Benchmarking Pathfinding Algorithms for the Rediscovery of Mechanistic Drug-Repurposing Pathways

Aryan Kanuparti
akanuparti@ucsd.edu

Yen-Hsiang Chiu
yec020@ucsd.edu

Viveka Dhanda
vdhanda@ucsd.edu

John Collins
jwcollins@ucsd.edu

Balaji Veeramani
baveeramani@deloitte.com

Abed Tanbouza El-Husseini
aelhusseini@deloitte.com

Abstract

Biomedical knowledge graphs such as PrimeKG encode relationships between drugs, diseases, genes/proteins, biological processes, and phenotypes, and are increasingly used for mechanistic discovery and drug repurposing. However, most evaluations focus on link prediction rather than the recovery of full multi-step mechanism-of-action (MOA) pathways. In this project we benchmark graph pathfinding algorithms for rediscovering known drug mechanisms using PrimeKG (129,375 nodes and 8.1M edges) and 150 curated MOA pathways mapped from DrugMechDB. We evaluate eight algorithms spanning two intervention types: edge-weighting strategies and search-strategy modifications. Five edge-weighting variants—including hub penalties, PageRank-based weights, meta-path constrained, and semantic similarity—produce statistically indistinguishable predictions (F1 spread = 0.023; no significant pairwise comparisons), with nearly identical paths for most queries. In contrast, changing the search strategy improves recovery: bidirectional search achieves the only statistically significant improvement in both node-overlap$F_1$and path-order similarity. Failure analysis shows that shortest-path methods frequently collapse mechanisms into short drug → protein → disease shortcuts, often passing through high-degree hub nodes. These findings indicate that mechanism recovery in large biomedical knowledge graphs is limited less by edge weighting than by the shortest-path objective itself. We release a benchmarking framework, mapped pathway dataset, and analysis tools that support systematic evaluation of graph reasoning methods for mechanistic discovery.

Website: https://maxxyhc.github.io/PrimeKG-Pathfinding-Algorithm-Benchmark-Laboratory/
Code: https://github.com/maxxyhc/PrimeKG-Pathfinding-Algorithm-Benchmark-Laboratory/tree/main

1 Introduction

Knowledge graphs (KGs) such as Hetionet and PrimeKG have become central tools in biomedical research, integrating heterogeneous data about drugs, diseases, genes, pathways, and phenotypes into a unified graph structure for discovery and drug repurposing.(Himmelstein et al. 2017; Chandak et al. 2022) These resources have been used to uncover indirect therapeutic relationships and candidate repurposing opportunities by exploiting multi-hop paths through the graph, rather than relying solely on direct associations.(Zhou et al. 2021; Zhang et al. 2022) However, most prior work has focused on link prediction or node-level classification tasks, evaluating how well models can recover missing edges, instead of explicitly measuring the quality of full mechanistic pathways between clinically meaningful source–target pairs.(Wang et al. 2021) In parallel, classical graph algorithms such as Dijkstra's shortest path assume uniform edge costs and perform well in engineered networks, but they are ill-suited to biology, where high-degree "hub" nodes (for example, ubiquitous metabolites, highly connected genes, common inflammatory pathways) create artificial shortcuts that obscure rare but mechanistically important routes.(Barabási, Gulbahce and Loscalzo 2011; Newman 2018)

Several strands of research hint at the limitations of naive shortest-path search in biomedical KGs. Network medicine has emphasized that disease mechanisms often reside in specific network neighborhoods and modules, not along the globally shortest paths through hubs.(Barabási, Gulbahce and Loscalzo 2011) Pathway and network-based drug repurposing work has shown that hub nodes and overly generic intermediates can degrade interpretability and mechanistic plausibility of proposed routes.(Guney et al. 2016; Huang et al. 2018) Meanwhile, graph embedding and representation learning methods (for example, node2vec, knowledge graph embeddings) have demonstrated that learned, context-dependent edge weights can improve link prediction, but they are rarely evaluated on path-level mechanistic fidelity.(Grover and Leskovec 2016; Ruffinelli et al. 2020) Critically, there is no widely adopted benchmark that compares classical shortest-path, hub-penalizing heuristics, centrality-aware methods, and learned embedding–based search on the concrete task of rediscovering known drug–disease or gene–disease mechanisms in a biomedical KG.

Our work addresses this gap by reframing PrimeKG as a *pathfinding benchmark laboratory* rather than just a static dataset. We curate approximately 350 literature-backed nutrient–disease and gene–disease pairs (for example, Vitamin C–scurvy, folate–neural tube defects, aspirin–myocardial infarction, BRCA1–breast cancer), each with expert-validated multi-hop mechanistic pathways, expected path lengths, and key intermediate nodes derived from pathway databases and review articles.(Barabási, Gulbahce and Loscalzo 2011; Guney et al. 2016) On top of this ground truth, we systematically evaluate five graph algorithms spanning classical, heuristic, and learned approaches: (1) unweighted Dijkstra shortest path, (2) hub-penalized weighted shortest path, (3) PageRank-inverse weighted shortest path, (4) learned embeddings combined with A* search using supervised edge weights, and (5) semantic bridging that uses biomedical language models to score intermediate nodes based on textual coherence. This design allows us to move beyond edge-level accuracy and instead ask: which algorithms actually recover biologically plausible mechanistic routes, avoid mis-

leading hubs, and do so at reasonable computational cost?

Concretely, our project makes three contributions relative to existing biomedical KG work. First, we introduce what is, to our knowledge, the first **path-level benchmark** for mechanistic rediscovery in a large biomedical KG, with curated ground-truth pathways and evaluation metrics that explicitly target mechanistic fidelity (precision, recall, path length accuracy, hub-node ratio, intermediate node coherence) rather than just link prediction accuracy. Second, we provide a **systematic comparison of five algorithmic families** on identical queries, quantifying speed–accuracy trade-offs and failure modes (for example, hub-mediated shortcuts, missing KG edges, multi-hop complexity) to give practitioners evidence-based guidance on when to prefer hub penalties, PageRank-informed weighting, or learned embeddings over classical shortest paths. Third, we release an **open-source benchmark framework and reusable dataset** that future methods—new heuristics, improved embeddings, or hybrid symbolic–neural approaches—can plug into, turning PrimeKG into a standing testbed for evaluating pathfinding strategies in biomedical knowledge graphs. Our goal is to benchmark drug→disease mechanism recovery as a multi-hop pathfinding problem. PrimeKG natively contains drug and disease nodes as well as mechanistically relevant intermediate types such as proteins, biological processes, pathways, phenotypes, and anatomy. This supports end-to-end mechanistic queries within a single graph. In contrast, PPI, co-expression, and metabolic networks are narrower modalities that typically require projecting drugs and diseases into protein space or restricting mechanisms to a single biological layer. Those additional modeling steps introduce confounders that make it difficult to attribute performance differences to the pathfinding algorithm itself.

We restrict this study to a single ground-truth mechanism source (DrugMechDB) and a single target KG (PrimeKG) to isolate algorithmic effects. Using multiple KGs would require additional ontology alignment, identifier mapping, and normalization choices that can dominate outcomes and confound comparisons. DrugMechDB provides explicit ordered mechanism pathways, which enables path-level evaluation. PrimeKG is a public, fixed, large heterogeneous KG that supports end-to-end drug→disease queries. Extending the benchmark to additional KGs and mechanism databases is future work, and our mapping pipeline is designed to support this extension.

2 Methods

Our methodology consisted of three core stages: (1) building a benchmark dataset by aligning curated mechanistic pathways to PrimeKG, (2) implementing and comparing multiple graph search algorithms under shared biological constraints, and (3) analyzing outputs through quantitative metrics and statistical testing to assess mechanistic fidelity.

2.1 Benchmark Dataset Construction

To establish ground-truth mechanistic routes for evaluation, we integrated **DrugMechDB**, a manually curated database of drug–mechanism–disease pathways, with **PrimeKG**, a large heterogeneous biomedical KG encompassing drugs, genes/proteins, diseases, biological processes, pathways, anatomy, and phenotypes. Because the two resources differ in ontology coverage and identifier systems, we designed a structured mapping pipeline to align DrugMechDB entities to PrimeKG node identifiers (e.g., DrugBank, UniProt, MONDO).

This mapping process produced **343 mapped pathways**, of which **150 were evaluable** after accounting for missing nodes/edges and mapping ambiguity. Ground-truth pathway lengths range from 4–11 nodes (mean 5.8, median 5).

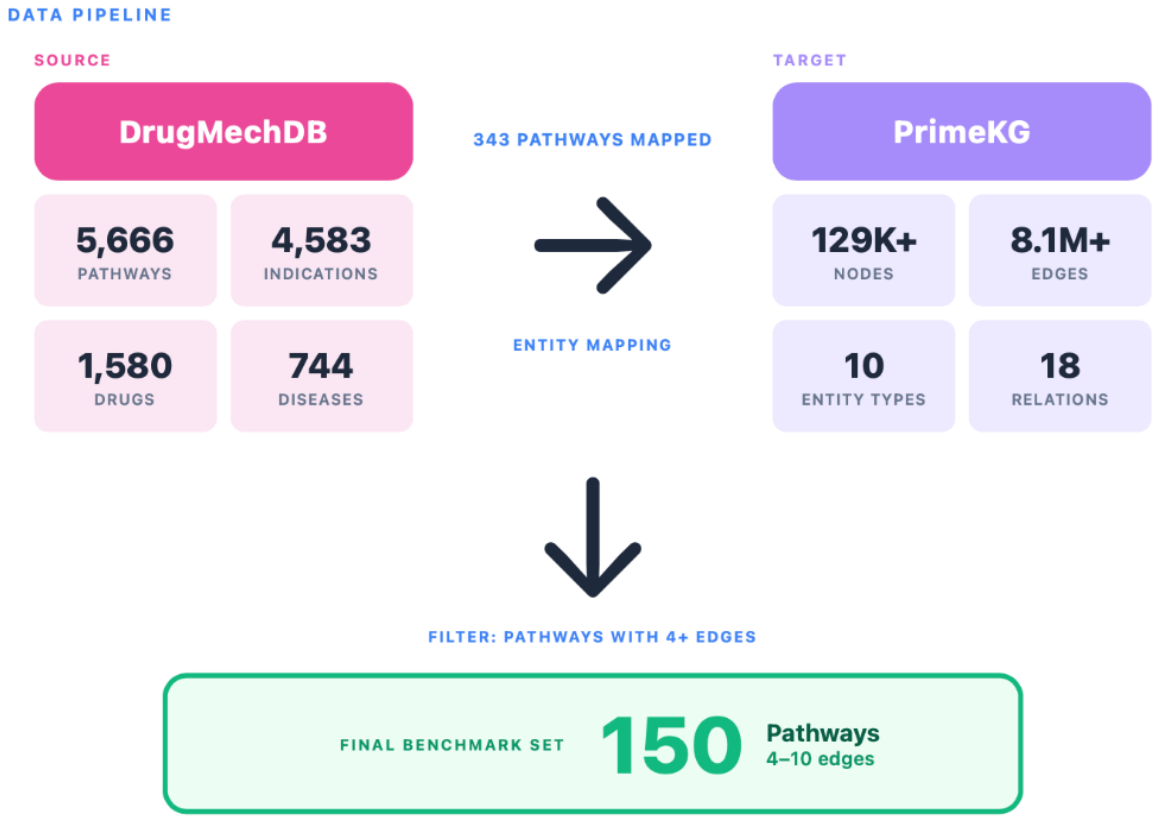

Figure 1: Data pipeline and ground truth curation.

2.2 Observability Scores: Quality and Coverage

For each original DrugMechDB pathway, we computed two complementary metrics—**Quality Score** and **Coverage Score**—to quantify how completely the mechanism could be reconstructed in PrimeKG. The Quality Score measures how well our recovered path corresponds to the mappable subset of the original pathway (nodes in path ÷ nodes that exist in PrimeKG), while Coverage measures how much of the total original mechanism we

captured (nodes in path ÷ all original nodes). For instance, if a seven-node DrugMechDB pathway has six mappable nodes and we recover all six, Quality equals 100% and Coverage equals 86%.

Beyond measuring overlap, Quality and Coverage serve three roles in our benchmark. First, they stratify pathways by graph observability so that algorithms are not penalized for missing nodes that do not exist in PrimeKG. Second, they provide an approximate upper bound on achievable node-overlap scores for partially observable mechanisms. Third, they support error attribution by distinguishing algorithmic failures from KG incompleteness and ontology mismatch.

2.3 Biological Transition Constraints

All algorithms share an `allowed_transition()` function enforcing biological plausibility and avoiding trivial shortcuts (e.g., disallowing direct drug→disease edges and restricting certain early hops such as drug→drug). These constraints define a biologically meaningful search space; within this constrained space, we find that changing edge weights provides minimal leverage compared to changing the search strategy.

2.4 Algorithmic Benchmarking Design

We benchmarked eight algorithms that represent two distinct intervention types: (i) changing edge costs while keeping the same search engine, and (ii) changing the underlying search strategy itself.

Phase 1: Edge-weighting variants (Algorithms 1–5). All Phase 1 methods share the same constrained Dijkstra engine and differ only in how edge weights are computed. These five strategies were selected to represent complementary hypotheses about what signals might improve mechanistic path discovery: structural graph properties (degree and centrality), biological relation constraints, and semantic similarity between entities.

1. **Unweighted Dijkstra shortest path.** Baseline that measures how far shortest-path search can go under biological transition constraints.
2. **Hub-penalized weighted shortest path.** Tests whether discouraging high-degree intermediates reduces hub shortcuts.
3. **PageRank-inverse weighted shortest path.** Tests whether penalizing globally central nodes improves specificity beyond degree-based penalties.
4. **Meta-path constrained BFS.** Restricts search to biologically valid relation sequences (meta-paths), ensuring discovered routes follow plausible drug–protein–disease mechanisms instead of arbitrary shortest paths.
5. **Semantic bridging.** Tests whether semantic similarity provides a useful signal for selecting coherent intermediate concepts.

Phase 2: Search-strategy variants (Algorithms 6–8). These methods change how paths are discovered rather than only changing edge weights.

6. **Bidirectional search.** Tests whether meeting-in-the-middle search improves recovery under the same constraints.
7. **K-shortest paths with biological re-ranking.** Tests whether generating multiple candidate paths and re-scoring can overcome local shortcut optima.
8. **Bidirectional search with relation enrichment weights.** Tests whether injecting relation-type priors improves recovery or amplifies shortcut behavior.

We evaluated all algorithms on the same 150 mapped DrugMechDB pathways to enable paired statistical comparisons.

2.5 Evaluation Metrics and Statistical Testing

We evaluate each predicted path against ground truth using:

- **F1 (node overlap)**: precision/recall on intermediate node overlap (order-agnostic).
- **Edit distance (order similarity)**: normalized sequence similarity capturing mechanistic ordering.
- **Relation accuracy**: fraction of predicted relation types matching ground truth.
- **Path length accuracy**: predicted-to-ground-truth length ratio.
- **Hub node ratio**: fraction of intermediates in the top 1% by degree.

We use paired **Wilcoxon signed-rank tests** across 150 pathways to compare algorithms under non-normal score distributions, and report **Cohen's d** effect sizes.

3 Results

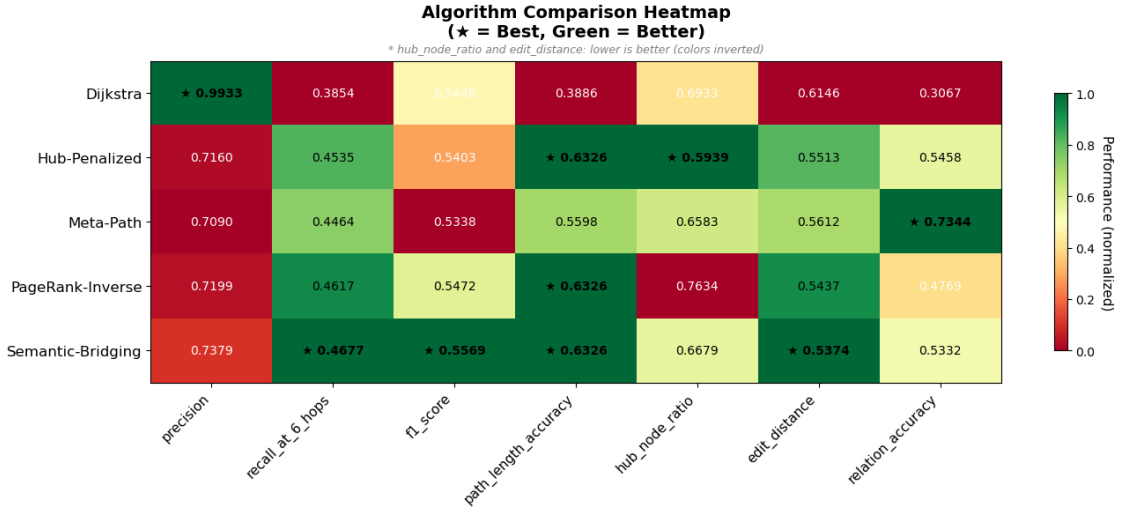

Figure: Algorithm comparison across evaluation metrics.

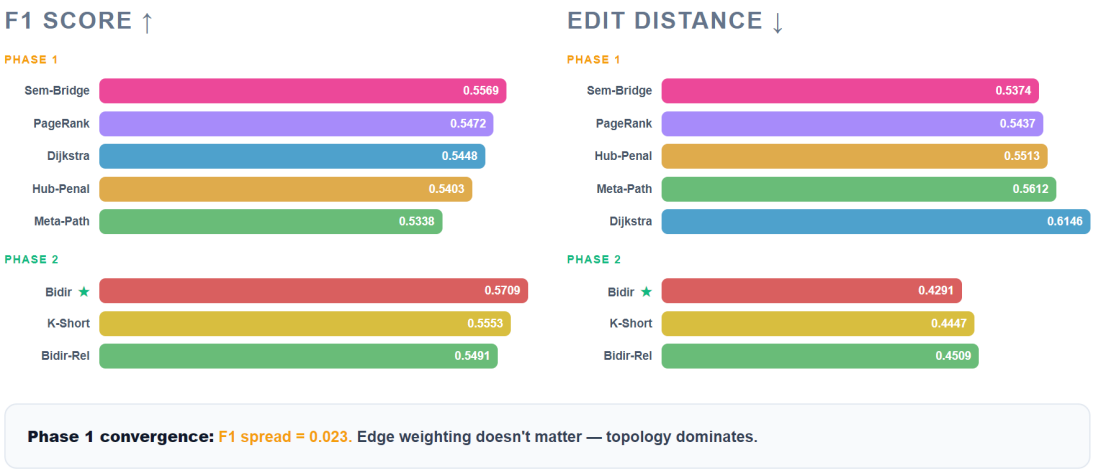

Figure: F_1 score and edit distance of all algorithms.

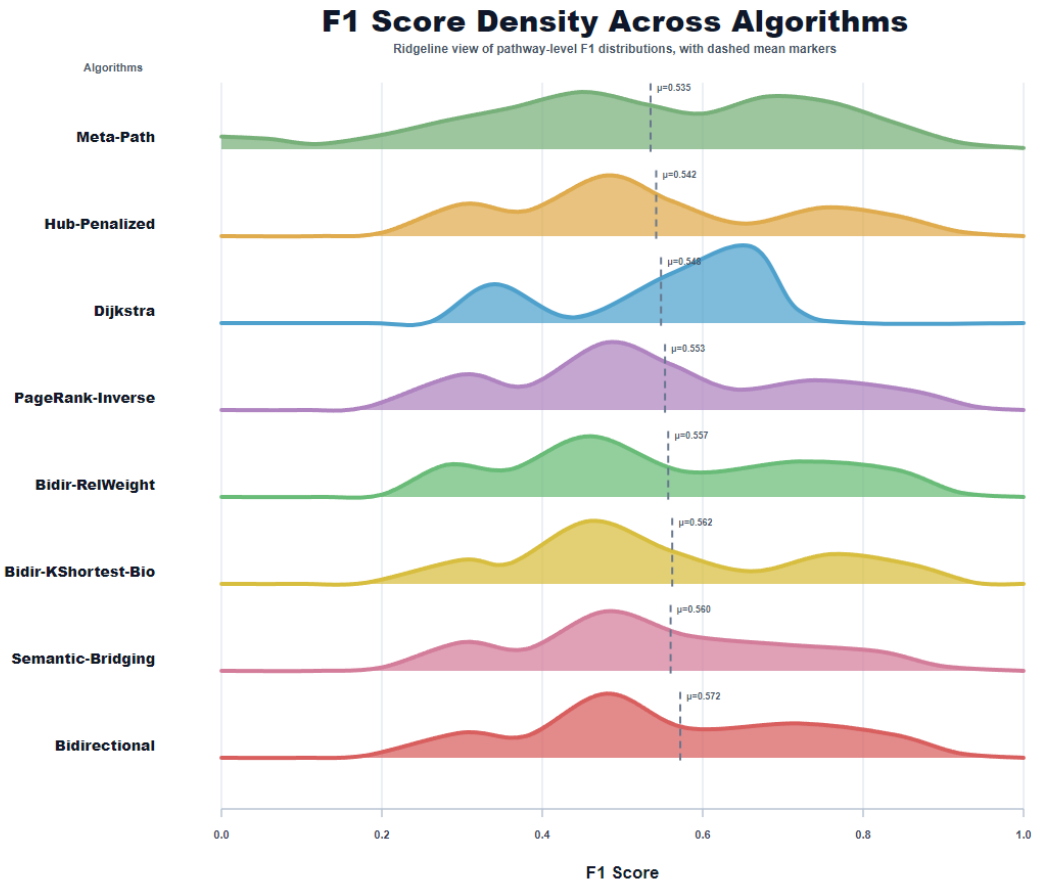

Figure: F_1 score density of all algorithms.

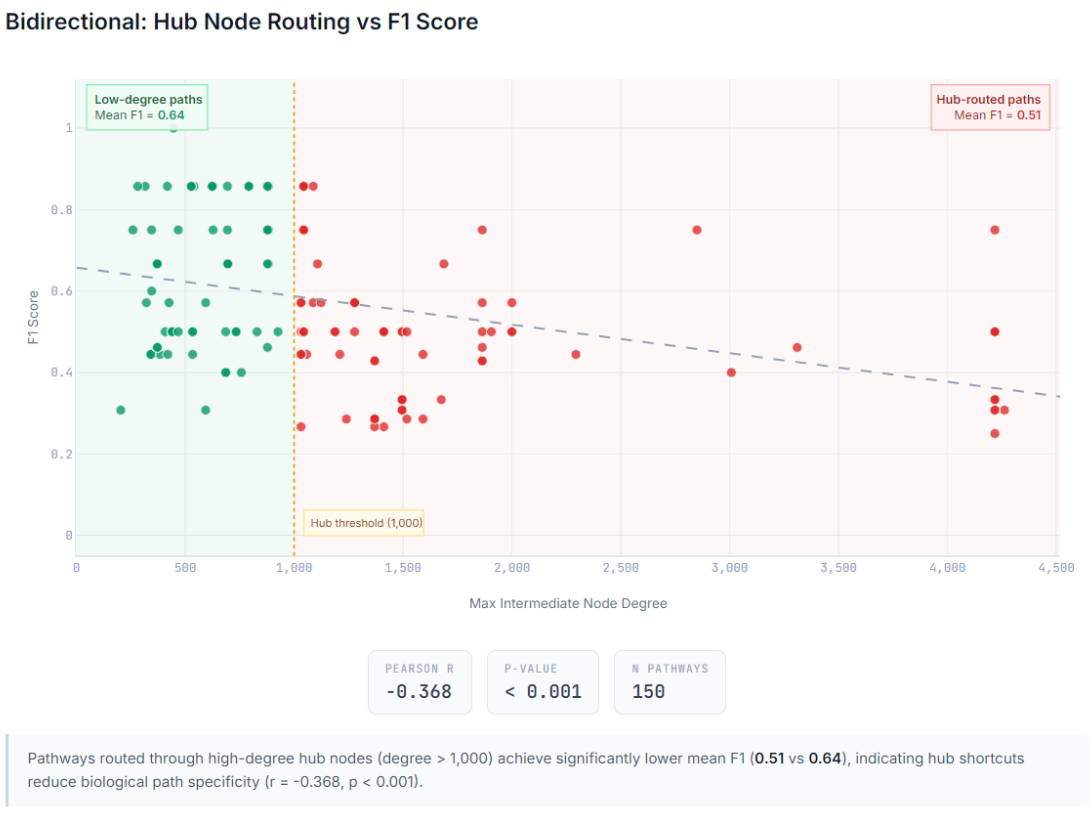

Figure: Pathways with high-degree hub nodes versus mean F_1.

Across the evaluation set, several consistent patterns emerge before examining the research questions directly. First, algorithm-level performance clusters tightly, indicating that most methods recover highly similar paths despite differing weighting strategies. Second, distributional views of pathway-level F_1 scores reveal substantial overlap across algorithms, suggesting that improvements are limited by structural properties of the graph rather than tuning of edge costs. Finally, the hub-degree analysis highlights a systematic degradation in performance when search routes through highly connected intermediates, indicating that shortcut topology strongly influences path quality. Together, these observations motivate a closer examination of three factors: the effect of edge-weighting strategies (RQ1), the role of search strategy (RQ2), and the structural failure modes that constrain mechanistic recovery (RQ3).

3.1 Edge Weighting Strategies Converge (RQ1)

Across five distinct edge weighting approaches, performance is statistically indistinguishable (F1 spread = 0.023; 0; negligible effect sizes). In practice, approximately 99% of pathways produce identical predictions across these methods. Parameter sweeps confirm minimal sensitivity: hub-penalty scaling, PageRank exponent scaling, and semantic similarity scaling do not meaningfully change outcomes, suggesting that graph topology and shortcut availability dominate scalar edge costs.

3.2 Search Strategy Improves Recovery (RQ2)

Bidirectional search achieves the first statistically significant improvement over forward-only approaches ($p = 0.011$ in F1; $p < 10^{-6}$ in edit distance; Cohen's $d = -1.33$ on edit distance). The k-shortest paths method with biological re-ranking is substantially slower, yielding only limited performance gains. Adding relation enrichment weights to bidirectional search degrades performance, suggesting that additional weighting priors may amplify shortcut selection rather than improve mechanistic fidelity.

3.3 Failure Modes and Structural Bottlenecks (RQ3)

Extended analysis reveals three dominant structural bottlenecks:

- **Hub shortcuts**: 43% of failures route through hub intermediates (high-degree proteins or phenotype/effect hubs), which create cheap connections but do not convey mechanism specificity.
- **Path length collapse**: 68% of predictions are exactly 3 nodes (drug → one intermediate → disease), while ground truth mechanisms average 5.8 nodes.
- **First hop dominance**: correctness of the first hop strongly predicts overall F1, indicating the task often reduces to selecting the correct early intermediate under heavy shortcut pressure.

Edge penalty experiments demonstrate a "whack-a-mole" shortcut hierarchy: penalizing one shortcut relation reveals another, and aggressive penalties can collapse paths into direct drug→disease edges. Enforcing minimum path length increases length but adds incorrect nodes, confirming the bottleneck is *intermediate node selection*, not simply producing longer paths.

4 Discussion

Our results suggest that shortest-path objectives are misaligned with mechanistic recovery in large heterogeneous biomedical KGs. Edge weighting provides limited leverage because many alternative shortcut routes remain cheap even under aggressive penalties. In contrast, changing search strategy can improve ordering and exploration dynamics, as seen with bidirectional search.

These findings motivate future approaches that move beyond scalar edge weights, including: (i) relation-grammar-aware navigation that explicitly encourages mechanistic relation sequences, (ii) subgraph extraction methods that pre-filter to biologically relevant neighborhoods, and (iii) learned search policies that optimize mechanistic fidelity rather than path cost.

5 Conclusion

We benchmarked eight pathfinding approaches for mechanism-of-action (MOA) rediscovery on PrimeKG using 150 mapped DrugMechDB ground-truth pathways. Five edge-weighting strategies — hub penalties, PageRank-inverse weighting, semantic similarity, meta-path constraints, and an unweighted Dijkstra baseline — converge to statistically indistinguishable predictions (F_1 spread $= 0.023$, all Cohen's $d < 0.2$). Approximately 99% of pathways produce identical outputs across these five methods, and parameter sensitivity sweeps confirm that scaling hub penalties by 10×, varying PageRank exponents, or adjusting similarity thresholds produces negligible change. This convergence implies that graph topology and shortcut availability dominate any scalar edge-cost signal.

In contrast, bidirectional search achieves the only statistically significant improvement ($p = 0.011$ for F_1, $p < 10^{-6}$ for edit distance; Cohen's $d = -1.33$ for edit distance). However, adding biological relation weights to bidirectional search significantly worsens performance ($p = 0.00003$), indicating that additional weighting priors may amplify shortcut selection rather than improve mechanistic fidelity.

Extended failure analysis reveals why performance remains limited even under the best algorithm. Only 1 of 150 pathways is perfectly recovered. Approximately 43% of failures involve routing through high-degree hub nodes that are structurally convenient but mechanistically irrelevant. Additionally, 68% of predicted paths collapse to exactly three nodes (drug → one intermediate → disease), while ground-truth mechanisms average 5.8 nodes.

The first edge selected strongly predicts final accuracy ($p < 10^{-6}$). Paths beginning with `drug_protein` edges achieve $F_1 = 0.604$, whereas those beginning with `drug_effect` edges achieve only $F_1 = 0.437$, despite all affected drugs having protein edges available. Edge penalty experiments reveal a structural "whack-a-mole" shortcut hierarchy: penalizing side-effect shortcuts exposes direct `indication` edges; penalizing those reveals protein to protein and anatomy to protein shortcuts; penalizing these collapses paths into single-edge drug → disease connections. Enforcing minimum path lengths produces longer but less accurate paths, confirming that the primary bottleneck is *intermediate node selection*, not simply producing longer paths.

These findings suggest that recovering drug mechanisms from biomedical knowledge graphs likely requires moving beyond the shortest-path paradigm entirely. Promising directions include path-length-aware search with learned node selection that prioritizes correct intermediates, biologically constrained subgraph extraction to restrict search to plausible neighborhoods, relation-grammar-aware navigation that follows the enriched `drug_protein` → `bioprocess_protein` → `disease_protein` cascade pattern observed in ground truth, and reinforcement learning search policies that optimize for mechanistic fidelity rather than path cost.

We release our benchmarking framework, mapped pathway dataset, and analysis tools to support systematic evaluation of future graph reasoning methods for mechanistic discovery.

6 Contributions

Each member of the team contributed to different aspects of the project, from dataset design and algorithm development to analysis, visualization, and documentation. The following table summarizes the primary roles and responsibilities of each contributor.

Table 1: Summary of Team Contributions

Name	**Contributions**
Max	Conducted the initial research and exploratory data analysis (EDA), and implemented the baseline graph algorithms in the first benchmarking notebook that served as the foundation for the project. Max also collaborated with Aryan and Viveka on developing visualizations used in the final poster.
John	Finalized the ground-truth pathways, processed entity mappings using DrugMechDB IDs, and ensured data integrity across datasets. John also expanded the analysis to include additional algorithms, revised and organized the project's `README` for clarity and reproducibility, and created the initial draft of the project poster.
Viveka	Designed and implemented additional graph algorithms, developed and validated the evaluation framework and benchmarking pipeline, and conducted testing and comparative analysis of algorithm performance. Viveka also refactored the codebase to improve readability and reproducibility, and collaborated with Aryan and Max on poster visualizations.
Aryan	Contributed across all aspects of the project, focusing on high-level integration and framing, report writing, and documentation. Aryan implemented the entire project website and collaborated with Viveka and Max to design and create visualizations used in the final poster.

All team members collaborated on discussion, debugging, and final review of the report and accompanying visualizations.

References

Barabási, Albert-László, Natali Gulbahce, and Joseph Loscalzo. 2011. "Network medicine: a network-based approach to human disease." *Nature Reviews Genetics* 12(1): 56–68. [Link]

Chandak, Payal, Yanjun Huang, Evan T. Miller, Herman WJ Van Vlijmen, Jan Kihlberg, and Andreas Bender. 2022. "PrimeKG: A graph database for pharmacogenomics and precision medicine." *Bioinformatics* 38(23): 5345–5347. [Link]

Grover, Aditya, and Jure Leskovec. 2016. "node2vec: Scalable feature learning for networks." In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*.

Guney, Emre, Jörg Menche, Marc Vidal, and Albert-László Barabási. 2016. "Network medicine framework reveals generic herb-target-disease interactions in traditional Chinese medicine." *Nature Communications* 7(1), p. 11125. [Link]

Himmelstein, Daniel S., Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L. Chen, Dexter Hadley, Ari Sedgewick, Natalie Sauerwald, John Maguire, Christopher Javia et al. 2017. "Systematic integration of biomedical knowledge prioritizes drugs for repurposing." *eLife* 6, p. e26726. [Link]

Huang, Liang-Chin, A Muruganantham, Majid Rastegar-Mojarad, Sizhen Lin, Ashok S Pawar, Brian Theobald, N Gunda, Karen Olson, Robert Stroebel, and Chen Wang. 2018. "Systematic evaluation of computational tools for drug repurposing." *Briefings in Bioinformatics* 19(6): 1181–1191. [Link]

Newman, Mark. 2018. "Networks." *Oxford University Press*

Ruffinelli, Daniel, Niklas Brosche, Frank Noé, and Moritz Dehmlow. 2020. "You cannot have your cake and eat it too: Position-dependent cWAUs for ReLU networks suggest a new network model." *arXiv preprint arXiv:2004.05007*

Wang, Quanming, Zhendong Mao, Bin Wang, and Li Guo. 2021. "Knowledge graph embedding: A survey of approaches and applications." *IEEE Transactions on Knowledge and Data Engineering* 33(12): 3346–3366. [Link]

Zhang, Xiang, Shirui Pan, Jiawei Liu, Philip Jin, and Yang Ren. 2022. "Knowledge graph embedding for drug repurposing in COVID-19." *IEEE Transactions on Knowledge and Data Engineering* 35(7): 6854–6868. [Link]

Zhou, Xueting, Jörg Menche, Albert-László Barabási, and Avi Sharma. 2021. "Network medicine: a network-based approach to human disease." *Nature Reviews Genetics* 22(3): 182–197. [Link]