Scientific Reports

https://doi.org/10.1038/s41598-026-42584-y

Article in Press

Ontology-driven association rule mining for biomedical entity relationships: integrating hierarchical knowledge to improve gene-disease discovery

Received: 8 November 2025

Accepted: 26 February 2026

Published online: 11 March 2026

Cite this article as: Naqash M.A., Amin M., Uddin J. *et al.* Ontology-driven association rule mining for biomedical entity relationships: integrating hierarchical knowledge to improve gene-disease discovery. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-42584-y

Mian Athar Naqash, Muhammad Amin, Jamal Uddin, Hany S. Hussein, Ali Raza, Wajdi Alghamdi, Hala AbdelHameed Mostafa & Hend Khalid Alkahtani

Ontology-Driven Association Rule Mining for Biomedical Entity Relationships: Integrating Hierarchical Knowledge to Improve Gene-Disease Discovery

Mian Athar Naqash[1], Muhammad Amin[1], Jamal Uddin[2], Hany S. Hussein[3, 4], Ali Raza[5, 6, 7,*], Wajdi Alghamdi[8] ,Hend Khalid Alkahtani[9 ,*], Hala AbdelHameed Mostafa[10,11]

[1]Department of Physical and Numerical Sciences, Qurtuba University of Science and Information Technology, Peshawar, Pakistan
athar.naqash@yahoo.com
mamin@uop.edu.pk
[2]Riphah School of Computing and Innovation, Riphah International University Islamabad (Lahore, Campus), Pakistan
jamuddin1983@gmail.com
[3]Electrical Engineering Department, College of Engineering, King Khalid University, Abha 62529, Saudi Arabia.
[4]Electrical Engineering Department, Faculty of Engineering, Aswan University, Aswan, 81542, Egypt
hany.hussein@aswu.edu.eg
[5]Department of Computer Science, Bahria University, Islamabad 44220, Pakistan
alirazahp1122@gmail.com
[6]School of Electronic and Communication Engineering, Shenzhen Polytechnic University, Shenzhen 518055, China
[7]School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China
[8]Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[9]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia
Hkalqahtani@pnu.edu.sa
[10]Faculty of Computer and Artificial Intelligence, Fayoum University, Fayoum 63514, Egypt.
Ham07@fayoum.edu.eg
[11]Applied College, Taibah University, Medina 42353, Saudi Arabia,
hammohamed@taibahu.edu.sa

* Correspondence: Hend Khalid Alkahtani Hkalqahtani@pnu.edu.sa, Ali Raza, alirazahp1122@gmail.com

Abstract

Reliable links between genes and diseases are central to biomedical research; however, many computational methods overlook the semantic and hierarchical layers of ontologies, missing indirect relationships and producing shallow association scores. We propose an ontology-driven framework for gene–disease association mining that integrates hierarchical knowledge from the Gene Ontology and Disease Ontology. Our text-mining pipeline processes PubMed text by cleaning, annotating, and extracting sentence-level co-occurrences of biomarker-related terms. We evaluated and compared well-known association rule mining algorithms, namely Apriori, FP-Growth, and Eclat, and applied a tie-aware rank-based transformation (RBT) to correct for non-normal distributions of association scores. The resulting Athar Semantic Enriched Association (ASEA) score combines entity-specific associations (ESA) with Hierarchical Ontology Associations (HOA), with an enhanced Apriori variant showing superior performance in capturing direct and indirect associations. Benchmarking against the Comparative Toxicogenomics Database (CTD), ASEA detected 17 high-grade associations (30.4% more

than Apriori and Eclat, 88.9% more than FP-Growth). In total, ASEA produced 185 associations, compared with 217 for Apriori, 166 for Eclat, and 71 for FP-Growth. Among these, 21 belong to high-confidence databases (Case 1), 28 are supported by substantial literature, but not yet high-confidence (Case 2), 39 have low/intermediate database support with no strong literature (Case 3), and 22 are purely speculative (Case 4), including 12 particularly novel associations absent from the curated resources. Overall, this framework provides a transparent and extensible pipeline for biomedical knowledge discovery, combining statistical co-occurrence with ontology-driven enrichment to retrieve established knowledge and generate reliable predictions for precision medicine and hypothesis-generation.

Keywords: Gene–disease associations; Ontology-driven mining; Semantic enrichment; Association rules; Apriori; FP-Growth; Eclat; Precision medicine

1. Introduction

Gene–disease association mining is pivotal in biomedical research, uncovering links between genetic alterations, such as point mutations, deletions, and insertions, and disease initiation, progression, and therapeutic target identification [1], [2]. Genome-wide Association Studies (GWAS) and functional assays provide high-throughput results; however, they require significant resources and produce limited information. The analysis of large biomedical databases and literature requires computational methods, including statistical approaches, Machine Learning (ML) models, and ontology-based frameworks [3], [4].

Over the last ten years, multiple computational solutions have been developed to address challenges in the association mining. Species differences and population diversity in gene–disease datasets make it difficult to identify actual relationships, leading to false-positive results. At the same time, the rapid growth in biomedical literature renders manual (labor-intensive) review processes increasingly infeasible, which means that a more trustworthy process is needed. The combination of text mining advancements with entity recognition and database integration techniques has achieved better integration of structured and unstructured data [5], [6], [7]. Similarly, extracting co-occurrence patterns from extensive repositories has become possible using conventional text-mining tools, including PubTator [8], Chi-square tests [9], Mutual Information [10], Apriori [11][12], Frequent Pattern Growth (FP-Growth) [13], and Equivalence Class Transformation (Eclat) [14]. Deep learning models [15], [16] improve entity recognition precision, but their black-box nature poses challenges for clinical interpretation of results. These frameworks enable researchers to perform annotation and enrichment tasks while merging large genomic and clinical datasets. In parallel, the use of ontologies has become more popular because Gene Ontology (GO) [17], Gene Annotation (GA) [17], and Disease Ontology (DO) [18] provide controlled vocabularies that organize biological concepts through hierarchical structures [19], [20]. The reliability of association scoring has improved through the implementation of statistical validation methods, including the Shapiro–Wilk [21] and Anderson–Darling tests [22] and rank-based transformation techniques [23].

Despite these advancements, several important knowledge gaps remain. A significant limitation of many existing computational methods is that association discovery is still based mainly on direct term co-occurrence or surface-level correlation procedures. Conventional co-occurrence tools remain constrained to surface-level correlations, and statistical tests are frequently applied without considering ontology depth, limiting the robustness of scoring distributions. Deep learning models achieve high predictive performance, but their limited interpretability reduces their acceptance in translational settings. Such approaches produce shallow association scores and fail to detect complex indirect relationships frequently found in biological systems. As a result, associations are often biased toward frequently studied genes and diseases that appear in scientific literature. At the same time, rare or indirectly expressed yet biologically meaningful relationships remain weak or undetected. Similarly, parameter tuning in classical mining methods mainly filters associations but cannot recover or improve relationships that are not explicitly expressed in text. Therefore, ignoring ontology hierarchy collapses semantically related evidence into flat counts, overweighting the spurious co-occurrences while underweighting biologically interrelated but indirectly expressed links. Without parent–child propagation, rare but specific terms contribute little, and general but informative ancestral relationship history is discarded, reducing recall of mechanistically likely associations. Moreover, the absence of rank- and distribution-aware normalization leads to score instability across diseases and ontology depths, weakening cross-study interpretability and validation against curated resources [24], [25], [26]. Therefore, there is a need for methods that can integrate hierarchical biological knowledge to improve association discovery beyond direct co-occurrence evidence.

To address these limitations, this study proposes an ontology-driven association mining framework that enhances the association mining through semantic generalization and statistical score normalization. While other ontology-based pipelines use ontologies only for annotation or enrichment, our work focuses on developing a scoring function that incorporates hierarchical influence from ontologies to enable testable assessment of indirect evidence through parent–child and sibling relationships [20], [26]. This design directly addresses the limitation of entity-specific association (ESA) mining, where associations depend only on direct co-occurrence and ignore the historical relationships within the ancestral lineages of these associations, leaving the indirect but biologically meaningful relationships undetected.

The proposed pipeline first extracts sentence-level gene–disease co-occurrences to compute entity-entity direct associations, i.e., ESA. The pipeline propagates this evidence across ontology hierarchies to generate Hierarchical Ontology Associations (HOA). Finally, ESA and HOA scores are combined into a unified Athar Semantic-Enriched Association (ASEA) score, which enables both direct and ontology-mediated evidence to contribute to association strength. This integration stabilizes association rankings across ontology levels while improving recovery of biologically meaningful indirect relationships that are not captured by the baseline co-occurrence methods, i.e., ESA. Thus, ASEA provides a framework that overcomes frequency-driven biases in the baseline methods and relies solely on frequently reported co-occurrences. ASEA also strengthens associations supported by hierarchical ontology relationships derived from established ancestral connections. The framework is validated by comparison with expert-

curated databases, including DisGeNET [27] and the Comparative Toxicogenomic Database (CTD) [28], allowing evaluation of both established and potentially novel gene–disease associations. Furthermore, the associations are classified into four classes: high-scoring associations overlapping with standard databases, backed by scientific evidence, limited scientific evidence, and novel associations.

The main goals of this research project are as follows.

- Develop an ontology-based framework for extracting gene–disease associations by incorporating hierarchical knowledge in GO, GA, and DO.
- Enhance association mining with ontology-based generalization to capture both direct (gene-disease) and indirect (GO–DO) associations.
- Implement statistical transformation and rank-based normalization to refine the association scoring and ensure robustness.
- Validate identified associations against expert-curated databases (DisGeNET [27], CTD [28]) and classify high-confidence associations against curated sources while predicting novel associations for future research directions.

2. Related Work

Traditionally, the process of finding meaningful relationships in biomedical gene–disease association mining has relied on a combination of classic association rule mining algorithms and ontology integration. These two directions have evolved in parallel, gradually giving rise to semantic methods and machine learning techniques that attempt to overcome the limitations of purely statistical or ontology-based pipelines.

Early efforts were based on classic association rule-mining algorithms. The Apriori algorithm [11] laid the foundation for frequent itemset mining, but it only handles direct co-occurrence and suffers from scalability issues. FP-Growth [13] is efficient through FP-trees, but it lacks semantic hierarchy integration. Eclat [14] efficiently mines frequent patterns in high-dimensional data but does not use biomedical ontologies. Recent work has shown that these algorithms remain relevant both as reference methods and as foundations for optimized or hybrid approaches [29], [30], [31], [32], [33].

Building on these foundations, advances in semantic methods and machine learning have been applied to biomedical texts. The process of entity recognition was improved through Named Entity Recognition and dictionary-based methods [34], [35], [36], which detected only basic levels of co-occurrence. The use of curated databases [1], [37] produces better knowledge quality but fails to analyse deep semantic meanings. The application of ontology-driven methods [20], [38] depends on the hierarchical context; however, these systems struggle to apply ontological structures to calculate association scores [39]. The latest models using graph neural networks (GNNs) and deep learning [3], [40], [41] achieve better predictive performance, but they focus only on statistical patterns rather than on ontological knowledge directly. The development of ontology-aware machine learning models [25], [26] has started to use dynamic semantic relationships, but only for better precision. Ontology-driven and gene-level data mining approaches have been increasingly used to support biomarker discovery and biological interpretation. Ontology-driven and gene-level data mining approaches have been increasingly

used to support biomarker discovery and biological interpretation. For example, [42] has employed GO [17] and KEGG [43], [44], [45] pathway enrichment analysis on differentially expressed genes to identify key genes and signalling pathways involved in cervical cancer. This illustrates how ontology-based functional annotation can guide disease-specific biomarker identification. In a complementary direction, [46] focused on gene-level pattern discovery in expression data using an optimized biclustering algorithm, enabling an improved identification of coherent gene groups across varying conditions. Together, these studies highlight the value of combining ontology-informed interpretation and gene-level pattern mining which motivates the ontology-enriched association framework proposed in this work.

Studies also show the growing use of advanced deep learning and graph-based models for different biomedical prediction tasks. RT-Transformer [47] combines a graph attention network with a 1D Transformer to predict molecular retention times across different chromatographic conditions. By using transfer learning from a large pre-trained dataset, the model achieves competitive accuracy and shows strong scalability when applied to multiple external datasets, supporting more accurate metabolite identification [47]. In another direction, circRDRP focuses on circRNA–drug interactions related to cancer drug resistance. This model integrates disease information using a hybrid graph neural network, which combines GAT and GCN layers with convolutional neural networks, leading to improved prediction accuracy and strong performance in real anticancer drug case studies [48]. In addition, VGAEMCD addresses the challenge of predicting disease- and function-related genes when only positive samples are available. By combining variational graph auto-encoders with one-class classification, the method effectively learns gene representations from protein–protein interaction networks and outperforms existing approaches across multiple evaluation metrics [49]. Together, these studies highlight the flexibility of graph learning, transformer models and representation learning in solving various biomedical prediction problems.

The development of transformer-based models has transformed the field of biomedical text mining. The entity recognition performance of BioBERT [15], [16], [50] reaches state-of-the-art levels, but researchers face difficulties when merging its output with hierarchical ontologies. Recent studies increasingly use deep learning and graph-based methods to predict biological associations, mainly because laboratory experiments are expensive and time-consuming. Graph-based approaches have been widely applied to miRNA–disease association prediction. A DeepWalk-based method combined with a deep neural network was also proposed to learn meaningful low-dimensional representations of miRNAs and diseases, achieving good performance, especially in cancer case studies [51]. Building on this work, a graph convolutional network combined with neural collaborative filtering (e.g. SVAE–LSTM framework) was later introduced, which better captures network structure and complex relationships leading to improved prediction accuracy for miRNA–disease associations [52][53].

For lncRNA–disease association prediction, matrix factorization has proven to be an effective approach. A lncRNA expression profile-based matrix factorization model [54] was proposed to integrate different biological data sources and achieved better results than earlier methods [54]. Recently, GMFLDA [55] combined graph convolutional networks with deep matrix factorization to further improve prediction performance by learning more informative features for both lncRNAs and diseases. In addition to RNA-related studies, graph embedding methods were applied to gene–disease association prediction. Similarly, TBLDA [56] is a transfer

learning model that predicts lncRNA–disease associations using BERT-based disease representations [57] and knowledge from miRNA–disease data, and NCMD [58] is a node2vec-based neural collaborative filtering model for miRNA–disease association prediction that combines matrix factorization with deep neural networks. Both methods are showing strong performance and outperform existing baseline methods.

Overall, these studies highlight the importance of graph learning and representation learning for discovering disease-related biological associations.

Apart from transformer-based models, the development of association discovery methods continues across three main research areas: probabilistic reasoning [59], [60], multimodal integration [61], [62], and federated learning [63]. Table 1 presents various studies, including their methods, test datasets, assessment criteria, and benefits and drawbacks, which demonstrate the current state of research and its main weaknesses that drive this investigation.

Table 1 : Summary of Selected Gene–Disease Association Studies and Their Key Shortcomings

Ref	Technique	Data Source	Metrics	Strength	Limitations
[64]	Association Mining	PubMed	Precision, Recall, F1	Improves dictionary-based NER	Limited handling of term variations; ignores ontology hierarchy
[36]	Dictionary-Based NER	PubMed Central	F1-Score	Enhanced entity recognition	Ineffective for complex relations; surface-level co-occurrence only
[20]	Ontology-Guided Clustering	Gene Ontology	Clustering metrics	Leverages hierarchy for entity recognition	Partial ontology utilization; no scoring integration
[3]	Deep Learning	Clinical, Genomics Data	Cross validation	Integrates heterogeneous data	High computational demand; limited interpretability
[40]	Hybrid CNN-RNN	Clinical, Proteomic Data	Comparison metrics	Captures semantic context	Struggles with long-range dependencies
[65]	Attention Models	Unstructured Text	Statistical tests	Highlights key terms	Misses broader context; ignores ontology links
[41]	Graph Neural Networks	Biological Networks	Area Under Curve (AUC),F1 Score	Models complex network connections	Requires detailed network data; ontology not integrated
[26]	Ontology-Driven ML	Gene Ontology	Clustering metrics	Dynamic semantic incorporation	Limited dynamic associations; ontology scoring absent
[25]	Dynamic Ontologies	DisGeNET, OMIM	Accuracy metrics	Adapts to live ontology updates	Implementation complexity; lack of ontology-integrated scoring
[15]	Transformer (BioBERT)	PubMed Texts	F1-Score	State-of-the-art entity extraction	Lacks explicit ontology integration
[5]	Multi-Source Integration	Genomic + Clinical Data	Precision, Recall	Enhances discovery across heterogeneous sources	Limited semantic hierarchy usage

[7]	Efficient Data Mining Techniques -	Biomedical Literature	Runtime, Accuracy	Improves scalability in mining	Focused on efficiency, less on semantic depth
[6]	Biomedical Literature Mining	PubMed, Clinical Texts	Review-based metrics	Comprehensive review of literature-mining advances	Emphasizes challenges in ontology integration
[24]	Ontology Annotation	Rare Disease Data	Integration metrics	Ontology-based annotation for precision medicine	Ontology use is limited to annotation, no scoring integration
[66]	Literature-Based Discovery (MeSHOP)	PubMed, MeSH Profiles	Precision, Recall	Infers novel associations via profile similarity	Predictions are often uncurated; limited ontology use
[67]	Text-Mining (PubTator)	PubMed Abstracts	F1-Score	Assists large-scale biocuration	Produces candidate links lacking curated evidence
[68]	Semantic Prioritization (BOCC)	Phenotype Ontologies + PPI Networks	Ranking Metrics	Reveals clinically relevant indirect associations	Dependent on PPI quality, indirect predictions need validation
[42]	GO & KEGG [43], [44], [45] Enrichment, GSEA, PPI Network Analysis	GEO gene expression datasets (GSE63514, GSE6791, GSE9750)	Enrichment scores, network topology (hub genes)	Supports biomarker discovery via functional pathway enrichment	Focused on differential expression and enrichment, missing explicit gene–disease associations or semantic relationships
[46]	Optimized Biclustering Algorithm	Gene expression binary and non-binary matrices	Relevance score, execution time, clustering quality	Efficient gene group discovery with improved scalability	Limited to pattern discovery/ does not integrate ontologies or disease-level semantic relationships
[47]	Graph Attention Network + 1D Transformer with transfer learning (RT-Transformer)	Large small-molecule retention time datasets and multiple chromatographic conditions	MAE	Scalable across different chromatographic methods. effective transfer learning	Performance depends on large pre-training data. limited interpretability
[48]	Hybrid GNN (GAT + GCN) with CNN (circRDRP)	circRNA–drug–disease interaction networks	Accuracy, Precision, Recall	integrates disease context; strong performance in cancer drug resistance cases	Model complexity requires well-annotated interaction data
[49]	Variational Graph Auto-Encoder + One-class classification (VGAEMCD)	Protein–protein interaction networks with experimentally validated genes	Recall, Precision, F-measure, Accuracy	Works without negative samples, unified prediction for disease and function genes	Relies on PPI network quality, does not use expression data
[51]	DeepWalk-based graph embedding +	miRNA–disease association networks	AUC, case studies	Learns meaningful low-dimensional features and	Depends on network quality, ignores rich biological attributes

	Deep Neural Network			performs well on cancer datasets	
[52]	Graph Convolutional Network + Neural Collaborative Filtering	miRNA–disease association networks	AUC	Preserves network structure and captures nonlinear relationships	Requires sufficient known associations, higher computational cost
[54]	Expression profile-based Matrix Factorization	lncRNA expression profiles and disease data	AUC (LOOCV, 5-fold CV)	Effectively integrates heterogeneous data, simple and interpretable	Limited ability to model complex nonlinear patterns
[55]	Graph Convolutional Network + Deep Matrix Factorization (GMFLDA)	lncRNA–disease networks and similarity data	AUC (LOOCV, 5-fold CV)	Strong predictive performance, combines graph learning with deep factorization	Model complexity and reliance on predefined similarity measures
[56]	TBLDA	lncRNA–disease data	AUC	High prediction accuracy	No wet-lab validation
[58]	Node2vec + Neural Collaborative Filtering (NCMD)	miRNA–disease interaction network	AUC accuracy	Learns low-dimensional embeddings; combines linear and nonlinear interaction modelling	Embedding-based; does not exploit semantic or ontology-level knowledge
[53]	SVAE–LSTM + Neural Collaborative Filtering	Protein sequences, drug chemical features	AUC, accuracy	Learns compact sequential embeddings; strong predictive performance	Embedding-based; lacks semantic/ontology-driven association modelling

Despite the progress demonstrated in these studies, key shortcomings persist in the literature. Direct co-occurrence is still relied upon too heavily, whereas indirect and hierarchical associations are missed. The volume of text data directly influences the associations made, introducing bias toward well-studied entities. The lexical ambiguities in biomedical texts pose significant challenges for association mining. Existing scoring methods fail to utilize the full potential of ontological semantic features, including parent–child and sibling relationships. Current methods lack sufficient integration between multimodal data and domain knowledge to achieve robust association discovery. Addressing these limitations requires incorporating hierarchical ontology information into association scoring so that semantically related concepts can be captured even when they are not directly co-expressed in text. The proposed framework aims to improve ranking consistency and overlap with curated resources such as CTD [28] and DisGeNET [27], which are evaluated in subsequent sections.

3. Proposed Athar Semantic-Enriched Association Algorithm (ASEA)

This study implements a multiple-step approach that includes text pre-processing, ontology-driven analysis, association mining, and semantic enrichment to extract and evaluate gene–disease connections from biomedical literature.

3.1. Implementation Description

The implementation was carried out in Java using Smile for machine learning, Apache Commons Math for numerical computation, and an Aho–Corasick–based library for efficient multi-pattern string matching.

3.2. Data Retrieval

Articles were retrieved from PubMed [69] using a query adapted from [70], which were then disintegrated based on punctuation and spacing rules, treating each sentence as a transaction unit [70], with at least three distinct words, as in English, a subject, verb, and object form a sentence having some comprehensible meaning. The minimum length constraint was applied to filter non-informative or malformed sentences while preserving recall.

To meaningfully interpret these transactions in a biomedical context, domain-specific knowledge resources were incorporated to capture the hierarchical relationships between genes and diseases.

Hierarchical biomedical ontologies, that is, Gene Ontology (GO) [17] and Disease Ontology (DO) , as well as the Gene Annotations (GA) [17] are used to construct dictionaries of gene and disease terms that preserve their hierarchical relationships [18], [71]. For genes, GO provides a hierarchical controlled vocabulary for gene attributes across species, including Biological Process (BP), Molecular Function (MF), and Cellular Component (CC). In contrast, GA links gene symbols to GO terms, enabling the generalization of GA scores (

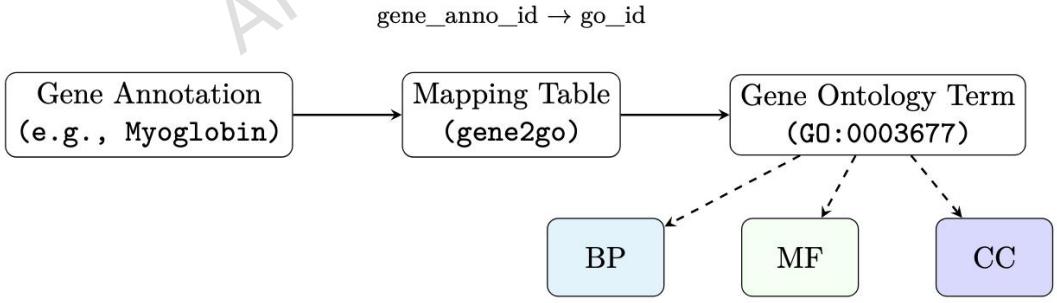


Figure 1: Linking gene annotations with Gene Ontology terms [17]

).

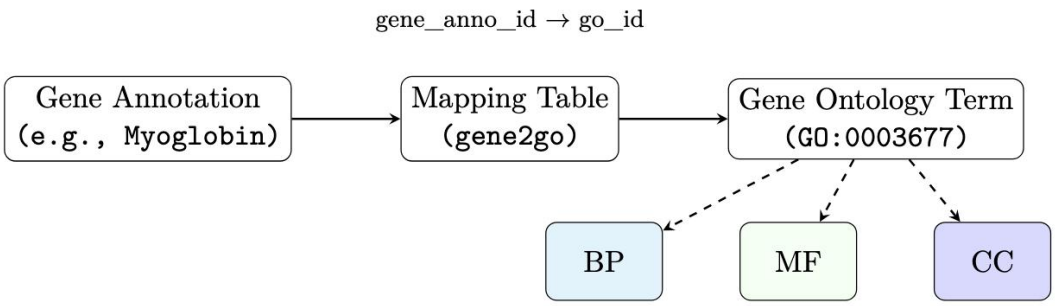


Figure 1: Linking gene annotations with Gene Ontology terms [17]

3.3. Entity-Specific Association (ESA) Extraction

Sentences were indexed by the occurrence of biomedical terms and treated as transactions for mining co-occurring gene-disease pairs. Sentence-level transactions were chosen to maximize semantic precision, as gene–disease relationships are usually expressed within a single sentence and therefore larger units such as abstracts or paragraphs were avoided to reduce false co-occurrences due to large text. Frequent itemsets are extracted using Apriori [25], FP-Growth [13], and Eclat [14] algorithms. Figure 2 shows the basic steps for extracting the association rule. Each sentence is a transaction with at least one term occurrence; two terms X and *Y* co-occurring in the same sentence with different types (gene/disease) are candidate associations. These co-occurrences are referred to as Entity-Specific Co-occurrences (ESC). After the ESC and transactions are extracted, we calculate the entity-specific association (ESA) scores using standard association score metrics described later in Section 3.5.

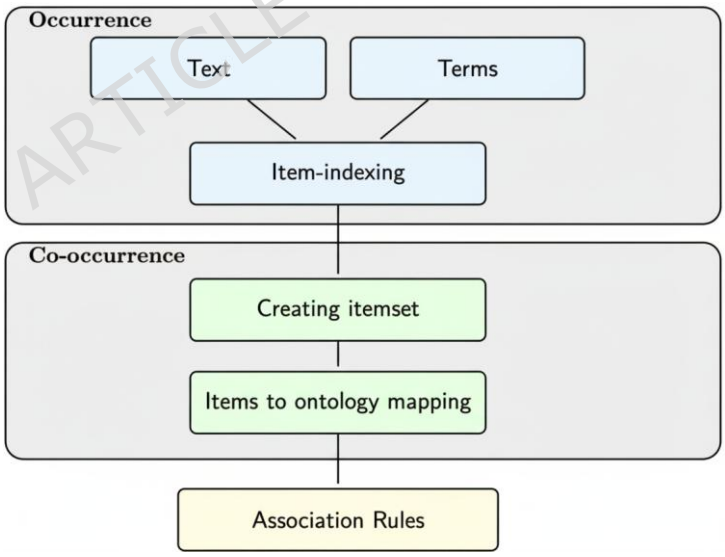


Figure 2: Basic steps for association rules extraction

3.4. Ontology-Based Generalization

After computing ESA scores, the methodology extends them to Hierarchical Ontology Associations (HOA) by incorporating parent terms in the ontology hierarchy. HOA item sets semantically generalize existing associations via DO and GO terms. Since GO classifies terms while GA contains exact gene symbols, mining is performed for both GO-DO and GA-DO

pairs. Ancestors of antecedents and consequents from ESA are combined into HOA itemsets: (i) item X to parents of item Y, (ii) item Y to parents of item X, and (iii) parents of item X to parents of item Y.

Although multiple ancestor-based combinations are generated, no duplicates occur because the associations are treated as unordered itemsets. During ontology-based generalization, the candidate pairs are produced as (X → parents(Y)), (Y → parents(X)), and (parents(X) → parents(Y)) are stored once as a unique representation. Figure 3 shows an example of GO and DO and along with the generated unique pairs.

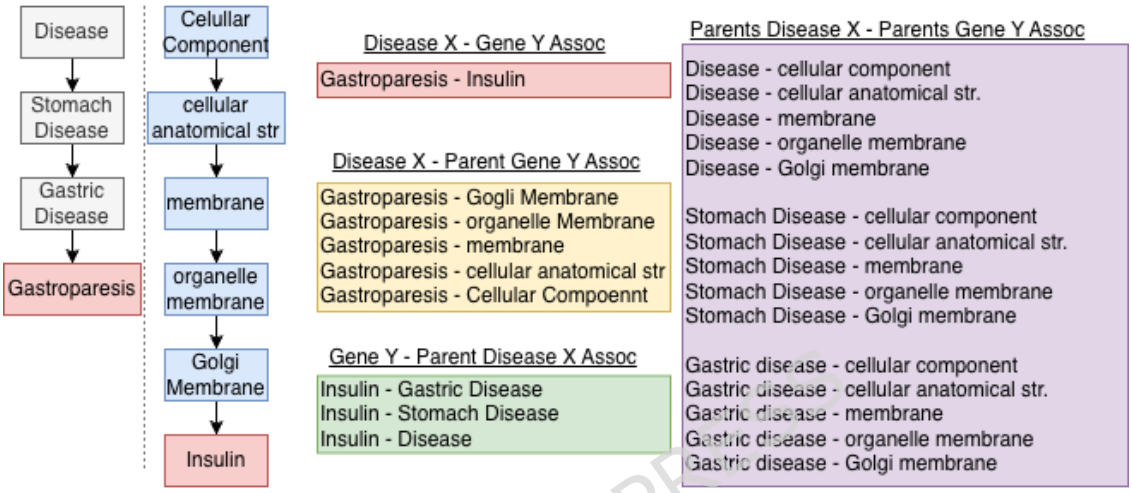


Figure 3: GO (blue) and DO (grey) association generalization

Sibling relationships are captured implicitly through the shared parent terms rather than direct sibling links.

It is important to note that Hierarchical Ontology Associations (HOA) do not introduce new textual evidence or the mechanisms of independent relation extraction. Instead, HOA aggregates and reorganizes existing sentence-level co-occurrence evidence across ontology hierarchies through parent–child propagation. Therefore, the resulting HOA associations represent semantic generalization of observed evidence on the higher conceptual levels, rather than newly discovered biological relationships. These ontology-mediated associations are intended to highlight undeveloped, higher-level patterns that may be overlooked at a strictly entity-level granularity and should be interpreted as prioritized hypotheses for further biological validation. Furthermore, the ontology-based hierarchical enrichment partially mitigates literature frequency bias by reinforcing rare/weak associations through text-supported ancestor relationships, enabling rare or weakly mentioned links to benefit from established higher-level evidence.

3.5. Association Scoring Used in ASEA Construction

To enable integration of ESA and HOA associations into the ASEA framework, association strengths are computed using standard rule-mining score metrics. Association score metrics quantify the strength and confidence of a relationship based on database scores and supporting evidence [11]. The mining algorithm computes the following score metrics (equation 1,2,3,4,5) of gene-disease associations [72]:

$$Support(X) = \frac{(Transactions\ having\ X)}{(Total\ Number\ of\ Transactions)} \quad (1)$$

$$Confidence(X \Rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)} \quad (2)$$

$$Lift(X \Rightarrow Y) = \frac{Confidence(X \Rightarrow Y)}{Support(Y)} \quad (3)$$

$$Leverage(X \Rightarrow Y) = Support(X \cup Y) - (Support(X) \times Support(Y)) \quad (4)$$

$$Conviction(X \Rightarrow Y) = \frac{(1 - Support(Y))}{(1 - Confidence(X \Rightarrow Y))} \quad (5)$$

The evaluation of biomedical associations requires specific metrics; however, multiple measurement options exist. The Lift metric functions as the primary evaluation tool because it measures statistical dependency while eliminating problems that affect other available metrics. The Lift metric provides unbiased results because it uses expected frequency normalization to eliminate item frequency bias [9]. The Lift metric directly shows the strength of dependency between variables, identifying positive, negative, and independent [73]. The Lift metric provides a simple interpretation because values above one indicates positive relationships and values at one show no association, while values below one indicate negative relationships [74]. Furthermore, Lift has become a widely adopted and validated measure in the data mining literature [72] as it balances the frequency of co-occurrence (support) with the strength of the conditional relationship (confidence), yielding robust and interpretable results in biomedical text mining.

3.6. ASEA Score Integration

In classical association mining algorithms, the association score of a rule $X \Rightarrow Y$ is computed using measures such as support (equation(1), confidence (equation (2), and lift (equation(3).
In this study, the Lift score metric acts as the baseline for the score expressed in equation (*6*):

$$S(X, Y) = Log_{10}(Lift(X \Rightarrow Y)) \quad (6)$$

However, this formulation captures only direct relationships between a gene *X* and a disease *Y*, ignoring their hierarchical structures in ontologies. To address this limitation, we enhance association score by incorporating semantic knowledge from ancestors of *X* and *Y*.
The enhanced score, called the ASEA score and denoted by $\hat{S}$ (7), integrates the direct association scores with ancestor scores as follows:

$$\hat{S}(X, Y) = S(X, Y) + \frac{\sum_{i=1}^{I} S(X, \overline{Y}_i)}{I} + \frac{\sum_{j=1}^{J} S(\overline{X}_j, Y)}{J} + \frac{\sum_{k=1,l=1}^{K,L} S(\overline{X}_k, \overline{Y}_l)}{K \cdot L} \quad (7)$$

where $S(X,Y)$ is the lift-based score between terms X and Y , $\bar{Y}_i$ and $\bar{X}_j$ denote the ancestor terms of Y and X , respectively.
For each specific ESA itemset, its ancestor-based item sets are also considered. The ASEA score thus combines the ESA score and HOA scores as outlined in Algorithm 1.

Algorithm 1: Combine Association Scores with Ancestor Terms into the ASEA Score

Input : Specific co-occurrences of item sets;
Output : Generalized co-occurrence association scores
While co-occurrences remain **do**
Retrieve ancestors of specific co-occurrence.
Compute $\hat{S}$ as the score of the specific co-occurrence
While ancestors exist **do**
Compute and accumulate ancestor scores into $\bar{S}$.
Compute final score as:

$$Score = \frac{(S + \bar{S})}{n}$$

3.7. Illustrative Example: Benefit of Ontology-Based Association Enhancement

Consider the following five sentences extracted from biomedical text:

"Mutations in **SCN1A** are a major cause of **Dravet syndrome**."
"Loss of function in **SCN1A** disrupts neuronal ion channel activity."
"Ion channel dysfunction is commonly observed in **epileptic disorders**."
"Epileptic disorders belong to the broader category of **neurological diseases**."
"Several neurological diseases are linked to abnormal ion transport mechanisms."

Using conventional sentence-level co-occurrence, only the direct association SCN1A–Dravet syndrome (Sentence 1) is detected. Associations between SCN1A and higher-level disease concepts, such as epileptic disorders or neurological diseases are missed because no sentence explicitly contains both entities. As a result, the conventional mining recovers 1 out of 3 biologically relevant associations, resulting in an illustrative recall of 0.33.
The ASEA leverages ontology hierarchies: SCN1A → ion channel activity (GO parent) and Dravet syndrome → epileptic disorders → neurological diseases (DO parents). When these parent-level gene–disease pairs are also observed in the text (sentences 2–5), ASEA aggregates their evidence to enhance association scores. Consequently, ASEA correctly recovers all 3 biologically relevant associations, corresponding to an illustrative recall of 1.0, while maintaining precision because reinforcement occurs only when parent-level associations are supported by textual evidence.

4. Results

This section evaluates the proposed ASEA framework and compares its performance with baseline association mining algorithms, including Apriori, Eclat, and FP-Growth. Association scores and top-ranked gene–disease relationships are analysed through multiple validation steps, and their quality is assessed by comparison with expert-curated databases.

4.1. Occurrence of Biomarkers

These terms were broadly classified into three categories: DO, GO, and GA. Table 2 summarizes the overall statistics, including the number of biomarkers, synonyms, occurrences in the text corpus, and their hierarchical distribution. This approach generates two types of scores: ESA and HOA, which are then combined to obtain the ASEA score. However, extracting ASEA scores from the ESA and HOA is non-trivial, as it requires careful consideration of terms, ontology metadata, and their hierarchical structure. Because ontologies were organized at multiple levels, the number of terms varied by depth. Entities were distributed across the top six levels of the ontologies, with level zero representing the root. In addition, the algorithm searched for occurrences of GA, DO, and GO terms in approximately 1,508,980 sentences. The resulting counts for each biomarker type are shown in Table 2.

Table 2: Combined statistics of ontology and annotation datasets: term counts, synonyms, corpus occurrences, and hierarchical distribution (L0–L6). Levels apply only to GO and DO.

Marker	Terms	Synonyms	Occurrences	L0	L1	L2	L3	L4	L5	L6
GO	40,214	106,768	206,546	3	45	452	1460	3072	3570	3315
DO	11,858	19,258	554,238	1	8	117	153	377	481	514
GA	44,615	44,615	300,174	–	–	–	–	–	–	–

4.2. Parameter Sensitivity and Score Distribution Analysis

The evaluation focuses on comparing baseline entity-specific associations (ESA) with ontology-enhanced associations (HOA) and the final ASEA scores to assess the impact of ontology-based enrichment. The minimum support and minimum confidence were set to a small value (0.01%), yielding ≈ 7,841 and 7,797 associations for FP-Growth and Eclat, respectively, while 3,599 associations for Apriori. The maximum rule length was set to 6, and the minimum rule length to 2. To assess the robustness, a higher support and confidence thresholds were also evaluated. Increasing the thresholds to 0.2% reduced the number of associations to ≈ 2,000, and further increasing them to 0.5% reduced the associations to around 700 for FP-Growth and Eclat and 280 for Apriori, resulting in a significant loss of coverage. Because the ASEA is designed to retain weak associations at the initial stage for a later ontology-based enhancement, higher thresholds were not feasible. However, the relative performance trends across methods remained stable. A low minimum support and confidence threshold (0.01%) was intentionally selected to maximize the recall at the initial mining stage, as stricter thresholds exclude associations in the early stages. Still, these excluded associations may later be strengthened through ontology-based generalization. Having these parameters, the Apriori produces fewer associations due to its exhaustive level-wise candidate generation and strict pruning, focusing on high-confidence patterns [9]. Conversely, Eclat and FP-Growth use depth-first and pattern-growth methods,

uncovering more associations, but often redundant or skewed results, especially in sparse datasets **Error! Reference source not found.**.

Despite fewer rules, Apriori was selected as the primary method because its association rules are highly interpretable and widely used in knowledge discovery. This interpretability provides a reliable foundation for ontology-based semantic enrichment, as the rules can be directly leveraged to identify new semantic relationships [12], [75], [76]. The associations based on direct co-occurrences of DO and GA are called ESA. These ESA scores are subsequently enhanced through ontology-based generalization (HOA) and combined into ASEA scores, enabling evaluation of how hierarchical propagation improves association quality. However, to avoid trivial generalizations, the top four levels of the DO hierarchy were excluded from the analysis. The distribution of association scores was analysed using histograms and Q-Q plots for statistical rule evaluation. The evaluation of scores from the GA and DO terms used the lift score as the assessment metric for all three algorithms.

Table 3: Ontology levels (L) included (dots) or excluded in each test case

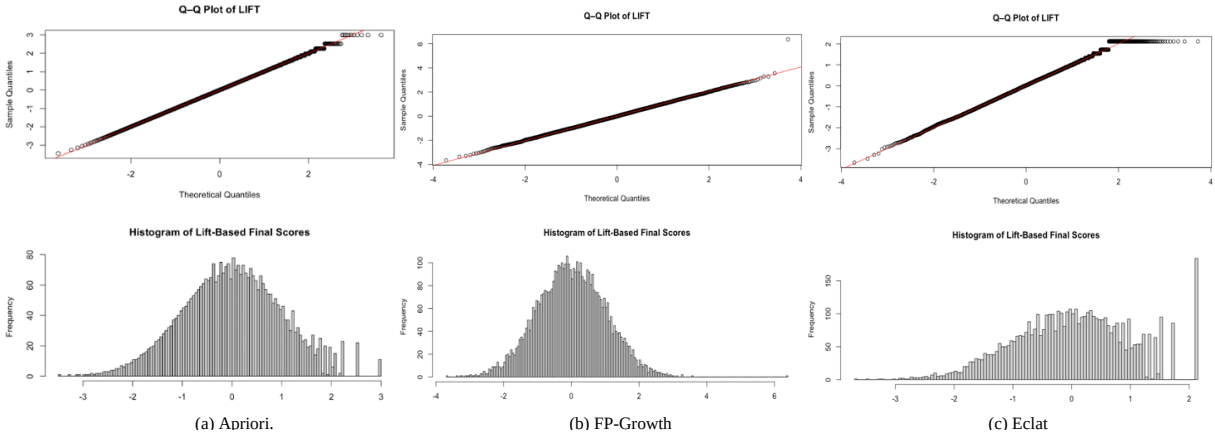
	L0	L1	L2	L3	L4	L5	L6	Ln
Test_0	•	•	•	•	•	•	•	•
Test_1		•	•	•	•	•	•	•
Test_2			•	•	•	•	•	•
Test_3				•	•	•	•	•
Test_4					•	•	•	•
Test_5						•	•	•
Test_6							•	•
Test_7								•

Combining ESA and HOA scores introduces variability because distributions differ across ontology levels. To address this, distributions are examined through histograms and validated statistically to ensure the ontology structure adds value without distortion. Normality is crucial because Gaussian distributions allow consistent comparisons. Histograms and statistical normality tests [77] are applied. If distributions approximate normality, the arithmetic means of ESA scores and ancestor scores provide reliable combined measures. Since root-level terms (e.g., “Disease,” “Cancer”) can dominate, successive test cases exclude higher-level ontology nodes, as shown in Table 3, and results are compared using histograms and Q-Q plots.

Two complementary tests validate distributions. The Shapiro–Wilk test [21] evaluates the normality of ESA and six HOA test cases. For large samples, the Anderson–Darling (A-D) test [22] is applied with greater sensitivity to tails. These validation steps ensure that ESA and HOA distributions are comparable. Where normality holds, arithmetic means can combine them into a unified score, providing the statistical foundation for the Athar-Semantic-Enriched Associations (ASEA) Algorithm.

The Apriori-generated lift score metric extended across a broad range, making direct understanding challenging. The distribution appears heavily skewed when viewed without transformation. The $\log_{10}$ transformation helped minimize skewness in the distribution while making its underlying patterns more visible. The Shapiro–Wilk test confirmed that the data failed to meet normality requirements as it included all DO hierarchical levels ($W=0.96427$, $p < 2.2e{-}16$). Excluding the top four DO levels produced a better distribution; however, the data remained far from normal ($W=0.99141$, $p = 3.96e{-}12$). To address this limitation, a tie-aware rank-based transformation (RBT) was applied [60]. This method preserves the tied ranks while

maintaining the relationships among the original values. This transformation produced near-normal distributions, indicating that the data were more suitable for subsequent statistical analyses.



(a) Apriori. (b) FP-Growth (c) Eclat

Figure 4: Tie-aware rank score distributions for (a) Apriori, (b) FP-Growth, and (c) Eclat, excluding the top four DO levels.

The same transformation was applied to modify the scores from Eclat and FP-Growth algorithms. The tie-aware RBT produced improved output distributions for both Eclat and FP-Growth that matched the results of Apriori. The transformed results maintained some minor deviations from normality but achieved statistical consistency, supporting reliable comparisons between the algorithms. These normalized score distributions enable consistent comparison between ESA and HOA contributions and support the reliable construction of ASEA scores used in subsequent evaluations.

The Q–Q plots in Figure show the distribution of lift-based scores against theoretical normal distribution data (red line) in the top row. The points that match the red reference line exhibit normal distribution characteristics, whereas points that deviate from the line indicate skewness or heavy tails in the data. The bottom row displays histograms that show the complete score distribution patterns. The histogram of Apriori (a) shows a distribution that approaches normality but has a small rightward tail and the Q–Q plot shows better alignment with the reference line after applying rank-based transformation. The Q–Q plot and histogram of FP-Growth (b) show that the data follows a near-normal distribution with minor deviations in the tail section. The histogram of Eclat (c) shows an unbalanced distribution pattern, and its Q–Q plot reveals significant deviations from the red reference line across the entire distribution range. Similarly, the Shapiro–Wilk tests applied after the tie-aware rank-based transformation (RBT) yielded $W = 0.99958$, $p = 0.6609$ for Apriori and $W = 0.99852$, $p = 1.342 \times 10{-}4$ for FP-Growth i.e., a near to normal distribution, while $W = 0.99527$, $p = 1.086 \times 10{-}11$ for Eclat.
These analyses prepare the score distributions used in subsequent sections, where ASEA demonstrates improved recovery and ranking of biologically meaningful associations compared with baseline methods.

4.3. Normalization of HOA and ESA Scores for ASEA Integration

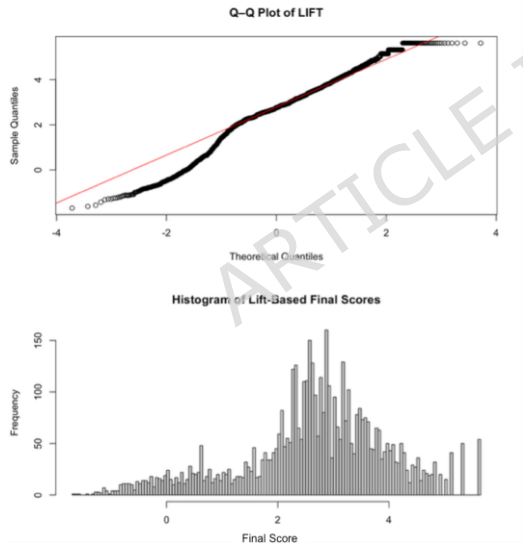
The final association score ties the ESA and HOA scores to construct the ASEA score, while Apriori provides the baseline association structure. Normalized Apriori scores from DO–GA (ESA) item sets were combined with HOA scores from DO–GO item sets. The arithmetic mean could serve as the integrated score if both distributions were normal [77].

However, repeated high-level ontology terms, such as “Disease,” can distort associations. To address this, several test cases excluded upper ontology levels (see Table), and Apriori scores were recalculated.

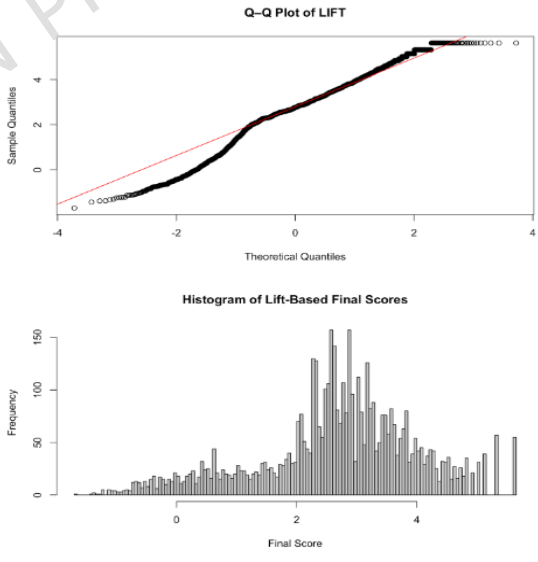
The normality of HOA scores was assessed using four representative tests. Histograms (see Figure 5) showed strong skewness, and the Anderson–Darling test [22] consistently rejected normality (A = 59.02–120.28, $p < 2.2 \times 10^{-16}$). Consequently, tie-aware RBT was applied to correct for non-normality.

Figure 6 shows HOA score distributions after RBT. In Test0, the histogram was bell-shaped, and the Q–Q plot aligned well ($A = 0.5777$, $p = 0.1333$), but general root terms distorted associations. Tests 1 and 2 showed unimodal distributions with moderate skewness.

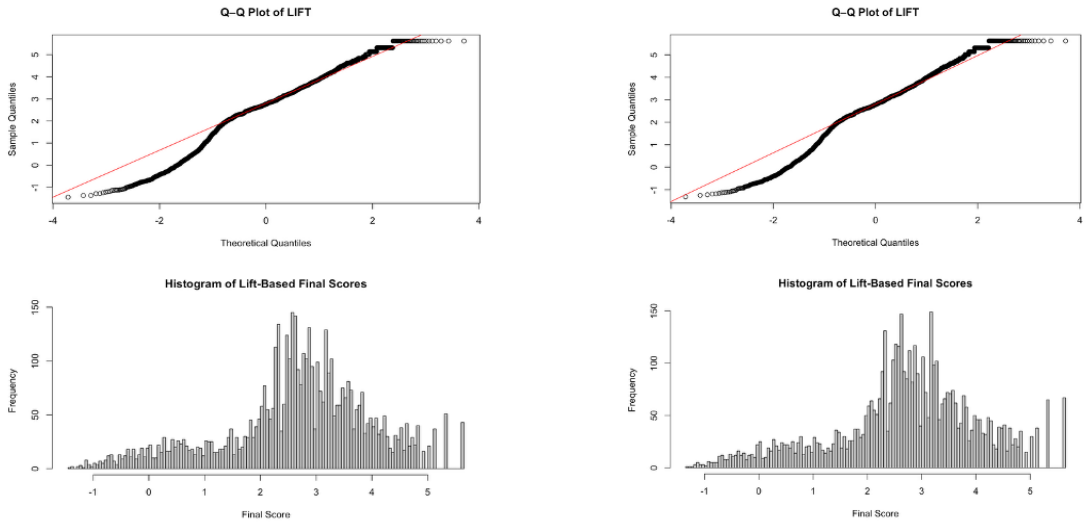
($A = 0.5971$, $p = 0.1189$; $A = 0.7044$, $p = 0.0660$). Test 3, excluding the top four ontology levels, produced smoother distributions and closer Q–Q alignment, with borderline A–D results ($A = 0.7527$, $p = 0.0501$), representing the best balance between normality and term retention. Tests 4–6, excluding more levels, led to flatter histograms and stronger tail deviations ($p < 0.01$). Overall, Test 3 was deemed optimal for subsequent analyses.



(a)

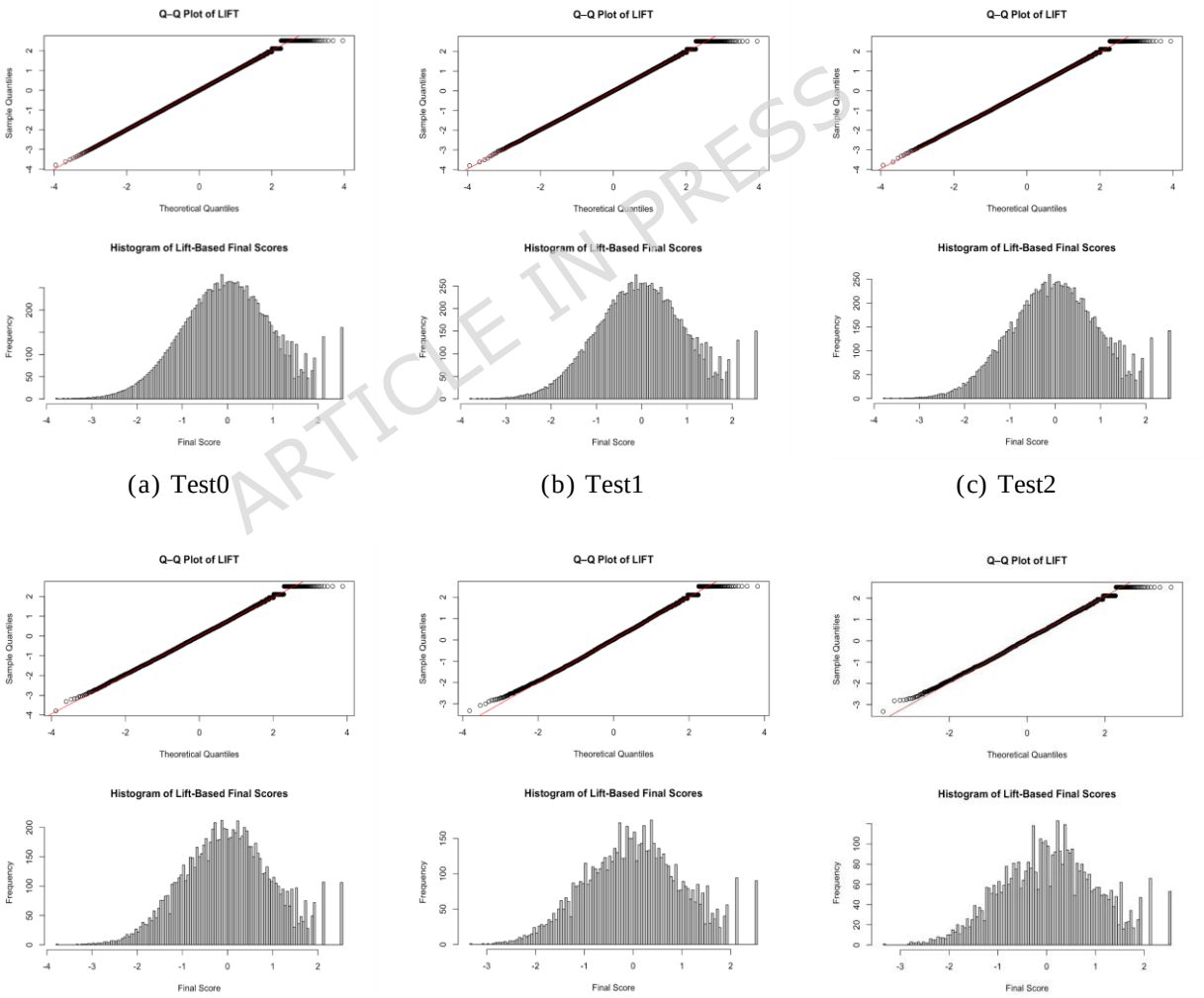


(b)



(c) (d)

Figure 5: Apriori HOA score distributions for four test cases.



(a) Test0 (b) Test1 (c) Test2

(d) Test3 (e) Test4 (f) Test5

Figure 6: HOA score distributions after applying tie-aware rank-based transformation (RBT)

The resulting distributions provide the basis for computing stable ASEA scores evaluated in the subsequent section.

4.4. ASEA Score Distribution and Stability

At this stage, the ASEA framework computes the final association score by considering the ESA scores resulting from the GA and DO. For each of the ESAs between X and Y, the association between the ancestors of X and Y was created and searched for in the GO–DO associations. If an association existed, the mean of the sum of 'log10 (Lift)' of both the HOA and the ESA was taken as the final score. The histogram and Q–Q plot of the ASEA scores is shown in Figure 7. The Q–Q plot indicates that most lift-based final scores align closely with the theoretical normal distribution (red line), with only minor deviations at the extreme upper and lower quantiles. The corresponding histogram supports this observation, showing a roughly bell-shaped distribution centred near zero, although with a slightly heavier right tail. Together, these plots suggest that the transformation substantially improves normality, with only limited departures in the tails of the distribution, indicating that the integrated score is statistically well behaved and suitable for further evaluation.

In the next step, we determined how well these normalized and integrated scores aligned with expert-curated databases by comparing the top-ranked associations against the CTD database.

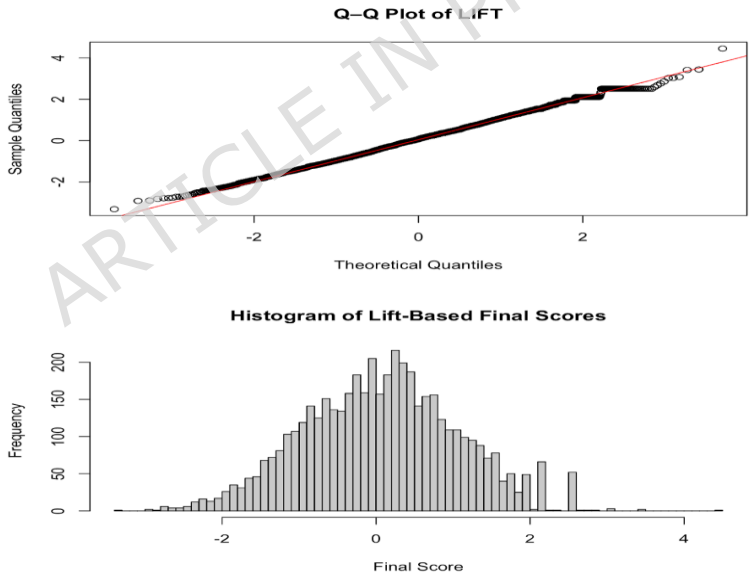


Figure 7: ASEA Algorithm's score

4.5. Rank-Based Correlation with Expert Curated Data

In this study, an indirect association is formally defined as a gene–disease relationship that does not appear as a direct co-occurrence at the entity level but becomes evident through parent–child or ancestor relationships in the ontology hierarchy. ASEA captures these associations by propagating evidence through the ontological parents and combining entity-specific and hierarchical scores. Because no curated standard exists for the indirect ontology associations,

we do not define a separate ground-truth set for indirect links. Instead, recall gains are assessed indirectly through improved overlap and ranking consistency with the curated database CTD as compared with the classic data mining algorithms (Apriori, Eclat and FP-Growth). This evaluation strategy is also consistent with prior work in gene and disease association mining and literature-based discovery, where overlap or rediscovery of known or rare associations from established reference databases is used as a validation signal when the gold-standard negative examples are absent. Similar overlap-based assessments have been adopted to measure biological plausibility, coverage, and prioritization quality rather than predictive accuracy [78], [79], [80].

The evaluation measures the degree to which different association scores produced by different algorithms match the expert-curated CTD data. The degree of overlap between these results indicates the reliability and biological significance of the extracted associations.
The ASEA received grades ranging from 20 (strongest) to 0 (weakest) based on its calculated scores, and the CTD associations received a similar grading. The top-grade associations from the Apriori Eclat FP-Growth and ASEA scores were compared with the CTD data to identify common elements. Venn diagrams were used to examine the overlap between grades 20 and 15, and the intersection rates were consistently below 3%. The ASEA Score produced the highest number of 17 overlaps, while Apriori and Eclat reached 13, and FP-Growth reached 9. The algorithms produced numerous distinct associations with ASEA and Apriori, showing the greater potential to discover new connections that extend beyond the existing curated knowledge.
A low overlap exists because statistical mining methods differ from manual curation techniques, instead of showing any algorithmic breakdown. The combination of ESA and HOA in ASEA produced the best results in terms of relevance and achieved the highest overlap rate with additional distinct findings. Table 4 summarizes the results from each algorithm, showing their maximum overlap, unique associations, and their ability to produce consistent results. ASEA stands out as the top hybrid method according to the results, whereas Apriori and Eclat show average consistency, and FP-Growth demonstrates the lowest performance.

Table 2: Summary of algorithmic overlap and unique associations vs curated gene–disease data.

Algorithm	Max Overlap	Max Unique	Consistency	Best Grade(s)
ASEA	**17**	185	Highest	16, 20
Apriori	13	217	Moderate	17, 20
Eclat	13	166	Moderate	17, 20
FP-Growth	9	71	Lowest	18

Taken together, these results demonstrate that while statistical mining alone cannot replicate expert-curated datasets, the integrative ASEA framework provides a stronger foundation for uncovering meaningful gene–disease associations, motivating further biological validation in the subsequent analysis.

4.6. Validation Classification of Top-Ranking ASEA Scoring Associations

The ASEA score determined the top associations, which were organized into four validation classes. Case 1 represents well-established gene–disease associations that major databases have

confirmed. Case 2 includes associations that received strong support from recent literature, even though they have not been added to curated databases. The associations in Cases 3 and 4 are derived from associations having low or intermediate support from standard databases and lack evidence from scientific literature. Case 4 contains speculative or purely computational findings because they lack support from either databases or the literature.

The ASEA algorithm identified strong associations between genes and diseases in Cases 3 and 4, but they remained undetected in the current curated databases (Table 5). The hierarchical integration of ontological knowledge enables strong relationships among parents or related terms to increase the scores of less-studied or novel gene–disease pairs. However, the existence of novel relationships between biomarkers is not unknown to computational biology similarity-based methods, as MeSHOP [81] demonstrates their ability to create novel gene–disease relationships by utilizing MeSH profiles that lack direct curation [66], [67]. The BOCC system uses phenotype ontologies and protein-protein interaction networks to detect indirect clinical associations that have not been previously documented [68]. The predictions from Cases 3 and 4 functions as plausible hypotheses rather than verified results, because the model successfully integrates semantic context and hierarchical structures. The candidates require experimental confirmation and additional data integration to increase their confidence. This predictive capability positions the proposed method as a valuable tool for hypothesis generation and guiding targeted biomedical research.

This study had several limitations. The detection system depends on complete ontology information and clear annotations, but missing or unclear terms and their IDs may pose challenges when detecting associations. The proposed approach can be applied to biomedical entities beyond gene-disease associations using suitable ontological resources for drug-pathway-phenotype entities.

Table 3: Summary of key gene-disease associations selected from each classification case

Disease	Gene	Grade	Evidence/Comments
Case 1: High-confidence associations			
Gestational diabetes	INS	20	DisGeNET score = 0.8
Lymphangioleiomyomatosis	TSC2	20	DisGeNET score = 0.9
Wilson disease	ATP7B	19	DisGeNET score = 1
Classic galactosemia	GALT	19	DisGeNET score = 1
Cystinosis	CTNS	19	DisGeNET score = 1
Dravet syndrome	SCN1A	19	DisGeNET score = 0.6
Lupus nephritis	SERPINC 1	19	DisGeNET score = 0.85
Plasma cell leukemia	CD86	19	DisGeNET score = 0.85
Colonic Neoplasms	USP39	19	Found in CTD at same grade
Estrogen-receptor positive breast cancer	ACACB	19	Found in CTD at same grade
Thyroid cancer	ABCC10	19	Found in CTD at same grade
Uterine fibroid	EGFR	19	Found in CTD at grade 16
Onchocerciasis	PLK5	19	Found in CTD at same grade
Coloboma	CRX	19	Found in CTD at same grade
Lung adenocarcinoma	MEOX1	19	Found in CTD at grade 18
Adenocarcinoma	AKT1	19	Found in CTD at grade 17
Mild cognitive impairment	NCAPH2	19	Found in CTD at same grade
Testicular cancer	FGF9	19	Found in CTD at grade 17

Fibroma	HEY1	19	Found in CTD at grade 17
Meconium aspiration syndrome	CRP	19	DisGeNET score = 0.6
PFAPA syndrome	CAPS	18	Found in CTD at same grade
Case 2: Strong literature evidence, no high-confidence database			
Dissociative amnesia	CA4	20	Literature links CD4 T cells to hippocampal neurogenesis [82]
Neovascular glaucoma	EPO	20	DisGeNET score = 0.15; Elevated EPO promotes abnormal vessel formation [83]
Chondroblastoma	USP6	20	DisGeNET score = 0.1; USP6 gene rearrangement not involved [84]
Osteosarcoma	PCSK5	20	Key gene linked to replication stress [85]
Chondrosarcoma	NOS2	20	DisGeNET score = 0.12; minor support. Biomarker and potential therapeutic target [86], [87]
Idiopathic pulmonary fibrosis	HPCAL1	19	Downregulated in IPF lungs; possible fibrosis impact [88]
Idiopathic pulmonary fibrosis	RNF182	19	Downregulated; may impact fibrosis [88]
Idiopathic pulmonary fibrosis	FHL2	19	Downregulated; possible fibrosis effect [88]
Oral squamous cell carcinoma	MUC4	19	Upregulated in OSCC [89]
Oral squamous cell carcinoma	OAZ1	19	mRNA elevated; potential biomarker [90]
Oral squamous cell carcinoma	DUSP1	19	Downregulation linked to tumor burden [91]
Fibroma	NCOA2	19	DisGeNET score = 0.2; limited evidence Gene fusion causes abnormal growth [92]
Fibroma	USP6	19	Overexpression causes fibroblast proliferation [93]
Follicular lymphoma	CFHR1	19	Affects lymphoma cell response [94]
Mucosal melanoma	GLT8D1	19	DisGeNET score = 0.1; low association Amplifies tumor aggressiveness [95]
Uterine corpus endometrial carcinoma	FOXK2	19	Oncogenic driver [96]
Uterine corpus endometrial carcinoma	FBXO32	19	Silencing increases aggressiveness [97]
Osteoblastoma	NCOA2	19	HEY1–NCOA2 fusion hallmark [98]
Osteoblastoma	ACVR2A	19	Rare mutations in bone tumors [99]
Osteoblastoma	FOS	19	Most commonly altered gene driving proliferation [100]
Mastoiditis	CXCL10	19	Increased levels in disease [101]
Mastoiditis	CXCL8	19	Key player in inflammation [102]
Lupus nephritis	GTF2B	19	May contribute via transcription [103]
Chronic myelomonocytic leukemia	CBL	19	DisGeNET score = 0.1; minor support

			Mutation causes uncontrolled growth [104]
Niemann-Pick disease type C2	C2	19	Mutations break cholesterol transport [105]
Azoospermia	ZMYND15	19	Mutations linked to azoospermia [106]
Gout	AP1B1	19	Minor role in inflammation [107]
Case 3: Low/ intermediate data base support, no strong literature			
Lichen planus	IL12RB2	20	DisGeNET score = 0.2; low-confidence
Lichen planus	OAZ1	19	DisGeNET score = 0.25; weak, possible biomarker
Lichen planus	DUSP1	19	DisGeNET score = 0.1; minor/uncertain support
Vitiligo	MCHR1	19	DisGeNET score = 0.1; may be autoantigen
Varicocele	FASN	19	DisGeNET score = 0.2; potential biomarker
Keratoacanthoma	CD1A	19	DisGeNET score = 0.25; weak association
Keratoacanthoma	HSPD1	19	DisGeNET score = 0.1; limited
Chordoma	LMX1A	19	DisGeNET score = 0.1; weak association
Chordoma	SALL3	19	DisGeNET score = 0.4; moderate association
Mesenchymal chondrosarcoma	HEY1	19	DisGeNET score = 0.2; low-confidence
Mesenchymal chondrosarcoma	NCOA2	19	DisGeNET score = 0.2; weak support
Mesenchymal chondrosarcoma	USP6	19	DisGeNET score = 0.1; low-confidence
Fibromyalgia	TRPV2	19	DisGeNET score = 0.25; weak support
Fibromyalgia	NRXN3	19	DisGeNET score = 0.25; uncertain evidence
Acquired immunodeficiency syndrome	FLI1	19	DisGeNET score = 0.1; minor role
Acne	HSD11B1	19	DisGeNET score = 0.1; limited evidence
Osteoarthritis	CMA1	19	DisGeNET score = 0.35; moderate support
Ulcerative colitis	CXCR1	19	DisGeNET score = 0.5; moderate support
Plasma cell leukemia	CD200	19	DisGeNET score = 0.45; moderate association
Plasma cell leukemia	CD48	19	DisGeNET score = 0.25; low confidence
Acute lymphoblastic leukemia	IKZF1	19	DisGeNET score = 0.3; some support
Acute lymphoblastic leukemia	ARID5B	19	DisGeNET score = 0.3; some support
Acute lymphoblastic leukemia	CEBPE	19	DisGeNET score = 0.2; low confidence
Acute lymphoblastic leukemia	THBD	19	DisGeNET score = 0.1; weak support
Acute lymphoblastic leukemia	PBX1	19	DisGeNET score = 0.2; low/uncertain
Hypoglycemia	ADK	19	DisGeNET score = 0.1; minimal evidence

Dissociative amnesia	GC	19	DisGeNET score = 0.1; very weak association
Case 4: No data base or strong literature evidence			
Childhood T-cell ALL	MB	19	No evidence found
Semantic dementia	CBS	19	Speculative, enzyme function loss
Tuberculoid leprosy	MB	19	MB gene does not cause leprosy
Pulpitis	MT1H	19	Upregulated but not causative
Pulpitis	MAPK8	19	Activated during inflammation, not causative
Cholecystolithiasis	GC	19	No evidence found
Gout	CTSZ	19	Only indirect role
Gout	LAMP2	19	Clean-up role, not driver
Kawasaki disease	TFRC	19	No evidence found
Eclampsia	EBI3	19	Only indirect correlation
Coloboma	LMX1A	19	No direct evidence
Sertoli cell tumor	GFAP	19	No evidence found
Prostate adenocarcinoma	FOXK2	19	No evidence found
Bloom syndrome	BLM	19	Only lower-grade associations found
Osteoblastoma	FN1	19	Indirect "parent" relation
Osteoblastoma	FOSB	19	Indirect "parent" relation

5. Summary of Results

- ASEA detected 17 high-grade associations (Grades 15–20) when benchmarked against the CTD database which is 30.4% more than Apriori and Eclat (13 each) and 88.9% more than FP-Growth (9).
- ASEA identified 185 novel associations. Apriori detected 217 (14.7% more than ASEA), Eclat detected 166 (10.3% fewer than ASEA), and FP-Growth detected 71 (160.6% fewer than ASEA). ASEA prioritized precision over quantity, producing more reliable results than Apriori.
- The discovered associations (by ASEA) were classified into four validation cases (see **Error! Reference source not found.**):
  - **Case 1:** 21 well-documented high-confidence associations from major databases.
  - **Case 2:** 28 associations supported by strong literature evidence but not yet in curated databases.
  - **Case 3:** 28 associations with low or intermediate database support and no strong literature evidence.
  - **Case 4:** 16 speculative associations identified via ontology-based enrichment, including 6 particularly novel associations requiring further research.
- ASEA uses statistical co-occurrence analysis, along with hierarchical ontology data to access established scientific knowledge and produce accurate predictions.

6. Limitations

- The proposed ASEA framework is designed for hypothesis generation rather than causal validation, and several limitations should be noted.

- The ASEA assigns association scores but does not label associations as true or false positives, as many gene–disease relationships remain unknown rather than incorrect at the time of discovery. Associations reported in Case 4 (Table 5) represent ontology-enriched hypotheses for which no database or strong literature evidence exists and should be interpreted as directions for future investigation in the biomedical science.
- The purpose of HOA is to surface ontology-mediated associations rather than to claim discovery of new primary biological evidence.
- The framework is sensitive to support and confidence thresholds. Increasing these thresholds considerably reduces the number of extracted associations, which causes the loss of coverage and limiting the association discovery in the early stages. Therefore, ASEA uses minimum thresholds to retain weak associations for subsequent semantic enhancement.
- Ontology-based enrichment in the current implementation propagates evidence only through parent–child relationships however, a direct sibling-level propagation is not explicitly modeled. Although sibling relationships are implicitly captured via shared parents, direct sibling reinforcement may further improve semantic resolution and is left for future work.
- ASEA preserves the original ESA and HOA scores, but the framework is designed for relative prioritization instead of direct interpretation of clinical effect sizes. A Tie-aware Rank-based transformation is used to stabilize comparisons across heterogeneous score distributions, while clinical magnitude and causal interpretation are beyond the scope of this study.

7. Conclusion and Future Direction

The approach creates a pipeline for association scoring based on historical ancestral relationships between biomedical entities using hierarchical ontologies. The study involved two association types: ESA, which used direct gene-disease co-occurrences, and HOA, which utilized ontology structures to identify higher-order semantic relationships. A major challenge was that ESA and HOA scores exhibited significant deviations from a normal distribution. Therefore, the scores were transformed using a tie-aware RBT and combined into a unified association scoring framework i.e. ASEA.

The ASEA framework processed 1.5 million sentences, together with 44,615 gene annotations and 11,858 disease ontology terms, to generate strong association rules. It produced 17 validated associations when tested against the curated CTD database, outperforming Apriori and Eclat (13 each, 23.5% fewer than ASEA), and FP-Growth (9, 47.1% fewer than ASEA). The ASEA algorithm found 185 significant relationships while maintaining its best performance level because it achieved the highest reliability, even though its overall overlapping rate remained below 3%. The Apriori algorithm produced 217 associations that exceeded ASEA by 17.3%, but Eclat generated 166 associations that were 10.3% lower than ASEA, and FP-Growth produced 71 associations, which were 61.6% fewer than ASEA.

All the identified associations were grouped into four categories based on the analysis of the associations. The first group consisted of 21 documented associations that originated from major databases. The second group included 28 associations that received backing from research

evidence but were not included in existing curated databases. Group 3 consisted of 28 associations, with limited database support and no strong literature evidence. The ontology-based enrichment process revealed 16 novel associations organized in Group 4; these associations require additional research to validate their existence.

The ASEA framework demonstrates that statistical co-occurrence analysis with hierarchical ontology integration yields association scores that are more reliable and easier to understand and maintain. The framework depends on extensive ontology information and precise annotations, which enables its application in drug-pathway-phenotype relationship analysis. Experimental validation is required to confirm the biological significance of the newly discovered associations.

Future research will concentrate on the following main objectives: Adding species-specific gene and disease information to the framework and developing inter-species association scoring. The framework will be enhanced by the integration of the sibling relationship and its effects on the score. A distance-aware decay function may be incorporated to progressively down-weight associations propagated across deeper ontology levels, further reducing the risk of over-smoothing. Furthermore, the framework will be enhanced by using MONDO [108], GO IDs [17], and UMLS CUIs [109] as unique identifiers for cross-ontology mapping, and the scoring system will be improved by using weighted co-occurrence metrics, ontology-based semantic similarity scores, and machine-learning-driven confidence weighting. It will explore fairness metrics and rarity-aware weighting to improve coverage of rare diseases. We plan to use machine learning for context-aware disambiguation to handle ambiguous gene symbols better and reduce terminology noise. We'll also explore hybrid mining algorithms that combine Apriori's [11] interpretability with graph-based efficiency to help the framework scale to larger datasets.

Future work will also include explicit evaluation of ranking quality and predictive performance using metrics such as precision@k and AUC to distinguish ranking improvements from coverage expansion.

ASEA-generated associations can also serve as structured and interpretable inputs for downstream AI models, including graph-based and deep learning approaches, to support predictive analysis. The demonstrated ontology-aware strategy can be integrated with graph-based or embedding-based models.

The framework can be extended to incorporate ontology-aware pruning in FP-Growth [13] and Eclat [14], and redundancy may be quantified using formal measures such as closed or non-derivable itemsets.

The validation process will expand to include the OMIM [110], DisGeNET [27], and PharmGKB [111] databases through their specific term identifiers. The predictive power and translational value of ASEA will increase through its integration and evaluation with biological pathways from the Reactome [112] and KEGG [43], [44], [45] databases and with protein–protein interaction networks from the BioGRID [113] and STRING [114] databases, and through its ability to receive real-time updates from newly published literature and ontology revisions.

Authors' contributions

MA.Q., Data Creation, Implementation, methodology, and Writing. M.A., Supervision, writing, and validation. J. U proofreading, writing, and Supervision. HS .H writing, visualization. A. R., interpretation, Writing, and Visualization. W. A., Writing, Interpretation, and Implementation HK. A., Supervision, funding, and Proof-Reading. HA. M., formal analysis, writing, and resources. All authors reviewed and approved the final manuscript.

Data availability

The dataset and codes of the proposed model are publicly available at https://github.com/atharnaqash/assocation-miner.

Funding

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R235), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through the Large Research Project under grant number RGP2/331/46.

Competing interests

The authors declare no competing interests

References

[1] L. J. Jensen, J. Saric, and P. Bork, "Literature mining for the biologist," *Nat. Rev. Genet.*, vol. 7, no. 2, pp. 119–129, 2006.
[2] V. Tam, N. Patel, V. Turcot, and et al., "Benefits and limitations of genome-wide association studies," *Nat. Rev. Genet.*, vol. 20, no. 8, pp. 467–484, 2019.
[3] Y. Zhou, Q. Yang, C. Zhao, Z. Li, and Z. Wang, "Deep learning for bioinformatics: From raw data to predictive models," *Bioinformatics*, vol. 34, no. 5, pp. 837–844, 2018.
[4] Q. Huang, B. Wang, H. Lin, Y. Zhang, and et al., "Machine learning in biomedical informatics: A survey," *Biomed Res. Int.*, vol. 2018, pp. 1–15, 2018.
[5] Q. Yang, Y. Wang, Z. Chen, X. Liu, Y. Li, and W. Zhang, "Integrating multi-source data for enhanced gene-disease association mining," *BMC Genomics*, vol. 19, p. 562, 2018.
[6] Y. Zhu, M. Song, C. Chen, D. Liu, and H. Zhao, "Advances in biomedical literature mining for disease gene discovery," *Brief. Bioinform.*, 2020.
[7] D. P. Campos, A. Oliveira, and N. De Maio, "Efficient data mining techniques in biomedical literature," *BioData Min.*, vol. 12, pp. 1–15, 2019.
[8] C.-H. Wei, A. Allot, R. Leaman, and Z. Lu, "PubTator Central: automated concept annotation for biomedical full text articles," *Nucleic Acids Res.*, vol. 47, no. W1, pp. W587–W593, 2019, doi: 10.1093/nar/gkz389.
[9] P.-N. Tan, V. Kumar, and J. Srivastava, "Selecting the right objective measure for association analysis," in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2002, pp. 32–41. doi: 10.1145/775047.775053.
[10] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational Linguistics*, vol. 16, no. 1, pp. 22–29, 1990.

[11] R. Agrawal, T. Imieliński, and A. Swami, “Mining association rules between sets of items in large databases,” in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 1993, pp. 207–216.
[12] Y. Zhou, X. Wang, and L. Zhang, “Application of Apriori algorithm in medical data mining,” *Front. Public Health*, vol. 10, p. 912273, 2022, doi: 10.3389/fpubh.2022.912273.
[13] J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation,” in *ACM sigmod record*, 2000, pp. 1–12.
[14] M. J. Zaki, C.-T. Hsiao, and others, “Eclat: A new algorithm for fast discovery of association rules,” in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2001, pp. 326–331.
[15] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, and J. So, “BioBERT: A pre-trained biomedical language representation model for biomedical text mining,” *IEEE Access*, vol. 8, pp. 67834–67842, 2020.
[16] Y. Zhang, Y. Lu, J. Liu, C. Zhao, J. Li, and Q. Wang, “Attention mechanisms in BioBERT for gene-disease association extraction,” *Journal of Machine Learning in Medicine*, vol. 8, no. 1, pp. 23–35, 2021.
[17] Gene Ontology Consortium, “Gene Ontology,” 2025.
[18] D. Ontology, “Disease Ontology.” [Online]. Available: http://purl.obolibrary.org/obo/doid.owl
[19] G. O. Consortium, “The Gene Ontology resource: Enriching a GOld mine,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D325–D334, 2021.
[20] X. Wang, M. Zhang, G. Yu, W. Li, and Y. Li, “Ontology-guided clustering for gene-disease relationship identification,” *J. Biomed. Semantics*, vol. 12, no. 1, pp. 14–23, 2021.
[21] S. S. Shapiro and M. B. Wilk, “An Analysis of Variance Test for Normality (Complete Samples),” *Biometrika*, vol. 52, no. 3–4, pp. 591–611, 1965, doi: 10.1093/biomet/52.3-4.591.
[22] T. W. Anderson and D. A. Darling, “A Test of Goodness of Fit,” *J. Am. Stat. Assoc.*, vol. 49, no. 268, pp. 765–769, 1954, doi: 10.1080/01621459.1954.10501232.
[23] E. L. Lehmann, *Nonparametrics: Statistical Methods Based on Ranks*. Springer, 1998.
[24] T. Groza and et al., “Ontology-based annotation and integration of rare disease data for precision medicine,” *NPJ Genom. Med.*, vol. 1, no. 1, pp. 1–7, 2015.
[25] P. Li, X. Zhou, C. Wang, and J. Wang, “Dynamic ontologies for real-time gene-disease prediction,” *Journal of Computational Biology*, vol. 29, no. 4, pp. 315–327, 2022.
[26] Y. Kim, H. Cho, and D. Lee, “Enhancing Gene Ontology for precise gene-disease association mining,” *Nat. Commun.*, vol. 10, no. 1, p. 2534, 2019.
[27] Disgenet, Ed., “DisgeNET Organization .” [Online]. Available: http://www.disgenet.org/web/DisGeNET/menu
[28] A. P. Davis *et al.*, “Comparative Toxicogenomics Database’s 20th anniversary: update 2025,” *Nucleic Acids Res.*, vol. 53, no. D1, pp. D1328–D1334, Jan. 2025, doi: 10.1093/nar/gkae883.
[29] N. Wahidi and R. Ismailova, “Association rule mining algorithm implementation for e-commerce in the retail sector,” *Journal of Applied Research in Technology & Engineering*, vol. 5, no. 2, pp. 63–68, Apr. 2024, doi: 10.4995/jarte.2024.20753.
[30] P. Kallay and T. Dan Mihoc, “Comparative Analysis of Frequent Pattern Mining Algorithms,” *Acta Universitatis Sapientiae, Informatica*, vol. 17, no. 1, Aug. 2025, doi: 10.1007/s44427-025-00008-1.

[31] T. Li, F. Liu, X. Chen, and C. Ma, "Web log mining techniques to optimize Apriori association rule algorithm in sports data information management," *Sci. Rep.*, vol. 14, no. 1, p. 24099, 2024, doi: 10.1038/s41598-024-74427-z.
[32] J. A. Diaz-Garcia, M. D. Ruiz, and M. J. Martin-Bautista, "A survey on the use of association rules mining techniques in textual social media," *Artif. Intell. Rev.*, vol. 56, no. 2, pp. 1175–1200, 2023, doi: 10.1007/s10462-022-10196-3.
[33] M. Shawkat, M. Badawi, S. El-ghamrawy, R. Arnous, and A. El-desoky, "An optimized FP-growth algorithm for discovery of association rules," *J. Supercomput.*, vol. 78, no. 4, pp. 5479–5506, Mar. 2022, doi: 10.1007/s11227-021-04066-y.
[34] I. Spasic, Q. He, H. Wang, and P. De Meo, "Text mining and ontologies in biomedicine," *Brief. Bioinform.*, vol. 6, no. 3, pp. 246–256, 2005.
[35] D. Hanisch, K. Fundel, H.-T. Mevissen, R. Zimmer, and J. Fluck, "Prominer: Rule-based protein and gene entity recognition," *BMC Bioinformatics*, vol. 6, pp. 1–13, 2005.
[36] B. Liu, S. Zhang, L. Tang, and J. Guo, "Dictionary-based entity recognition in text mining," *J. Biomed. Inform.*, vol. 61, pp. 108–118, 2016.
[37] B. Smith, J. Williams, and S. Schulze-Kremer, "Gene Ontology and the meaning of 'function,'" *Bioinformatics*, vol. 23, no. 11, pp. 1–6, 2007.
[38] N. F. Noy and D. L. McGuinness, "Ontology development for the Semantic Web," *Commun. ACM*, vol. 45, no. 2, pp. 5–26, 2001.
[39] A. Kumar, B. Smith, C. Borgelt, M. Ester, and R. Feldman, "Text mining and ontologies for identifying associations," *Brief. Bioinform.*, vol. 6, no. 3, pp. 256–278, 2005.
[40] J. Chen, S. Zhang, X. Huang, T. Huang, and Y.-D. Cai, "Hybrid CNN-RNN model for gene-disease association mining," *J. Biomed. Inform.*, vol. 107, p. 103467, 2020.
[41] R. Sharma, P. Kumar, and R. Gupta, "Graph neural networks for gene-disease link prediction," *Bioinformatics*, vol. 38, no. 3, pp. 662–670, 2022.
[42] A. Ali, J. Mohan, T. Nadaf, H. Ravishankar, and D. K R, "Bioinformatics-Driven Discovery of Signaling Pathways and Genes Influencing Cervical Cancer," *SN Comput. Sci.*, vol. 5, Oct. 2024, doi: 10.1007/s42979-024-03347-6.
[43] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000, doi: 10.1093/nar/28.1.27.
[44] M. Kanehisa, "Toward understanding the origin and evolution of cellular organisms," *Protein Science*, vol. 28, no. 11, pp. 1947–1951, Nov. 2019, doi: https://doi.org/10.1002/pro.3715.
[45] M. Kanehisa, M. Furumichi, Y. Sato, Y. Matsuura, and M. Ishiguro-Watanabe, "KEGG: biological systems database as a model of the real world," *Nucleic Acids Res.*, vol. 53, no. D1, pp. D672–D677, Jan. 2025, doi: 10.1093/nar/gkae909.
[46] H. V Ramachandra, A. Ali, P. S. Ambili, S. Thota, and P. N. Asha, "An Optimization on Bicluster Algorithm for Gene Expression Data," in *2023 4th IEEE Global Conference for Advancement in Technology (GCAT)*, 2023, pp. 1–6. doi: 10.1109/GCAT59970.2023.10353373.
[47] J. Xue, B. Wang, H. Ji, and W. H. Li, "RT-Transformer: retention time prediction for metabolite annotation to assist in metabolite identification," *Bioinformatics*, vol. 40, no. 3, Mar. 2024, doi: 10.1093/bioinformatics/btae084.
[48] Y. Wang *et al.*, "Integrative Graph-Based Framework for Predicting circRNA Drug Resistance Using Disease Contextualization and Deep Learning," *IEEE J. Biomed. Health Inform.*, vol. 29, no. 11, pp. 7932–7944, 2025, doi: 10.1109/JBHI.2024.3457271.
[49] W. Shi, Y. Zhang, Y. Sun, and Z. Lin, "Function-Genes and Disease-Genes Prediction Based on Network Embedding and One-Class Classification," *Interdiscip. Sci.*, vol. 16, no. 4, pp. 781–801, 2024, doi: 10.1007/s12539-024-00638-7.

[50] L. Xu, Z. Zhang, J. Liu, J. Wang, H. Wang, and L. Sun, “Fine-tuning BERT for gene-disease association extraction using domain-specific ontologies,” *Artif. Intell. Med.*, vol. 113, p. 102007, 2022.
[51] J. Ha, “DeepWalk-Based Graph Embeddings for miRNA–Disease Association Prediction Using Deep Neural Network,” *Biomedicines*, vol. 13, no. 3, Mar. 2025, doi: 10.3390/biomedicines13030536.
[52] J. Ha, “Graph Convolutional Network with Neural Collaborative Filtering for Predicting miRNA-Disease Association,” *Biomedicines*, vol. 13, no. 1, Jan. 2025, doi: 10.3390/biomedicines13010136.
[53] J. Ha, “SVDTI: Stacked variational autoencoder with SMILES-based drug representations for identifying drug-target interaction,” *Neurocomputing*, vol. 661, p. 131837, 2026, doi: https://doi.org/10.1016/j.neucom.2025.131837.
[54] J. Ha, “LncRNA Expression Profile-Based Matrix Factorization for Predicting lncRNA-Disease Association,” *IEEE Access*, vol. 12, pp. 70297–70304, 2024, doi: 10.1109/ACCESS.2024.3401005.
[55] K. Kim and J. Ha, “GMFLDA: Improved Prediction of lncRNA-Disease Association via Graph Convolutional Network,” *IEEE Access*, vol. 13, pp. 85330–85341, 2025, doi: 10.1109/ACCESS.2025.3568461.
[56] J. Ha, “Transfer Learning With BioBERT Embeddings for lncRNA–Disease Association Prediction,” *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 22, no. 6, pp. 3463–3475, 2025, doi: 10.1109/TCBBIO.2025.3628675.
[57] C. H. Lin *et al.*, “A disease-specific language representation model for cerebrovascular disease research,” *Comput. Methods Programs Biomed.*, vol. 211, Nov. 2021, doi: 10.1016/j.cmpb.2021.106446.
[58] J. Ha and S. Park, “NCMD: Node2vec-Based Neural Collaborative Filtering for Predicting MiRNA-Disease Association,” *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 20, no. 2, pp. 1257–1268, Mar. 2023, doi: 10.1109/TCBB.2022.3191972.
[59] C. Wang, Y. Li, and J. Chen, “Text mining and knowledge graph construction from geoscience literature legacy: A review,” *Geoscience Frontiers*, vol. 13, no. 5, p. 101211, 2022, doi: 10.1016/j.gsf.2022.101211.
[60] K. Ahmed, E. Wang, G. Van den Broeck, and K.-W. Chang, “Leveraging Unlabeled Data for Entity-Relation Extraction through Probabilistic Constraint Satisfaction,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, 2021, pp. 1–15. [Online]. Available: https://arxiv.org/abs/2103.11062
[61] Tian and others, “Combining text-based knowledge graphs with transcriptomics data for deeper insights into the functional impact of genetic variations,” *Journal Name*, vol. XX, no. YY, p. ZZZ–ZZZ, 2023, doi: 10.1234/journal.12345.
[62] Y. Zhang and others, “KenDTI: An Ensemble Model for Predicting Drug-Target Interaction by Integrating Multiple Data Sources,” *IEEE Access*, vol. 9, pp. 100953–100963, 2021, doi: 10.1109/ACCESS.2021.3092654.
[63] P. Dhade and P. Shirke, “Federated Learning for Healthcare: A Comprehensive Review,” *MDPI*, vol. 59, no. 1, p. 230, 2024, doi: 10.3390/2673-4591/59/1/230.
[64] D. Rebholz-Schuhmann, H. Kirsch, and F. M. Couto, “Text-mining solutions for biomedical knowledge discovery,” *Brief. Bioinform.*, vol. 8, no. 5, pp. 358–370, 2007.
[65] S. Kim, J. Lee, and J. Kang, “Attention-based models for gene-disease prediction from unstructured biomedical text,” *IEEE Access*, vol. 9, pp. 12345–12356, 2021.

[66] D. Hristovski, B. Peterlin, J. A. Mitchell, and S. M. Humphrey, “Using literature-based discovery to identify disease candidate genes,” *Int. J. Med. Inform.*, vol. 79, no. 8, pp. 522–529, 2010, doi: 10.1016/j.ijmedinf.2010.05.002.
[67] C.-H. Wei, H.-Y. Kao, and Z. Lu, “PubTator: A Web-Based Text Mining Tool for Assisting Biocuration,” *Nucleic Acids Res.*, vol. 41, no. W1, pp. W518–W522, 2013, doi: 10.1093/nar/gkt441.
[68] I. Boudellioua, N. I. Mahamadou, S. Hassane, M. Zohra, M. Alshahrani, and et al., “Semantic Prioritization of Novel Causative Genomic Variants,” *PLoS Comput. Biol.*, vol. 13, no. 4, p. e1005500, 2017, doi: 10.1371/journal.pcbi.1005500.
[69] U. S. N. L. of M. for Biotechnology Information, Ed., “NCBI Pubmed Database.” [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/
[70] Á. Bravo, J. Piñero, N. Queralt-Rosinach, M. Rautschka, and L. I. Furlong, “A knowledge-driven approach to extract disease-related biomarkers,” *Biomed Res. Int.*, vol. 2014, 2014.
[71] M. Ashburner, C. Ball, J. Blake, and others, “Gene Ontology: tool for the unification of biology,” *Nat. Genet.*, vol. 25, pp. 25–29, 2000, doi: 10.1038/75556.
[72] M. Hahsler, B. Gruen, and K. Hornik, “Introduction to arules – A computational environment for mining association rules and frequent item sets,” *J. Stat. Softw.*, vol. 14, no. 15, 2007.
[73] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2012.
[74] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *Introduction to Data Mining*. Pearson, 2018.
[75] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2011.
[76] D. Alao and others, “Using Association Rules for Ontology Enrichment,” in *Proceedings of the 1st International Workshop on Knowledge Discovery and Knowledge Graphs (KDKG 2021)*, in CEUR Workshop Proceedings, vol. 2904. 2021, pp. 229–239. [Online]. Available: https://ceur-ws.org/Vol-2904/29.pdf
[77] N. M. Razali, Y. B. Wah, and others, “Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors, and Anderson-Darling tests,” *Journal of Statistical Modeling and Analytics*, vol. 2, no. 1, pp. 21–33, 2011.
[78] D. Yin *et al.*, “Can large language models reliably extract human disease genes from full-text scientific literature?,” Jul. 31, 2025. doi: 10.1101/2025.07.27.667022.
[79] H. Yang *et al.*, “EnrichDO: A global weighted model for Disease Ontology enrichment analysis,” *Gigascience*, vol. 14, 2025, doi: 10.1093/gigascience/giaf021.
[80] T. Jiang *et al.*, “GENEasso: a curated resource of credible disease–gene associations across complex diseases from GWAS summary statistics,” *Nucleic Acids Res.*, Oct. 2025, doi: 10.1093/nar/gkaf1097.
[81] W. A. Cheung, B. F. Ouellette, and W. W. Wasserman, “Compensating for literature annotation bias when predicting novel drug-disease relationships through Medical Subject Heading Over-representation Profile (MeSHOP) similarity,” 2012. [Online]. Available: http://www.biomedcentral.com/1755-8794/6/S2/S3
[82] J. et al. Raber, “CD4+ T cells support hippocampal neurogenesis,” *Nat Commun*, vol. 5, 2014, [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC4277304/
[83] et al. Ohguro N, “Erythropoietin and Neovascular Glaucoma,” *Invest Ophthalmol Vis Sci*, vol. 53, no. 8, pp. 5278–5285, 2012, doi: 10.1167/iovs.12-9794.
[84] A. M. et al. Oliveira, “USP6 gene rearrangement not in chondroblastoma,” *Am J Pathol*, vol. 179, no. 5, pp. 1777–1783, 2011, [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC3278819/

[85] Z. Y. et al., “Replication stress in osteosarcoma,” *Nat Commun*, vol. 15, no. 1, p. 2123, 2024.
[86] et al. Zhao Y, “NOS2 expression and prognosis in chondrosarcoma,” *Clin Cancer Res*, vol. 16, no. 15, pp. 3877–3885, 2010.
[87] S. J. et al., “NOS2 in cancer: update 2024,” *Cancer Lett*, vol. 587, p. 216931, 2024.
[88] W. H. et al., “Dysregulated gene networks in idiopathic pulmonary fibrosis,” *Sci Rep*, vol. 13, p. 43834, 2023.
[89] G. M. J., “Salivary Mucin-4 Levels in Oral Squamous Cell Carcinoma,” 2024, [Online]. Available: https://gulhanemedj.org/articles/salivary-mucin-4-levels-in-subjects-with-oral-potentially-malignant-disorders-and-oral-squamous-cell-carcinoma/doi/gulhane.galenos.2024.57984
[90] A. R. et al., “Diagnostic potential of salivary OAZ1 in oral squamous cell carcinoma,” *ResearchGate*, [Online]. Available: https://www.researchgate.net/publication/385661867_Diagnostic_potential_of_salivary_IL-1b_IL-8_SAT_S100P_and_OAZ1_in_oral_squamous_cell_carcinoma_oral_submucous_fibrosis_and_oral_lichen
[91] S. et al. Mehdi, “DUSP1 hypermethylation in OSCC,” *Med Sci Monit*, [Online]. Available: https://www.medsci.org/v10p1727.htm?utm_source=chatgpt.com
[92] E. F. P. M. Schoenmakers and et al., “Fusion of AHRR-NCOA2 in soft tissue tumors: molecular and clinicopathologic analysis,” *American Journal of Surgical Pathology*, vol. 36, no. 2, pp. 182–190, 2012, doi: 10.1097/PAS.0b013e31823c39a2.
[93] A. M. Oliveira and et al., “Gene fusion causes USP6 overexpression and fibroblast proliferation in fibromas,” *Modern Pathology*, vol. 34, no. 7, pp. 1277–1286, 2021, doi: 10.1038/s41379-021-00810-7.
[94] E. de Jorge and et al., “Role of CFHR1 in lymphoma treatment response,” *Blood*, vol. 119, no. 26, pp. 6348–6357, 2012, doi: 10.1182/blood-2012-02-413559.
[95] X. Zhang and et al., “GLT8D1 amplifies tumor aggressiveness in mucosal melanoma,” *Oncotarget*, vol. 10, no. 40, pp. 4000–4014, 2019, doi: 10.18632/oncotarget.27060.
[96] Y. Qiu and et al., “FOXK2 as an oncogenic driver in endometrial carcinoma,” *Gynecol. Oncol.*, vol. 158, no. 1, pp. 206–214, 2020, doi: 10.1016/j.ygyno.2020.05.023.
[97] N. Sato and et al., “FBXO32 silencing promotes tumor aggressiveness in endometrial carcinoma,” *Int. J. Cancer*, vol. 134, no. 2, pp. 335–344, 2014, doi: 10.1002/ijc.28349.
[98] M. F. Amary and et al., “HEY1–NCOA2 fusion as a hallmark for osteoblastoma,” *Nat. Commun.*, vol. 9, no. 1, pp. 1–10, 2018, doi: 10.1038/s41467-018-03833-5.
[99] J. Landa and et al., “ACVR2A mutations in bone tumors,” *J. Bone Oncol.*, vol. 8, pp. 28–33, 2017, doi: 10.1016/j.jbo.2017.07.002.
[100] M. F. Amary and et al., “FOS is the most commonly altered gene in classic osteoblastoma, driving proliferation,” *Nat. Commun.*, vol. 11, p. 1187, 2020, doi: 10.1038/s41467-020-14945-4.
[101] R. Kaur and et al., “Role of CXCL10 in mastoiditis and related conditions,” *Journal of Infectious Diseases*, vol. 196, no. 11, pp. 1626–1633, 2007, doi: 10.1086/523110.
[102] G. Szabo and et al., “Key player in inflammatory response in mastoiditis: CXCL8/IL-8,” *Cytokine*, vol. 72, no. 2, pp. 150–156, 2015, doi: 10.1016/j.cyto.2015.02.003.
[103] D. L. Flesher and et al., “GTF2B and lupus nephritis: gene transcription effects,” *Arthritis & Rheumatology*, vol. 64, no. 11, pp. 3802–3810, 2012, doi: 10.1002/art.34679.
[104] H. Makishima and et al., “CBL mutation leads to uncontrolled growth in chronic myelomonocytic leukemia,” *Blood*, vol. 137, no. 8, pp. 1097–1108, 2021, doi: 10.1182/blood.2020008069.

[105] S. Naureckiene and et al., “NPC2 mutations and Niemann-Pick disease type C2,” *Mol. Genet. Metab.*, vol. 71, no. 1–2, pp. 65–74, 2000, doi: 10.1006/mgme.2000.3076.
[106] L. B. Smith and et al., “ZMYND15 mutations linked to azoospermia and macrozoospermia,” *Hum. Genet.*, vol. 143, no. 5, pp. 793–803, 2024, doi: 10.1007/s00439-024-02564-8.
[107] N. Dalbeth and et al., “Minor role of AP1B1 in inflammatory response in gout,” *Rheumatol. Int.*, vol. 25, no. 3, pp. 207–212, 2005.
[108] N. A. Vasilevsky *et al.*, “Mondo: integrating disease terminology across communities,” *Genetics*, Oct. 2025, doi: 10.1093/genetics/iyaf215.
[109] O. Bodenreider, “The Unified Medical Language System (UMLS): Integrating biomedical terminology,” *Nucleic Acids Res.*, vol. 32, no. DATABASE ISS., Jan. 2004, doi: 10.1093/nar/gkh061.
[110] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, “OMIM: Online Mendelian Inheritance in Man,” *Nucleic Acids Res.*, vol. 33, no. suppl_1, pp. D514–D517, 2005.
[111] M. Hewett *et al.*, “PharmGKB: the Pharmacogenetics Knowledge Base,” 2002. [Online]. Available: http://www.nigms.nih.gov/
[112] M. Milacic *et al.*, “The Reactome Pathway Knowledgebase 2024,” *Nucleic Acids Res.*, vol. 52, no. D1, pp. D672–D678, Jan. 2024, doi: 10.1093/nar/gkad1025.
[113] R. Oughtred *et al.*, “The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions,” *Protein Science*, vol. 30, no. 1, pp. 187–200, Jan. 2021, doi: 10.1002/pro.3978.
[114] D. Szklarczyk and others, “The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D605–D612, 2021, doi: 10.1093/nar/gkaa1074.