

# UCLA

## UCLA Previously Published Works

### Title

A review of feature selection strategies utilizing graph data structures and Knowledge Graphs.

### Permalink

<https://escholarship.org/uc/item/7g84p0r5>

### Journal

Briefings in Bioinformatics, 25(6)

### Authors

Shao, Sisi

Henrique Ribeiro, Pedro

Ramirez, Christina

et al.

### Publication Date




2024-09-23

### DOI

10.1093/bib/bbae521

Peer reviewed

# A review of feature selection strategies utilizing graph data structures and Knowledge Graphs

Sisi Shao <sup>1</sup>, Pedro Henrique Ribeiro<sup>2</sup>, Christina M. Ramirez <sup>1</sup>, Jason H. Moore <sup>1,2,\*</sup>

<sup>1</sup>Department of Biostatistics, Fielding School of Public Health at University of California, Los Angeles, 650 Charles E Young Dr S, Los Angeles, CA 90095-1772, United States

<sup>2</sup>Department of Computational Biomedicine, Cedars-Sinai Medical Center, 8700 Beverly Blvd, Los Angeles, CA 90048, United States

\*Corresponding author. Jason.Moore@csmc.edu

## Abstract

Feature selection in Knowledge Graphs (KGs) is increasingly utilized in diverse domains, including biomedical research, Natural Language Processing (NLP), and personalized recommendation systems. This paper delves into the methodologies for feature selection (FS) within KGs, emphasizing their roles in enhancing machine learning (ML) model efficacy, hypothesis generation, and interpretability. Through this comprehensive review, we aim to catalyze further innovation in FS for KGs, paving the way for more insightful, efficient, and interpretable analytical models across various domains. Our exploration reveals the critical importance of scalability, accuracy, and interpretability in FS techniques, advocating for the integration of domain knowledge to refine the selection process. We highlight the burgeoning potential of multi-objective optimization and interdisciplinary collaboration in advancing KG FS, underscoring the transformative impact of such methodologies on precision medicine, among other fields. The paper concludes by charting future directions, including the development of scalable, dynamic FS algorithms and the integration of explainable AI principles to foster transparency and trust in KG-driven models.

**Keywords:** feature selection; Knowledge Graphs; deep learning; precision medicine; explainable AI

## Introduction

### Brief introduction to Knowledge Graphs

In the era of large-scale digital information, Knowledge Graphs (KGs) are an increasingly popular tool to organize data and information [1]. At their core, KGs are an organized representation of entities, such as objects, events, situations, or concepts that illustrate how those entities are related—through triplets (subject-predicate-object). For instance, a triplet like ‘Cyclophosphamide - treats - Cancer’ could be used to guide a KG in drug discovery and repurposing. KGs allow for in-depth data analysis and the development of personalized care strategies. KG platforms like Bio2RDF, for instance, have been instrumental in exploring the complex relationships between genetics, diseases, and environmental factors (see [Figure 1](#)). KGs can facilitate a comprehensive approach to healthcare supporting a wide range of applications, from advanced decision-support systems to personalized medicine and innovative drug discovery methods [2, 3].

One of the most well-known uses for KGs is in the development of web-based technologies, including search engines and the Semantic Web (an extension of the World Wide Web that enables data to be shared and reused across applications). Google KG, DBpedia, and Yet Another Great Ontology (YAGO) utilize the principles of the Semantic Web and Linked Open Data (a method of publishing structured data so that it can be interlinked and become more useful) to create extensive networks of nodes and edges, that represent the intricate relationships within vast datasets, and enable enhanced query processing and analytics

capabilities. The contributions of scholars such as Fensel *et al.* [4], Bonner *et al.* [5], and Yang *et al.* [6] have been crucial in shedding light on the foundational aspects and ongoing evolution of these systems.

As technology advances at an incredible pace, we are accumulating a vast amount of knowledge about genes, proteins, chemicals, cells, diseases, and other biological entities, along with their complex interactions[7]. To make sense of this complexity, KGs have emerged as powerful tools for organizing and connecting this intricate and multifaceted information in meaningful ways. In the realm of precision medicine, KGs have been used to consolidate disparate biomedical data, and thereby systematically utilize genetic, environmental, and lifestyle information to improve the effectiveness of personalized patient care. This is exemplified by PrimeKG, which significantly contributes to creating a comprehensive medical knowledge base by integrating a wide ontology with data from various sources, including genomic databases, thereby supporting detailed medical research and personalized care planning [8].

At their core, KGs are characterized by representing entities and their relationships through triplets (subject-predicate-object), allowing for in-depth data analysis and the development of personalized care strategies. For instance, a triplet like ‘Cyclophosphamide - treats - Cancer’ demonstrates KGs’ potential in drug discovery and repurposing. Platforms like Bio2RDF have been instrumental in exploring the complex relationships between genetics, diseases, and environmental factors. KGs

Received: June 21, 2024. Revised: August 1, 2024. Accepted: October 2, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

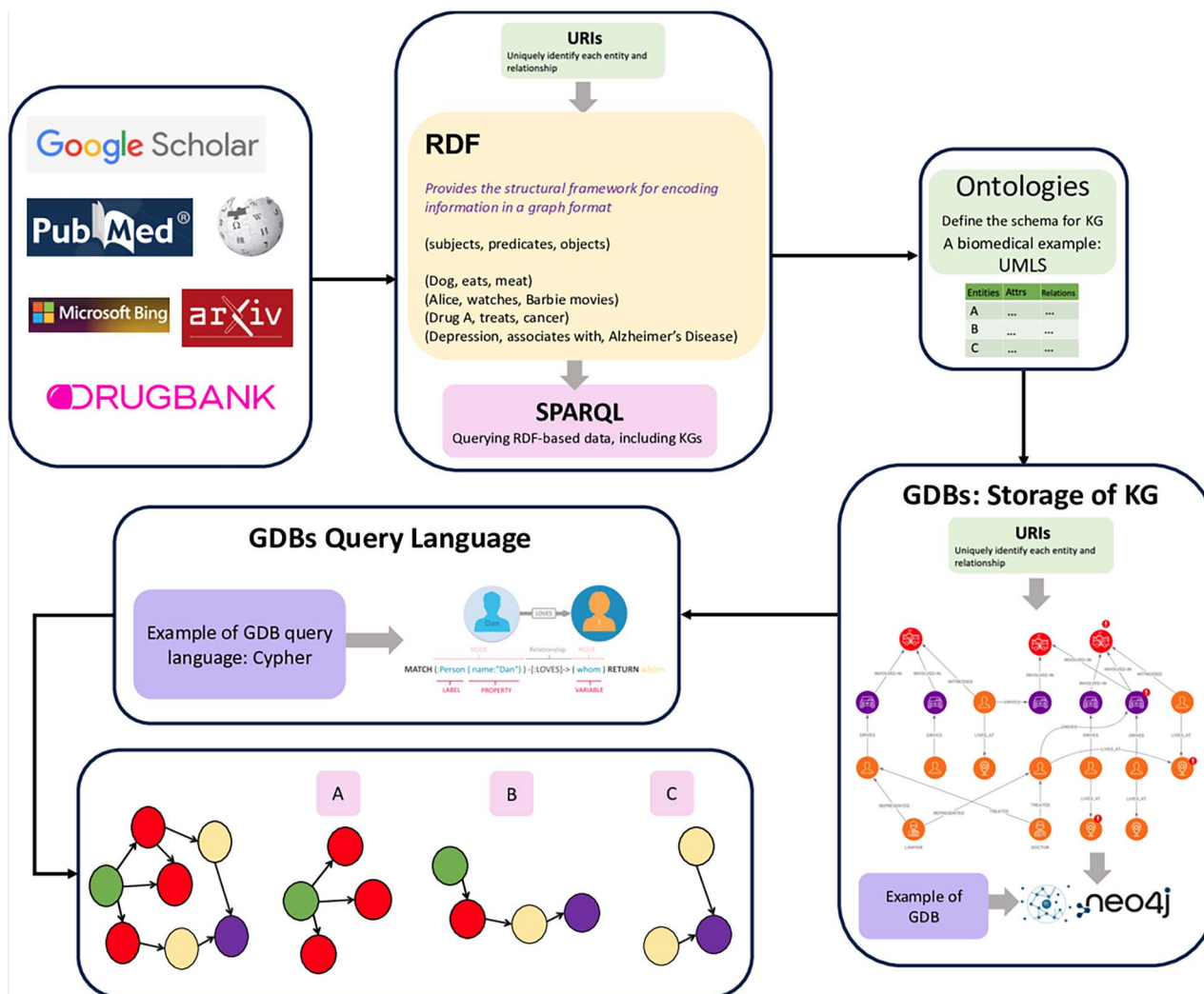


Figure 1. An integrated overview of KGs encompassing RDF structuring, Ontological frameworks, and GDB management, illustrating the flow from data sources to semantic querying and storage. Figure 1 delineates the contribution of varied scholarly and scientific data sources—such as Google Scholar, PubMed, arXiv, and DrugBank—in providing raw data inputs. These inputs are then semantically encoded via the RDF, using triples that consist of subjects, predicates, and objects, alongside URIs that ensure the unique identification and integration of data entities across the KG. At the heart of the semantic structure are ontologies, exemplified here by the Unified Medical Language System, which defines the schema for the KG by outlining the essential relationships and attributes of the domain-specific entities. This ontology-based schema informs the organization and representation of knowledge within GDBs, such as Neo4j, which are specialized for storing and operationalizing the complex relational data of KGs. The central rounded box showcases the role of query languages, with Cypher portrayed as a model for extracting information from GDBs through its intuitive syntax and pattern matching capabilities. The graphic elucidation of the query output illustrates a network of nodes and edges, representing the intricate interrelations and potential analytical insights derived from KGs. Each cluster within the network, designated as A, B, and C, symbolizes distinct subsets or aspects of the graph database that have been queried.

thereby facilitate a comprehensive approach to healthcare; this approach supports a wide range of applications, from advanced decision-support systems to personalized medicine and innovative drug discovery methods [2, 3].

The integration and analysis of data from biomedical research and clinical practice through KGs provide a dynamic platform for advancements in understanding and treating diseases. The academic discourse on feature selection (FS) methods applied to KGs, as highlighted by the studies referenced, underscores their transformative potential in various domains, particularly in advancing personalized medicine and healthcare outcomes.

### Importance of FS

Feature selection involves choosing the subset of input variables that are most relevant for analysis. It is a crucial step in any type of modern research that uses machine learning (ML) models.

As datasets grow in size and complexity (ranging from petabytes to exabytes), robust FS is essential for preventing the ‘curse of dimensionality’ [9], which can degrade model performance. Reducing a model’s feature set helps to mitigate overfitting and improves computational efficiency [10]. This reduction aids ML model interpretability in critical domains like healthcare and finance [11, 12], and enhances a model’s generalizability to new data, a cornerstone for practical applications [13, 14]. Streamlined ML models require fewer computational resources, and are beneficial in resource-constrained scenarios like edge computing [15, 16]. With big data’s growing influence, especially in healthcare where it is projected to reach \$79.23 billion by 2028, FS is increasingly vital for ensuring robust and applicable ML models.

In regards to ML, FS most often refers to selecting particular columns from a tabular dataset. In this paper, we take a broader view, whereby FS also includes the selection of specific nodes or

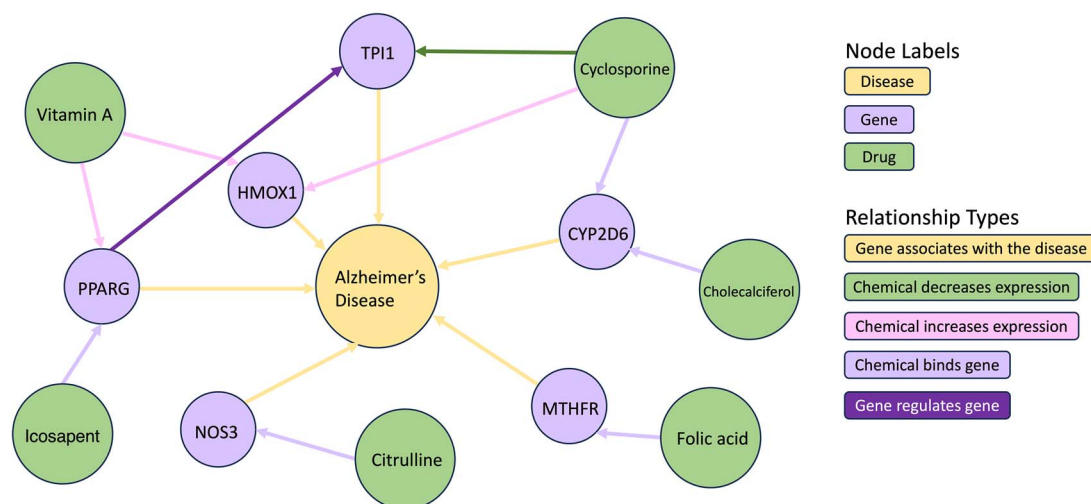


Figure 2. A tiny-sized ADKG (yellow node: AD; purple nodes: genes; green nodes: drugs) [28]. There are five instances of the 'Chemical binds gene' relationship (light purple arrows), where a chemical is shown to interact directly with a gene; six instances of the 'Gene associates with disease' relationship (yellow arrows), representing genes that have an association with AD; one instance of the 'Chemical decreases expression' relationship (dark green arrow), indicating a chemical that downregulates or decreases the expression of a gene; one instance of 'Gene regulates gene' (purple arrow), suggesting a regulatory interaction between two genes, PPARG and TPI1. More detailed information on genes and drugs is given in the Appendix B.

entities for hypothesis generation and further investigation. With this broader view, for example, a KG with genes and diseases can hypothesize new subsets of genes related to a specific disease.

Recognizing various FS methods, such as algorithmic techniques, statistical analyses [17], and expert insights, this review will explore the relationship between KGs and FS, highlighting how these frameworks can enhance the FS process.

## Overview of the relationship between KGs and FS

Integrating KGs with FS enhances ML models by incorporating domain-specific knowledge often overlooked in AI/ML systems. KGs provide structured representations of entities, attributes, and interconnections, aiding in precise FS across domains like the Semantic Web, Natural Language Processing (NLP), and data integration. This improves model performance, reduces overfitting, and enhances interpretability.

However, challenges include scalability, KG integrity, and domain adaptation. Research efforts are needed to develop scalable algorithms, improve KG completeness, and integrate diverse data sources. Combining knowledge representation, ML, and domain expertise is essential.

Innovative methods like embedding-based FS and graph neural networks (GNNs) leverage KGs' unique characteristics for effective FS. These approaches manage high-dimensional spaces in KGs, enabling comprehensive data analysis.

KGs' dynamic nature requires adaptive, real-time FS methods to ensure relevant features and maintain ML models' integrity in rapidly changing scenarios.

## Background and key concepts

### Definition and structure of KGs

KGs categorize and link data for domain-specific knowledge discovery.

### Ontologies

KGs use ontologies to define relationships and model semantics [18]. Ontologies categorize concepts to allow for flexible queries.

Bio2RDF, for example, defines classes like 'proteins' and 'chemical entities,' and their relationships using resource description framework (RDF) triples.

### Example: Bio2RDF

Bio2RDF integrates datasets like DrugBank [19], SIDER [20], and KEGG [21] into a unified RDF structure, thus enhancing data interoperability and supporting complex queries.

- **Nodes:** tagged with URIs, representing biomedical entities like genes and drugs.
- **Relationships:** include 'targets' and 'is affected by,' illustrating drug-protein interactions and genetic influences.

## Structuring domain knowledge with RDF RDF

RDF provides a structure for semantic representation in KGs [22]. It formalizes relationships as triplets (subject-predicate-object) forming a graph  $G = \{(s, p, o)\}$ . RDF enhances data interlinking and queryability [23, 24].

### Ontologies

Ontologies in KGs categorize and describe concepts with flexible relationships. They enhance querying capabilities by defining both specific and abstract relationships, as seen in Bio2RDF and AlzKB.

### Leveraging graph databases

Graph databases (GDBs), like Neo4j, manage complex data relationships within KGs, enabling efficient semantic analysis [25]. Freebase and query languages like Cypher and SPARQL extend GDB functionality for intuitive querying [26, 27].

### Visual demonstration of ADKGs of varying sizes

We use AlzKB, an Alzheimer's disease KG, as an example to demonstrate KGs of various sizes. Figures represent tiny (Cypher query limit 8), small (Cypher query limit 15), and medium (Cypher query limit 200) KGs. A tiny KG example is shown in Fig. 2.

## Feature selection on KGs

In this section, we will categorize and evaluate the methodological frameworks delineated within the referenced manuscripts. Below, we elaborate on four distinct KG FS methods, including search algorithms, similarity-based methods, vector embeddings, and advanced network representation learning—all available in the most current literature to the best of our present knowledge.

One particular application of FS on KGs is drug repurposing with selected feature sets using AlzKB [28]. Here, feature sets refer to the genes that are targeted by a given drug. For instance, flubendazole is an anthelmintic that targets many different genes including PCNA, CDK4, etc., [29]. To investigate new drugs that have potential value in treating Alzheimer's disease, we can select genes that are potentially related to Alzheimer's disease and use them to form feature sets for analysis. This is an ongoing research and was funded by the National Institutes of Health [U01 AG066833].

### Causal discovery-search algorithm

The goal of causal discovery is to move beyond merely describing correlated events to identifying the direction of influence between observed phenomena. The challenge in causality analysis lies in capturing the complex interactions between variables. Typically, these relationships are formalized using causal graphs, where nodes represent variables and directed edges denote causal effects.

In medicine, the gold standard for establishing causal relationships, including confounding, collider, mediation, moderation, reverse causality, effect modification, causal chain, and causal graph, is through randomized controlled trials. However, various analytical methods can infer causal relationships from observational data. In this analytical approach, researchers must consider other measured or unmeasured variables that may act as confounders, mediators, or colliders. For a comprehensive review of causal discovery, we recommend this survey paper by Zanga et al. [30].

There has been a lot of work recently on building automated methods, generally utilizing NLP techniques, to extract causal relations from the scientific literature. KGs can be used to consolidate knowledge and form inferences and hypotheses about how different variables interact. Causal analysis can then be used to identify features that have causal effects on downstream variables.

The study by Malec et al. [31] introduced a novel causal FS framework using the 'ADKG' KG. This ADKG was constructed from post-2010 PubMed biomedical literature and an ontology-grounded KG via the PheKnowLator workflow [32]. The authors used PubMed identifiers and machine reading systems like EIDOS, REACH, and SemRep within the INDRA ecosystem [33] to extract data. INDRA assembles knowledge into a model of causal molecular interactions [34], resulting in an OWL ontology [35].

The Malec study aimed to enhance causal FS with the ADKG. The authors performed hygiene steps, and omitted logical entailments. They then map predicates to the relation ontology (RO) to provide logical definitions and infer additional knowledge. Their forward-chaining inference used CLIPS to generate new triples based on RO properties, after assigning belief scores. The authors then integrated PheKnowLator to facilitate path search algorithms, thereby reweighting edges with hierarchical relationships for optimized path searches. Competency questions, such as causal relationships between depression and AD, were

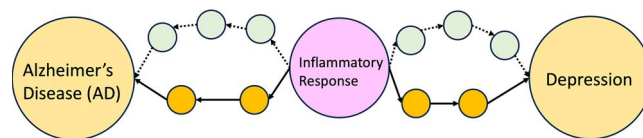


Figure 3. Illustration of inflammatory response (pink node) as a potential confounder in the association between AD (left yellow node) and depression (right yellow node). The diagram represents the shortest paths (through orange nodes) identified by Dijkstra's algorithm. The two green paths also connect inflammatory response with AD and depression, but both of them are one unit longer than the orange ones. Consequently, Dijkstra's algorithm picks the shortest path.

addressed using SPARQL queries and Dijkstra's shortest path algorithm [34, 36].

When applied to ADKG, Dijkstra's algorithm identified the shortest paths connecting genes and diseases, highlighting direct relationships [31]. These paths were analyzed to identify potential confounders, colliders, and mediators. Confounders influence both exposure and outcome, colliders are influenced by both, and mediators act as intermediaries. Figure 3 illustrates identifying a potential confounder between AD and depression using Dijkstra's algorithm. The study identified 126 unique potential confounders, 29 colliders, and 18 potential mediators, showcasing the ADKG's ability to uncover intricate relationships that traditional searches might miss.

### Feature selection-dimensionality reduction

KGs can be utilized to perform FS for high-dimensional tabular datasets. In this scenario, nodes in the graph may relate to the columns, or features, of the tabular dataset. These subsets of features can then be analyzed using methods like ML models. Below, we outline a few examples of graph-based methods for selecting subsets of features.

- Fang et al. [37] developed an information theory approach, informed by a KG, to select features for training ML models. The goal of their study was to develop a predictive model of chronic obstructive pulmonary disease (COPD) from a tabular dataset including 28 features representing medical tests and patient symptoms. First, the authors constructed a KG by integrating electronic medical records and domain-specific biomedical knowledge to identify and represent relationships among diseases, symptoms, causes, risk factors, drugs, side effects, and more (see Figure 5). The features of the tabular dataset corresponded to nodes in the KG. Their algorithm, CMFS- $\eta$ , used the weights between features in the KG to iteratively add or remove features from the set according to an information-theory-based heuristic. The study used this approach to select subsets of the corresponding features of the tabular dataset to train an SVM model.
- Ma et al. [38] sought to develop a model to predict whether a given Android app contained malware based on the Android API calls contained in the source code. First, they used the official documentation to construct a KG containing all API entities, such as classes and methods, as well as relationships between entities, such as return types and inheritance. Next, they identified a set of permissions considered to be highly sensitive that was required for each API entity. The study created a binary feature vector for each application based on whether or not a given entity was present in the code. To reduce the size of the binary feature vector, the authors



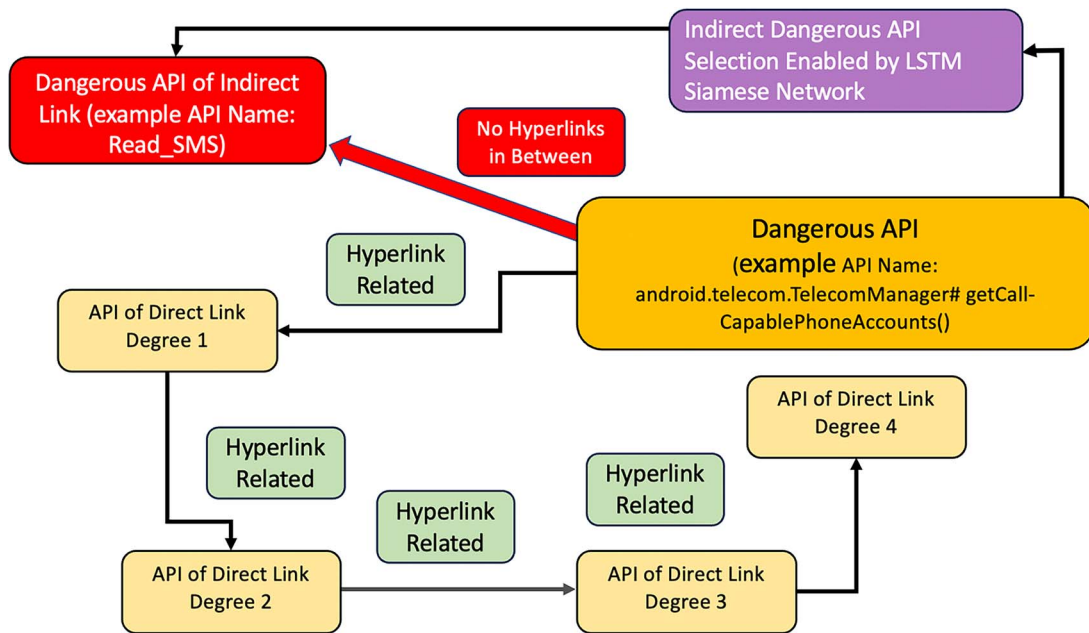


Figure 4. Example of direct and indirect dangerous API selection, as enabled by the Android API KG. The golden-orange rounded rectangle in the figure signifies a dangerous API called ‘getCall-CapablePhoneAccounts,’ which facilitates the retrieval of Phone Account Handles for making and receiving calls. The light-yellow rounded rectangles are APIs directly connected to the Dangerous API, up to four degrees of separation through hyperlinks, with the understanding that links beyond this do not markedly enhance classification accuracy. The Siamese-BiLSTM network comes into play by identifying indirectly connected, potentially dangerous APIs—represented by the red rounded rectangle, such as ‘READ SMS,’ which allows reading SMS messages but lacks a direct hyperlink or descriptive connection to other APIs. By embedding API descriptions into a vector space using Word2Vec and processing them through a Bidirectional LSTM, the network encodes the APIs’ textual data from both directions for a full context capture. These encoded vectors are then condensed through a dense layer into a final representation. Comparing these representations enables the network to detect hidden APIs that, while not directly linked, share sensitive characteristics with the known dangerous API, thereby revealing hidden dangers through textual similarity rather than explicit interlinking.

selected only entities that were between one to four hops from a node requiring sensitive permission. As not all entities contained explicit links in the documentation, an LSTM model was used to identify an additional subset of entities that shared similar descriptions with entities that require sensitive permissions. This feature vector was then used to train a classification model. A detailed description of how sensitive APIs, or nodes in the KGs, were selected is shown in Fig. 4.

- In the Hadith Corpus KG created by Mohammed et al. [41], nodes represent distinct features and semantic categories derived from Hadith texts. Features include specific Islamic terms like ‘prayer’ or ‘fasting,’ while categories encompass broader thematic areas like rituals, ethics, jurisprudence, and other domains of Islamic scholarship. Edges in this KG quantify associations between features and categories based on co-occurrence frequency.

Feature selection for text classification is guided by Ant Colony Optimization (ACO) [42–44]. ACO is a probabilistic technique for solving computational problems that can be reduced to finding good paths through graphs. Inspired by the behavior of ants, which find the shortest path from their colony to food sources, ACO is a type of swarm intelligence method and a subset of evolutionary algorithms. Initially, several paths are randomly constructed, and after traversing a path, an ant deposits pheromones along it (typically inversely proportional to path length), so shorter paths receive more pheromones. Over time, the pheromones evaporate, reducing their attractive strength to prevent premature convergence. When choosing their paths, ants probabilistically

prefer paths with stronger pheromone concentrations while also exploring new paths to avoid local optima. The process is repeated until convergence. In this way, ACO balances between exploring new feature paths (exploration) and intensifying the search around promising features found in previous iterations (exploitation), thus adapting dynamically to find optimal feature sets for text classification [45–48]. The pheromone trail and PageRank-like heuristic measure guide this optimization. We provide a graphical illustration of the ACO FS process in Fig. 6.

This study demonstrates that integrating ACO into Arabic text classification yields a notable 3% average increase in accuracy, F1 score, recall, and precision compared to conventional methods like Naive Bayes, Random Forest, Decision Trees, and XGBoost, thus contributing significantly to the field of Arabic text classification.

### Data linking and data integration-similarity based methods

- Data linkage and data integration refer to the process of combining different sources of data [49]. As KGs are developed to summarize large amounts of data, they can be great, easy-to-use tools for adding additional data and context to make ML workflows. For example, features of a given dataset can be expanded to include additional information per sample based on what we know about a given feature. In Li et al. [50], the authors collected data on self-reported student anxiety levels as well as basic information such as age, gender, grade, and home address. They then used the ‘Own-Think KG’ (see

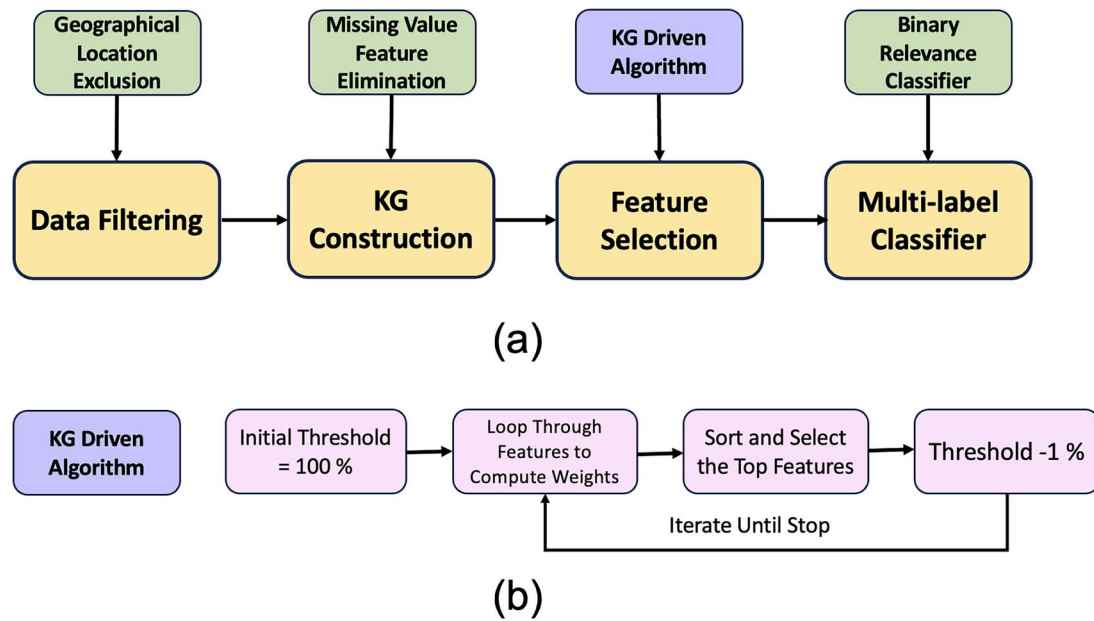


Figure 5. Illustration of interrelated FS procedure. (1) In the data filtering step as shown in part (a), states lacking lung cancer cases are excluded after referencing previous surveys spanning several years. (2) Features with over 50% missing values are eliminated. Then a KG is constructed from the remaining features. (3) A KG driven algorithm is used to transform the health survey question list to a data set with significantly interrelated features. (4) Finally, a binary relevance classifier (a special case of multi-label classifier) is proposed to predict the likelihood of multiple diseases by identifying one-to-many cancer relationship. In part (b), the KG driven algorithm starts with the initial threshold 100%. Then it loops through existing features and computes weights for each (features with more edges will get more weights). By sorting the weights, the features with highest weights are kept and the threshold is subtracted by 1%. The algorithm is iterated until the stopping criterion is met.

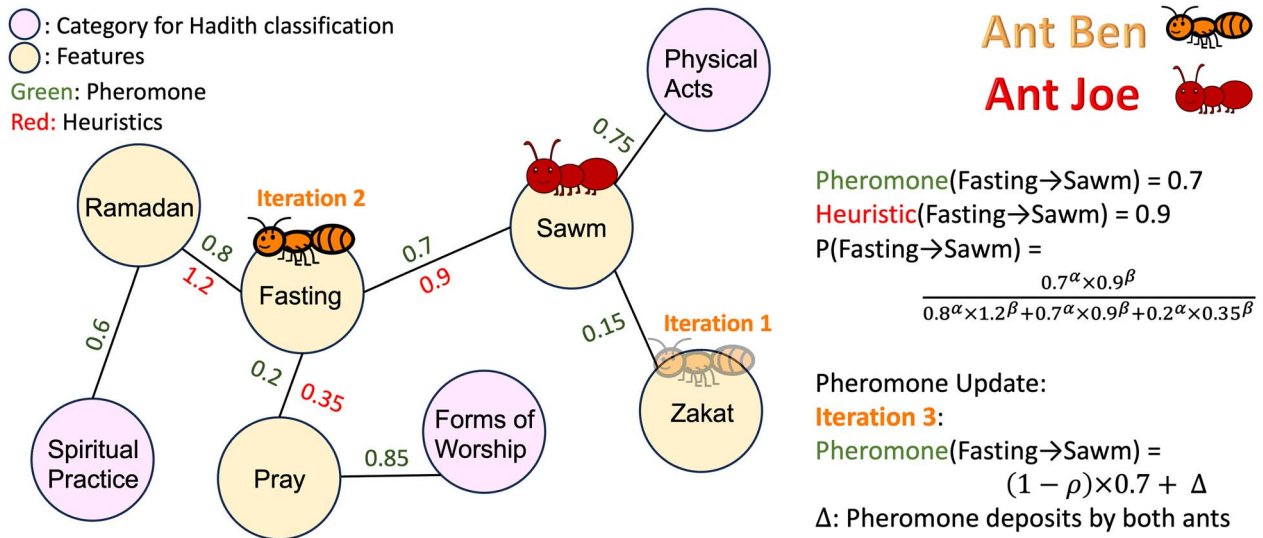


Figure 6. A demonstration of simplified ACO FS on Hadith Corpus KG. Here, two ants named Ben and Joe traverse the KG, with Ben starting at the 'Zakat' node and moving to 'Fasting' across iterations, and Joe beginning his journey at a randomly selected node 'Sawm'. The pheromone and heuristic values, represented by the green and red numbers above and below the edges, are aggregated outcomes of the explorations conducted by all ants in the system. Parameters  $\alpha$  and  $\beta$  determine the relative influence of pheromone trails and heuristic information respectively, while the evaporation rate  $\rho$  ensures flexibility in pathfinding, preventing premature convergence on suboptimal routes. The collective pheromone deposit  $\Delta$  between 'Fasting' and 'Sawm' by Ben and Joe is a cumulative measure reflecting the alignment of the Hadith content with specific categories, denoted by the pink nodes. The probability that Ben chooses 'Sawm' as the next feature is computed as a normalized version of Pheromone $^\alpha \times$  Heuristics $^\beta$  (see the middle right of the figure). In this instance, the focus is to reinforce the linkage between fasting-related Hadiths and the 'Physical Acts' category, differentiating it from the 'Spiritual Practice' category and the 'Forms of Worship' category, which are more aligned with spiritual benefits and devotional acts.

Figure 7), as well as 'DBpedia,' both known for their credibility and encyclopedic nature, to identify other features for their analysis based on the home address, including weather, population size, and GDP at both the district and regional area levels (see Figure 8). These KGs follow a clear and explainable

three-tuple storage structure, consisting of entities, attributes, and values, making them suitable for non-numerical feature generation. Importantly, they offer online querying capabilities, eliminating the need to download extensive datasets [51].

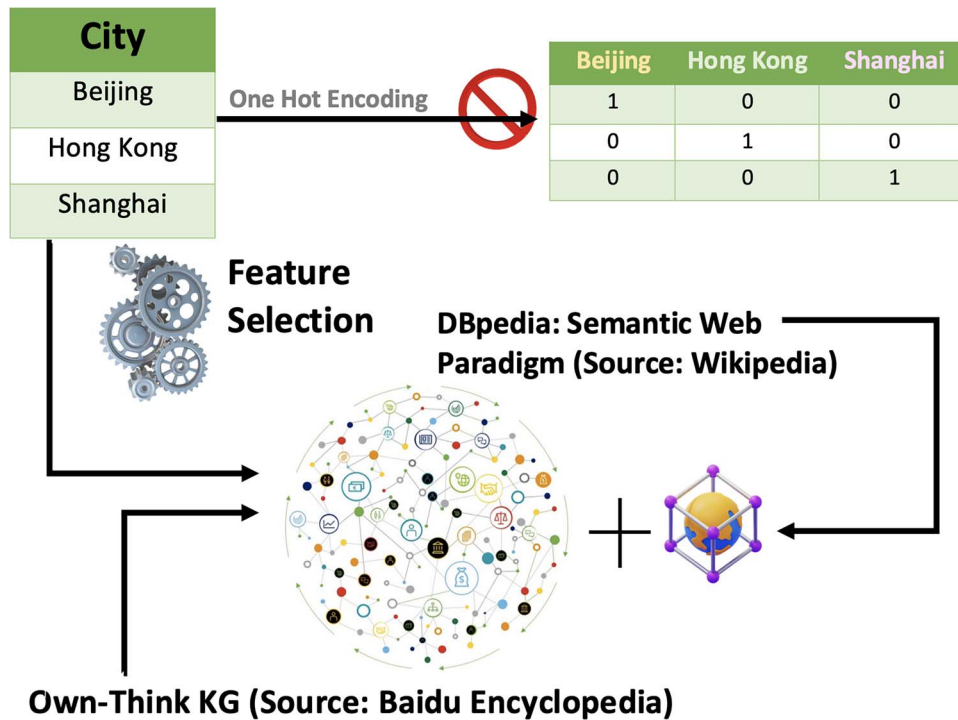


Figure 7. Own-think KG advantage over tradition one-hot encoding. Consider a dataset that includes information about various cities, like Beijing, Shanghai, and Hong Kong, where each city is represented by non-numerical discrete features such as its name. In a traditional dataset, the city name might be converted into a numerical form using techniques like one-hot encoding. However, this process strips the city's name of any contextual information about the city itself. Using a KG like the Own-Think KG, we can query additional information about each city to enrich the features, such as geography, economy, demography, culture, and so on, to enrich the features.

## Knowledge graph embeddings-vector embeddings

The embedding-focused approach in FS, exemplified by methods like DistMult [52], ComplEX [53], TransE [54], RESCAL [55], and FeaBI [56], RippleNet [57] seeks to represent nodes in a continuous vector space that captures deep semantic relationships and properties. This is a similar concept to word embeddings. Whereas in word embeddings, similar vectors capture similar semantic meaning, with similar words having similar representation, graph node embeddings capture relationship similarity within the graph network. The approach is popular for various applications, including link prediction [58] and entity classification [59]. Link prediction serves several purposes, from selecting movies a user would be interested in, to predicting drug-target interactions. Several methods have been developed to leverage embeddings for recommendation algorithms.

*Embedding via DistMult:*

1. The DistMult method, designed to predict missing relationships or facts within a KG [60], embeds entities and their interactions as vectors, inherently performing FS by
  - *Capturing Semantic Similarities:* Entities with closer interactional kinship within the KG are embedded proximately, emphasizing features underlying these semantic similarities.
  - *Highlighting Relevant Interactions:* DistMult accentuates features defining the interactions, such as biological pathways or chemical properties relevant to the interaction.
2. *Optimization of Feature Representation:* the DistMult training process fine-tunes the entity and relation representations

in the vector space, adjusting the significance of various attributes to enhance model accuracy.

- One relatively simple strategy for edge prediction is to first create embeddings for each node and then to train a classification algorithm to predict whether or not a connection exists between two nodes given their embeddings. For example, Wang et al. [61] utilized this strategy to predict drug-target interactions. In that study, the authors created node embeddings from a KG that contained known drug-target interactions. Next, they trained a deep learning model that took in a pair of embeddings (one drug and one target) to predict whether or not this pair was an existing edge in the graph. The authors showed that the model was able to identify some known interactions that were removed from the training set.
- A unique example comes from Wang et al., who proposed a hybrid KG embedding and path-based method in a recommendation algorithm they named RippleNet [57]. In this context, the KG contains nodes representing items that can be recommended, e.g. movies, along with other nodes that represent other features associated with each item, such as actors, genres, and release date. Edges highlight associations between items and features, e.g. a movie and its actors. In addition, there is a separate matrix that contains the interactions between each user and item. The goal is to predict the likelihood of a user selecting an item given the KG and the user's prior interactions. The algorithm begins by initializing the representation of each item based on the user's click history. Next, the algorithm iterates over items that



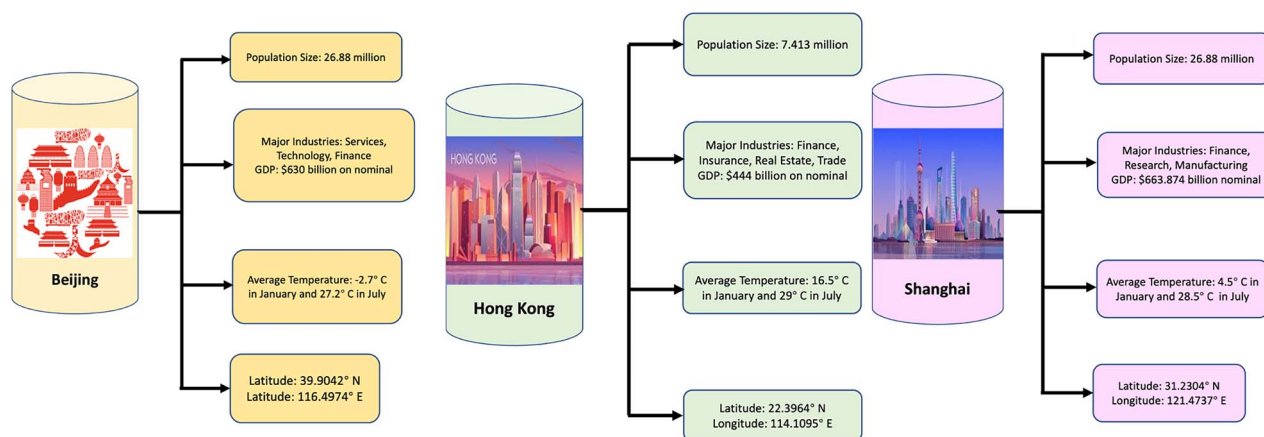


Figure 8. Demonstration of non-numeric discrete features enrichment and selection by Own-think KG. The figure includes enriched information for Beijing, Hong Kong, and Shanghai. For example, the additional features for Shanghai provided by the Own-Think KG (see Figure 7) detail Shanghai's population size, average temperature, latitude, longitude, and GDP. This thus contributes to a richer, more nuanced profile of Shanghai, compared to a one-hot encoding representation of each city, and offers additional insights as to how each aspect of a city may relate to the analysis at hand.

are increasing hops from items the user had already interacted with. The end result is an embedding of the relevance of each item that is combined with the initial vector representation with a model for the final prediction of the likelihood of selecting that item. This was later extended by Wang et al. [62] by having a combined deep framework that is simultaneously trained on a KG embedding task in addition to learning the recommendation task. The model architecture features shared latent features between the two tasks, with the idea being that the inclusion of the embedding task will enhance the latent representations. We give an illustration of Ripp-MKR in Fig. 9.

- Ismaeil et al. [56] introduced a method, FeaBI, to generate interpretable KG entity embeddings. First, a standard KG embedding is calculated. Additionally, a few categories of features for each node are extracted to form a vector, including the types of edges or relations it has, the types of nodes it is connected to, sequences of edge types of a certain length, and graph structural statistics. Next, random forest models are trained to predict each of the original embedding dimensions from its extracted feature vector. The random forest model ranks features based on their importance for the reconstruction task. These rankings can be used to better understand the information captured by embeddings. Additionally, a smaller subset of the feature vector can be selected for the most important features and used in place of the original embedding for more interpretable analysis.

## Deep learning-advanced network representation learning

Deep Learning models are designed to capture high-level, abstract representations of data. This ability allows them to capture meaningful insights from KGs, thereby enhancing applications in various domains, including personalized recommendations and predictive healthcare analytics.

- Anelli et al. [63] proposes KGFlex, a recommendation system [64] that integrates KG-based FS to improve the personalization and accuracy of recommendations. They use the notion

of multi-hop predicates [65] (i.e. considering chains of predicates that connect two entities at a high depth) to construct the semantic features on a KG. For instance,  $A \rightarrow B \rightarrow C$  is a 2-hop predicate.

In the FS step, KGFlex utilizes the concepts of entropy and information gain [66, 67] to assess how significant and relevant a feature is to a user when determining whether to engage with an item or not, i.e. to watch a movie or not. The features, represented as (predicate,entity) pairs, are then embedded in a latent space to construct the user-item interaction along with user embeddings via DL methods. For a particular user, the items with higher user-item interactions are recommended. All the embeddings and model parameters in KGFlex are learned from the Bayesian Personalized Ranking (BPR) optimization criterion [68]. The whole procedure is visualized in Fig. 10.

The performance of KGFlex is evaluated on three datasets from various domains, *Yahoo! Movies*, *MovieLens*, and *Facebook Books*. The experiments are designed to test the performance of KGFlex in terms of the Gini Index [69, 70]. KGFlex outperforms certain latent factor models such as kaHFM [71], ItemKNN [72], NeuMF [73], and BPR-MF [68] by an average of 18%. It also surpasses other key metrics, such as Item Coverage [74], in the recommendations it generates. Additionally, it excels in metrics like ACLT [75], PopREO, and PopRSP [76], which measure recommendation performance concerning the underrepresentation of rare items. It is occasionally outperformed only by kaHFM in top-10 recommendations.

- Su et al. [77] presented an attention-based KG representation learning framework, named DDKG, which aimed at feature representation and selection to improve drug-drug interaction (DDI) prediction. This approach allows for end-to-end prediction of DDIs. We summarize the DDKG into the below four main parts:

- a. *KG Construction*: the KG construction amalgamates the Simplified Molecular Input Line Entry System (SMILES), SMILES-associated triple facts, and entities such as proteins and diseases. For example, we have two drugs, A and B, and we integrate their SMILES sequences alongside their relationships (e.g. 'targets') with diseases into the KG.

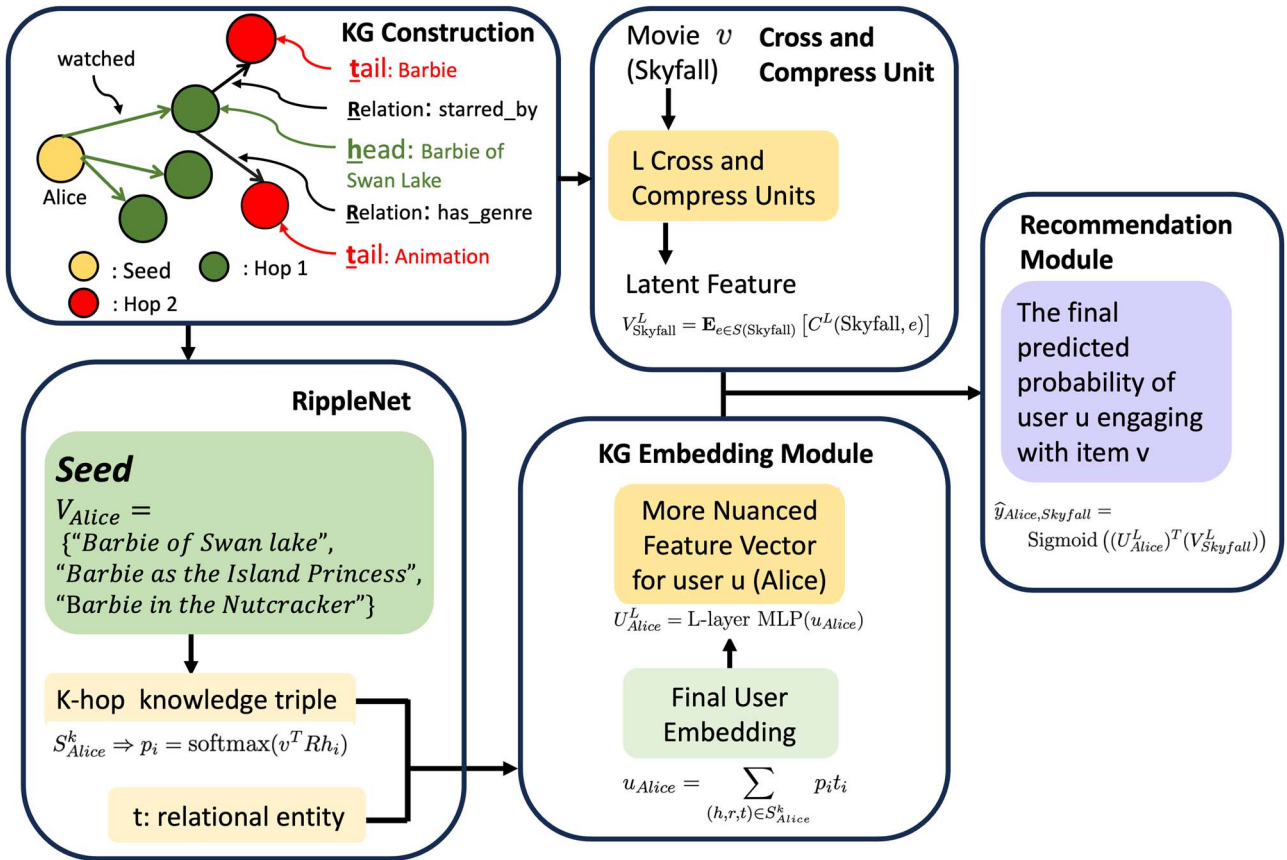


Figure 9. Illustration of Ripp-MKR feature learning mechanisms. The Ripp-MKR model involves a recommendation system KG with nodes representing users, movies, genres, and actors. In this KG, relationships such as 'Alice watched Barbie of Swan Lake,' 'Barbie of Swan Lake is starred by Barbie,' and 'Barbie of Swan Lake has genre Animation' are examples of how the system is structured (see **KG Construction**). Taking Alice as the initial point, we construct a historical set,  $V_{Alice}$ , comprising Alice's movie-watching history, which includes three movies (see **RippleNet**). RippleNet then extends Alice's preference for the Barbie series to other movies with similar genres and actors, like 'Barbie.' The **KG Embedding Module** (KGE) refines Alice's embedding,  $u_{Alice}$ , by aggregating all the  $k$ -hop softmax-weighted tail embeddings  $t_i$ , for instance, 'Animation' and 'Barbie' (see **KG Construction**). This refined embedding,  $u_{Alice}$ , is processed through an  $L$ -layer MLP to derive a nuanced user vector,  $U_{Alice}^L$ . The KGE is informed by the interactions among movies, genres, and actors. The **Cross and Compress Unit** examines the interactions between different genres by calculating the outer product of the movie vector  $v$  (e.g. 'Skyfall') and an entity vector  $e$  from the set  $S_{Skyfall}$ , which includes entities related to 'Skyfall' in the KG. After performing the outer product between  $v$  and each  $e \in S_{Skyfall}$   $L$  times, the final latent feature vector,  $V_{Skyfall}^L$ , for 'Skyfall' is obtained by taking the expectation over the  $L$  outer products. The **Recommendation Module** then selects the movie with the highest sigmoid probability from the inner product of  $U_{Alice}^L$  and  $V_{Skyfall}^L$ , denoted by  $\hat{y}_{Alice, Skyfall}$ . From potential next movies like 'Skyfall,' 'Inception,' and 'Barbie: Fairytopia,' Ripp-MKR recommends 'Barbie: Fairytopia' to Alice as it has the highest probability value, indicating it as the most suitable next watch.

- b. *Drug Embedding Initialization:* DDKG uses an encoder-decoder layer to learn the initial embeddings of drug nodes, mainly from the SMILES sequences in the KG. This step transforms the SMILES sequences of drugs A and B into vector representations that capture their chemical structure and properties.
- c. *Drug Representation Learning:* this part, consisting of three elements, serves as the key FS step in DDKG.
  - *Neighborhood Sampling:* for each drug node, a fixed-size set of neighboring nodes is selected. The significance of each neighbor is determined by *attention weights*, which are calculated based on the embeddings of the nodes and the types of relationships among them. This step ensures only the most relevant neighbors (in terms of both graph structure and drug relationships) are considered for further computation.
  - *Information Propagation:* The next step involves calculating a weighted sum of the neighbor embeddings. The attention weights (calculated in

the previous step) are used to determine how much each neighbor's information should contribute to the drug node's new representation. This ensures that more relevant neighbors have a bigger impact on the final representation.

- *Information Aggregation:* in the final step, the weighted sum of the neighbor embeddings is combined with the drug node's initial embedding and a final global representation of a drug node is obtained.

- d. *DDI Prediction:* for a queried pair of drugs, DDKG estimates their interaction probability by simply multiplying their final respective representations derived in c.

- In the work by Hsieh et al. [78], a GNN [79] was employed to advance the FS (drug selection) process for COVID-19 treatment from a drug-target interaction network (see Figure 11). The authors first constructed a COVID-19 KG (see the top-left region in Fig. 11) and generated embeddings using a GNN. The method involved transferring knowledge

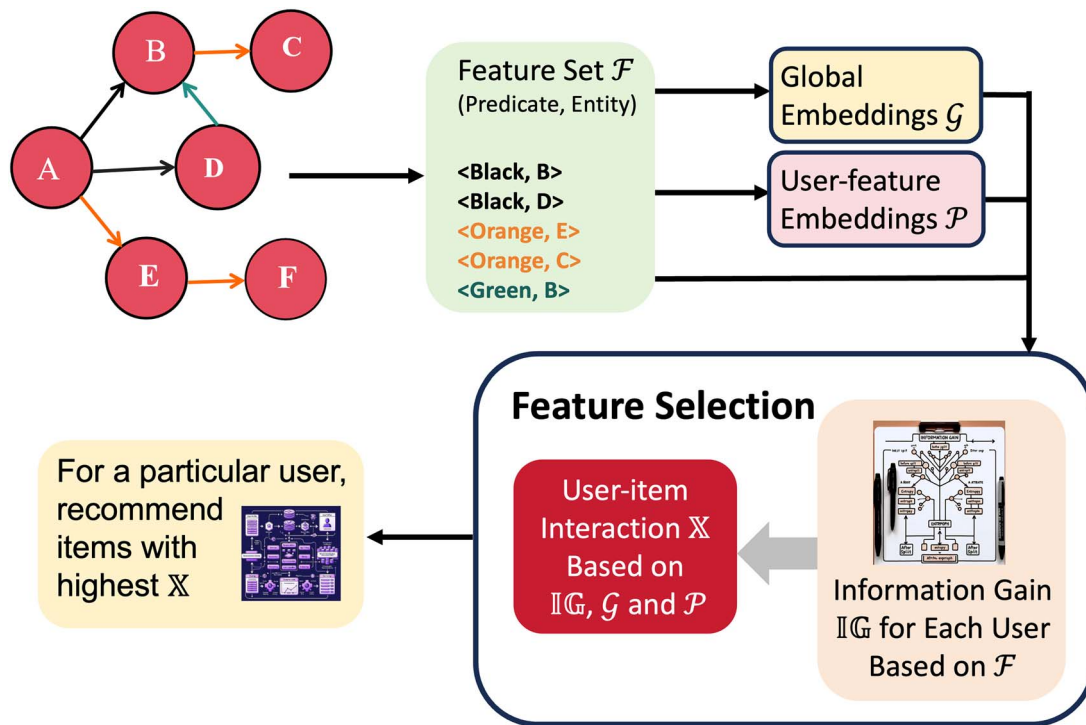


Figure 10. Illustration of KGFlex FS and recommendation procedure. We start with a KG with six nodes and six predicates (edges/relations). A feature set  $\mathcal{F}$  is constructed where each element is of form (predicate, node); for instance, from node A we can get to node B via a black predicate, then a feature is constructed as (Black, B). We construct a global embedding set  $\mathcal{G}$  representing each feature in  $\mathcal{F}$ , and a user-feature embedding set  $\mathcal{P}$  for each pair of user and feature. All embeddings and parameters in KGFlex are learned via DL methods with the BPR optimization criterion. We then associate each user-feature pair with an information gain  $\mathbb{IG}$ , which measures the expected reduction in information entropy from a prior node to a new node that acquires some information. For instance, suppose a user is currently at node A. The computed information gain  $\mathbb{IG}((\text{Black}, B))=1$ ,  $\mathbb{IG}((\text{Orange}, E))=0$  and  $\mathbb{IG}((\text{Black}, D))=1$  means the nodes B, D and the predicate ‘Black’ have influential impacts on the user’s next move. Finally, for each user, we compute the user-item interaction  $\mathbb{X}$  and recommend items to him with the highest  $\mathbb{X}$  values.

from another drug-repurposing KG (see top-right region) and learning high-dimensional embeddings for drugs that encapsulate the desired set of complex pharmacological characteristics of drugs (see middle region). By utilizing a ranking model informed by Bayesian pairwise ranking loss, this approach prioritizes potential drug candidates for downstream tasks such as gene set enrichment analysis (see middle-left region), and retrospective in vitro drug screening (see middle-right region). The top 22 most-promising drugs including aspirin, acetaminophen, and teicoplanin that are highlighted in the paper, demonstrate the rapid identification of candidate drugs for COVID-19 treatment.

## Comparative analysis of different approaches

Next, we will evaluate the methodologies from referenced manuscripts, focusing on their advantages and disadvantages. We have summarized this information in Table 1.

**1. Search Algorithms** Used in the Hadith Corpus KG [16] with the ACO algorithm and in COPD diagnosis [37] with the CMFS- $\eta$  algorithm. These methods highlight the importance of selecting appropriate strategies based on specific dataset requirements.

**2. Vector Embeddings** This approach, exemplified by the DistMult Algorithm and FeaBI, moves away from explicit path searches to embedding entities in a continuous vector space. It captures deep semantic relationships, facilitating the identification of intricate patterns relevant to complex domains like drug discovery [41, 44].

**3. Similarity-based Methods** These methods compare entities within a graph to identify similarities using metrics like cosine

similarity or Jaccard index. They are beneficial for clustering or recommendation systems, as demonstrated by Ma et al. [38] in Android malware classification and Jaworsky et al. [39] in health survey datasets.

**4. Advanced Network Representation Learning** Utilizes deep learning models to interpret and analyze KGs, capturing high-level data representations. Examples include KGFlex for optimizing recommendation systems and DDKG for drug-drug interaction predictions, showcasing the power of GNN frameworks in FS [37].

**Comparison and Contrast** Search algorithms and similarity-based methods provide direct, interpretable insights into KG structures, making them suitable for applications requiring clarity and precision. In contrast, vector embeddings and advanced network representation learning offer a nuanced understanding of data, identifying complex patterns and relationships. These latter methods are valuable for scenarios where data relationships are not straightforward, enabling flexible and powerful KG modeling for predictive analytics. The drug ranking technique by Hsieh et al. [78] demonstrates the intersection of vector embeddings and advanced network learning, highlighting their transformative potential in FS.

## Challenges and opportunities in KG FS

KGs are transforming data-driven fields like biomedical research, bioinformatics, and recommendation systems. They offer significant analytical capabilities but also present challenges and opportunities, especially in FS for ML models.

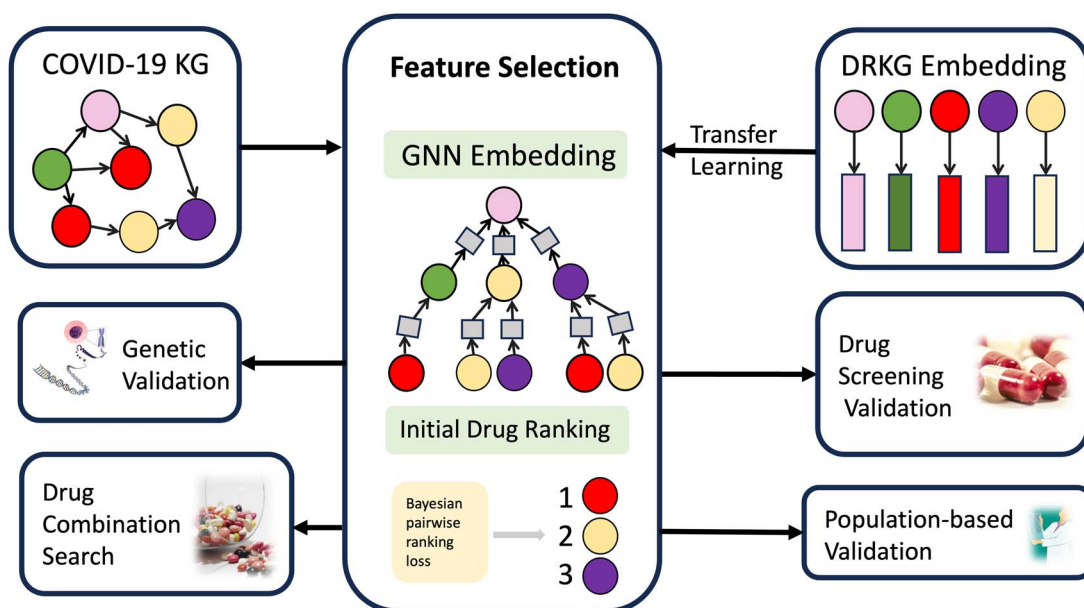


Figure 11. Feature selection (drug selection) via GNN embedding and drug ranking. The authors first constructed a **COVID-19 KG** containing different types of nodes (including 3,635 drugs) and interactions. The variational graph autoencoder with GraphSAGE messages passing [80, 81], a specific type of GNN, was used to derive the drug embedding (the grey squares in **Feature Selection**) by transferring a drug repurposing KG [82] to boost the representativeness. Initial drug ranking using Bayesian pairwise ranking loss was then applied to rank and select possibly potent drugs out of all candidates, hence serving as a FS step. The model efficacy was demonstrated using different validations. For instance, the authors performed **Genetic Validation** by identifying significant associations between SARS-CoV-2 and selected drugs. **Drug Screening Validation** is also performed by retrospectively comparing selected drugs with effective drugs in various in vitro drug screening experiments. In the **Population-based Validation**, the proposed method identified six drugs administered to the COVID-19 patients out of ten positive drugs that were effective in the electronic health records. In addition, **Drug Combination Search** for improving the COVID-19 treatment efficacy is conducted on the selected drugs. All validation results testify the capability of the proposed method speeding up the discovery of candidate drugs for treating COVID-19.

Table 1. Comparison of Feature Selection Methods for KGs

Method	Computational Complexity	Scalability	Practical Applicability	Pros and Cons
Search Algorithms	High	Limited	High (well-defined domains)	<b>Pros:</b> Efficient and precise in known domains. Straightforward implementation. <b>Cons:</b> May miss novel connections. Less adaptive to new patterns.
Similarity-based Methods	Moderate	Manageable	High (clustering, recommendation systems)	<b>Pros:</b> Easy to understand. Efficient for clustering/recommendations. <b>Cons:</b> Reliant on similarity metric quality. Computational challenges with large KGs.
Vector Embeddings	High (training), Low (inference)	High	Extensive (link prediction, drug discovery)	<b>Pros:</b> Captures deep semantic relationships. Scalable to large KGs. Enhances predictive power. <b>Cons:</b> Challenges in interpretability. High initial training cost.
Advanced Network Representation Learning	Significant	Challenging	High (complex pattern recognition)	<b>Pros:</b> Learns complex representations. Integrates heterogeneous data. Versatile in application. <b>Cons:</b> Computationally intensive. Complex model structure.

## Challenges

Feature selection in KGs faces several hurdles:

1. **High Dimensionality and Complexity:** KGs encompass numerous entities and relationships, creating high-dimensional spaces that challenge traditional FS methods.
2. **Data Heterogeneity:** KGs integrate diverse data types (numerical, categorical, textual) from various sources, necessitating robust FS techniques.
3. **Interpretability:** enhancing interpretability is crucial, especially in fields like healthcare, where understanding why features are selected is essential.

## Future directions

Several promising research avenues could redefine KG FS:

- **Causal Inference Techniques:** applying causal inference techniques to KGs can refine FS strategies [31].
- **Embedding KGs into Feature Matrices:** creating feature matrices from KGs facilitates downstream tasks and enhances model performance [83].
- **Novel Algorithms:** exploring the use of algorithms like ACO would introduce new approaches to FS within KGs [41, 44].
- **Multi-objective Optimization:** using multi-objective optimization techniques would offer a refined methodology



for FS, balancing criteria like redundancy and relevance [84].

- **Interdisciplinary Integration:** combining KGs with quantum computing, reinforcement learning (RL), and federated learning (FL) can enhance FS. Quantum-enhanced selection addresses scalability, RL refines the process based on feedback, and FL enables decentralized selection, preserving privacy [85, 86].
- **Semantic Enrichment and XAI:** leveraging the semantic information in KGs and applying Explainable AI principles can improve FS and model interpretability. Incorporating XAI principles into FS for KGs can be achieved through various methods, including attention mechanisms, interpretable models, and visualization techniques. Attention mechanisms in models such as Graph Attention Networks (GATs) allow for the identification of important features by assigning different weights to different parts of the input data, making it easier to understand which features significantly impact the model's predictions. Interpretable models, such as decision trees or rule-based systems, can be employed to provide clear and understandable decision paths that explain why certain features were selected. Additionally, visualization techniques, such as heatmaps and graph visualizations, can help users intuitively understand the relationships and significance of different features within the KG. These methods not only enhance the transparency of the FS process but also build trust in the model's predictions by providing insights into its underlying decision-making process. However, integrating XAI principles into KG FS comes with challenges, including ensuring scalability and maintaining interpretability in complex models. Scalability issues arise as the size and complexity of KGs increase, and necessitate efficient algorithms that can handle large datasets without compromising interpretability. Balancing model complexity with the need for transparency is crucial, as overly complex models may offer better performance but at the cost of reduced interpretability.
- **Domain Knowledge Integration:** integrating domain-specific knowledge into the FS process results in more effective selections, particularly in specialized fields like genomics and pharmacology.
- **Multi-modal Data Fusion:** combining various data sources into KGs offers a holistic view and unlocks new insights and applications.
- **Dynamic KGs and Real-time Feature Selection:** developing methods for real-time FS as KGs evolve can lead to more agile models, critical in rapidly changing domains like social media analysis.
- **Collaborative KG Frameworks:** creating frameworks for sharing and integrating KGs can enhance feature diversity and quality, fostering standardized protocols and benchmarks.
- **Ethical Considerations:** prioritizing ethical considerations and bias mitigation in KG FS ensures fairness and equity in applications. KGs can inherit biases from their data sources, leading to skewed outcomes. Addressing these biases requires diverse datasets and fairness-aware algorithms. Privacy is crucial, especially in sensitive domains like healthcare, necessitating robust data anonymization techniques and secure methods such as differential privacy and encryption. Ethical implications include the need for transparency and accountability in decision-making, especially in healthcare, where explainable AI principles and regulatory frameworks can prevent data misuse and discrimination. This expanded discussion ensures a responsible approach to KG FS.

## Conclusion

Examining KG methodologies underscores the importance of scalability, accuracy, and interpretability in FS processes. As KGs grow, developing scalable algorithms that efficiently process large-scale KGs without losing information granularity is paramount. This requires a balanced approach that leverages KGs' rich semantic relationships while addressing computational challenges.

### Key Points

- Emphasizes combining feature selection techniques with KGs to enhance predictive modeling in biomedical research.
- Shows significant applications in bioinformatics, improving disease prediction and drug discovery processes.
- Discusses challenges like computational complexity and the need for comprehensive KGs, proposing future research to develop efficient algorithms and integrate additional data sources.

## Acknowledgements

We would like to thank Alixanna M. Norris for her editorial support and Elvis Han Cui for his aesthetic suggestions and valuable insights on the theoretical papers reviewed.

## Funding

This work was funded by the National Institutes of Health (NIH) [U01AG066833].

## References

1. Chicaiza J, Valdiviezo-Diaz P. A comprehensive survey of knowledge graph-based recommender systems: technologies, development, and contributions. *Information* 2021;**12**:232. <https://doi.org/10.3390/info12060232>.
2. Belleau F, Nolin M-A, Tourigny N. et al. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform* 2008;**41**:706–16. <https://doi.org/10.1016/j.jbi.2008.03.004>.
3. Shamimul Hasan SM, Rivera D, Xiao-Cheng W. et al. Knowledge graph-enabled cancer data analytics. *IEEE J Biomed Health Inform* 2020;**24**:1952–67. <https://doi.org/10.1109/JBHI.2020.2990797>.
4. Fensel D, Şimşek U, Angele K. et al. Introduction: what is a knowledge graph? In: *Knowledge Graphs: Methodology, Tools and Selected Use Cases*, pp. 1–10. Charm: Springer, 2020. [https://doi.org/10.1007/978-3-030-37439-6\\_1](https://doi.org/10.1007/978-3-030-37439-6_1).
5. Bonner S, Barrett IP, Ye C. et al. Understanding the performance of knowledge graph embeddings in drug discovery. *Artif Intell Life Sci* 2022;**2**:100036. <https://doi.org/10.1016/j.aillsci.2022.100036>.
6. Yang Y, Yuwei L, Yan W. A comprehensive review on knowledge graphs for complex diseases. *Brief Bioinform* 2023;**24**:1–10. <https://doi.org/10.1093/bib/bbac543>.
7. Levine B, Kroemer G. Biological functions of autophagy genes: a disease perspective. *Cell* 2019;**176**:11–42. <https://doi.org/10.1016/j.cell.2018.09.048>.
8. Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. *Scientific Data* 2023;**10**:67. <https://doi.org/10.1038/s41597-023-01960-3>.
9. Bellman R. Dynamic programming. *Science* 1966;**153**:34–7. <https://doi.org/10.1126/science.153.3731.34>.



10. Ferreira AJ, Figueiredo MAT. Efficient feature selection filters for high-dimensional data. *Pattern Recogn Lett* 2012;**33**:1794–804. <https://doi.org/10.1016/j.patrec.2012.05.019>.
11. Lahmiri S. Features selection, data mining and financial risk classification: a comparative study. *Intell Syst Account Finance Manag* 2016;**23**:265–75. <https://doi.org/10.1002/isaf.1395>.
12. Huda S, Yearwood J, Jelinek HF. et al. A hybrid feature selection with ensemble classification for imbalanced healthcare data: a case study for brain tumor diagnosis. *IEEE Access* 2016;**4**:9145–54. <https://doi.org/10.1109/ACCESS.2016.2647238>.
13. Forster MR. Key concepts in model selection: Performance and generalizability. *J Math Psychol* 2000;**44**:205–31. <https://doi.org/10.1006/jmps.1999.1284>.
14. Saari P, Eerola T, Lartillot O. Generalizability and simplicity as criteria in feature selection: application to mood classification in music. *IEEE Trans Audio Speech Lang Process* 2010;**19**:1802–12.
15. Thulasi Bikku N, Rao S, Akepogu AR. Hadoop based feature selection and decision making models on big data. *Indian J Sci Technol* 2016;**9**:1–6. <https://doi.org/10.17485/ijst/2016/v9i10/88905>.
16. Mohammed B, Hamdan M, Bassi JS. et al. Edge computing intelligence using robust feature selection for network traffic classification in internet-of-things. *IEEE Access* 2020;**8**:224059–70. <https://doi.org/10.1109/ACCESS.2020.3037492>.
17. Jović A, Brkić K, Bogunović N. A review of feature selection methods with applications. In: *2015 38th international convention on information and communication technology, electronics and micro-electronics (MIPRO)*, pp. 1200–5. Piscataway, NJ: IEEE, 2015.
18. Staab S, Studer R. *Handbook on Ontologies*. Berlin, Heidelberg: Springer Science & Business Media, 2010. <https://doi.org/10.1007/978-3-540-92673-3>.
19. Wishart DS, Feunang YD, Guo AC. et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Res* 2018;**46**:D1074–82. <https://doi.org/10.1093/nar/gkx1037>.
20. Kuhn M, Letunic I, Jensen LJ. et al. The sider database of drugs and side effects. *Nucleic Acids Res* 2016;**44**:D1075–9. <https://doi.org/10.1093/nar/gkv1075>.
21. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000;**28**:27–30. <https://doi.org/10.1093/nar/28.1.27>.
22. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**:267D–270. <https://doi.org/10.1093/nar/gkh061>.
23. Donnelly K. et al. SNOMED-CT: the advanced terminology and coding system for ehealth. *Stud Health Technol Inform* 2006;**121**:279–90.
24. Nelson SJ, Zeng K, Kilbourne J. et al. Normalized names for clinical drugs: Rxnorm at 6 years. *J Am Med Inform Assoc* 2011;**18**:441–8. <https://doi.org/10.1136/amiajnl-2011-000116>.
25. Miller JJ. Graph database applications and concepts with neo4j. In: *Proceedings of the southern association for information systems conference, Atlanta, GA, USA, Vol. 2324*, pp. 141–7. Atlanta, GA: Southern Association for Information Systems, 2013.
26. Bollacker K, Evans C, Paritosh P. et al. Freebase: a collaboratively created graph database for structuring human knowledge. In: *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 1247–50. New York, NY: Association for Computing Machinery (ACM), 2008.
27. Francis N, Green A, Guagliardo P. et al. Cypher: an evolving query language for property graphs. In: *Proceedings of the 2018 international conference on management of data*, pp. 1433–45. New York, NY: Association for Computing Machinery (ACM), 2018.
28. Romano JD, Truong V, Kumar R. et al. The Alzheimer's knowledge base - a knowledge graph for therapeutic discovery in Alzheimer's disease research. *J Med Internet Res* 2024;**26**. <https://doi.org/10.2196/46777>.
29. Geary TG, Mackenzie CD, Silber SA. Flubendazole as a macrofilaricide: history and background. *PLoS Negl Trop Dis* 2019;**13**:e0006436. <https://doi.org/10.1371/journal.pntd.0006436>.
30. Zanga A, Ozkirimli E, Stella F. A survey on causal discovery: theory and practice. *International Journal of Approximate Reasoning* 2022;**151**:101–29. <https://doi.org/10.1016/j.ijar.2022.09.004>.
31. Malec SA, Taneja SB, Albert SM. et al. Causal feature selection using a knowledge graph combining structured knowledge from the biomedical literature and ontologies: a use case studying depression as a risk factor for Alzheimer's disease. *J Biomed Inform* 2023;**142**:104368. <https://doi.org/10.1016/j.jbi.2023.104368>.
32. Callahan TJ, Tripodi IJ, Stefanski AL. et al. An open source knowledge graph ecosystem for the life sciences. *Sci Data* 2024;**11**:363. <https://doi.org/10.1038/s41597-024-03171-w>.
33. Gyori BM, Bachman JA, Subramanian K. et al. From word models to executable models of signaling networks using automated assembly. *Mol Syst Biol* 2017;**13**:954–80. <https://doi.org/10.15252/msb.20177651>.
34. Pérez J, Arenas M, Gutierrez C. Semantics and complexity of SPARQL. *ACM Trans Database Syst (TODS)* 2009;**34**:1–45. <https://doi.org/10.1145/1567274.1567278>.
35. Horrocks I, Patel-Schneider PF, Bechhofer S. et al. Owl rules: a proposal and prototype implementation. *J Web Semant* 2005;**3**:23–40. <https://doi.org/10.1016/j.websem.2005.05.003>.
36. DuCharme B. *Learning SPARQL: Querying and Updating with SPARQL 1.1*, 1, 38, 41. Sebastopol, CA: O'Reilly Media, Inc., 2013. <https://doi.org/10.1089/big.2012.0004>.
37. Fang Y, Wang H, Wang L. et al. Diagnosis of COPD based on a knowledge graph and integrated model. *IEEE Access* 2019;**7**:46004–13. <https://doi.org/10.1109/ACCESS.2019.2909069>.
38. Ma D, Bai Y, Xing Z. et al. A knowledge graph-based sensitive feature selection for android malware classification. In: *2020 27th Asia-Pacific Software Engineering Conference (APSEC)*, pp. 188–197. Singapore: IEEE, 2020.
39. Jaworsky M, Tao X, Pan L. et al. Interrelated feature selection from health surveys using domain knowledge graph. *Health Inf Sci Syst* 2023;**11**:54. <https://doi.org/10.1007/s13755-023-00254-7>.
40. Pierannunzi C, Shaohua Sean H, Balluz L. A systematic review of publications assessing reliability and validity of the behavioral risk factor surveillance system (BRFSS), 2004–2011. *BMC Med Res Methodol* 2013;**13**:1–14. <https://doi.org/10.1186/1471-2288-13-49>.
41. Mosa MA. Feature selection based on ACO and knowledge graph for Arabic text classification. *J Exp Theor Artif Intell* 2022;**34**:1–18. <https://doi.org/10.2139/ssrn.4040689>.
42. Dorigo M, Blum C. Ant colony optimization theory: a survey. *Theoretical computer science* 2005;**344**:243–78. <https://doi.org/10.1016/j.tcs.2005.05.020>.
43. Dorigo M, Birattari M, Stützle T. Ant colony optimization. *IEEE Comput Intell Mag* 2006;**1**:28–39. <https://doi.org/10.1109/MCI.2006.329691>.
44. Dorigo M, Stützle T. Ant colony optimization: overview and recent advances. In: Gendreau M, Potvin JY (eds). *Handbook of Metaheuristics. International Series in Operations Research & Management Science*, vol 272. Cham: Springer, 2019. [https://doi.org/10.1007/978-3-319-91086-4\\_10](https://doi.org/10.1007/978-3-319-91086-4_10).
45. Parpinelli RS, Lopes HS, Freitas AA. Data mining with an ant colony optimization algorithm. *IEEE Trans Evol Comput* 2002;**6**:321–32. <https://doi.org/10.1109/TEVC.2002.802452>.
46. Martens D, De Backer M, Haesen R. et al. Classification with ant colony optimization. *IEEE Trans Evol Comput* 2007;**11**:651–65. <https://doi.org/10.1109/TEVC.2006.890229>.

47. Aghdam MH, Ghasem-Aghae N, Basiri ME. Text feature selection using ant colony optimization. *Exp Syst Appl* 2009;**36**: 6843–53. <https://doi.org/10.1016/j.eswa.2008.08.022>.
48. Onan A. SRL-ACO: a text augmentation framework based on semantic role labeling and ant colony optimization. *J King Saud Univ-Comput InfSci* 2023;**35**:101611. <https://doi.org/10.1016/j.jksuci.2023.101611>.
49. Chang H. Making sense of the big picture: data linkage and integration in the era of big data. *Healthc Inform Res* 2018;**24**: 251–2. <https://doi.org/10.4258/hir.2018.24.4.251>.
50. Li L, Yang H, Jiao Y. et al. Feature generation based on knowledge graph. *IFAC-PapersOnLine* 2020;**53**:774–9. <https://doi.org/10.1016/j.ifacol.2021.04.172>.
51. Auer S, Bizer C, Kobilarov G. et al. DBpedia: a nucleus for a web of open data. In: *International semantic web conference*, pp. 722–735. Berlin, Heidelberg: Springer, 2007. [https://doi.org/10.1007/978-3-540-76298-0\\_52](https://doi.org/10.1007/978-3-540-76298-0_52).
52. Yang B, Yih W-T, He X. et al. Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint arXiv:1412.6575. 2014. <https://doi.org/10.48550/arXiv.1412.6575>.
53. Trouillon T, Welbl J, Riedel S. et al. Complex embeddings for simple link prediction. In: *International conference on machine learning*, pp. 2071–80. New York, NY: PMLR, 2016.
54. Bordes A, Usunier N, Garcia-Duran A. et al. Translating embeddings for modeling multi-relational data. *Adv Neural Inf Process Syst* 2013;**26**:2787–95. <https://doi.org/10.48550/arXiv.1412.6575>.
55. Nickel M, Tresp V, Kriegel H-P. et al. A three-way model for collective learning on multi-relational data. In: *ICML*, Vol. **11**, pp. 809–16. New York, NY, USA: ACM Digital Library, 2011.
56. Ismaeil Y, Stepanova D, Tran T-K. et al. FeaBI: a feature selection-based framework for interpreting KG embeddings. In: *The Semantic Web – ISWC 2023: 22nd International Semantic Web Conference, Athens, Greece, November 6–10, 2023, Proceedings, Part*, pp. 599–617. Springer, Cham, 2023. [https://doi.org/10.1007/978-3-031-47240-4\\_32](https://doi.org/10.1007/978-3-031-47240-4_32).
57. Wang H, Zhang F, Wang J. et al. RippleNet: propagating user preferences on the knowledge graph for recommender systems. In: *Proceedings of the 27th ACM international conference on information and knowledge management*, pp. 417–26. New York, NY, USA: Association for Computing Machinery, 2018. <https://doi.org/10.1145/3269206.3271739>.
58. Kumar A, Singh SS, Singh K. et al. Link prediction techniques, applications, and performance: a survey. *Phys A: Stat Mech Appl* 2020;**553**:124289. <https://doi.org/10.1016/j.physa.2020.124289>.
59. Al-Moslemi T, Ocaña MG, Opdahl AL. et al. Named entity extraction for knowledge graphs: a literature overview. *IEEE Access* 2020;**8**:32862–81. <https://doi.org/10.1109/ACCESS.2020.2973928>.
60. Chen Z, Wang Y, Zhao B. et al. Knowledge graph completion: a review. *IEEE Access* 2020;**8**:192435–56. <https://doi.org/10.1109/ACCESS.2020.3030076>.
61. Wang S, Zhenzhen D, Ding M. et al. KG-DTI: a knowledge graph based deep learning method for drug-target interaction predictions and Alzheimer's disease drug repositions. *Appl Intell* 2022;**52**:846–57. <https://doi.org/10.1007/s10489-021-02454-8>.
62. Wang YQ, Dong LY, Li YL. et al. Multitask feature learning approach for knowledge graph enhanced recommendations with ripplenet. *PLoS One* 2021;**16**:e0251162. <https://doi.org/10.1371/journal.pone.0251162>.
63. Anelli VW, Di Noia T, Di Sciascio E. et al. Sparse feature factorization for recommender systems with knowledge graphs. In: *RecSys'21: Proceedings of the 15th ACM Conference on Recommender Systems*, pp. 154–65. New York, NY, USA: ACM, 2021. <https://doi.org/10.1145/3460231.3474243>.
64. Shani G, Gunawardana A. Evaluating recommendation systems. In: *Recommender Systems Handbook*, pp. 257–97. New York: Springer, 2011. [https://doi.org/10.1007/978-0-387-85820-3\\_8](https://doi.org/10.1007/978-0-387-85820-3_8).
65. Zhang Z, Wang J, Chen J. et al. Cone: Cone embeddings for multi-hop reasoning over knowledge graphs. *Advances in Neural Information Processing Systems* 2021;**34**:19172–83.
66. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J* 1948;**27**:379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>.
67. Rokach L, Maimon O. Top-down induction of decision trees classifiers—a survey. *IEEE Trans Syst Man Cybern C (Appl Rev)* 2005;**35**:476–87. <https://doi.org/10.1109/TSMCC.2004.843247>.
68. Rendle S, Freudenthaler C, Gantner Z. et al. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 452–61, ACM, 2009.
69. Gastwirth JL. The estimation of the Lorenz curve and Gini index. *Rev Econ Stat* 1972;**54**:306–16. <https://doi.org/10.2307/1937992>.
70. Castells P, Hurley NJ, Vargas S. Novelty and diversity in recommender systems. In: Ricci F, Rokach L, Shapira B (eds) *Recommender Systems Handbook*, pp. 603–46. Boston, MA: Springer, 2015. [https://doi.org/10.1007/978-1-4899-7637-6\\_26](https://doi.org/10.1007/978-1-4899-7637-6_26).
71. Anelli VW, Di Noia T, Di Sciascio E. et al. How to make latent factors interpretable by feeding factorization machines with knowledge graphs. In: Ghidini C, et al. (eds.) *The Semantic Web – ISWC 2019, Vol 11778*. Springer, Cham: Lecture Notes in Computer Science, 2019. [https://doi.org/10.1007/978-3-030-30793-6\\_3](https://doi.org/10.1007/978-3-030-30793-6_3).
72. Koren Y. Factor in the neighbors: scalable and accurate collaborative filtering. *ACM Trans Knowl Discov Data (TKDD)* 2010;**4**:1–24. <https://doi.org/10.1145/1644873.1644874>.
73. He X, Chua TS. Neural factorization machines for sparse predictive analytics. In: *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 355–64. Tokyo, Japan: ACM, 2017.
74. Adomavicius G, Kwon YO. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Trans Knowl Data Eng* 2011;**24**:896–911. <https://doi.org/10.1109/TKDE.2011.15>.
75. Abdollahpour H, Burke R, Mobasher B. Managing popularity bias in recommender systems with personalized re-ranking. *Proceedings of the Thirty-Second International Florida Artificial Intelligence Research Society Conference (FLAIRS 2019)*. Sarasota, Florida: AAAI Press, 2019.
76. Zhu Z, Wang J, Caverlee J. Measuring and mitigating item under-recommendation bias in personalized ranking systems. In: *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pp. 449–58. New York, NY, USA: ACM, 2020. The conference was held virtually from Xi'an, China, July 25–30, 2020.
77. Xiaorui S, Lun H, You Z. et al. Attention-based knowledge graph representation learning for predicting drug-drug interactions. *Brief Bioinform* 2022;**23**. <https://doi.org/10.1093/bib/bbac140>.
78. Hsieh K, Wang Y, Chen L. et al. Drug repurposing for Covid-19 using graph neural network and harmonizing multiple evidence. *Sci Rep* 2021;**11**:23179. <https://doi.org/10.1038/s41598-021-02353-5>.
79. Zhou J, Cui G, Shengding H. et al. Graph neural networks: a review of methods and applications. *AI Open* 2020;**1**:57–81. <https://doi.org/10.1016/j.aiopen.2021.01.001>.
80. Kipf TN, Welling M. Variational graph auto-encoders. arXiv preprint arXiv:1611.07308. 2016 Nov 21. Available at: <https://arxiv.org/abs/1611.07308>.

81. Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. *Adv Neural Inf Process Syst* 2017;**30**:1025–35.
82. Zeng X, Song X, Ma T. et al. Repurpose open data to discover therapeutics for Covid-19 using deep learning. *J Proteome Res* 2020;**19**:4624–36. <https://doi.org/10.1021/acs.jproteome.0c00316>.
83. Strande NT, Riggs ER, Buchanan AH. et al. Evaluating the clinical validity of gene-disease associations: an evidence-based framework developed by the clinical genome resource. *Am J Hum Genet* 2017;**100**:895–906. <https://doi.org/10.1016/j.ajhg.2017.04.015>.
84. Mouret J-B, Clune J. Illuminating search spaces by mapping elites. arXiv preprint arXiv:1504.04909. 2015. <https://arxiv.org/abs/1504.04909>.
85. Ma Y, Tresp V. Quantum machine learning algorithm for knowledge graphs. *ACM Trans Quant Comput* 2021;**2**:1–28. <https://doi.org/10.1145/3467982>.
86. Huang W, Liu J, Li T. et al. FedCKE: cross-domain knowledge graph embedding in federated learning. *IEEE Trans Big Data* 2022;**9**:792–804. <https://doi.org/10.1109/TBDATA.2022.3205705>.

## Appendix

### Appendix A. Table of Acronyms

Table A.1 lists the Table of Acronyms for this paper.

Table A.1. Table of Acronyms

Abbreviation	Definition
ACLT	Average Coverage of Long Tail items
ACO	Ant Colony Optimization
AD	Alzheimer's Disease
ADKG	Alzheimer's Disease Knowledge Graph
AI	Artificial Intelligence
AlzKb	Alzheimer's Disease Knowledge Base
APOE	Apolipoprotein E
AUC	Area Under the Curve
Bi-LSTM	Bidirectional Long Short-Term Memory
BPR	Bayesian Personalized Ranking
COPD	Chronic Obstructive Pulmonary Disease
CYP2D6	Cytochrome P450 2D6
DDI	Drug-Drug Interaction
DistMult	The Distributed Multinomial Method
DL	Deep Learning
DR	Dimensionality/Dimension Reduction
DSA-SVM	Direct Search Simulated Annealing with Support Vector Machine
DTP	Drug-Target Pairs
FS	Feature Selection
GDB	Graph Database
GNN	Graph Neural Network
HMOX1	Heme Oxygenase 1
KEGG	Kyoto Encyclopedia of Genes and Genomes
KG	Knowledge Graph
LDA	Linear Discriminant Analysis
LLE	Local Linear Embedding
ML	Machine Learning
MLP	Multiple Layer Perceptron
MQL	Metaweb Query Language
MTHFR	Methylenetetrahydrofolate Reductase
RDF	Resource Description Framework
RFE	Recursive Feature Elimination
nDCG	Normalized Discount Cumulative Gain
NLP	Natural Language Processing
NOS3	Nitric Oxide Synthase 3
OWL	The Web Ontology Language
PCA	Principal Component Analysis
PPARG	Peroxisome Proliferator-Activated Receptor Gamma

(Continued)

Table A.1. Continued

Abbreviation	Definition
RDF	Resource Description Framework
RFE	Recursive Feature Elimination
RO	Relation Ontology
TPI1	Triosephosphate Isomerase 1
URIs	Uniform Resource Identifiers
UMLS	Unified Medical Language System
W3C	World Wide Web Consortium
YAGO	Yet Another Great Ontology

### Appendix B. A more detailed description of KGs of sizes tiny, small, and medium

Within each of the three graphs (see Figs 2, A.1, and A2), the nodes and their connections are represented by distinct colors and arrow types to convey different biological relationships:

Orange (see Figs A.1 and A2) and Yellow (see Figure 2) nodes represent the disease entity, with AD positioned as the central node, highlighting it as the primary focus of this network.

Purple nodes signify genes, which are implicated in AD through various associations such as genetic risk factors, differential gene expression, or other genetic interactions.

Green nodes denote chemicals, encompassing drugs, vitamins, or other bioactive molecules. These external agents are potential modulators of gene function or disease pathology.

- There are five instances of the 'Chemical binds gene' relationship (light purple arrows in Fig. 2 and brown arrows in Figs A.1 and A2), where a chemical is shown to interact directly with a gene. This does not necessarily indicate an increase or decrease in gene expression, but rather a physical or functional interaction. For example, one of the edges indicates that folic acid, a form of vitamin B that is vital for making DNA and other genetic material, binds the MTHFR gene. MTHFR plays a crucial role in processing amino acids, the building blocks of proteins. Variants of this gene can affect homocysteine levels in the blood. Deficiencies in folic acid are linked to elevated homocysteine levels, which may increase the risk of AD.
- There are six instances of the 'Gene associates with disease' relationship (yellow arrows in Fig. 2 and red arrows in Figs A.1 and A2), representing genes that have an association with AD. These relationships might represent genetic risk factors, genes involved in the pathology of the disease, or genes that could be potential targets for therapeutic intervention. For instance, the NOS3 gene is associated with AD. It is involved in the generation of nitric oxide, a molecule that aids in blood vessel dilation. Impairment in NOS3 function can affect blood flow in the brain, potentially impacting Alzheimer's disease pathology.
- There are three instances of the 'Chemical increases expression' relationship (pink arrows), denoting chemicals that are known to upregulate or increase the expression of certain genes. For instance, vitamin A increases the expression of HMOX1, a gene that encodes an enzyme in response to oxidative stress, and that is also a contributing factor in neuronal damage observed in AD.
- There is one instance of the 'Chemical decreases expression' relationship (green arrow), indicating a chemical that down-regulates or decreases the expression of a gene. Namely, cyclosporine, an immunosuppressant that may inhibit the formation of the amyloid plaques, that are a hallmark of AD, decreases the expression of TPI1, an enzyme that plays a



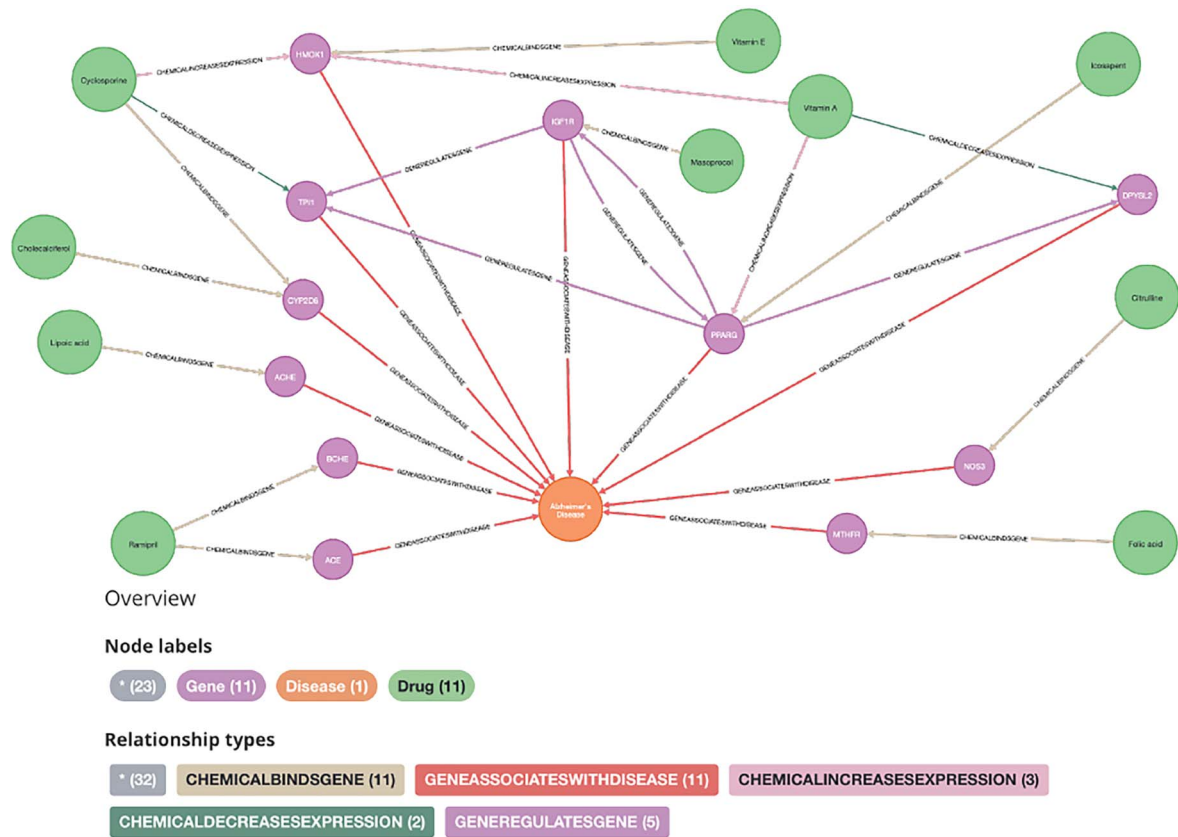


Figure A.1. A Small-sized ADKG (Orange Node: AD; Purple Nodes: Genes; Green Nodes: Drugs) [28].

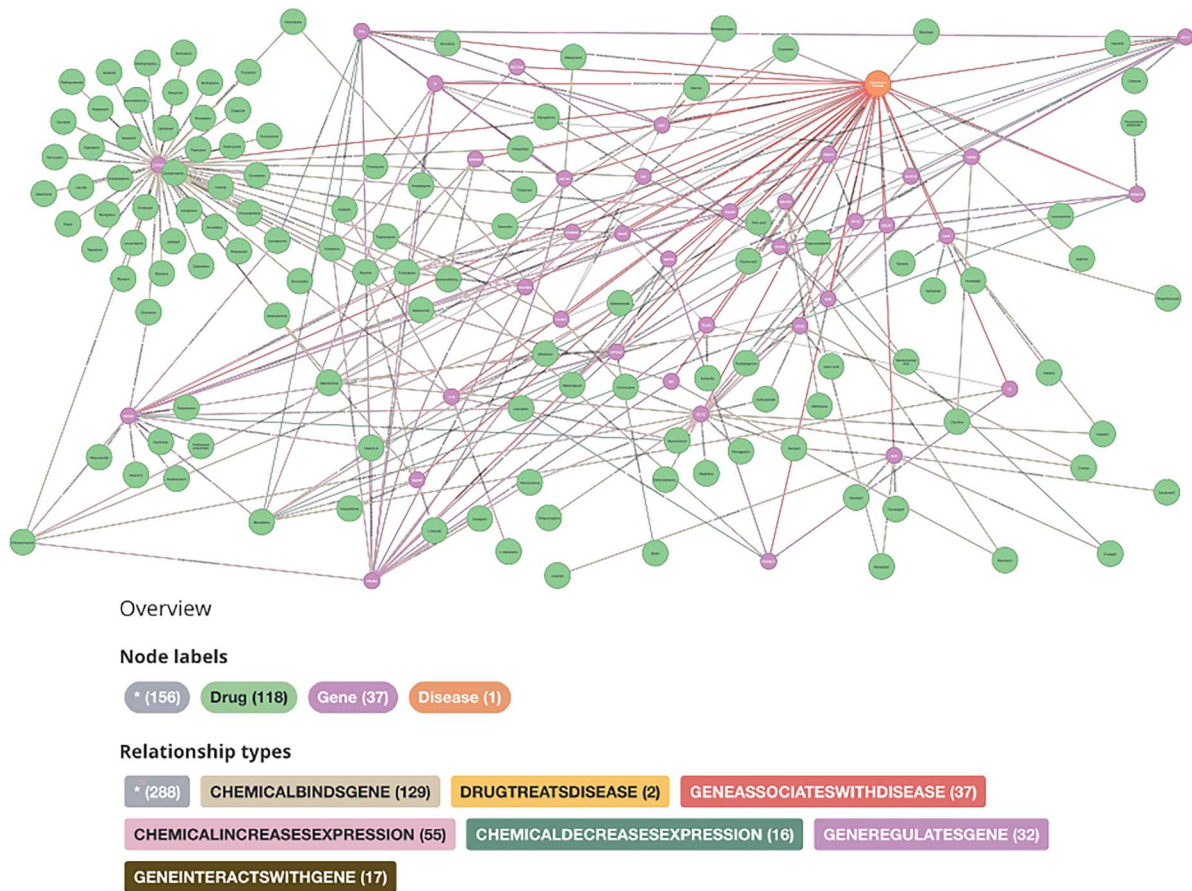


Figure A2. A Medium-sized ADKG (Orange Node: AD; Purple Nodes: Genes; Green Nodes: Drugs) [28].

crucial role in glycolysis, a metabolic pathway that occurs in the cytoplasm of cells.

- There is one instance of 'Gene regulates gene' (purple arrow), suggesting a regulatory interaction between two genes, PPARG and TPI1. For context, PPARG is a gene that codes for a protein that regulates fatty acid storage and glucose metabolism. It is a target for some drugs that might influence Alzheimer's disease progression.

Figure A.1 provides an example of a small-sized ADKG with 23 nodes and 32 edges (setting the Cypher limit clause to 15) and

Fig. A2 provides an example of a medium-sized ADKG with 156 nodes and 288 edges (setting the Cypher limit clause to 200). In addition to the relationship types described above, the medium-sized ADKG also demonstrates the 'DRUGTREATDISEASE' (gold arrows) and 'GENEINTERACTSWITHGENE' (brown arrows) relationships. As the size of the KGs continues to expand, the challenge of comprehending the intricate web of entities and relationships within them becomes daunting for human observers. Consequently, there arises an urgent need for the development of sophisticated computational tools capable of effectively managing these vast KGs.