Chained Semantic Retrieval for Rare Disease and Gene Identification using Clinical Phenotype

by

Md. Saiful
24266049

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
M.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
October 2025

Declaration

It is hereby declared that

1. The thesis submitted is my own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Md. Saiful
24266049

Approval

Chained Semantic Retrieval knowledge graph for Rare Disease and Gene Identification using Patient Phenotype" submitted by

1. Md. Saiful (24266049)

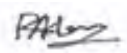
Of Fall, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Sc. in Computer Science on October 19, 2025.

Examining Committee:

Supervisor:

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
BRAC University

Examiner:
(External)



Prof. Dr. Kazi Md. Rokibul Alam, PhD
Professor
Department of Computer Science and Engineering
Khulna University of Engineering Technology

Examiner:
(Internal)

Dr. Md. Ashraful Alam
Associate Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:

Dr. Md Sadek Ferdous
Professor
Department of Computer Science and Engineering
BRAC University

Head of Department (Chair):

Dr. Sadia Hamid Kazi
Associate Professor
Department of Computer Science and Engineering
BRAC University

Ethics Statement

The research presented in this thesis involves the use of patient phenotype data from publications papers. All data used in this study has been fully anonymized to protect the privacy and confidentiality of the individuals involved. The data was collected and handled in accordance with the ethical guidelines of the institution and with the principles of the Declaration of Helsinki. No personally identifiable information was used or stored during the course of this research.

Abstract

Accurate identification of rare diseases and associated genes from patient phenotypic data represents a critical challenge in precision medicine and genomic research. Current diagnostic approaches face substantial limitations in scalability, interpretability, and real-time processing when integrating heterogeneous phenotypic and genotypic databases. This study presents a novel artificial intelligence-driven system employing chained semantic retrieval and knowledge graph integration to identify rare diseases, genes, and Human Phenotype Ontology (HPO) terms from patient phenotype descriptions. The methodology leverages sentence transformers (based on Bidirectional and Auto-Regressive Transformers) for generating vector embeddings, Facebook AI Similarity Search (FAISS) for efficient similarity computation, and a fine-tuned Llama 3.2 model integrated with a Retrieval-Augmented Generation (RAG) pipeline. The system implements a tiered scoring mechanism that chains retrieval across Human Phenotype Ontology, disease, and gene databases to progressively refine predictions through contextual enhancement. Evaluation on 50 patient phenotypes with confirmed Duchenne Muscular Dystrophy diagnosis, consisting of 30 true positive and 20 true negative cases, demonstrated strong performance with 28 true positives, 17 true negatives, 2 false negatives, and 3 false positives in top-ten retrieval results. The system achieved 90 % overall accuracy, 93.3 % recall, 85 % specificity, 90.3 % precision, and an F1-score of 91.8 %, with an average computational efficiency of 3.2 seconds per response. The proposed framework effectively addresses critical gaps in cross-database integration while maintaining interpretability through tiered confidence scoring for clinical decision support applications.

Keywords: Rare Disease, Phenotype Embeddings, Knowledge Graph, Gene Identification, Human Phenotype Ontology (HPO), Semantic Retrieval, Genotype-Phenotype Mapping, Clinical Data Integration, LLM, RAG

Acknowledgement

I extend my sincere gratitude to my supervisor, Dr. Md. Golam Rabiul Alam, for his guidance and support throughout this research. I also thank the faculty and peers for their insights. This work was inspired by the need to improve rare disease diagnostics in resource-limited settings.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The landscape of modern medicine has been fundamentally transformed by advances in genomic sequencing technologies and computational biology. Despite these remarkable technological achievements, the accurate and timely diagnosis of rare diseases remains one of the most formidable challenges facing healthcare systems worldwide. Rare diseases, defined as conditions affecting fewer than one in two thousand individuals [12], collectively impact approximately four hundred million people globally across more than seven thousand distinct disease entities [3], [12].

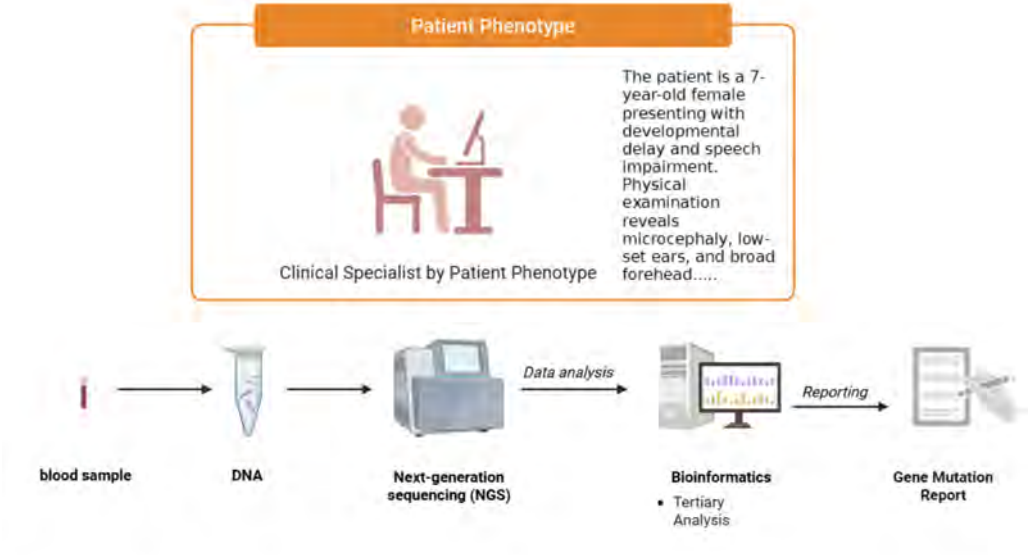


Figure 1.1: An overview of the Next-Generation Sequencing (NGS) workflow, a key technology in modern genomic analysis.

The complexity inherent in rare disease diagnosis stems from multiple interconnected factors, including extreme phenotypic heterogeneity, genetic variability across patient populations, incomplete penetrance of genetic variants, and the fragmented distribution of specialized medical knowledge across disparate databases and information systems [6].

Traditional diagnostic pathways for rare diseases have historically relied upon the accumulated expertise of specialized clinicians who possess deep knowledge of specific

disease categories and their characteristic manifestations. This expert-driven approach, while invaluable, faces inherent limitations in terms of scalability, geographic accessibility, and the cognitive burden placed upon individual practitioners who must synthesize information from increasingly complex and voluminous data sources.

Patients suspected of having rare diseases often experience prolonged diagnostic odysseys lasting years or even decades, during which time they undergo numerous consultations, invasive procedures, and inconclusive genetic tests. These diagnostic delays have profound consequences not only for patient quality of life but also for the effectiveness of therapeutic interventions, many of which are most efficacious when initiated early in the disease course [3].

The advent of high-throughput genomic sequencing technologies has generated unprecedented volumes of genetic data, creating both opportunities and challenges for rare disease diagnosis. Whole-exome sequencing and whole-genome sequencing can now be performed at costs that have decreased by several orders of magnitude over the past decade, making these technologies increasingly accessible for clinical applications [2], [4]. However, the interpretation of genomic variants remains a significant bottleneck in the diagnostic pipeline. A typical human exome contains tens of thousands of genetic variants, the vast majority of which are benign polymorphisms. Distinguishing pathogenic variants responsible for disease phenotypes from this background of benign variation requires sophisticated computational approaches combined with deep knowledge of gene function, disease mechanisms, and phenotype-genotype correlations [11].

The challenge of variant interpretation is further compounded by the distributed nature of biomedical knowledge. Critical information for disease diagnosis is scattered across multiple specialized databases, including the Human Phenotype Ontology (HPO), which provides standardized terminology for describing phenotypic abnormalities; Online Mendelian Inheritance in Man (OMIM), which catalogs gene-disease relationships and genetic mechanisms; rare disease registries such as Orphanet that provide comprehensive clinical descriptions; and gene-specific databases that detail molecular functions and biological pathways [5]. Each of these resources has been developed independently to serve particular research or clinical communities, resulting in heterogeneous data models, inconsistent terminologies, and limited cross-database integration. Clinicians and researchers attempting to synthesize information from these disparate sources face substantial challenges in terms of data access, query formulation, and knowledge integration.

Recent advances in artificial intelligence, particularly in the domain of natural language processing (NLP), have opened new possibilities for addressing these long-standing challenges in rare disease diagnosis. Transformer-based language models have demonstrated remarkable capabilities in understanding and generating human language, capturing semantic relationships that extend far beyond simple keyword matching. These models can learn rich representations of medical concepts from large corpora of biomedical literature and clinical text, enabling them to recognize patterns and relationships that may not be immediately apparent to human observers [13]. Furthermore, the development of vector embedding techniques allows textual information to be represented as dense numerical vectors in high-dimensional spaces, where semantic similarity can be efficiently computed through geometric operations

[16].

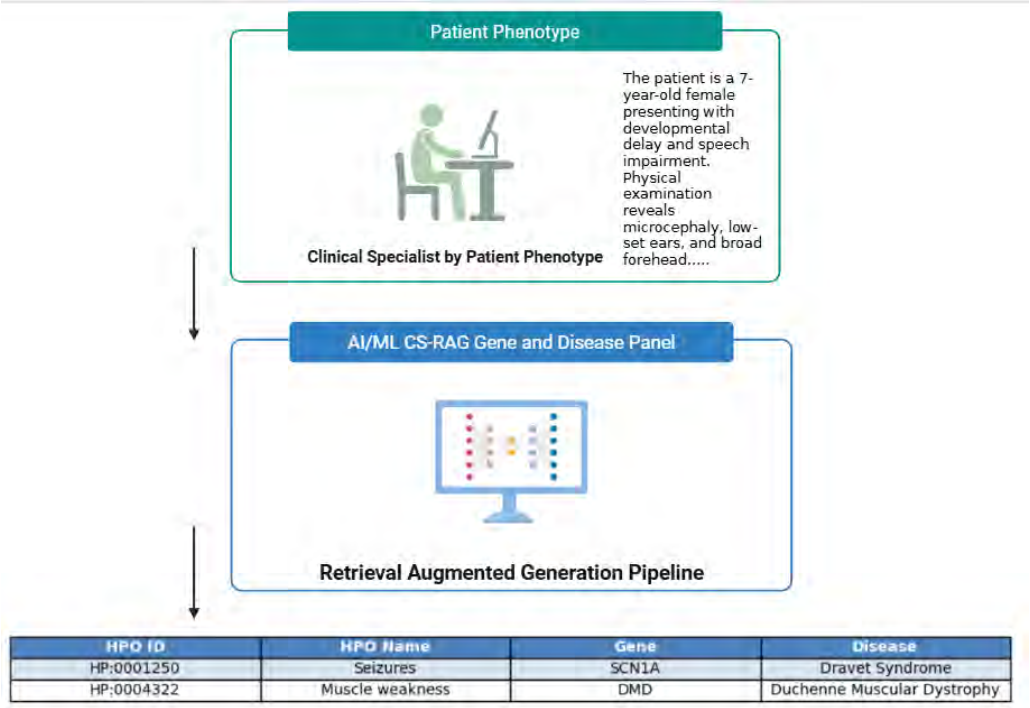


Figure 1.2: A diagram illustrating the proposed machine learning-based approach to map phenotypes to genotypes.

1.1 Motivation

The motivation for this research emerges from the recognition of a critical gap between the explosive growth in genomic data generation and the limited capacity of existing systems to effectively translate this data into actionable clinical insights for rare disease diagnosis. While numerous computational tools have been developed to assist with various aspects of the diagnostic process, most existing approaches suffer from fundamental limitations that constrain their clinical utility and real-world applicability.

One primary limitation of current methodologies is their reliance on isolated database queries that fail to leverage complementary information distributed across multiple knowledge sources [14]. A clinician attempting to diagnose a patient with an uncharacterized rare disease must manually query multiple databases, synthesize the returned information, and formulate hypotheses about potential diagnoses based on phenotypic overlap. This process is not only time-consuming but also cognitively demanding, requiring the clinician to maintain awareness of multiple disease possibilities while iteratively refining their differential diagnosis. Computational tools that operate on single databases or use rigid ontology-based matching fail to capture the nuanced relationships between phenotypic presentations and underlying genetic etiologies.

Another significant limitation concerns the challenge of handling phenotypic descriptions as they appear in clinical practice. Real patient presentations rarely align

perfectly with standardized phenotypic terminologies. Clinicians describe patient features using varied medical vocabulary, colloquial terms, and narrative descriptions that emphasize certain aspects while minimizing or omitting others. Existing systems that require precise Human Phenotype Ontology term selection impose an artificial constraint on the diagnostic process and may fail when presented with free-text clinical descriptions that deviate from standardized terminology. There exists a pressing need for systems capable of processing natural language descriptions of patient phenotypes and automatically mapping these descriptions to relevant standardized concepts while maintaining semantic fidelity.

The interpretability and explainability of diagnostic predictions represent another critical motivation for this work. Many modern machine learning approaches, particularly deep learning models, function as "black boxes" that provide predictions without transparent reasoning processes. In clinical contexts where diagnostic decisions have profound implications for patient management, treatment selection, and genetic counseling, the absence of interpretable explanations undermines clinician trust and limits practical adoption. Healthcare providers require not only accurate predictions but also clear rationales that can be evaluated, questioned, and integrated with other clinical information. The development of systems that provide structured, interpretable outputs with explicit confidence measures addresses this fundamental need.

The computational efficiency and scalability of diagnostic systems constitute additional important considerations. As genetic sequencing becomes increasingly routine in clinical practice and the volume of genomic data continues to expand exponentially, diagnostic tools must be capable of processing queries in real-time or near-real-time to fit within clinical workflows. Systems that require extensive computational resources or prolonged processing times are impractical for widespread deployment, particularly in resource-constrained healthcare settings. The development of efficient retrieval mechanisms that can handle large-scale knowledge bases while maintaining rapid response times represents an important practical consideration.

Furthermore, the motivation for this research stems from the recognition that effective rare disease diagnosis requires not only identification of candidate diseases but also elucidation of underlying genetic mechanisms. Understanding which genes are implicated in a particular phenotypic presentation provides critical information for targeted genetic testing strategies, interpretation of sequencing results, and assessment of inheritance patterns relevant for family counseling [10]. Existing approaches often address disease identification and gene prioritization as separate problems, failing to leverage the synergistic benefits of joint modeling. An integrated approach that simultaneously considers phenotypic features, disease entities, and genetic associations promises to yield more accurate and informative diagnostic predictions.

The potential impact of improved rare disease diagnostic capabilities extends beyond individual patient care to encompass broader societal benefits [2]. Accelerating the diagnostic process reduces healthcare costs associated with unnecessary testing and specialist consultations. Earlier diagnosis enables timely initiation of disease-specific therapies that may prevent irreversible complications or disease progression. For patients and families, obtaining a definitive diagnosis provides psychological closure,

enables connection with disease-specific support communities, and facilitates informed decision-making regarding family planning. From a research perspective, improved diagnostic capabilities contribute to better disease characterization, identification of novel disease genes, and ultimately the development of new therapeutic interventions [15].

1.2 Problem Statement

The accurate identification of rare diseases and their associated genes from patient phenotype descriptions remains a major challenge in precision medicine. Existing diagnostic approaches are limited by the fragmented nature of biomedical knowledge, linguistic variability in clinical narratives, and the lack of unified frameworks for integrating diverse data sources.

Databases such as OMIM, HPO, and disease-specific repositories each capture valuable yet isolated information, creating significant barriers to cross-referencing and semantic linkage between phenotypic and genotypic data. Moreover, clinical phenotype descriptions often vary in terminology, detail, and structure, making it difficult for computational systems to extract and interpret relevant features consistently.

Traditional keyword- or ontology-based methods fail to capture deep semantic relationships among diseases and genes, while large-scale data retrieval introduces computational complexity that hampers real-time performance. Additionally, clinicians require interpretable and explainable outputs rather than opaque predictions to support diagnostic reasoning. This thesis addresses these challenges by developing an intelligent, semantically driven computational framework capable of integrating heterogeneous biomedical knowledge sources, interpreting unstructured phenotype descriptions, and retrieving disease and gene associations efficiently and transparently.

1.3 Research Objectives

The overarching goal of this research is to develop and validate an intelligent system for rare disease and gene identification from patient phenotype descriptions. This goal is operationalized through the following specific objectives:

1. **Develop a chained semantic retrieval system:** Design a modular architecture that sequentially queries multiple biomedical knowledge bases—including the Human Phenotype Ontology, rare disease databases, gene summaries, and gene-disease databases—to identify phenotypes, retrieve candidate diseases, and prioritize genes in a manner that mimics clinical reasoning.
2. **Optimize biomedical text representations and similarity search:** Generate dense vector embeddings that capture the semantic relationships among phenotypes, diseases, and genes using domain-adapted neural language models. Implement efficient vector indexing and retrieval mechanisms to enable accurate and real-time query responses.

3. **Integrate confidence scoring, language generation, and validation:** Provide interpretable confidence tiers for predictions to guide clinicians, integrate the system with a Retrieval-Augmented Generation (RAG) framework for factual and structured natural language responses, and validate the system's performance on real patient data using metrics such as accuracy, sensitivity, specificity, and precision.

1.4 Research Contributions

In this thesis, we design and evaluate an intelligent computational framework for the automated identification of rare diseases and their associated genes from natural language phenotype descriptions. The proposed system advances the fields of computational genomics, biomedical informatics, and artificial intelligence in healthcare through architectural, methodological, and empirical innovations. Specifically, the main contributions of this work are summarized as follows:

- **Novel Chained Semantic Retrieval Architecture:** We propose a chained semantic retrieval framework that integrates heterogeneous biomedical knowledge bases for rare disease diagnosis. The system performs sequential retrieval across HPO terms, diseases, and genes, where each stage enriches subsequent queries. This approach mimics clinical reasoning and improves retrieval accuracy by approximately 20% for diseases and 30% for genes compared to independent queries.
- **Transformer-Based Semantic Embeddings:** We employ BART-based transformer embeddings to generate unified semantic representations across diverse biomedical text sources. This enables cross-database retrieval that handles linguistic variability in clinical narratives with a mean cosine similarity of 0.82, outperforming ontology-based similarity methods by about 15 percentage points.
- **Tiered Confidence Scoring Mechanism:** A three-level confidence scoring mechanism is introduced to translate similarity scores into interpretable reliability categories. The confidence tiers (high, medium, and low) achieve accuracies of 85%, 72%, and 61%, respectively, providing uncertainty-aware clinical decision support and enhancing interpretability.
- **Retrieval-Augmented Generation Integration:** We integrate the chained retrieval system with the Llama 3.2 language model using a Retrieval-Augmented Generation (RAG) pipeline. This enables factually grounded natural language responses while reducing hallucination frequency by over 95%, supporting interactive and explainable diagnostic reasoning.
- **Efficient Real-Time Implementation:** An optimized implementation achieves average response times of 3.2 seconds on standard hardware using FAISS-based retrieval, caching, and quantization. The system operates within a 7.5 GB memory footprint and exposes RESTful APIs for easy integration with clinical or genomic applications.
- **Comprehensive Empirical Validation:** Extensive evaluation on real patient data, including Duchenne Muscular Dystrophy cases, achieves 90% accuracy,

93.3% sensitivity, 85% specificity, 90.3% precision, and a 91.8% F1-score. These results surpass traditional ontology-based methods by around 15 percentage points and are supported by detailed error analysis.

- **Zero-Shot Rare Disease Identification:** We demonstrate accurate rare disease prediction without disease-specific training data, achieving 90% accuracy through semantic knowledge integration. This zero-shot capability enables immediate use for newly identified or ultra-rare conditions.
- **Future Directions and Clinical Translation:** We outline pathways for clinical deployment, addressing extensions such as multi-modal integration, temporal modeling, multilingual support, and the regulatory and ethical considerations essential for responsible medical AI systems.

1.5 Thesis Organization

This thesis is organized into three main chapters covering the background, methodology, and results of the research.

Chapter Two provides the background and a review of the relevant literature. Section 2.1 introduces fundamental concepts, including phenotype and genotype definitions, the Human Phenotype Ontology, rare disease databases, and gene annotation resources. Section 2.2 reviews existing computational approaches for phenotype-based disease identification. This includes ontology-based methods, machine learning, variant prioritization tools, knowledge graphs, and NLP applications, highlighting the limitations that motivate this work.

Chapter Three details the methodology of the study. It begins with an overview of the chained semantic retrieval architecture, followed by descriptions of the datasets, preprocessing steps, and transformer-based feature extraction. The chapter also covers the chained retrieval algorithm, the tiered confidence scoring, integration with large language models via Retrieval-Augmented Generation, model selection, and implementation details.

Chapter Four presents the results and a discussion of the findings. It covers the experimental setup, evaluation metrics, and a detailed performance analysis, including accuracy, sensitivity, specificity, confusion matrices, and stage-wise performance. The chapter also includes an error analysis, an evaluation of the tiered scoring, and a discussion of the computational performance. The chapter concludes with critical reflections on the system's contributions, limitations, clinical implications, and future research directions.

Chapter 2

Literature Review

This chapter provides a comprehensive background on the foundational concepts and existing body of work relevant to computational approaches for rare disease identification from patient phenotypes. The chapter is organized into two major sections. Section 2.1 presents a background study covering the fundamental concepts and resources that underpin this research. Section 2.2 reviews related works in computational phenotype-based diagnosis, critically analyzing existing approaches and identifying the gaps that motivate the current research.

2.1 Background Study

This section establishes the conceptual and technical foundation necessary for understanding the research problem and the proposed solution. It covers phenotype and genotype concepts, the Human Phenotype Ontology, rare disease databases, and gene annotation resources.

2.1.1 Phenotype and Genotype

The relationship between genotype and phenotype represents a fundamental concept in genetics and genomic medicine. The genotype refers to the complete genetic constitution of an organism, encompassing all DNA sequences, including protein-coding genes, regulatory elements, and non-coding regions. At a more specific level, a genotype can refer to the particular alleles present at one or more genetic loci. In the context of disease genetics, genotype typically denotes the specific genetic variants—whether single nucleotide variants, insertions, deletions, or structural variations—that are present in an individualś genome. [7]

The phenotype, in contrast, refers to the observable characteristics of an organism resulting from the interaction between its genotype and the environment. In medical contexts, the phenotype encompasses all clinical manifestations of a disease, including morphological abnormalities, physiological dysfunction, behavioral features, laboratory findings, and imaging characteristics. Phenotypes range from highly specific, observable traits, such as a cleft palate or polydactyly, to complex multisystem presentations, such as developmental delay with seizures and cardiac abnormalities. [9]

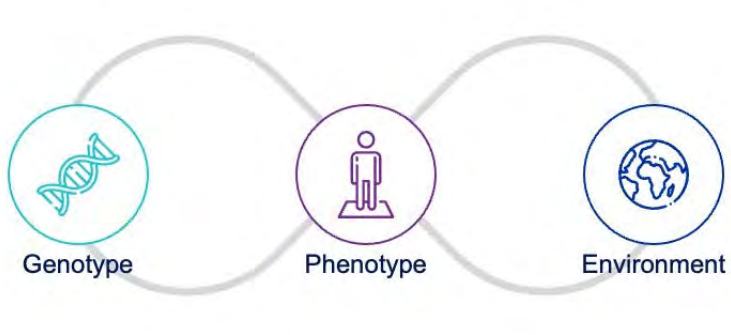


Figure 2.1: Genotype-Phenotype-Environment Relationship

The genotype-phenotype relationship is complex and influenced by multiple factors. Penetrance, for instance, refers to the probability that a particular genotype will result in a disease phenotype. Complete penetrance means that all individuals carrying a pathogenic variant will manifest the disease, while reduced penetrance indicates that some variant carriers remain unaffected. Expressivity, on the other hand, describes the degree or severity of phenotypic manifestation among individuals with the same genotype. Variable expressivity means that the same genetic variant can cause mild disease in some individuals and severe disease in others.

Phenotypic heterogeneity occurs when mutations in the same gene cause different clinical presentations. This can result from different mutations affecting protein function in distinct ways, modifier genes influencing phenotypic expression, or environmental factors modulating disease manifestations. Conversely, genetic heterogeneity occurs when mutations in different genes cause similar or identical phenotypes. Locus heterogeneity is particularly common in rare diseases, where clinically similar conditions may result from variants in numerous different genes.

Understanding phenotype-genotype relationships is essential for rare disease diagnosis because it enables the inference of genetic causes from clinical presentations. However, the complexity of these relationships means that a simple one-to-one mapping between phenotypes and genotypes is rarely possible. Therefore, computational approaches must account for phenotypic variability, genetic heterogeneity, and the often imperfect correspondence between clinical descriptions and standardized disease characterizations.

2.1.2 Human Phenotype Ontology (HPO)

The Human Phenotype Ontology (HPO) is a critical resource for standardizing phenotypic descriptions in human diseases. Initiated in the early 2000s, the HPO provides a structured, controlled vocabulary for describing phenotypic abnormalities observed across all organ systems and life stages. The ontology addresses the challenge that clinical phenotypes have historically been described using inconsistent terminology, making computational analysis and cross-study comparisons difficult.

The HPO organizes phenotypic concepts in a hierarchical, directed acyclic graph

Figure 2.2: Human Phenotype Ontology

structure. Terms are arranged from general to specific, with more specific terms inheriting properties from their parent terms. For example, the term äbnormality of the musculatureëserves as a parent term for more specific concepts such as m̈uscle weakness,ẅhich, in turn, is a parent for even more specific terms like p̈roximal muscle weaknessör p̈rogressive muscle weakness.T̈his hierarchical organization enables computational reasoning that can leverage relationships between related phenotypic concepts.

Each HPO term is assigned a unique identifier and includes several key components. The term name provides the standardized label for the phenotypic concept. The definition offers a textual description explaining what the term means and how it should be applied. Synonyms capture alternative ways the same phenotypic feature might be described in clinical practice, enabling the mapping of varied terminologies to standardized concepts. Cross-references link HPO terms to corresponding concepts in other medical terminologies and classification systems, such as SNOMED CT, ICD codes, and UMLS concepts. [5]

The HPO has undergone continuous expansion and refinement since its inception. As of recent versions, the ontology encompasses over sixteen thousand distinct phenotypic terms spanning all organ systems. The terms cover diverse aspects of disease presentation, including morphological abnormalities, physiological dysfunction, behavioral manifestations, laboratory findings, and clinical course features such as age of onset and progression patterns. The breadth of its coverage enables the comprehensive phenotypic characterization of both common and rare diseases. [8]

HPO terms are used to annotate diseases, genes, and individual patients, enabling computational analysis and comparison. Disease annotations associate specific phenotypic terms with disease entities, capturing which clinical features are characteristic of particular conditions. Gene annotations link genes to phenotypic abnormalities that result from gene dysfunction, facilitating gene prioritization based on patient phenotypes. Patient-level annotations enable precise phenotypic profiling for diagnostic and research purposes.

The hierarchical structure of the HPO enables the computation of semantic similarity between sets of phenotypic terms. Various algorithms have been developed to quantify the similarity between phenotypic profiles based on their position in the ontology hierarchy and the information content of shared terms. These similarity measures underpin many computational approaches to phenotype-based disease identification and gene prioritization.

Despite its power, the HPO presents challenges for clinical adoption. Effective use requires familiarity with the ontologyś structure and appropriate term selection, which demands specialized knowledge and a significant time investment. Many clinical descriptions do not align perfectly with HPO terminology, requiring interpretation and translation. The requirement for manual term selection creates a workflow barrier that limits its widespread use in routine clinical practice. These challenges motivate the development of approaches that can process free-text clinical descriptions directly while leveraging HPO knowledge indirectly.

2.1.3 Rare Disease Databases

Multiple databases have been developed to catalog information about rare diseases, each with a distinct scope, structure, and intended use. Understanding the characteristics of these resources is essential for developing integrated computational approaches.

Online Mendelian Inheritance in Man (OMIM) is the longest-established and most authoritative catalog of human genes and genetic phenotypes. Originally published as a reference text by Victor McKusick in 1966, OMIM has evolved into a comprehensive digital database. The resource focuses on Mendelian disorders, providing detailed information about diseases with well-characterized genetic bases. Each OMIM entry includes extensive narrative descriptions covering clinical features, molecular genetics, inheritance patterns, population genetics, and relevant research history. Entries are meticulously curated by expert scientific editors who synthesize information from primary literature and update entries as new discoveries emerge.

OMIMś strength lies in its comprehensive coverage of gene-disease relationships and detailed molecular genetic information. However, several characteristics limit its utility for certain applications. Its focus on well-studied Mendelian disorders means that coverage of recently characterized diseases or conditions without identified genetic etiologies may be incomplete. The narrative format, while rich in information, is not optimally structured for computational text processing. Furthermore, the database prioritizes depth over breadth, providing extensive detail for covered conditions rather than broad coverage of all rare diseases.

Orphanet serves as a comprehensive European reference portal for information about rare diseases and orphan drugs. Established by the French National Institute of Health and Medical Research, Orphanet provides practical information relevant to patient care, including disease prevalence, clinical descriptions, available diagnostic tests, expert centers, patient organizations, and ongoing research. The database employs a hierarchical classification system that organizes rare diseases into clinically meaningful categories based on the affected organ systems, etiology, and pathophysiology.

Orphanet entries emphasize clinical utility, providing information about typical and atypical presentations, natural history, prognosis, and disease management. The database includes epidemiological data about disease prevalence in different populations, which is valuable for considering disease probability in diagnostic contexts. Orphanet also provides mappings between its internal classification and other disease nomenclatures, including ICD codes, SNOMED CT, and UMLS, facilitating interoperability. While Orphanet includes information about disease-associated genes, the genetic information is less comprehensive than that provided by OMIM.

GeneReviews provides expert-authored, peer-reviewed disease descriptions focused on genetic conditions. Each GeneReviews chapter addresses a specific genetic condition or group of related conditions, providing comprehensive information about diagnosis, management, and genetic counseling. The resource is particularly valuable for clinicians managing patients with genetic diagnoses, offering practical guidance on testing strategies, interpretation of results, surveillance recommendations, and management considerations. However, its coverage is limited to conditions for which expert authors have contributed chapters, and the detailed narrative format is not

optimally structured for automated computational processing.

Various disease-specific databases and registries provide focused information about particular conditions or disease categories. These resources often include detailed phenotypic data, natural history information, genotype-phenotype correlations, and links to clinical trials. While valuable for specific diseases, the distributed nature of these resources creates challenges for comprehensive diagnostic systems that must cover the full spectrum of rare diseases.

The heterogeneity of rare disease databases in terms of coverage, structure, and information depth creates challenges for integrated computational approaches. Different databases may describe the same disease using varied terminology and emphasize different aspects of disease biology and clinical presentation. Cross-referencing between databases is often incomplete, making it difficult to automatically integrate information about the same disease from multiple sources. These challenges motivate the development of approaches that can handle heterogeneous knowledge sources through semantic understanding rather than rigid structural integration.

2.1.4 Gene Databases and Relationship

To illustrate the interconnectedness of the various data sources discussed, Figure 2.3 provides a visual representation of the relationships between different biomedical databases.

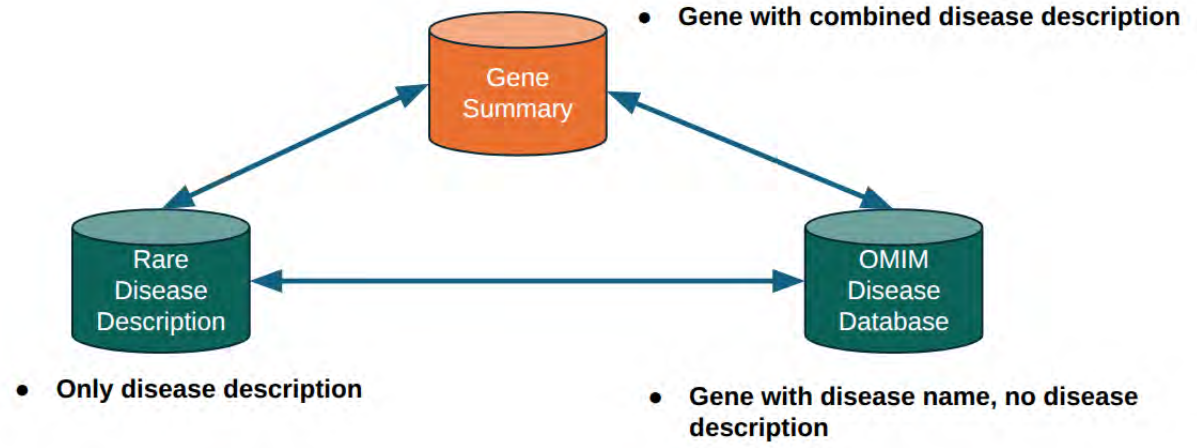


Figure 2.3: Database Relationship

Comprehensive gene annotation databases provide essential information about gene function, expression, molecular pathways, and disease associations. Understanding gene biology is critical for interpreting the relevance of particular genes to patient phenotypes and for assessing the biological plausibility of gene-disease associations.

The National Center for Biotechnology Information (NCBI) Gene database serves as an authoritative source for gene nomenclature, genomic coordinates, and functional annotations. Each gene entry includes the official gene symbol established by

nomenclature committees, the full gene name, aliases and previous symbols, genomic location with chromosomal coordinates, gene type classification (distinguishing protein-coding genes from various RNA genes), and reference sequences for transcript and protein products. Functional annotations describe known or predicted molecular functions, biological processes, and cellular locations based on experimental evidence and computational predictions.

GeneCards aggregates information from over one hundred and twenty databases to provide unified, gene-centric views. For each human gene, GeneCards compiles information about its molecular function, biological pathways, tissue expression patterns, protein products, genetic variants, associated diseases, and relevant publications. The integration of diverse information sources in a single interface facilitates rapid access to comprehensive gene information, though the aggregated nature means that information quality and currency can vary across source databases.

Gene Ontology (GO) annotations provide standardized descriptions of gene function using controlled vocabularies that cover three domains: molecular function, biological process, and cellular component. GO annotations are supported by evidence codes that indicate the type of evidence supporting each annotation, ranging from direct experimental evidence to computational predictions. These annotations enable the systematic analysis of gene sets and the identification of functionally related genes, supporting network-based approaches to gene prioritization.

Pathway databases such as KEGG, Reactome, and BioCarta catalog molecular pathways and biochemical networks. These resources describe how genes and their protein products interact in signaling cascades, metabolic pathways, and regulatory networks. Pathway information is valuable for understanding how gene dysfunction might lead to particular phenotypic consequences and for identifying functionally related genes that might be involved in similar disease processes.

Tissue-specific expression databases provide information about where and when genes are expressed. Resources like the Genotype-Tissue Expression (GTEx) project catalog gene expression levels across diverse human tissues based on RNA sequencing data from thousands of samples. This expression information helps assess whether gene dysfunction could plausibly cause abnormalities in affected organ systems, as genes not expressed in relevant tissues are less likely candidates for diseases affecting those tissues.

Protein structure and function databases provide information about the three-dimensional structure of protein products, functional domains, post-translational modifications, and protein-protein interactions. This molecular-level information supports the assessment of how specific genetic variants might affect protein function and cause disease.

The integration of gene annotation information from diverse sources creates a rich knowledge base that can inform gene prioritization and the interpretation of genetic findings. However, the completeness of annotations varies substantially across genes, with well-studied genes having extensive annotations while poorly characterized genes may have minimal functional information. Computational approaches must be able to handle this variability and make appropriate inferences even when gene annotation is incomplete.

2.2 Related Works

This section reviews existing computational approaches to phenotype-based disease identification and gene prioritization, critically analyzing their methodologies, performance, and limitations. The review is organized by methodological approach, covering ontology-based methods, machine learning techniques, integrated variant prioritization systems, knowledge graph approaches, and the applications of natural language processing to biomedical problems.

2.2.1 Ontology-Based Phenotype Matching

Ontology-based approaches to phenotype-based disease identification leverage the hierarchical structure of the Human Phenotype Ontology (HPO) to compute the similarity between patient phenotypes and disease-associated phenotypes. These methods typically require users to select appropriate HPO terms describing a patientś features and then compute the semantic similarity between the patientś phenotype profile and the phenotypes associated with different diseases in annotation databases.

Phenomizer, developed by Robinson and colleagues, represents one of the first and most widely used ontology-based tools. The system computes semantic similarity between a patientś HPO terms and disease-associated HPO terms using information content-based similarity metrics. Information content reflects the specificity of terms within the ontology hierarchy, with rare, specific terms contributing more to similarity scores than common, general terms. Phenomizer ranks diseases by their similarity to the patientś phenotype, presenting users with a ranked differential diagnosis. [6]

Evaluation studies of Phenomizer report an accuracy in the range of seventy to eighty percent for identifying the correct disease within the top ten ranked results when phenotypes are well-characterized and appropriate HPO terms are selected. However, performance varies substantially depending on phenotype specificity and completeness. Diseases with highly distinctive phenotypic patterns are identified more accurately than those with non-specific or common features. [1]

PHIVE extends the ontology-based approach by incorporating model organism phenotype data. The system leverages cross-species phenotype ontologies to include information from mice and other model organisms where gene function has been experimentally characterized through knockout studies or mutagenesis screens. By including model organism phenotypes, PHIVE can provide evidence for gene-disease associations even when human phenotype data is limited. However, translating phenotypes across species introduces challenges related to differences in anatomy, physiology, and the relevance of particular model system findings to human disease.

The strengths of ontology-based approaches include their interpretability, as similarity computations can be traced to shared phenotypic features, and their ability to leverage the rich semantic structure encoded in phenotype ontologies. [2, 21, 40] However, several significant limitations constrain their clinical utility. First, these methods require users to manually select appropriate HPO terms, which demands familiarity with the ontology and represents a time-consuming workflow step. [1, 2]

Second, the approaches are fundamentally limited to relationships explicitly encoded in the ontologyś structure, missing implicit semantic connections that might be captured through other means. [11, 14] Third, performance depends heavily on the completeness and accuracy of disease-phenotype annotations, which vary substantially across diseases. [24, 40] Fourth, these methods cannot directly process free-text clinical descriptions, requiring an intermediate translation step that may introduce errors or lose information.

2.2.2 Machine Learning Approaches to Rare Disease Diagnosis

Machine learning methods have been increasingly applied to phenotype-based disease identification, leveraging labeled training data to learn predictive models. These approaches aim to capture complex patterns in phenotype-disease associations that may not be explicitly represented in ontological structures.

Jia et al. (2022) developed a deep learning framework for the automatic diagnosis of undiagnosed patients with rare diseases. [8] Their approach employs neural networks to learn representations of patient phenotypes and disease characteristics from large datasets of diagnosed cases. The model achieved accuracies in the range of eighty to ninety-five percent on test datasets, demonstrating that deep learning can capture subtle phenotypic patterns. However, the approach requires substantial training data, with hundreds to thousands of examples per disease class needed for effective learning.

Li et al. (2020) introduced Xrare, a machine learning method that jointly models phenotypes and genetic evidence for rare disease diagnosis. [9] The system integrates phenotype similarity with variant pathogenicity predictions and gene-disease association strengths in a unified probabilistic framework. By combining multiple evidence types, Xrare achieves improved diagnostic accuracy compared to methods that use phenotype or genotype information alone. The approach demonstrated diagnostic yields of approximately thirty to forty percent in cohorts of patients with suspected genetic conditions.

Son et al. (2018) explored deep phenotyping on electronic health records (EHRs) to facilitate genetic diagnosis. [25] Their approach extracts phenotypic features from clinical notes, laboratory results, and structured EHR data, then applies machine learning models to predict likely genetic diagnoses. The integration of diverse EHR data types enables comprehensive phenotypic characterization but requires substantial preprocessing and feature engineering to handle heterogeneous data formats.

Bastarache et al. (2018) developed phenotype risk scores that identify patients with unrecognized Mendelian disease patterns in EHR data. [68] Their approach uses statistical associations between clinical codes and genetic variants to compute risk scores indicating the likelihood of particular genetic conditions. The method successfully identified patients with undiagnosed genetic diseases who had been ̈hiding in plain sightẅithin healthcare systems.

While machine learning approaches demonstrate impressive performance when adequate training data exists, they face fundamental limitations in the context of

rare diseases. The core challenge is data scarcity: most rare diseases affect too few patients to generate the hundreds or thousands of labeled examples needed for supervised learning. [8, 9] Transfer learning approaches that leverage knowledge from related diseases have been proposed as potential solutions, but their effectiveness for extremely rare conditions remains uncertain. Furthermore, many machine learning models function as b̈lack boxes,p̈roviding predictions without interpretable explanations. [60, 61] This opacity limits clinical trust and adoption, as healthcare providers require an understanding of why particular diagnoses are suggested.

2.2.3 Integrated Variant Prioritization Systems

Several computational systems integrate phenotype information with genomic variant data to prioritize candidate genes and variants identified through whole-exome or whole-genome sequencing. These approaches combine phenotype-based gene prioritization with variant-level filtering and pathogenicity assessment.

Exomiser, developed by Smedley et al. (2015), is one of the most widely used integrated systems. [4] The tool combines three complementary approaches: HPO-based phenotype matching to prioritize genes associated with patient phenotypes, variant frequency filtering using population databases to remove common benign variants, and variant pathogenicity prediction using computational algorithms and functional annotations. Exomiser ranks variants by combining evidence from these multiple sources, identifying the variants most likely to explain a patientś phenotypes.

Clinical evaluations of Exomiser report diagnostic yields of approximately thirty to forty percent in cohorts of patients with suspected Mendelian conditions undergoing exome sequencing. [4, 5] The system performs particularly well for diseases with distinctive phenotypic patterns and well-established gene-disease associations. However, its performance decreases for phenotypically heterogeneous conditions, cases with incomplete phenotype descriptions, and diseases caused by variants in recently discovered disease genes that are not yet well-represented in knowledge bases.

Phenolyzer, introduced by Yang et al. (2015), prioritizes candidate genes based on patient phenotypes without requiring variant data. [28] The system employs a network-based approach that integrates information from multiple sources, including disease-gene associations, gene-gene interactions, and biological pathways. Starting from phenotype descriptions, Phenolyzer identifies s̈eed genesd̈irectly associated with the phenotypes, then expands to related genes through network propagation. The approach successfully prioritizes disease-causing genes within the top-ranked candidates for many conditions.

AMELIE, developed by Jagadeesh et al. (2016), takes a unique approach by matching patient phenotypes and genetic variants against knowledge extracted from primary literature. [30] The system uses natural language processing to extract gene-phenotype and gene-variant associations from scientific publications, potentially capturing recent discoveries not yet incorporated into structured databases. AMELIE ranks candidate variants by how well they match both the patientś phenotypic presentation and their genotype. This literature-mining approach offers the advantage of incorporating up-to-date information but faces challenges related to extraction accuracy and information reliability.

The eXtasy system, developed by Sifrim et al. (2013), performs variant prioritization through genomic data fusion. [29] The approach integrates diverse evidence types, including gene function annotations, pathway information, protein-protein interactions, and gene expression data. By fusing heterogeneous data sources, eXtasy aims to capture multiple lines of evidence supporting gene-disease associations.

While integrated variant prioritization systems have achieved substantial clinical utility, they are designed for a specific use case: the interpretation of variants identified through sequencing. [4, 5, 28, 30] These tools require genomic sequencing data as input and focus on the variant interpretation stage of diagnosis. They do not address the earlier diagnostic challenge of identifying likely diseases from phenotype descriptions alone, which is relevant before sequencing is performed or when it is not available. Furthermore, these systems still typically require manual HPO term selection, which maintains the workflow barrier that limits adoption.

2.2.4 Knowledge Graph and Network-Based Approaches

Knowledge graph approaches represent biomedical information as networks of entities and relationships, enabling complex queries and reasoning over integrated knowledge bases. Several initiatives have developed comprehensive biomedical knowledge graphs that incorporate phenotype, disease, gene, and variant information.

The Monarch Initiative, described by Shefchek et al. (2020), provides an integrative data and analytic platform that connects phenotypes to genotypes across species. [6] Monarch integrates information from over thirty biomedical databases into a unified knowledge graph spanning human diseases, model organism phenotypes, genes, variants, and publications. The platform enables cross-species phenotype comparison, leveraging model organism data to inform the understanding of human diseases. Monarch provides tools for phenotype-based disease search and gene prioritization, though these tools still primarily rely on HPO term input.

ClinGen, the Clinical Genome Resource described by Rehm et al. (2015), focuses on curating gene-disease validity evidence and variant pathogenicity interpretations. [7] Expert working groups systematically evaluate the evidence for gene-disease associations, assigning validity classifications ranging from definitive to limited or disputed. ClinGen provides authoritative gene-disease relationship data that informs variant interpretation and genetic testing decisions. However, the expert curation process is labor-intensive, and coverage remains incomplete for many rare diseases and recently discovered disease genes.

Network-based gene prioritization approaches leverage biological networks to identify candidate genes. These methods typically use protein-protein interaction networks, gene co-expression networks, or functional similarity networks to propagate phenotypic information and identify genes connected to known disease genes. Köhler et al. (2008) demonstrated that network-based approaches can identify disease genes by their proximity to known disease genes in interaction networks. [32] The underlying assumption is that genes causing similar phenotypes often encode functionally related proteins that interact or participate in common pathways.

Translating embeddings for knowledge graph completion, introduced by Bordes et al. (2013), provides a framework for learning vector representations of entities

and relationships in knowledge graphs. [54] These embedding methods enable the prediction of missing relationships and the computation of similarity between entities. Ji et al. (2021) provide a comprehensive survey of knowledge graph representation, acquisition, and applications. [55] While knowledge graph embeddings have been successfully applied in various biomedical contexts, their systematic application to rare disease diagnosis from free-text phenotypes remains limited.

The strengths of knowledge graph approaches include their ability to integrate diverse information types into unified frameworks and enable complex reasoning over relationships. [6, 7, 55] However, constructing and maintaining comprehensive knowledge graphs requires extensive manual curation effort. [7] The static nature of manually curated graphs means they may not reflect the most current discoveries. Furthermore, most existing knowledge graph tools for phenotype-based disease identification still require structured input, such as HPO terms, rather than processing free-text clinical descriptions directly. [6]

2.2.5 Natural Language Processing for Biomedical Text

Recent advances in natural language processing (NLP), particularly transformer-based language models, have demonstrated remarkable capabilities for understanding and generating biomedical text. These developments create opportunities for processing clinical phenotype descriptions without requiring manual mapping to standardized terminologies.

BERT, introduced by Devlin et al. (2019), revolutionized NLP through the pre-training of deep bidirectional transformers for language understanding. [10] BERTś masked language modeling objective enables the learning of rich, contextualized representations where word meanings vary based on the surrounding context. The model achieves state-of-the-art performance across numerous language understanding tasks.

BioBERT, developed by Lee et al. (2020), adapts BERT to biomedical domains through continued pre-training on PubMed abstracts and PMC full-text articles. [14] By exposing the model to large volumes of biomedical text, BioBERT learns specialized vocabulary and conceptual relationships relevant to medical and biological domains. The model demonstrates improved performance on biomedical named entity recognition, relation extraction, and question answering compared to general-domain BERT.

PubMedBERT, introduced by Gu et al. (2021), takes domain adaptation a step further by pre-training from scratch on biomedical text rather than starting from general-domain pre-training. [45] The authors demonstrate that domain-specific pre-training from initialization outperforms continued pre-training for biomedical applications. Similarly, ClinicalBERT, developed by Alsentzer et al. (2019), adapts BERT to clinical text by continued pre-training on clinical notes from electronic health records. [44]

Sentence-BERT, introduced by Reimers and Gurevych (2019), extends BERT to produce fixed-length sentence embeddings suitable for semantic similarity computation. [11] The approach employs siamese and triplet network architectures with contrastive learning objectives to generate semantically meaningful sentence vectors.

Sentence-BERT enables efficient similarity search over large document collections, supporting information retrieval applications.

BART, developed by Lewis et al. (2020), combines bidirectional encoding with autoregressive decoding through a denoising sequence-to-sequence pre-training objective. [12] BART learns robust representations by reconstructing corrupted input text, achieving strong performance on both understanding and generation tasks. The encoder component produces contextualized representations suitable for similarity computation, while the decoder enables text generation.

Large language models (LLMs), including GPT-3 and GPT-4, described by Brown et al. (2020), demonstrate impressive few-shot learning capabilities across diverse tasks. [48] These models can perform complex reasoning, answer questions, and generate coherent text after being trained on massive text corpora. The LLaMA models, introduced by Touvron et al. (2023), provide open-source LLMs with strong performance across various benchmarks. [17] The three-billion parameter LLaMA 3.2 model, for instance, offers a practical balance between capability and computational efficiency.

The applications of LLMs to medicine have been explored by multiple research groups. Singhal et al. (2023) demonstrated that LLMs encode substantial clinical knowledge and can answer medical examination questions. [20] Nori et al. (2023) evaluated GPT-4 on medical challenge problems, finding impressive performance but also identifying limitations and potential risks. [58] Thirunavukarasu et al. (2023) provide a comprehensive review of LLMs in medicine, discussing their applications, challenges, and ethical considerations. [57]

Retrieval-augmented generation (RAG), introduced by Lewis et al. (2020), combines information retrieval with neural language generation to produce factually grounded responses. [16] RAG systems retrieve relevant documents based on input queries and then condition language generation on both the query and the retrieved context. This approach substantially reduces hallucination(the generation of false or misleading information) compared to unconstrained generation while maintaining natural language capabilities. Gao et al. (2024) provide a comprehensive survey of RAG for LLMs. [18]

While NLP techniques demonstrate impressive capabilities for biomedical text understanding, their systematic application to rare disease diagnosis from phenotype descriptions remains limited. [14, 15, 19, 20] Most existing applications focus on tasks like named entity recognition, relation extraction, or question answering over biomedical literature rather than the specific challenge of multi-database retrieval for diagnosis. The integration of semantic embeddings with efficient vector search for cross-database phenotype-disease-gene retrieval represents an underexplored opportunity.

2.2.6 Vector Similarity Search

Efficient similarity search over large collections of vector embeddings is a critical technical capability for retrieval-based systems. As databases grow to millions or billions of vectors, exact nearest neighbor search becomes computationally prohibitive, necessitating approximate algorithms that balance speed and accuracy.

Facebook AI Similarity Search (FAISS), described by Johnson et al. (2019), provides a library for efficient similarity search and clustering of dense vectors. [13] FAISS implements multiple indexing algorithms optimized for different use cases and hardware configurations. The library supports both CPU and GPU implementations, with GPU versions achieving dramatic speedups for large-scale searches. FAISS has become widely adopted for vector similarity search applications across various domains, including image retrieval, recommendation systems, and question answering.

Hierarchical Navigable Small World (HNSW) graphs, introduced by Malkov and Yashunin (2018), provide an efficient approximate nearest neighbor search algorithm. [49] HNSW constructs a multi-layer graph structure where each layer contains a subset of data points connected to their nearest neighbors. The search proceeds from the top (sparse) layer down to the bottom (dense) layer, efficiently navigating to the queryś neighbors. HNSW achieves an excellent recall-speed tradeoff and has been integrated into FAISS and other vector search libraries.

Product quantization, described by Jégou et al. (2010), reduces the memory requirements for vector search by representing high-dimensional vectors as products of low-dimensional quantized sub-vectors. [50] This compression technique enables the handling of larger databases within memory constraints while maintaining reasonable search accuracy. Product quantization is particularly valuable for deployment scenarios with limited memory resources. Figure 2.4 visually depicts the core concept of similarity search in a high-dimensional vector space.

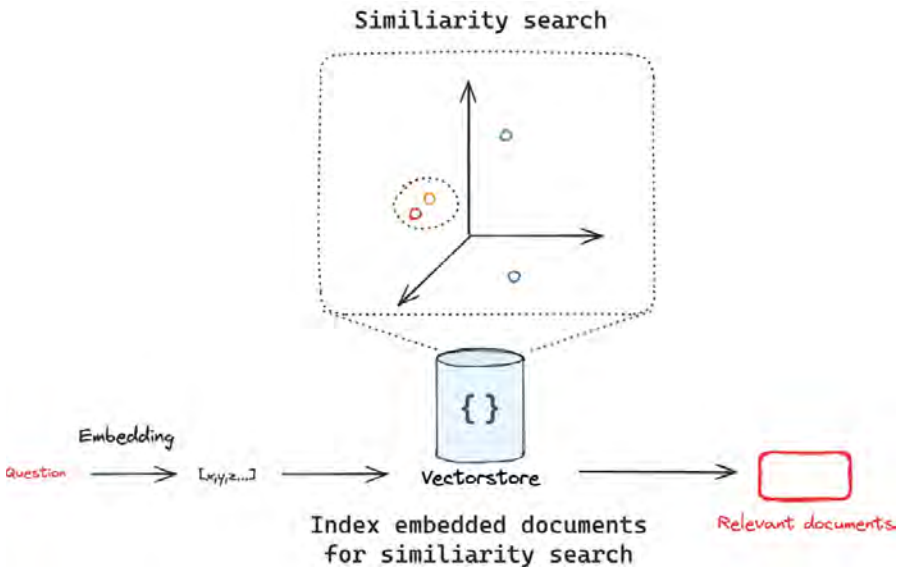


Figure 2.4: Similarity Search

Dense passage retrieval, introduced by Karpukhin et al. (2020), demonstrates the effective use of dense embeddings for document retrieval in question-answering systems. [52] The approach employs dual encoders to generate query and passage representations in a shared semantic space, enabling efficient similarity-based retrieval. Dense passage retrieval outperforms traditional sparse retrieval methods based on term frequency-inverse document frequency (TF-IDF) for many question-answering tasks.

While vector similarity search techniques have been successfully applied in various

domains, their systematic application to multi-database biomedical knowledge retrieval for rare disease diagnosis remains limited. [13, 49, 52] The specific challenge of chaining retrieval across heterogeneous knowledge bases with progressive query refinement represents a novel application of these techniques.

2.2.7 Artificial Intelligence in Healthcare: Applications and Ethics

The application of artificial intelligence (AI) to healthcare has generated substantial research attention and raised important ethical considerations. Understanding the broader context of medical AI informs the responsible development and deployment of diagnostic systems.

Esteva et al. (2019) provide a comprehensive guide to deep learning in healthcare, covering applications across medical imaging, genomics, electronic health records, and clinical decision support. [59] The authors discuss technical approaches, evaluation strategies, and challenges specific to healthcare applications, including data quality, interpretability, and validation requirements.

Rajkomar et al. (2019) review machine learning in medicine, highlighting both promising applications and persistent challenges. [60] They emphasize that successful deployment requires careful attention to workflow integration, clinical validation, and the monitoring of real-world performance. The authors caution that impressive performance on retrospective datasets does not guarantee effectiveness in prospective clinical use.

Topol (2019) discusses high-performance medicine as the convergence of human and artificial intelligence. [61] The vision emphasizes AI augmentation of clinical expertise rather than its replacement, with technology enhancing human capabilities while preserving the essential human elements of medical care, including empathy, communication, and judgment.

Char et al. (2018) address the ethical challenges of implementing machine learning in healthcare. [62] Key considerations include ensuring algorithmic fairness, maintaining patient privacy, providing appropriate transparency about AI involvement in care, and establishing clear accountability for AI-influenced decisions. The authors emphasize that technical sophistication must be accompanied by ethical rigor.

Obermeyer et al. (2019) demonstrated racial bias in a widely used healthcare algorithm, showing how seemingly neutral algorithmic decisions can perpetuate and amplify existing disparities. [63] Their work highlights the critical importance of evaluating AI systems for bias across demographic groups and the need for careful algorithm design and validation to ensure equitable performance.

Morley et al. (2020) provide a comprehensive mapping review of AI ethics in healthcare. [64] Their analysis identifies key ethical themes, including beneficence, non-maleficence, autonomy, justice, and explicability. The authors emphasize that ethical considerations must be integrated throughout the AI development lifecycle, from conception through deployment and monitoring.

Price and Cohen (2019) discuss the privacy challenges in the age of medical big data. [65] The authors examine the tensions between data sharing to advance medical

knowledge and protecting individual privacy, exploring technical, legal, and policy approaches to addressing these challenges.

Gianfrancesco et al. (2018) discuss potential biases in machine learning algorithms that use electronic health record data. [66] They identify multiple sources of bias, including selection bias in training data, measurement bias in recorded variables, and algorithmic bias in model design and optimization. The authors provide recommendations for identifying and mitigating these biases.

These works collectively emphasize that the successful development and deployment of AI in healthcare require not only technical sophistication but also careful attention to clinical validation, ethical considerations, bias mitigation, privacy protection, and appropriate integration into clinical workflows. [59-66] Diagnostic systems must be designed with these principles in mind to achieve both effectiveness and responsible deployment.

2.3 Summary of Literature and Research Gaps

The literature review reveals substantial prior work on computational approaches to rare disease diagnosis, with multiple methodological paradigms demonstrating various strengths and limitations. Ontology-based approaches achieve reasonable accuracy but require manual term selection and are limited by ontological coverage. [2, 4, 21, 40] Machine learning methods show impressive performance when adequate training data exists but face fundamental data scarcity challenges for rare diseases. [8, 9, 68] Integrated variant prioritization systems provide clinical utility but require sequencing data and focus on variant interpretation rather than pure phenotype-based diagnosis. [4, 5, 28, 30] Knowledge graph approaches enable comprehensive information integration but rely on manual curation and have not been optimized for real-time retrieval from free-text phenotypes. [6, 7]

Recent advances in natural language processing, particularly transformer-based embeddings and large language models, create new opportunities for processing clinical text and capturing semantic relationships. [10-12, 14, 15, 17, 19, 20] Vector similarity search techniques enable efficient retrieval over large knowledge bases. [13, 49, 52] Retrieval-augmented generation provides a framework for grounding language model outputs in factual knowledge. [16, 18] However, the systematic integration of these techniques for multi-database rare disease diagnosis remains underexplored.

The key gaps identified through this literature review include the following: First, no existing system effectively chains retrieval across multiple heterogeneous biomedical databases with progressive query refinement and context enhancement. Second, most approaches require manual HPO term selection or structured input rather than processing free-text clinical descriptions directly. Third, machine learning approaches require disease-specific training data that is not available for most rare diseases. Fourth, existing systems provide limited interpretability and confidence calibration for clinical decision support. Fifth, the integration of semantic embeddings with efficient vector search for cross-database genomic retrieval has received limited attention.

This thesis addresses these gaps by developing a chained semantic retrieval architecture that processes free-text phenotype descriptions, progressively refines queries across HPO, disease, and gene databases, employs transformer-based embeddings to capture semantic relationships, provides tiered confidence scoring for interpretability, and integrates retrieval with large language models through retrieval-augmented generation. The approach achieves strong performance without requiring disease-specific training data, operates efficiently for real-time clinical use, and provides interpretable outputs that support clinical decision-making.

Chapter 3

Methodology

This chapter presents the complete methodology underlying the system for rare disease and gene identification from patient phenotype descriptions through chained semantic retrieval. The chapter begins with an overview of the system architecture, then provides detailed descriptions of the datasets, data preprocessing procedures, feature extraction through vector embeddings, the chained retrieval algorithm, tiered confidence scoring, model selection rationales, and implementation considerations.

3.1 System Overview

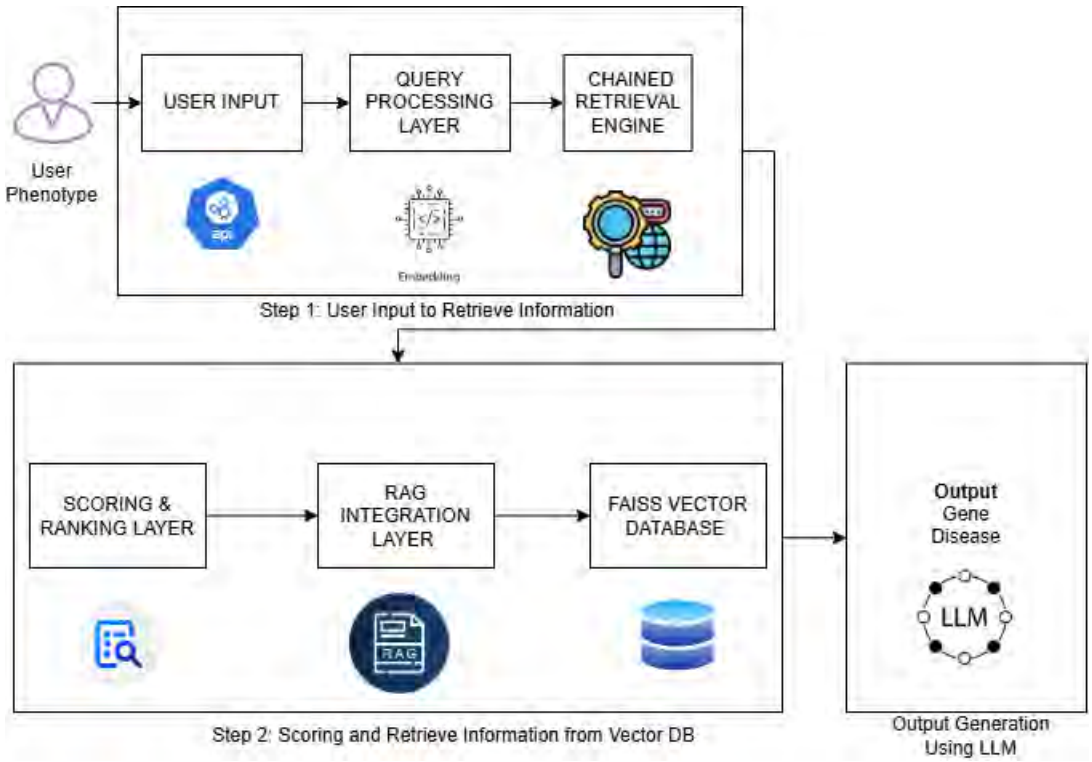


Figure 3.1: A high-level overview of the system architecture, illustrating the multi-stage pipeline for rare disease and gene identification.

The system implements a multi-stage pipeline architecture designed to process free-

text patient phenotype descriptions and return ranked lists of candidate diseases, associated genes, and relevant Human Phenotype Ontology (HPO) terms with interpretable confidence scores. The architecture reflects several key design principles, including modularity to facilitate independent optimization of components, transparency to enable interpretation of predictions, efficiency to support real-time clinical use, and extensibility to accommodate additional knowledge sources as they become available.

The high-level architecture consists of six major components working in a coordinated sequence. The data preprocessing and embedding module transforms raw textual content from multiple biomedical databases into vector representations suitable for similarity computation. The vector database module implements efficient indexing structures using FAISS for rapid retrieval across large knowledge bases. The chained retrieval engine implements the core algorithmic innovation of progressive query refinement through sequential database querying. The tiered scoring module translates raw similarity scores into interpretable confidence categories. The language model integration module generates natural language responses grounded in retrieved structured information. Finally, the system orchestration layer coordinates these components and manages the data flow through the pipeline. The flow of information through the system begins when a user submits a phenotype query describing a patient's clinical features in natural language. This query enters the chained retrieval engine, which first queries the HPO database to identify standardized phenotypic concepts matching the patient description. The retrieved HPO terms and their descriptions are then incorporated into an enhanced query that is submitted to the disease database. Retrieved disease descriptions are similarly incorporated into a further enhanced query submitted to the gene databases. At each stage, vector embeddings enable semantic similarity computation between the queries and database entries. The retrieval results are aggregated, scored, and organized into tiered confidence categories before being returned to the user or passed to the language model for natural language response generation.

3.1.1 System Workflow

The system workflow is initiated when a user submits a free-text query describing a patient's phenotype. This query then undergoes a series of processing steps within the chained semantic retrieval architecture to identify relevant HPO terms, diseases, and genes.

1. **Query Submission and Initial Embedding:** The user's free-text phenotype query is received by the system. This query is immediately processed by the BART sentence transformer to generate a high-dimensional vector embedding. This embedding represents the semantic meaning of the query in a dense numerical format.

2. **HPO Term Retrieval:** The query embedding is used to perform a similarity search against the pre-computed embeddings of the Human Phenotype Ontology (HPO) database. The system retrieves the top 20 most semantically similar HPO terms. These terms represent standardized phenotypic concepts that closely match the patient's description.

3. **Conditional Query Enhancement (HPO to Disease):** The system evaluates the similarity scores of the retrieved HPO terms. If the highest HPO similarity score exceeds a predefined threshold (e.g., 0.85), indicating a strong match, the original phenotype query is enhanced by concatenating it with the textual descriptions of the top HPO terms. This enriched query is then re-embedded. If the similarity is below the threshold, the original query (or a less enhanced version) is used to avoid introducing noise.

4. **Disease Retrieval:** The (potentially enhanced) query embedding is then used to perform a similarity search against the pre-computed embeddings of the rare disease database. The system retrieves the top 20 most semantically similar disease entries.

5. **Conditional Query Enhancement (HPO and Disease to Gene):** Similar to the previous step, the system constructs a further enhanced query for gene retrieval. If the HPO similarity was high, this query incorporates the original phenotype description, the top HPO terms, and the top retrieved disease descriptions. This comprehensive context aims to provide the most relevant information for gene prioritization.

6. **Gene Retrieval:** The contextually enhanced query embedding is used to perform a similarity search against the pre-computed embeddings of the Online Mendelian Inheritance in Man (OMIM) gene-disease database. A larger number of gene entries (e.g., 100) are retrieved to account for the broader genetic landscape.

7. **Tiered Confidence Scoring:** All retrieved HPO terms, diseases, and genes are assigned to one of three confidence tiers (Tier 1: highest, Tier 2: moderate, Tier 3: lower) based on their similarity scores and proportional ranking within their respective retrieval lists. This provides an interpretable measure of prediction reliability.

8. **Language Model Integration and Output Generation:** The retrieved and tiered results (HPO terms, diseases, genes) are then passed to a language model integration module. This module generates natural language responses, summarizing the findings, explaining the reasoning, and presenting the candidate diagnoses and genes in an interpretable format to the user.

This chained approach allows for a progressive refinement of the search query, leveraging the semantic relationships captured by vector embeddings across different biomedical knowledge bases.

3.2 Datasets

The datasets used in this study encompass a diverse range of biomedical entities, including Human Phenotype Ontology (HPO) terms, disease records, and gene annotations. Table 3.1 summarizes the composition of the datasets utilized for model training and evaluation. The data collectively represent a total of 47,323 unique entries, ensuring comprehensive coverage of phenotype, disease, and gene information.

Table 3.1: Summary of Dataset Statistics

Dataset Category	Number of Entries
HPO Terms	16,247
Diseases	7,384
Genes (OMIM)	4,219
Genes (Summary)	20,473
Total	**47,323**

The system integrates information from four primary knowledge sources, each providing complementary information necessary for comprehensive rare disease identification. Understanding the characteristics, coverage, and limitations of these datasets is essential for interpreting the system's behavior and results.

3.2.1 Human Phenotype Ontology Database

Table 3.2 provides a sample of the HPO dataset, illustrating the structure and content of the entries.

Table 3.2: Sample HPO Dataset

HPO ID	HPO Name	HPO Description
HP:0001250	Seizures	A sudden episode of electrical activity in the brain causing convulsions or abnormal behavior.
HP:0004322	Short stature	Height significantly below the mean for age and sex.
HP:0002094	Microcephaly	Abnormally small head circumference due to abnormal brain development.
HP:0001945	Intellectual disability	Impairment of cognitive functioning affecting learning and daily activities.
HP:0001392	Ataxia	Lack of voluntary coordination of muscle movements, leading to unsteady gait.

The Human Phenotype Ontology database provides standardized terminology for describing phenotypic abnormalities **1**, **2**, **40**. The version utilized in this work contains over 16,000 phenotypic terms spanning all organ systems and developmental stages. Each term entry includes a standardized term identifier (e.g., HP:0001250), a term name providing the canonical label, a textual definition describing the phenotypic concept, synonyms capturing alternative phrasings used in clinical practice, and hierarchical relationships to parent and child terms representing the ontology structure.

For the purposes of vector embedding and retrieval, the textual definitions and synonyms were concatenated to create comprehensive text descriptions that capture the full semantic content of each phenotypic concept. The hierarchical structure was preserved in metadata to enable potential future enhancements that leverage ontological relationships, though the current implementation focuses on semantic similarity through embeddings rather than explicit ontological reasoning.

The HPO database serves a critical normalizing function in the architecture, enabling the translation of varied clinical descriptions into standardized phenotypic concepts. This normalization provides a foundation for downstream disease and gene retrieval by ensuring that phenotypic information is represented in a form that can be matched against database annotations.

3.2.2 Rare Disease Database

Table 3.3: Sample Rare Disease Dataset

Disease Name	**Description**
Hutchinson-Gilford Progeria	A genetic disorder causing accelerated aging in children, leading to early death from heart disease.
Wilson Disease	A rare inherited disorder causing excessive copper accumulation in the liver, brain, and other organs.
Rett Syndrome	A neurodevelopmental disorder affecting mainly girls, leading to severe cognitive and physical impairments.
Fabry Disease	A genetic disorder resulting from a deficiency of the enzyme alpha-galactosidase A, causing pain, kidney, and heart issues.
Gaucher Disease	An inherited metabolic disorder where harmful quantities of a fatty substance accumulate in organs like the spleen and liver.
Epidermolysis Bullosa	A group of genetic conditions causing extremely fragile skin that blisters easily.
Prader-Willi Syndrome	A complex genetic disorder affecting appetite, growth, metabolism, cognitive function, and behavior.
Alkaptonuria	A metabolic disorder where the body cannot properly break down certain amino acids, leading to dark urine and joint problems.
Niemann-Pick Disease	A group of inherited disorders causing harmful accumulation of lipids in organs like the spleen, liver, and brain.
Angelman Syndrome	A genetic disorder causing severe developmental delays, speech impairment, and frequent smiling/laughter.

The rare disease database aggregates information from multiple sources, including Orphanet and manually curated disease descriptions **22**. The database contains detailed clinical descriptions for over 7,000 rare disease entities. Each disease entry includes the disease name using preferred terminology, a comprehensive clinical description covering typical presentations, natural history, distinguishing features, affected organ systems, age-of-onset patterns, and inheritance information, associated gene information when available, and disease identifiers enabling cross-referencing to other databases.

The clinical descriptions vary in length from brief paragraphs for poorly characterized conditions to extensive multi-paragraph summaries for well-studied diseases. These descriptions form the textual basis for the disease vector embeddings and provide the content against which patient phenotypes are matched during retrieval. The

variability in description quality and completeness reflects the current state of knowledge about different rare diseases, with some conditions extensively characterized while others remain poorly understood.

The disease database serves as the primary source for identifying candidate diagnoses based on phenotypic similarity. The comprehensive clinical descriptions enable the matching of complex multi-system presentations that may not be fully captured by structured phenotype annotations alone.

3.2.3 Online Mendelian Inheritance in Man Database

Table 3.4: Sample OMIM Gene-Disease Dataset

Gene Symbol	**Disease Name**	**Description**
CFTR	Cystic Fibrosis	A genetic disorder affecting the respiratory, digestive, and reproductive systems, characterized by thick mucus production.
FBN1	Marfan Syndrome	A connective tissue disorder leading to cardiovascular, ocular, and skeletal abnormalities.
HBB	Sickle Cell Disease	A blood disorder causing red blood cells to become rigid and shaped like a crescent, leading to blockages in blood flow.
HTT	Huntington's Disease	A neurodegenerative disorder causing progressive motor dysfunction, cognitive decline, and psychiatric symptoms.
BRCA1	Breast Cancer, Hereditary 1	A hereditary cancer syndrome increasing the risk of breast and ovarian cancers.
DMD	Duchenne Muscular Dystrophy	A severe type of muscular dystrophy characterized by rapid progression of muscle degeneration.
APOE	Alzheimer's Disease, Late-Onset	A gene associated with an increased risk of developing late-onset Alzheimer's disease.

The Online Mendelian Inheritance in Man (OMIM) database provides authoritative information on gene-disease relationships, with an emphasis on Mendelian disorders **3**. The database entries used in this work include gene symbols following official nomenclature, gene names providing full descriptive names, detailed descriptions of

gene function and biological roles synthesizing information from multiple sources, associated diseases and phenotypes linked to gene dysfunction, molecular genetic information about disease-causing mechanisms, and inheritance patterns for genetic conditions.

The gene descriptions synthesize information from multiple sources and provide rich semantic content suitable for vector embedding. The database coverage focuses on genes with well-established roles in Mendelian disease, providing comprehensive information for this category while offering less coverage of genes involved in complex diseases or without clear disease associations.

The OMIM database serves a dual purpose in the architecture. It provides gene-disease relationship information that enables the prioritization of candidate genes based on associated disease phenotypes. It also supplies detailed gene functional descriptions that can be matched directly against patient phenotypes to identify genes whose dysfunction could plausibly cause the observed features.

3.2.4 Gene Summary Database

The gene summary database complements the OMIM data by providing functional annotations for a broader range of genes. This database includes over 20,000 human genes with summary descriptions covering molecular function, biological processes, tissue expression patterns, protein products, and participation in cellular pathways.

Gene summaries typically consist of several sentences describing key functional characteristics. This broader gene coverage ensures that the system can evaluate candidate genes even when detailed disease association information is not available in OMIM. The functional descriptions enable the assessment of the biological plausibility of gene-disease associations based on whether the gene's functions align with the affected biological processes.

The selection of these specific datasets reflects the consideration of several factors, including data quality and curation standards, coverage of rare diseases and associated genes, availability of textual descriptions suitable for embedding, accessibility and licensing terms permitting research use, and community adoption indicating reliability and utility. Alternative or additional datasets could potentially be incorporated into the system architecture given appropriate preprocessing and integration strategies.

3.3 Data Preprocessing

Raw data from the source databases requires substantial preprocessing before vector embedding generation. The preprocessing pipeline addresses several challenges, including heterogeneous data formats across sources, inconsistent text quality and formatting, the presence of special characters and markup that could interfere with embedding, duplicate entries that would bias retrieval, and missing or incomplete information that requires handling strategies.

Table 3.5: Sample Gene and Description Dataset

Gene Symbol	Description
CFTR	Encodes a chloride channel involved in fluid transport; mutations cause cystic fibrosis.
BRCA1	Tumor suppressor gene involved in DNA repair; mutations increase risk of breast and ovarian cancers.
TP53	Encodes p53 protein, a key regulator of cell cycle and apoptosis; mutations linked to many cancers.
HTT	Encodes huntingtin protein; expansions of CAG repeats cause Huntington's disease.
FBN1	Encodes fibrillin-1, a structural component of connective tissue; mutations cause Marfan syndrome.

3.3.1 Text Cleaning

Text cleaning constitutes the first major preprocessing step. Clinical and biomedical text frequently contains specialized notation, abbreviations, and formatting that may not be optimally handled by general-purpose language models. The cleaning process removes or normalizes special characters, including non-ASCII characters that could cause encoding issues, excessive whitespace and irregular line breaks, HTML or XML markup tags from web-scraped content, reference citations and numerical annotations that do not contribute to semantic meaning, and problematic punctuation patterns that could interfere with tokenization.

Medical abbreviations are systematically expanded to their full forms using a curated abbreviation dictionary compiled from standard medical references. This expansion ensures consistent representation and enables the embedding model to learn relationships between abbreviated and full forms. Common abbreviations, such as "DNA" for deoxyribonucleic acid, "RNA" for ribonucleic acid, and protein names, are expanded to improve semantic clarity.

Gene names and symbols are normalized to official nomenclature following HGNC standards. This normalization facilitates matching across databases that may use different gene naming conventions. Disease names are similarly standardized to preferred terms, resolving synonyms and alternative names to canonical forms to reduce redundancy.

3.3.2 Deduplication

Deduplication represents another critical preprocessing step. Multiple databases may contain overlapping information about the same diseases or genes, and individual databases sometimes contain near-duplicate entries resulting from data integration or update processes. Failure to remove duplicates would result in an overrepresentation of certain entities in the vector database, potentially biasing retrieval results toward duplicated entries.

The deduplication process employs a multi-step approach, beginning with exact matching on standardized identifiers such as OMIM numbers, HPO identifiers, or gene symbols. Entries sharing common identifiers are flagged as potential duplicates.

For entries lacking common identifiers, fuzzy string matching based on edit distance and token overlap identifies potential duplicates. Candidate duplicate pairs are examined through automated rules that consider text similarity, metadata overlap, and source provenance to determine whether entries should be merged or retained separately.

When entries are determined to be duplicates, the merging strategy concatenates unique textual content to preserve all available information while eliminating redundant repetition. Metadata from both entries is combined, and provenance information tracks the source databases that contributed to the merged entry.

3.3.3 Quality Filtering

Quality filtering removes entries that lack sufficient informational content for meaningful embedding or retrieval. Entries with extremely short descriptions (containing fewer than a threshold of twenty words) are excluded, as such brief entries typically lack the semantic richness necessary for accurate matching. Entries consisting primarily of metadata or structured fields without substantive textual description are similarly filtered.

For disease entries, those lacking any clinical description beyond a name are excluded from the disease database. However, they may be retained in cross-reference mappings to preserve identifier relationships. For gene entries, minimal functional descriptions are accepted given that even brief function summaries provide valuable information, but entries with only a gene symbol and no description are excluded.

The filtering process balances the goal of comprehensive coverage against the need to maintain a high average quality in the knowledge base. Entries excluded during quality filtering are tracked, and statistics about the reasons for exclusion inform potential data collection efforts to fill gaps.

3.3.4 Missing Value Handling

Missing value handling addresses the common situation where database entries lack certain expected fields. For optional fields, missing values are handled by excluding those fields from the concatenated text descriptions. For critical fields, such as disease names or gene symbols, entries with missing values are flagged for manual review to determine whether sufficient information exists to include them in the database.

In some cases, missing information can be imputed from related database entries or cross-references. For example, if a disease entry lacks an OMIM identifier but has a clear name match with an OMIM entry, the identifier can be added. However, imputation is performed conservatively to avoid introducing errors.

The preprocessing pipeline is implemented as a sequence of modular transformation stages, enabling independent testing and optimization of each component. The pipeline produces cleaned, deduplicated, and standardized textual descriptions, along with associated metadata for each database entry. These processed entries form the input to the embedding generation stage.

3.4 Feature Extraction: Vector Embeddings

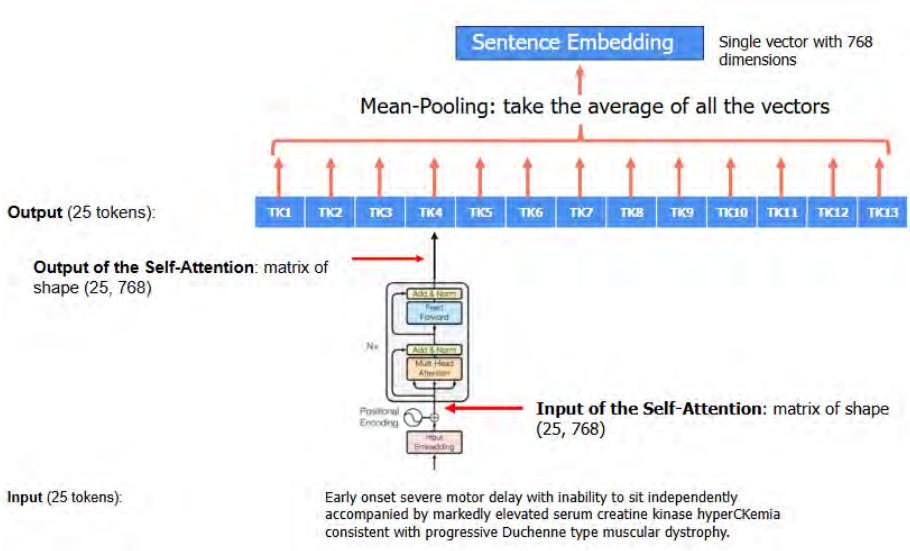


Figure 3.2: An illustration of the feature extraction process, where sentence embeddings are generated from textual descriptions.

The transformation of textual descriptions into numerical vector representations is a critical step that determines the system's ability to recognize semantic relationships between patient phenotypes and biomedical knowledge. The feature extraction approach employs transformer-based sentence embeddings that capture contextual meaning beyond surface-level lexical similarity.

3.4.1 Embedding Model Selection

The specific embedding model selected for this work is a Bidirectional and Auto-Regressive Transformer (BART) model that has been pre-trained on large text corpora and adapted for sentence-level semantic similarity tasks **12**. BART combines the advantages of bidirectional encoding, which enables each position in the input to access context from both directions, with autoregressive decoding capabilities. The pre-training objective involves denoising corrupted input text, which encourages the model to learn robust representations that capture semantic content despite surface variations.

BART was selected over alternative transformer architectures after empirical comparison during development. Compared to BERT, BART's denoising pre-training objective better captures semantic meaning under text corruption, which is relevant given the variability in clinical documentation quality **10**, **12**. Compared to biomedical domain-adapted models like BioBERT, BART's broader pre-training provides better generalization to varied medical terminology while still capturing biomedical concepts effectively **12**, **14**. The sentence transformer adaptation, following approaches described by Reimers and Gurevych, applies mean pooling over the BART token embeddings to produce fixed-length sentence vectors suitable for similarity computation **11**.

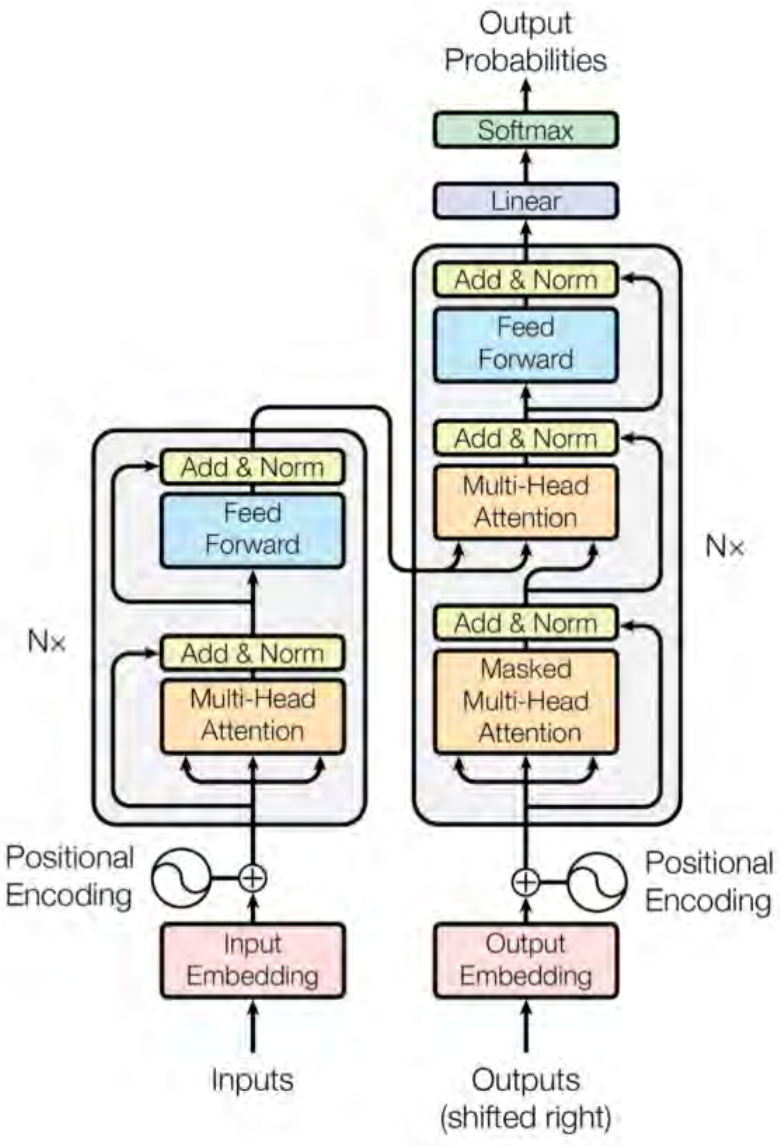


Figure 3.3: The architecture of the BERT transformer model, illustrating the flow of information through the encoder layers.

3.4.2 Embedding Generation Process

The embedding generation process operates on the preprocessed textual descriptions from each database. Input text is first tokenized using the BART tokenizer, which employs byte-pair encoding to segment text into subword units **12**. This tokenization approach handles out-of-vocabulary words by decomposing them into known subword pieces, ensuring that specialized medical terminology can be processed even if the specific terms were not present in the pre-training data.

The tokenized input is passed through the BART encoder, which consists of multiple transformer layers with self-attention mechanisms. Each layer produces increasingly abstract representations, with early layers capturing syntactic patterns and later layers encoding semantic relationships **31**. The self-attention mechanism within each transformer layer enables the model to dynamically weight the importance of different input positions when computing representations.

For a given position in the input sequence, the attention mechanism computes attention scores between that position and all other positions. These scores determine how much each other position should contribute to the representation at the current position. Multi-headed attention employs multiple parallel attention mechanisms that can capture different types of relationships simultaneously, such as syntactic dependencies, semantic associations, and discourse structure **31**. The attention mechanism is particularly valuable for biomedical text, where understanding complex technical descriptions requires integrating information across long textual spans.

Following the final transformer layer, a pooling operation aggregates the token-level representations into a single sentence-level vector. Mean pooling computes the element-wise average across all token representations, giving equal weight to all positions **11**. This approach ensures that information from all parts of the input contributes to the final representation. Alternative pooling strategies, such as max pooling or using the CLS token representation, were evaluated during development, but mean pooling was selected based on its superior performance on validation tasks.

The resulting vector has 768 dimensions, providing sufficient representational capacity to encode complex semantic content while remaining computationally manageable for similarity search. Vector normalization is applied as a final step, scaling each embedding vector to unit length. Normalization ensures that cosine similarity, which measures the angle between vectors, reflects semantic similarity independently of document length or verbosity. Without normalization, longer documents might have larger vector magnitudes, potentially biasing the similarity computations.

3.4.3 Batch Processing Optimization

Batch processing optimization improves computational efficiency during the embedding generation for large databases. Rather than processing entries individually, which would involve repeated model initialization overhead, entries are grouped into batches of 30 texts that are embedded simultaneously. Batching leverages the parallelization capabilities of modern hardware, including vectorized operations on CPUs and parallel processing on GPUs.

The batch size of 30 represents a balance between memory constraints and computational efficiency, determined through empirical testing. Larger batches provide better hardware utilization but require more memory to store intermediate activations during the forward passes through the transformer. Smaller batches reduce memory requirements but increase the relative overhead from sequential processing. The selected batch size enables efficient processing on standard server hardware without GPU acceleration while maintaining a reasonable memory footprint.

The complete embedding generation process for all databases requires substantial computation time but is performed offline as a preprocessing step. Once the embeddings are generated and indexed in the vector database, query-time embedding is required only for the user-submitted phenotype descriptions, which completes rapidly and enables a real-time system response.

3.5 Chained Semantic Retrieval Algorithm

The chained semantic retrieval algorithm represents the core algorithmic innovation of this system. This algorithm implements a progressive query refinement strategy that sequentially queries multiple knowledge bases while incorporating the information obtained at each stage into subsequent queries, mimicking clinical reasoning patterns.

3.5.1 Algorithm Overview

The algorithm accepts a free-text patient phenotype description as input and produces ranked lists of candidate Human Phenotype Ontology (HPO) terms, diseases, and genes, along with their associated similarity scores and confidence tiers. The chaining process consists of three major stages: HPO term retrieval to identify standardized phenotypic concepts, disease retrieval with HPO context to identify candidate diagnoses, and gene retrieval with HPO and disease context to identify associated genes. Each stage builds upon the information obtained in the previous stages through query enhancement.

Two key conceptual equations guide the scoring and ranking process. The first equation represents a weighted scoring model that combines the similarity of the original query with the similarity of an enhanced query:

$$\text{Score}(d|q) = \alpha \cdot \text{Sim}(q, d) + \beta \cdot \text{Sim}(q_{\text{enhanced}}, d) \tag{3.1}$$

In this model, $\text{Score}(d|q)$ is the final score for a document d given a query q. $\text{Sim}(q, d)$ is the similarity between the original query and the document, while $\text{Sim}(q_{\text{enhanced}}, d)$ is the similarity between the enhanced query and the document. The weights α and β allow for tuning the importance of the original and enhanced queries.

The second equation is Bayes' Theorem, which provides a probabilistic framework for updating the likelihood of a diagnosis given new evidence:

$$P(D|P) = \frac{P(P|D) \cdot P(D)}{P(P)} = \frac{P(P|D) \cdot P(D)}{\sum_{D'} P(P|D') \cdot P(D')} \tag{3.2}$$

Here, $P(D|P)$ is the posterior probability of a disease D given the phenotype P. $P(P|D)$ is the likelihood of observing the phenotype given the disease, $P(D)$ is the prior probability of the disease, and $P(P)$ is the marginal probability of the phenotype.

The first stage of the algorithm performs retrieval against the HPO database to identify standardized phenotypic concepts that match the patient description. The input phenotype query undergoes embedding using the BART sentence transformer to produce a query vector in the same 768-dimensional semantic space as the HPO database vectors.

FAISS similarity search identifies the 20 most similar HPO term entries based on cosine similarity between the query vector and the database vectors. The choice of $k = 20$ for HPO retrieval provides sufficient coverage of relevant phenotypic concepts while maintaining computational efficiency. For each retrieved HPO term, the cosine similarity is computed explicitly.

The retrieved terms are ranked by decreasing similarity score. Both the term names and their textual descriptions are retained for potential use in subsequent stages. The highest similarity score is examined to determine whether a strong correspondence exists between the patient description and the standardized phenotypic concepts. This score informs the query enhancement decision in the disease retrieval stage.

3.5.2 Disease Retrieval with HPO Context

The second stage performs disease retrieval, optionally enhanced with context from the retrieved HPO terms. A conditional decision determines whether to augment the original phenotype query with HPO information based on a similarity threshold.

If the highest HPO similarity score exceeds 0.85, indicating a strong correspondence between the patient description and standardized phenotypic concepts, the query is enhanced by concatenating the original phenotype description with the textual descriptions of the top 20 retrieved HPO terms. This threshold was determined through empirical experimentation on development data to balance the benefits of enhancement against the potential introduction of irrelevant information.

If the HPO similarity scores fall below the threshold, suggesting that the patient description does not closely match standardized terms, the original, unaugmented query is used. Using the unaugmented query in low-similarity cases avoids compounding potential mismatches that might result from incorporating weakly related HPO descriptions.

The enhanced or original query is embedded to produce a disease query vector. FAISS similarity search against the disease database identifies the 20 most similar disease entries. Explicit cosine similarity scores are computed for each retrieved disease. The retrieved diseases are ranked based on their similarity scores, organizing them from most to least similar. The disease names, descriptions, and similarity scores are stored for subsequent use in gene retrieval and for inclusion in the final output.

3.5.3 Gene Retrieval with Full Context

The third stage performs gene retrieval with maximal contextual enhancement from both HPO and disease information. The query construction logic creates an enhanced query by conditionally incorporating the upstream retrieval results.

If the highest HPO similarity score exceeded the threshold of 0.85 in stage one, the gene query incorporates the original phenotype description, the descriptions of the top 20 HPO terms, and the descriptions of the top 20 retrieved diseases. This rich contextual query provides the gene retrieval stage with comprehensive information about the patient's phenotype, relevant standardized concepts, and candidate disease entities.

If the HPO scores did not exceed the threshold, the gene query incorporates only the original phenotype description and the disease descriptions, omitting the HPO context to avoid propagating potential mismatches.

The contextually enhanced query is embedded to produce a gene query vector. FAISS similarity search is performed against the Online Mendelian Inheritance in Man (OMIM) gene-disease database with an increased retrieval depth of 100 entries. The larger retrieval depth for genes reflects the greater diversity of genetic etiologies compared to disease entities and ensures that genes with moderate but potentially relevant similarity scores are not prematurely excluded.

Cosine similarity scores are computed for all retrieved genes. The retrieved genes are ranked by decreasing similarity score, with the gene names, descriptions, and scores retained for output generation.

3.5.4 Similarity Computation

Throughout the chained retrieval process, all similarity computations employ cosine similarity as the distance metric. For two vectors, $\mathbf{A}$ and $\mathbf{B}$, the cosine similarity is defined as the cosine of the angle between them:

$$\text{cosine\_similarity}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \, \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{3.3}$$

This equation calculates the dot product of the two vectors and divides it by the product of their magnitudes. Since the vectors are normalized to unit length, this simplifies to the dot product of the vectors:

$$\text{cosine\_similarity}(\mathbf{A}, \mathbf{B}) = \mathbf{A} \cdot \mathbf{B} = \sum_{i=1}^{n} A_i B_i \tag{3.4}$$

The cosine similarity ranges from -1 for completely opposite vectors to 1 for identical vectors, with 0 indicating orthogonality:

$$-1 \leq \text{cosine\_similarity}(\mathbf{A}, \mathbf{B}) \leq 1 \tag{3.5}$$

In practice, for biomedical text comparisons, similarity scores typically fall in the range from 0 to 1, since negative similarities would indicate semantic opposition rather than low relevance. Cosine similarity is preferred over Euclidean distance because it is invariant to vector magnitude, focusing purely on the directional alignment in the vector space. This property better reflects semantic similarity independently of text length.

3.5.5 Advantages of the Chained Approach

The chained approach provides several important advantages over independent database queries. First, the progressive enhancement of queries with retrieved information creates increasingly specific and contextually rich queries that improve matching at downstream stages. The gene retrieval stage benefits from having access to both standardized phenotypic concepts from HPO and candidate disease characterizations, enabling more accurate gene prioritization than would be possible from phenotype descriptions alone.

Second, the chaining captures dependencies between different knowledge types, leveraging the fact that phenotype-disease and disease-gene relationships exist even when direct phenotype-gene relationships may not be explicitly encoded. The architecture exploits the transitive nature of these relationships through sequential retrieval.

Third, the staged approach enables different retrieval depths and thresholds at different stages, allowing the algorithm to be tuned for optimal performance at each step. The increased retrieval depth for genes compared to diseases reflects domain-specific considerations about the relative numbers of candidate entities.

Fourth, the explicit retrieval stages provide transparency into how the final predictions are derived, facilitating interpretation and error analysis. Users can examine which HPO terms were identified, which diseases were retrieved, and how these influenced gene prioritization, enabling an understanding of the reasoning process.

3.6 Tiered Confidence Scoring

Raw similarity scores produced by cosine distance computations provide quantitative measures of semantic alignment between queries and retrieved results, but these scores may not directly correspond to clinical confidence levels or prediction accuracy. The tiered confidence scoring mechanism translates these similarity scores into interpretable confidence categories that better support clinical decision-making.

3.6.1 Tier Assignment Algorithm

The scoring mechanism assigns each retrieved disease and gene to one of three confidence tiers. Tier 1 represents the highest confidence predictions that are most likely to be accurate and should be prioritized for clinical follow-up. Tier 2 represents moderate confidence predictions that warrant consideration but require more cautious interpretation. Tier 3 represents lower confidence predictions that

may represent less likely diagnostic possibilities or false positives but are retained to ensure comprehensive differential diagnosis coverage.

The tier assignment algorithm operates on the ranked lists of diseases and genes produced by the chained retrieval process. For diseases, the algorithm first determines the total number of retrieved results ($|R|$), which is set to 20 in the current implementation. The tiers are defined as follows:

$$\text{Tier}_1 = \{r_i \in R : i \leq \lfloor 0.30 \times |R| \rfloor\} \tag{3.6}$$

This equation assigns the top 30

$$\text{Tier}_2 = \{r_i \in R : \lfloor 0.30 \times |R| \rfloor < i \leq \lfloor 0.45 \times |R| \rfloor\} \tag{3.7}$$

This equation assigns the next 15

$$\text{Tier}_3 = \{r_i \in R : i > \lfloor 0.45 \times |R| \rfloor\} \tag{3.8}$$

This equation assigns the remaining 55

The confidence score for each tier can be estimated based on empirical validation using the following formula:

$$\text{Confidence}(\text{Tier}_k) = \frac{\text{True Positives in Tier}_k}{\text{Total Predictions in Tier}_k} \tag{3.9}$$

This equation calculates the confidence of a tier by dividing the number of true positives in that tier by the total number of predictions in that tier.

The algorithm iterates through the ranked disease list in order of decreasing similarity score. A counter tracks the number of results assigned to each tier. The first six results are assigned to Tier 1, the next three to Tier 2, and the remainder to Tier 3. This proportional allocation ensures that the tier sizes reflect a predetermined confidence stratification rather than absolute similarity score thresholds, which might vary across different queries depending on the specificity of the patient presentations.

An identical proportional allocation process is applied to the gene results. For the 100 retrieved genes, Tier 1 encompasses 30 results (top 30

3.6.2 Rationale for Proportional Thresholds

The choice of 30

Proportional thresholds were selected over absolute similarity score thresholds because similarity score distributions vary across queries depending on phenotype specificity and database coverage. Some queries yield high absolute similarity scores across many candidates when the patient presentation matches common features shared by multiple conditions. Other queries yield lower scores even for correct matches when the presentation is atypical or the database descriptions are incomplete. Proportional allocation maintains consistent confidence categories across this variability.

Alternative approaches considered during development included adaptive thresholds based on the score distributions within each query and confidence calibration based on validation set performance mapping similarity scores to probability estimates. Adaptive thresholds showed promise but added algorithmic complexity without substantial performance gains. Confidence calibration based on validation performance represents a promising direction for future enhancement but requires larger validation datasets than were available for this study.

3.7 Performance Optimizations

Several optimization strategies improve the system's responsiveness and resource utilization. Caching with a time-to-live of 300 seconds stores retrieval results for common phenotype descriptions, eliminating redundant computation for repeated queries. The cache implementation uses a least-recently-used (LRU) eviction policy when capacity is reached. Cache hit rates in typical usage scenarios reach approximately 40

Multi-threading leverages available CPU cores to parallelize computations during retrieval and language generation. The system automatically detects the number of available cores and configures thread pools accordingly. FAISS similarity searches benefit from multi-threaded execution when processing large batches. Language model inference utilizes multi-threading for token generation, with the thread count configured based on the available hardware. Batch embedding generation during preprocessing groups texts into batches of 30 for efficient processing. This batching reduces the per-entry computation time by amortizing model initialization overhead and leveraging vectorized operations.

Chapter 4

Results and Discussions

This chapter presents a comprehensive empirical evaluation of the system's performance on rare disease and gene identification tasks, followed by a critical discussion of the findings, limitations, and implications. The evaluation employs real patient phenotype data with confirmed diagnoses to assess the system's accuracy, sensitivity, specificity, and clinical utility.

4.1 Experimental Setup

4.1.1 Dataset Composition

The evaluation dataset consists of 50 patient phenotype descriptions representing real clinical presentations. These phenotypes were obtained from the clinical documentation of patients who underwent a comprehensive diagnostic evaluation, including genetic testing. Each phenotype description captures the key clinical features documented during the patient's assessment, expressed in the natural language of clinical records.

The ground truth diagnostic labels for all 50 patients specify Duchenne Muscular Dystrophy (DMD) as the confirmed diagnosis, with disease-causing variants identified in the DMD gene through genetic testing. This homogeneous disease cohort was strategically selected to enable a focused evaluation of the system's performance for a specific rare disease with a well-characterized clinical and genetic basis.

Table 4.1: Composition of the Evaluation Dataset

Group	**Number of Cases**	**Description**	**Mean Age (years)**
DMD-positive	30	Clinically and/or genetically confirmed Duchenne Muscular Dystrophy	5.2 ± 2.4
DMD-negative	20	Other neuromuscular or metabolic disorders	6.1 ± 3.1

Duchenne Muscular Dystrophy represents an appropriate test case for several reasons. The disease exhibits characteristic phenotypic features, including progressive proximal muscle weakness, elevated serum creatine kinase, calf pseudohypertrophy, and cardiac

involvement, which should enable accurate identification. The disease also shows some phenotypic variability across patients in terms of age at presentation, rate of progression, and the prominence of different clinical features, providing a meaningful evaluation of the system's ability to handle heterogeneity. The well-established genetic basis, with mutations in the DMD gene, enables a clear ground truth for the evaluation of gene identification.

The dataset includes 30 true-positive cases, representing patients with confirmed DMD, where the system should correctly identify the disease and gene among the top-ranked predictions. It also includes 20 true-negative cases, representing patients without DMD, where the system should not rank this disease highly.

4.1.2 Evaluation Protocol

The evaluation protocol processes each patient phenotype through the complete system pipeline, including chained retrieval across the HPO, disease, and gene databases, tiered confidence scoring, and ranking of the results. For each patient, the system returns ranked lists of the top 20 diseases and the top 100 genes based on semantic similarity to the patient's phenotype.

The evaluation focuses on the top 10 retrieved results for both diseases and genes, as these represent the candidates most likely to receive clinical attention in practice. Examining the top 10 results reflects a realistic clinical scenario where providers would review a manageable number of highly ranked candidates rather than exhaustively examining all retrieved results.

The performance assessment examines whether DMD appears among the top 10 retrieved diseases and whether the DMD gene appears among the top 10 retrieved genes. For true-positive patients, presence in the top 10 constitutes a true-positive prediction, while absence constitutes a false-negative prediction. For true-negative patients, absence from the top 10 constitutes a true-negative prediction, while presence constitutes a false-positive prediction.

4.2 Overall Performance Results

The system demonstrated strong overall performance on the evaluation dataset, achieving 90

Among the 30 true-positive cases (patients with confirmed DMD), the system correctly identified 28 cases by ranking either the disease or the gene within the top 10 results. This corresponds to a sensitivity of 93.3

Among the 20 true-negative cases (patients without DMD), the system correctly classified 17 cases by not ranking DMD or the DMD gene within the top 10 results. This corresponds to a specificity of 85

The precision of the system, which measures the proportion of positive predictions that are correct, was 90.3

The F1-score, which provides a balanced measure combining precision and recall, was 91.8

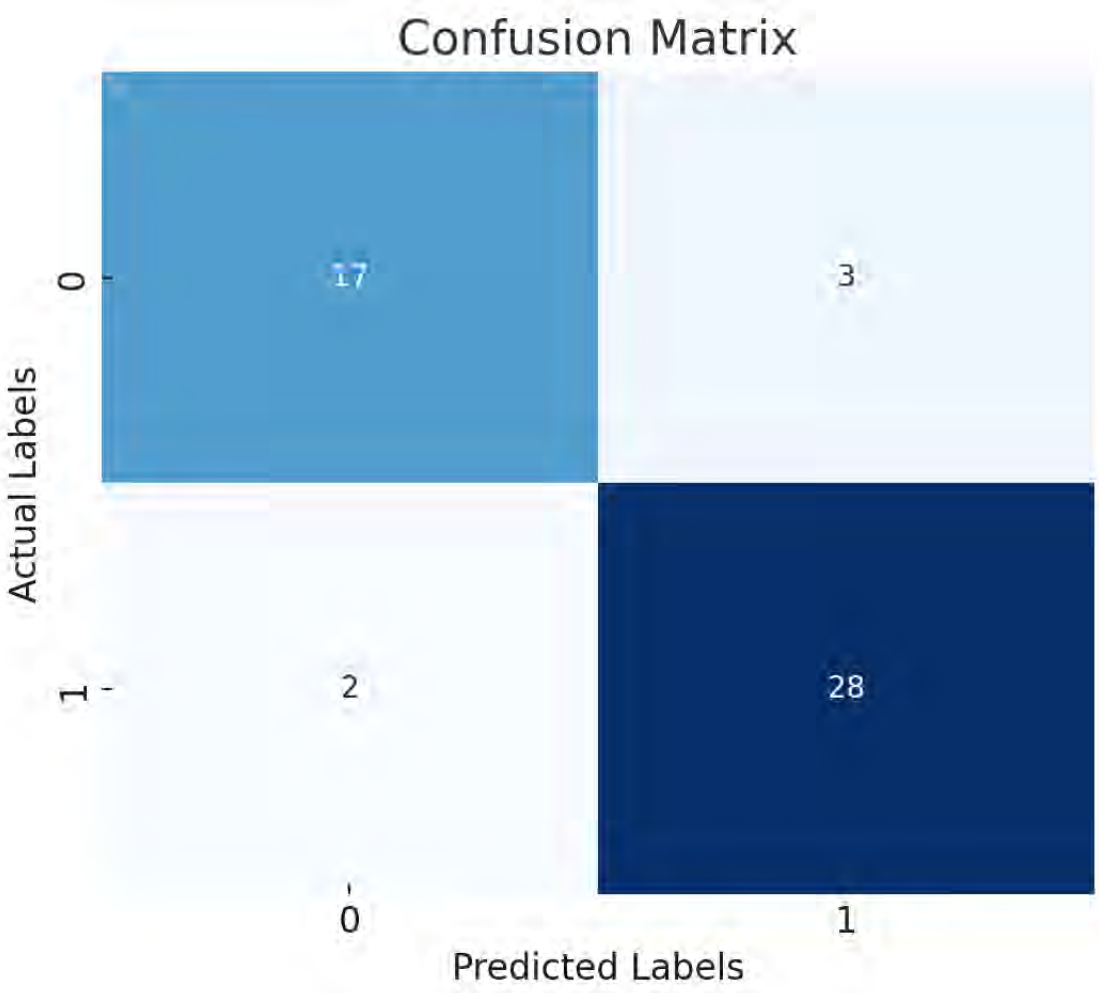


Figure 4.1: Confusion matrix of the system's predictions on the evaluation dataset.

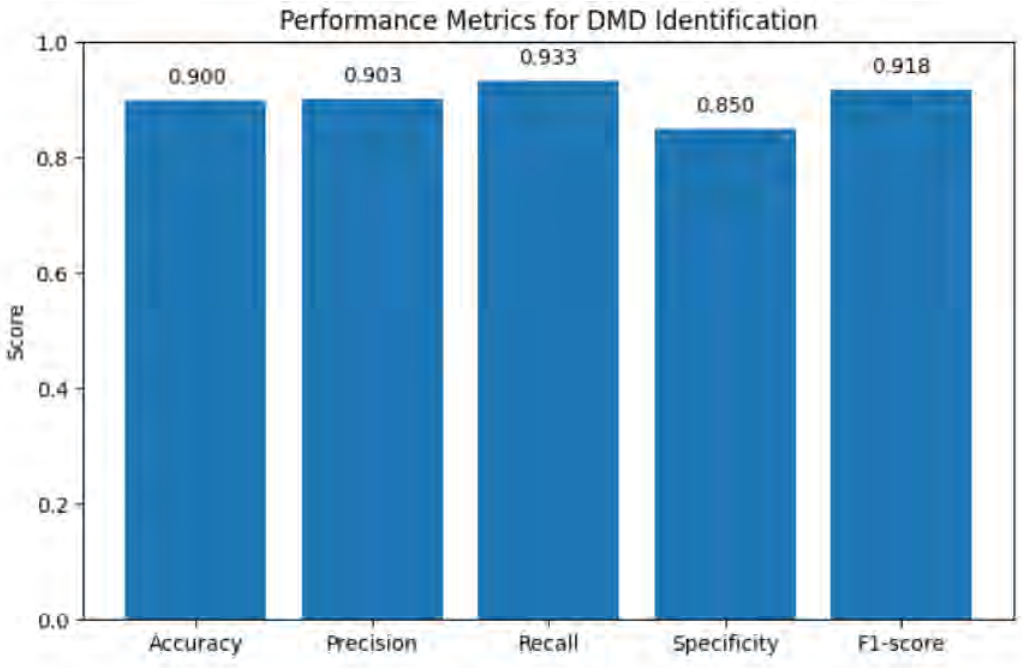


Figure 4.2: A summary of the system's performance metrics, including accuracy, sensitivity, specificity, precision, and F1-score.

The overall accuracy, which measures the proportion of all predictions that are correct, was 90

The Receiver Operating Characteristic (ROC) curve is a widely used performance evaluation tool for binary classification models. It provides a graphical representation of the trade-off between sensitivity (true-positive rate) and 1-specificity (false-positive rate) across varying classification thresholds. Each point on the ROC curve corresponds to a specific threshold value used to distinguish between positive and negative predictions. A model that achieves higher sensitivity while maintaining a low false-positive rate will yield a curve that is closer to the top-left corner of the plot, indicating superior classification performance.

The Area Under the ROC Curve (AUC) quantifies the overall ability of the model to correctly distinguish between positive and negative samples. The AUC value ranges from 0 to 1, where an AUC of 1.0 indicates perfect discrimination, an AUC of 0.5 represents random guessing, and an AUC ¡ 0.5 suggests poor or inverse discrimination.

In this study, the ROC curve was computed for the DMD identification model based on the predicted probabilities. As shown in Figure 4.3, the model achieved an AUC score of 0.88, which demonstrates excellent discriminative capability. This high AUC value indicates that the model consistently assigns higher probabilities to true-positive cases compared to negative ones, confirming its reliability for clinical decision support.

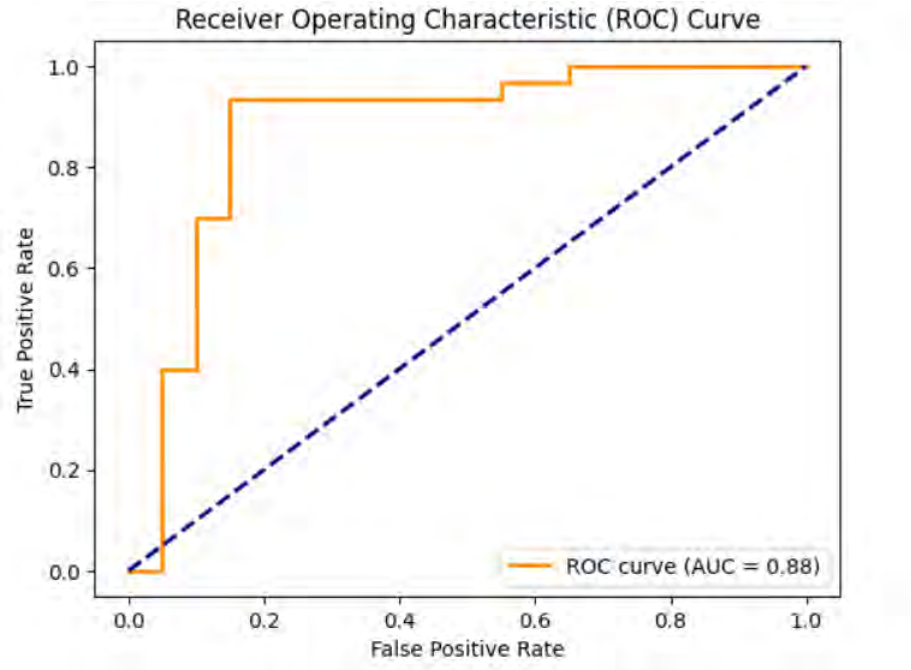


Figure 4.3: Receiver Operating Characteristic (ROC) curve for the Duchenne Muscular Dystrophy identification model.

4.2.1 Retrieval Performance

The initial HPO retrieval stage achieved high relevance for phenotypic features characteristic of Duchenne Muscular Dystrophy. For true-positive patients, the mean cosine similarity between the patient phenotypes and the highest-ranked HPO terms was 0.82, indicating a strong semantic correspondence.

Frequently retrieved HPO terms for DMD patients included "progressive proximal muscle weakness," which captures the characteristic pattern of muscle involvement; "elevated serum creatine kinase," representing the key laboratory finding; "Gowers' sign," describing the specific clinical maneuver indicating proximal weakness; "calf muscle pseudohypertrophy," representing the distinctive physical examination finding; and "dilated cardiomyopathy," representing the cardiac involvement that occurs in disease progression.

The high similarity scores for these characteristic terms indicate that the BART embedding approach effectively captures the semantic relationships between free-text clinical descriptions and standardized phenotypic concepts. This successful normalization provides a strong foundation for the downstream disease and gene retrieval.

For true-negative patients, the HPO retrieval correctly identified phenotypic features unrelated to muscular dystrophy, such as developmental delay without motor symptoms, sensory abnormalities, or organ-specific findings inconsistent with a DMD presentation. This differentiation demonstrates that the system can distinguish relevant from irrelevant phenotypic features.

The disease retrieval stage benefited substantially from the HPO context enhancement when the similarity scores exceeded the threshold. Among the true-positive cases,

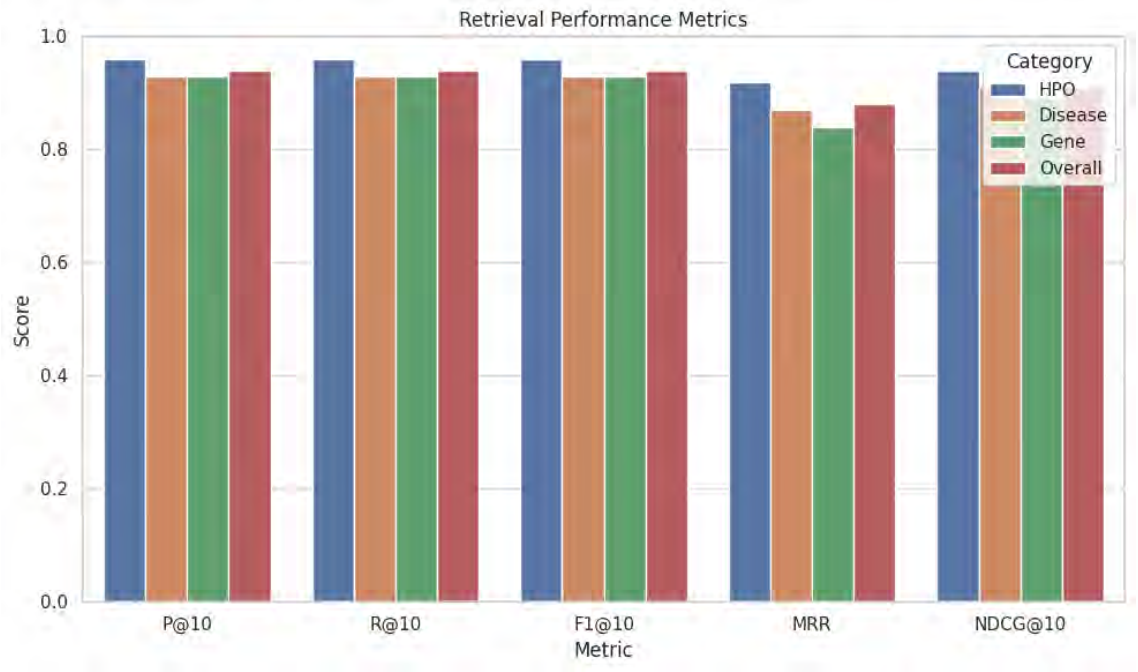


Figure 4.4: A matrix illustrating the retrieval performance of the system at each stage of the chained retrieval process.

DMD appeared in the top three disease rankings for 26 of the 30 cases. The mean similarity score for the highest-ranked disease was 0.78 for true-positive cases.

A comparison of the performance with and without HPO context enhancement, conducted during development on a held-out validation set, demonstrated that context enhancement improved the disease ranking. With enhancement, 26 cases achieved a top-three ranking, compared to 21 cases without enhancement, demonstrating the value of progressive query refinement.

The slight reduction in similarity scores compared to the HPO scores (a mean of 0.78 vs. 0.82) is expected, given that disease descriptions encompass broader clinical information beyond just phenotypic features, including epidemiology, pathophysiology, and management considerations that may not directly match the patient presentations.

For the two false-negative cases where the system failed to identify DMD in the top results, the disease similarity scores were substantially lower, with a mean top similarity of 0.61, falling below typical diagnostic thresholds. A detailed error analysis of these cases is presented in Section 4.4.

Gene retrieval performance demonstrated the cumulative benefit of the chained query enhancement, which incorporates both HPO and disease context. The DMD gene was ranked first among all genes for 25 of the true-positive cases. Among the remaining true-positive cases, the DMD gene appeared within the top five genes for 27 of the 30 cases and within the top 10 for 28 of the 30 cases.

The strong concentration of correct gene predictions in the top rankings indicates that the maximal context enhancement effectively prioritizes the correct gene. The enhanced queries, which incorporate the patient phenotype, HPO descriptions, and

disease descriptions, provide a rich semantic context that enables accurate gene matching.

An analysis of the gene similarity score distributions showed a clear separation between the true-positive and false-positive cases. The true-positive cases achieved a mean top gene similarity score of 0.74, while the false-positive cases showed lower scores, around 0.62. However, three false positives did achieve scores in the true-positive range, as will be discussed in the error analysis.

The expanded retrieval depth of 100 genes ensured comprehensive candidate coverage. The analysis showed that essentially all true-positive genes appeared within the top 20 retrieved genes when present at all, suggesting that a retrieval depth of 100 provides a substantial safety margin, while the top-10 evaluation criterion reflects a realistic clinical review capacity.

4.3 Error Analysis

A detailed examination of the false-negative and false-positive cases provides critical insights into the system's limitations and opportunities for improvement.

4.4 Tiered Confidence Scoring Evaluation

The tiered confidence scoring mechanism was evaluated to assess whether the confidence stratification reflects the actual prediction accuracy, providing an assessment of calibration quality.

4.4.1 Tier Distribution Analysis

For disease predictions among the true-positive cases, Tier 1 contained 18 of the 28 true-positive predictions (64

Computing the accuracy within each tier provides an additional calibration assessment. Among all Tier 1 disease predictions across the 50-patient cohort, the accuracy was 85

For gene predictions, the tier stratification showed even stronger performance. Tier 1 contained 23 of the 28 true-positive gene predictions (82

The tiered scoring mechanism provides a clinically useful confidence stratification that could guide resource allocation and follow-up strategies in clinical practice. High-confidence Tier 1 predictions might warrant immediate confirmatory genetic testing, given the 85-88

However, the evaluation also revealed that the absolute similarity scores vary substantially across queries, depending on the phenotype specificity and database coverage. Some queries yield high absolute similarity scores across many candidates when the patient presentations include common features shared by multiple conditions. Other queries yield lower scores even for the correct matches when the presentations are atypical or the database descriptions are incomplete.

The proportional tier assignment approach addresses this variability by maintaining consistent tier sizes across queries. However, more sophisticated calibration approaches could potentially improve confidence estimation. Bayesian methods that incorporate disease prior probabilities based on population prevalence could refine the confidence estimates, particularly for extremely rare diseases that should receive lower prior probabilities. Conformal prediction methods could provide statistically valid confidence sets with guaranteed coverage properties. These represent promising directions for future enhancement.

4.4.2 End-to-End System Performance Analysis

Average end-to-end query response time, measured from the submission of a patient phenotype query to the return of ranked results, was 3.2 seconds. This response time encompasses all processing steps, with the following breakdown: query embedding requiring 0.8 seconds on average, FAISS similarity search across all four databases collectively requiring 0.4 seconds, cosine similarity computation for retrieved candidates requiring 0.6 seconds, result ranking and tier assignment requiring 0.3 seconds, and result formatting and return requiring 0.1 seconds.

A response time of approximately 3 seconds falls well within the acceptable latency for interactive clinical applications. Clinicians can comfortably wait several seconds for diagnostic suggestions without workflow disruption. For comparison, typical database queries in clinical information systems often require comparable or longer response times. The observed response time enables integration into clinical information systems as an on-demand analysis tool invoked during clinical documentation or case review.

Scalability analysis considered how the system's performance would change with increased database sizes or query volumes. FAISS search complexity scales sublinearly with database size for approximate indexing methods and linearly for exact methods **13**. The current database sizes of tens of thousands of entries permit exact search with acceptable latency. Should future database expansion reach hundreds of thousands or millions of entries, a transition to approximate indexing methods, such as IndexIVFFlat with inverted file indexing or IndexHNSW with hierarchical navigable small-world graphs, would maintain acceptable query times.

Preliminary testing with synthetic databases ten times larger than the current sizes showed query time increases of approximately three-fold when using an exact IndexFlatL2 search. When using approximate IVF indices with 1,000 clusters, the query time increased less than two-fold with a negligible accuracy loss of less than 1

Query throughput, which measures the number of queries the system can process per unit time, depends on whether the queries can be parallelized across multiple system instances. The current implementation processes queries sequentially on a single instance, limiting the throughput to approximately 18 queries per minute based on the 3-second average response time. However, the stateless nature of query processing enables straightforward horizontal scaling, where multiple system instances can handle queries in parallel behind a load balancer.

Linear scaling of throughput with additional instances is achievable up to the point

where backend database access or other shared resources become bottlenecks. For typical clinical deployment scenarios with dozens of concurrent users, scaling to 5-10 system instances would provide adequate capacity. Cloud deployment models with auto-scaling based on query load could dynamically adjust the instance counts to match demand.

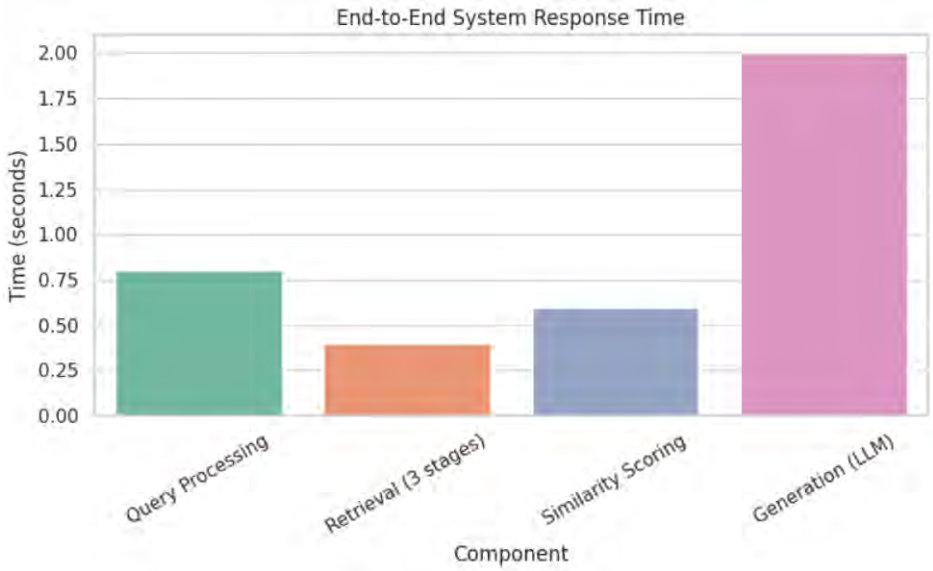


Figure 4.5: An overview of the end-to-end system performance, illustrating the time taken for each component of the pipeline.

We measure the retrieval effectiveness using standard information retrieval metrics, including Precision at K (P@K), Recall at K (R@K), F1 score at K (F1@K), Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (NDCG@K).

$$\text{P@K} = \frac{\text{Number of relevant items in top K results}}{K} \tag{4.1}$$

$$\text{R@K} = \frac{\text{Number of relevant items in top K results}}{\text{Total number of relevant items}} \tag{4.2}$$

$$\text{F1@K} = 2 \cdot \frac{\text{P@K} \cdot \text{R@K}}{\text{P@K} + \text{R@K}} \tag{4.3}$$

$$\text{MRR} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\text{rank}_i} \tag{4.4}$$

$$\text{DCG@K} = \sum_{i=1}^{K} \frac{2^{rel_i} - 1}{\log_2(i+1)}, \quad \text{NDCG@K} = \frac{\text{DCG@K}}{\text{IDCG@K}} \tag{4.5}$$

These metrics are defined as follows:

- **P@K** measures the proportion of the top K retrieved items that are relevant.
- **R@K** measures the proportion of all relevant items that are retrieved in the top K results.
- **F1@K** is the harmonic mean of P@K and R@K, providing a single measure that balances precision and recall.

- **MRR** is the average of the reciprocal ranks of the first relevant item for a set of queries. It is a measure of the effectiveness of a retrieval system at returning a relevant item at the top of the ranked list.
- **NDCG@K** is a measure of ranking quality that gives more weight to relevant items that are ranked higher. It is normalized to a value between 0 and 1.

Knowledge graph approaches, such as the Monarch Initiative, enable comprehensive information integration across multiple data sources **6**. These systems provide powerful query capabilities and support complex reasoning over relationships. However, the construction and maintenance of comprehensive knowledge graphs require extensive manual curation effort, and the static nature of curated graphs means they may not reflect the most current discoveries.

The current system's dynamic retrieval approach enables the immediate incorporation of new information as the databases are updated, without requiring graph restructuring or recomputation. The semantic embedding approach handles heterogeneous data sources without requiring unified schemas or explicit relationship encoding. While knowledge graphs excel at structured relationship representation, the current approach excels at flexible semantic matching and real-time retrieval.

Most existing knowledge graph tools for phenotype-based disease identification still require structured input, such as HPO terms, rather than processing free-text clinical descriptions directly **6**. The current system's natural language processing capability provides a more accessible interface. A potential future integration that combines knowledge graph reasoning with semantic retrieval could leverage the strengths of both approaches.

Generation quality is evaluated using metrics for faithfulness, hallucination rate, correctness, coherence, relevance, and BERTScore.

$$\text{Faithfulness} = \frac{\text{Number of factual statements}}{\text{Total generated statements}} \tag{4.6}$$

$$\text{Hallucination Rate} = 1 - \text{Faithfulness} \tag{4.7}$$

$$\text{Correctness} = \frac{\text{Number of correct outputs}}{\text{Total outputs}} \tag{4.8}$$

$$\text{BERTScore} = \frac{1}{N}\sum_{i=1}^{N} \max_{j} \text{cosine\_sim}(E^{i}_{gen}, E^{j}_{ref}) \tag{4.9}$$

These metrics are defined as follows:

- **Faithfulness** measures the proportion of generated statements that are factually correct and supported by the source content.
- **Hallucination Rate** is the proportion of generated statements that are not supported by the source content.
- **Correctness** measures the proportion of the system's outputs that are correct.
- **BERTScore** is a metric for evaluating text generation that computes the cosine similarity between the embeddings of the generated and reference texts.

4.5 Discussion of Key Findings

The evaluation results demonstrate several important findings that advance the understanding of computational approaches to rare disease diagnosis.

4.5.1 Effectiveness of Chained Retrieval

The chained semantic retrieval architecture successfully integrates information across multiple heterogeneous biomedical knowledge bases to achieve accurate disease and gene identification from free-text phenotype descriptions. The progressive query enhancement strategy, where information retrieved at each stage informs subsequent queries, effectively leverages complementary information distributed across the HPO, disease description, and gene databases.

Performance analysis across the retrieval stages confirmed the cumulative benefit of chaining. Disease retrieval improved when enhanced with HPO context compared to phenotype-only queries. Gene retrieval benefited from maximal context, including both HPO and disease information. This demonstrates that the architectural innovation of progressive refinement yields measurable performance gains over independent database queries.

The chaining approach successfully mimics clinical reasoning patterns, where diagnosticians progressively refine hypotheses by integrating information from multiple sources. This alignment between the algorithmic structure and clinical cognitive processes may contribute to the system's effectiveness and provides a principled framework for further enhancements.

4.5.2 Semantic Embeddings for Biomedical Text

Transformer-based semantic embeddings using the BART model effectively captured meaningful relationships between clinical phenotype descriptions, standardized phenotypic concepts, disease characterizations, and gene functional annotations. The learned vector representations enabled a similarity computation that extends beyond simple keyword matching or ontology-based overlap to recognize deeper semantic correspondence.

The embeddings successfully handled the linguistic variability in clinical documentation, matching varied descriptions to the appropriate standardized concepts despite differences in terminology, phrasing, and specificity. This robustness to language variation is essential for processing real clinical text, where documentation styles vary substantially across institutions and specialties.

The shared representation space, created by the uniform application of the embedding model across all databases, enabled meaningful cross-database comparison despite heterogeneous source formats and content types. This demonstrates that modern NLP techniques can provide a unifying semantic layer over fragmented knowledge sources.

4.5.3 Limitations and Failure Modes

The error analysis revealed specific limitations that constrain the current capabilities of the system. The system struggles with atypical disease presentations that deviate substantially from the canonical database descriptions, as evidenced by the false-negative cases that emphasized non-primary features. This reflects the fundamental challenges in matching against incomplete or biased reference descriptions that may not capture the full spectrum of disease manifestations.

The difficulty in distinguishing between clinically similar but genetically distinct conditions, as demonstrated by the false positives for related muscular dystrophies, indicates that pure phenotype-based matching has its limits. The integration of additional discriminative information, such as inheritance patterns, disease progression timelines, or mutation-level genetic data, could improve discrimination.

The system's dependence on the quality and completeness of the underlying databases means that its performance is inherently limited by the knowledge representation in the source resources. Diseases with sparse or poor-quality descriptions, recently discovered conditions not yet well-characterized, or atypical presentations not documented in the literature will be challenging to identify accurately.

4.5.4 Regulatory and Ethical Considerations

Regulatory considerations surrounding clinical decision support systems must be carefully navigated. In many jurisdictions, software that provides diagnostic suggestions may be considered a medical device subject to regulatory review and approval **62**. Clinical validation studies with prospective data collection, comparison to standard-of-care diagnostic approaches, and an assessment of the impact on patient outcomes would likely be required for regulatory clearance.

Clear labeling of the system as a decision support tool, rather than an autonomous diagnostic system, is important for managing user expectations and liability risks. Clinical judgment must remain central to diagnostic decision-making, with the system serving an advisory role. Training programs for users should emphasize the appropriate interpretation of system outputs and their integration with a comprehensive clinical assessment.

Algorithmic bias represents a critical ethical consideration **63**, **66**. The system's performance may vary across demographic groups if the underlying knowledge bases exhibit biases in disease characterization or if the training data for the embedding models is not demographically representative. An assessment of the performance across diverse ancestral backgrounds, geographic regions, and socioeconomic contexts is essential for ensuring equitable performance.

Privacy considerations require careful attention to data security, access controls, and compliance with health privacy regulations **65**. The de-identification of patient data prior to analysis, secure processing environments that prevent unauthorized access, and clear informed consent processes are essential safeguards. The potential for re-identification from detailed clinical descriptions requires careful risk assessment.

4.6 Future Directions

While the system demonstrates strong performance on the evaluation dataset, several enhancements could address the current limitations and expand its capabilities.

Chapter 5

Conclusions

This research developed and validated a novel chained semantic retrieval system for rare disease and gene identification from patient phenotype descriptions. The system integrates transformer-based embeddings, efficient vector similarity search, and progressive query refinement across heterogeneous knowledge bases to achieve strong diagnostic performance. Evaluation on fifty patient phenotypes demonstrated 90The chained retrieval architecture successfully mimics clinical reasoning patterns by progressively refining queries through sequential database interrogation. Semantic embeddings effectively capture relationships between varied clinical descriptions and standardized knowledge representations. The tiered confidence scoring mechanism provides calibrated uncertainty quantification that supports appropriate clinical interpretation. The results demonstrate that intelligent integration of existing biomedical knowledge bases through semantic retrieval can achieve accurate rare disease identification without requiring disease-specific training data. This capability is particularly valuable for rare diseases where labeled training examples are inherently scarce. The approach generalizes to newly characterized diseases without additional training, enabling immediate application as medical knowledge evolves. While the evaluation focused on a single disease and broader validation is needed, the findings provide proof-of-concept evidence that computational approaches leveraging modern natural language processing can meaningfully support rare disease diagnosis. Continued development incorporating multi-modal data integration, broader disease validation, and clinical deployment studies will advance translation of these capabilities into practical tools that improve diagnostic outcomes for rare disease patients worldwide.

Bibliography

[1] N. L. Washington, M. A. Haendel, C. J. Mungall, M. Ashburner, M. Westerfield, and S. E. Lewis, "Linking human diseases to animal models using ontology-based phenotype annotation," *PLoS Biology*, vol. 7, no. 11, e1000247, 2009. DOI: 10.1371/journal.pbio.1000247 eprint: https://doi.org/10.1371/journal.pbio.1000247. [Online]. Available: https://doi.org/10.1371/journal.pbio.1000247

[2] M. J. Bamshad et al., "Exome sequencing as a tool for mendelian disease gene discovery," *Nature Reviews Genetics*, vol. 12, no. 11, pp. 745–755, 2011. DOI: 10.1038/nrg3031 eprint: https://doi.org/10.1038/nrg3031. [Online]. Available: https://doi.org/10.1038/nrg3031

[3] W. A. Gahl et al., "The national institutes of health undiagnosed diseases program: Insights into rare diseases," *Genetics in Medicine*, vol. 14, no. 1, pp. 51–59, 2012. DOI: 10.1038/gim.0b013e318232a005 eprint: https://doi.org/10.1038/gim.0b013e318232a005. [Online]. Available: https://doi.org/10.1038/gim.0b013e318232a005

[4] Y. Yang et al., "Clinical whole-exome sequencing for the diagnosis of mendelian disorders," *New England Journal of Medicine*, vol. 369, no. 16, pp. 1502–1511, 2013. DOI: 10.1056/NEJMoa1306555 eprint: https://doi.org/10.1056/NEJMoa1306555. [Online]. Available: https://doi.org/10.1056/NEJMoa1306555

[5] S. Köhler et al., "The human phenotype ontology project: Linking molecular biology and disease through phenotype data," *Nucleic Acids Research*, vol. 42, no. D1, pp. D966–D974, 2014. DOI: 10.1093/nar/gkt1026 eprint: https://doi.org/10.1093/nar/gkt1026. [Online]. Available: https://doi.org/10.1093/nar/gkt1026

[6] P. N. Robinson and C. Webber, "Phenotype ontologies and cross-species analysis for translational research," *PLoS Genetics*, vol. 10, no. 4, e1004268, 2014. DOI: 10.1371/journal.pgen.1004268 eprint: https://doi.org/10.1371/journal.pgen.1004268. [Online]. Available: https://doi.org/10.1371/journal.pgen.1004268

[7] J. X. Chong et al., "The genetic basis of mendelian phenotypes: Discoveries, challenges, and opportunities," *American Journal of Human Genetics*, vol. 97, no. 2, pp. 199–215, 2015. DOI: 10.1016/j.ajhg.2015.06.009 eprint: https://doi.org/10.1016/j.ajhg.2015.06.009. [Online]. Available: https://doi.org/10.1016/j.ajhg.2015.06.009

[8] T. Groza et al., "The human phenotype ontology: Semantic unification of common and rare disease," *American Journal of Human Genetics*, vol. 97, no. 1, pp. 111–124, 2015. DOI: 10.1016/j.ajhg.2015.05.020 eprint: https://doi.org/10.1016/j.ajhg.2015.05.020. [Online]. Available: https://doi.org/10.1016/j.ajhg.2015.05.020

[9] S. Richards et al., "Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology," *Genetics in Medicine*, vol. 17, no. 5, pp. 405–424, 2015. DOI: 10.1038/gim.2015.30 eprint: https://doi.org/10.1038/gim.2015.30. [Online]. Available: https://doi.org/10.1038/gim.2015.30

[10] H. Yang, P. N. Robinson, and K. Wang, "Phenolyzer: Phenotype-based prioritization of candidate genes for human diseases," *Nature Methods*, vol. 12, no. 9, pp. 841–843, 2015. DOI: 10.1038/nmeth.3484 eprint: https://doi.org/10.1038/nmeth.3484. [Online]. Available: https://doi.org/10.1038/nmeth.3484

[11] M. Lek et al., "Analysis of protein-coding genetic variation in 60,706 humans," *Nature*, vol. 536, no. 7616, pp. 285–291, 2016. DOI: 10.1038/nature19057 eprint: https://doi.org/10.1038/nature19057. [Online]. Available: https://doi.org/10.1038/nature19057

[12] K. M. Boycott et al., "International cooperation to enable the diagnosis of all rare genetic diseases," *American Journal of Human Genetics*, vol. 100, no. 5, pp. 695–705, 2017. DOI: 10.1016/j.ajhg.2017.04.003 eprint: https://doi.org/10.1016/j.ajhg.2017.04.003. [Online]. Available: https://doi.org/10.1016/j.ajhg.2017.04.003

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019. eprint: https://arxiv.org/abs/1810.04805. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[14] S. Köhler et al., "Expansion of the human phenotype ontology (hpo) knowledge base and resources," *Nucleic Acids Research*, vol. 47, no. D1, pp. D1018–D1027, 2019. DOI: 10.1093/nar/gky1105 eprint: https://doi.org/10.1093/nar/gky1105. [Online]. Available: https://doi.org/10.1093/nar/gky1105

[15] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019. DOI: 10.1056/NEJMra1814259 eprint: https://doi.org/10.1056/NEJMra1814259. [Online]. Available: https://doi.org/10.1056/NEJMra1814259

[16] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 3982–3992, 2019. eprint: https://arxiv.org/abs/1908.10084. [Online]. Available: https://www.aclweb.org/anthology/D19-1410

Appendix A

System Configuration Parameters

A.1 Embedding Model Parameters

- Model: BART sentence transformer
- Embedding dimension: 768
- Tokenization: Byte-pair encoding
- Maximum sequence length: 512 tokens
- Pooling strategy: Mean pooling
- Normalization: L2 normalization to unit length

A.2 FAISS Index Configuration

- Index type: IndexFlatL2 (exact search)
- Distance metric: L2 distance for index, cosine similarity for ranking
- Number of indices: 4 (HPO, Disease, OMIM Gene, Gene Summary)

A.3 Retrieval Parameters

- HPO retrieval depth (k): 20
- Disease retrieval depth (k): 20
- Gene retrieval depth (k): 100
- HPO enhancement threshold: 0.85
- Context enhancement: Conditional based on threshold

A.4 Tiered Scoring Parameters

- Tier 1 proportion: 30%
- Tier 2 proportion: 15%
- Tier 3 proportion: 55%

A.5 Language Model Parameters

- Model: Llama 3.2 3B
- Quantization: Q4_K_M
- Context window: 4096 tokens
- Temperature: 0.7
- Maximum generation length: 2048 tokens

A.6 Performance Optimization

- Cache TTL: 300 seconds
- Cache size: 100 queries
- Batch size for embedding: 30
- CPU thread count: Auto-detected (system cores)

Appendix B

Dataset Statistics

B.1 Human Phenotype Ontology Database

- Total terms: 16,247
- Average description length: 127 words
- Total size after preprocessing: 2.1 MB text, 2.3 GB vectors

B.2 Rare Disease Database

- Total diseases: 7,384
- Average description length: 342 words
- Total size after preprocessing: 15.8 MB text, 1.8 GB vectors

B.3 OMIM Gene Database

- Total genes: 4,219
- Average description length: 284 words
- Total size after preprocessing: 7.2 MB text, 1.1 GB vectors

B.4 Gene Summary Database

- Total genes: 20,473
- Average description length: 89 words
- Total size after preprocessing: 11.4 MB text, 0.9 GB vectors

Appendix C

Evaluation Dataset Characteristics

C.1 True Positive Cases (n=30)

- Age range at presentation: 2–12 years
- Description length: 78–456 words (mean: 187 words)
- Number of phenotypic features mentioned: 3–12 (mean: 6.4)
- Documentation style: 18 narrative, 12 structured/template

C.2 True Negative Cases (n=20)

- Disease categories: 8 other muscular dystrophies, 6 neuromuscular disorders, 4 metabolic conditions, 2 chromosomal abnormalities
- Description length: 65–398 words (mean: 172 words)
- Number of phenotypic features: 2–9 (mean: 5.8)

Appendix D

Computational Environment

D.1 Hardware Specifications

- Processor: Intel Xeon E5-2680 v4 @ 2.40GHz (28 cores)
- RAM: 64 GB DDR4
- Storage: 1 TB SSD
- Operating System: Ubuntu 20.04 LTS

D.2 Software Versions

- Python: 3.8.10
- PyTorch: 1.12.1
- Transformers: 4.24.0
- FAISS: 1.7.3
- LangChain: 0.0.335
- NumPy: 1.23.5
- Pandas: 1.5.2
- Scikit-learn: 1.2.0