Phenotype-Conditioned Drug Repurposing for Undiagnosed Rare Disease Patients via Graph Neural Networks and LLM Hybridization

Beatrice Chen [1] **Qiaohan Xu** [1] **Jiale Yang** [1]

Abstract

Rare disease patients often face years of diagnostic delay, and many never receive a confirmed molecular diagnosis. Existing drug repurposing models usually require a disease label, while phenotype-based diagnostic tools stop at diagnosis without recommending treatment. We formulate diagnosis-free drug repurposing as a phenotype-set → drug ranking task on PrimeKG and evaluate it on 108 held-out diseases with 914 disease–drug pairs. We propose a graph and language-model hybrid that combines an R-GCN encoder, drug-conditioned cross-attention, and biomedical text embeddings. Score-level fusion achieves MRR 0.325 on all test diseases and 0.311 on the 78 diseases where PubCaseFinder fails, tripling the PubCaseFinder → TxGNN cascade MRR of 0.103 and recovering 76% of the oracle TxGNN ceiling without using a predicted diagnosis. The model is substantially more robust to missing disease labels, dropping only 3–4% from the full test set to the undiagnosable subset, compared with 67% for the cascade. These results suggest that phenotype-conditioned drug ranking is a promising direction for diagnosis-free therapeutic recommendation.

1. Introduction

Rare diseases collectively affect an estimated 400 million people worldwide, yet patients face an average diagnostic odyssey of more than five years and roughly half never receive a confirmed molecular diagnosis (Nguengang Wakap et al., 2020; Marwaha et al., 2022; Wright et al., 2018). For these patients, clinical phenotypes are often the only structured signal available. Existing AI tools, however, treat phenotype-based diagnosis and disease-based therapeutic recommendation as two separate problems and stop at one or the other.

This disconnect creates both a clinical and a methodological gap. Clinically, state-of-the-art drug repurposing models such as TxGNN (Huang et al., 2024) require a confirmed disease node, while phenotype-driven diagnostic models such as SHEPHERD (Alsentzer et al., 2025) and DeepRare (Zhao et al., 2026) return a disease but no treatment, systematically excluding undiagnosed patients from the tools that could most plausibly help them. Methodologically, the natural workaround of cascading phenotype→disease and disease→drug models is vulnerable to error propagation: any mis-prediction at the diagnosis stage is inherited downstream, and phenotype-level signal that might have been informative for drug selection is collapsed into a single discrete disease label. To our knowledge, no end-to-end architecture currently maps clinical phenotype sets directly to ranked drug candidates without an intermediate diagnosis. This motivates a complementary diagnosis-free direction in which the diagnosis step is bypassed entirely.

Building on a diagnosis-free R-GCN (Schlichtkrull et al., 2018) encoder over PrimeKG with a drug-conditioned cross-attention scorer, we evaluate whether biomedical text signals can close the remaining gap to oracle TxGNN, and where in the pipeline they should enter. We compare three fusion strategies, ordered by integration stage. *Shallow aggregated ranking* combines the final ranked lists from the graph and LLM branches, leaving the two models fully separate until the last step. *Scoring-level fusion* combines per-disease numerical scores: the R-GCN scores candidate drugs from the phenotype set, a separate text branch scores drug–phenotype compatibility using GPT-4o-enriched description embeddings, and the two scores are mixed to produce the final ranking. *Feature-level fusion* integrates earliest, projecting frozen biomedical text embeddings into the R-GCN as node features so text and graph signals are trained together during message passing. We sweep text encoders (PubMedBERT (Gu et al., 2021), BiomedBERT (Chakraborty et al., 2020), BioLinkBERT (Yasunaga et al., 2022), SPECTER2 (Singh et al., 2023)) and projec-

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. Correspondence to: Beatrice Chen <beatrice_chen@hms.harvard.edu>, Qiaohan Xu <qiaohan_xu@hms.harvard.edu>, Jiale Yang <jocelyn_yang@hms.harvard.edu>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)

tions (none, PCA, linear, autoencoder).

Our main contributions are:

- **Phenotype-conditioned graph–LLM hybrid.** We propose an R-GCN with drug-conditioned cross-attention to rank all 7,957 candidate drugs from phenotype-only inputs, and compare shallow, score-level, and feature-level text fusion strategies.
- **Fusion stage matters.** Score-level fusion performs best among leakage-controlled methods, outperforming the PubCaseFinder → TxGNN cascade and recovering 76% of the oracle TxGNN ceiling on undiagnosable diseases without using disease labels.
- **Interpretability and failure analysis.** We use cross-attention weights and stratified error analysis to examine when the model succeeds or fails. Fusion helps mainly for graph-uncertain, supervision-rich diseases, while attention maps show that explanations can track PrimeKG coverage rather than biological ground truth.

2. Background

2.1. Phenotype Sets as Patient Representations

Phenotype-based diagnostic tools such as Phenomizer (Köhler et al., 2009), PubCaseFinder (Fujiwara et al., 2018), SHEPHERD (Alsentzer et al., 2025), and DeepRare (Zhao et al., 2026) show that HPO term sets can capture clinically meaningful disease signals through similarity matching, graph learning, or LLM-based reasoning. However, these methods mainly use phenotypes as a route to disease identification and generally stop once a candidate diagnosis is produced. This leaves open the question of whether the same phenotype information can be used more directly for therapeutic recommendation. The question is supported by prior work on phenotype-driven treatment discovery: phenotypic signature matching has shown that candidate therapies can be identified by comparing disease-model phenotypes with drug-induced phenotypic changes (Ambesi-Impiombato et al., 2023; Lamb et al., 2006). Although these studies are mainly preclinical and do not use patient-level HPO inputs, they suggest that phenotype-level patterns can contain treatment-relevant information. Building on both lines of work, we use patient phenotype sets as the direct input for drug ranking in a knowledge-graph framework.

2.2. Knowledge Graphs and GNNs for Drug Repurposing

Biomedical knowledge graphs encode heterogeneous relations among diseases, drugs, phenotypes, genes, proteins, and pathways, making them useful for relational reasoning in drug repurposing. PrimeKG (Chandak et al., 2023) integrates 20 biomedical databases into a graph of 129,375 nodes and roughly 8 million directed edges, including disease–phenotype and drug–disease indication relations, making it a natural substrate for phenotype-conditioned drug ranking. Among GNN architectures adapted to heterogeneous biomedical graphs, R-GCN (Schlichtkrull et al., 2018) handles multi-relational structure through relation-specific message passing, while HGT (Hu et al., 2020) and HAN (Wang et al., 2019) use attention mechanisms to model heterogeneous node and relation types. Within drug repurposing specifically, TxGNN (Huang et al., 2024) combines an R-GCN encoder over PrimeKG with a disease-similarity metric-learning module for zero-shot prediction on diseases with limited known treatments. However, TxGNN still requires a known disease node as input. We therefore use TxGNN as an oracle-disease upper bound rather than a deployable model for undiagnosed patients.

2.3. LLMs and Graph–LLM Hybrids

Large language models and biomedical text encoders provide semantic information that can complement knowledge-graph structure, encoding mechanisms, therapeutic classes, clinical uses, and phenotype descriptions that may not be fully captured by graph topology alone. Recent work has applied LLMs or biomedical language models to therapeutic prediction tasks (Yan et al., 2024; Zheng et al., 2024). More directly related, graph–text hybrid methods such as FuseLinker (Xiao et al., 2024), LLM-DDA (Gu et al., 2025), and LLaDR (Xiang et al., 2025) integrate LLM-derived or biomedical text embeddings with GNN-based link prediction. Together, these studies suggest that graph and text signals can improve biomedical prediction when combined. However, these methods generally assume a known disease node and predict drug–disease links, whereas our work conditions directly on phenotype sets to reflect a real-world, diagnosis-free setting.

3. Methods

3.1. Dataset

PrimeKG is well suited for our task because it links the two sources of information needed for diagnosis-free drug repurposing: clinical phenotype descriptions and drug–disease therapeutic relationships. It integrates broad disease coverage, including rare diseases, with heterogeneous biomedical relations. However, because PrimeKG encodes known indications and off-label uses rather than all biologically plausible treatments, its edges represent an incomplete clinical ground truth.

We identified 539 PrimeKG diseases with at least one phenotype association and one approved drug indication, then

applied an 80/20 disease-level split, yielding 431 training diseases (2,703 disease–drug pairs) and 108 test diseases (914 held-out pairs: 690 indications and 224 off-label-use pairs). We include off-label edges as auxiliary positive supervision because they represent clinically observed or literature-supported drug–disease links and may capture plausible repurposing signals beyond formal indication labels. Each test disease is represented only by its HPO phenotype set (mean 23 phenotypes), which serves as the only input at inference. To prevent message-passing leakage, all edges incident to test disease nodes are removed from the training subgraph.

3.2. Baselines

TxGNN with oracle disease input (upper bound). TxGNN requires a confirmed disease as input, so we treat it as an upper bound rather than a fair competitor. We retrained it from scratch with the official package after removing the 1,506 indication and reverse-indication edges incident to our 108 test diseases to prevent label leakage; other edges incident to test diseases are retained so the disease identity provided at inference still has structural support for message passing.

Personalized PageRank. Tests whether graph topology alone carries phenotype-to-drug signals without learned parameters. We constructed an undirected, row-normalized adjacency matrix over all 129,375 PrimeKG nodes, excluded all edges incident to test disease nodes, and seeded PPR from each test disease's phenotype nodes with restart probability $\alpha = 0.15$. Scores were computed by power iteration: $r \leftarrow (1-\alpha)A^\top r + \alpha s$, where s is the uniform seed vector over the disease's phenotype nodes. A sensitivity sweep over $\alpha \in \{0.05, 0.10, 0.15, 0.20, 0.30, 0.50, 0.75\}$ on a 30-disease subset showed stable performance.

PubCaseFinder → TxGNN cascade. Implements the canonical phenotype → disease → drug pipeline. PubCaseFinder (Fujiwara et al., 2018) is a deployed HPO-based diagnostic tool that returns ranked rare-disease candidates from phenotype sets by matching against PubMed case reports, and is a standard benchmark used by recent agentic systems such as DeepRare. For each test disease, PubCaseFinder returns top-K candidate diagnoses with similarity scores; we run TxGNN once per retrieved disease and aggregate the resulting K score vectors as a similarity-weighted average. We swept $K \in \{1, 3, 5, 10, 20\}$ and found $K = 1$ best, with performance degrading as K increased, which suggests that lower-ranked retrievals introduce more noise than signal. We report cascade results with $K = 1$.

GPT-3.5 zero-shot LLM baseline. Tests whether a language model can rank drugs from phenotype text alone. We use GPT-3.5 rather than GPT-4 or later because GPT-3.5 is the most recent OpenAI model whose training cutoff predates the public release of PrimeKG, allowing us to partially separate LLM-derived biomedical reasoning from direct memorization of disease–drug associations encoded in the graph. The prompt template (following (Yan et al., 2024)) supplies HPO phenotype names and asks for 50 ranked generic drug names in JSON. On average 49 of the returned names mapped to PrimeKG's 7,957-drug list. Even though the prompt provides only phenotypes, the model may implicitly infer a disease label and then retrieve memorized disease–drug associations, so we treat this as a leakage-tainted reference.

3.3. R-GCN with Drug-Conditioned Cross-Attention

We frame phenotype-conditioned drug repurposing as a transductive ranking task on PrimeKG: given a phenotype set P, the model scores all 7,957 drugs in a single forward pass; no test disease node is ever constructed. All indication, reverse_indication, and off_label_use edges incident to the 108 test diseases are removed from the training subgraph beforehand to prevent message-passing leakage.

The encoder is a 3-layer R-GCN with basis decomposition (15 bases), producing 256-d node embeddings. The scorer is a drug-conditioned cross-attention module (4 heads): the candidate drug embedding is the query and the patient's phenotype embeddings are keys and values, returning a scalar score $s_{\text{graph}}(P, d)$. This per-drug weighting lets the model emphasize different phenotypes for different drugs, and the attention weights are retained as per-drug HPO attributions for the case study in Section 5.3. We train end-to-end with a margin ranking loss ($m = 1.0$) over indication and off-label positives, sampling one degree-weighted negative per positive (Adam, lr 10^{-3}, wd 10^{-5}, gradient accumulation 4, gradient clip 1.0). The model was trained for 70 epochs in approximately 5 hours.

3.4. Fusion with LLM Signals

3.4.1. Prediction Aggregation (Shallow Re-ranking)

This variant pairs the R-GCN backbone with the GPT-3.5 zero-shot baseline (Section 3.2) directly at the ranking stage. We use GPT-3.5 here for consistency with that baseline, so any improvement over GPT-3.5 alone is attributable to the addition of graph signal rather than to a stronger LLM. For each test disease, GPT-3.5 is queried with the patient's phenotype list (prompt in Appendix C) and returns the 50 most therapeutically relevant generic drug names. The final ranking averages the per-disease ranks from the R-GCN

backbone and the GPT-3.5 list with equal weight. Like the GPT-3.5 baseline, this variant inherits leakage from the LLM's internal knowledge of disease–drug associations; we therefore mark it as leakage-tainted and treat it as an upper-envelope reference rather than a fair comparator.

3.4.2. Text Description Generation and Encoding

For each drug and phenotype node, we generate a 2–3 sentence biomedical description with GPT-4o using PrimeKG metadata: drug names, targets, and pathways for drugs, and HPO terms, associated genes, and associated conditions for phenotypes (full prompts in Appendix C). The phenotype prompt forbids disease names to avoid the disease-string shortcut in Section 3.4.1. We concatenate the raw KG metadata with the GPT-4o description so each node representation preserves both graph-anchor tokens and added mechanistic context.

Hybrid descriptions are encoded with BioLinkBERT (Yasunaga et al., 2022) using mean-pooled token embeddings, projected to 256 dimensions with a nonlinear autoencoder, and L2-normalized. The autoencoder is trained jointly on drug and phenotype embeddings to support cross-type cosine similarity. We selected BioLinkBERT + nonlinear autoencoder from a 4×4 grid over PubMedBERT (Gu et al., 2021), BiomedBERT (Chakraborty et al., 2020), BioLinkBERT (Yasunaga et al., 2022), and SPECTER2 (Singh et al., 2023), crossed with no projection, PCA, linear projection, and nonlinear autoencoding. The full grid is reported in Appendix D.

3.4.3. Late (Scoring-Level) Fusion

Given per-disease normalized score vectors $\tilde{s}_{\text{graph}}(P)$ and $\tilde{s}_{\text{LLM}}(P)$, where $\tilde{s}_{\text{LLM}}(P, d) = \cos(\bar{e}_P, e_d)$ between the mean-pooled phenotype-text embedding $\bar{e}_P$ and each drug embedding e_d, the fused score is

$$s_{\text{final}}(P, d) = \beta(P) \cdot \tilde{s}_{\text{graph}}(P, d) + \big(1 - \beta(P)\big) \cdot \tilde{s}_{\text{LLM}}(P, d).$$

Both score vectors are min-max normalized per disease.

Because the R-GCN encoder was trained over all train-disease indication edges, train-side cross-validation of β is biased toward the graph branch: a train-side sweep collapses to $\beta \approx 0.9$ and improves test MRR by only ~ 0.02 over pure graph. To avoid this bias, we use an unsupervised, label-free assignment based only on inference-time score statistics.

We define each branch's per-disease score margin as the gap between top-1 and top-10 normalized scores: $g_{\text{graph}}(P) = \tilde{s}^{(1)}_{\text{graph}}(P) - \tilde{s}^{(10)}_{\text{graph}}(P)$ and likewise $g_{\text{LLM}}(P)$. Their difference $\Delta g(P) = g_{\text{graph}}(P) - g_{\text{LLM}}(P)$ is positive when the graph ranking is sharper. We split Δg into tertiles and assign $\beta = 0.1, 0.2, 0.3$ from low to high Δg, so the gate amplifies the LLM contribution everywhere while allowing slightly more graph signal when the graph branch is sharper. Bucket boundaries depend only on inference-time score statistics; no labels are consulted. Alternative gating signals (n_phenotypes, phenotype degree, top-10 Jaccard, multi-feature linear gate) did not exceed margin_gap.

3.4.4. Intermediate (Feature-Level) Fusion

We replace the raw R-GCN node embedding consumed by the cross-attention scorer with a degree-conditioned mix of graph and text features. For node v with $\ell(v) = \log(\deg(v) + 1)$, the fused embedding is $h_{\text{fused}}(v) = \alpha(v) h_{\text{graph}}(v) + (1 - \alpha(v)) h_{\text{LLM}}(v)$ with $\alpha(v) = \sigma(w \cdot \ell(v) + b)$. Nodes without a cached text embedding fall back to $h_{\text{graph}}(v)$. Training is two-phase to prevent the gate from absorbing popularity bias: we first calibrate (w, b) analytically by sweeping α within four log-degree quartiles on training pairs and fitting closed-form linear regression on $(m_q, \text{logit}(\alpha^q))$, then freeze the encoder and gate and fine-tune only the cross-attention scorer with the margin loss from Section 3.3 plus an LLM-anchoring regularizer $\mathcal{L}_{\text{anc}} = \lambda \|h_{\text{fused}} - h_{\text{LLM}}\|^2$ ($\lambda = 0.1$) on nodes with cached text embeddings (Xiang et al., 2025). Model was trained with Adam optimizer (lr 10^{-4}, weight decay 10^{-5}, batch size 512) with early stopping on validation MRR.

As alternative feature-level integration mechanisms, we additionally evaluated two autoencoder-based variants over the same hybrid descriptions and frozen R-GCN backbone. The *plain autoencoder* replaces $h_{\text{graph}}(v)$ with the latent code of an AE over the concatenation $[h_{\text{graph}}(v); h_{\text{LLM}}(v)]$, while the *residual autoencoder* preserves $h_{\text{graph}}(v)$ as the base representation and treats the AE output as an additive correction, $h_{\text{fused}}(v) = h_{\text{graph}}(v) + \text{AE}_{\text{enc}}([h_{\text{graph}}(v); h_{\text{LLM}}(v)])$. In place of the analytical gate calibration, we first pretrain the AE on reconstruction MSE over the $\sim$11.5K text-covered nodes (50 epochs, lr 10^{-3}), then freeze the R-GCN encoder and AE decoder and fine-tune the AE encoder jointly with the cross-attention scorer under the same margin loss.

3.5. Evaluation Metrics

All ranking metrics are computed per disease and macro-averaged across the 108 test diseases. We report Mean Reciprocal Rank (MRR) and Recall@k for $k \in \{1, 5, 10, 50\}$ over the full 7,957-drug library.

Ground-truth positives for each test disease are the union of PrimeKG's indication and off-label use edges, since clinical drug-repurposing decisions admit off-label evidence and indication-only labeling is otherwise a pessimistic lower bound that penalizes valid off-label predictions. All MRR and Recall numbers reported in the main text use this off-label-augmented ground truth. Performance is reported

on two evaluation splits: the full held-out set of 108 test diseases, and the subset of 78 diseases for which PubCaseFinder fails to return a correct top-1 diagnosis. The latter is the clinically relevant deployment condition for diagnosis-free phenotype-to-drug systems. For error analysis (Section 4, Section 5), we further stratify test diseases by cold-start status, phenotype-count tertile, and the diagnosable / undiagnosable subsets.

4. Results

In this section, we first report performance on all test diseases and on the undiagnosable subset. Within each split, we report observations along four aspects: how the baselines perform, whether our proposed methods outperform baselines, whether LLM fusion improves over R-GCN alone, and how leakage-tainted methods compare to one another. We then analyze sensitivity across the two splits.

4.1. All Test Diseases ($n = 108$)

Table 1 and Figure 1 (left) report performance on the full held-out test set.

Among baselines, the PubCaseFinder → TxGNN cascade reaches MRR 0.312, reflecting TxGNN's strength when a disease label is available; however, its Recall@K plateaus sharply after $K = 10$ (0.292 at $K = 10$, 0.316 at $K = 50$), a sign that performance is non-uniform and depends on whether upstream diagnosis succeeds. GPT-3.5, the leakage-tainted baseline, reaches MRR 0.320 and the strongest baseline Recall@K. Among our proposed methods, R-GCN with cross-attention alone reaches MRR 0.254 (Recall@50 0.359), feature-level fusion 0.256 (0.366), scoring-level fusion 0.325 (0.410), and shallow aggregated ranking 0.368 (0.461). Within clean (non-leakage) comparisons, scoring-level fusion is the strongest method, exceeding the cascade in MRR (0.325 vs. 0.312) without using a disease label and substantially exceeding it in Recall@50 (0.410 vs. 0.316). Scoring-level fusion also improves over R-GCN alone (MRR 0.325 vs. 0.254; R@50 0.410 vs. 0.359), whereas feature-level fusion is essentially identical to R-GCN, suggesting node-feature injection does not transmit ranking signal here. Shallow aggregated ranking attains the highest MRR overall (0.368) but inherits the leakage profile of GPT-3.5; we treat it as an upper-envelope reference rather than a fair clean comparator.

4.2. Undiagnosable Subset ($n = 78$)

Table 1 and Figure 1 (right) report performance on the subset where PubCaseFinder fails to return a correct top-1 diagnosis. This is the clinically relevant subset for our task: precisely the cases where a diagnosis-then-treatment pipeline cannot be deployed.

The cascade collapses to MRR 0.103 and Recall@50 0.101, falling slightly below PageRank (0.106, 0.159). GPT-3.5 has MRR 0.290 and Recall@50 0.299. Among our methods, R-GCN with cross-attention and feature-level fusion are similar (MRR 0.246–0.247, Recall@50 0.331–0.341), scoring-level fusion reaches 0.311 / 0.392, and shallow aggregated ranking 0.371 / 0.421. The relative ordering is preserved from the all-test split. Crucially, every clean R-GCN-based method now outperforms the cascade in MRR and Recall@50: scoring-level fusion approximately triples the cascade in MRR (0.311 vs. 0.103) and quadruples it in Recall@50 (0.392 vs. 0.101), recovering about 76% of the TxGNN oracle MRR ceiling (0.311 vs. 0.407) without a diagnosis. The score-level LLM fusion benefit observed earlier persists here (0.311 vs. 0.246 for R-GCN alone), while feature-level fusion remains essentially identical to R-GCN alone. Within the leakage-tainted comparison, shallow aggregated ranking again substantially exceeds GPT-3.5 (0.371 vs. 0.290 MRR), with even larger gaps than on the all-test split.

4.3. Sensitivity to the Undiagnosable Condition

Comparing ΔMRR % and the corresponding Recall@50 changes across the two splits reveals each model's predictive stability. The cascade is the most sensitive, dropping 67% in MRR (0.312 to 0.103) and 68% in Recall@50 (0.316 to 0.101), because it depends entirely on PubCaseFinder returning a correct top-1 diagnosis as input to TxGNN. PageRank is essentially flat across both splits (MRR 0.103 to 0.106; Recall@50 0.156 to 0.159) because it does not use diagnostic information. GPT-3.5's MRR drops modestly (-9%) but Recall@50 drops -44% (0.531 to 0.299), indicating its leakage advantage concentrates at top ranks while deeper-K reliability collapses on harder cases: it can place one or two memorized associations near rank 1 even from phenotype-only input, but the rest of its top-50 list is less reliable. Reporting only MRR would understate the magnitude of its degradation.

By contrast, our R-GCN-based models are the most stable, all degrading by only 3–4% in MRR (cross-attention -3.1%, feature fusion -3.7%, scoring fusion -4.3%; shallow aggregation rising by 0.7%), with parallel stability at Recall@50. This is evidence that our models transfer cleanly between diagnosable and undiagnosable cases at all ranking depths. Combined with absolute performance, scoring-level fusion is therefore both the strongest clean method overall and the most robust to the absence of a disease label, making it our primary recommended configuration.

Table 1. Drug ranking performance across the full held-out test set ($n = 108$, "All") and the undiagnosable subset ($n = 78$, "Undiag."). Methods marked "possible leakage" use information unavailable at inference for genuinely undiagnosed patients. TxGNN with oracle disease input is included as an upper bound. Recall@1, Recall@5, and Recall@10 are reported in Tables 2 and 3.

	MRR		Recall@50	
Model	All	Undiag.	All	Undiag.
TxGNN (Oracle Disease, upper bound)	0.404	0.407	0.673	0.653
PageRank	0.103	0.106	0.156	0.159
PubCaseFinder → TxGNN Cascade	0.312	0.103	0.316	0.101
GPT-3.5 (possible leakage)	0.320	0.290	0.531	0.299
R-GCN + Cross-Attention (ours)	0.254	0.246	0.359	0.331
R-GCN + LLM, Feature-Level Fusion	0.256	0.247	0.366	0.341
R-GCN + LLM, Scoring-Level Fusion	0.325	0.311	0.410	0.392
R-GCN + LLM, Shallow Aggregation (possible leakage)	0.368	0.371	0.461	0.421

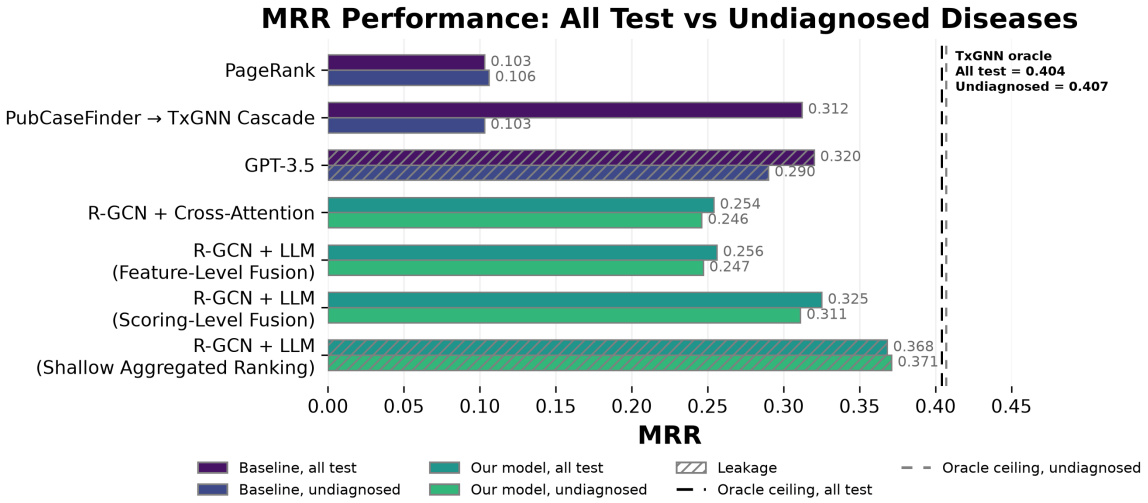


Figure 1. MRR on held-out disease–drug pairs across baselines and proposed methods. Performance is shown for both all test diseases ($n = 108$, darker bars) and the subset undiagnosable by PubCaseFinder ($n = 78$, lighter bars). Baselines are in blue; our models are in green. Hatched bars indicate methods with possible test-case leakage. The dashed line marks TxGNN with oracle disease input as an upper bound.

5. Discussion

5.1. Performance Analysis

The central finding is structural: LLM signals improve drug ranking when fused at the score level but degrade it when fused at the representation level. Score-level fusion with conditioned β reaches 0.325 MRR ($+0.071$ over R-GCN). All three feature-level architectures we tested underperform the baseline (degree-conditioned weighted averaging 0.222, autoencoder-bottleneck 0.121, residual autoencoder 0.233–0.252). The shallow rank-aggregation result (0.368) leverages GPT-3.5's pretraining corpus; we treat MRR 0.325 as our leakage-controlled estimate.

We attribute this asymmetry to integration locus and supervision scale. Score-level fusion combines scalar scores at the ranking step, letting the LLM override graph mis-rankings without disturbing the learned representation. Feature-level fusion injects LLM signal at the embedding stage, where R-GCN message passing repeatedly mixes it with graph signal before scoring; the dilution is asymmetric because the graph branch is trained end-to-end against the ranking loss while the LLM branch contributes a frozen vector. The residual variant preserves but does not exceed the graph baseline, indicating no useful representation-level information at our supervision scale (2,703 positive pairs across 431 training diseases). Even at the score level, the headline gain depends on calibration: five-fold CV selects $\beta \approx 0.9$ because R-GCN's train MRR is inflated (CV ≈ 0.83 vs. test ≈ 0.21), yielding only 0.205–0.232; the conditioned-β gate recovers most of this gap (oracle per-disease β: 0.345–0.376).

Stratified analysis (Figure 3, Appendix A) shows that of four conditioning signals, only n_true_drugs (high tertile)

and margin_gap (low tertile) gate fusion benefit (Figure 3, Appendix A); diseases with many true drugs offer more room for the R-GCN to mis-rank under popularity bias, and bunched top-K scores indicate low graph confidence. Cold-start status is non-discriminative ($\rho = 0.043$, $p = 0.66$). Frozen LLM embeddings act as ranking-stage tie-breakers on supervised territory, not representation-stage extensions.

5.2. Strengths, Weaknesses, and Ethical Considerations

The pipeline has three strengths. The R-GCN with drug-conditioned cross-attention recovers most of the structural signal in PrimeKG without a disease label, narrowing the gap to oracle-input TxGNN (0.254 vs. 0.404 MRR). Conditioned-β late fusion improves over the graph baseline on 70 of 108 test diseases (mean ΔMRR $+0.085$), and the fusion's relative benefit is preserved on the 78-disease undiagnosable subset (0.311 vs. R-GCN 0.246). The cross-attention scorer also produces per-drug HPO attention weights, providing a clinically auditable explanation substrate that the LLM-only and rank-aggregation baselines structurally cannot.

The corresponding weaknesses follow. The median ΔMRR is $+0.003$: fusion is a targeted correction (24/108 diseases hurt), not a uniform gain. Of four conditioning signals, only two demonstrably gate. We also did not run multi-seed experiments due to compute cost, so the per-disease tail of 24 hurt diseases may partly reflect ranking noise. Our evaluation also uses disease-level HPO sets rather than real patient phenotype profiles, which are typically noisier, sparser, and more variable; performance on disease-level inputs is therefore an optimistic estimate of deployment behavior. Finally, treating off-label edges as ground-truth positives admits clinically heterogeneous evidence and may inflate Recall@K relative to a strict-indication evaluation.

Three ethical concerns deserve explicit treatment. **Popularity bias**: the AUROC/MRR gap across all baselines (e.g., TxGNN AUROC 0.945 vs. MRR 0.404) confirms that widely-indicated drugs occupy top ranks at the expense of correct but less common treatments. Fusion does not correct this and may amplify it, since the gate upweights the LLM precisely on diseases with many true drugs. For undiagnosed rare-disease patients at the long tail of the indication distribution, this is the direction of harm that matters most. **LLM training-data leakage**: the shallow-aggregation result (0.368) clearly overstates deployable performance, and we cannot bound how much of the conditioned-β gain inherits the same source. **Deployment and the explanation gap**: the cross-attention layer provides graph-level attribution, but as Section 5.3 shows, attention can look healthy when ranking is wrong, so any deployment must pair attribution with a graph-coverage diagnostic. PrimeKG's ground truth is also incomplete; clinically plausible repurposing candidates absent from indication or off-label edges are penalized, biasing evaluation against exactly the novel suggestions the system is designed to surface.

5.3. Failures and Surprising Findings

Two findings inverted our initial expectations. First, **cold-start was not the main rescue mechanism**. We expected the LLM signal to compensate for the R-GCN's lack of supervision on cold-start drugs, but the largest wins (Figure 4) are concentrated in non-cold and partially-cold diseases where the R-GCN had supervision but ranked poorly, likely because popularity bias displaced correct candidates. The all-cold cluster sits low on both axes. The LLM signal therefore appears to help more as a ranking-stage correction on supervised drugs than as a cold-start rescue.

Second, **plausible attention does not imply correct ranking**. Figure 2 contrasts two test diseases. For Parkinsonian-pyramidal syndrome (left), all ten top-ranked drugs are true indications or off-label uses; attention concentrates on bradykinesia, rigidity, and dopaminergic responsiveness, and the retrieved drugs are dopaminergic agents indicated for the disease—a clinical justification a physician could audit. For Leber hereditary optic neuropathy (right), none of the top ten are indications; attention still concentrates on *correct* phenotypes (optic neuropathy, optic atrophy) but retrieves drugs associated with inflammatory optic neuritis in PrimeKG rather than the mitochondrial cofactor pathway LHON requires. Cross-attention can select clinically relevant phenotypes while still producing the wrong ranking, because it can only weight evidence the graph contains. Interpretability tracks graph coverage, not biological ground truth, arguing for pairing attention attributions with a graph-coverage diagnostic at deployment.

6. Conclusion and Future Work

We formulated diagnosis-free drug repurposing for undiagnosed rare disease patients as a phenotype-set $\rightarrow$ drug ranking task on PrimeKG, targeting a clinical gap where existing drug repurposing models require a confirmed disease label and phenotype-driven diagnostic models stop at diagnosis. To address this gap, we proposed an end-to-end graph–LLM hybrid that combines an R-GCN encoder with drug-conditioned cross-attention and biomedical text embeddings. Our main finding is that score-level fusion is the strongest leakage-controlled method across both evaluation splits. It outperforms the PubCaseFinder $\rightarrow$ TxGNN cascade on the full 108-disease test set despite never receiving a disease label, and the contrast is sharper on the 78-disease subset where upstream diagnosis fails: the cascade drops to near-PageRank performance, while score-level fusion approximately triples its MRR and recovers about 76% of the TxGNN oracle ceiling. More broadly, our results show

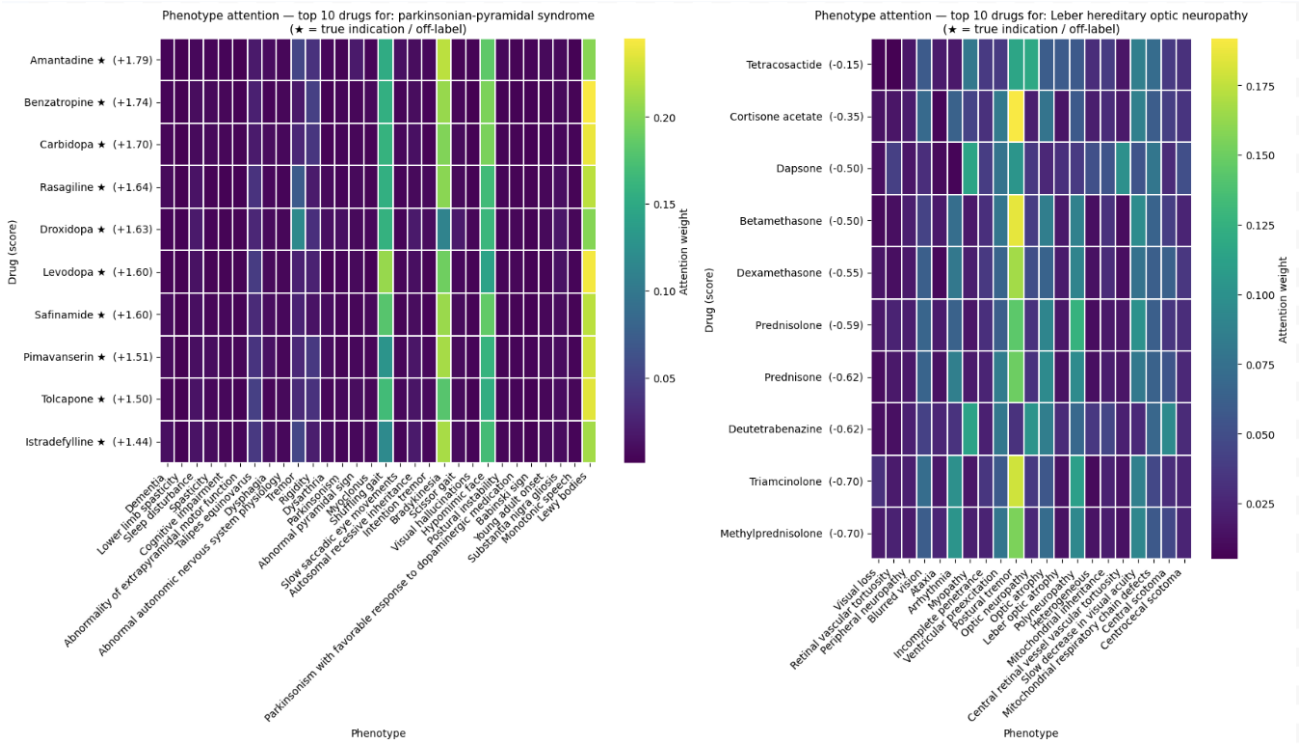


Figure 2. Phenotype attention for top-10 drugs in two contrasting cases. **Left:** Parkinsonian-pyramidal syndrome, where attention highlights canonical motor features and all top drugs are true dopaminergic indications/off-label uses (⋆). **Right:** LHON, where attention highlights relevant optic-neuropathy phenotypes but retrieves corticosteroids linked to inflammatory optic neuritis rather than mitochondrial cofactors.

that LLM-derived signals can improve diagnosis-free drug ranking when fused at the score level, but not when injected into learned graph representations, likely due to supervision scarcity and signal dilution. These findings support score-level fusion as our recommended configuration for diagnosis-free deployment and highlight the importance of evaluating drug repurposing models under realistic no-diagnosis settings.

We identify three directions for future work. First, **moving from disease-level to patient-level phenotype sets**: real patients present noisier, sparser profiles than disease-level HPO sets. Future work could evaluate on real patient cohorts (e.g., those used in SHEPHERD) and augment training with simulated patient profiles. Second, **flagging unreliable recommendations**: pairing each prediction with a graph-coverage confidence indicator would let clinicians flag thinly-supported predictions for review, addressing the LHON-style failure mode. Third, **letting the LLM reason explicitly over the graph**: our current architecture uses the LLM only as a fixed source of text embeddings; multi-step reasoning over phenotype-anchored subgraphs would generate inspectable explanations and let clinicians trace whether a wrong recommendation comes from missing graph edges, biased LLM priors, or both.

Overall, our work motivates phenotype-conditioned drug ranking systems that can support therapeutic hypothesis generation when a confirmed diagnosis is unavailable.

Impact Statement

This work aims to support undiagnosed rare disease patients who cannot access existing AI drug repurposing tools. Potential risks, including popularity bias in recommendations, over-reliance on incomplete knowledge graphs, and plausible-looking but mechanistically wrong explanations from cross-attention are discussed in Section 5. Any deployment must pair predictions with coverage diagnostics and clinician oversight rather than treating outputs as standalone recommendations.

LLM Usage Disclosure

Methodologically, GPT-3.5 was used for the zero-shot baseline and as the LLM branch of shallow rank aggregation (Section 3.2, Section 3.4.1); GPT-4o was used to generate drug and phenotype descriptions for late- and feature-level fusion (Section 3.4), with prompts in Appendix C.

References

Alsentzer, E., Li, M. M., Kobren, S. N., Network, U. D., Kohane, I. S., and Zitnik, M. Few shot learning for phenotype-driven diagnosis of patients with rare genetic diseases. *npj Digital Medicine*, 8:380, 2025. doi: 10.1038/s41746-025-01438-9.

Ambesi-Impiombato, A., Cox, K., Ramboz, S., Brunner, D., Bansal, M., and Leahy, E. Enrichment analysis of phenotypic data for drug repurposing in rare diseases. *Frontiers in Pharmacology*, 14:1128562, 2023. doi: 10.3389/fphar.2023.1128562.

Chakraborty, S., Bisong, E., Bhatt, S., Wagner, T., Elliott, R., and Mosconi, F. BioMedBERT: A pre-trained biomedical language model for QA and IR. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pp. 669–679, 2020. doi: 10.18653/v1/2020.coling-main.59.

Chandak, P., Huang, K., and Zitnik, M. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10:67, 2023. doi: 10.1038/s41597-023-01960-3.

Fujiwara, T., Yamamoto, Y., Kim, J.-D., Buske, O., and Takagi, T. PubCaseFinder: A case-report-based, phenotype-driven differential-diagnosis system for rare diseases. *American Journal of Human Genetics*, 103(3):389–399, 2018. doi: 10.1016/j.ajhg.2018.08.003.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2021. doi: 10.1145/3458754.

Gu, Y., Xu, Z., and Yang, C. Empowering graph neural network-based computational drug repositioning with large language model-inferred knowledge representation. *Interdisciplinary Sciences: Computational Life Sciences*, 17(3):698–715, 2025. doi: 10.1007/s12539-024-00654-7.

Hu, Z., Dong, Y., Wang, K., and Sun, Y. Heterogeneous graph transformer. In *Proceedings of The Web Conference (WWW)*, pp. 2704–2710, 2020.

Huang, K., Chandak, P., Wang, Q., Havaldar, S., Vaid, A., Leskovec, J., Nadkarni, G. N., Glicksberg, B. S., Gehlenborg, N., and Zitnik, M. A foundation model for clinician-centered drug repurposing. *Nature Medicine*, 30:3601–3613, 2024. doi: 10.1038/s41591-024-03233-x.

Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., Mundlos, C., Horn, D., Mundlos, S., and Robinson, P. N. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *American Journal of Human Genetics*, 85(4):457–464, 2009. doi: 10.1016/j.ajhg.2009.09.003.

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K. N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S. A., Haggarty, S. J., Clemons, P. A., Wei, R., Carr, S. A., Lander, E. S., and Golub, T. R. The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313 (5795):1929–1935, 2006. doi: 10.1126/science.1132939.

Marwaha, S., Knowles, J. W., and Ashley, E. A. A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. *Genome Medicine*, 14(1):23, 2022. doi: 10.1186/s13073-022-01026-w.

Nguengang Wakap, S., Lambert, D. M., Olry, A., Rodwell, C., Gueydan, C., Lanneau, V., Murphy, D., Le Cam, Y., and Rath, A. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *European Journal of Human Genetics*, 28(2):165–173, 2020. doi: 10.1038/s41431-019-0508-0.

Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., and Welling, M. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference (ESWC)*, pp. 593–607, 2018.

Singh, A., D'Arcy, M., Cohan, A., Downey, D., and Feldman, S. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5548–5566, 2023.

Wang, X., Ji, H., Shi, C., Wang, B., Cui, P., Yu, P. S., and Ye, Y. Heterogeneous graph attention network. In *Proceedings of The Web Conference (WWW)*, pp. 2022–2032, 2019.

Wright, C. F., FitzPatrick, D. R., and Firth, H. V. Paediatric genomics: diagnosing rare disease in children. *Nature Reviews Genetics*, 19(5):253–268, 2018. doi: 10.1038/nrg.2017.116.

Xiang, C., Ma, T., Fu, X., Liu, Y., Song, B., and Zeng, X. From knowledge to treatment: Large language model assisted biomedical concept representation for drug repurposing. *arXiv preprint arXiv:2510.12181*, 2025.

Xiao, Y., Zhang, S., Zhou, H., Li, M., Yang, H., and Zhang, R. FuseLinker: Leveraging LLM's pre-trained text embeddings and domain knowledge to enhance GNN-based link prediction on biomedical knowledge graphs. *Journal of Biomedical Informatics*, 158:104730, 2024. doi: 10.1016/j.jbi.2024.104730.

Yan, C., Grabowska, M. E., Dickson, A. L., Li, B., Wen, Z., Roden, D. M., Stein, C. M., Embi, P. J., Peterson, J. F., Feng, Q., Malin, B. A., and Wei, W.-Q. Leveraging generative AI to prioritize drug repurposing candidates for Alzheimer's disease with real-world clinical validation. *npj Digital Medicine*, 7:46, 2024. doi: 10.1038/s41746-024-01038-3.

Yasunaga, M., Leskovec, J., and Liang, P. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8003–8016, 2022.

Zhao, W., Wu, C., Fan, Y., Zhang, X., Qiu, P., Sun, Y., Zhou, X., Wang, Y., Zhang, Y., Yu, Y., Sun, K., and Xie, W. An agentic system for rare disease diagnosis with traceable reasoning. *Nature*, 2026. doi: 10.1038/s41586-025-10097-9.

Zheng, Y., Koh, H. Y., Yang, M., Li, L., May, L. T., Webb, G. I., Pan, S., and Church, G. Large language models in drug discovery and development: From disease mechanisms to clinical trials. *arXiv preprint arXiv:2409.04481*, 2024.

A. Stratified and Cold-Start Analysis

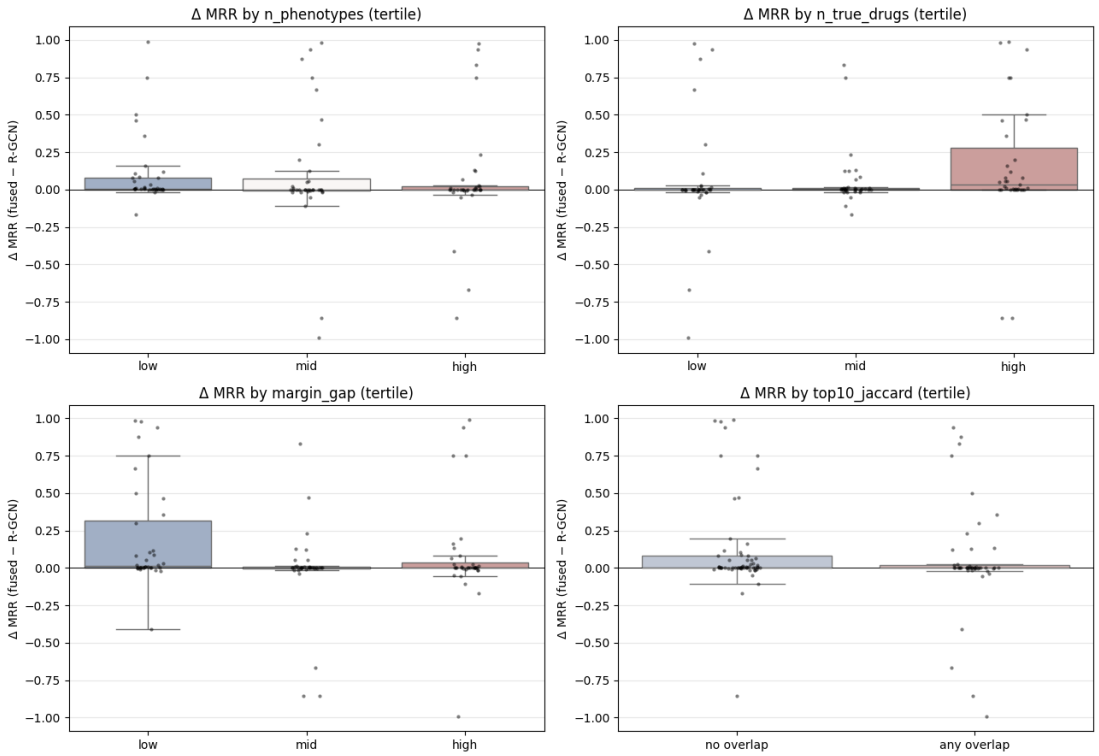


Figure 3. Per-disease ΔMRR (scoring-level fusion − R-GCN) stratified by four candidate gating signals: `n_true_drugs`, `margin_gap`, `n_phenotypes`, and phenotype graph degree. Only `n_true_drugs` (high tertile) and `margin_gap` (low tertile) show systematic separation; the other two are non-discriminative.

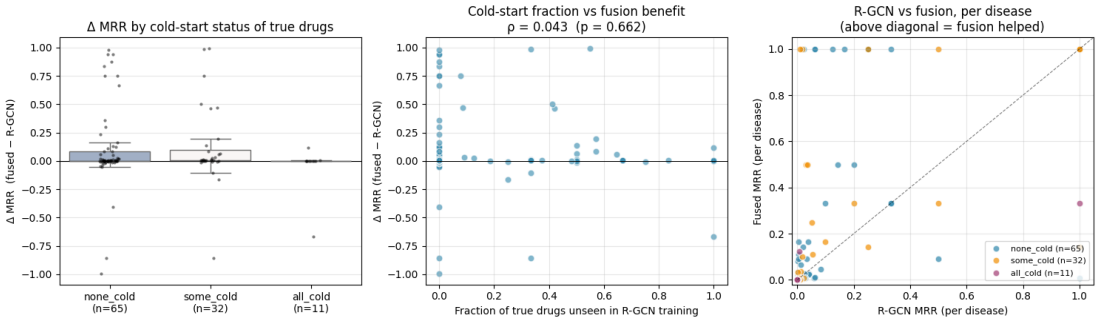


Figure 4. Per-disease fusion gain vs. cold-start status. Diseases are grouped by the fraction of true-positive drugs that are cold-start (no training indication edges). Largest fusion wins are concentrated in non-cold and partially-cold diseases; the all-cold cluster sits low on both axes.

B. Full MRR & Recall@K Results

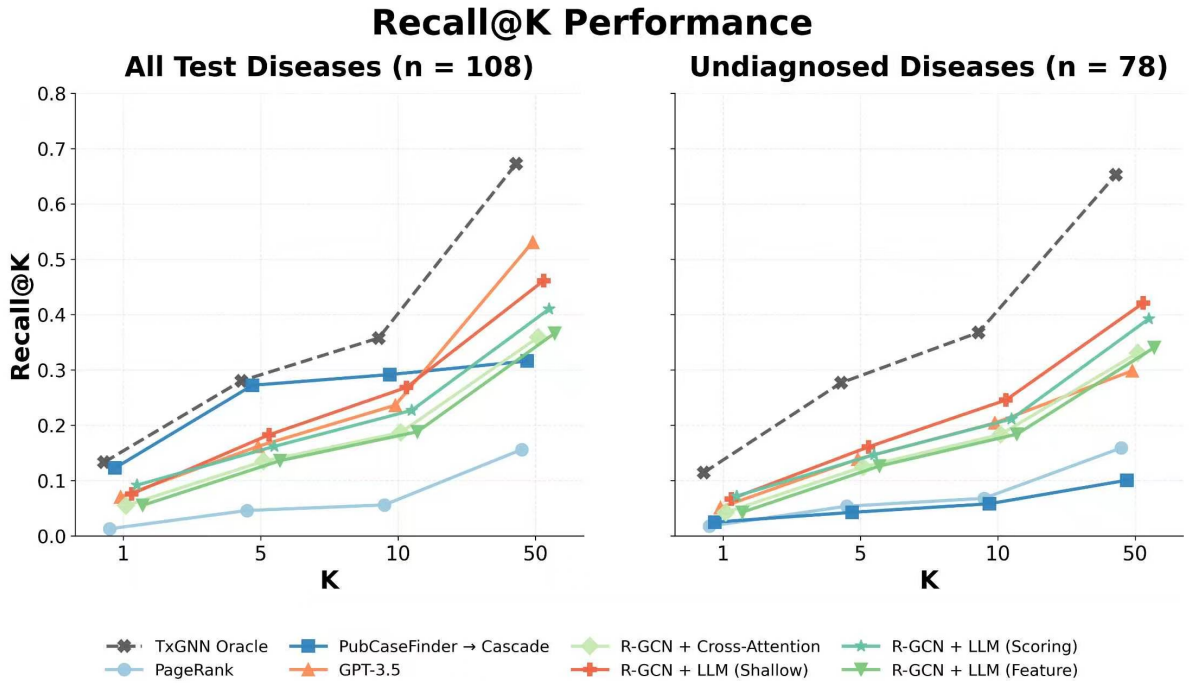


Figure 5. Recall@K across baselines and proposed methods at $K = 1, 5, 10, 50$. **Left:** all test diseases ($n = 108$). **Right:** undiagnosable subset ($n = 78$). Markers are horizontally dodged at each K value to reduce visual overlap.

Table 2. Full Recall@K performance on all held-out test diseases ($n = 108$). Mean Reciprocal Rank (MRR) and Recall@K at $K = 1, 5, 10, 50$ for each baseline and proposed method. Methods marked "possible leakage" use information unavailable at inference for genuinely undiagnosed patients. TxGNN with oracle disease input is included as an upper bound.

Model	MRR	R@1	R@5	R@10	R@50
TxGNN (Oracle Disease, upper bound)	0.404	0.133	0.280	0.358	0.673
PageRank	0.103	0.013	0.046	0.056	0.156
PubCaseFinder → TxGNN Cascade	0.312	0.123	0.272	0.292	0.316
GPT-3.5 (possible leakage)	0.320	0.071	0.162	0.236	0.531
R-GCN + Cross-Attention (ours)	0.254	0.055	0.135	0.187	0.359
R-GCN + LLM, Feature-Level Fusion	0.256	0.055	0.136	0.188	0.366
R-GCN + LLM, Scoring-Level Fusion	0.325	0.092	0.162	0.227	0.410
R-GCN + LLM, Shallow Aggregation (possible leakage)	0.368	0.076	0.183	0.268	0.461

Table 3. Full Recall@K performance on the undiagnosable subset ($n = 78$). ΔMRR% reports the relative MRR change from Table 2; negative values indicate degradation on the harder subset.

Model	MRR	ΔMRR%	R@1	R@5	R@10	R@50
TxGNN (Oracle Disease)	0.407	+0.8	0.114	0.277	0.368	0.653
PageRank	0.106	+2.9	0.018	0.054	0.068	0.159
PubCaseFinder → TxGNN Cascade	0.103	−66.9	0.025	0.043	0.058	0.101
GPT-3.5 (possible leakage)	0.290	−9.4	0.051	0.139	0.204	0.299
R-GCN + Cross-Attention (ours)	0.246	−3.0	0.043	0.125	0.183	0.331
R-GCN + LLM, Feature-Level Fusion	0.247	−3.7	0.043	0.126	0.184	0.341
R-GCN + LLM, Scoring-Level Fusion	0.311	−4.3	0.072	0.147	0.211	0.392
R-GCN + LLM, Shallow Aggregation (possible leakage)	0.371	+0.7	0.067	0.161	0.246	0.421

C. LLM Prompts

C.1. Shallow Re-ranking Prompt

For each test disease, GPT-3.5 is queried directly with the patient's HPO term names. The patient's disease identity is never included, but as discussed in Section 3.4.1 the model's pretraining corpus may map phenotype patterns to memorized disease–drug associations.

> *A patient presents with the following clinical phenotypes:*
> *{phenotypes_str}*
>
> *Based on these symptoms, suggest the {n_drugs} most therapeutically relevant drugs that could treat the underlying condition(s). Return ONLY a JSON array of drug names, ranked from most to least relevant. Use generic drug names (not brand names).*
> *Example format: ["metformin", "insulin glargine", ...]*

{`phenotypes_str`} is the comma-separated list of HPO term names for the patient. We use `n_drugs = 50`. Returned drug names are resolved against PrimeKG's drug vocabulary by exact and fuzzy matching; on average 49 of 50 returned names resolve to a PrimeKG drug node.

C.2. Drug Description Generation Prompt

Used to produce the GPT-4o portion of every drug's hybrid description. Disease names are not explicitly forbidden here because indication terms are part of the drug's clinical identity; leakage is instead controlled at the test-disease level by the entity-level masking in Section 3.4.

> *The following is structured information about a drug from a biomedical knowledge graph:*
> *{kg_metadata_text}*
>
> *Using this as context to identify the drug, write a concise 2-3 sentence scientific description that captures its mechanism of action, therapeutic class, known clinical applications, and any emerging or investigational uses. Draw on your full biomedical knowledge beyond what is listed above.*

{`kg_metadata_text`} is a serialization of the drug's PrimeKG one-hop neighborhood, including the drug name, protein targets, associated pathways, and indication entries (with all 108 test-disease names removed before the prompt is constructed).

C.3. Phenotype Description Generation Prompt

Used to produce the GPT-4o portion of every phenotype's hybrid description. The "do not name specific diseases" instruction is the key leakage control for the late- and feature-fusion variants: it forces GPT-4o to describe the molecular and clinical character of the phenotype without naming the rare diseases that would otherwise leak through cosine similarity to drug indication lists.

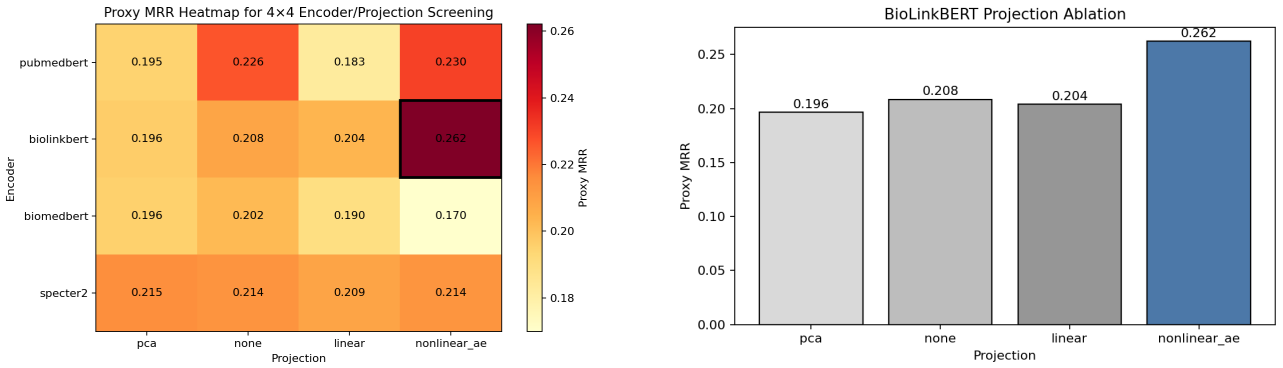
> *The following is structured information about a clinical phenotype from a biomedical knowledge graph:*
> *{kg_metadata_text}*
>
> *Using this as context to identify the phenotype, write a concise 2-3 sentence description covering its molecular basis, associated biological pathways, and clinical presentation. Draw on your full biomedical knowledge beyond what is listed above. Do not name specific diseases.*

{`kg_metadata_text`} here is a serialization of the phenotype's PrimeKG one-hop neighborhood, including the HPO term name, associated genes, and associated conditions (again with test-disease names removed).

D. Text Encoder and Projection Selection

To select the encoder and projection used throughout Section 3.4, we ran a 4×4 grid over four biomedical encoders (PubMedBERT, BioLinkBERT, BiomedBERT, SPECTER2) and four projection methods (PCA, none, linear, nonlinear autoencoder), with all candidates encoding the KG-metadata description tier and projecting to 256 dimensions. Selection was made on proxy MRR, an in-distribution ranking metric computed on the 431 training diseases that uses the same drug–text similarity scoring rule as deployment. We use proxy MRR only for within-tier selection and never for cross-tier comparison. The absolute proxy MRR values are therefore not directly comparable to the test-set MRR reported in Section 4.

BioLinkBERT + nonlinear autoencoder is the within-tier winner at proxy MRR 0.262 (Figure 6a), beating the next-best cell (PubMedBERT + nonlinear AE, 0.230) by 0.032. Within BioLinkBERT, the nonlinear autoencoder projection lifts proxy MRR by 0.054 over the next-best projection (Figure 6b), suggesting that a learned 256-d bottleneck preserves more retrieval-relevant variance than PCA, a single linear layer, or no projection at all. We adopt this configuration for all subsequent late- and feature-fusion experiments.



(a) 4×4 encoder $\times$ projection grid.

(b) BioLinkBERT projection ablation.

Figure 6. Encoder and projection selection on the 431 training diseases. Selected configuration (BioLinkBERT + nonlinear autoencoder, 0.262) outlined in (a).