Faculty of Medicine

SYSTEMS BIOLOGY MASTER PROGRAM

Vilnius University, Faculty of medicine

Ariadna Šamatovič

Master thesis

MeSH terminais paremtas fenotipo-geno sąsajų duomenų rinkinys skirtas genų prioritetizavimui

A MeSH-Term-Based Phenotype-Gene Association Dataset for Gene Prioritization

Supervisor Dr. Erinija Pranckevičienė

Head of the department Prof. Algirdas Utkus, PhD

Vilnius, 2026

Student's email ariadna.samatovic@mf.stud.vu.lt

CONTENTS

LIST OF ABBREVIATIONS **3**

INTRODUCTION **4**

LAY DESCRIPTION **6**

LITERATURE REVIEW **7**

1.1 MeSH Ontology 7

1.2. Literature mining for gene-phenotype relations 7

1.3. Gene Ontology 9

1.4. Integrative disease-chemical-GO relationships 10

1.5. Biomedical databases for gene-phenotype associations 11

1.6. Gene prioritization methods and computational tools 12

METHODS **15**

2.1 Data sources and data collection 15

2.2 Program architecture and workflow 16

2.3 Calculation of association scores and gene prioritization 18

2.4 Gene prioritization procedure 21

2.5 Gene network visualization 22

RESULTS **23**

3.1 Gene prioritization tool and graphical user interface 23

3.2 Data analysis 25

3.3 Evaluation of gene prioritization performance 28

3.4. Validation and biological interpretation of prioritized genes 29

DISCUSSION **33**

4.1. Overall performance of the gene prioritization tool 33

4.2. Effect of literature and database coverage 33

4.3. Interpretation of the scoring method 33

CONCLUSIONS **35**

ACKNOWLEDGEMENTS **36**

REFERENCES **37**

SUMMARY **41**

SUMMARY IN LITHUANIAN **42**

APPENDICES **43**

LIST OF ABBREVIATIONS

Gene Ontology (GO)
Biological Process (BP)
Cellular Component (CC)
Molecular Function (MF)
Inferred from Electronic Annotation (IEA)
Inferred from Experiment (EXP)
Inferred from Direct Assay (IDA)
Inferred from Mutant Phenotype (IMP)
Inferred from Sequence or Structural Similarity (ISS)
Comparative Toxicogenomic Database (CTD)
Human Phenotype Ontology (HPO)
Online Mendelian Inheritance in Man (OMIM)
DECIPHER (Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources)
Medical Subject Headings (MeSH)
Unified Medical Language System (UMLS)
Natural Language Processing (NLP)
MeSH Over-representation Profiles (MeSHOPs)
FACTA+ (Text-mining system for biomedical entity associations)
ToppGene (Gene prioritization tool using functional similarity and enrichment)
DISEASES (Database integrating curated and text-mined gene-disease associations)
Positional PubMed (PosMed)
PolySearch (Literature mining tool for biomedical entity associations)

INTRODUCTION

Understanding the relationship between genes and diseases is one of the main goals of biomedical research. Identifying these associations is essential for uncovering disease mechanisms and developing effective treatments. However, identifying such relationships using experimental methods is expensive and time-consuming [1]. Many meaningful relationships between genes and phenotypes can be revealed from biomedical research publications. Large databases such as PubMed, OMIM, and DisGeNET contain millions of scientific publications and biological associations. As a result, manual extraction of relevant information has become difficult and inefficient. Computational approaches such as text mining and ontology-based methods are widely used to overcome this challenge.

However, many biologically important associations are not directly reported in scientific articles or curated databases. Literature mining techniques are widely used to identify direct and indirect relationships between genes and diseases. These methods include co-occurrence analysis, natural language processing, and graph-based models. These approaches are also commonly applied in gene prioritization tools. Candidate genes are ranked according to their predicted relevance to a specific disease or phenotype. Using these approaches, hidden connections between genes and diseases can be revealed. Such relationships may be identified through intermediate entities, such as chemicals or biological processes. As a result, novel candidate genes can be detected even when no direct association has been previously reported [2-4].

Indirect phenotype-gene associations can be explored by combining different types of biological information. Chemicals are often involved in disease-related molecular mechanisms. They can serve as intermediate links between phenotypes and genes. Gene Ontology annotations provide additional functional context by describing the biological roles of gene products. Together, chemical associations and GO annotations can help connect phenotypes with genes through shared biological mechanisms. This approach can reveal biologically meaningful relationships that may not be detected through direct literature evidence alone [5-7].

This master's thesis is a continuation of previous studies performed by the supervisor Dr. Erinija Pranckevičienė. These studies introduced approaches for identifying indirect phenotype-gene associations. These approaches integrate MeSH-based literature mining with chemical intermediates and Gene Ontology annotations [8,9]. The present work further develops and applies these concepts for gene prioritization and exploration of phenotype-gene relationships.

In this master's thesis, a gene prioritization tool and database were developed to identify indirect links between diseases and genes. The system uses scientific publications from PubMed and combines MeSH disease (Category C) and chemical compound (Category D) terms with Gene Ontology (GO) annotations. Chemicals are used as intermediate links, while GO annotations provide information about gene function. By integrating these data, the system identifies and ranks genes that are most likely related to a selected disease. This approach aims to improve the discovery of disease-associated genes and provide an interpretable framework for exploring underlying biological relationships.

Aim of the study - to develop a database and a gene prioritization tool that enables the identification of indirect associations between phenotypes and genes, using MeSH disease (Category C) and chemical compound (Category D) terms, as well as Gene Ontology (GO) annotations.

Tasks:

1. To develop a structured database for storing phenotype, chemical, gene, and Gene Ontology annotation relationships.
2. To develop and evaluate a gene prioritization pipeline for identifying indirect phenotype-gene associations and ranking candidate genes.
3. To implement a graphical user interface and visualization module for interpretation of prioritization results and exploration of underlying biological relationships.

LAY DESCRIPTION

Investigation of relationships between genes and diseases is essential for improving diagnosis and developing new treatments. However, experimental methods used to identify these relationships are expensive and time-consuming. As a result, biomedical literature has become the main knowledge source. Millions of scientific articles are available in databases such as PubMed. However, extracting useful knowledge from them manually is extremely difficult.

According to that gene prioritization tool was developed. Implemented pipeline analyzes scientific literature from PubMed to uncover hidden connections between diseases and genes. Instead of relying only on direct evidence, the tool identifies indirect links. This is done by connecting diseases to chemicals (from MeSH terms in PubMed), and then linking these to genes and their biological functions using Gene Ontology annotations. The system integrates multiple data sources such as PubMed, Gene2GO, and Gene2PubMed. It ranks genes based on how strongly they are associated with a selected disease. It also provides simple visualizations that explain how each gene is linked through intermediate biological steps.

Overall, this tool offers a faster and more efficient way to identify genes related to diseases from PubMed. It does this by analyzing connections through intermediate steps involving chemicals and biological functions. It helps improve understanding of disease mechanisms by highlighting the key pathways linking diseases to genes. This can support future biomedical research and potential treatment development.

LITERATURE REVIEW

1.1 MeSH Ontology

Scientific publications in the biomedical field are continuously increasing over time [10]. In order to organize this information efficiently and make it accessible to millions of users, a standardized indexing system is required. The PubMed database currently holds over 39 million citations in biomedical literature and is based on the MEDLINE database [11]. PubMed employs the Medical Subject Headings (MeSH) ontology to index and categorize its articles [10]. Each article is manually annotated by MEDLINE experts with approximately 10-15 MeSH terms. The MeSH vocabulary, maintained by the National Library of Medicine, is updated annually and included over 40,000 terms as of 2025 [12].

MeSH is organized into 16 categories and follows a hierarchical structure. Each MeSH term has a tree number indicating its position in the hierarchy [13]. Terms can appear in multiple branches and have more specific child terms. Among these categories, “Diseases [C]” and “Chemicals and Drugs [D]” are particularly important for biomedical research. They provide a standardized way to represent phenotypes and chemicals. In biomedical literature mining, MeSH “Diseases [C]” terms are commonly used as a reliable source for phenotype identification. Free-text disease mentions, introduce variability in disease naming, synonym redundancy, and ambiguity arising from overlapping or context-dependent terminology. As a result, MeSH-based indexing offers higher precision and interpretability due to expert curation and the use of a controlled vocabulary.

MeSH encodes biomedical concepts in a hierarchical structure. The co-occurrence of MeSH terms within the same PubMed article can be used to identify relationships between diseases, chemicals, and genes [1, 2, 14]. Associations identified through MeSH annotations can be classified as direct or indirect. Direct associations arise when two entities, are mentioned together within the same publication. Indirect associations occur when entities are connected through intermediate concepts, such as chemicals or biological processes. Together, these associations enable the exploration of previously unrecognized phenotype-gene relationships through large-scale literature mining [15].

Despite its strengths as a controlled vocabulary, MeSH has several limitations that can negatively impact large-scale knowledge extraction. For instance, MeSH terms may not exist for new or emerging biomedical concepts. In addition, a delay commonly occurs between an article’s inclusion in PubMed and the assignment of MeSH indexing. Furthermore, many MeSH terms are associated with only a limited number of publications. Such patterns may reduce their effectiveness in topic modeling and relationship mining. Another limitation is that MeSH is an English language vocabulary and is mainly used for the MEDLINE database, which consists primarily of English language publications. Consequently, biomedical literature in other languages may be underrepresented in MeSH-based analyses [16][17].

1.2. Literature mining for gene-phenotype relations

Phenotypes are observable traits of an organism that arise from the interaction between genetic and environmental factors. To understand disease mechanisms and develop treatments, it is essential to analyze the relationships between genes and phenotypes. Many such associations are identified through experimental studies and are made publicly available through standardized databases [18]. As new biomedical literature is published, the amount of information is growing

exponentially. Manual extraction of the relevant phenotype-gene relationships is becoming increasingly challenging. That is why automated approaches, such as text mining, are necessary to capture these associations.

In recent decades, various text-mining approaches have been developed to investigate phenotype-gene associations [1]. These approaches include co-occurrence analysis, natural language processing and graph-based methods. Co-occurrence analysis is based on detecting how often genes, diseases, or chemicals appear together in the same article, abstract, or sentence. Natural language processing (NLP) methods analyze the meaning and context of words in biomedical text. By understanding the occurrence patterns of these terms, NLP can find more detailed relationships, like causal or regulatory links. Graph-based approaches create a network with entities as nodes and relationships as edges. By studying the formed network, complex patterns present in the data can be found. Machine learning-based text mining approaches can also be applied to discover relationships from biomedical literature [19]. One of the main advantages is that machine learning-based methods can combine and analyze multiple types of information. However, the performance of machine learning depends on the quality of the input data. High-quality and well-structured datasets are needed, especially for less-studied genes or diseases [1].

Direct and indirect relationships are investigated for phenotype-gene associations using text mining bioinformatics tools. Direct relationships are identified when terms co-occur frequently in the same articles or abstracts. Indirect associations, in contrast, are not based on direct co-occurrence and are built from hidden patterns or shared connections. For instance, intermediate entities such as chemicals, biological processes, or functional annotations are commonly used. These entities help link diseases and genes through indirect relationships. In biomedical literature, indirect relationships can provide meaningful associations that have not been explicitly reported in individual publications. As a result, new and previously unrecognized gene-disease associations can be discovered.

For example, approaches like MeSH Over-representation Profiles (MeSHOPs) allow the prediction of novel gene-disease associations. This method is based on comparing the MeSH term patterns associated with genes and diseases across PubMed abstracts. A profile is generated for each gene or disease. It is calculated based on which MeSH terms appear more frequently than expected across all related PubMed articles. These profiles are then compared using similarity scores. When a gene and a disease share similar MeSH term patterns, they are inferred to be related. This approach helps find indirect gene-disease links that are not explicitly reported in the literature [2].

Another indirect relationship-based text mining system is FACTA+. The system was developed to detect, rank, and visualize associations between biomedical entities. Information about genes, diseases (phenotypes), chemicals, and biological processes is collected from PubMed abstracts. Using controlled vocabularies and dictionary-based entity recognition techniques, named entities are identified in the text. FACTA+ then measures how often pairs of entities co-occur across a large collection of PubMed abstracts. Based on these co-occurrence patterns, associations between entities are statistically ranked. FACTA+ has shown that gene-disease associations can be detected even when they are not explicitly mentioned in the literature, through shared pathways or chemical compounds [3].

Graph-based approaches for indirect connections through intermediate nodes are presented in Indirect association and ranking hypotheses for literature-based discovery [4]. In this research, biomedical entities such as genes, diseases, and chemicals were extracted from PubMed literature and represented as nodes in a graph. Edges were constructed based on co-occurrence relationships observed across publications. Both direct and indirect connections between entities were captured in the resulting network. Graph-based scoring and ranking strategies were applied to prioritize hypotheses supported by multiple indirect paths. This approach allowed generating ranked gene-disease hypotheses that were not explicitly reported in individual articles [4].

It is important to mention that literature mining has several limitations that affect the reliability and completeness of extracted gene-phenotype associations. For example, extraction methods based on co-occurrence can produce results with high false positive rates. This occurs when the proximity of terms, such as gene and disease, is falsely interpreted as evidence. However, they can be mentioned together in unrelated contexts [20]. Publication bias in biomedical literature can also affect text mining results. Well-known diseases and genes are studied and reported more frequently than rare or less-studied conditions. As a result, potentially novel associations appear less often in the literature. This limits the discovery of novel or understudied disease mechanisms. Moreover, automated methods are limited by unclear or inconsistent terminology and varied naming conventions for genes and phenotypes. This includes the use of synonyms and abbreviations, which can lead to both false positives and false negatives in the extracted data [21]. These limitations should be considered when interpreting results and combined with other evidence.

1.3. Gene Ontology

The Gene Ontology (GO) provides a structured representation of gene (or gene product) function. This framework enables systematic functional analysis of genes from any cellular organism or virus [22]. GO terms are standardized labels assigned to genes. By using these terms, it is possible to compare gene functions across different studies and organisms. GO annotations are widely used in ontology-based text mining methods. Structured functional knowledge defined in GO can clarify how gene functions relate to observable traits and disease processes.

The Gene Ontology classifies gene functions into three main categories: molecular function (MF), biological process (BP) and cellular component (CC). Molecular function (MF) may involve a single protein, an RNA molecule, or a larger molecular complex. It describes what a gene product does at the molecular level, such as catalysis or regulation of transcription. Biological process (BP) indicates the role of a gene product in the overall activity of the cell or organism. These processes often involve multiple molecular functions working together. Cellular component (CC) indicates the location within the cell where the gene product is active. Structures like the plasma membrane, cytoskeleton, and membrane-bound compartments such as mitochondria are included [23][24].

GO annotations associate a gene or gene product with a GO term supported by evidence. The evidence indicate how the association was derived. In the GO knowledgebase, annotations include both a reference source and an evidence code. The reference sources are represented by the publication (PubMed ID) or database entry from which the information was obtained. The evidence code defines the type of support of such as experimental data (e.g., EXP, IDA, IMP) or computational inference (e.g., ISS, IEA) [23]. Because each annotation is associated with its supporting evidence, the assigned function can be traced back to the original literature or curated

database. Experimentally supported annotations provide high-confidence curated links between genes and functions, derived from published studies [22].

1.4. Integrative disease-chemical-GO relationships

Many genes that are known to cause diseases are not directly annotated in databases or mentioned in the literature as disease-causing. This creates a knowledge gap and makes it difficult to identify novel targets for therapy. To overcome this problem, intermediate biological relationships are used. They link diseases to genes through shared biological mechanisms. Network and integrative approaches can identify hidden gene-disease associations. This is achieved by integrating functional, phenotypic, and molecular interaction data [25].

Diseases are often linked to chemical entities like drugs, environmental exposures, toxins, or metabolites. Environmental exposures contribute to many human diseases and can affect molecular and biological processes. Chemicals affect biological systems by binding to proteins, disrupting signaling pathways, or affecting gene expression. Environmental chemicals can disrupt biological processes and cause diseases like inflammation, metabolic disorders, and endocrine disruption. This makes chemicals involved in disease-related processes [26]. As a result, chemicals can act as intermediate links between diseases and related genes. Even if a gene is not directly linked to a disease, its interaction with a disease-related chemical can help provide indirect evidence of its role in the same disease mechanism.

Gene Ontology (GO) annotations enable functional annotation of genes by describing their role in biological processes, molecular functions, and cellular components. Gene Ontology is a systematic and standardized vocabulary that facilitates functional annotation of genes in different biological databases and studies [5]. Shared biological processes between a disease and a chemical can help identify potential candidate disease genes. For instance, a disease could be linked to inflammation, and a chemical could be known to influence inflammation. Genes that are annotated to these pathways can then be considered potential disease genes. GO annotations can also be used to group genes according to their functions and identify relationships between genes and diseases [5]. This functional information helps interpret indirect relationships beyond simple co-occurrence.

Disease-chemical associations can be combined with functional Gene Ontology annotations. This allows diseases and genes to be linked through shared biological mechanisms. In this approach, a disease can be associated with a chemical, the chemical with a biological process, and the biological process with genes involved in that process. This multi-step method allows the identification of genes involved in disease mechanisms even if they have not been directly associated with the disease. Ontology-based integration allows functional annotations to be compared in a systematic way. This helps identify biologically relevant novel candidate genes [6].

Using multiple data sources increases confidence in indirect gene-disease associations. Literature annotations, biological databases, functional ontologies, and interaction networks can be integrated to provide stronger evidence. The more independent sources that support an association used, the higher the confidence in predicted relationship. This is especially important for complex diseases, where many pathways and environmental factors contribute [7].

1.5. Biomedical databases for gene-phenotype associations

Vast amounts of genotype and phenotype data are spread across numerous published studies, making them challenging to efficiently access and use [27]. The identification of gene-phenotype relationships depends on curated biomedical databases. They collect, standardize, and integrate information from experimental studies and large-scale genomic analyses. Such resources provide structured and accessible knowledge about genes, diseases, and their functional relationships. They enable direct gene-disease association studies. They also support computational methods for identifying new candidate genes.

Many biomedical databases of gene-disease and gene-phenotype relationships integrate heterogeneous data from multiple sources. These include experimental studies, clinical observations, genome-wide association studies, functional genomics research, and curated literature. By transforming heterogenous data into standardized formats, these databases enable systematic analysis of disease mechanisms and facilitate further investigations. The use of controlled vocabularies and ontologies, such as MeSH and Gene Ontology, ensures consistent representation of phenotypes and biological functions across datasets. As a result, curated databases provide a high-quality foundation for computational approaches aimed at identifying both known and novel disease-associated genes [28][29].

Online Mendelian Inheritance in Man (OMIM) is one of the most widely used databases for investigating connections between genes and diseases. It contains detailed, manually curated information on over 24 600 entries, including more than 16 000 genes and about 8 600 phenotypes. Manually curated entries are based on detailed clinical descriptions, molecular mechanisms, and references to supporting literature. Gene- diseases relationships are validated by expert curators based on experimental and clinical evidence. The database mainly contains information about Mendelian disorders. It also provides limited access to data on complex diseases that result from multiple genetic and environmental factors. OMIM offers a complete research and clinical platform by integrating clinical summaries, standardized phenotype classifications, and links to external resources. Its structured data and programmatic access allow large-scale computational analyses, making it possible to uncover novel gene-phenotype associations and patterns across many conditions [30].

In order to analyze associations across a broader range of diseases and evidence types, several integrative databases were developed. DisGeNET is one such resource that aggregates gene-disease associations from curated databases, genome-wide studies, animal models, and text-mining approaches. DisGeNET contains hundreds of thousands of associations linking genes to a wide range of human diseases, traits, and phenotypes. Controlled vocabularies and ontologies are used to keep the data standardized. The system assigns confidence scores to associations based on the type and number of supporting evidence sources. This scoring system allows to distinguish between well-established associations and less reliable links derived from computational predictions or limited data. By combining curated and automatically extracted information, DisGeNET enables exploration of both direct and indirect gene-phenotype relationships and supports large-scale network analyses [29].

Resources that integrate environmental factors are essential for understanding how chemicals interact with genes to cause diseases. The Comparative Toxicogenomic Database (CTD) focuses

on those interactions. The CTD stores manually curated information showing how various chemicals and drugs impact human gene expression. It also includes data on how environmental exposures influence disease development. The data are collected from peer-reviewed scientific publications, experimental studies, and public databases. In addition to manual curation, the database uses text-mining tools to help identify and prioritize relevant studies for review. The database currently includes over 1.6 million chemical-gene interactions, 400,000 gene-disease associations, and more than 200,000 chemical-disease links. Chemicals frequently serve as intermediaries that link gene activity to disease development. By capturing these connections, CTD enables the discovery of hidden gene-phenotype relationships. The database uses controlled vocabularies and ontologies, such as MeSH and Gene Ontology. This provides that chemicals, genes, and diseases are represented and stored in a standardized way [31].

Another essential resource for gene-phenotype associations is the Human Phenotype Ontology (HPO). It is a structured vocabulary and database that stores phenotypic abnormalities observed in human diseases. HPO standardizes clinical feature descriptions using a set of terms that include developmental abnormalities, organ system dysfunctions, and other observable traits. The database contains tens of thousands of phenotype terms and more than 250,000 phenotype annotations linked to over 10,000 diseases. These annotations link both rare and common diseases using data from OMIM, Orphanet, DECIPHER, and medical research studies. HPO annotations include metadata such as frequency, onset and references. This enables the study of how specific phenotypes develop in various medical conditions. The ontology organizes phenotypes in a hierarchy, allowing computational tools to perform phenotype-based and diagnosis-based analyses. HPO connects with other ontologies and biomedical resources using equivalence mappings, such as UMLS and MedDRA. This allows it to link genotype information with disease data for research and clinical use [32].

1.6. Gene prioritization methods and computational tools

The identification of genes causing genetic human disease results in large sets of candidate genes [33]. Experimental validation of all potential interactions is limited by time and financial resources [34]. To address this challenge, computational gene prioritization tools have been developed. These tools rank candidate genes according to their predicted relevance to a specific disease. Gene prioritization is widely used in disease gene discovery, variant interpretation, and hypothesis generation.

Gene prioritization methods rely on the principle that genes with matching phenotypes often share functional similarities. These genes may also participate in related biological pathways and disease processes. Tools for gene prioritization evaluate candidate genes using multiple types of biological data. The datasets includes molecular interaction networks, functional annotations, phenotype ontologies and literature-derived associations. Additional information sources such as sequence features, protein interactions, and gene expression profiles can also be used to improve candidate gene ranking. The process of prioritizing candidate genes depends on their degree of similarity to existing disease genes [35]. Combining multiple biological data sources enables identification of genes that are biologically relevant to a phenotype. Such relationships can be identified even if genes have not been directly linked to the disease in individual studies.

Multiple biological information sources are used to assess and rank candidate genes through gene prioritization methods. Different methods use different main information sources for their operations. Network-based methods use molecular interaction networks to analyze similarities between candidate genes and known disease genes. In such networks, genes located close to known disease genes have a higher probability to participate in similar biological processes [36].

Ontology-based methods compare functional annotations of candidate genes such as Gene Ontology or phenotype ontologies. They prioritize genes based on similar functional annotations or phenotype profiles. Highly ranked genes tend to be involved in similar biological functions. Ontology-based approaches are particularly useful for studying rare diseases and Mendelian disorders where phenotype descriptions are well defined.

Literature-based approaches identify gene-disease relationships through information extraction from scientific publications. These methods use text-mining techniques and controlled vocabularies, such as MeSH terms. This allows them to capture both direct and indirect associations. These approaches have been implemented in a number of gene prioritization tools. One widely used example is Endeavour. The system uses candidate genes by comparing them with a training set of known disease genes. The method integrates multiple data sources including gene expression data, functional annotations, sequence information, protein-protein interactions, and literature data. Each type of data generates a separate score for candidate genes based on their similarity to the training genes. The overall score is calculated by combining the individual scores to generate a global score. Endeavour allows identification of disease genes in genome-wide association studies and linkage analyses [37].

Another commonly used tool is ToppGene. The system ranks genes according to their functional similarities to known disease genes. ToppGene looks for important annotations in the training genes using different biological databases. These include Gene Ontology terms, biological pathways, protein interactions, gene expression data, phenotype terms, transcription factor binding sites, drugs, and microRNA targets. The algorithm takes a training gene set and a test gene set as inputs. Candidate genes are ranked based on the similarity of their annotations to those enriched in the training genes. Similarity scores are computed using statistical metrics such as correlation-based or fuzzy similarity methods. ToppGene also provides functional enrichment analysis, helping to interpret the biological significance of the selected genes [38].

The DISEASES database provides gene prioritization based on integrated evidence from curated databases, genome-wide studies, and text-mining approaches. The database integrates information from curated sources with gene-to-disease associations using text-mining approaches. Evidence from the different sources is combined. This provides confidence scores that indicate the strength of each gene-disease association. The DISEASES is an effective tool for literature-based prioritization. It integrates automatically extracted associations from biomedical publications with curated data sources [39].

Phenolyzer is a phenotype-driven gene prioritization tool designed for disease gene discovery. The system uses disease or phenotype terms as input and identifies candidate genes by integrating information from multiple biomedical databases. The resources used include gene-disease association databases, gene function annotations, molecular interaction networks, and pathway databases. Phenolyzer first identifies known disease genes. It then uses network-based predictions and functional similarity analysis to expand the list of candidate genes. The final ranking is based

on both direct gene-disease associations and predicted associations from biological networks. This approach allows the identification of both well-known disease genes and novel candidate genes [40].

Another important tool is GeneMANIA. GeneMANIA is a network approach for gene prioritization, which prioritize genes functionally similar to a given set of input genes. . GeneMANIA employs large-scale biological networks built from protein-protein interactions, genetic interactions, co-expression data, pathway data, and protein domain similarity. GeneMANIA weighs different types of networks automatically based on their relevance to the input genes. GeneMANIA expands the input gene set by finding genes that are strongly connected in biological networks. It then ranks all genes based on their connectivity to the original query gene set in the integrated network [41].

METHODS

This study presents a literature-based gene prioritization system for phenotype analysis. This section describes the developed model data sources, data collection procedure, computational workflow, score calculation, gene prioritization procedure, and gene visualization module.

2.1 Data sources and data collection

In this project several publicly available biological datasets where integrated. The datasets were used to retrieve phenotype-related publications, identify chemical MeSH terms, map publications to genes, and assign Gene Ontology (GO) annotations. Phenotype-related publications were retrieved from PubMed up to March 2026. Additional datasets, including Gene2PubMed, Gene2GO, MeSH vocabulary, and gene annotation files, were obtained from publicly available NCBI resources. Table 1 summarizes the datasets and data types used.

Table 1. Data sources used

Data source	Purpose	Data type
PubMed	Phenotype publications	PMIDs, MeSH terms
MeSH database	Chemical terms	MeSH D-terms
Gene2PubMed	Gene mapping	GeneID-PMID links
Gene2GO	GO annotations	GeneID-GO term links
Gene annotation file	Gene names	Gene symbols, gene links

PubMed Dataset. Phenotype-related publications were retrieved from PubMed database. Phenotype terms entered by the user through the graphical user interface were used to collect the data. Searches were performed using MeSH queries using the format "Phenotype"[MeSH Terms]. Phenotype-associated publications were retrieved from PubMed for the period 1940 to 2026**.** Publications were downloaded using the NCBI Entrez Programming Utilities [42]. For each publication the PubMed identifier (PMID) and associated MeSH terms were extracted. Each MeSH term was preprocessed by removing special characters and retaining the base term before any qualifier. This step produced phenotype-specific literature data with all MeSH terms assigned to the publications.

MeSH (D-Terms). Chemical MeSH terms were obtained from the Medical Subject Headings vocabulary [43]. Only MeSH terms belonging to category D (Chemicals and Drugs) were identified from extracted MeSH terms and used in further analysis. These terms were used as intermediate entities linking phenotype-related publications to genes and GO terms.

Gene2PubMed Dataset. Genes associated with phenotype-related publications were identified using the Gene2PubMed dataset. This dataset provides mappings between PubMed identifiers and

NCBI Gene identifiers. Only human genes (taxonomy ID 9606) were used. If phenotype-related publication is present in Gene2PubMed and linked to a GeneID, that gene is considered to be related to publication. For example, if a retrieved publication with PMID 12345678 is linked in Gene2PubMed to GeneID 7157, that gene is added to the set of genes associated with the phenotype. These gene-publication relationships allows identification of genes linked to phenotype-associated chemical MeSH terms.

Gene2GO Dataset. Functional annotations of genes were obtained from the Gene Ontology through the NCBI Gene resource [44]. This dataset links GeneIDs to Gene Ontology terms describing biological process, molecular function, and cellular component. These annotations were used to evaluate functional relationships between genes and chemical MeSH terms. For example, if GeneID 7157 is annotated in Gene2GO with GO term GO:0006915, this annotation is assigned to the gene and included in the analysis. These annotations were later used to evaluate how strongly each chemical term is connected to particular GO terms through shared genes.

The general data processing pipeline used in this study is illustrated in Figure 1. Starting from a phenotype MeSH term provided by the user, the system retrieves phenotype-related publications from PubMed, extracts chemical MeSH terms, and maps publications to genes and Gene Ontology annotations. These data are then used to calculate association scores and produce a ranked list of candidate genes.

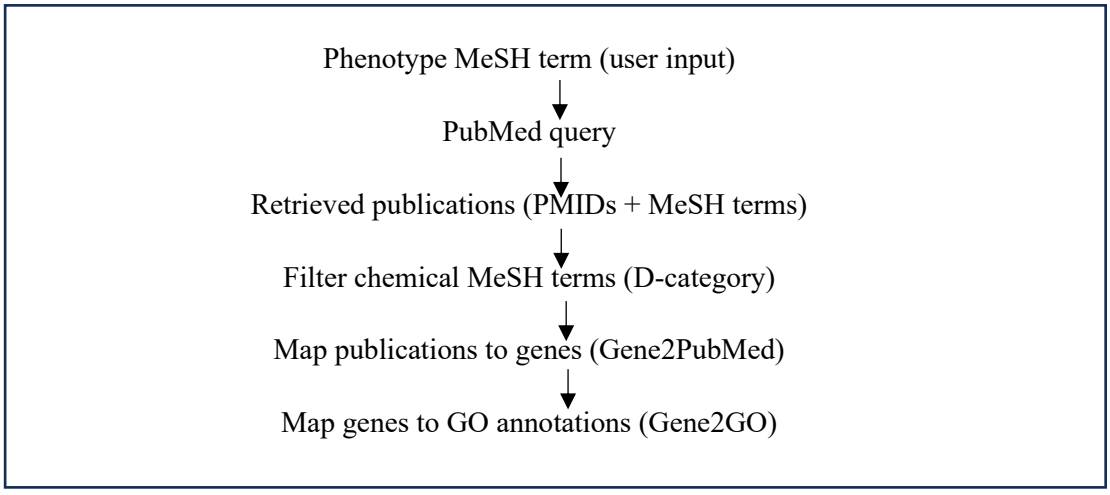

Figure 1. Data processing workflow used for gene prioritization.

2.2 Program architecture and workflow

The gene prioritization system was implemented in Python and accessed through a graphical user interface developed with Tkinter. The program accepts a phenotype term from the user, automatically retrieves phenotype-related publications from PubMed, extracts chemical MeSH terms, maps publications to genes, maps genes to GO annotations, calculates association scores, and returns a ranked list of genes and provides network visualization module.

The full workflow is shown in Figure 2. First, the user enters a phenotype term using the format - "Phenotype"[MeSH Terms], the number of top genes to display, and user email. If the entered

phenotype is not already stored in the local SQLite database, the program initiates the pipeline. The phenotype term is normalized for internal processing and used as a PubMed MeSH query. The pipeline collects all matched publications, extracts their PMID and MeSH terms. From these terms, only MeSH terms belonging to the D-category (chemicals and drugs) are selected. If a MeSH D-category term is present among publications extracted for phenotype defined, it is recorded as a chemical term associated with that publication. For each phenotype, all associated publications were analyzed to extract MeSH D-category terms. For each chemical term, global publication counts *nd* were obtained across the entire PubMed dataset, independent of the selected phenotype. Next, the program groups the retrieved publications by phenotype and chemical term. Then calculates the number of shared publications for each phenotype-chemical pair. These values are stored in the *mterm* table of a local SQLite database. The corresponding PMIDs are then mapped to human genes using Gene2PubMed. This produces, for each phenotype-associated chemical term, a set of genes supported by the relevant publications.

In the next step, the genes are mapped to GO annotations using Gene2GO. The resulting chemical-gene-GO combinations are stored in the *dterm* table. The GO identifiers, descriptions, and ontology categories are stored in the *go_term* table. After scoring, the program combines phenotype-chemical and chemical-GO association strengths(scores) to compute a final score for each gene. Genes are then ranked in descending order and stored in the *phenotype_genes* table.

The program also supports result visualization. Ranked genes are shown in the graphical user interface together with their normalized score, most influential chemical term, most influential GO term, and a NCBI gene link. In addition, the user can generate a network view for a selected gene. By entering gene name desired number of top-scoring chemicals and GO terms network is visualized. (Figure 2)

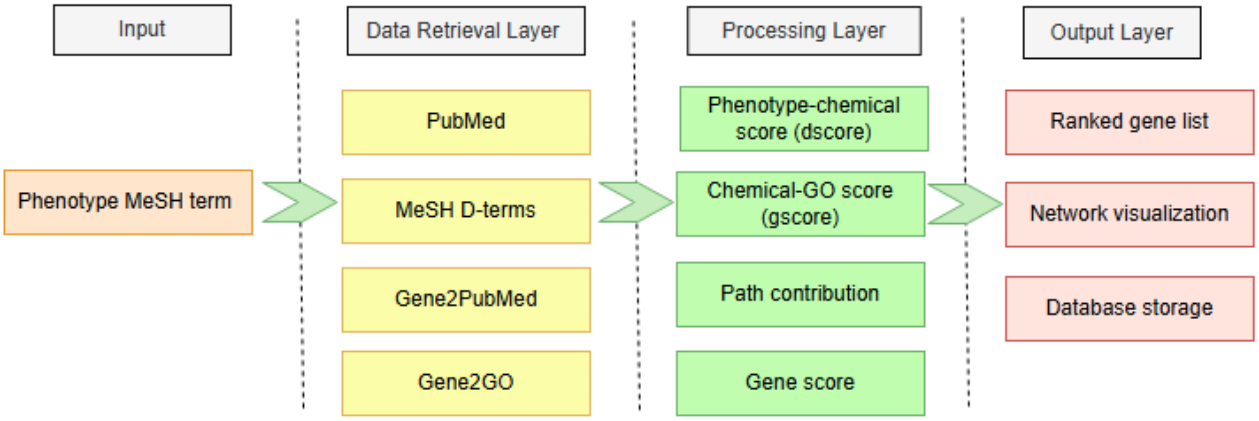

Figure 2. Program workflow

The SQLite database was used to store data and final gene prioritization results. For these purposes four tables: mterm, dterm, go_term and phenotype_genes were created. The mterm table stores phenotype-chemical associations and chemical scores. The dterm table stores relationships between chemicals, genes and GO terms. The go_term table contains Gene Ontology term descriptions. The phenotype_genes table stores calculated gene prioritization results. Database structure is illiustrated in Figure 3.

The **mterm** table stores phenotype-chemical relationships and chemical scores. Each row represents one phenotype-chemical pair. The column *mid* is a unique identifier of the phenotype-chemical relationship. The column *mterm* contains the phenotype MeSH term and column *dterm* contains the chemical MeSH term. The column *inters* stores the number of publications shared between phenotype and chemical terms, while *pmids* contains the list of shared PubMed identifiers. The column *nm* represents the number of publications associated with the phenotype and *nd* represents the number of publications associated with the chemical term. The column *unio* stores the union of phenotype and chemical publications. The column *dscore* contains the calculated chemical score and *dtid* is a unique identifier of the chemical term.

The **dterm** table stores relationships between chemicals, genes and GO annotations. Each row represents a chemical-GO association. The column *id* is a unique record identifier and *dtid* links the table to the mterm table. The column *dterm* contains the chemical MeSH term and *gene* contains GeneID identifiers associated with the chemical term. The column *goterm* contains Gene Ontology identifiers and *gokey* links the record to the go_term table. The column *genenum* represents the number of genes shared between the chemical term and GO annotation. The column *gogenes* represents the number of genes annotated by the GO term and *genetot* represents the total number of genes associated with the chemical term. The column *gscore* contains the calculated chemical-GO association score.

The **go_term** table stores Gene Ontology term descriptions. The column *gokey* is a unique identifier of a GO term. The column *goterm* contains the GO identifier, *description* contains the GO term name, and *category* indicates the ontology category.

The **phenotype_genes** table stores final gene prioritization results. The column *mid_val* links the results to the phenotype record in the mterm table. The column *phenotype* contains the phenotype name and *genes_txt* stores ranked gene lists.

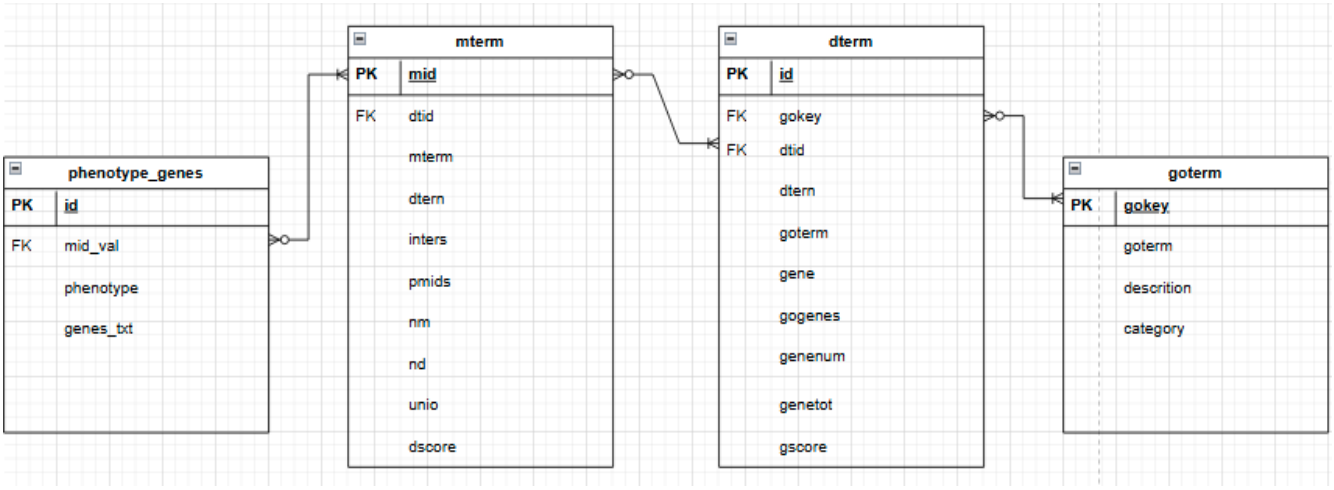

Figure 3. Database structure

2.3 Calculation of association scores and gene prioritization

All scores were calculated from the chemicals, publications, genes, and GO annotations obtained for the input phenotype during the pipeline run. The score definitions correspond to the variables

stored in the database tables. In the *mterm* table, the variable *nm* represents the number of publications associated with the phenotype, *inters* represents the number of publications shared between the phenotype and a chemical MeSH term, and *nd* represents the total number of PubMed publications annotated with the corresponding chemical MeSH term. The calculated phenotype-chemical association score is stored as *dscore*. In the *dterm* table, *genenum* represents the number of genes shared between a chemical term and a GO annotation, *gogenes* represents the number of genes annotated with the GO term, and *genetot* represents the number of genes associated with the chemical term. The resulting chemical-GO association score is stored as *gscore*.

For each phenotype-chemical pair, the number of publications containing both the phenotype MeSH term and the chemical MeSH term is calculated as:

$$inters = |PMID_{phenotype} \cap PMID_{chemical}| \quad (1)$$

where *inters* represents the number of publications shared by the phenotype and the chemical term. The total number of publications retrieved for the phenotype is stored as:

$$nm = |PMID_{phenotype}| \quad (2)$$

and the number of publications associated with the chemical term is stored as:

$$nd = |PMID_{chemical}| \quad (3)$$

Using these values, the phenotype-chemical association score (*dscore*) is calculated as:

$$dscore = \frac{inters}{nd} \quad (4)$$

This score reflects how frequently a chemical term co-occurs with the phenotype within the analyzed literature set. An example of phenotype-chemical *dscore* calculation is shown in Figure 4.

Genes associated with each chemical term are identified by mapping phenotype-chemical shared publications to genes using the Gene2PubMed dataset. These genes are then mapped to Gene Ontology annotations using the Gene2GO dataset. Only GO terms assigned to the identified genes are considered in the subsequent analysis. If G_d denotes the set of genes associated with a chemical term d, and G_{go} denotes the set of all genes annotated with a GO term go in the Gene2GO dataset, then the number of genes shared by the chemical term and the GO annotation is calculated as:

$$genenum = |G_d \cap Ggo| \quad (5)$$

The total number of genes annotated with the GO term is:

$$gogenes = |Ggo| \quad (6)$$

and the total number of genes associated with the chemical term is:

$$genetot = |G_d| \quad (7)$$

Using these values, the chemical-GO association *gscore* is calculated as:

$$gscore = \frac{genenum}{\sqrt{gogenes \cdot genetot}} \tag{8}$$

An example of chemical-GO-gene *gscore* calculation used for gene prioritization is illustrated in Figure 5.

The *gscore* score reflects the strength of functional association between a chemical term and a GO annotation through their shared genes. The square-root denominator serves as a size-normalization factor, reducing the influence of very large chemical-associated or GO-annotated gene sets on the final score.

Gene prioritization is performed by integrating phenotype-chemical and chemical-GO relationships. For each chemical-GO pair, the contribution of the path is calculated as:

$$Contribution = dscore \times gscore \tag{9}$$

Each gene may be connected to multiple chemical terms and GO annotations. Therefore, the final score for a gene is obtained by summing all contributions associated with that gene:

$$GeneScore(g) = \sum (dscore \times gscore) \tag{10}$$

where the summation includes all chemical-GO pairs connected to gene *g*.

Summation was used to integrate evidence across multiple phenotype-chemical-GO paths. This allows genes supported by several independent associations to receive higher scores compared to genes supported by a single strong association.

To enable comparison between genes within the same phenotype analysis, the final gene scores are normalized by the highest gene score obtained for that phenotype:

$$GeneScore_{norm}(g) = \frac{GeneScore(g)}{\max(GeneScore)} \tag{11}$$

Genes are ranked in descending order according to the normalized score.

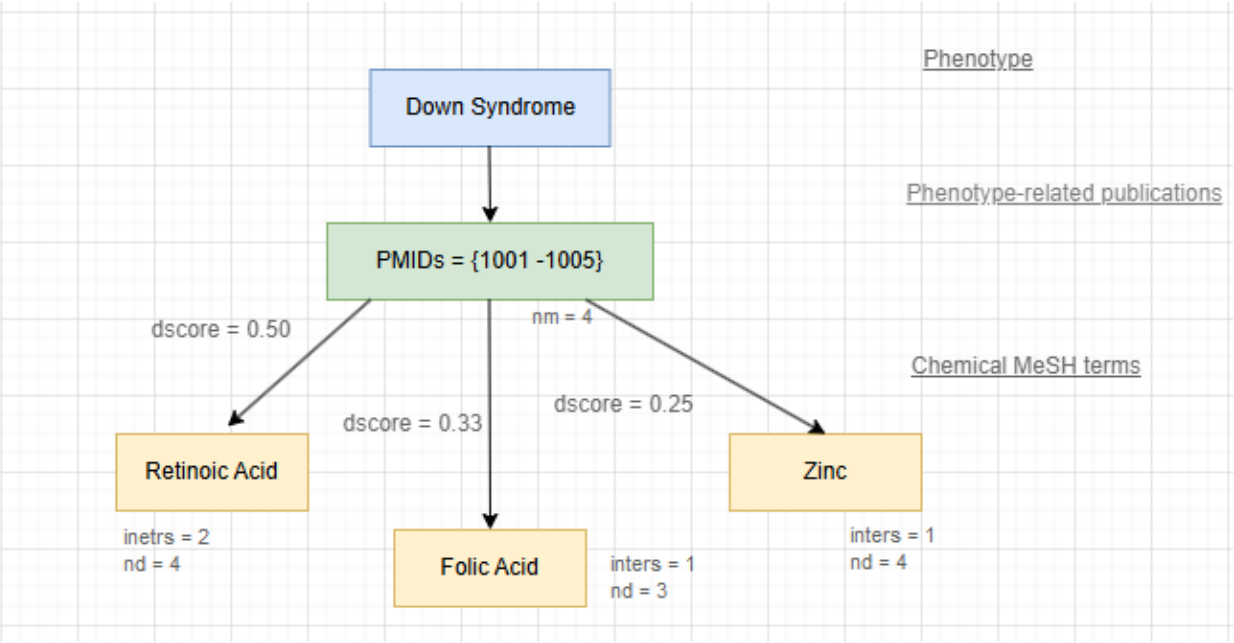

Figure 4. Phenotype-chemical association score calculation (dscore)

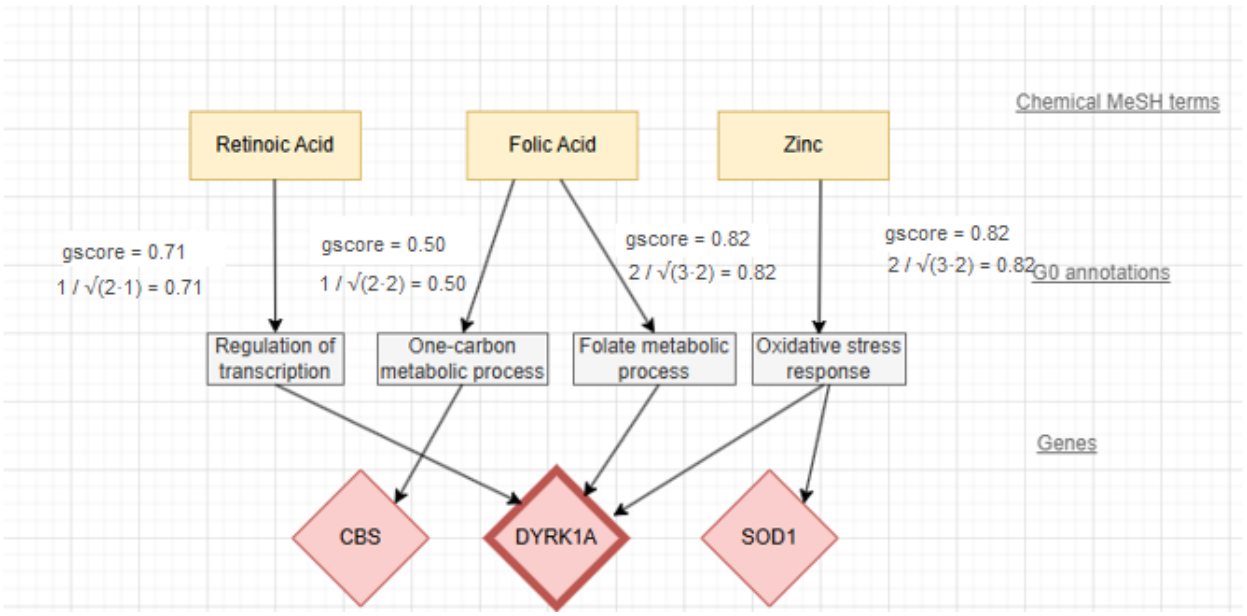

Figure 5. Chemical-GO-gene association score calculation (gscore)

The phenotype, chemicals, genes, and scores shown in Figures 4 and 5 are illustrative examples used to explain the scoring procedure and do not represent the actual output of the implemented system.

2.4 Gene prioritization procedure

Gene prioritization was implemented as an automated Python pipeline. After the user enters a phenotype MeSH term, the program normalizes the input and converts it into a corresponding MeSH-based query("Phenotype [MeSH Terms]"). Then program retrieves phenotype-associated

PubMed publications and extracts all MeSH annotations from those records. Chemical MeSH terms are identified by filtering for MeSH D-category terms. The retrieved PMIDs are then matched to human genes using Gene2PubMed. The identified genes are further mapped to GO annotations using Gene2GO.

For each phenotype, the program calculates phenotype-chemical association scores and chemical-GO association scores. The values are combined into path contributions, and all contributions are summed for each gene. (See section 2.3) The final ranked table contains the gene symbol, the normalized score, the dominant chemical term, the dominant GO term, and a direct NCBI gene link. Results are saved in the SQLite database and also can be exported.

2.5 Gene network visualization

A gene network visualization was implemented to support interpretation or relationships of ranked results. After gene prioritization, a gene from the ranked list can be selected and a network is generated. The visualization is created from the calculated scores stored in the database. For the selected gene, the program retrieves specified number of associated chemical and GO paths. Contribution of each path is calculated as the product of the phenotype-chemical score and the chemical-GO score, and top-scoring chemical terms and GO annotations are presented in ranked results table. A directed graph is then created with four types of nodes: the phenotype, chemical terms, GO annotations, and the selected gene. Edges represent the inferred path from phenotype to chemical, from chemical to GO annotation, and from GO annotation to gene.

The graph is displayed using NetworkX and Matplotlib libraries. To maintain readability of the network, the number of displayed chemical terms and GO annotations per chemical can be specified as visualization parameters. This visualization helps the user understand why a gene was prioritized by showing the main intermediate links contributing to its final score.

RESULTS

3.1 Gene prioritization tool and graphical user interface

The gene prioritization software tool with a graphical user interface was implemented using the Python programming language. The program integrates multiple processes into a single automated workflow, including: literature retrieval, data integration, gene prioritization, and network visualization. The graphical user interface was implemented using the Tkinter library, while SQLite was used to store datasets and final prioritization results. The program source code and documentation are publicly available through a GitHub repository (link: https://github.com/Ariadna21/gene_prioritization_tool_). The program can be installed on a personal computer and executed after the required Python libraries are installed. After the program is launched, users can perform phenotype-based gene prioritization through interaction with the graphical user interface.

Figure 6 illustrates the main program interface. The user is required to enter a phenotype defined by a MeSH disease term and specify the number of top genes to display in the prioritization results table. In this example, Down Syndrome is entered as the phenotype and the number of prioritized genes is set to 10. User email is also specified.

After the input parameters are provided and the analysis is started. Program checks whether information for the specified phenotype has already been computed and stored in the SQLite database. If the phenotype has not been processed previously, the program automatically starts collecting the required data. During execution, the program retrieves phenotype-associated publications from PubMed and extracts their MeSH annotations. Chemical MeSH terms are identified and linked to phenotype-associated publications. The retrieved publications are then mapped to human genes using the Gene2PubMed dataset. The identified genes are then connected to Gene Ontology annotations through the Gene2GO dataset. The detailed prioritization procedure and scoring methodology are described in 2.3 and 2.4 sections.

After the analysis is completed, a second interface window displays the ranked list of genes associated with the selected phenotype and network visualization settings. (Figure 7) The results table includes the following information for each prioritized gene: the gene symbol, normalized prioritization score, dominant chemical term contributing to the score, dominant Gene Ontology annotation, and a direct link to the corresponding NCBI Gene database entry. The results can also be exported from the interface in .csv format for further analysis. In network visualization feature, user can select a gene from the prioritized list and specify the number of dominant chemical terms and Gene Ontology annotations to include in the network visualization. After the visualization command is executed, the program generates a network. It represents the relationships between the phenotype, chemical terms, GO annotations, and the selected gene. Figure 8 illustrates a network visualization of the DYRK1A gene of Down Syndrome, showing the top three dominant chemicals and the three dominant GO terms for each chemical. The network illustrates the strongest intermediate links contributing to the prioritization score and highlights the main biological pathways connecting the phenotype to the gene.

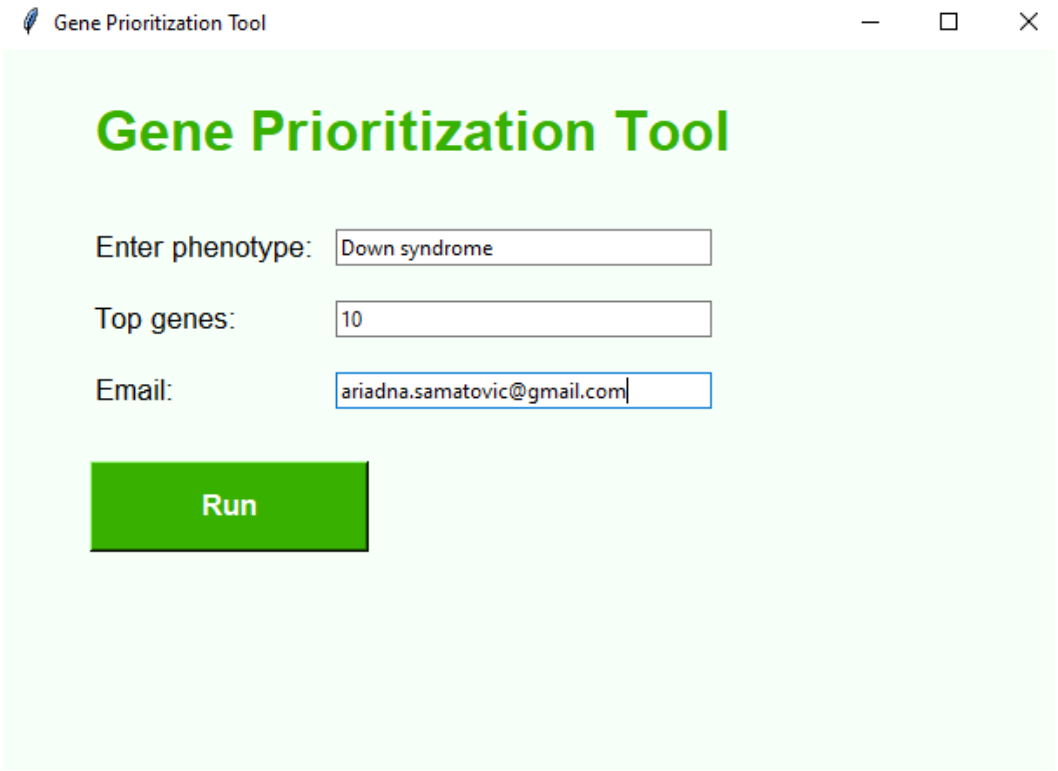

Figure 6. Graphical user interface for gene prioritization

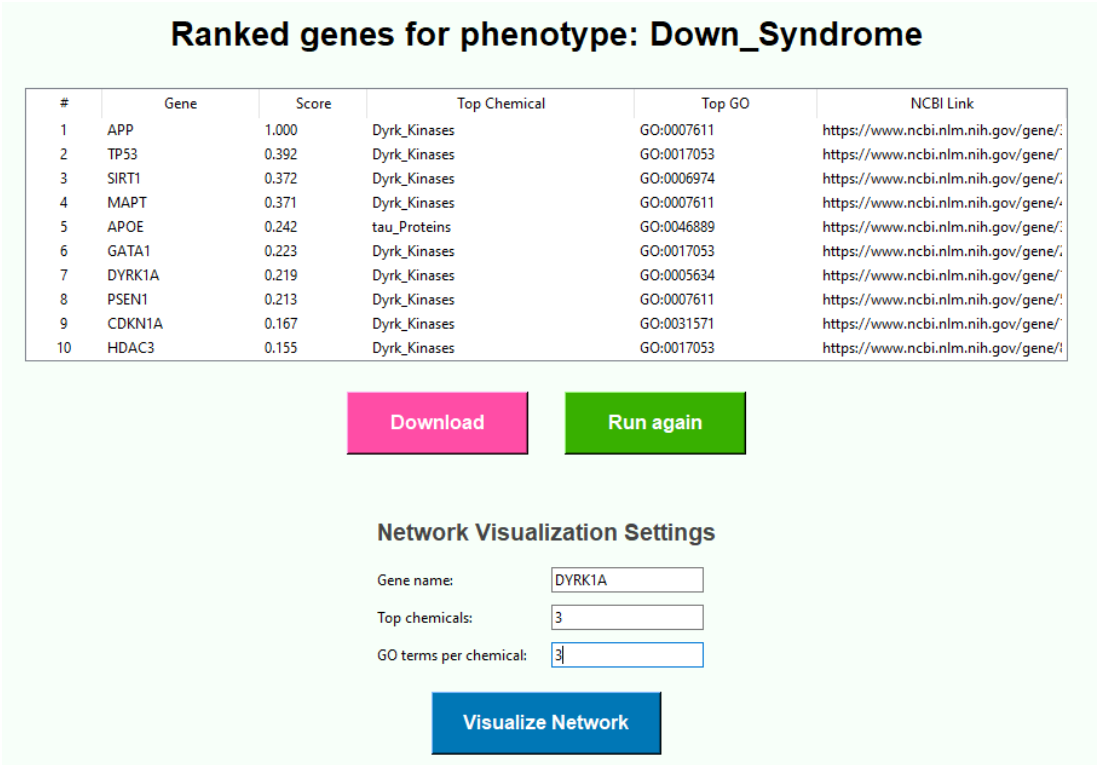

Figure 7. Gene prioritization results and network visualization settings in the graphical user interface.

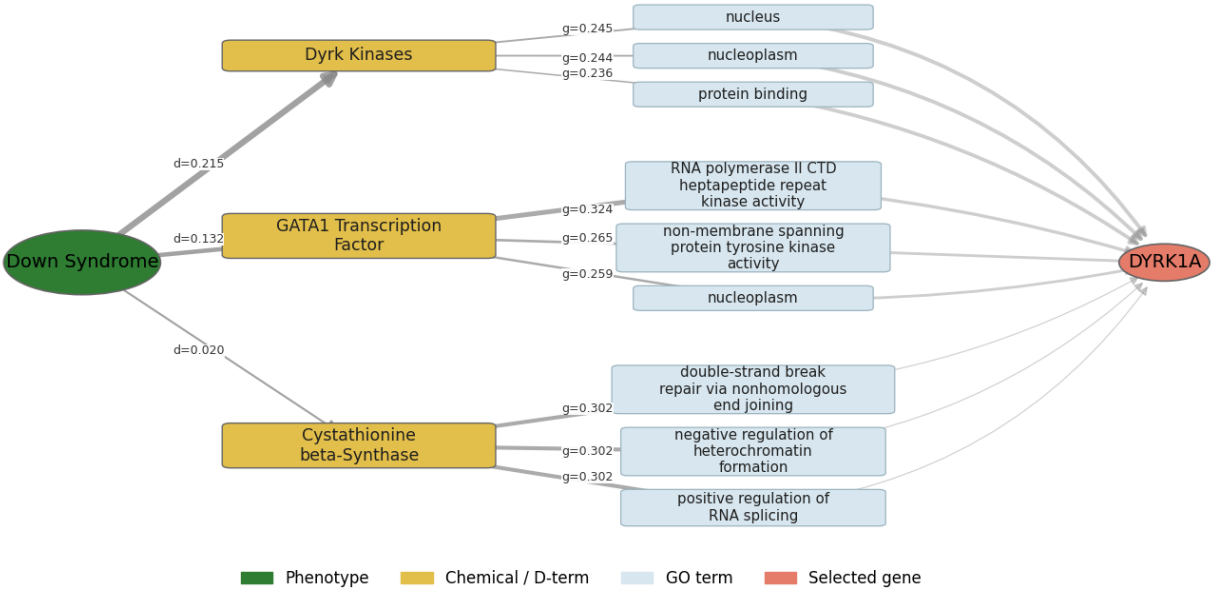

Figure 8. Network visualization of selected DYRK1A gene

3.2 Data analysis

A dataset of 75 phenotypes related to neurological disorders was constructed using MeSH disease terms from the Nervous System Diseases (C10) category. Including Retinitis Pigmentosa due to its involvement in neurodegenerative processes affecting photoreceptor neurons. For each phenotype, phenotype-associated publications were retrieved from the PubMed database using MeSH-based queries. In total, 83,635 publications were collected across all analyzed phenotypes. The complete dataset statistics for each phenotype, including the number of retrieved publications, unique chemical terms, genes, and GO annotations, are provided in Appendices Table 2. The complete list of analyzed phenotypes is also presented in Appendices Table 1. An example of the collected data is shown in Table 2.

Table 2. Example statistics of the collected dataset

Phenotype	Publications (nm)	Unique chemicals	Unique genes	Unique GO terms
Williams Syndrome	542	465	606	2399
WAGR Syndrome	61	87	11	87
Retinitis Pigmentosa	6583	2117	1992	5127

Further statistical analysis of the collected data was performed. Figure 9 presents the distribution of the number of PubMed publications across selected phenotypes included in the dataset. The figure shows substantial variability in literature coverage, with some phenotypes associated with

thousands of publications while others have relatively limited available literature. This variation reflects differences in research attention and data availability among diseases.

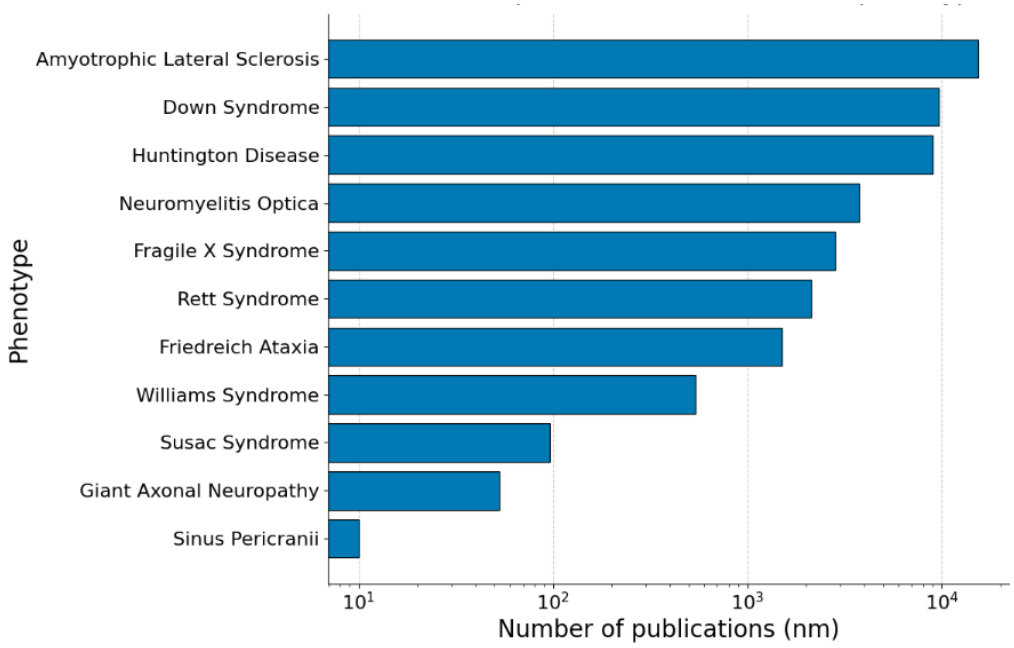

Figure 9. Distribution of publications across phenotypes

Relationship between the number of phenotype-associated publications and the number of retrieved genes is presented in Figure 10. Each point represents a phenotype. Positive correlation is observed. The calculated correlation coefficients indicate a moderately strong relationship between literature volume and gene coverage (Pearson $r = 0.74$, Spearman $\rho = 0.68$). This suggests that phenotypes with more extensive literature tend to have a larger number of associated genes.

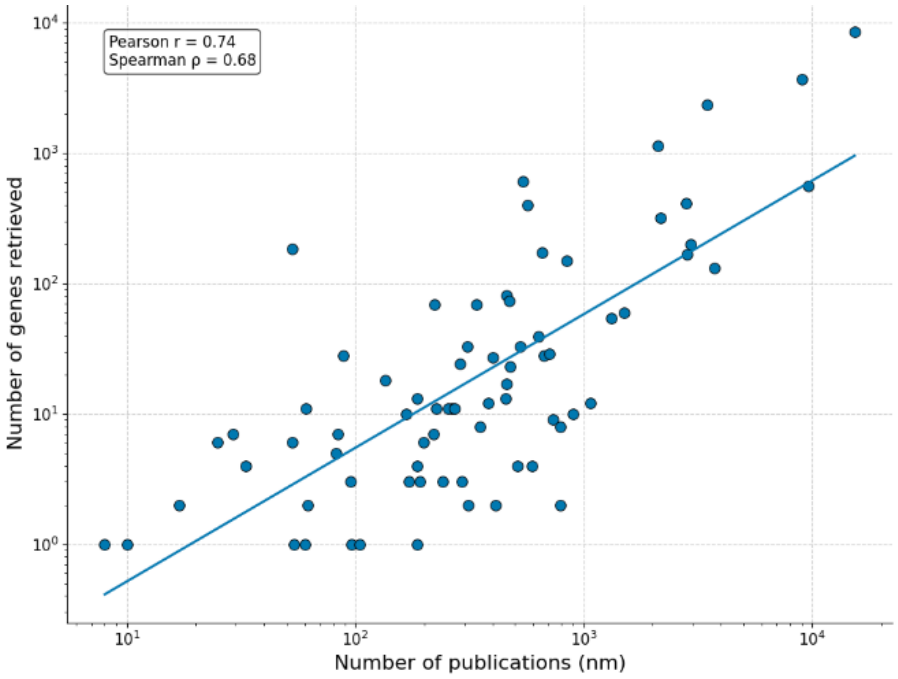

Figure 10. Number of publications and retrieved genes per phenotype

Figure 11 presents the relationship between the number of unique genes and the number of associated Gene Ontology annotations for each phenotype. A very strong positive correlation is observed between these variables (Pearson r = 0.96). This result indicates that phenotypes with a larger number of retrieved genes also tend to have a higher number of functional annotations.

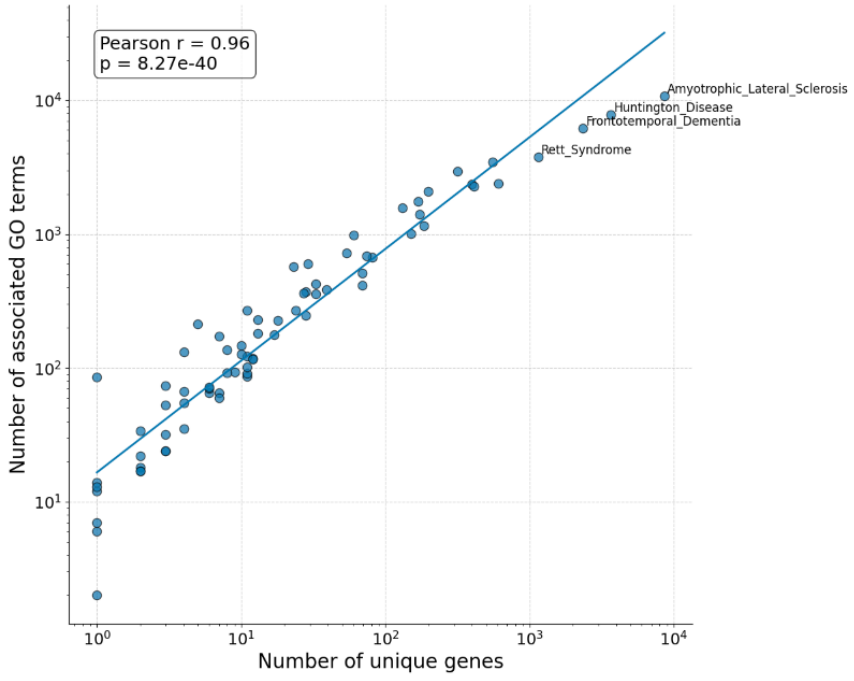

Figure 11. Number of unique genes and associated GO terms per phenotype

3.3 Evaluation of gene prioritization performance

The performance of the gene prioritization tool was evaluated by comparing the ranked gene lists with known disease genes obtained from the OMIM and DisGeNET databases. Genes present in either OMIM or DisGeNET were treated as known disease genes and used as the gold standard for evaluation.

For each phenotype, a ranked list of candidate genes was generated. The evaluation was performed using the top 5 ranked genes. All phenotypes were included in the evaluation. To ensure unbiased results, phenotypes where no known genes were identified among the top-ranked results were also included.

Several evaluation metrics were calculated to assess the prioritization performance. Recall@5 represents the fraction of known disease genes identified within the top 5 ranked candidates. In addition, the first true gene rank was recorded to determine the position of the first correctly predicted disease gene in the ranking. Based on this, the Reciprocal Rank (RR) was calculated as the inverse of the rank of the first true gene. The mean reciprocal rank (MRR) was computed across phenotypes where at least one known gene was identified. A summary of the evaluation results for all analyzed phenotypes is provided in Appendices Table 3. An example of the calculated evaluation values is shown in Table 3.

Table 3. Example evaluation results of gene prioritization performance

Phenotype	Gold standard genes	True genes in top-5	Recall@5	First true gene rank	RR
Williams Syndrome	8	4	0.5	2	0.5
WAGR Syndrome	6	3	0.5	1	1
Vein of Galen Malformations	2	2	1	1	1
Trisomy 13 Syndrome	5	1	0.2	2	0.5

Figures 12 show the distribution of Recall@5 across phenotypes. The results demonstrate variability in model performance. A subset of phenotypes exhibit values equal to 0. This indicates that no known disease genes were identified among the top five ranked candidates. This was observed for Thalamic diseases, Susac syndrome, Polymicrogyria, and several other phenotypes. In contrast, several phenotypes achieved moderate to high recall values. For example, Williams syndrome (Recall@5 = 0.5), WAGR syndrome (Recall@5 = 0.5), and Small fiber neuropathy (Recall@5 = 0.75) demonstrate effective prioritization. Performance varied significantly across phenotype categories. The model achieved higher performance for monogenic and well-characterized diseases. However, performance was limited for heterogeneous or non-genetic phenotypes. This highlights the dependence of literature-based prioritization on well-defined gene-disease relationships.

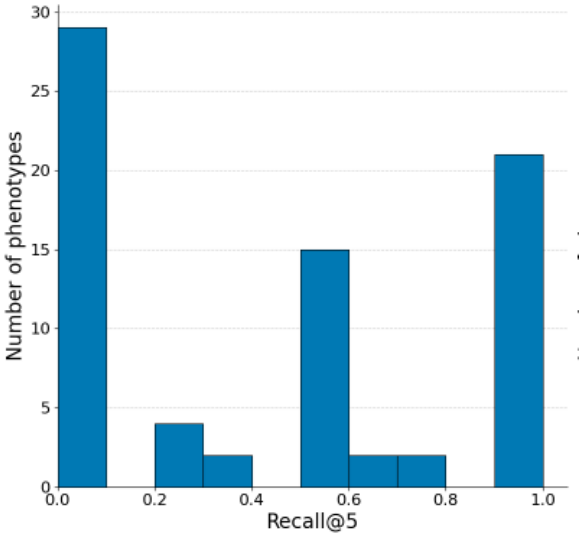

Figure 12. Distribution of Recall@5 across phenotypes

Figure 13 presents the distribution of the first true gene rank across all evaluated phenotypes. In most cases, the first correctly predicted disease gene appears at rank 1, resulting in reciprocal rank values close to 1.0. This result indicates that the prioritization algorithm frequently places known disease genes at the top of the ranking. The mean reciprocal rank (MRR) = 0.73 was calculated over phenotypes with at least one correctly identified gene. The result indicates that relevant genes were typically ranked among the top positions.

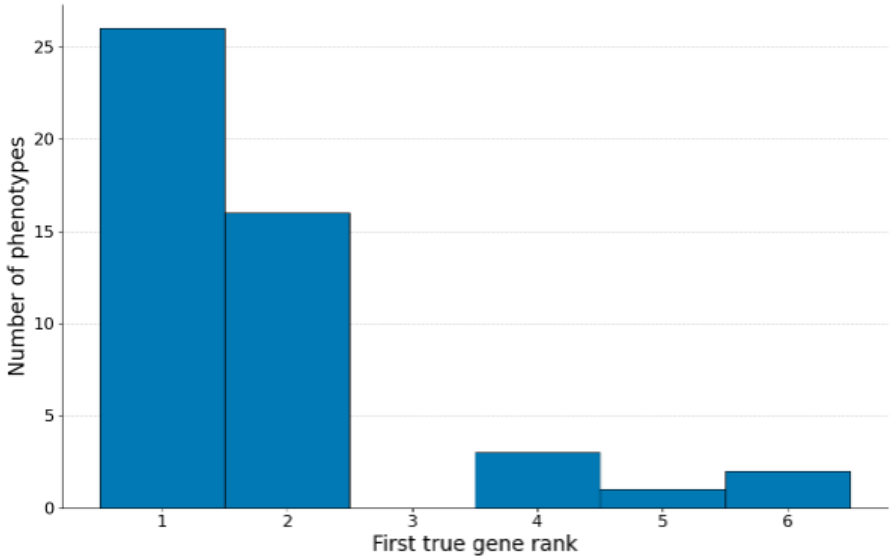

Figure 13. Distribution of first true gene rank across phenotypes

3.4. Validation and biological interpretation of prioritized genes

For validation and interpretation of the developed tool results, the phenotype Retinitis Pigmentosa was analyzed. Gene prioritization results are shown in Figure 14. The analysis of the top 10 genes

was made. Validation against OMIM and DisGeNET showed that 8 out of 10 genes are known disease-associated genes. Four genes (RHO, ABCA4, USH2A, PRPF31) are directly linked to non-syndromic retinitis pigmentosa. Another four (MYO7A, USH1C, USH1G, ADGRV1) are associated with syndromic retinal degeneration (mainly Usher syndrome). The remaining two genes (VCP, GJB2) are not directly associated with retinitis pigmentosa. Therefore, they were considered candidate genes.

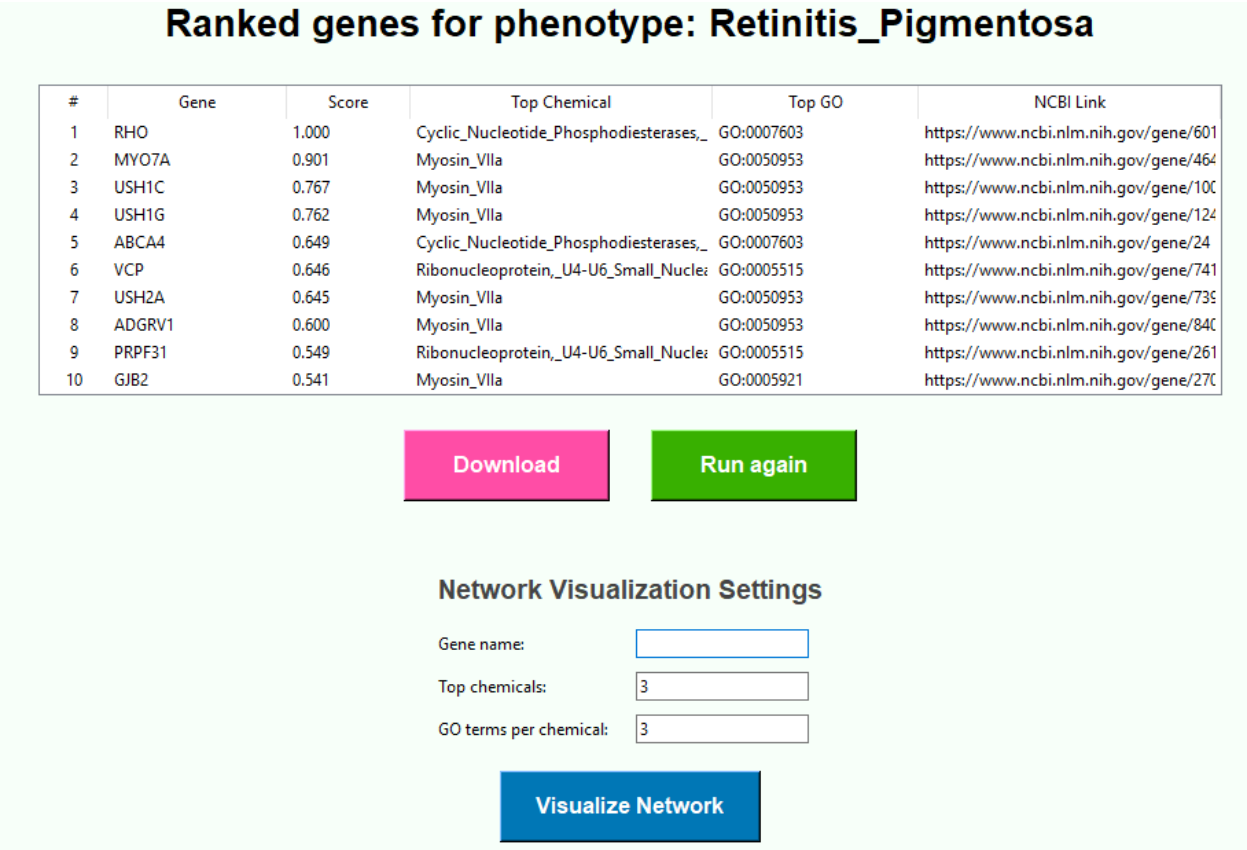

Figure 14. Gene prioritization results of Retinitis Pigmentosa

To interpret gene prioritization results, chemical associations and Gene Ontology (GO) annotations were analyzed. For each gene, three dominant chemical terms and three GO terms were selected. Known disease-associated genes showed strong association with retinal biological processes. The gene ranked first, RHO, demonstrates a biologically meaningful path. (Figure 15) The phenotype is associated with retinal chemicals such as rhodopsin, cyclic nucleotide phosphodiesterases, and peripherin. These chemicals are directly involved in photoreceptor function. Associated GO annotations include phototransduction, photoreceptor outer segment, and photoreceptor disc membrane. These GO terms describe the cellular structures and processes where these molecules act. Since RHO encodes rhodopsin, it plays a key role in phototransduction. Rhodopsin is located in the photoreceptor outer segment. Both the chemical terms and GO annotations point to the same retinal mechanism.

For the candidate gene GJB2, the network is illustrated in Figure 16. GJB2 is not directly associated with Retinitis Pigmentosa. However its connections suggest a plausible biological role.

The associated chemicals include peripherin and cadherin-related proteins, which help maintain retinal cell structure. They connect to GO terms such as gap junction channel activity, gap junction assembly, and connexin complex. These processes defined by GO annotations are responsible for cell-cell communication and ion exchange in tissues. The relationship between these elements explains the connection to GJB2. The structural chemicals are associated with processes that require coordinated communication between retinal cells. These processes depend on gap junctions, which enable direct signaling between cells. Since GJB2 encodes a connexin protein that forms gap junctions, it is functionally involved in these processes.

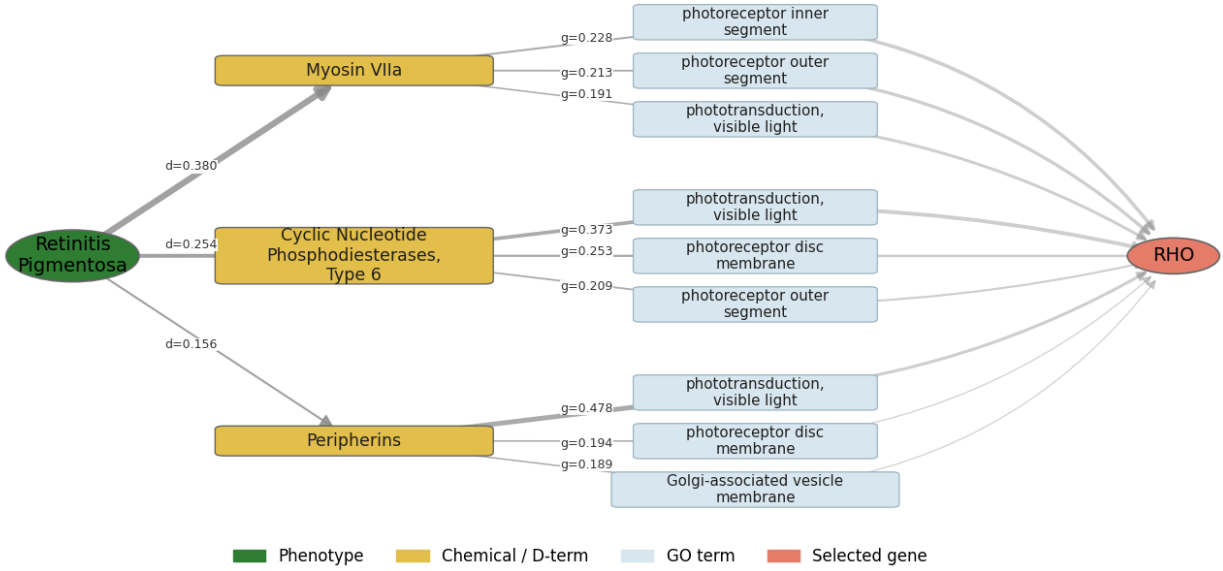

Figure 15. RHO gene network for Retinitis Pigmentosa

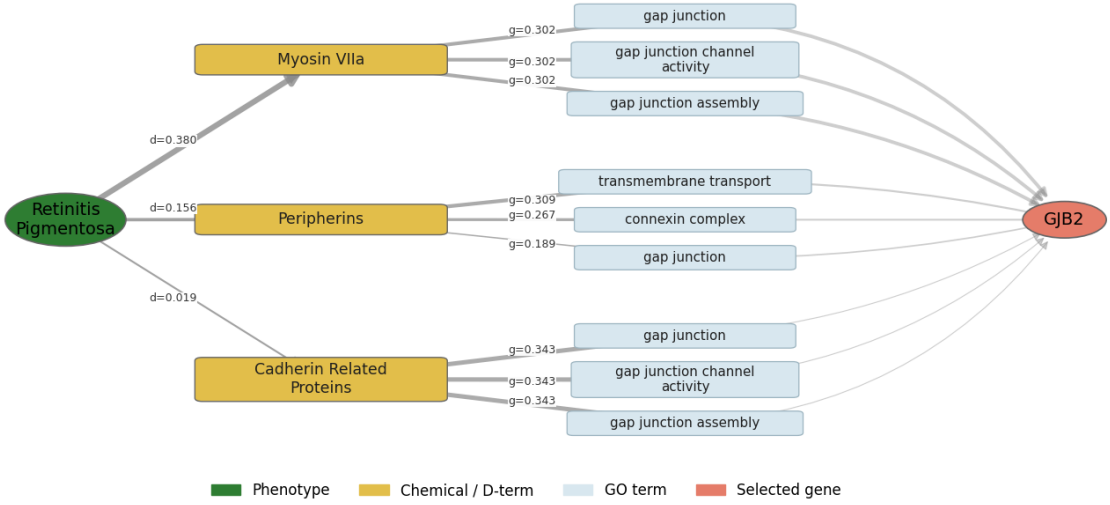

Figure 16. GJB2 gene network for Retinitis Pigmentosa

Another candidate gene VCP is mainly connected through chemicals and GO terms related to general processes. These processes include protein folding and endoplasmic reticulum stress.

These pathways are not specific to retinal biology, resulting in weaker and non-specific phenotype-gene connections.

The results showed that chemicals and GO associations reflect biologically meaningful retinal pathways. This demonstrates that developed gene prioritization tools can uncover biologically plausible candidate genes. supported by real functional mechanisms.

The primary evaluation dataset consisted of neurological phenotypes. However, the developed tool was also explored outside the primary evaluation framework. It was also applied to the analysis of complex multifactorial phenotypes. Sarcopenia was analyzed as an age-related multifactorial phenotype. It is associated with muscle loss, reduced physical function, inflammation, and immune system aging [45].

The developed literature-driven prioritization approach identified biological pathways related to muscle biology, cytoskeletal organization, cellular stress, inflammation, and aging processes. The developed approach was additionally applied in collaborative work submitted to the ASHG2026 conference, “Telomere Shortening, Epigenetic Remodeling of CCT5, and Immune System Aging Converge to Drive Muscle Wasting in Older Adults” [46]. In this study, the prioritization algorithm was used to investigate indirect relationships between the CCT5 gene and Sarcopenia phenotype through Gene Ontology and MeSH-based associations.

DISCUSSION

4.1. Overall performance of the gene prioritization tool

According to the results, the developed gene prioritization tool demonstrated its ability to identify biologically meaningful relationships. Based on Recall@5, and MRR values, the method showed moderately good performance. This indicates that the proposed tool is able to prioritize biologically relevant genes through indirect relationships. These relationships are based on intermediate entities, including MeSH chemical terms and Gene Ontology annotations. An interpretable workflow was created by combining several steps: PubMed literature retrieval, MeSH term extraction, Gene2PubMed mapping, Gene2GO annotation, score calculation, and network visualization. Together, these components form a biologically meaningful prioritization system. The implemented network visualization module allows investigation of why a gene was prioritized. The user can investigate how a gene is connected to the phenotype through chemicals and GO terms. This improves interpretability and helps to understand the ranking results. However, several limitations must be considered. The program depends on the availability of annotations in Gene2PubMed and Gene2GO. If a gene is not well annotated or not present in these databases, the pipeline may stop or return incomplete results. The method is not fully robust for poorly annotated genes.

4.2. Effect of literature and database coverage

Data analysis (Section 3.2) showed that phenotypes with more associated PubMed publications tend to have more retrieved genes. In addition, phenotypes with more genes have more associated GO annotations. These results demonstrate that the performance of the pipeline is influenced by literature density and database coverage. The method performs best for phenotypes that are well represented in the literature. This dependence introduces an important bias. The pipeline starts from PubMed MeSH-indexed publications. Therefore, phenotypes with limited literature may produce weaker results. Delayed indexing and sparse annotation also affect the output. Phenotypes with fewer publications will have fewer chemical links, fewer mapped genes, and fewer GO annotations. Because of this, the scores depend not only on biology but also on how complete the databases are. A low-ranked gene is not always unimportant. It may just have little information available in the databases or literature. This is a known limitation in text-mining methods

4.3. Interpretation of the scoring method

The gene prioritization ranking results depends on the scoring system used in this thesis. The phenotype-chemical score (dscore) reflects how frequently a chemical co-occurs with the phenotype in the literature. The chemical-GO score (gscore) estimates the strength of functional association through shared genes. Their product is summed across all paths to obtain the final gene score. Based on results, known disease genes are often ranked at the top, indicating that the method works reasonably well. However, several limitations must be considered. First, the dscore formulation (inters / nd) may favor chemicals that are highly specific to a phenotype, while more general but biologically important chemicals may be underrepresented. Second, the scoring is based on co-occurrence, which does not imply causality. Entities may co-occur in the literature without direct biological involvement. As a result, some high-ranking genes may reflect strong literature association rather than true biological importance. Another important limitation is that

missing annotations directly affect the score. If a gene lacks Gene2PubMed or Gene2GO links, it will have fewer or no contributing paths, which leads to lower ranking or exclusion from results. Future improvements could include testing alternative scoring strategies.

4.4. Biological interpretation, practical value, and future improvements

The results in Section 3.4 show that the method can identify biologically meaningful genes and also suggest new candidate genes. For Retinitis Pigmentosa, most of the top-ranked genes (8 out of 10) are already known disease-associated genes, which supports the validity of the method. At the same time, genes such as GJB2 and VCP were not directly linked to the disease but showed biologically plausible connections through chemicals and GO terms.

The network analysis also supports interpretation of results. For example, the RHO gene is connected through retinal-specific chemicals and GO terms related to phototransduction. This reflects Retinitis Pigmentosa disease mechanisms. In contrast, candidate genes showed weaker or more general pathways, but still biologically reasonable links. This shows that the method can explain why a gene is prioritized.

Lower recall values observed in some phenotypes should not always be considered as poor performance. Some genes that are not found in OMIM or DisGeNET may represent potential novel candidates rather than false positives. This is important because these databases are not complete and are biased toward well-studied genes. Overall, the method has practical value both for identifying known disease genes and for generating new hypotheses. Future improvements could include better filtering of non-specific pathways, integration of additional databases, and weighting of evidence to improve biological relevance and precision.

CONCLUSIONS

1. The developed system successfully integrates multiple biological databases and enables identification of indirect associations between phenotypes and genes.
2. The proposed scoring method effectively prioritizes genes, with most known disease genes ranked among the top positions, demonstrating strong performance.
3. The tool allows tracking of intermediate relationships (phenotype-chemical-GO-gene), providing interpretability and explaining how each gene was identified.
4. The system is capable of identifying biologically relevant candidate genes not present in existing databases, indicating its potential for discovering novel gene-disease associations.
5. The source code of the developed system is publicly available on GitHub (*link:https://github.com/Ariadna21/gene_prioritization_tool_*).

ACKNOWLEDGEMENTS

The author acknowledges the use of ChatGPT (OpenAI) for language editing and structural suggestions during thesis preparation. All content was reviewed and verified by the author.

REFERENCES

1. Zhou, J. & Fu, B.Q., 2018. The research on gene-disease association based on text-mining of PubMed. BMC Bioinformatics, 19, 37. https://doi.org/10.1186/s12859-018-2048-y
2. Cheung, W.A., Ouellette, B.F.F. & Wasserman, W.W., 2012. Inferring novel gene-disease associations using Medical Subject Heading Over-representation Profiles. Genome Medicine, 4, 75. https://doi.org/10.1186/gm376
3. Tsuruoka, Y., Miwa, M., Hamamoto, K., Tsujii, J. & Ananiadou, S., 2011. FACTA+: A text search engine for finding associated biomedical concepts. Bioinformatics, 27(13), i111-i119. https://academic.oup.com/bioinformatics/article/27/13/i111/178661
4. Alaa, A.M., Zanin, M., Hutter, F. & van der Schaar, M., 2019. Indirect association and ranking hypotheses for literature-based discovery. BMC Bioinformatics, 20, 298. https://doi.org/10.1186/s12859-019-2989-9
5. The Gene Ontology Consortium, 2021. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Research, 49(D1), pp.D325-D334. https://doi.org/10.1093/nar/gkaa1113
6. Zhang, Y., Chen, M., Li, X., Wang, J., 2023. Multi-domain knowledge graph embeddings for gene-disease association prediction. Journal of Biomedical Semantics, 14, 11. https://doi.org/10.1186/s13326-023-00291-x
7. Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.C., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P., Moreau, Y., 2014. Identifying disease genes by integrating multiple data sources. BMC Medical Genomics, 7(Suppl 2), S2. https://doi.org/10.1186/1755-8794-7-S2-S2
8. Pranckevičienė, E., 2015. Procedure and datasets to compute links between genes and phenotypes defined by MeSH keywords. F1000Research, 4, 47. https://doi.org/10.12688/f1000research.6004.1
9. Pranckevičienė, E., 2014. Analysis of indirect associations between genes and phenotypes using literature mining and biomedical ontologies. University of Ottawa. https://ruor.uottawa.ca/server/api/core/bitstreams/a9844619-6121-4635-8c7a-f6fb60bc8c22/content
10. Hassani, S., Huang, X., Silva, E., 2023. Insights from PubMed for biomedical literature analysis. ACL Anthology. https://aclanthology.org/2023.insights-1.9.pdf
11. National Library of Medicine, 2026. PubMed overview. https://pubmed.ncbi.nlm.nih.gov/
12. National Library of Medicine, 2025. Medical Subject Headings (MeSH) 2025 highlights. Available at: https://www.nlm.nih.gov/oet/ed/mesh/2025/mesh_highlights.html
13. National Library of Medicine, 2026. Introduction to MeSH. https://www.nlm.nih.gov/oet/ed/pubmed/mesh/mod01/04-100.html
14. Gangemi, A., Hristovski, D., Kastrin, A., Milic Frayling, N. & Peterlin, B., 2014. Network based analysis reveals distinct association patterns in a semantic MEDLINE based drug-

disease-gene network. Studies in Health Technology and Informatics, 205, pp.579-583. https://pubmed.ncbi.nlm.nih.gov/25170419/

15. Kastrin, A., Rindflesch, T.C. & Hristovski, D., 2014. Link prediction in a MeSH co-occurrence network: preliminary results. Studies in Health Technology and Informatics, 205. https://pubmed.ncbi.nlm.nih.gov/25160252/
16. Yu, Z., Bernstam, E.V., Cohen, T., Wallace, B.C. & Johnson, T.R., 2016. Improving the utility of MeSH terms using the TopicalMeSH representation. Journal of Biomedical Informatics, 61, pp.77-86. https://doi.org/10.1016/j.jbi.2016.03.005
17. Sakamoto, T., Yamaguchi, A., Kato, T. & Kinoshita, R., 2022. Automated recommendation of research keywords from PubMed using MeSH co-occurrence data. Metabolites, 12(2), 133. https://doi.org/10.3390/metabo12020133
18. Chen, Q., Shi, Z. & Xu, L., 2019. A MeSH-based database of gene-disease associations. Database, 2019, baz019. https://doi.org/10.1093/database/baz019
19. Singhal, A., Simmons, M. & Lu, Z., 2016. Text mining genotype-phenotype relationships from biomedical literature for database curation and precision medicine. PLoS Computational Biology, 12(11), e1005017. https://doi.org/10.1371/journal.pcbi.1005017
20. Nguyen, T. & Shatkay, H., 2018. Challenges in extracting gene-disease associations from biomedical literature. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5931382/
21. Wei, C., Kao, H.Y. & Lu, Z., 2014. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Research, 42(W1), W518-W522. https://doi.org/10.1093/nar/gku498
22. The Gene Ontology Consortium, 2023. The Gene Ontology knowledgebase in 2023. Nucleic Acids Research. https://doi.org/10.1093/nar/gkac1018
23. Gene Ontology Consortium, 2026. Gene Ontology documentation. https://geneontology.org/docs/ontology-documentation/
24. Feuermann, M., Mi, H., Gaudet, P., Muruganujan, A., Lewis, S.E., Ebert, D., Mushayahama, T., Gene Ontology Consortium & Thomas, P.D., 2025. A compendium of human gene functions derived from evolutionary modelling. Nature, 640, pp.146-154. https://doi.org/10.1038/s41586-025-08592-0
25. Linghu, B., Snitkin, E.S., Hu, Z., Xia, Y., DeLisi, C., 2009. Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. Genome Biology, 10(9), R91. https://doi.org/10.1186/gb-2009-10-9-r91
26. Toscano, W.A., Oehlke, K.P., 2005. Systems biology: new approaches to old environmental health problems. International Journal of Environmental Research and Public Health, 2(1), pp.4-9.
27. Li, B., Wang, Z., Chen, Q., Li, K., Wang, X., Wang, Y., Zeng, Q., Han, Y., Lu, B., Zhao, Y., Zhang, R., Jiang, L., Pan, H., Luo, T., Zhang, Y., Fang, Z., Xiao, X., 2021. GPCards: an integrated database of genotype-phenotype correlations in human genetic diseases. Database, 2021, baab043. https://doi.org/10.1093/database/baab043

28. Grissa, D., Junge, A., Oprea, T.I., Jensen, L.J., 2022. Diseases 2.0: a weekly updated database of disease-gene associations from text mining and data integration. Database, 2022, baac019. https://doi.org/10.1093/database/baac019
29. Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., Furlong, L.I., 2017. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Research, 45(D1), pp.D833-D839. https://doi.org/10.1093/nar/gkw943
30. Amberger, J.S., Bocchini, C.A., Scott, A.F., Hamosh, A., 2019. OMIM.org: leveraging knowledge across phenotype-gene relationships. Nucleic Acids Research, 47(D1), pp.D1038-D1043. https://doi.org/10.1093/nar/gky1151
31. Davis, A.P., Murphy, C.G., Saraceni-Richards, C.A., Rosenstein, M.C., Wiegers, T.C., Mattingly, C.J., 2009. Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. BMC Bioinformatics, 10, 326. https://doi.org/10.1186/1471-2105-10-326
32. Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L.C., Lewis-Smith, D., Vasilevsky, N.A., Danis, D., Balagura, G., Baynam, G., Brower, A.M., et al., 2021. The Human Phenotype Ontology in 2021. Nucleic Acids Research, 49(D1), pp.D1207-D1217. https://doi.org/10.1093/nar/gkaa1043
33. Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.C., De Moor, B., Marynen, P., Hassan, B. & Moreau, Y., 2015. GeneTIER: prioritization of candidate disease genes using tissue-specific gene expression profiles. Bioinformatics, 31(17), pp.2810-2817. https://doi.org/10.1093/bioinformatics/btv225
34. Zolotareva, O. & Kleine, M., 2019. A survey of gene prioritization tools for Mendelian and complex human diseases. Journal of Integrative Bioinformatics, 16(4). https://doi.org/10.1515/jib-2018-0069
35. Chen, Y., Wang, W., Zhou, Y., Shields, R., Chanda, S.K., Elston, R.C. & Li, J., 2011. In silico gene prioritization by integrating multiple data sources. PLOS ONE, 6(6), e21137. https://doi.org/10.1371/journal.pone.0021137
36. Köhler, S., Bauer, S., Horn, D. & Robinson, P.N., 2008. Walking the interactome for prioritization of candidate disease genes. American Journal of Human Genetics, 82(4), pp.949-958. https://doi.org/10.1016/j.ajhg.2008.02.013
37. Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.C., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P., Moreau, Y., 2006. Gene prioritization through genomic data fusion. Nature Biotechnology, 24(5), pp.537-544. https://doi.org/10.1038/nbt1203
38. Chen, J., Bardes, E.E., Aronow, B.J., Jegga, A.G., 2009. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Research, 37(Web Server issue), pp.W305-W311. https://doi.org/10.1093/nar/gkp427

39. Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J.X., Jensen, L.J., 2015. DISEASES: text mining and data integration of disease-gene associations. Nucleic Acids Research, 43(D1), pp.D1071-D1078. https://doi.org/10.1093/nar/gku1010
40. Yang, H., Robinson, P.N., Wang, K., 2015. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nature Methods, 12(9), pp.841-843. https://doi.org/10.1038/nmeth.3484
41. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., Morris, Q., 2008. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biology, 9(Suppl 1), S4. https://doi.org/10.1186/gb-2008-9-s1-s4
42. Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., Dicuccio, M., Federhen, S., Feolo, M., et al., 2009. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research, 37(Database issue), pp.D5-D15. https://doi.org/10.1093/nar/gkn741
43. Lipscomb, C.E., 2000. Medical Subject Headings (MeSH). Bulletin of the Medical Library Association, 88(3), pp.265-266.
44. Brown, G.R., Hem, V., Katz, K.S., Ovetsky, M., Wallin, C., Ermolaeva, O., Tolstoy, I., Tatusova, T., Pruitt, K.D., Maglott, D.R., Murphy, T.D., 2015. Gene: a gene-centered information resource at NCBI. Nucleic Acids Research, 43(Database issue), pp.D36-D42. https://doi.org/10.1093/nar/gku1055
45. Kilaitė, J., Pranckevičienė, E., Ginevičienė, V., Urnikytė, A., Dadelienė, R., Mastavičiūtė, A., Jamontaitė, I.E., Alekna, V. & Ahmetov, I.I., 2025. Causal relationships between sarcopenia, frailty, and health outcomes: A systematic review of Mendelian randomization studies. Experimental Gerontology, 212, p.112953. https://doi.org/10.1016/j.exger.2025.112953
46. Pranckevičienė, E., Šamatovič, A., Kilaitė, J., Alekna, V. & Ginevičienė, V., 2026. Telomere Shortening, Epigenetic Remodeling of CCT5, and Immune System Aging Converge to Drive Muscle Wasting in Older Adults. ASHG 2026 Annual Meeting Abstract, American Society of Human Genetics
47. OpenAI, 2025. ChatGPT (May 2025 version) [Large language model]. https://chat.openai.com/ [Accessed 14 May 2026].

SUMMARY

This master's thesis, „A MeSH-Term-Based Phenotype-Gene Association Dataset for Gene Prioritization" by Ariadna Šamatovič, was completed within the Systems Biology study program at Vilnius University.

Understanding how genes are linked to diseases is essential for improving diagnosis and developing new treatments. However, identifying these relationships through experimental methods is time-consuming and expensive. Rapidly growing volume of biomedical literature makes manual extraction extremely difficult. This master's thesis aims to develop a gene prioritization tool and structured database for identifying and storing disease-gene associations. A gene prioritization tool and structured dataset were created using information from PubMed, Gene2PubMed, and Gene Ontology (Gene2GO). The method focuses on identifying indirect relationships. This is done by connecting diseases to chemicals through MeSH terms, and then linking these chemicals to genes and their biological functions. These connections are quantified using a scoring system, and ranked based on their relevance to a selected disease. The developed system integrates multiple data sources into a unified pipeline. It also includes visualization features that illustrate the intermediate biological steps linking diseases to genes. Overall, this work provides an efficient and interpretable method for exploring disease-gene relationships. It has the potential to identify novel candidate genes and support future biomedical research.

SUMMARY IN LITHUANIAN

Šis magistro darbas „ *MeSH terminais paremtas fenotipo-geno sąsajų duomenų rinkinys skirtas genų prioritetizavimui*“, kurio autorė yra Ariadna Šamatovič, buvo parengtas Vilniaus universiteto Sistemų biologijos studijų programoje.

Genų ir fenotipų ryšių supratimas yra svarbus siekiant pagerinti ligų diagnostiką ir kurti naujus gydymo metodus. Tačiau šių ryšių nustatymas eksperimentiniais metodais yra brangus ir reikalauja daug laiko. Sparčiai augantis biomedicininės literatūros kiekis daro rankinį informacijos išgavimą itin sudėtingą. Šio magistro darbo tikslas - sukurti genų prioritetizavimo įrankį ir struktūrizuotą duomenų bazę, skirtą ligų ir genų sąsajoms nustatyti ir saugoti. Genų prioritetizavimo įrankis ir struktūrizuotas duomenų rinkinys sukurti naudojant PubMed, Gene2PubMed ir Gene Ontology (Gene2GO) duomenis. Metodas orientuotas į netiesioginių ryšių nustatymą. Tai atliekama susiejant fenotipus su cheminėmis medžiagomis per MeSH terminus, o vėliau šias medžiagas susiejant su genais ir jų biologinėmis funkcijomis. Ryšiai vertinami taikant vertinimo metodą ir rikiuojami pagal jų reikšmingumą pasirinktam fenotipui. Sistema sujungia kelis duomenų šaltinius į bendrą apdorojimo sistemą ir leidžia vizualizuoti tarpinius biologinius žingsnius, jungiančius fenotipus su genais. Apibendrinant, šiame darbe sukurtas efektyvus ir interpretuojamas įrankis, skirtas fenotipų ir genų ryšiams tirti. Jis leidžia nustatyti naujus kandidatinius genus ir gali būti naudingas būsimuose biomedicininiuose tyrimuose.

APPENDICES

Table 1. List of phenotypes included in the dataset.

No.	Phenotype
1	Williams Syndrome
2	WAGR Syndrome
3	Vein of Galen Malformations
4	Trisomy 13 Syndrome
5	Thalamic Diseases
6	Tangier Disease
7	Susac Syndrome
8	Spina Bifida Occulta
9	Sneddon Syndrome
10	Small Fiber Neuropathy
11	Sinus Pericranii
12	Schizencephaly
13	Rett Syndrome
14	Refsum Disease
15	Radial Neuropathy
16	Pyruvate Dehydrogenase Complex Deficiency Disease
17	Postpoliomyelitis Syndrome
18	Posterior Leukoencephalopathy Syndrome
19	Polymicrogyria
20	Peroneal Neuropathies
21	Pentalogy of Cantrell
22	Paraneoplastic Polyneuropathy
23	Paraneoplastic Cerebellar Degeneration
24	POEMS Syndrome
25	Olivopontocerebellar Atrophies
26	Neuromyelitis Optica
27	Neurofibromatosis 2
28	Neurofibromatosis 1
29	Neurocytoma
30	Neuroaxonal Dystrophies
31	Myotonia Congenita
32	Mucopolysaccharidosis II
33	Miller Fisher Syndrome
34	Menkes Kinky Hair Syndrome
35	Meningeal Carcinomatosis
36	Lissencephaly
37	Limbic Encephalitis

38	Lewy Body Disease
39	Lafora Disease
40	Kuru
41	Isaacs Syndrome
42	Intracranial Hypotension
43	Hyperekplexia
44	Hydranencephaly
45	Huntington Disease
46	Holoprosencephaly
47	Glycogen Storage Disease Type IIb
48	Giant Axonal Neuropathy
49	Frontotemporal Dementia
50	Friedreich Ataxia
51	Fragile X Syndrome
52	Focal Cortical Dysplasia
53	Empty Sella Syndrome
54	Dystonia Musculorum Deformans
55	Down Syndrome
56	Diffuse Cerebral Sclerosis of Schilder
57	De Lange Syndrome
58	Corticobasal Degeneration
59	Congenital Cranial Dysinnervation Disorders
60	Cockayne Syndrome
61	Canavan Disease
62	CHARGE Syndrome
63	CADASIL
64	Brachial Plexus Neuritis
65	Angelman Syndrome
66	Amyotrophic Lateral Sclerosis
67	Alstrom Syndrome
68	Alexander Disease
69	Aicardi Syndrome
70	Adrenoleukodystrophy
71	Acute Febrile Encephalopathy
72	Acrodynia
73	Acrocallosal Syndrome
74	AIDS Dementia Complex
75	Retinitis Pigmentosa

Table 2. Statistics of the collected dataset

Phenotype	Publications (nm)	Unique chemicals	Unique genes	Unique GO terms
Williams Syndrome	542	465	606	2399
WAGR Syndrome	61	87	11	87
Vein of Galen Malformations	33	45	4	67
Trisomy 13 Syndrome	199	78	6	65
Thalamic Diseases	191	264	3	53
Tangier Disease	477	426	23	572
Susac Syndrome	96	70	1	12
Spina Bifida Occulta	313	234	2	34
Sneddon Syndrome	95	95	3	24
Small Fiber Neuropathy	186	187	13	228
Sinus Pericranii	10	11	1	14
Schizencephaly	17	12	2	18
Rett Syndrome	2117	1026	1147	3776
Refsum Disease	460	363	17	178
Radial Neuropathy	54	69	1	6
Pyruvate Dehydrogenase Complex Deficiency Disease	382	273	12	118
Postpoliomyelitis Syndrome	172	169	3	24
Posterior Leukoencephalopathy Syndrome	790	472	2	22
Polymicrogyria	89	97	28	246
Peroneal Neuropathies	105	172	1	13
Pentalogy of Cantrell	8	15	1	2
Paraneoplastic Polyneuropathy	186	152	4	35
Paraneoplastic Cerebellar Degeneration	264	170	11	91
POEMS Syndrome	592	306	4	132
Olivopontocerebellar Atrophies	288	367	24	270
Neuromyelitis Optica	3744	909	132	1570
Neurofibromatosis 2	569	459	398	2361
Neurofibromatosis 1	2827	1442	168	1753
Neurocytoma	227	168	11	122
Neuroaxonal Dystrophies	667	474	28	373
Myotonia Congenita	632	450	39	383
Mucopolysaccharidosis II	731	341	9	93
Miller Fisher Syndrome	512	257	4	55
Menkes Kinky Hair Syndrome	788	440	8	137
Meningeal Carcinomatosis	457	242	13	182
Lissencephaly	460	353	81	670
Limbic Encephalitis	900	287	10	127
Lewy Body Disease	2946	1089	199	2087

Lafora Disease	310	238	33	423
Kuru	256	121	11	102
Isaacs Syndrome	221	168	7	173
Intracranial Hypotension	293	214	3	74
Hyperekplexia	53	34	6	70
Hydranencephaly	84	80	7	65
Huntington Disease	9012	2888	3666	7818
Holoprosencephaly	470	349	74	689
Glycogen Storage Disease Type IIb	167	106	10	147
Giant Axonal Neuropathy	53	51	184	1149
Frontotemporal Dementia	3471	1184	2349	6204
Friedreich Ataxia	1508	999	60	984
Fragile X Syndrome	2818	1310	410	2267
Focal Cortical Dysplasia	60	71	1	85
Empty Sella Syndrome	412	214	2	17
Dystonia Musculorum Deformans	529	416	33	359
Down Syndrome	9661	3059	557	3483
Diffuse Cerebral Sclerosis of Schilder	1073	602	12	116
De Lange Syndrome	399	263	27	363
Corticobasal Degeneration	82	69	5	214
Congenital Cranial Dysinnervation Disorders	339	260	69	416
Cockayne Syndrome	656	569	172	1404
Canavan Disease	242	204	3	32
CHARGE Syndrome	222	155	69	511
CADASIL	705	329	29	597
Brachial Plexus Neuritis	352	336	8	92
Angelman Syndrome	840	525	150	1008
Amyotrophic Lateral Sclerosis	15370	3559	8604	10812
Alstrom Syndrome	135	141	18	227
Alexander Disease	273	189	11	269
Aicardi Syndrome	29	45	7	60
Adrenoleukodystrophy	1318	809	54	721
Acute Febrile Encephalopathy	62	73	2	17
Acrodynia	186	136	1	7
Acrocallosal Syndrome	25	29	6	72
AIDS Dementia Complex	2179	1222	317	2953
Retinitis Pigmentosa	6583	2117	1992	5127

Table 3. Evaluation results of gene prioritization performance

Phenotype	Gold standard genes	True genes in top-5	Recall@5	First true gene rank	RR
Williams Syndrome	8	4	0.5	2	0.5
WAGR Syndrome	6	3	0.5	1	1
Vein of Galen Malformations	2	2	1	1	1
Trisomy 13 Syndrome	5	1	0.2	2	0.5
Thalamic Diseases	5	0	0	0	0
Tangier Disease	1	1	1	2	0.5
Susac Syndrome	2	0	0	0	0
Spina Bifida Occulta	4	1	0.25	2	0.5
Sneddon Syndrome	2	1	0.5	2	0.5
Small Fiber Neuropathy	4	3	0.75	1	1
Sinus Pericranii	2	0	0	0	0
Schizencephaly	3	1	0.33	1	1
Rett Syndrome	3	0	0	0	0
Refsum Disease	2	2	1	2	0.5
Radial Neuropathy	2	0	0	0	0
Pyruvate Dehydrogenase Complex Deficiency Disease	4	2	0.5	2	0.5
Postpoliomyelitis Syndrome	2	0	0	0	0
Posterior Leukoencephalopathy Syndrome	3	0	0	0	0
Polymicrogyria	4	0	0	0	0
Peroneal Neuropathies	2	0	0	0	0
Pentalogy of Cantrell	2	0	0	0	0
Paraneoplastic Polyneuropathy	2	0	0	0	0
Paraneoplastic Cerebellar Degeneration	2	0	0	0	0
POEMS Syndrome	2	1	0.5	1	1
Olivopontocerebellar Atrophies	4	0	0	0	0
Neuromyelitis Optica	2	1	0.5	4	0.25
Neurofibromatosis 2	1	0	0	0	0
Neurofibromatosis 1	1	0	0	0	0
Neurocytoma	2	0	0	0	0
Neuroaxonal Dystrophies	2	1	0.5	2	0.5
Myotonia Congenita	1	1	1	2	0.5
Mucopolysaccharidosis II	1	1	1	1	1
Miller Fisher Syndrome	2	0	0	0	0
Menkes Kinky Hair Syndrome	1	1	1	2	0.5

Meningeal Carcinomatosis	2	1	0.5	1	1
Lissencephaly	3	1	0.33	2	0.5
Limbic Encephalitis	2	2	1	1	1
Lewy Body Disease	2	1	0.5	1	1
Lafora Disease	2	1	0.5	2	0.5
Kuru	1	1	1	1	1
Isaacs Syndrome	2	2	1	1	1
Intracranial Hypotension	2	0	0	0	0
Hyperekplexia	3	3	1	1	1
Hydranencephaly	1	1	1	2	0.5
Huntington Disease	1	1	1	1	1
Holoprosencephaly	4	2	0.5	1	1
Glycogen Storage Disease Type IIb	1	0	0	0	0
Giant Axonal Neuropathy	1	1	1	4	0.25
Frontotemporal Dementia	4	1	0.25	1	1
Friedreich Ataxia	1	0	0	0	0
Fragile X Syndrome	1	1	1	1	1
Focal Cortical Dysplasia	2	1	0.5	1	1
Empty Sella Syndrome	2	0	0	0	0
Dystonia Musculorum Deformans	1	1	1	1	1
Down Syndrome	3	2	0.7	1	1
Diffuse Cerebral Sclerosis of Schilder	2	1	0.5	5	0.2
De Lange Syndrome	4	3	0.75	2	0.5
Corticobasal Degeneration	2	1	0.5	1	1
Congenital Cranial Dysinnervation Disorders	3	0	0	0	0
Cockayne Syndrome	2	1	0.5	1	1
Canavan Disease	1	1	1	1	1
CHARGE Syndrome	1	1	1	1	1
CADASIL	1	0	0	0	0
Brachial Plexus Neuritis	1	0	0	0	0
Angelman Syndrome	1	1	1	1	1
Amyotrophic Lateral Sclerosis	5	1	0.2	1	1
Alstrom Syndrome	1	0	0	6	0.17
Alexander Disease	1	0	0	6	0.17
Aicardi Syndrome	1	1	1	2	0.5
Adrenoleukodystrophy	1	1	1	2	0.5
Acute Febrile Encephalopathy	2	0	0	0	0
Acrodynia	2	0	0	0	0
Acrocallosal Syndrome	1	1	1	4	0.25
AIDS Dementia Complex	2	0	0	0	0

Retinitis Pigmentosa	8	5	0.625	1	1