# Exploring Uncertainty in Knowledge Graphs for Drug Indication Prediction

**Bachelor Thesis Exposé**

Tom Frey

2025-10-05

## Contents

# 1 Introduction/Motivation

The development of novel drugs is associated with high costs and long development timelines, while many diseases remain without approved therapies. Drug repurposing, the use of already approved drugs for new therapeutic purposes, offers an alternative with reduced development costs and shorter approval times. Combined with computational and bioinformatics approaches, this enables the systematic analysis of large biomedical datasets, but require the integration of multiple heterogeneous data sources.

Knowledge graphs integrate heterogeneous biomedical data sources into unified graph structures, enabling computational drug repurposing through interpretable entity relationships. Graph based machine learning methods can then analyze these structures to predict novel drug-disease associations.

However, it remains unclear which types of connections in such large scale knowledge graphs are actually relevant for predicting drug-disease associations. Current approaches integrate all available data assuming more data leads to better predictions but leave unexplored if some edge types contribute noise rather than signal.

# 2 Background/Related Work

Heterogeneous biomedical data represent a central challenge in biomedical computer science. As Sujansky [1, p. 285] notes, databases are "highly heterogeneous with respect to the data models they employ, the data schemas they specify, the query languages they support, and the terminologies they recognize." The heterogeneity spans both patient-level clinical information and molecular genomic information [1] and can be classified into two categories [2]. Structural heterogeneity is the difference in database architecture and organizational frameworks [2]. This is also referenced as representational heterogeneity [1]. Semantic, or sometimes terminological, heterogeneity involves inconsistencies in naming conventions and coding systems (terminologies) [2]. This lack of standardized disease representations and multimodal nature of datasets makes it challenging to harmonize biomedical knowledge [1], [3].

Given the resource intensive nature of traditional drug development, using previously approved drugs to treat diseases other than those initially intended (drug repurposing) has emerged as an alternative [4]. This strategy reduces both financial investments and development timelines, particularly when supported by computational methods that systematically analyze biomedical data [4]. These computational methods require integrating multiple heterogeneous biomedical data sources [3]. These data sources include, for example, drug databases, disease ontologies, gene expression repositories, protein interaction networks and clinical knowledge sources [3]. The integration process is complicated due to fundamental differences in database architectures and naming conventions [1]. Additionally, relevant medical knowledge itself exists in a fragmented state across multiple organizational scales and is distributed among diverse sources [3]. These structural and representational challenges limit the ability to identify promising drug-disease associations for drug repurposing [4].

Knowledge graphs are a promising solution to address these integration challenges. Hogan, Blomqvist, Cochez, *et al.* [5, p. 2] define a knowledge graph as "a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities." This representation allows the integration of biomedical entities such as genes, drugs, diseases and exposures through relationships like "indication", "contraindication" or "off-label use" [3].

The construction of a knowledge graph involves integrating multiple heterogeneous data sources with varying reliability. Besides the varying reliability, the variety of terminologies and semantic representations used across biomedical databases is a big challenge [1]. For example identical diseases may be referenced by different naming conventions and therefore make the entity resolution difficult [1].

Knowledge graph projects typically use multiple sources to generate a knowledge graph [5]. For instance Chandak, Huang, and Zitnik [3] integrated diverse biomedical sources into their PrimeKG knowledge graph. These sources include different types of biomedical knowledge, including expert-curated databases, standardized ontologies and experimental measurement repositories [3]. The construction process involves selecting ontologies for each entity type, harmonizing datasets into standardized formats and merging the processed resources into a unified graph structure [3].

The different types of sources integrated contribute varying qualities of knowledge. Chandak, Huang, and Zitnik [3, p. 2] categorizes biomedical data sources into three types: expertly curated annotations (e.g. DisGeNet gene-disease associations), standardized ontologies (e.g. MONDO Disease Ontology) and direct experimental measurements (e.g. DrugBank drug properties). The quality differs because expert curation can include biases (e.g. geographic or linguistic bias), automated extraction can make mistakes (e.g. imprecise extraction results) and experimental measurements depend on strict methods and thresholds [3], [5]. As a result, not all facts in the knowledge graph are equally trustworthy [5].

Chandak, Huang, and Zitnik [3] constructed PrimeKG by harmonizing data from over 20 different biomedical sources, resulting in a heterogeneous network containing approximately 129000 nodes connected by over 4 million edges. Compared to predecessor graphs like SPOKE, which covered only 137 diseases, or GARDs rare disease focus, PrimeKG expands coverage by 10-100 times while distinguishing between therapeutic applications (approved indications), adverse interactions (contraindication) and exploratory uses (off-label) for drug-disease pairs alongside clinical textual descriptions.

Graph neural networks (GNNs) are defined as "neural models that capture the dependence of graphs via message passing between the nodes of graphs" [6, p. 57]. Unlike knowledge graph embeddings that focus on static representations, GNNs are able to perform specific prediction tasks with supervised learning [5]. GNNs are particularly suitable for biomedical applications as they are able to handle unstructured data types as input [7].

One specific biomedical application is drug repurposing, where GNNs can analyze molecular compounds within graph structures and predict drug-disease associations [4], [5].

Biomedical GNNs face structural challenges from the dominance of protein-protein interaction networks, which cover over 90% of network connections and cause bias toward protein interactions rather than drug-disease associations [4].

Successful biomedical applications include TxGNN, a graph neural network trained on a medical knowledge graph with 17080 diseases and 7957 drugs [8]. TxGNN demonstrates effective prediction of drug repurposing candidates across diverse disease categories [8].

The high dimensionality of biomedical data, where features substantially outnumber samples, leads to overfitting in machine learning models [9]. This is called curse of dimensionality (CoD) and leads to excellent performance on training data but poor results on unseen data [9].

In PrimeKG, protein-protein interactions (642150 edges) outnumber drug-disease indications (18776 edges) [3]. This imbalance introduces systematic bias in graph-based learning algorithms towards protein networks rather than drug-disease associations [4].

Drug repurposing tasks face severe class imbalance, where positive drug-disease associations represent only 0.2% of all edges in biomedical knowledge graphs [3]. Models need to balance the influence of positive and negative samples during training to handle this unequal distribution [10].

Knowledge graph pruning involves the intentional removal of edges from a graph structure to improve predictive performance. In contrast to completion methods, that add missing relationships [5]. For drug prediction tasks, pruning removes known drug-disease associations and evaluates whether models can successfully rediscover those connections through link prediction [5].

Pruning addresses the trade-off between removing edges to improve generalization and preserving sufficient connectivity for accurate predictions [11].

The theoretical foundation for pruning lies in compositional sparsity, where reducing network connections can enhance model generalization by decreasing the number of trainable parameters [11]. Caranzano, Pancotti, Rollo, *et al.* [11] demonstrate that performance improvements in pathway-informed neural networks originate from structural sparsity rather than biological knowledge, as randomized sparse models achieved comparable or superior results. This suggests that systematic edge removal may improve drug indication prediction by optimizing the sparsity level for the specific task rather than relying on predefined biological structures [11].

Since sparsity itself improves performance regardless of biological relevance, indicating that sparsity level and structure matter more than biological meaning [11].

The uncertainty in knowledge graphs can be categorized into three types, each providing a different rationale for systematic edge removal.

Building on the established variation in source quality, pruning can exploid differences in data provenance. Multi-source knowledge graphs contain edges with inherently different reliability levels [5]. Source-based pruning strategies remove edges originating from sources with known limitations, for example automated extraction systems prone to imprecision [5].

Graph topology provides source-independent indicators of edge criticality. Node degree quantifies local connectivity with robust measures like median degree accounting for hub node effects [11]. Centrality measures including betweenness and closeness iden-

tify structurally important nodes whose connections are essential for graph coherence [5]. Topology-based pruning targets low-degree or low-centrality edges as candidates for removal.

Another type is the confidence based uncertainty, where explicit confidence metadata enables direct quantification of edge reliability [12]. Knowledge graphs like NELL and ConceptNet implement numerical confidence scores representing the likelihood of relational assertions [12].

These uncertainty types can be used as strategy to systematically remove edges from a knowledge graph and to evaluate the impact of edge removal on drug indication prediction. This represents a well-defined task as multiple studies employ standardized metrics such as AUROC and AUPR for performance evaluation [4], [10].

Random removal serves as a baseline control strategy that eliminates edges uniformly and provides a performance benchmark for evaluating uncertainty-aware pruning strategies [11].

While source based and topology based pruning provide static strategies for edge removal, dynamic approaches adjust sparsity during model training. Evci, Gale, Menick, *et al.* [13] show that networks with adaptive connectivity during optimization avoid performance limitations of fixed sparse structures by removing minimally contributing connections while activating high gradient connections. Theoretical work on sparse subnetwork discovery [14] and joint graph model optimization suggests that optimal edge selection for knowledge graph based drug prediction may require training driven refinement rather than static analysis alone.

# 3 Research Objectives/Problem Statement

Drug repurposing requires integrating multiple heterogeneous biomedical data sources. Knowledge graphs like PrimeKG enable this integration by representing drugs, diseases, genes and their relationships in a single graph structure. GNNs can then be trained to analyze those relationships and predict novel drug-disease associations for repurposing candidates. However, with knowledge graphs integrating millions of edges from heterogeneous sources, it is unclear, which relationships are relevant and which are noise.

The primary research question is: How does systematic removal of different edge types affect the performance of drug indication prediction in medical knowledge graphs? The goal of this work is to quantify the relevance of different edge types in medical knowledge graphs like PrimeKG for the specific task of drug indication prediction. This work will explore different strategies for edge removal and measure their impact on prediction performance to identify which connections are critical for accurate drug repurposing predictions.

Primary objective: Develop a systematic framework for edge relevance quantification and pruning in medical knowledge graphs using all available diseases in the dataset.

Secondary objective: Quantify the impact of different edge types in PrimeKG on drug indication perfomance, compare pruning strategies (random baseline, source based, topology based), establish optimal trade-offs between graph size and prediction accuracy.

The evaluation will use Graph Neural Networks for link prediction and shortest path algorithms as fallback. Performance will be evaluated using AUROC and precision-recall metrics by comparing predictions on the pruned graph variants to predictions on the complete graph.

# 4 Methodology/Approach/Evaluation

## 4.1 Phase 1: Graph Setup and Exploration

The first step involves setting up the environment with PyTorch Geometric (PyG), loading the Knowledge Graph and exploring its structure. It is unclear which resources are needed to work with the PrimeKG knowledge graph. Therefore it is possible that the graph is too large to be loaded into memory on a personal computer and it requires a more powerful setup like a compute server (e.g. computeserver@HU) or a cluster (e.g. HPC@HU, Charité Cluster). In case the PrimeKG graph is still too large or too complicated to be worked with, there are several other smaller medical knowledge graphs available, for example Hetionet [15] or OREGANO [16].

After the graph is loaded, the structure of the knowledge graph will be explored to understand the graph topology, entity/relationship distributions and domain specific medical terminology and relations. This exploration will include network measures such as node degree distributions, centrality measures (betweenness, closeness), path lengths between drug-disease pairs. The gained insights will be crucial for the next steps as a foundation for the selection of an appropriate pruning method and evaluation metrics.

## 4.2 Phase 2: Baseline Evaluation Framework

A baseline evaluation framework will be established to compare the performance of different pruning methods. The evaluation will focus on the drug indication prediction task, which involves predicting drug-disease associations that have received regulatory approval by agencies such as FDA or EMA. These regulatory approved associations serve as ground truth for evaluation. The task systematically removes known drug-indication edges and trains a GNN to rediscover them through link prediction, which simultaneously generates numerical embeddings that encode biomedical relationships.

This will evaluate to what extent predictions remain possible when using different pruning strategies that remove certain types of nodes and edges. The performance will be measured with standard metrics such as AUROC, precision and recall. A set of fixed random seeds will be used to ensure that performance differences are not due to random initialization effects.

If this approach is too difficult to implement, an alternative would be to use tree-based models by converting the network into a tabular data set or use shortest path algorithms as a simpler evaluation method.

The baseline performance on the complete knowledge graph will be used as reference for measuring the impact of pruning in the next phase.

### 4.3 Phase 3: Systematic Pruning

This phase will involve systematically removing edges from the knowledge graph using different strategies: random removal (baseline uncertainty), confidence based pruning using PrimeKG metadata (data source uncertainty) and structural pruning based on graph topology (topological uncertainty). For each of these strategies a subset of edges (e.g. 5%, 10%, 15%, 20%) will be removed and the impact on the drug indication prediction performance will be measured. Additionally a fixed set of true-positive drug-indication edges will be held out as test data using cross-validation or bootstrap sampling to be able to obtain measures of uncertainty. This will identify which connections are most critical for predictive accuracy and which contribute to uncertainty and noise in the model.

## References


[1] W. Sujansky, "Heterogeneous database integration in biomedicine," *Journal of Biomedical Informatics*, vol. 34, no. 4, pp. 285–298, 2001, ISSN: 1532-0464. DOI: https://doi.org/10.1006/jbin.2001.1024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S153204640191024X.

[2] J.-F. Ethier, O. Dameron, V. Curcin, *et al.*, "A unified structural/terminological interoperability framework based on lexevs: Application to transform," *Journal of the American Medical Informatics Association*, vol. 20, no. 5, pp. 986–994, Sep. 2013, ISSN: 1527-974X. DOI: 10.1136/amiajnl-2012-001312. [Online]. Available: http://dx.doi.org/10.1136/amiajnl-2012-001312.

[3] P. Chandak, K. Huang, and M. Zitnik, "Building a knowledge graph to enable precision medicine," *Scientific Data*, vol. 10, no. 1, Feb. 2023, ISSN: 2052-4463. DOI: 10.1038/s41597-023-01960-3. [Online]. Available: http://dx.doi.org/10.1038/s41597-023-01960-3.

[4] D. Bang, S. Lim, S. Lee, and S. Kim, "Biomedical knowledge graph learning for drug repurposing by extending guilt-by-association to multiple layers," *Nature Communications*, vol. 14, no. 1, Jun. 2023, ISSN: 2041-1723. DOI: 10.1038/s41467-023-39301-y. [Online]. Available: http://dx.doi.org/10.1038/s41467-023-39301-y.

[5] A. Hogan, E. Blomqvist, M. Cochez, *et al.*, "Knowledge graphs," *ACM Computing Surveys*, vol. 54, no. 4, pp. 1–37, Jul. 2021, ISSN: 1557-7341. DOI: 10.1145/3447772. [Online]. Available: http://dx.doi.org/10.1145/3447772.

[6] J. Zhou, G. Cui, S. Hu, *et al.*, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020, ISSN: 2666-6510. DOI: https://doi.org/10.1016/j.aiopen.2021.01.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666651021000012.

[7] I. R. Ward, J. Joyner, C. Lickfold, Y. Guo, and M. Bennamoun, *A practical tutorial on graph neural networks*, 2021. arXiv: 2010.05234 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2010.05234.

[8] K. Huang, P. Chandak, Q. Wang, *et al.*, "A foundation model for clinician-centered drug repurposing," *Nature Medicine*, vol. 30, no. 12, pp. 3601–3613, Sep. 2024, ISSN: 1546-170X. DOI: 10.1038/s41591-024-03233-x. [Online]. Available: http://dx.doi.org/10.1038/s41591-024-03233-x.

[9] M. Martínez-García and E. Hernández-Lemus, "Data integration challenges for machine learning in precision medicine," en, *Front. Med. (Lausanne)*, vol. 8, p. 784 455, 2021.

[10] W. Li, W. Ma, M. Yang, and X. Tang, "Drug repurposing based on the dtd-gnn graph neural network: Revealing the relationships among drugs, targets and diseases," *BMC Genomics*, vol. 25, no. 1, Jun. 2024, ISSN: 1471-2164. DOI: 10.1186/s12864-024-10499-5. [Online]. Available: http://dx.doi.org/10.1186/s12864-024-10499-5.

[11] I. Caranzano, C. Pancotti, C. Rollo, *et al.*, *Sparsity is all you need: Rethinking biological pathway-informed approaches in deep learning*, 2025. arXiv: 2505.04300 [q-bio.QM]. [Online]. Available: https://arxiv.org/abs/2505.04300.

[12] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 494–514, Feb. 2022, ISSN: 2162-2388. DOI: 10.1109/tnnls.2021.3070843. [Online]. Available: http://dx.doi.org/10.1109/TNNLS.2021.3070843.

[13] U. Evci, T. Gale, J. Menick, P. S. Castro, and E. Elsen, *Rigging the lottery: Making all tickets winners*, 2021. arXiv: 1911.11134 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1911.11134.

[14] Y.-D. Sui, X. Wang, T. Chen, M. Wang, X.-N. He, and T.-S. Chua, "Inductive lottery ticket learning for graph neural networks," *Journal of Computer Science and Technology*, vol. 39, no. 6, pp. 1223–1237, Nov. 2024, ISSN: 1860-4749. DOI: 10.1007/s11390-023-2583-5. [Online]. Available: http://dx.doi.org/10.1007/s11390-023-2583-5.

[15] D. S. Himmelstein, A. Lizee, C. Hessler, *et al.*, "Systematic integration of biomedical knowledge prioritizes drugs for repurposing," *eLife*, vol. 6, A. Valencia, Ed., e26726, Sep. 2017, ISSN: 2050-084X. DOI: 10.7554/eLife.26726. [Online]. Available: https://doi.org/10.7554/eLife.26726.

[16] M. Boudin, G. Diallo, M. Drancé, and F. Mougin, "The oregano knowledge graph for computational drug repurposing," *Scientific Data*, vol. 10, no. 1, Dec. 2023, ISSN: 2052-4463. DOI: 10.1038/s41597-023-02757-0. [Online]. Available: http://dx.doi.org/10.1038/s41597-023-02757-0.