neighbor, the edge label is checked to determine whether it matches the target relationship.
- **Matching Relationships:** If a direct match is found, the algorithm records the path from the start node to the target node in the results dictionary.
- **Applying Reasoning Rules:** If the target relationship has a predefined reasoning rule, the algorithm attempts to infer the relationship by following the sequence of relationships defined in the rule.
- **Applying the Depth Limit:** At each step, the algorithm tracks the current depth of traversal. If the depth exceeds the predefined limit, the exploration of that path is terminated.
- **Results:** After completing the traversal, the algorithm compiles all valid nodes and paths that satisfy the target relationship and returns them as the output.

Example Execution

For instance, consider the query:

- **Start Node:** BMI
- **Target Relationship:** ASSOCIATED_WITH
- **Depth Limit:** 4

The algorithm starts at BMI and explores its direct neighbors (e.g., DXA visceral fat, CRP) and their successors, up to a depth of 4. If it encounters nodes that satisfy the target relationship, they are recorded, along with the paths leading to them.

The depth-limited search helped to infer both direct and indirect relationships. Direct associations such as BMI being ASSOCIATED_WITH DXA visceral fat. Indirect associations such as connections mediated by intermediate nodes.

This algorithm proved particularly useful for Identifying clusters of biomarkers related to a specific entity (e.g., BMI), Identifying patterns, Supporting efficient and focused graph traversal within computational constraints.

The depth-limited search algorithm complemented the knowledge graph's capabilities, allowing users to query and reason over biomarker data effectively.

3.4.4 Calculating Weights Between Relationships

To find the strength of relationships between biomarkers, we used Pearson's correlation coefficient, a statistical measure that captures the linear relationship between two variables. This analysis allowed us to assign weights to the relationships in the knowledge graph, reflecting the degree of association between biomarkers.

Pearson's Correlation Coefficient

Pearson's correlation coefficient is a statistical measure used to quantify the strength and direction of a linear relationship between two continuous variables. It provides a value between -1 and 1, where a value of 1 indicates a perfect positive relationship (as one variable increases, the other also increases proportionally), -1 indicates a perfect negative relationship (as one variable increases, the other decreases proportionally), and 0 indicates no linear

relationship (changes in one variable do not predict changes in the other). In this study, Pearson's correlation was used to evaluate the relationships between various biomarkers, helping to uncover significant associations.

The formula for Pearson's correlation coefficient (r) is expressed as:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \cdot \sqrt{\sum(y_i - \bar{y})^2}}$$

In this formula, x_i and y_i are the data points for the two variables being compared, while $\bar{x}$ and $\bar{y}$ represent their respective averages (means). To calculate r we first determine how far each data point is from its mean (known as the deviation). Then, we multiply the deviations for each pair of data points (x_i and y_i) and sum these products. This measures how the two variables change together. Next, we calculate the squared deviations for each variable, sum them, and take the square root, which provides a measure of the overall variability (or spread) in each variable. Finally, the sum of the products is divided by the product of the variability for the two variables, resulting in the correlation coefficient.

The value of r indicates not only the strength but also the direction of the relationship between the two variables. A strong positive correlation (close to 1) implies that as one variable increases, the other also increases. A strong negative correlation (close to -1) means that as one variable increases, the other decreases. A value near zero suggests no linear relationship. By analyzing these values, we could identify biomarkers with significant correlations and focus on these relationships in our knowledge graph.

In this study, Pearson's correlation was crucial for assigning weights to relationships between biomarkers. The correlation coefficients served as indicators of relationship strength, enabling us to prioritize stronger connections and refine the reasoning in our depth-limited search algorithm. This allowed us to not only represent the relationships between biomarkers visually but also explore them analytically based on their statistical significance.

Data Preparation

The initial step involved preparing the biomarker dataset. The dataset combinedDataframe.csv was used and filtered to include only numerical columns, as Pearson's correlation requires continuous variables. Non-numerical columns, such as identifiers patient_id, were excluded. Missing values were handled by replacing them with the mean value of the respective column, ensuring a complete dataset for correlation calculation.

Correlation Matrix

Using the cleaned dataset, a correlation matrix was computed to evaluate pairwise Pearson's correlation coefficients between all biomarkers. This matrix represents the correlation coefficients as values ranging from -1 to 1. Here, 1 is perfect positive linear relationship, 0 is no linear relationship and -1 is perfect negative linear relationship. This correlation matrix captures the direction and magnitude of relationships between biomarkers.

The matrix was visualized as a heatmap using the seaborn library to provide an overview of correlations, where the color intensity indicates the strength and direction of relationships.

To focus on meaningful relationships, a correlation threshold of +-0.4 was applied. Biomarker pairs with absolute correlation values above the threshold were considered to have significant relationships. For each biomarker, correlations meeting this criterion were extracted and sorted in descending order of magnitude. All this data was created in a csv file.

After calculating the Pearson correlation coefficients to measure the relationships between biomarkers, the next step was to load this data into the Neo4j knowledge graph. This process involved categorizing biomarkers, creating nodes for them, and assigning weighted relationships based on their correlation coefficients.

To organize the data meaningfully, each biomarker was categorized based on expert knowledge. These categories were, Metabolism, Fatty Acids, Iron, Antioxidants, Damage, Folates, Cell Count, Thyroid, Caffeine, PFAs, Nicotine, Metals and Orphan Biomarkers. This categorization helped structure the graph and made the relationships more interpretable

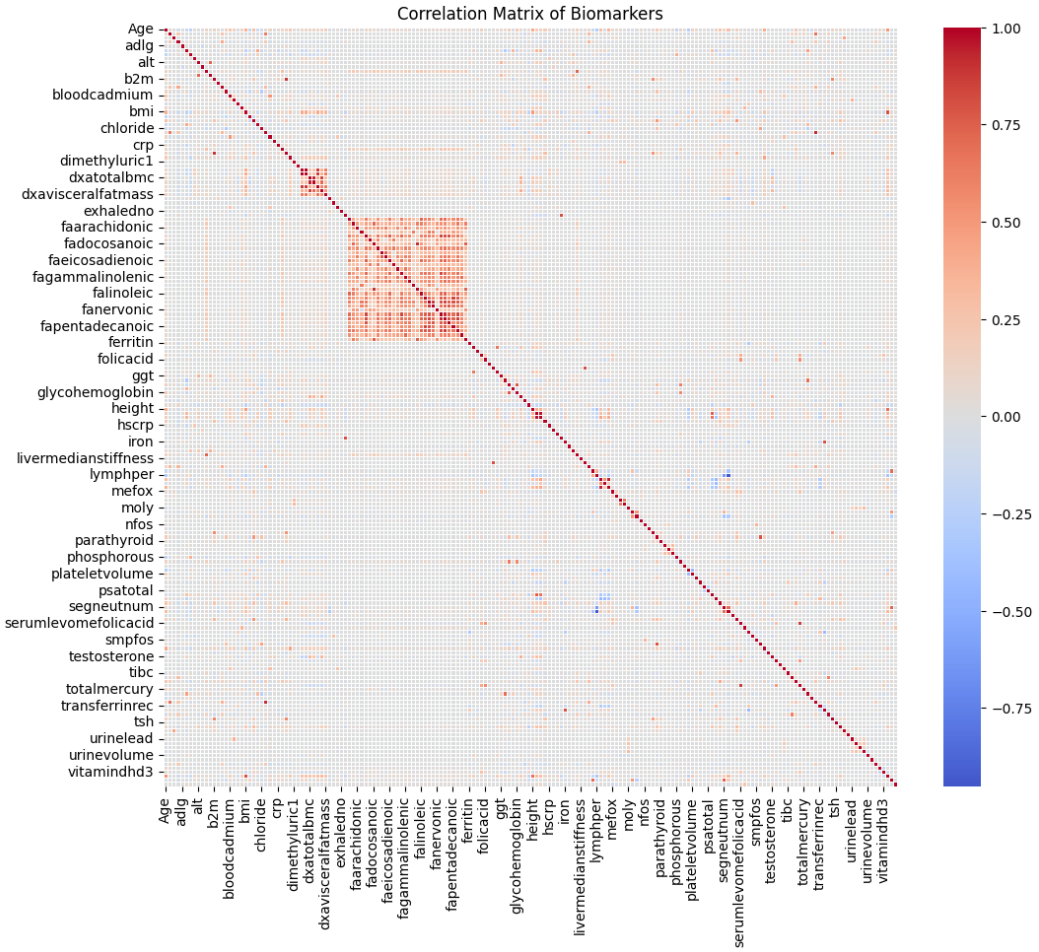

Figure 3.9: Correlation Matrix

for querying and reasoning. For each biomarker, a node was created in Neo4j and assigned to its respective category. Using the correlation coefficients calculated earlier, relationships were established between biomarker nodes. The correlation coefficient served as a weight on the relationship, indicating the strength and direction of the association between the two biomarkers.

3.4.5 Weighted Depth-Limited Search

We combined the correlation data with the depth-limited search algorithm to identify and infer stronger relationships between biomarkers in the knowledge graph. By doing this we prioritized the connections and only traversed through the relations which weigh more.

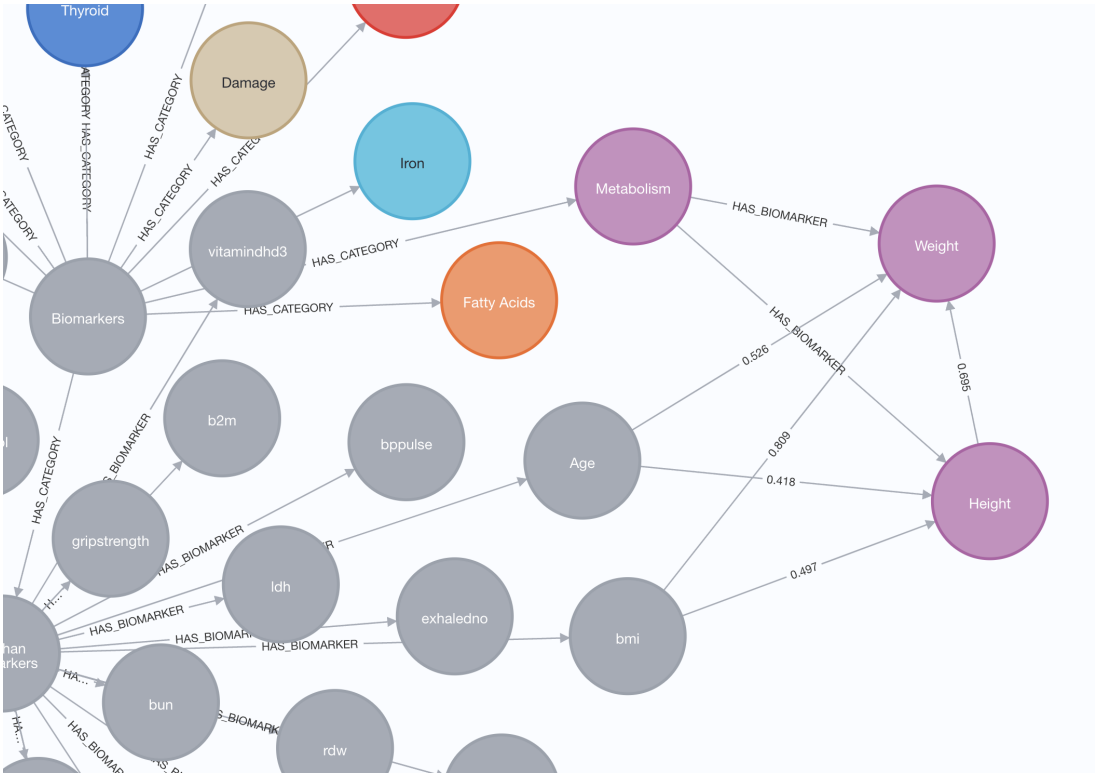

Figure 3.10: Neo4j graph with Weighted Relations

Key Concepts of the Weighted Depth-Limited Search

- Cumulative Weight: The probability of reaching a node along a path was calculated by multiplying the weights of all edges in the path.
- Weight Threshold: A minimum cumulative weight was set to filter paths and include only those with sufficiently strong relationships.
- Depth Limit: The maximum depth of traversal restricted how far the algorithm explored from the starting node, ensuring computational efficiency.

The algorithm started from a given node. For example we will give the starting node as BMI. The algorithm explored its neighbors up to a depth limit of 3. It evaluated each edge's weight to calculate the cumulative weight of reaching a neighbor. If the cumulative weight exceeded the threshold, the algorithm recorded the path and its cumulative probability and continued exploring the neighbor nodes recursively as long as the depth limit was not exceeded.

The algorithm produced paths showing how nodes (biomarkers) were connected in the graph and their cumulative probabilities. For example, a path like bmi → hsCRP → Triglycerides with a cumulative probability of 0.6 indicated that these biomarkers were strongly related through direct and indirect correlations. Paths with lower probabilities (below the threshold) were ignored, focusing the analysis on meaningful relationships.

From the Neo4j figure above we can see that as per our expert categorization Weight comes under the category Metabolism and BMI was not categorised and put them under orphan Biomarkers. As per the correlations bmi is related to Weight with strong correlation 0.809 and it is related to Metabolism. We can also infer bmi is also Metabolism.

3.5 Graph-Based Analysis Using Embeddings

In this section, we describe how we used graph-based methods and machine learning techniques to analyze biomarker relationships. Our approach involves constructing a biomarker knowledge graph, converting it into a numerical representation using embeddings, and then using machine learning to predict new associations. We also validate these predictions using scientific literature. This process helped us to speed up the analysis.

3.5.1 Building the Biomarker Knowledge Graph

The first step was to create a Knowledge Graph where each biomarker is represented as a node (point), and relationships between biomarkers (such as "Associated_With") are represented as edges (connections between points). We extracted relationships from an ontology file (OWL format), which contains structured biomedical information.

A knowledge graph helps in representing complex relationships between different biomarkers in an intuitive and searchable format. For example, if biomarker A is linked to biomarker B, and biomarker B is linked to biomarker C, we can infer potential indirect relationships between A and C that were not explicitly defined in the data.

To create this graph, we used a Python package called NetworkX, which allows us to

work with complex graphs and analyze their properties. After constructing the graph, we visualized it to understand how different biomarkers are connected. This helped us identify clusters of closely related biomarkers and recognize which ones might play a crucial role in aging-related studies.

3.5.2 Building the Adjacency Matrix

A graph can be converted into a mathematical structure called an adjacency matrix. This is a table (or matrix) that represents which nodes (biomarkers) are connected to each other. In this table:

- Each row and column represents a biomarker.
- If two biomarkers are connected, the corresponding cell in the table contains a 1.
- If they are not connected, the cell contains a 0.

For example, if we have three biomarkers A, B, and C, and only A is connected to B, the adjacency matrix would look like this:

	A	B	C
A	0	1	0
B	1	0	0
C	0	0	0

Table 3.1: Adjacency matrix representation of a simple graph

This means that A is connected to B, but neither A nor B are connected to C. Converting a knowledge graph into an adjacency matrix allows us to apply mathematical operations and machine learning techniques that are not possible with just a list of connections.

3.5.3 Reducing Complexity with Singular Value Decomposition (SVD)

The adjacency matrix of a large biomarker graph can be very high-dimensional, meaning it has a lot of information but is difficult to process efficiently. To make this data more manageable, we use a dimensionality reduction technique called Singular Value Decomposition (SVD).

SVD is a mathematical method that compresses the information in the adjacency matrix while keeping the most important patterns. It reduces the number of columns and rows but retains the essential relationships between biomarkers. This allows us to visualize biomarker relationships in a 2D or 3D space, making it easier to detect patterns.

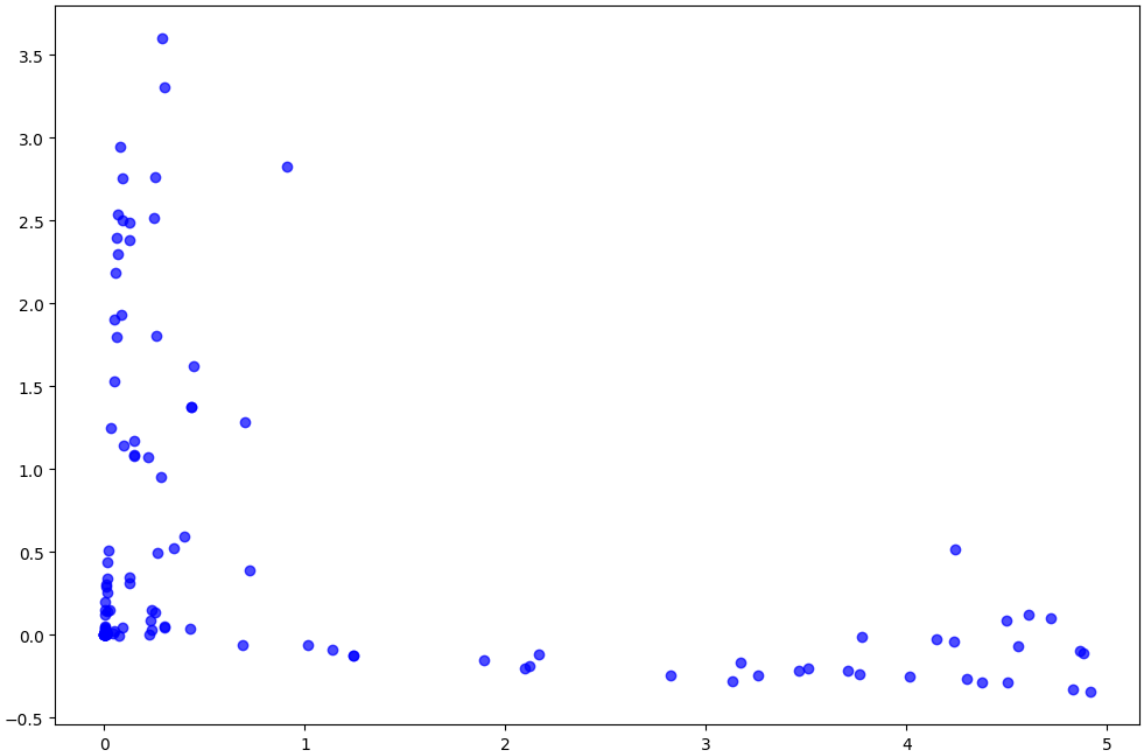

Figure 3.11: Result of Graph Embeddings

For example, if two biomarkers are often found together in biological studies, their embeddings (reduced representations) will be close to each other in the new space. This helps us group similar biomarkers and find potential new associations.

3.5.4 Using Cosine Similarity to Predict New Associations

Once we have the reduced biomarker embeddings from SVD, we need a way to measure how similar two biomarkers are. A commonly used method for this is cosine similarity.

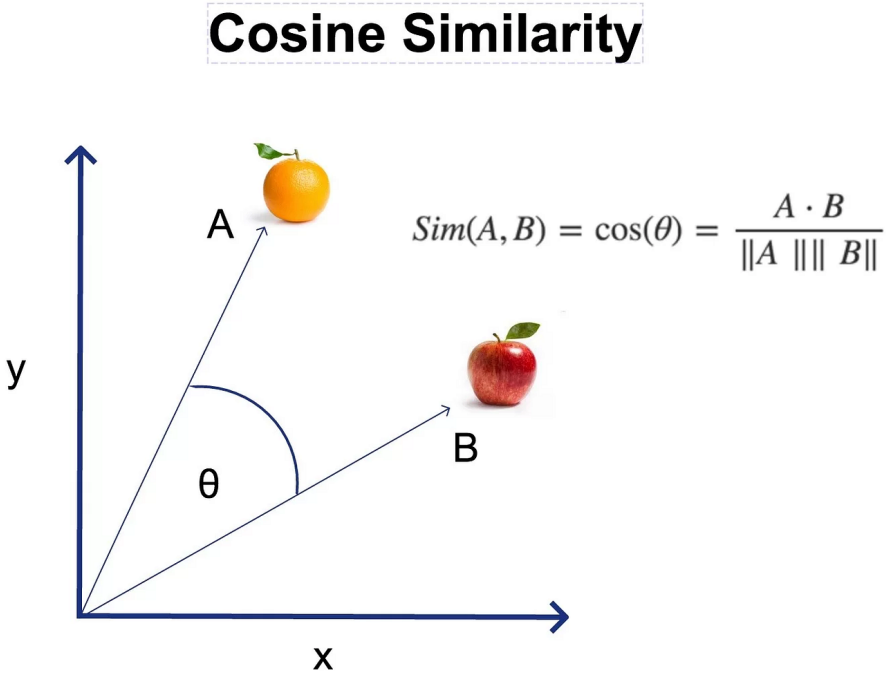

Figure 3.12: Cosine Similarity

Cosine similarity is a mathematical measure that checks how close two numerical vectors (representations of biomarkers) are in space. If two biomarkers have a high cosine similarity score (close to 1), it means they are likely to be biologically related, even if they were not explicitly connected in the original graph. If their score is low (close to 0 or negative), they are not likely to be related.

For each biomarker, we computed the five most similar biomarkers based on their cosine similarity scores. These newly identified connections represent potential new relationships that were not directly present in the original dataset.

3.5.5 Validating Predictions

As a preliminary step toward automating the validation process, we retrieved abstracts from Google Scholar using search queries composed of the two biomarker names which were potentially related to each other. This allowed us to gather contextual information about whether a relationship between them has been discussed in the scientific literature. We then manually reviewed the abstracts to check for any meaningful connection. If at least two abstracts mentioned both biomarkers in a relevant biological context, we marked the relationship as tentatively supported. While this approach does not replace domain expert validation, it provides an initial, scalable method to highlight potentially meaningful associations for further investigation. In future work, we plan to refine this process by incorporating more specialized databases such as PubMed to improve the accuracy and reliability of our literature-based checks.

Chapter 4

Conclusion and Future Work

This chapter presents the results obtained from the analysis and methodologies detailed in the preceding chapters. The findings are structured around key aspects of the study: the statistical analysis of biomarker data, the construction and utility of the knowledge graph, and the application of the depth-limited search algorithm for predictions of new biomarkers and their connections, as well as the use of graph embeddings to predict new biomarker relationships faster. These results highlight the effectiveness of combining statistical approaches with knowledge graphs and machine learning techniques to extract meaningful insights from biomedical datasets.

To achieve these results, we first collected and preprocessed biomarker data from NHANES dataset, ensuring data consistency by handling missing values and structuring the dataset for analysis. We used an approach based on the Interquartile Range (IQR) to identify significant biomarker differences. This method compared the IQR and medians of different biomarker groups to determine whether significant differences existed, ensuring robustness against missing data and non-normal distributions. By applying this method, we discovered several biomarkers that were related to one-another based on statistical significance differences.

The analysis process was significantly improved through the development of an interactive user interface, which introduced advanced filtering options and visualization tools, making

it easier to explore biomarker relationships and gain insights from the data. Throughout the development process, our industry collaborator actively used the interface and provided continuous feedback, helping refine its functionality. According to their feedback, the interface significantly streamlined their workflow, accelerating their findings by months. While this represents limited feedback, it suggests that this tool could be valuable for broader research teams working on biomarkers of aging. Furthermore, through presentations at the Biomarkers of Aging Symposium and AAAI, positive feedback was received from the research community regarding the potential impact of our tools and methodologies. Researchers expressed interest in applying our approach to their own biomarker discovery efforts, further validating the usefulness and relevance of our work in the field of aging research. Using this approach, we identified 715 statistically significant relationships between 171 biomarkers. These findings provided a foundational set of biomarker relationships that were used to construct a structured knowledge graph for further exploration.

After identifying statistically relevant biomarkers, we constructed a biomarker knowledge graph using Neo4j, categorizing biomarkers into structured ontological classes based on domain knowledge. This knowledge graph served as a structured framework for querying and exploring biomarker relationships. As part of this effort, the Biomarkers of Aging Ontology was developed, which is a novel model for defining biomarkers of aging that provides a systematic approach to organizing and categorizing aging-related biomarkers. This helped managing the aging biomarker data and improving the interpretability of associations. To support this ontology-driven model, we implemented methods to automate the database back-end generation using Neo4j using Python, significantly increasing the efficiency of ontology-based knowledge graph construction.

To further analyze biomarker relationships efficiently, a Depth Limited Search (DLS) algorithm was used, which allows traversal of the biomarker knowledge graph while restricting exploration to a predefined depth. This method allowed us to discover additional 499 indirect relationships with depth limit of 3 that were not immediately visible through statistical comparisons alone. The DLS algorithm works by initializing a traversal from a starting biomarker, keeping track of visited nodes to avoid redundant searches. By applying weight-

ing mechanisms based on statistical correlations, we ensured that stronger connections were prioritized. This method ensures that searches remain computationally efficient, preventing excessive expansion into less relevant connections while still identifying both direct and indirect relationships. DLS starts from a given biomarker node and explores its immediate neighbors, following relationships within the graph up to a specified limit. This approach prioritizes stronger, localized associations while maintaining the ability to uncover indirect connections that might not be immediately visible through statistical analysis alone. DLS revealed deeper biomarker connections, expanding upon the statistically identified relationships by uncovering previously unknown interactions in the structured knowledge graph.

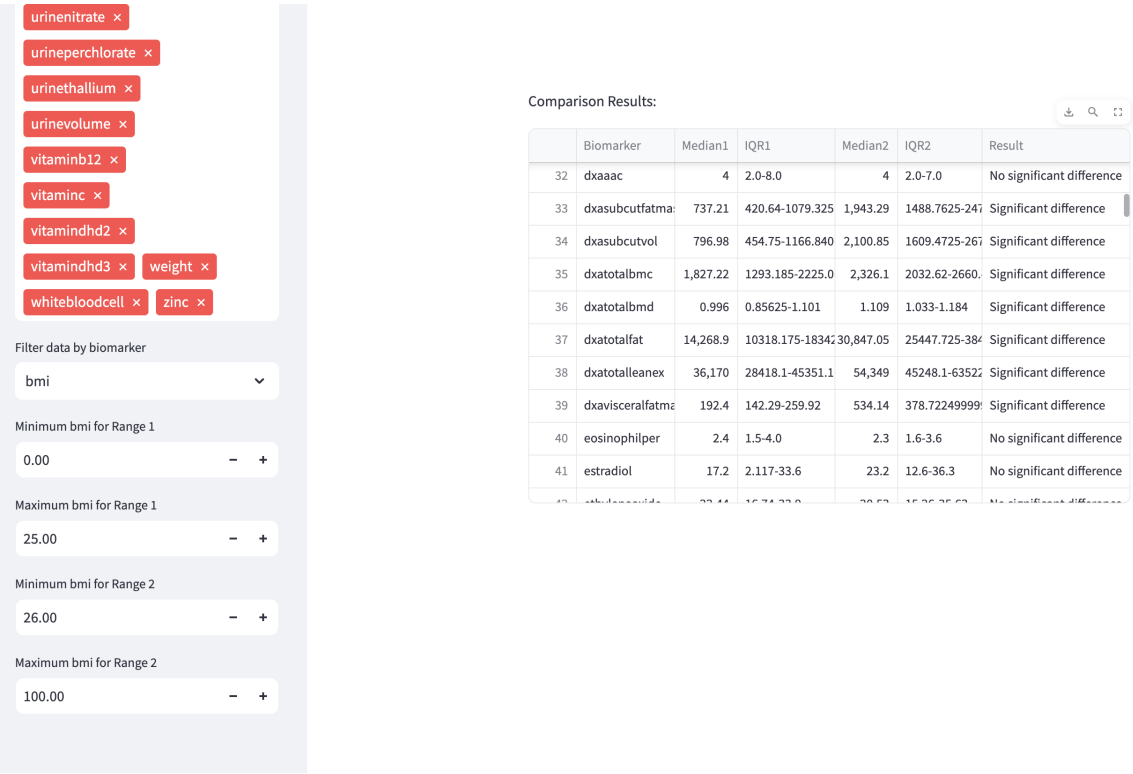

Figure 4.1: Biomarkers which shows Significant Differences for BMI

To enhance the analysis, we applied graph embeddings to uncover hidden relationships between biomarkers that were not explicitly present in the knowledge graphs. We transformed the adjacency matrix of the biomarker knowledge graph into a lower-dimensional numerical representation using Singular Value Decomposition (SVD). This technique reduced the complexity of the data while preserving key structural relationships, allowing us to analyze

biomarker associations more effectively. By computing cosine similarity between the embedded representations, we identified an additional 147 biomarker associations that had not been previously detected by either statistical methods or the DLS algorithm. These newly uncovered relationships suggest potential indirect connections between biomarkers that were not explicitly represented in the structured graph but were inferred through numerical similarity analysis. This approach helped to find new connections between biomarkers that were not known before, giving important insight into how they interact. This is a new method used in this research and the findings have not been published yet. This approach improved the overall biomarker discovery process, providing a complementary perspective to both statistical analysis and graph traversal.

Figure 4.2: Results obtained when traversing through BMI

A key observation from this comparison is that DLS identified explicit biomarker relationships that were structurally connected in the graph but did not necessarily have high similarity scores in the SVD-based embeddings. For example, DLS discovered relationships such as:

- Body Mass Index → DXA Total Fat → Fatty Acid-Palmitoleic → Fatty Acid-cis-Vaccenic
- Body Mass Index → DXA Total Fat → Weight → DXA Total Bone Mineral Content

These associations were not found using SVD similarity scores, indicating that SVD, while effective in capturing indirect associations, may overlook explicit relationships present in the knowledge graph due to its reliance on numerical similarity measures rather than direct connectivity.

Biomarker	**Relation**	**Predicted Biomarker**	**Similarity Score**
Body Mass Index	Potentially Associated With	Weight	0.954
Body Mass Index	Potentially Associated With	DXA Total Fat	0.8954
Body Mass Index	Potentially Associated With	Alpha-1-Acid	0.8625
Body Mass Index	Potentially Associated With	Cholesterol HDL	0.8577
Body Mass Index	Potentially Associated With	DXA Subcutaneous Fat Volume	0.8145

Table 4.1: SVD-Predicted Associations for Body Mass Index

In contrast, SVD was able to capture relationships that were not explicitly present in the graph but were inferred through latent semantic patterns. Identified biomarker pairs with strong similarity scores despite lacking direct graph connections, suggesting an alternative perspective on biomarker associations that might require further validation.

The similarity scores presented in the 4.1 table reflect the strength of association between biomarkers based on their embedding proximity, without considering the direction (positive or negative) of the relationship. Specifically, we used the absolute values of correlation-based similarity measures, meaning both positive and negative associations appear as high similarity. For instance, although Body Mass Index (BMI) is known to have a negative relationship with HDL cholesterol ("good cholesterol"), our method identified a strong connection based on embedding similarity, not directionality. Incorporating the direction of correlation into the similarity calculation is a potential improvement for future work.

The predicted relationships from the embeddings were further validated using Google Scholar searches. For each newly predicted pair of biomarkers, we retrieved scientific articles discussing their relationship. If multiple research studies supported the predicted association, it was considered validated for initial results. Further validation studies are part of our future research plans. We aim to partner with our industry collaborator for support and feedback in the validation steps going forward.

As described above, the combination of biomarker ontology, statistical methods, knowledge graph, and graph embeddings allowed deeper exploration of connections through a

depth-limited search algorithm. This approach uncovered additional relationships among biomarkers by efficiently navigating the structured graph and prioritizing meaningful connections. This combined approach might be particularly useful for studying aging biomarkers, where the relationships between different biological factors are complex and difficult to analyze using traditional methods. Aging often defies simplification, as other computational methods tend to overlook subtle, multidimensional relationships. By addressing these challenges, our framework demonstrated the value of integrating structured ontologies, graph-based reasoning, and machine learning algorithms. Collaboration with a domain expert further validated the findings, confirming that this approach provides an accurate and interpretable understanding of biomarker relationships. The contrast between DLS and SVD highlights the necessity of using multiple analytical approaches to ensure comprehensive biomarker discovery and validation.

As part of our future research plans, we aim to conduct further validation studies to ensure the accuracy and reliability of our findings. We plan to collaborate with our industry partner to refine and validate our biomarker predictions, using expert feedback to enhance the usability and effectiveness of our approach. This partnership will play a crucial role in evaluating the real-world applicability of our knowledge graph and machine learning methodologies, ensuring that the identified biomarker relationships are biologically meaningful and clinically relevant. A key aspect of our ongoing work is the integration of aging biomarkers into the broader biomarker research community.

Bibliography

[1] Robert M Califf. Biomarker definitions and their applications. *Experimental biology and medicine*, 243(3):213–221, 2018.

[2] Ali Bodaghi, Nadia Fattahi, and Ali Ramazani. Biomarkers: Promising and valuable tools towards diagnosis, prognosis and treatment of covid-19 and other diseases. *Heliyon*, 9(2), 2023.

[3] Srikar Reddy Gadusu, Yiğit Küçük, Vania Santillana, Aaron King, Hande Küçük McGinty. Enhancing Aging Biomarker Research through Large Language Models and Knowledge Graphs. Student Abstract, 2025.

[4] Ran Duan, Qiaoyu Fu, Yu Sun, and Qingfeng Li. Epigenetic clock: A promising biomarker and practical tool in aging. *Ageing Research Reviews*, 81:101743, 2022.

[5] Laurent Perrin, Ursula Loizides-Mangold, Stéphanie Chanon, Cédric Gobet, Nicolas Hulo, Laura Isenegger, Benjamin D Weger, Eugenia Migliavacca, Aline Charpagne, James A Betts, et al. Transcriptomic analyses reveal rhythmic and clock-driven pathways in human skeletal muscle. *Elife*, 7:e34114, 2018.

[6] M Austin Argentieri, Sihao Xiao, Derrick Bennett, Laura Winchester, Alejo J Nevado-Holgado, Upamanyu Ghose, Ashwag Albukhari, Pang Yao, Mohsen Mazidi, Jun Lv, et al. Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations. *Nature medicine*, 30(9):2450–2460, 2024.

[7] Zhang-Zhi Hu, Hongzhan Huang, Cathy H Wu, Mira Jung, Anatoly Dritschilo, Anna T Riegel, and Anton Wellstein. Omics-based molecular target and biomarker identification. *Bioinformatics for Omics Data: Methods and Protocols*, pages 547–571, 2011.

[8] Masahiko Gosho, Kengo Nagashima, and Yasunori Sato. Study designs and statistical analyses for biomarker research. *Sensors*, 12(7):8966–8986, 2012.

[9] Jon R Sobus, Robert S DeWoskin, Yu-Mei Tan, Joachim D Pleil, Martin Blake Phillips, Barbara Jane George, Krista Christensen, Dina M Schreinemachers, Marc A Williams, Elaine A Cohen Hubal, et al. Uses of nhanes biomarker data for chemical risk assessment: trends, challenges, and opportunities. *Environmental health perspectives*, 123(10):919–927, 2015.

[10] Jordan Bortz, Andrea Guariglia, Lucija Klaric, David Tang, Peter Ward, Michael Geer, Marc Chadeau-Hyam, Dragana Vuckovic, and Peter K Joshi. Biological age estimation using circulating blood biomarkers. *Communications biology*, 6(1):1089, 2023.

[11] Shichao Zhang, Chengqi Zhang, and Qiang Yang. Data preparation for data mining. *Applied artificial intelligence*, 17(5-6):375–381, 2003.

[12] Arjun K Manrai, Chirag J Patel, and John PA Ioannidis. In the era of precision medicine and big data, who is normal? *Jama*, 319(19):1981–1982, 2018.

[13] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[14] Richard Mayeux. Biomarkers: potential uses and limitations. *NeuroRx*, 1:182–188, 2004.

[15] Shannon E Ellis and Jeffrey T Leek. How to share data for collaboration. *The American Statistician*, 72(1):53–57, 2018.

[16] M Cristina Polidori. Aging hallmarks, biomarkers, and clocks for personalized medicine:(re) positioning the limelight. *Free Radical Biology and Medicine*, 2024.

[17] Adiv A Johnson and Maxim N Shokhirev. Contextualizing aging clocks and properly describing biological age. *Aging Cell*, 23(12):e14377, 2024.

[18] Mahdi Moqri, Chiara Herzog, Jesse R Poganik, Kejun Ying, Jamie N Justice, Daniel W Belsky, Albert T Higgins-Chen, Brian H Chen, Alan A Cohen, Georg Fuellen, et al. Validation of biomarkers of aging. *Nature medicine*, 30(2):360–372, 2024.

[19] Stephanie Bauer, Michael J Lambert, and Steven Lars Nielsen. Clinical significance methods: A comparison of statistical techniques. *Journal of personality assessment*, 82 (1):60–70, 2004.

[20] B Rosner and Douglas Grove. Use of the mann–whitney u-test for clustered data. *Statistics in medicine*, 18(11):1387–1400, 1999.

[21] Jacob Benesty, Jingdong Chen, and Yiteng Huang. On the importance of the pearson correlation coefficient in noise reduction. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(4):757–765, 2008.

[22] David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. Detecting novel associations in large data sets. *science*, 334(6062):1518–1524, 2011.

[23] Trevor Hastie. The elements of statistical learning: data mining, inference, and prediction, 2009.

[24] Martin Bland. *An introduction to medical statistics*. Oxford university press, 2015.

[25] Pascal Hitzler, Markus Krötzsch, Bijan Parsia, Peter F Patel-Schneider, Sebastian Rudolph, et al. Owl 2 web ontology language primer. *W3C recommendation*, 27(1):123, 2009.

[26] Carl Yang, Hejie Cui, Jiaying Lu, Shiyu Wang, Ran Xu, Wenjing Ma, Yue Yu, Shaojun Yu, Xuan Kan, Chen Ling, et al. A review on knowledge graphs for healthcare: Resources, applications, and promises. *arXiv preprint arXiv:2306.04802*, 2023.

[27] Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife*, 6: e26726, 2017.

[28] Johannes Maron and et al. Using knowledge graphs for biomarker discovery. *Frontiers in Genetics*, 12:x–y, 2021. doi: 10.xxxx/fgene.2021.xxxxxx.

[29] Hande Küçük McGinty. *KNowledge Acquisition and Representation Methodology (KNARM) and its applications*. PhD thesis, University of Miami, 2018.

[30] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, et al. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255, 2007.

[31] Renzo Angles and Claudio Gutierrez. Survey of graph database models. *ACM Computing Surveys (CSUR)*, 40(1):1–39, 2008.

[32] Daniel Rivas-Barragan, Sarah Mubeen, Francesc Guim Bernat, Martin Hofmann-Apitius, and Daniel Domingo-Fernández. Drug2ways: Reasoning over causal paths in biological networks for drug discovery. *PLoS computational biology*, 16(12):e1008464, 2020.

[33] Dan Kalifa, Uriel Singer, and Kira Radinsky. Goproteingnn: Leveraging protein knowledge graphs for protein representation learning. *arXiv preprint arXiv:2408.00057*, 2024.

[34] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009. doi: 10.4018/jswis.2009081901.

[35] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1): 11–33, 2016. doi: 10.1109/JPROC.2015.2483592.

[36] Dongwoo Kim, Lexing Xie, and Cheng Soon Ong. Probabilistic knowledge graph construction: Compositional and incremental approaches. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 2257–2262, 2016.

[37] Xiang Yue, Zhen Wang, Jingong Huang, Srinivasan Parthasarathy, Soheil Moosavinasab, Yungui Huang, Simon M Lin, Wen Zhang, Ping Zhang, and Huan Sun. Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics*, 36(4):1241–1251, 2020.

[38] Anton Tsitsulin, Davide Mottin, Panagiotis Karras, and Emmanuel Müller. Verse: Versatile graph embeddings from similarity measures. In *Proceedings of the 2018 world wide web conference*, pages 539–548, 2018.