Research Project: Graph in Graph for Rare Disease Detection

Kai-Ze Deng Büşra Nur Zeybek Tuna Karacan

March 2025

Abstract

Rare disease diagnosis presents significant challenges in modern healthcare, with patients typically experiencing a diagnostic delay of 5-7 years. This research introduces a novel Graph-in-Graph (GiG) framework for rare disease detection that addresses the fundamental data limitations inherent to rare conditions. Our approach operates at two levels: a node-level that processes individual patient phenotype subgraphs through Graph Convolutional Networks, and a population-level that establishes connections between similar patients. This hierarchical structure enables knowledge transfer across rare disease cases while preserving interpretable connections. We evaluate our approach on the Shepherd Dataset, which contains over 40,000 simulated patient records spanning 2,405 distinct gene classes. Our model demonstrates effective classification performance despite the challenges of data scarcity and class imbalance. Results indicate that the framework's ability to integrate biomedical knowledge graphs and enable cross-patient information sharing offers significant advantages over traditional methods for complex rare disease classification tasks.

1 Introduction

1.1 Definition of Rare Disease

According to the European Commission on Public Health, rare diseases are defined as life-threatening or chronically debilitating conditions with a prevalence of no more than 5 per 10,000 people in the European Union. In the United States, the Orphan Drug Act defines a rare disease as a condition affecting fewer than 200,000 Americans.

1.2 Impact of Rare Disease

There are over 7,000 distinct identified rare diseases to date. While individual rare diseases affect small populations, collectively, they affect approximately 300 million people worldwide, causing a significant public health challenge. Moreover, the average patient with a rare disease faces a diagnostic delay of 5-7 years, during which they often undergo multiple tests, specialist referrals, and misdiagnoses, which causes not only waste of health resources, but also delays critical treatment time for patients.

1.3 Complexity of Detecting Rare Disease

The complexity of detecting rare diseases stems from several key factors:

- **Symptom overlap** — Rare disease symptoms frequently resemble those of common diseases, making accurate diagnose difficult.
- **Phenotypic heterogeneity** — Symptoms differing markedly from patient to patient, even within the same disease, making diagnosis even more challenging as the clinical presentation may not match textbook descriptions or previous experience.
- **Limited clinical experience** — Healthcare providers typically encounter few rare disease cases during their careers, which restricts their ability to recognize distinct clinical patterns.
- **Research constraints** — Due to the inadequate funding and small patient populations, scientific investigation into rare diseases remains limited.

These challenges highlight the need for advanced computational approaches that can identify patterns across limited data and provide interpretable results to support clinical decision-making. Furthermore, if we can discover hidden patterns between patients with the same diseases, it will facilitate further disease research and potentially lead to new insights into rare disease mechanisms and treatments.

1.4 Our Approach

Our research addresses the challenges of rare disease detection through a novel Graph-in-Graph (GiG) framework. This hierarchical approach is specifically designed to overcome data limitations inherent in rare disease classification. The GiG framework consists of three interconnected modules:

- **Node-Level Module:** At the foundation, our approach leverages existing biomedical knowledge graphs to establish meaningful connections between different phenotypes within individual patients. This first-level module processes these individual patient subgraphs through Graph Convolutional Networks (GCNs) and outputs an embedding vector for each patient that encapsulates their phenotypic profile and the complex relationships between symptoms.
- **Population-Level Module:** The second-level module constructs a population-level graph by calculating similarities between patient embeddings and establishing connections between phenotypically similar cases. This structure enables knowledge transfer across patients with similar disease manifestations, effectively addressing the data scarcity problem.
- **Classifier Module:** Finally, this module integrates information from both the node and population levels for the designed downstream task.

This end-to-end learning approach offers several advantages over traditional methods: (1) it preserves the structural relationships between phenotypes rather than treating them as independent features, (2) it enables knowledge sharing between different modules, and (3) it produces interpretable patient-to-patient connections that may reveal previously unrecognized patterns across rare diseases.

1.5 Evaluation

For evaluating our model, we set a downstream task where we classify patients based on the true gene ID from the dataset. This task directly corresponds to diagnosing the underlying rare genetic condition that causes the observed phenotypes.

2 Benefits of Graph Structures for Rare Disease Detection

Graph-based models offer several intrinsic advantages that make them particularly suitable for addressing the complex challenges of rare disease detection.

2.1 Integration of Heterogeneous Biomedical Knowledge

Traditional machine learning approaches face significant limitations when applied to rare disease detection due to sparse training examples and high-dimensional biomedical data. These methods typically treat features as independent dimensions, failing to capture complex relationships between biological entities.

Graph-based models offer a more appropriate framework for biomedical data because they explicitly model the interconnected nature of biological information. In biological systems, entities such as genes, proteins, phenotypes, and diseases form complex networks of interactions and relationships rather than existing as isolated elements. Graph models naturally represent these connections, preserving the inherent structure of biomedical knowledge.

2.2 Leverage Existing Biomedical Knowledge into Training Data

Furthermore, graph-based approaches enable us to leverage information from similar patient profiles. By connecting patients with similar phenotypic presentation, our model can transfer knowledge between related cases, effectively expanding the available training signal beyond individual examples.

2.3 Connecting Similar Patient Profiles

The results of our model establish connections between similar patients, serving two critical functions: leveraging patient data more effectively and uncovering hidden disease patterns.

By establishing these patient-to-patient connections, our approach amplifies the available training signal, enabling information flow between cases that share underlying biological mechanisms despite potentially divergent symptomatic presentations. This data leverage is particularly crucial for rare diseases, where individual examples are inherently scarce.

Additionally, these connections enable the discovery of subtle, previously unrecognized patterns across patient cases. Even when patients present with ostensibly different clinical manifestations, our graph-based approach can recognize latent similarities in their profiles, potentially revealing novel disease subtypes or shared pathological pathways that might remain undetected through conventional analysis methods.

All in all, we believe that the graph-based model is uniquely positioned to address the challenges of rare disease detection by simultaneously representing the nature of biological information and enabling knowledge transfer through the different modules.

3 Methodology

3.1 Overview of Graph-in-Graph Framework

Our approach utilizes existing knowledge graphs to establish connections of individual patient phenotypes. The framework consists of three key interconnected modules:

1. **Node-Level Module (F1)**: Processes individual patient data represented as phenotype subgraphs through graph convolution layers to generate patient embeddings.
2. **Population-Level Module (F2)**: Creates connections between patients with similar profiles, enabling knowledge transfer across rare disease cases.
3. **Classifier Module (F3)**: Combines outputs from F1 and F2 modules for the following downstream task.

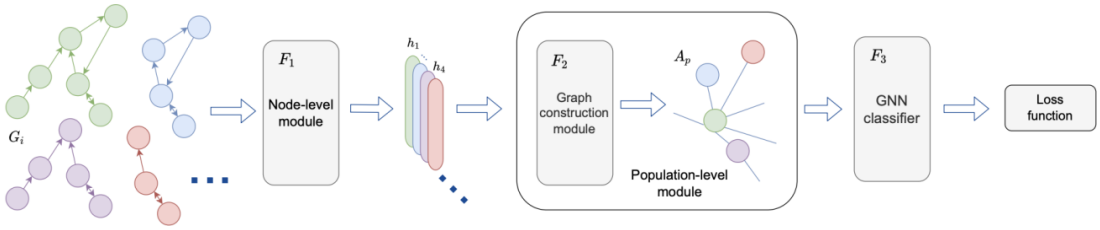

Figure 1: GiG Framework

3.2 Knowledge Graph

A knowledge graph is the known relationships between entities (phenotypes, genes, diseases), organized as a graph where each node represents one entity and edges represent the relationships between them.

In our approach, we utilize knowledge graphs to establish connections between entities to construct the individual patient subgraph.

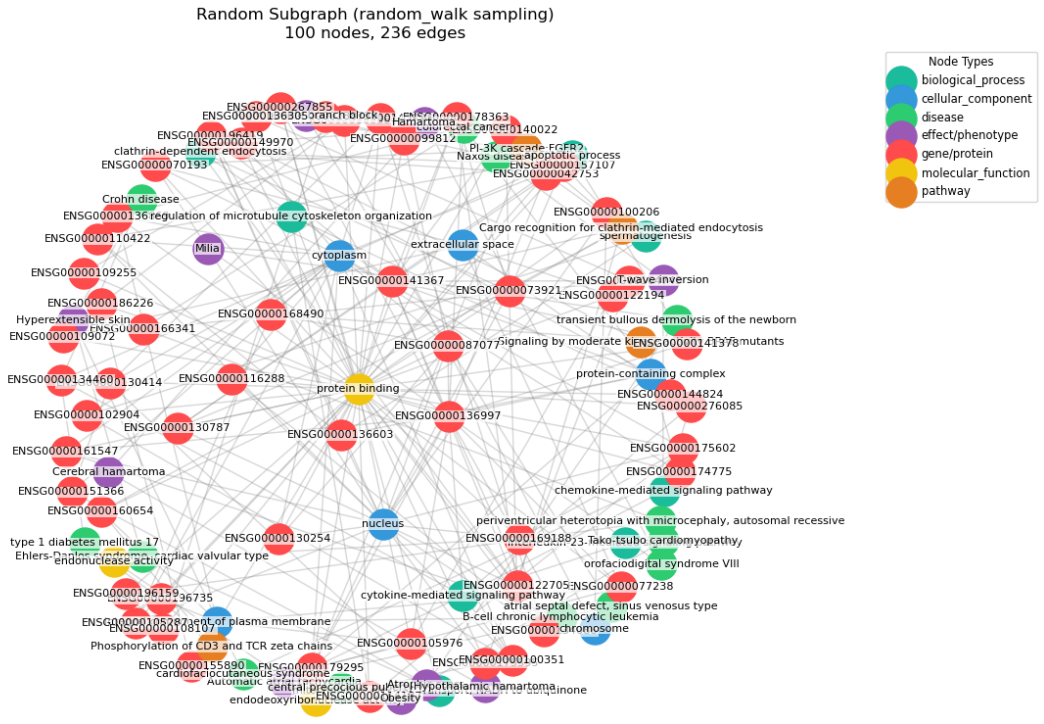

Figure 2: Visualization of Knowledge Graph

3.3 Graph Components

The input to our model is a set of N graphs, $G = \{G_1, G_2, ..., G_N\}$, where each graph $G_i = (V_i, E_i, X_i)$ represents a patient case with V_i and E_i being vertices and edges of the graph, and $X_i \in \mathbb{R}^{|V_i| \times D}$ being the node feature matrix with D representing the number of features.

3.4 Node-level Module(F1)

The node level module is designed to process the individual patient graphs. Mathematically, F1 function is defined as $h_i = F_1(G_i)$, where $h = [h_1, \ldots, h_N]$ are output graph representations in an H dimensional latent space with $h_i \in \mathbb{R}^{1 \times H}$

F1 implements graph convolutional network (GCN) layers. These layers aggregate information from neighboring nodes to generate enriched representations that capture both the phenotypic features and their relationships, subsequently updating each node's representation. Following the GCN layers, a global mean pooling operation aggregates the node-level features to obtain a single embedding vector for each patient.

3.5 Population-level Module(F2)

The population-level module learns the latent connections between patients based on their embedding representations from F1. Formally, F2 is defined as $A_p = F_2(h)$ where $A_p \in (0, 1)^{N \times N}$ is the weighted adjacency matrix encoding the relationships between patients.

To construct this population-level graph, we first project each patient's representation h_i into a latent space using a multi-layer perceptron function g:

$$\tilde{h}_i = g(h_i)$$

We then calculate the cosine similarity between each pair of patient representations.

$$\text{sim}(\tilde{h}_i, \tilde{h}_j) = \frac{\tilde{h}_i \cdot \tilde{h}_j}{\|\tilde{h}_i\| \|\tilde{h}_j\|}$$

Based on the similarity scores, the edge weight a_{ij} is computed as

$$a_{ij} = \frac{1}{1 + e^{t\|\tilde{h}_i - \tilde{h}_j\|^2 - \theta_1}}$$

To build meaningful connections between patients, we employ two key parameters: temperature(t) and threshold(θ). Temperature(t) serves as a scaling factor for the similarity metric, higher temperature values will scale up the similarity value, allow easier connections between patients. Threshold defines the decision boundary for establishing connections, only similarity values exceeding the threshold result in meaningful connections, a higher threshold values enforce more stringent criteria for connection formation, which means less connections.

3.5.1 Node Degree Distribution Loss (NDDL)

In the F2 module, beyond learning the adjacency matrix based on similarity, we implement a specialized loss function to improve the quality of connections. When optimizing solely with classification loss, the population graph can become either overly dense or sparse, which diminishes its utility for capturing meaningful patient relationships. We utilized a Node Degree Distribution Loss (NDDL) that regularizes the connections. This loss function applies Kullback-Leibler divergence between the computed degree distribution over the adjacency matrix A_p and a target Gaussian distribution.

Mathematically, NDDL is defined as the Kullback–Leibler (KL) divergence between the empirical degree distribution of the learned adjacency matrix A_p and a target Gaussian distribution:

$$\text{NDDL} = D_{\text{KL}}(q \,\|\, r) = \sum_i q(i) \log \frac{q(i)}{r(i)},$$

where:

- $q(i)$ is the empirical distribution of node degrees in the learned population graph, normalized to form a valid probability distribution;
- $r(i)$ is the target Gaussian distribution with mean μ and standard deviation σ, treated as hyperparameters.

The empirical degree distribution q is computed from the weighted adjacency matrix A_p as:

$$\text{degree}(i) = \sum_{j=1}^{N} A_p(i, j).$$

We adopt a Gaussian target distribution motivated by empirical observations in biomedical networks, which typically exhibit moderate rather than extreme connectivity. By regularizing toward this target, NDDL promotes:

- Prevention of overly dense connections that could dilute meaningful relationships;
- Avoidance of isolated nodes that cannot benefit from population-level knowledge transfer.

3.6 Classifier Module(F3)

The F3 Classifier Module integrates the outputs from both the node-level module (F1) and the population-level module (F2) to predict the gene responsible for the patient's condition. Formally, F3 is defined as $p = F_3(h, A_p)$, where p represents the probability distribution over gene classes for each patient.

Our implementation consists of several components:

- We first augment each patient's representation from F_1 with gene embeddings:

  $$h'_i = [h_i, e_g]$$

  where e_g is a learnable embedding vector for each gene class.

- We then apply graph convolutions on the population-level graph (from F_2):

$$h^{(l+1)} = \text{GNN}(h^{(l)}, A_p)$$

  followed by layer normalization, ReLU activation, and dropout.

- After GNN processing, a global mean pooling operation aggregates node features to obtain graph-level representations:

$$z_i = \text{GlobalMeanPool}(\{h_j \mid j \in \text{batch}_i\})$$

- When the population adjacency matrix A_p is provided, we apply softmax normalization and refine the embeddings:

$$z_i' = \sum_{j=1}^{N} \text{softmax}(A_p)_{ij} \cdot z_j$$

- The final classification layer maps these processed representations to gene class probabilities using the softmax activation function:

$$p_i = \text{softmax}(W \cdot z_i' + b)$$

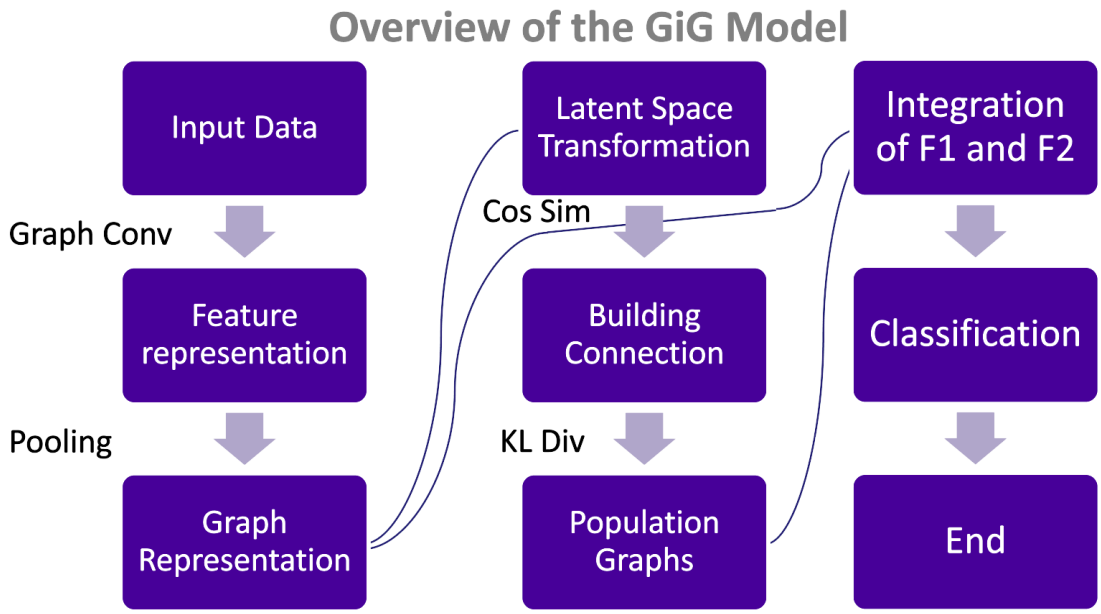

Figure 3: Workflow of GiG Model

3.7 Loss funtion

The final loss function combines the NDDL regularization term with the standard cross-entropy loss for classification:

$$\text{loss} = \text{CE}_{\text{loss}} + \alpha \cdot \text{NDDL}$$

Where α is a hyperparameter controlling the contribution of the regularization term.

4 Introduction of Dataset and Data Preprocessing

4.1 Introduction of Shepherd Dataset

This study leverages the Shepherd Dataset, a valuable resource from the Harvard Dataverse, tailored for rare disease detection research. The dataset comprises phenotype-genotype data for patients

with rare diseases, featuring clinical phenotypes encoded as Human Phenotype Ontology (HPO) terms—standardized descriptions of symptoms and characteristics.

There are more than 40 thousand simulated patient data instances with 2,405 unique true gene IDs. The dataset exhibits significant class imbalance, with some gene IDs appearing only once or twice in the entire training dataset, which reflects the real-world scenario of rare disease occurrence.

4.2 Data Preprocessing

Our reprocessing pipeline consists of several key steps:

- **Label Transformation**: the dataset contains 2405 different true gene ids, therefore we have turn discrete 2405 different classes into continuous labels $N \in [0, 2404]$.
- **Patient Graph Construction**: We utilized existing knowledge graphs to construct individual patient subgraphs, where the knowledge graph provided connection information for each patient's phenotypic features.
- **Embedding Approach**: We utilized specialized embedding techniques to transform the HPO terms into dense vector representations that capture semantic relationships between clinical phenotypes.

5 Results

6 Discussion

Through our experiments, we achieved promising results in classifying rare genetic diseases despite data limitations and computational constraints. In this section, we discuss key challenges encountered during implementation and how they informed our methodological decisions.

6.1 Model Capacity and Architecture Design

The classification task presented significant challenges due to the high-dimensional nature of the data, comprising over 70 phenotypes per patient across more than 40,000 patients and 2,405 distinct gene classes. Our initial architecture employed a relatively simple design: a shallow Multi-Layer Perceptron (MLP) with a single layer (128 hidden units and 128 output units) and a minimal Graph Convolutional Network (GCN) with only one layer (32 hidden units). This limited configuration failed to capture the complex patient aggregated information, resulting in poor convergence during training and restricted generalization capability. The model exhibited persistently high training loss and near-random classification accuracy, indicating its inability to learn meaningful representations from the complex data structure.

Through systematic experimentation, we determined that significantly increasing model capacity was essential for addressing the complexity of rare disease classification. We expanded the GCN component to incorporate deeper architectures (3 layers) with wider hidden dimensions (256 units per layer), enabling more effective aggregation and transformation of neighborhood information within patient phenotype graphs. Similarly, we enhanced the MLP components in the population-level module with wider layers (256 units), allowing for more expressive similarity calculations between patient embeddings.

Unfortunately, due to computational resource limitations on our local hardware, we were constrained in our ability to further extend the model capacity beyond these parameters. While larger architectures might have yielded additional performance improvements, the enhanced model demonstrated substantial gains over the initial configuration, validating our approach to addressing the data complexity through increased model capacity.

This architectural expansion substantially improved the model's ability to learn meaningful patterns. In particular, we observed that:

- Increased GNN hidden dimensions were crucial for capturing the complex relationships between phenotypes, with performance metrics improving significantly as we scaled from 32 to 256 units

- Additional MLP layers enhanced the model's capacity to learn non-linear transformations between the node-level and population-level representations, enabling more sophisticated patient similarity calculations
- The combination of deeper GCN layers and wider MLP components created synergistic effects, with the greatest performance improvements observed when both were scaled concurrently

These architectural insights highlight the importance of sufficient model capacity when dealing with rare disease classification tasks that involve complex, heterogeneous phenotypic data spanning thousands of potential genetic causes.

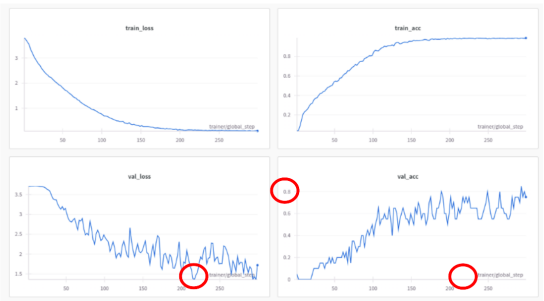

Figure 4: Three layers MLP Projection

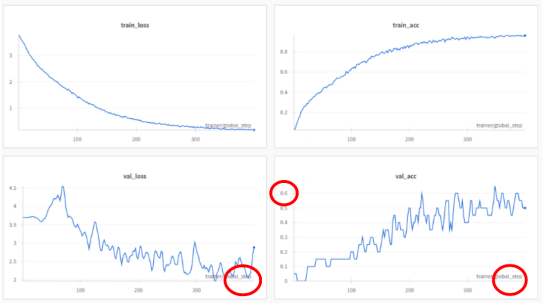

Figure 5: Single layer MLP Project

Figure 6: *
Comparison of training metrics between model architectures. Left: Initial model with limited capacity. Right: Enhanced model with increased hidden dimensions and deeper layers.

6.2 Computational Constraints and Scalability

Due to computational limitations, our presented results are based on a subset of 10,000 patients rather than the full 105,220 training samples. This constraint reflects a common challenge in graph neural network applications, where memory requirements scale with graph size and connectivity. Despite this limitation, our experiments demonstrate the feasibility and potential of the GiG approach, and the results obtained from the subset provide valuable insights into the model's behavior.

Future implementations could address this challenge through more memory-efficient graph representations or distributed computing approaches, potentially enabling analysis of the complete dataset.

6.3 Dataset Representativeness for Rare Disease Classification

Given the nature of rare diseases, we initially questioned whether patient phenotypes would be sufficiently representative of their underlying genetic causes. This concern is particularly relevant in rare disease contexts, where the genotype-phenotype relationship may be complex and incompletely understood.

To address this concern, we conducted two validation analyses. First, we analyzed cosine similarities between embeddings of patients with the same gene ID, revealing clusters of meaningful phenotypic patterns that supported the feasibility of our classification approach. Second, we examined prior research using this dataset, confirming its suitability for our downstream genetic classification task.

6.4 Feature Selection for Patient Data

The original data representation presented significant challenges for model learning. It included only categorical phenotype identifiers (genes, diseases, etc.) without meaningful semantic representation of patient conditions. With this limited representation, the model could not effectively distinguish between patients with different diseases or identify patterns in gene manifestations. Consequently, initial training attempts resulted in persistently high loss values and near-zero accuracy, indicating the model's inability to learn meaningful patterns from the raw categorical data.

To address this fundamental challenge, we implemented a learnable embedding layer for phenotype IDs. This embedding approach effectively transformed discrete phenotype IDs into continuous vector representations that captured semantic relationships between clinically related phenotypes. These embedding vectors functioned as a dynamic lookup table that could be refined during training, significantly enhancing the graph convolution's ability to aggregate and process biologically relevant information.

The effectiveness of this approach was demonstrated by substantial improvements in model performance metrics. After implementing the embedding layer, training loss decreased, while classification accuracy improved across the test set. This marked improvement indicates that the learned embeddings successfully captured biological relationships not explicitly encoded in the original data representation, enabling the model to identify meaningful patterns across phenotypically diverse presentations of the same genetic conditions.

7 Conclusion

7.1 Problem Addressed

This research presented a Graph-in-Graph (GiG) framework for rare disease detection, addressing the challenge of classifying 2,405 distinct gene classes associated with rare genetic conditions. Our approach successfully overcomes fundamental data limitation challenges inherent in rare disease detection through a two-level graph representation approach. The GiG framework demonstrated several key advantages:

- **Integration of biomedical knowledge**: By leveraging established knowledge graphs, our approach incorporated valuable domain expertise that enhanced phenotype connections within patient subgraphs.
- **Patient similarity for knowledge transfer**: The population-level module enabled knowledge transfer between patients with similar phenotypic presentations, effectively expanding the available training signal beyond individual examples.
- **Interpretable disease patterns**: The connections established between patients revealed meaningful patterns across rare diseases, potentially uncovering subtle similarities in their profiles that might remain undetected through conventional analysis methods.
- **Effective rare disease classification**: Despite class imbalance challenges, our model demonstrated the capability to classify rare genetic conditions by learning meaningful representations of patient phenotypes.

The hierarchical approach—from phenotypes to patients to population-level connections—provides both predictive power and interpretability, two critical requirements for clinical applications in rare disease diagnosis.

7.2 Future Work

Several promising directions for future research emerge from this work. First, incorporating multimodal data, such as medical images or clinical summaries, could enhance diagnostic accuracy by capturing phenotypic manifestations not represented in structured HPO terms. Second, addressing the computational challenges posed by the high model capacity required for this complex data remains an important area for improvement. Future iterations could implement more efficient architectures to mitigate model complexity while maintaining performance.

Furthermore, extending the framework to identify potential therapeutic targets based on similarities in genetic and phenotypic profiles represents a valuable application. By integrating molecular pathway information, the model could potentially bridge the gap between diagnosis and treatment recommendation.

8 Reference

Alsentzer, E., Li, M. M., Kobren, S. N., Noori, A., Undiagnosed Diseases Network, Kohane, I. S., & Zitnik, M. (2023). Few-shot learning for phenotype-driven diagnosis of rare genetic diseases. *Nature Medicine*, *29*(10), 2504–2514. https://doi.org/10.1038/s41591-023-02504-3

Mullakaeva, K., Cosmo, L., Kazi, A., Ahmadi, S. A., Bronstein, M. M., & Navab, N. (2023). Graph-in-Graph (GiG): Learning interpretable latent graphs in non-Euclidean domain for biological and healthcare applications. *Medical Image Analysis*, *88*, 102839. https://doi.org/10.1016/j.media.2023.102839