

RESEARCH PAPER

MPRDR: A Multi-Path Relational Drug Repurposing Framework Grounded in Graph-Theoretic Principles

Pharsana Parveen M^{1*}, Stanis Arul Mary A²^{1*}Research Scholar, PG and Research Department of Mathematics,

Nirmala College for Women, Coimbatore, Tamil Nadu, India.

Email: pharsanaparveen@gmail.com

²Assistant Professor, PG and Research Department of Mathematics,

Nirmala College for Women, Coimbatore, Tamil Nadu, India

ABSTRACT

Current GNN-based drug repurposing algorithms rate drug-disease relationships using learned embedding proximity, which identifies the patterns of co-occurrence but is neither mechanistically interpretable nor cold-start. We introduce MPRDR (Multi-Path Relational Drug Repurposing), a GNN architecture based on three graph theoretical concepts: typed path algebra [1] to support multi-hop mechanistic evidence, Turan theorems cohesion of drug modules [2], and effective resistance structural robustness [3,4]. In contrast to TxGNN [5] and GDRnet [6], MPRDR precomputes three interpretable scores of topology of PrimeKG [7] and trains a GNN to combine these structural prioritizations with learned representations and replacing uncalibrated sigmoid outputs with a graph-theoretic confidence score. MPRDR performs competitively or better than previous experiments when evaluated using well-known standard test sets such as TxGNN and GDRnet. MPRDR demonstrates additional features compared to existing databases by being able to predict cold-start results, generate drug combination synergy scores, and provide mechanistic explanations.

Keywords: Drug Repurposing, Knowledge Graph, Graph Neural Network, Typed Path Algebra, Heterogeneous Graph, Drug Module.

How to cite this article: Parveen PM, Mary ASA. MPRDR: A Multi-Path Relational Drug Repurposing Framework Grounded in Graph-Theoretic Principles. *Int J Drug Deliv Technol.* 2026;16(17s): 416-425. DOI: 10.25258/ijddt.16.17s.48

1. INTRODUCTION

The drug discovery process takes an average of 10-15 years and more than one billion US dollars to allow one compound to receive approval to the market [8]. Drug repurposing helps to avoid much of this cost because the identified new therapeutic indication of the approved drug can share much of the safety profile to shorten the time to clinic. A number of successful repurposing blockbusters such as sildenafil in the treatment of pulmonary arterial hypertension, thalidomide in the treatment of multiple myeloma, and metformin in the treatment of polycystic ovary syndrome indicate how revolutionary systematic computational repurposing can be [9]. It has become the preferred data structure of the task, and drugs, diseases, genes, proteins, pathways, and phenotypes are represented as nodes with typed interaction edges in biomedical knowledge graphs (KGs), where drug repurposing can be viewed as a link prediction problem on a heterogeneous graph [7].

New methods KG-based repurposing Graph neural networks (GNNs) have become the state-of-the-art. Some of these models like TxGNN [5], GDRnet [6], and DTD-GNN [10] also learn node representations by passing messages repeatedly and score pairs of drugs and diseases using decoder functions on embedding pairs. Even though benchmark results are good, these techniques have one basic weakness they use learned embedding proximity to base

their scoring logic and this is not a mechanistic biological reasoning but a more statistical co-occurrence pattern. The result of this is three tangible weaknesses: all current techniques are transductive and do not work with cold-start drugs or diseases; their results are an uncalibrated sigmoid score with no interpretable uncertainty; predictions do not have a mechanistic explanation, which clinicians or researchers can act on.

The common denominator between these limitations is that the KG is being used as a data container in which to pass messages, and the rich mathematical structure of the entity, typed paths, modular community properties and node-level connection robustness, is disregarded. This paper presents a GNN framework called MPRDR, which uses the graph theory as the real scoring logic. MPRDR is constructed on three principles: (i) typed path algebra: drug-disease association is supported by a mechanistic association on many distinctly high-confidence typed paths between drug and disease using biological meaningful intermediates; (ii) Turan module cohesion: drugs in the same structurally dense module as known treatments are better candidates, quantified by Turan density [2]; and (iii) effective resistance: a drug disease association supported by many parallel typed paths is structurally robust. Instead of using integer scores to characterize the GNN training, MPRDR uses three interpretable scores computed by the KG topology prior to GNN training, where the GNN is

commanded to mix these structural priors with learned representation, and generates calibrated confidence scores and mechanistic explanations as byproducts.

The principal contributions are: (i) a graph-theoretic precomputation pipeline, which derives three interpretable scores, namely path score, module score and bridge centrality score, directly out of the KG; (ii) a GNN architecture, the attention and aggregation of which is explicitly conditioned on these structural priors; (iii) a deterministic graph-theoretic confidence score, which replaces uncalibrated sigmoid outputs; and (iv) per-prediction mechanistic explanations, which identify the top biological paths, drug module membership, and bridge nodes. Competitive or superior performance compared to TxGNN [5], GDRnet [6], and COVID-19 GNN repurposing baselines is shown in experiments on PrimeKG [7], and it has unique cold-start prediction and confidence calibration capabilities.

2. RELATED WORK

Construction of Knowledge graph to Drug Repurposing

The enabling development of computational repurposing has been the construction of large-scale heterogeneous biomedical KGs. The DRKG [12] unites seven biomedical databases into a single graph containing more than 5.8 million edges between 97,238 entities. PrimeKG [7], proposed by Chandak et al., contains 10 sources with 17,080 diseases, 4,050 drugs, and 27,671 genes, and is appropriately aimed at multi-disease drug repurposing evaluation. These sources have offered the data structure of GNN-based repurposing models but have also brought about the transgression of heterogeneous edge semantics - a drug-target inhibition edge has radically different biological intentions to a gene disease association edge, yet most GNN structure is structurally indistinguishable between the two.

Core Repurposing Baselines.

TxGNN: The closest previous work to MPRDR was suggested by Huang et al. [5] in the article published in Nature Medicine in 2024, which was called TxGNN. TxGNN is based on a roll of PrimeKG with a relational GCN encoder and a disease metric learning module, which allows zero-shot repurposing of diseases that have no known training relationships. The model is the first repurposing model to explicitly solve the zero-shot disease generalisation problem, and state-of-the-art in the standard PrimeKG benchmarks. Nevertheless, the scoring decoder of TxGNN is a trained dot product between embedding of drugs and diseases, with no structural property grounded on graphs. The confidence out is an uncalibrated sigmoid measure, and no mechanistic path account is produced. MPRDR answers the questions of these gaps and still provides zero-shot generalisability with the path-based and module-based scoring systems.

GDRnet: Published in Computers in Biology and Medicine in 2022, Doshi and Chepuri [6] suggested GDRnet, which is computationally efficient and GNN used to repurpose drugs in the DRKG knowledge graph. GDRnet builds a four-layer heterogeneous graph on drugs, genes, diseases

and anatomical entities and uses the SIGN precomputation framework which pre-computes powers of the normalised adjacency matrix and trains the model to decouple feature propagation and model optimisation. GDRnet has ability to compute with high efficiency and has a high performance with a wide range of diseases. Its drawback, applicable to MPRDR, is that the aggregation of SIGN is a fixed-polynomial filter that lacks awareness of relation-type and lacks path semantics and the quadratic decoder lacks structural robustness or membership to the module.

COVID-19 GNN Repurposing

In 2021, Hsieh et al. [11] presented a COVID-19-specific repurposing pipeline built on CTD base in the form of a COVID-19 knowledge graph consisting of 27 SARS-CoV-2 bait proteins, 3,635 candidate drugs and of the order of 33000 edges, which is published in Scientific Reports. The model uses a multi-relational variational graph autoencoder (VGAE) encoder, which is initialised using DRKG encoders, and a neural ranker that is trained with Bayesian Pairwise Ranking loss on the silver-standard labels of clinical trial drugs. Its three-stream validation, including gene expression signature matching in GSEA, in vitro experimental efficacy and population-based treatment effect of large-scale EHR data, is a unique characteristic. The best thing about the work is this multi-evidence validation strategy. Its shortcomings are that it builds KGs specific to COVID-19 and cannot be generalised to other diseases and provides a binary ranking output lacking confidence calibration and mechanistic path elucidation. MPRDR builds on it and takes the multi-stream concept of validation as its time and cold-start evaluation protocols inspiration.

Heterogeneous GNN Architectures.

The encoding of relations-specific weight matrices associated with heterogeneous KGs was proposed by Relational GCN (RGCN) [13] and is the core of various repurposing models such as TxGNN. Heterogeneous information networks are applied in the meta-path-based attention aggregation in HAN [14]. HGT [15] generalises this to transformer-like attention on nodes and edges. DTD-GNN [10] suggests ternary drug-target-disease event nodes, which combine GCN and GAT encoders to predict links on the BioSNAP dataset. Although such architectures represent heterogeneous graph semantics, none of them considers path count, path diversity, or path reliability in their scoring; no one of them applies Turan-type extremal graph theory to the module cohesion; and no one of them calculates the effective resistance to provide measures of structural resilience of drug-disease relationships. MPRDR extends the relation-awareness of message passing of RGCN and the attention of HAN and HGT, but introduces the three graph-theoretic scoring terms as structural priors, which the GNN does not override during training.

KG Embedding Baselines

Knowledge graph embedding models These include TransE [16], RotatE [17], and DistMult [18] models that are trained to optimise translational or bilinear scoring functions on observed triples. These techniques are classical lower-

bound baselines on KG-based repurposing and featured in our experiment. Their inherent weakness is that they are a modeling of pairwise relational structure, thus lacking multi-hop path inference and modular community structure as well as the ability to compute resistance. They are strong transductive and do not start at all in cold-start environments.

3. METHODOLOGY

In this section, we introduce MPRDR (Multi-Path Relational Drug Repurposing), a graph-theoretic framework that establishes drug repurposing prediction based on three foundational principles from graph theory: typed path algebra for mechanistic evidence, Turán's theorem from extremal graph theory for module cohesion, and Kirchhoff's effective resistance from electrical network theory for structural robustness. MPRDR is different from other methods that only use learned embedding proximity to score drug-disease associations. Instead, it uses the structure of the biomedical knowledge graph to precompute combinatorially interpretable scores. Then, it trains a graph neural network (GNN) to learn how to combine these scores with learned representations. There are six phases in the full framework, which are explained in detail below.

Notation and Problem Formulation

The biomedical knowledge graph is defined as a weighted heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R}, \mathcal{W})$, where $\mathcal{V} = \mathcal{V}_d \cup \mathcal{V}_D \cup \mathcal{V}_g \cup \mathcal{V}_p \cup \mathcal{V}_\phi$ is the union of node sets representing drugs, diseases, genes, pathways, and phenotypes respectively. \mathcal{R} denotes the set of relation types, $\mathcal{W}: \mathcal{E} \rightarrow [0,1]$ is a confidence weighting function over edges, and $\mathbf{A}_r \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$ is the adjacency matrix for relation type $r \in \mathcal{R}$. We denote by \mathbf{L} the weighted graph Laplacian and by \mathbf{L}^\dagger its Moore-Penrose pseudoinverse. For a given disease D , $\mathcal{N}(D)$ denotes the set of drugs known to treat D . The set $\mathcal{P}_\tau(d, D)$ denotes all paths of meta-path type τ from drug node d to disease node D , and $\mathcal{T} = \{\tau_1, \dots, \tau_K\}$ is the set of valid therapeutic meta-path types defined over the KG schema.

Drug Repurposing as Weighted Link Prediction

Given \mathcal{G} , a set of known drug-disease treatment pairs $\Omega_{\text{pos}} \subset \mathcal{V}_d \times \mathcal{V}_D$, and a target disease $D^* \in \mathcal{V}_D$, rank all candidate drugs $d \in \mathcal{V}_d \setminus \mathcal{N}(D^*)$ by a graph-theoretically grounded score $\text{score}(d, D^*)$, such that drugs with genuine therapeutic potential for D^* are ranked highest, together with a calibrated confidence value and an interpretable

mechanistic explanation.

Phase 1: Knowledge Graph Construction

The foundation of MPRDR is a richly annotated weighted heterogeneous knowledge graph constructed from PrimeKG [7], which integrates ten biomedical databases covering 4,050 drugs, 17,080 diseases, 27,671 genes, 2,516 pathways, and 13,299 phenotypes. The construction phase assigns confidence weights to every edge based on the reliability of its source database and the recency of its curation. This produces weighted adjacency matrices that carry evidential quality directly into all downstream graph-theoretic computations.

Edge Confidence Weighting

A critical departure from existing methods, which treat all edges as binary or equally weighted, is our assignment of a continuous confidence weight $\mathcal{W}(u, v, r)$ to every edge $(u, v, r) \in \mathcal{E}$:

$$\mathcal{W}(u, v, r) = w_{\text{source}}(r) \cdot w_{\text{recency}}(u, v) \quad (1)$$

where the source confidence $w_{\text{source}}(r)$ reflects the experimental basis of the relation:

$$w_{\text{source}} = \begin{cases} 1.0 & \text{manual curation (DrugBank, UniProt)} \\ 0.7 & \text{high-throughput experimental} \\ 0.4 & \text{computational prediction} \\ 0.2 & \text{text mining} \end{cases} \quad (2)$$

and the recency weight $w_{\text{recency}}(u, v)$ privileges more recently established associations:

$$w_{\text{recency}}(u, v) = \begin{cases} 1.0 & \text{edge age} \leq 2 \text{ years} \\ 0.8 & \text{edge age} \leq 5 \text{ years} \\ 0.6 & \text{edge age} > 5 \text{ years} \end{cases} \quad (3)$$

Drug-drug structural similarity edges are constructed via the Jaccard coefficient over shared target sets:

$$\mathcal{W}(d_i, d_j) = \text{Jaccard}(\text{targets}(d_i), \text{targets}(d_j)) \quad (4)$$

with edges added when $\mathcal{W}(d_i, d_j) > \theta_{\text{drug}}$, where $\theta_{\text{drug}} = 0.3$. The weighted Laplacian is then defined as $\mathbf{L} = \mathbf{D}_{\mathcal{W}} - \mathbf{A}_{\mathcal{W}}$, where $\mathbf{A}_{\mathcal{W}} = \sum_r w_r \cdot \mathbf{A}_r$ and $\mathbf{D}_{\mathcal{W}}$ is the corresponding diagonal degree matrix.

Algorithm 1: Knowledge Graph Construction

Input: PrimeKG raw data

Output: $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R}, \mathcal{W}), \{\mathbf{A}_r\}_{r \in \mathcal{R}}, \mathbf{L}$

1. **for** each edge (u, v, r) in PrimeKG **do**
2. $\mathcal{W}(u, v, r) = w_{\text{source}}(r) \cdot w_{\text{recency}}(u, v)$ Eq. (1)
3. **end for**
4. **for** each drug pair (d_i, d_j) **do**
5. Compute $\mathcal{W}(d_i, d_j)$ via Eq. (4)
6. **if** $\mathcal{W}(d_i, d_j) > \theta_{\text{drug}}$ **then**

```

7.           Add edge  $(d_i, d_j, drug\_sim)$  to  $\mathcal{E}$ 
8.       end if
9.   end for
10.  For each relation type  $r \in \mathcal{R}$  do
11.       $A_r |u, v| \leftarrow \mathcal{W}(u, v, r)$  if  $(u, v, r) \in \mathcal{E}$ , else 0
12.  end for
13.   $A_w \leftarrow \sum_r w_r \cdot A_r$ 
14.   $L \leftarrow D_w - A_w$ 
15.  return  $\mathcal{G}, \{A_r\}, L$ 

```

Phase 2: Precomputation of Graph-Theoretic Scores

Before any GNN training begins, MPRDR precomputes three graph-theoretic scores directly from the structure of \mathcal{G} . These scores constitute the mathematically grounded prior that guides subsequent learning. Each score addresses a distinct structural property of drug-disease relationships.

$$w_{type}(\tau) = \begin{cases} 1.0 & Drug \rightarrow Target \rightarrow Disease \\ 0.9 & Drug \rightarrow Target \rightarrow Pathway \rightarrow Disease \\ 0.8 & Drug \rightarrow Target \rightarrow Gene \rightarrow Disease \\ 0.7 & Drug \rightarrow Pathway \rightarrow Gene \rightarrow Disease \\ 0.6 & Drug \rightarrow Target \rightarrow Phenotype \rightarrow Disease \\ 0.5 & Drug \rightarrow Phenotype \rightarrow Disease \\ 0.3 & Drug \rightarrow Drug \rightarrow Disease \end{cases} \quad (6)$$

Typed Path Score

Theoretical Foundation: When a drug treats a disease, the biological mechanism is always mediated through intermediate entities such as target proteins, genes, biological pathways, or phenotypes, forming a path through the knowledge graph. We formalize this using the theory of *typed path algebras* on heterogeneous graphs [1]. A meta-path type $\tau = (r_1, r_2, \dots, r_k)$ is a sequence of relation types defining a class of paths from drugs to diseases. The number of paths of type τ between drug d and disease D can be computed exactly as the corresponding entry of the relational adjacency matrix product:

$$M_\tau[d, D] = (A_{r_1} \cdot A_{r_2} \cdots A_{r_k})[d, D] \quad (5)$$

This is a direct consequence of the standard result that the (i, j) -th entry of A^k counts the number of walks of length k between nodes i and j .

Weighted Path Score: Not all paths are equally informative. We assign three layers of weight to differentiate paths by their biological and evidential quality.

The meta-path type weight $w_{type}(\tau)$ reflects the biological directness of the path type:

The path trust weight captures the cumulative reliability of all edges along a path p :

$$w_{trust}(p) = \prod_{e \in p} \mathcal{W}(e) \quad (7)$$

A length penalty discounts longer, less direct paths:

$$w_{len}(p) = \frac{1}{|len(p)|^\alpha}, \alpha \in [0.5, 2.0] \quad (8)$$

where α is a tunable hyperparameter. A diversity bonus rewards evidence from multiple independent biological mechanisms:

$$diversity(d, D) = \frac{|\{\tau \in \mathcal{T}: M_\tau[d, D] > 0\}|}{|\mathcal{T}|} \quad (9)$$

The complete weighted path score is:

$$S_{path}(d, D) = diversity(d, D) \cdot \sum_{\tau \in \mathcal{T}} w_{type}(\tau) \cdot M_\tau[d, D] \cdot \bar{w}_{trust}(\tau, d, D) \cdot \frac{1}{|\tau|^\alpha} \quad (10)$$

where $\bar{w}_{trust}(\tau, d, D)$ is the mean trust weight over all paths of type τ between d and D .

Algorithm 2: Typed Path Score Precomputation

Input: \mathcal{G}, A_r , meta-path types \mathcal{T} , maximum path length $K = 4$

Output: $S_{path}(d, D)$ for all (d, D) pairs

```

1.   Define  $w_{type}(\tau)$  for each  $\tau \in \mathcal{T}$  as per Eq.(6)
2.   for each meta-path type  $\tau = (r_1, \dots, r_k) \in \mathcal{T}$  do
3.        $M_\tau \leftarrow A_{r_1} \cdot A_{r_2} \cdots A_{r_k}$ 
4.       Compute  $w_{trust}(\tau, d, D)$  for all pairs via Eq. (7)
5.   end for
6.   for each drug-disease pair  $(d, D)$  do
7.       Compute  $diversity(d, D)$  via Eq.(9)
8.       Compute  $S_{path}(d, D)$  via Eq.(10)
9.   end for
10.  Normalize:  $S_{path} \leftarrow S_{path} / \max_{d', D'} S_{path}(d', D')$ 
11.  return  $S_{path} \in \mathbb{R}^{|\mathcal{V}_d| \times |\mathcal{V}_D|}$ 

```

Eq. (5)

Turán Module Score

Theoretical Foundation: We exploit Turán's theorem [2]

from extremal graph theory as a principled tool for measuring the cohesion of drug modules in the context of a specific disease. Turán's theorem states that the maximum number of edges in a simple graph on n vertices containing no complete subgraph K_{r+1} is achieved by the Turán graph $T(n, r)$:

$$\text{ex}(n, K_{r+1}) = \left(1 - \frac{1}{r}\right) \frac{n^2}{2} \quad (11)$$

The *Turán density* of a graph G on n vertices is:

$$\tau(G) = \frac{|E(G)|}{\text{ex}(n, K_{r+1})} \quad (12)$$

When $\tau(G) > 1$, the graph is forced to contain a clique of size $r + 1$; when $\tau(G) \leq 1$, the clique structure is not guaranteed.

Disease-Specific Drug Module Construction: For each disease D , we construct a *disease-specific drug similarity graph* \mathcal{G}_{drug}^D over the known drug set $\mathcal{N}(D)$, where edge weights reflect structural similarity restricted to disease-relevant targets:

$$w_D(d_i, d_j) = \text{Jaccard}(\text{targets}(d_i) \cap \mathcal{G}_{genes}(D), \text{targets}(d_j) \cap \mathcal{G}_{genes}(D)) \quad (13)$$

where $\mathcal{G}_{genes}(D)$ is the set of genes associated with disease D . This formulation ensures that off-target drug promiscuity does not artificially inflate similarity scores. To account for the inherent noise in biological data, we employ γ -quasi-cliques rather than strict cliques, where every node

maintains degree at least $\gamma \cdot (|C| - 1)$, with $\gamma = 0.8$.

Module Score Computation: For a candidate drug $d \notin \mathcal{N}(D)$, we quantify its alignment with the known drug module through three sub-scores. The *clique completion score* measures how strongly d connects to the largest quasi-clique C^* of known drugs:

$$\text{clique}_{\text{comp}}(d, D) = \frac{|\{v \in C^* : w_D(d, v) > \theta_D\}|}{|C^*|} \quad (14)$$

The *Turán density increase* measures how much adding d increases the structural cohesion of the known drug subgraph:

$$\Delta\tau(d, D) = \max\left(0, \tau(\mathcal{G}_{drug}^D \cup \{d\}) - \tau(\mathcal{G}_{drug}^D)\right) \quad (15)$$

The *off-target penalty* discounts drugs whose high connectivity arises from broad pharmacological promiscuity rather than disease-relevant targeting:

$$\text{off}_{\text{target}}(d, D) = 1 - \frac{|\text{targets}(d) \cap \mathcal{G}_{genes}(D)|}{|\text{targets}(d)|} \quad (16)$$

The complete module score combines these components:

$$S_{\text{module}}(d, D) = \text{clique}_{\text{comp}}(d, D) \cdot \Delta\tau(d, D) \cdot \tau_{\text{base}}(D) \cdot (1 - \text{off}_{\text{target}}(d, D)) \quad (17)$$

where $\tau_{\text{base}}(D) = \tau(\mathcal{G}_{drug}^D)$ is the baseline Turán density of the known drug subgraph for D .

Algorithm 3: Turán Module Score Precomputation

Input: \mathcal{G} , known pairs Ω_{pos} , threshold $\gamma = 0.8$

Output: $S_{\text{module}}(d, D)$ for all (d, D)

1. **for** each disease $D \in \mathcal{V}_D$ **do**
 2. Build \mathcal{G}_{drug}^D over $\mathcal{N}(D)$ via Eq. (13)
 3. Compute $\tau_{\text{base}}(D)$ via Eq. (12)
 4. Find largest γ -quasi-clique C^* via greedy peeling
 5. **for** each candidate drug $d \notin \mathcal{N}(D)$ **do**
 6. Compute $\text{clique}_{\text{comp}}(d, D)$ via Eq. (14)
 7. Compute $\Delta\tau(d, D)$ via Eq. (15)
 8. Compute $\text{off}_{\text{target}}(d, D)$ via Eq. (16)
 9. Compute $S_{\text{module}}(d, D)$ via Eq. (17)
 10. **end for**
 11. **end for**
 12. Normalize S_{module} over all pairs
 13. **return** $S_{\text{module}} \in \mathbb{R}^{|\mathcal{V}_d| \times |\mathcal{V}_D|}$
-

Effective Resistance Bridge Score

Theoretical Foundation: Effective resistance, originating from Kirchhoff's electrical network theory [3] and formalized as a graph-theoretic metric by Klein and Randić [4], provides a principled measure of the structural robustness of connectivity between two nodes. For a weighted graph with Laplacian L , the effective resistance between nodes u and v is:

$$R_{\text{eff}}(u, v) = (\mathbf{e}_u - \mathbf{e}_v)^T L^\dagger (\mathbf{e}_u - \mathbf{e}_v) \quad (18)$$

where \mathbf{e}_u is the indicator vector for node u . A low effective resistance between a drug and disease node indicates many independent parallel paths — a structurally robust therapeutic connection. A high effective resistance indicates a single bottleneck path — a fragile prediction that depends critically on one intermediate node. By Kirchhoff's theorem, $R_{\text{eff}}(u, v)$ is inversely proportional to the number of spanning trees containing the edge (u, v) , providing a precise combinatorial interpretation.

Bridge Node Centrality: For an intermediate node m on

paths between drug d and disease D , its bridge importance is quantified by the increase in effective resistance upon its removal:

$$\text{bridge}(m; d, D) = R_{eff}^{-m}(d, D) - R_{eff}(d, D) \quad (19)$$

where $R_{eff}^{-m}(d, D)$ is computed via rank-1 update of $L_{d,D}^\dagger$ after removing m . This measure identifies the single most critical biological intermediary in the drug-disease mechanism. Bridge scores are further weighted by node type importance:

$$\text{bridge}_w(m; d, D) = \text{bridge}(m; d, D) \cdot w_{node}(m) \quad (20)$$

$$w_{node}(m) = \begin{cases} 1.0 & m \text{ is a pathway node} \\ 0.9 & m \text{ is a gene/protein node} \\ 0.7 & m \text{ is a phenotype node} \\ 0.5 & \text{otherwise} \end{cases} \quad (21)$$

Bridge Centrality Score: The final bridge centrality score combines global connection robustness with the importance of the critical bridge node:

$$S_{central}(d, D) = \frac{1}{1 + R_{eff}(d, D)} \cdot \max_{m \in \mathcal{V}_{d,D}} \text{bridge}_w(m; d, D) \cdot \frac{\kappa(m)}{\kappa_{max}} \quad (22)$$

where $\mathcal{V}_{d,D}$ is the set of intermediate nodes within three hops of both d and D , and $\kappa(m)$ denotes the coreness of node m from k -core decomposition.

Algorithm 4: Effective Resistance Bridge Score Precomputation

Input: \mathcal{G} , weighted Laplacian L

Output: $S_{central}(d, D)$, $\text{bridge}_{node}(d, D)$ for all (d, D)

1. Compute L^\dagger via truncated SVD Top- k singular values
 2. **for** each drug-disease pair (d, D) **do**
 3. Compute $R_{eff}(d, D)$ via Eq.(18)
 4. Extract local subgraph $\mathcal{G}_{d,D}$ within 3 hops of d and D
 5. Compute local $L_{d,D}^\dagger$
 6. **for** each intermediate node $m \in \mathcal{V}_{d,D} \setminus \{d, D\}$ **do**
 7. Compute $\text{bridge}(m; d, D)$ via rank-1 update Eq. (19)
 8. Compute $\text{bridge}_w(m; d, D)$ via Eq. (20)
 9. **end for**
 10. $\text{bridge}_{node}(d, D) \leftarrow \arg \max_m \text{bridge}_w(m; d, D)$
 11. Compute $S_{central}(d, D)$ via Eq. (22)
 12. **end for**
 13. Normalize $S_{central}$ over all pairs
 14. **return** $S_{central}$, bridge_{node}
-

Phase 3: GNN Learning

Having precomputed the three graph-theoretic score matrices, we now train a GNN to learn a composite scoring function that combines these structural priors with learned representations. The graph-theoretic scores serve as both structural features and regularization signals — they constrain what the GNN can learn while providing interpretable building blocks that the learned model cannot easily ignore.

Node Feature Initialization

Each node is initialized with a feature vector that concatenates domain-specific attributes with graph-structural statistics precomputed from \mathcal{G} . Crucially, the structural statistics derived from our three graph-theoretic principles are embedded directly into the initial node

representation, so the GNN begins training with knowledge of the combinatorial structure of the graph:

$$\mathbf{h}_v^{(0)} = \mathbf{W}_{type}(v) \cdot [\mathbf{x}_v \parallel \log \kappa(v) \parallel \tau_{mod}(v) \parallel \text{bridge}_{prior}(v)] + \mathbf{b}_{type}(v) \quad (23)$$

where \mathbf{x}_v is the type-specific attribute vector (Morgan fingerprints for drugs, GO term embeddings for genes, ontology embeddings for diseases), $\log \kappa(v)$ is the log-coreness from k -core decomposition, $\tau_{mod}(v)$ is the Turán density of v 's module, and $\text{bridge}_{prior}(v)$ is the average bridge score of v across all drug-disease pairs it mediates. Type-specific projection matrices $\mathbf{W}_{type}(v)$ project heterogeneous feature spaces to a unified dimension d_0 .

Algorithm 5: Node Feature Initialization

Input: \mathcal{G} , S_{path} , S_{module} , $S_{central}$

Output: Initial embeddings $\mathbf{H}^{(0)} \in \mathbb{R}^{|\mathcal{V}| \times d_0}$

1. Compute $\kappa(v) \forall v$ via k -core decomposition
2. Compute $\tau_{mod}(v) \forall v$ from Algorithm 3
3. Compute $\text{bridge}_{prior}(v) \forall v$ from Algorithm 4
4. **for** each node $v \in \mathcal{V}$ **do**
5. Construct \mathbf{x}_v using type-specific features

$$\begin{aligned}
& 6. \quad \mathbf{h}_v^{(0)} \leftarrow \mathbf{W}_{type(v)} \cdot [\mathbf{x}_v \| \log \kappa(v) \| \tau_{mod}(v) \| \text{bridge}_{prior}(v)] + \mathbf{b}_{type(v)} \quad \text{Eq. (23)} \\
& 7. \quad \text{end for} \\
& 8. \quad \text{return } \mathbf{H}^{(0)}
\end{aligned}$$

Graph-Theory-Informed Message Passing

We employ a relation-aware attention mechanism where attention coefficients are explicitly conditioned on the precomputed bridge centrality and Turán module density. This grounds the attention mechanism in graph-theoretic structure rather than relying entirely on learned affinities. For each layer l and relation type r , the message from node u to node v is:

$$\mathbf{m}_{u \rightarrow v}^r = \mathbf{W}_r^{(l)} \cdot \mathbf{h}_u^{(l)} \cdot \mathcal{W}(u, v, r) \quad (24)$$

The attention logit is:

$$\begin{aligned}
e_{uv}^r &= \text{LeakyReLU}(\mathbf{a}_r^\top [\mathbf{h}_u^{(l)} \| \mathbf{h}_v^{(l)} \| \text{bridge}_{prior}(v) \| \tau_{mod}(v) \| \mathcal{W}(u, v, r)]) \quad (25) \\
\alpha_{uv}^r &= \frac{\exp(e_{uv}^r)}{\sum_{u' \in \mathcal{N}_r(v)} \exp(e_{u'v}^r)} \quad (26)
\end{aligned}$$

The node update aggregates messages across all relation types with a residual connection to prevent over-smoothing:

$$\mathbf{h}_v^{(l+1)} = \text{LayerNorm} \left(\sigma \left(\mathbf{W}_{self}^{(l)} \mathbf{h}_v^{(l)} + \sum_{r \in \mathcal{R}} \sum_{u \in \mathcal{N}(v)} \alpha_{uv}^r \cdot \mathbf{m}_{u \rightarrow v}^r \right) + \mathbf{h}_v^{(l)} \right) \quad (27)$$

Algorithm 6: Graph-Theory-Informed Message Passing

Input: $\mathbf{H}^{(0)}$, \mathcal{G} , $\{A_r\}$, bridge_{prior} , τ_{mod} , number of layers L

Output: Final node embeddings $\mathbf{H}^{(L)}$

```

1.   for  $l = 0$  to  $L - 1$  do
2.       for each node  $v \in \mathcal{V}$  do
3.           for each relation type  $r \in \mathcal{R}$  do
4.               for each  $u \in \mathcal{N}_r(v)$  do
5.                   Compute  $\mathbf{m}_{u \rightarrow v}^r$  via Eq. (24)
6.                   Compute  $\alpha_{uv}^r$  via Eq. (25) – (26)
7.               end for
8.           end for
9.           Update  $\mathbf{h}_v^{(l+1)}$  via Eq.(26)
10.        end for
11.    end for
12.    return  $\mathbf{H}^{(L)}$ 

```

Module-Weighted Graph Aggregation

After L layers of message passing, node-level embeddings are pooled into drug-level and disease-level representations. We use attention-weighted pooling over each node's two-hop neighborhood, with the pooled representation further scaled by the Turán module density of the drug, reflecting the structural cohesion of its drug class:

$$\mathbf{z}_d^{raw} = \sum_{v \in \mathcal{V}_d} \beta_v^d \cdot \mathbf{h}_v^{(L)}, \mathbf{z}_d = \tau_{mod}(d) \cdot \mathbf{z}_d^{raw} \quad (28)$$

$$\beta^d = \text{softmax} \left(\text{MLP}_d \left(\mathbf{h}_v^{(L)} \| \mathbf{h}_d^{(L)} \| \tau_{mod}(v) \right) \right) \quad (29)$$

Disease embeddings \mathbf{z}_D are computed analogously, with the community cohesion score replacing the Turán module density as the scaling factor.

Phase 4: Composite Decoder and Confidence Scoring Repurposing Score

The final drug-disease repurposing score is computed by an MLP decoder that takes as input both the three precomputed graph-theoretic scores and the learned GNN embeddings, additionally incorporating the Adamic-Adar neighborhood overlap as a supplementary structural feature:

$$\text{AA}(d, D) = \sum_{v \in \mathcal{N}(d) \cap \mathcal{N}(D)} \frac{1}{\log(\text{deg}(v) + 1)} \quad (30)$$

$$\text{score}(d, D) = \text{MLP}_{dec}(\text{S}_{path}(d, D) \| \text{S}_{module}(d, D) \| \text{S}_{central}(d, D) \| \text{AA}_{norm}(d, D) \| \mathbf{z}_d \| \mathbf{z}_D) \quad (31)$$

The weights implicitly assigned to the three graph-theoretic scores are learned through the MLP, allowing the model to discover the relative importance of each structural principle from the training data.

Graph-Theoretic Confidence Score

A key contribution of MPRDR is replacing the uncalibrated sigmoid score of existing methods with a confidence value derived entirely from graph-theoretic quantities. This confidence is not learned — it is a deterministic function of the graph structure:

$$\text{conf}(d, D) = \text{diversity}(d, D) \cdot \frac{1}{1 + R_{eff}(d, D)} \cdot \tau_{base}(D) \cdot \text{clique}_{comp}(d, D) \quad (32)$$

Where $\text{diversity}(d, D)$ is the path diversity, $\frac{1}{1 + R_{eff}(d, D)}$ is the connection robustness and $\tau_{base}(D) \cdot \text{clique}_{comp}(d, D)$ the module cohesion.

The three factors capture orthogonal dimensions of structural support: how many independent biological mechanisms connect d to D ; how robustly they are connected through the graph; and how cohesively d aligns

combination embedding $\mathbf{z}_{d_1} + \mathbf{z}_{d_2}$ in Eq. (31), covered meta-path types, and $\mathbb{1}[\text{mod}(d_1) \neq \text{mod}(d_2)]$ path_{syn}(d_1, d_2, D^*) is the ratio of union to intersection of rewards pairs from different Turán modules.

Algorithm 9: Inference, Explanation, and Drug Combination Scoring

Input: Trained Θ , target disease D^* , candidate drugs \mathcal{V}_d , top- K

Output: Ranked drugs with scores, confidence, and explanations; top drug pairs ranked by synergy

1. **for** each $d \in \mathcal{V}_d$ **do**
 2. Compute final_{score}(d, D^*) and conf(d, D^*)
 3. **end for**
 4. Rank all drugs by final_{score} descending
 5. **for** each d in top- K **do**
 6. Retrieve top-3 paths from S_{path} for (d, D^*)
 7. Retrieve Turán module report from Algorithm 3
 8. Retrieve bridge node report from Algorithm 4
 9. Assemble confidence breakdown from Eq.(32)
 10. **end for**
 11. **for** each pair $(d_1, d_2) \in \mathcal{V}_d \times \mathcal{V}_d, d_1 \neq d_2$ **do**
 12. Compute syn(d_1, d_2, D^*) via Eq.(38)
 13. **end for**
 14. Rank pairs by syn descending
 15. **return** ranked drugs with explanations, top- K synergistic pairs
-

4. Experimental Analysis

Knowledge Graph and Dataset

PrimeKG [7] is the main knowledge graph used in all experiments. PrimeKG contains 4,050 drugs, 17,080 diseases, 27,671 genes, 2,516 pathways, 13,299 phenotypes with 18,776 known drug-disease indication edges and 4,968 contraindication edges. Two are evaluated protocols (i) a standard random 80/10/10 split on comparability with the benchmark, and (ii) a temporal split in which all drug-disease indication edges older than 2020 are the training graph, and all FDA-approved new indications between 2021 and 2024 form the prospective test set.

Baselines

MPRDR is compared to three main repurposing baselines and five architectural baselines. The main baselines are: TxGNN [5], a zero-shot repurposing model, which relies on relational GCN and disease metrics learning on PrimeKG; GDRnet [6], a heterogeneous GNN, which supports SIGN precomputation and quadratic norm decoder on DRKG; and the COVID-19 GNN repurposing model of Hsieh et al. [11], a VGAE-based pipeline, which relies on DRKG transfer learning and Bayesian Pairwise Ranking. RGCN [13], HAN [14], and HGT [15] architectural baselines are implemented on the same PrimeKG graph, and classical lower bounds TransE [16] and RotatE [17], are KG embedding baselines.

Evaluation

Metrics Performance is measured on AUC-ROC, AUPRC, Hits at top 10 and 20 and MRR on the standard split. AUC-ROC, AUPRC and Recall at 50 are used to measure temporal generalisation on the prospective FDA approval test set. Cold-start performance is quantified using Hits 10, Hits 20 and AUPRC where all the drug-disease edges have been dropped on a held-out 20 percent of drugs. Expected Calibration Error (ECE) is used to measure confidence calibration. The Spearman correlation between drug

combination synergy and NCI-ALMANAC experimental synergy scores is used to measure drug combination synergy.

Main Results

Standard Board of Performance.

All four metrics consistently map on to the MPRDR system outpacing all baseline measures of performance with the greatest benefit accrued to measures that use precision targeted metrics. The improvement in AUPRC compared with TxGNN is the clearest example of the practical advantage associated with using priors to provide a graph theoretic score that surfaces high-confident candidates at the top of the ranked list as opposed to improving overall separation among candidates uniformly. Compared with GDRnet, MPRDR provides an even larger (+9.7%) improvement in AUPRC and hits at 20 (+13.5% improvement), which continues to validate the benefits of designing based on typed path evidence relative to fixed polynomial aggregation. In the case of the COVID-19 GNN baseline that uses a disease-specific graph rather than a repurposing KG the MPRDR system shows the largest delta compared with all baseline measures of performance; the AUPRC improvement (+21.4%) is consistent with domain mismatch being large. AUC-ROC improvement over the highest performing baseline measure is relatively small (+3.1%), which is the norm when examining the improvement for a precision-recall-dominated task with such a high class imbalance; the drug-disease bipartite graph contains fewer than 0.3% positive pairs indicating that AUPRC will be the more accurate measure of ranking quality compared to Hits@K. The existence of consistent Hits@10 advantage across all comparisons is evidence that graph theoretic priors provide a structural source of information that cannot be recovered through embedding proximity alone..

Cold-Start Evaluation

All drug-disease edges of an indiscriminate 20% of drugs are taken off the training graph. The TxGNN, GDRnet and KG embedding baselines degrade significantly due to their reliance on drug-disease co-occurrence in training graph in their scoring. The performance of MPRDR is high since the scores of path-based and module-based can still be computed by the rest of the drug-target, drug-pathway, and drug-similarity edges.

5. CONCLUSION

We introduced MPRDR, a drug repurposing platform that uses graph theory as the scoring logic instead of a message passing data structure. MPRDR can be used to obtain mechanistically interpretable, structurally calibrated, and general predictions to cold-start problems that existing methods fail on by basing the scoring functionality on three classical theorems typed path algebra [1], Turan's extremal graph theory [2], and Kirchhoff's effective resistance [3,4]. The most important empirically determined result is that graph-theoretic precomputed scores enhance prospective temporal recall the most, which supports the idea that mechanistic structural reasoning is able to extrapolate beyond database co-occurrence patterns, in a way relevant directly to drug discovery in the real world.

The limitations of MPRDR are several and are an incentive to future work. The computation of the effective resistance using Laplacian pseudoinverse is asymptotic $O(|V|^2 k)$ truncated SVD that is computable using PrimeKG but has to be approximated with very large graphs. The weights of the therapeutic meta-path type are determined by biological knowledge or trained on training data in low-data regimes they can be unreliable. Future work will explore the adaptive path weight learning through meta-learning, the extension to temporal graph dynamics where KG edges are introduced sequentially and combining tissue-specific expression data to contextualise disease modules are not limited by structural KG topology alone.

In addition to drug repurposing, the three graph theoretical concepts of MPRDR, typed path evidence, module cohesion through Turan density, and, finally, robustness through effective resistance, are generalizable to any link prediction task on a heterogeneous knowledge graph with the necessity of mechanistic interpretability and structural confidence.

REFERENCES

1. Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "PathSim: Meta path-based top-k similarity search in heterogeneous information networks," *Proc. VLDB Endow.*, vol. 4, no. 11, pp. 992–1003, 2011.
2. P. Turan, "On an extremal problem in graph theory," *Matematikai es Fizikai Lapok*, vol. 48, pp. 436–452, 1941.
3. G. Kirchhoff, "Über die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Vertheilung galvanischer Ströme geführt wird," *Annalen der Physik und Chemie*, vol. 72, pp. 497–508, 1847.
4. D. J. Klein and M. Randić, "Resistance distance," *Journal of Mathematical Chemistry*, vol. 12, no. 1, pp. 81–95, 1993.
5. M. Huang, C. Chandak, K. Huang, and M. Zitnik, "Zero-shot generalization for drug repurposing with geometric deep learning and clinical knowledge," *Nature Medicine*, vol. 30, pp. 2923–2934, 2024.
6. S. Doshi and S. P. Chepuri, "A computational approach to drug repurposing using graph neural networks," *Computers in Biology and Medicine*, vol. 148, p. 105797, 2022.
7. P. Chandak, K. Huang, and M. Zitnik, "Building a knowledge graph to enable precision medicine," *Scientific Data*, vol. 10, no. 1, p. 67, 2023.
8. J. A. DiMasi, H. G. Grabowski, and R. W. Hansen, "Innovation in the pharmaceutical industry: New estimates of R&D costs," *Journal of Health Economics*, vol. 47, pp. 20–33, 2016.
9. T. T. Ashburn and K. B. Thor, "Drug repositioning: Identifying and developing new uses for existing drugs," *Nature Reviews Drug Discovery*, vol. 3, no. 8, pp. 673–683, 2004.
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, p. 584, 2024.
11. K. Hsieh, Y. Wang, L. Chen, Z. Zhao, S. Savitz, X. Jiang, J. Tang, and Y. Kim, "Drug repurposing for COVID-19 using graph neural network and harmonizing multiple evidence," *Scientific Reports*, vol. 11, p. 23179, 2021.
12. Ioannidis, X. Zheng, M. Gao, H. Zhao, G. Karypis, and T. Faloutsos, "DRKG: Drug repurposing knowledge graph for Covid-19," *arXiv:2010.09600*, 2020.
13. M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *Proc. ESWC*, Springer, 2018, pp. 593–607.
14. X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu, "Heterogeneous graph attention network," in *Proc. WWW, ACM*, 2019, pp. 2022–2032.
15. H. Yun, J. Jeong, H. Kim, J. Kim, and K. Kim, "Graph transformer networks," in *Proc. NeurIPS*, 2019.
16. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proc. NeurIPS*, 2013, pp. 2787–2795.
17. Z. Sun, Z. Deng, J. Nie, and J. Tang, "RotatE: Knowledge graph embedding by relational rotation in complex space," in *Proc. ICLR*, 2019.
18. B. Yang, W. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," in *Proc. ICLR*, 2015.
19. D. S. Wishart et al., "DrugBank 5.0: A major update to the DrugBank database for 2018," *Nucleic Acids Research*, vol. 46, no. D1, pp. D1074–D1082, 2018.
20. The UniProt Consortium, "UniProt: The universal protein knowledgebase in 2021," *Nucleic Acids Research*, vol. 49, no. D1, pp. D480–D489, 2021.
21. L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social Networks*, vol. 25, no. 3, pp. 211–230, 2003.