PrimeKG: A Multimodal Knowledge Graph for Precision Medicine

Roshan Raj Shah

RWTH Aachen University, Aachen, Germany

Abstract. PrimeKG is a multimodal knowledge graph created to integrate fragmented biomedical knowledge into a unified, clinically relevant framework. It brings together 20 high-quality resources into a graph containing 129,375 nodes and 4,050,249 edges, spanning ten biological scales that include diseases, proteins, phenotypes, drugs, pathways, and environmental exposures. PrimeKG goes beyond prior biomedical knowledge graphs by explicitly modeling drug–disease relationships through indications, contraindications, and off-label uses, and by augmenting nodes with natural-language clinical guidelines from trusted sources such as Mayo Clinic and Orphanet. These features make PrimeKG both broad in coverage and rich in contextual detail, supporting analyses that combine structured graph reasoning with textual information. By bridging molecular data and clinical knowledge, PrimeKG provides a comprehensive resource for precision medicine research and AI-driven biomedical discovery.

Keywords: Knowledge Graph · Precision Medicine · Multimodal · Biomedical Data Integration · Artificial Intelligence

1 Introduction

The last two decades have witnessed an unprecedented acceleration in biomedical data generation. High-throughput sequencing, large-scale proteomic profiling, medical imaging, and the routine use of electronic health records have produced massive volumes of data spanning nearly every biological scale. These developments have fundamentally expanded the potential for understanding disease mechanisms, discovering novel therapeutic targets, and designing individualized treatments. However, the biomedical knowledge derived from such data remains scattered across siloed databases, unstructured literature, and domain-specific ontologies. For any single disease, the relevant information about genetic drivers, molecular pathways, phenotypic manifestations, environmental risk factors, and therapeutic options is rarely co-located. This fragmentation creates barriers to holistic reasoning and complicates the translation of data-driven discoveries into actionable insights for clinical care [1,9,8].

The central ambition of precision medicine is to move beyond one-size-fits-all approaches and deliver interventions that are tailored to the individual patient.

Achieving this requires not only the identification of relevant molecular biomarkers but also the integration of knowledge across different biological scales. A patient's genomic variants must be contextualized with molecular interactions, phenotypic outcomes, drug responses, and even environmental exposures. Without such integration, precision medicine risks being reduced to isolated insights rather than a coherent, actionable framework for care. The question, therefore, is not whether sufficient biomedical data exist, but whether they can be effectively unified, harmonized, and analyzed in ways that are clinically meaningful [6].

Knowledge graphs (KGs) offer a compelling answer to this challenge. A KG encodes entities such as genes, diseases, and drugs as nodes, with edges representing relationships like associations, interactions, or therapeutic effects. This network-based structure provides three critical advantages. First, it allows integration across heterogeneous sources, unifying disparate datasets into a shared schema. Second, it is semantically rich: edges can encode different relationship types, enabling nuanced reasoning beyond simple co-occurrence. Third, KGs are naturally compatible with modern machine learning techniques, including graph neural networks and embedding models, which can leverage the graph structure to uncover hidden associations and generate predictive insights. These qualities make KGs uniquely positioned to bridge the gap between fragmented biomedical knowledge and the demands of precision medicine [7,2].

Several biomedical KGs have already demonstrated their potential to advance research. The Human Disease Network (HDN) [4] revealed the "diseasome" by linking diseases through shared gene associations, showing that seemingly distinct conditions often share molecular roots. The Human Symptoms–Disease Network (HSDN) [13] complemented this approach by uncovering disease connections based on symptom overlap, providing an early demonstration of how patient-facing phenotypic information could be systematized. SPOKE [5] significantly expanded the field by integrating dozens of biomedical datasets into a large, heterogeneous network, enabling researchers to overlay patient-level data onto graph structures for translational applications. The Genetic and Rare Diseases Information Center (GARD) [15] addressed an important gap by curating detailed information on rare and underdiagnosed conditions. More recently, the COVID-19 Open Research Dataset (CORD-19) [12] illustrated the adaptability of KGs during a global health emergency, supporting large-scale literature mining and drug repurposing efforts in real time. Together, these examples underscore the versatility of KGs in synthesizing data and supporting biomedical discovery.

Nevertheless, the impact of earlier biomedical KGs has been constrained by several persistent limitations. Many rely on expert manual curation, a process that ensures quality but severely limits scalability in the face of exponentially growing data. Others struggle with ontology inconsistency: databases often adopt vocabularies tailored to their own purposes, such as ICD for billing, Orphanet for rare diseases, or MedGen for genetic disorders, resulting in redundancies and conflicts when integrated. Perhaps most critically, disease representations in many KGs fail to align with clinical reality. Ontological disease concepts are

frequently fragmented into dozens or hundreds of sub-entries that do not map neatly onto recognizable clinical subtypes. Autism spectrum disorder (ASD) exemplifies this problem: MONDO includes 37 ASD-related entries, UMLS records 192, and Orphanet only 6. Such discrepancies complicate computational analysis, obstruct efforts at harmonization, and reduce the clinical interpretability of the graph [3].

To overcome these challenges, Chandak, Huang, and Zitnik introduced the **Precision Medicine Knowledge Graph (PrimeKG)**, a multimodal KG that integrates 20 high-quality resources into a unified, clinically relevant framework [3]. PrimeKG encompasses 129,375 nodes and 4,050,249 relationships spanning ten biological scales, including diseases, proteins, phenotypes, drugs, exposures, and pathways [3]. Two key innovations distinguish PrimeKG from its predecessors. First, it explicitly models drug–disease interactions across three categories: indications, contraindications, and off-label uses. These edges are particularly valuable for translational applications such as drug repurposing, where knowledge of contraindications and off-label patterns can reveal opportunities that would be invisible in graphs limited to approved indications. Second, PrimeKG supplements graph structure with natural-language clinical guidelines from authoritative sources such as Mayo Clinic and Orphanet. This multimodal integration allows analyses that combine graph connectivity with contextual information from text, bridging molecular data with real-world clinical insights [3].

By addressing the scalability, consistency, and clinical alignment issues of prior KGs, PrimeKG positions itself as a transformative platform for precision medicine. Its breadth of coverage improves disease representation by one to two orders of magnitude over earlier resources, while its multimodal design expands the analytical toolkit available to researchers and clinicians [3,6].

The remainder of this report is structured as follows. Section 2 reviews related biomedical knowledge graphs and articulates the specific need for PrimeKG. Section 3 details PrimeKG's design and integration strategy. Section 4 describes the methodology of its construction, including harmonization and quality control. Section 5 outlines the data processing flow and embedding strategies. Section 6 explores key applications and their impact. Section 7 presents a detailed case study on autism spectrum disorder. Section 8 reflects on current limitations and potential future directions, and Section 9 concludes.

2 The Landscape of Biomedical Knowledge Graphs

The idea of using network-based representations to study diseases has developed steadily over the past two decades. Biomedical knowledge graphs (KGs) aim to unify heterogeneous resources and uncover hidden relationships across biological scales. Before introducing PrimeKG, it is essential to review prior efforts, their contributions, and the limitations that motivated the development of a more comprehensive multimodal framework [7].

2.1 Early Disease-Centric Networks

The first influential attempts to map diseases using a network perspective came from the Human Disease Network (HDN) [4]. HDN connected diseases that shared associated genes, producing the so-called "diseasome." This network revealed that many apparently unrelated diseases actually share common molecular roots. Soon after, the Human Symptoms–Disease Network (HSDN) [13] provided a complementary view by connecting diseases through overlapping symptom profiles. Together, HDN and HSDN established the principle that diseases can be understood not in isolation but as part of interconnected systems.

These early resources demonstrated the potential of knowledge graphs to generate new insights. However, they were limited in scope: HDN relied primarily on gene–disease associations, while HSDN depended on phenotypic co-occurrence. Both lacked broad coverage of drugs, pathways, or higher-level biological contexts, and neither was designed with scalability in mind [4,13].

2.2 Integration of Heterogeneous Data

As the volume of biomedical data increased, efforts shifted toward integrating multiple resource types into a single knowledge graph. A notable example is the Scalable Precision Medicine Open Knowledge Engine (SPOKE) [5]. SPOKE integrates dozens of biomedical databases covering genes, diseases, drugs, pathways, and ontologies. Importantly, SPOKE allows researchers to overlay patient-level data, enabling translational studies that link individual clinical records to large-scale biomedical knowledge. Despite its innovative scope, SPOKE provides only limited disease coverage and focuses more on structural integration than on harmonizing clinically meaningful disease representations [5].

Another important initiative is the Genetic and Rare Diseases Information Center (GARD) [15], which curates detailed knowledge about rare conditions. Rare diseases represent a critical but often neglected area of medicine, and GARD has been valuable for centralizing knowledge for conditions that are typically underrepresented in large-scale biomedical databases. However, while GARD addresses an important gap, it is not designed for large-scale integration across multiple biological modalities [15].

2.3 Rapid-Response Knowledge Graphs

The COVID-19 pandemic illustrated both the importance and adaptability of biomedical KGs. The COVID-19 Open Research Dataset (CORD-19) [12] was released as a large corpus of literature and subsequently structured into knowledge graphs that enabled machine reading, search, and drug repurposing analyses. CORD-19 demonstrated how KGs can support urgent, time-sensitive biomedical challenges by rapidly integrating new knowledge. Nevertheless, such focused efforts are narrow in scope, designed to address specific crises rather than to provide a broad foundation for precision medicine [12].

2.4 Persistent Challenges

Although existing biomedical knowledge graphs have advanced the field considerably, they continue to face several persistent challenges that limit their scalability and clinical utility. A first challenge lies in scalability. Many prior KGs rely heavily on expert manual curation to ensure high-quality integration. While this approach provides accurate and reliable data, it cannot keep pace with the accelerating growth of biomedical knowledge. As new sequencing studies, proteomic datasets, and clinical reports emerge daily, manually curated graphs quickly fall out of date, leaving significant gaps in coverage [7].

A second challenge is the inconsistency of ontologies across resources. Databases such as ICD, Orphanet, MedGen, and MONDO were each designed with different objectives in mind, ranging from billing and administrative coding to research-oriented disease classification. As a result, these ontologies use vocabularies that are not always directly compatible. When integrated into a single KG, they often create redundancies, conflicting identifiers, or mismatched levels of granularity. These inconsistencies make harmonization difficult and sometimes obscure the very relationships that the graph is intended to reveal [3].

Beyond technical integration, a third limitation concerns clinical alignment. Disease representations in many KGs do not correspond to categories that are clinically meaningful. Autism spectrum disorder (ASD) is a striking example. In MONDO, ASD is represented by 37 separate entries; in UMLS, by 192; and in Orphanet, by only 6. Such fragmentation introduces ambiguity that complicates computational analyses and makes it difficult for the graph to reflect real-world diagnostic practice. Without harmonization, these discrepancies risk producing misleading results in downstream applications such as subtype classification or treatment recommendation [3].

Finally, most earlier KGs provide only limited multimodality. They emphasize structured relationships—such as gene–disease associations or drug–target interactions—while neglecting unstructured information that could provide richer context. Clinical guidelines, textual disease descriptions, and patient-facing resources are rarely incorporated, even though they capture valuable information about disease presentation, management strategies, and real-world variability. The absence of these modalities constrains the utility of prior KGs for tasks that require bridging molecular insights with clinical decision-making [3,14].

Together, these challenges highlight why existing biomedical KGs, though valuable, remain insufficient as foundations for precision medicine. Addressing scalability, ontology consistency, clinical alignment, and multimodality is essential for building the next generation of resources, a gap that PrimeKG was designed to fill [3].

2.5 The Need for PrimeKG

The limitations of earlier KGs highlight the need for a new resource that combines the breadth of heterogeneous integration with the depth of clinically meaningful disease representation. PrimeKG was designed to address these gaps by integrating 20 curated resources, harmonizing disease concepts, expanding drug–disease

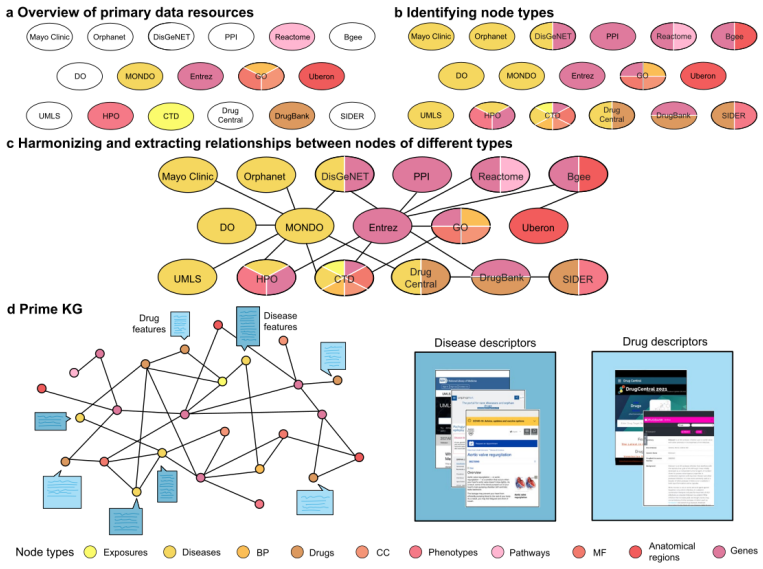

Fig. 1: Workflow for constructing PrimeKG, adapted from [3]. The diagram illustrates the process from raw data ingestion and ontology mapping through harmonization, graph assembly, text augmentation, and embedding generation, leading to downstream applications.

relationships to include contraindications and off-label uses, and supplementing graph structure with natural-language clinical guidelines. In doing so, PrimeKG establishes itself as a multimodal, clinically aligned knowledge graph capable of supporting both mechanistic discovery and translational applications in precision medicine [3,6].

3 Methodology of PrimeKG Construction

The construction of PrimeKG followed a systematic methodology designed to ensure scale, consistency, and clinical relevance [3]. This section describes the data processing pipeline, beginning with identifier standardization, continuing through harmonization of overlapping entities, incorporation of multimodal textual data, and culminating in the assembly of a unified graph. The result is a knowledge graph that integrates twenty heterogeneous biomedical resources into a coherent multimodal framework [3,7].

3.1 Integrated Resources

PrimeKG draws from 20 curated datasets spanning molecular biology, phenotypic abnormalities, clinical guidelines, pharmacology, exposures, and ontologies.

Node Type	Count	Percent (%)	Data Sources
Biological process	28,642	22.1	CTD, Entrez Gene, Gene Ontology
Protein	27,671	21.4	Bgee, CTD, DisGeNET, DrugBank, Entrez Gene, Human Phenotype Ontology, Human PPI Network, Reactome, UMLS
Disease	17,080	13.2	CTD, DisGeNET, Disease Ontology, Drug Central, Human Phenotype Ontology, Mayo Clinic, MONDO Disease Ontology, Orphanet
Phenotype	15,311	11.8	DisGeNET, Human Phenotype Ontology, SIDER
Anatomy	14,035	10.8	Bgee, UBERON
Molecular function	11,169	8.6	CTD, Entrez Gene, Gene Ontology
Drug	7,957	6.2	DrugBank, Drug Central, SIDER
Cellular component	4,176	3.2	CTD, Entrez Gene, Gene Ontology
Pathway	2,516	1.9	Reactome
Exposure	818	0.6	CTD
Total	129,375	100.0	20 primary data sources

Fig. 2: Integrated resources in PrimeKG, adapted from [3]. The table lists representative datasets, their biomedical domains, and the scale of nodes and edges contributed.

Together these resources provide the foundation for 129,375 nodes and 4,050,249 edges across ten biological scales. Each dataset contributes a unique dimension: Bgee offers tissue-specific gene expression; CTD links environmental exposures to genes and diseases; DisGeNET and ClinVar provide gene–disease associations; DrugBank and DrugCentral enrich the drug–disease space; STRING and Reactome contribute networks of protein interactions and pathways; SIDER encodes drug side effects; while Orphanet and Mayo Clinic provide natural-language clinical descriptions. Ontologies such as MONDO, HPO, UBERON, GO, and UMLS enable harmonization across datasets [3].

3.2 Standardization and Ontology Mapping

The first stage of construction was to standardize identifiers across the twenty source resources, each of which used its own schema (e.g., genes by Entrez or Ensembl IDs, diseases by OMIM or Orphanet codes, drugs by DrugBank identifiers). Without harmonization, these mismatches would lead to duplication and fragmentation. PrimeKG resolved this by mapping entities to common ontologies: diseases to MONDO, phenotypes to HPO, anatomy to UBERON, proteins to Entrez Gene IDs, biological processes to GO, and drugs to DrugBank/DrugCentral IDs [3]. This process involved handling deprecated identifiers, resolving mismatches in granularity (e.g., ICD's broad categories vs. Orphanet's fine-grained subtypes), and disambiguating cases where a condition could map to multiple terms, such as "juvenile myoclonic epilepsy" being treated as either a disease or a phenotype. By prioritizing MONDO as the canonical disease ontology and using UMLS as a bridge across vocabularies, PrimeKG ensured that entities were consistently aligned, establishing a coherent backbone for the graph [3].

3.3 Harmonization of Disease Concepts

After ontology mapping, PrimeKG addressed redundancy in disease concepts, which often appeared under slightly different labels across ontologies. For example, "autistic disorder" (UMLS), "autism spectrum disorder" (MONDO), and "autism with seizures" (Orphanet) would otherwise be treated as separate nodes, fragmenting the representation of autism. To resolve this, PrimeKG combined string-based matching of names and synonyms with ClinicalBERT embeddings of disease descriptions, enabling the grouping of lexically or semantically similar entries such as "autism with epilepsy" and "autism spectrum disorder with seizures." This process reduced 22,205 MONDO disease nodes to 17,080 harmonized entities. Rather than merely merging duplicates, the method created clinically coherent subgroups. In the case of autism, fragmented entries were consolidated into clusters reflecting known subtypes such as seizure-associated autism, psychiatric comorbid autism, or gastrointestinal-associated autism. This ensured that disease representation remained clinically meaningful and aligned with real-world diagnostic practice [3,6].

3.4 Graph Assembly

With entities standardized and harmonized, PrimeKG proceeded to graph assembly. All knowledge was converted into typed triplets `(head, relation, tail)`. Relationships covered a wide range of biomedical associations: gene–disease, protein–protein interactions, drug–target, pathway membership, disease–phenotype, and drug–disease interactions (indications, contraindications, off-label uses) [3].

A critical aspect of assembly was maintaining provenance. If multiple resources supported the same edge (e.g., DisGeNET and CTD both linking *TP53* to lung cancer), the edge was consolidated but annotated with both sources. This preserved traceability without inflating edge counts.

Another challenge was directionality. Many databases differ in how they encode relationships, for example: some list drug indications as `(disease, treated_by, drug)` while others use `(drug, treats, disease)`. PrimeKG standardized relation directionality across all sources, ensuring consistency and interpretability.

Finally, the graph was pruned to its largest connected component. This step was important to remove isolated nodes that were unlikely to provide meaningful context for analyses. The largest component retained more than 99.9% of nodes and edges, confirming that PrimeKG is highly connected [3,7].

3.5 Clinical Text Augmentation

To move beyond purely structural relationships, PrimeKG enriched nodes with textual descriptions. Disease nodes were annotated with information such as definitions, causes, symptoms, complications, and management guidelines from sources including Mayo Clinic, Orphanet, MONDO, and UMLS. Drug nodes were supplemented with mechanisms of action, pharmacodynamics, pharmacokinetics, and warnings from DrugBank and DrugCentral [3].

This textual layer adds clinical realism. For example, risperidone is connected to autism through an edge in the graph. The text field expands on this by stating: *"Risperidone is prescribed to reduce irritability in children with autism spectrum disorder, but long-term use may lead to significant weight gain and metabolic complications."* This allows researchers to analyze not just the structural association but also the contextual meaning of that relationship.

By integrating text alongside graph structure, PrimeKG enables multimodal analyses. Graph embeddings can capture network topology, while language models can capture semantic content. Combined, these features provide a richer basis for downstream machine learning tasks such as drug repurposing or disease similarity prediction [3,14].

3.6 Quality Control and Validation

The final stage of constructing PrimeKG involved extensive quality control to ensure that the graph was not only large and multimodal, but also consistent, reliable, and clinically meaningful. Validation was carried out in several complementary ways, beginning with coverage benchmarking. One important benchmark was Orphanet, which provides curated knowledge on rare diseases. PrimeKG was compared against this resource and shown to capture more than 90% of rare diseases, representing a significant improvement over earlier biomedical knowledge graphs that typically covered only a small subset. This confirmed that PrimeKG provides both breadth and depth in disease representation [3].

Another critical validation step focused on drug–disease associations. PrimeKG's edges describing indications, contraindications, and off-label uses were evaluated against FDA-approved therapies released since June 2021. The majority of these new approvals were already present in PrimeKG, demonstrating that the graph captures contemporary clinical knowledge and can support translational applications such as drug repurposing and therapeutic discovery. This analysis highlighted the value of explicitly including contraindication and off-label edges, which had been largely absent in prior biomedical graphs [3,11].

Beyond benchmarking against external references, the internal structure of the graph was also evaluated. Degree distributions and edge-type frequencies were examined and found to follow patterns consistent with other large-scale biological networks, such as scale-free behavior in protein–protein interactions. This suggested that the integration process had preserved the statistical properties expected in biological systems. Finally, additional integrity checks were carried out to ensure that edges were correctly directed, identifiers were consistently mapped to ontologies, and no major inconsistencies remained between different data sources. Taken together, these validation efforts confirmed that PrimeKG is a robust and clinically aligned resource that can be trusted as a foundation for precision medicine research [3].

4 Data Processing Flow and Embeddings

After the construction of PrimeKG, the next step is to transform the assembled knowledge graph into formats that can be directly used for computational analyses and machine learning tasks. This involves a data processing pipeline that prepares the graph, extracts features, and generates embeddings—vector representations of nodes and edges that preserve the structure and semantics of the graph in a numerical form [3,6].

4.1 Data Processing Workflow

The data processing begins with the assembled graph exported into standardized file formats such as CSV, TSV, or Parquet. Each edge is stored as a triplet `(head, relation, tail)`, while node attributes include ontology identifiers and, where available, textual descriptions. The repository accompanying PrimeKG is organized into modules that reflect this workflow: (i) `datasets/processing_scripts/` for downloading and cleaning raw resources, (ii) `knowledge_graph/` for assembling the harmonized graph, and (iii) `embeddings/` for learning numerical representations of nodes. This modular structure makes it possible to reproduce the graph and retrain embeddings as new versions of the resources are released [3].

Prior to embedding generation, several preprocessing steps are performed. These include removing isolated nodes not part of the largest connected component, normalizing edge directionality, and ensuring that provenance information is retained for traceability. Node features such as ontology mappings and textual attributes are also prepared at this stage, enabling both unimodal and multimodal embedding strategies [3].

4.2 Graph Embedding Methods

Embeddings map graph entities into low-dimensional vector spaces where geometric proximity reflects network similarity. Several embedding methods have been applied in the context of PrimeKG [3,6,2]:

node2vec. Node2vec is a random-walk based embedding algorithm that generates node sequences by simulating biased random walks across the graph. It then applies the skip-gram model, commonly used in natural language processing, to learn embeddings that preserve both local and global neighborhoods. In PrimeKG, node2vec can capture relationships such as the tendency of diseases to cluster by shared genetic drivers [6].

metapath2vec. Metapath2vec extends random-walk embeddings to heterogeneous graphs by constraining walks to follow specific types of edges, or "metapaths." For example, a walk might be restricted to the pattern `disease → gene → disease`, allowing embeddings to capture disease similarity based on genetic overlap. In PrimeKG, metapath2vec is particularly useful because of the graph's multimodal nature and the diversity of relation types [6].

GraphSAGE. GraphSAGE is an inductive embedding method based on graph neural networks (GNNs). It generates embeddings by aggregating feature information from a node's neighborhood. Unlike node2vec, which requires retraining when new nodes are added, GraphSAGE can generalize to unseen nodes, making it more scalable for large and evolving graphs like PrimeKG [6].

Graph Neural Networks (GNNs). Beyond GraphSAGE, several GNN architectures can be applied to PrimeKG, including Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs). These models propagate information across the graph using neural message passing, allowing them to capture complex dependencies among nodes. GNNs are well-suited for downstream tasks such as link prediction (e.g., predicting novel drug–disease associations) or node classification (e.g., categorizing disease subtypes) [6,10].

4.3 Multimodal Embeddings

One of PrimeKG's distinctive features is its integration of textual clinical knowledge alongside structured graph edges. This opens the possibility of multimodal embeddings, which combine graph-based and text-based representations. Disease and drug descriptions from Mayo Clinic, Orphanet, MONDO, DrugBank, and DrugCentral are embedded using transformer-based language models such as ClinicalBERT or BioBERT. These embeddings are then concatenated or integrated with graph embeddings to produce richer node representations [3,6].

For example, a disease node embedding may combine structural information—capturing its connectivity to genes, drugs, and phenotypes—with semantic information extracted from its textual description. This allows models not only to reason about the topology of the graph but also to incorporate clinical context. Multimodal embeddings are particularly valuable for tasks where text adds nuance to relationships, such as distinguishing between drugs with similar targets but different clinical guidelines or between diseases with overlapping phenotypes but distinct management strategies [3,14].

4.4 Embedding Pipeline

The complete embedding pipeline for PrimeKG thus consists of three layers: (i) graph construction and preprocessing, (ii) generation of unimodal embeddings using algorithms such as node2vec, metapath2vec, or GraphSAGE, and (iii) integration with textual embeddings to create multimodal representations. These embeddings are then used as inputs for downstream analyses, including drug repurposing, disease similarity, and phenotype clustering [3,6]. For a detailed visualization of the reconciliation of autism-related disease nodes, see Appendix A, Figure 3.

5 Applications and Case Study: Autism Spectrum Disorder

PrimeKG enables a broad range of applications in precision medicine, from therapeutic discovery to disease similarity analysis and safety monitoring. Its enriched graph structure, combined with multimodal textual context, makes it uniquely positioned for bridging molecular insights with clinical outcomes [3]. This section first discusses three representative applications—drug repurposing, disease similarity, and adverse event prediction—and then presents a detailed case study on autism spectrum disorder (ASD), including results from the public GitHub repository analysis.

5.1 Drug Repurposing

Drug repurposing, the identification of new therapeutic uses for existing drugs, is one of the most impactful applications of biomedical knowledge graphs. PrimeKG is well-suited for this task because it incorporates not only drug indications but also contraindications and off-label uses. To validate this capability, the authors compared PrimeKG against FDA-approved therapies released after June 2021, as shown in Appendix A, Figure 4. Among 40 new approvals, 11 involved repurposed drugs. Graph-based proximity analysis showed that 8 out of 10 repurposed drugs were significantly closer to their new disease indications in PrimeKG than expected by chance. This suggests that network features in PrimeKG can anticipate repurposing opportunities, reducing the search space for experimental validation. Oncology drugs in particular were shown to connect to new cancer indications through shared pathways and genetic drivers, mirroring subsequent FDA approvals [3,10].

5.2 Disease Similarity

PrimeKG also supports the analysis of disease similarity across scales. Because it integrates data from genes, pathways, phenotypes, and textual guidelines, diseases that share biological mechanisms or clinical features cluster together. For example, PrimeKG recapitulates the known similarity between breast and ovarian cancers via shared *BRCA1/2* mutations, and between Alzheimer's and Parkinson's disease through protein–protein interactions in neurodegeneration pathways. Beyond confirming known associations, such analyses can reveal novel clusters of related conditions, supporting comorbidity studies, subtyping, and therapeutic repositioning [3,14].

5.3 Adverse Event Prediction

Another application is the prediction of adverse drug events. PrimeKG integrates over 200,000 drug–phenotype links from SIDER, enabling explicit modeling of side effects. This allows researchers to examine not only potential benefits but

also safety risks of repurposing candidates. For example, clozapine's association with agranulocytosis and corticosteroids' link to osteoporosis are encoded directly in the graph. By training models on these relationships, it becomes possible to predict novel drug–phenotype edges, flagging potential adverse events for further investigation [3,11].

5.4 Autism Spectrum Disorder: A Case Study

Autism spectrum disorder (ASD) illustrates the strengths of PrimeKG in harmonizing fragmented disease ontologies and connecting them to molecular, phenotypic, and therapeutic knowledge. Ontological inconsistencies are especially striking in autism: MONDO lists 37 autism-related entries, UMLS 192, and Orphanet only 6. Without harmonization, these would appear as separate nodes, fragmenting the representation of autism. PrimeKG resolves this by combining lexical matching with ClinicalBERT embeddings of disease descriptions, grouping semantically similar entries into a single clinically coherent ASD node [3].

From the GitHub case study, clustered subtypes of autism emerged that reflected clinical variability, including seizure-associated autism, psychiatric comorbid autism, and gastrointestinal disorder-associated autism, as illustrated in Appendix A, Figure 6. This demonstrates how PrimeKG not only reduces redundancy but also highlights clinically meaningful heterogeneity [3].

The notebook further explores how ASD connects to therapeutic agents. Direct queries confirmed that risperidone is linked to ASD through a drug–disease indication edge. Analysis of the risperidone drug node revealed pharmacological details such as its dopamine D2 receptor mechanism and known side effects, demonstrating how drug features are captured alongside relationships [3].

PrimeKG also enables multi-hop reasoning. By analyzing one-hop and two-hop neighborhoods of the ASD node, the notebook revealed that risperidone can be reached via intermediate nodes representing epilepsy and related phenotypes, as shown in Appendix A, Figure 5. This reflects clinical reality: risperidone is often prescribed to reduce irritability in children with autism, particularly those with comorbid seizures. Further multi-hop queries uncovered paths connecting ASD to risperidone through proteins (e.g., ALB and dopamine receptors), drug–protein interactions, and anatomy–protein associations [3].

The case study highlights PrimeKG's ability to unify ontology terms, represent heterogeneity, and connect diseases to treatments through mechanistic intermediates. By linking autism to epilepsy, gastrointestinal phenotypes, and pharmacological agents like risperidone, PrimeKG illustrates how a multimodal knowledge graph can provide clinically interpretable paths that reflect real-world practice. The inclusion of scripts and visualizations in the GitHub repository ensures that these analyses are transparent and reproducible, allowing others to extend the case study to additional diseases [3].

6 Discussion

PrimeKG represents a substantial step forward in the development of multimodal biomedical knowledge graphs, but its evaluation also reveals both achievements and remaining challenges. The results presented in the original paper demonstrate that PrimeKG substantially outperforms prior efforts in terms of coverage and translational utility. Table 4 of the publication shows that PrimeKG captures 90.8% of Orphanet rare diseases, an order-of-magnitude improvement compared to SPOKE, HSDN, and GARD. This breadth ensures that the graph is not restricted to common disorders but also supports research into the long tail of rare diseases. Similarly, validation against FDA drug approvals since June 2021 confirmed that 8 out of 10 repurposed drugs were significantly closer to their new indications in PrimeKG than expected by chance. These results, summarized in Appendix A, Figure 4, demonstrate that PrimeKG is capable of anticipating clinically validated repurposing events and provide evidence of its translational value [3].

The autism spectrum disorder (ASD) case study further illustrates PrimeKG's clinical interpretability. Ontology inspection showed severe fragmentation of autism-related concepts, with 37 entries in MONDO, 192 in UMLS, and only 6 in Orphanet. PrimeKG harmonized these into a single coherent ASD node using ClinicalBERT embeddings, while still preserving clinically meaningful subtypes. The harmonized node was connected to phenotypes such as epilepsy and gastrointestinal abnormalities and to drugs such as risperidone. This structure reproduces real-world practice, where risperidone is prescribed for irritability in autism, especially in patients with comorbid seizures. The GitHub reference notebook (`autism.ipynb`) reproduced these results, confirming ontology mismatch counts, ClinicalBERT grouping, and path-based connections from autism to risperidone through intermediate phenotypes. The availability of these results in an open-source notebook reinforces PrimeKG's transparency and reproducibility [3].

Additional tests using my own notebooks revealed both strengths and limitations. For Hurler syndrome, a rare lysosomal storage disorder, the pipeline worked effectively. The generated subgraph recovered known therapies such as enzyme replacement therapy with Laronidase and hematopoietic stem cell transplantation, alongside associated phenotypes including skeletal abnormalities and developmental delay. This confirmed that PrimeKG's improved rare disease coverage is not only quantitative but also practically useful [3]. The independent autism notebook I developed (`PrimeKG_Autism_Case_Study.ipynb`) also produced consistent results, highlighting risperidone as a therapeutic node and revealing multi-hop paths through epilepsy and protein intermediates such as ALB. In contrast, the COVID-19 case study with Baricitinib was less successful. Although Baricitinib and relevant targets such as AAK1, GAK, and JAK1/2 were present in the graph, the expected mechanistic paths were incomplete or missing. This shortcoming reflects the temporal limitations of PrimeKG, which was constructed with data available before the end of 2021 and therefore does not capture later pandemic updates. The COVID-19 test highlights the importance

of timely updates: a static knowledge graph, no matter how comprehensive at release, risks becoming outdated in rapidly evolving biomedical domains [3,12].

These results emphasize several broader limitations. A first issue concerns scalability. PrimeKG integrates twenty resources and harmonizes over 17,000 diseases, yet biomedical data grow at a pace that static releases cannot match. Earlier graphs such as SPOKE and GARD encountered similar problems, as their updates lagged behind new discoveries. Incremental update pipelines will be essential to ensure that PrimeKG remains current. A second challenge involves ontology drift. Although PrimeKG harmonizes MONDO, UMLS, and Orphanet, these vocabularies evolve independently, introducing redundancies and mismatches over time. ClinicalBERT embeddings improve harmonization, but continuous alignment mechanisms will be required to prevent fragmentation as ontologies change [3,7].

Bias in data sources also persists. Many integrated datasets disproportionately emphasize well-studied conditions, meaning that common diseases still dominate relative to rare or underdiagnosed ones. While PrimeKG improves coverage of Orphanet rare diseases to above ninety percent, imbalances remain. Textual resources such as Mayo Clinic and Orphanet further introduce regional and cultural biases, as guidelines reflect local practice and may not generalize globally. This raises concerns about fairness when using PrimeKG to train predictive models. Ethical considerations are therefore central: predictions made from the graph are only as reliable as the data they are built upon. Without deliberate audits and inclusion of diverse resources, there is a risk that PrimeKG could perpetuate biases rather than mitigate them [3,6].

Finally, PrimeKG's multimodal design points toward future opportunities. At present, graph structure and textual features are combined primarily at the data level. Future research should develop multimodal graph neural networks capable of jointly reasoning over structure and text, capturing both topological similarity and semantic nuance. In addition, uncertainty quantification and provenance tracking should be built into edges to improve interpretability and trust. Federated update pipelines, where distributed data providers contribute updates in near real-time, could also help PrimeKG avoid the pitfalls of outdated snapshots. Such technical and ethical advances will be crucial to move PrimeKG from a proof-of-concept research graph to a robust and equitable platform for precision medicine [3,6].

Taken together, these reflections highlight both the successes of PrimeKG and the work that remains. Compared to earlier biomedical KGs, it achieves far greater coverage, integrates multimodality, and demonstrates translational applications in drug repurposing and disease analysis. Yet the challenges of scalability, ontology drift, data bias, and software stability remind us that the resource is still evolving. The continued development of PrimeKG will require both technical innovation and community engagement to ensure that it remains accurate, equitable, and clinically meaningful [3].

7 Conclusion

PrimeKG marks an important advance in the effort to integrate biomedical knowledge at scale. By unifying twenty diverse resources into a single multimodal framework, it demonstrates that it is possible to connect molecular mechanisms, phenotypic outcomes, and clinical practice in a coherent and clinically interpretable way. The breadth of its coverage, extending to more than 17,000 harmonized diseases, and the richness of its drug–disease layer establish PrimeKG as a resource that goes well beyond earlier knowledge graphs. Applications ranging from drug repurposing to disease similarity analysis confirm its translational potential, and case studies such as autism and Hurler syndrome highlight how it can resolve long-standing ontological inconsistencies while producing outputs that align with clinical reality [3].

At the same time, PrimeKG also illustrates the challenges of maintaining relevance in a rapidly evolving biomedical landscape. Its limitations—such as the outdated representation of COVID-19 pathways—underscore the importance of building mechanisms for continual updates, managing ontology drift, and addressing bias in both structured data and clinical text. These challenges do not diminish the value of PrimeKG; rather, they define the agenda for its next phase of development. Looking forward, progress will depend on automated update pipelines, multimodal graph learning methods, and community-driven curation efforts. If these challenges are met, PrimeKG will not only remain a state-of-the-art research resource but also evolve into a dynamic platform capable of supporting equitable and trustworthy precision medicine on a global scale [3,6].

References

1. Adams, S., Petersen, C.: Big data and precision medicine. Journal of Translational Medicine **14**(1), 36 (2016). https://doi.org/10.1186/s12967-016-0797-4
2. Alshahrani, M., et al.: Biological applications of knowledge graph embedding models. Briefings in Bioinformatics **22**(2), 1679–1693 (2020). https://doi.org/10.1093/bib/bbaa192, https://academic.oup.com/bib/article/22/2/1679/5739186
3. Chandak, P., Huang, K., Zitnik, M.: Building a knowledge graph to enable precision medicine. Scientific Data **10**(1), 67 (2023). https://doi.org/10.1038/s41597-023-01960-3, https://pubmed.ncbi.nlm.nih.gov/36732524
4. Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., Barabási, A.L.: The human disease network. Proceedings of the National Academy of Sciences **104**(21), 8685–8690 (2007). https://doi.org/10.1073/pnas.0701361104
5. Himmelstein, D.S., et al.: Systematic integration of biomedical knowledge with hetionet: A public knowledge graph for drug repurposing. eLife **6**, e26726 (2017). https://doi.org/10.7554/eLife.26726, https://pubmed.ncbi.nlm.nih.gov/29144497
6. Li, M.M., Huang, K., Zitnik, M.: Graph representation learning in biomedicine and healthcare. Nature Biomedical Engineering **6**(12), 1353–1369 (2022). https://doi.org/10.1038/s41551-022-00942-x, https://pubmed.ncbi.nlm.nih.gov/36316368

7. Nicholson, D.N., Greene, C.S.: Constructing knowledge graphs and their biomedical applications. Computational and Structural Biotechnology Journal **18**, 1414–1428 (2020). https://doi.org/10.1016/j.csbj.2020.05.017, https://www.sciencedirect.com/science/article/pii/S2001037020302978
8. Ping, P., Hermjakob, H., et al.: Integrating big data in precision medicine: Research and clinical applications. Journal of the American College of Cardiology **69**(9), 1130–1132 (2017). https://doi.org/10.1016/j.jacc.2016.12.014
9. Prosperi, M., Min, J., Bian, J., Modave, F.: Artificial intelligence and precision medicine: A narrative review of current status and future perspectives. Journal of the American Medical Informatics Association **25**(8), 923–929 (2018). https://doi.org/10.1093/jamia/ocy053
10. Rivas-Barragán, D., et al.: Ensembles of knowledge graph embedding models improve predictions for drug discovery. Briefings in Bioinformatics **23**(6), bbac462 (2022). https://doi.org/10.1093/bib/bbac462, https://academic.oup.com/bib/article/23/6/bbac462/6760141
11. Santos, A., et al.: The open targets platform: Systematic identification and prioritisation of drug targets. Nucleic Acids Research **48**(D1), D1046–D1053 (2020). https://doi.org/10.1093/nar/gkz1025, https://pubmed.ncbi.nlm.nih.gov/31691823
12. Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R., et al.: Cord-19: The covid-19 open research dataset. arXiv preprint arXiv:2004.10706 (2020), https://arxiv.org/abs/2004.10706
13. Zhou, X., Menche, J., Barabási, A.L., Sharma, A.: Human symptoms–disease network. Nature Communications **5**, 4212 (2014). https://doi.org/10.1038/ncomms5212
14. Zhu, C., et al.: Multimodal reasoning based on knowledge graph embedding for specific diseases. Bioinformatics **38**(8), 2235–2243 (2022). https://doi.org/10.1093/bioinformatics/btac100, https://academic.oup.com/bioinformatics/article/38/8/2235/6527626
15. Zhu, X., et al.: Genetic and rare diseases information center (gard) (2020), available at: https://rarediseases.info.nih.gov/ (Accessed: 2025-08-29)

A Appendix A : Supplementary Figures

This appendix provides additional figures referenced in the main text. They illustrate the harmonization of autism-related disease nodes, validation of drug repurposing, and example path-based analyses conducted with PrimeKG.

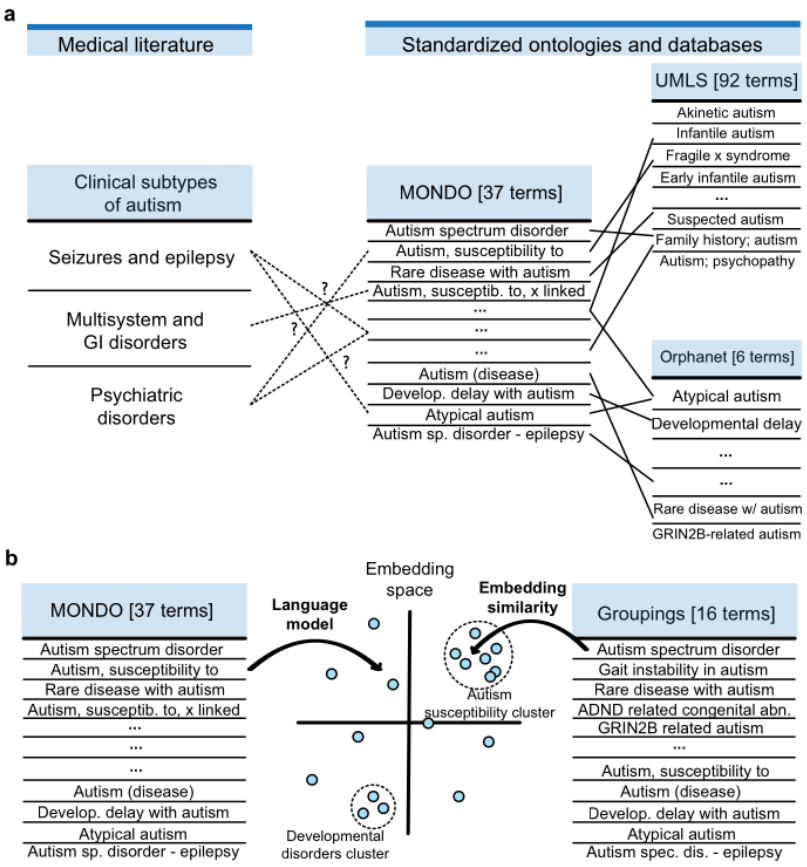

Fig. 3: Reconciling autism-related disease nodes into clinically meaningful entities. (a) Comparison of clinically determined autism subtypes with ontology-specific terms from MONDO, UMLS, and Orphanet. (b) Use of ClinicalBERT embeddings to cluster semantically similar MONDO terms, creating clinically coherent groupings. Adapted from [3].

Source	Type of feature	Count	Unique	Percent (%)
Drug Central[83]	Molecular weight	2,797	2,308	35.2
	TPSA	2,718	2,718	34.2
	cLogP	2,574	980	32.3
DrugBank[80]	Group	7,957	7,903	100.0
	State	6,517	6,463	81.9
	Category	5,431	5,431	68.3
	Description	4,591	4,565	57.7
	Indication	3,393	3,076	42.6
	Mechanism of action	3,242	3,161	40.7
	ATC 4	2,818	1,040	35.4
	ATC 3	2,818	2,818	35.4
	ATC 2	2,818	2,818	35.4
	ATC 1	2,818	2,818	35.4
	Pharmacodynamics	2,659	2,617	33.4
	Half life	2,063	1,893	25.9
	Protein binding	1,669	1,487	21.0
	Pathway	598	598	7.5

Fig. 4: Validation of drug repurposing using FDA-approved therapies since June 2021. PrimeKG significantly captures repurposed drug–disease associations, as shown in FDA benchmarking. Adapted from [3].

```
In [68]: kg.query('x_name=="ALB" and y_name=="Risperidone"')
```

Out[68]:

	relation	display_relation	x_index	x_id	x_type	x_name	x_source	y_index	y_id	y_type	y_name	y_source
5708102	drug_protein	carrier	4315	213	gene/protein	ALB	NCBI	14223	DB00734	drug	Risperidone	DrugBank

Fig. 5: Path-based analysis of autism in PrimeKG. Shortest paths connect ASD to risperidone through phenotypes and proteins, illustrating mechanistic and therapeutic links. Adapted from the PrimeKG GitHub repository.

```
In [14]: grouped_diseases = pd.read_csv('../datasets/kg/auxillary/kg_grouped_diseases.csv')
         x = grouped_diseases.query('{}.str.contains("autism")'.format('node_name'))
         x.get(['node_id','group_name_bert']).groupby('group_name_bert').count()
```

Out[14]:

group_name_bert	node_id
ADNP-related multiple congenital anomalies - intellectual disability - autism spectrum disorder	1
GRIN2B-related developmental delay, intellectual disability and autism spectrum disorder	1
atypical autism	1
autism (disease)	1
autism spectrum disorder	1
autism spectrum disorder - epilepsy - arthrogryposis syndrome	1
autism spectrum disorder due to AUTS2 deficiency	1
autism susceptibility 1	1
autism, susceptibility to	16
autism, susceptibility to, X-linked	5
autism, susceptiblity to	1
autism-facial port-wine stain syndrome	1
developmental delay with autism spectrum disorder and gait instability	1
developmental delay with or without dysmorphic facies and autism	1
intellectual developmental disorder with speech delay, autism, and dysmorphic facies	1
macrocephaly-autism syndrome	1

Fig. 6: Grouping of autism-related disease concepts using ClinicalBERT embeddings. Multiple ontology entries are consolidated into clinically coherent subtypes. Adapted from the PrimeKG GitHub repository.

B Appendix B : Repository Walkthrough and Testing Files

The official PrimeKG GitHub repository contains several directories:

- `datasets/processing_scripts/`: ingestion and preprocessing of raw resources
- `knowledge_graph/`: harmonization and graph assembly
- `embeddings/`: code for generating node embeddings
- `case_study/`: example Jupyter notebooks, including autism

In addition, all files and notebooks used in my own seminar testing are publicly available here:

`https://github.com/RoshanRajShah/PrimeKG_Testing/tree/main`

The repository contains small, reproducible case studies that mirror examples from the PrimeKG paper, adapted for hands-on learning. Below is a summary adapted from the repository README:

- **Objectives:** Load and inspect `kg.csv`; trace disease–gene–drug–pathway mechanisms; replicate case studies (Autism, Hurler syndrome, COVID-19 → Baricitinib); and test network analyses such as shortest paths and ego-subgraphs.
- **Structure:**

```
PrimeKG_Testing/
 notebooks/
   autism.ipynb (From : PrimeKG Repository)
   PrimeKG_Autism_Case_Study.ipynb
   PrimeKG_Hurler_Syndrome_Case_Study.ipynb
   PrimeKG_COVID19_Baricitinib_Case_Study.ipynb
 data/
   kg.csv   # <- put PrimeKG CSV here (gitignored)
 .gitignore
 README.md
```

- **Requirements:** Python $\geq$ 3.9, Jupyter, `pandas`, `networkx`, `matplotlib`, `numpy`.
- **Case Studies:**
  - Autism Spectrum Disorder — reproduces paper's autism analysis.
  - Hurler Syndrome — explores therapies and phenotypic context.
  - COVID-19 → Baricitinib — investigates mechanistic paths and performs a mini permutation test.
- **Reproducibility Tips:** Downsample large graphs for memory efficiency, filter by relevant node/edge types early, and handle CSV quirks with `engine="python"` in `pandas`.

This appendix ensures that all experiments and analyses reported in the main text can be independently reproduced using the notebooks in the linked repository.