CONFIDENCE-GATED MECHANISTIC BRIDGE RETRIEVAL FOR BIOMEDICAL KNOWLEDGE GRAPH REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Biomedical knowledge-graph reasoning requires recovering intermediate mechanistic evidence rather than only connecting a source and target. Semantic guidance can improve graph retrieval, but its value is conditional: useful candidates may be missing, semantically plausible candidates may provide limited downstream retrieval value, and the guidance strength may vary across queries. We introduce **Confidence-Gated Mechanistic Bridge Retrieval**, a graph-constrained framework that separates these challenges into candidate coverage, retrieval utility, and query-specific influence. The method constructs graph-linked Bridges from complementary biological views, estimates retrieval-aware reliability from downstream supervision, and adaptively controls their influence during graph-valid search. On DrugMechDB, the full system improves Path Success from 27.02% to 39.52% over graph-only retrieval. More importantly, controlled analyses show that the improvement does not arise from a single adaptive component: candidate coverage, downstream Bridge utility, and query-specific influence constitute distinct failure modes that require separate interventions. These results support a failure-oriented view of semantic graph guidance, in which availability, downstream value, and query-specific influence should be modeled separately.An anonymous repository for code and artifacts is available at `https://anonymous.4open.science/r/iclr2027-bridge/`.

1 INTRODUCTION

Retrieval-augmented generation (RAG) gives language models access to external evidence during reasoning. When the external knowledge is represented as a graph, retrieval can also exploit explicit entities, relations, and multi-hop connectivity. Recent graph-based RAG methods use graph traversal, query-conditioned subgraphs, hierarchical structures, and learned retrievers to improve structured evidence access (Sun et al., 2024; Ma et al., 2025b; He et al., 2024; Mavromatis & Karypis, 2025; Li et al., 2025; Zhang et al., 2025; Zhuang et al., 2026; Luo et al., 2026). This setting is especially relevant to biomedical reasoning, where knowledge spans drugs, genes, proteins, pathways, biological processes, diseases, and phenotypes (Himmelstein et al., 2017; Chandak et al., 2023). Mechanistic questions place an additional demand on retrieval. They require not only relevant facts, but also the intermediate biological relations that explain how a source entity leads to a target (Gonzalez-Cavazos et al., 2023; Callaghan et al., 2023; Joy & Su, 2026).

Existing methods improve graph retrieval by selecting relations or entities, ranking query-relevant graph regions, and adapting graph organization or search to the query (Zhou et al., 2026; Zhao et al., 2026). Recent work studies graph constraints, relation-aware retrieval, and medical evidence (Luo et al., 2025; Wu et al., 2025; Um et al., 2026). These advances improve which evidence is retrieved. A different issue arises when intermediate semantic information is used to guide graph search. Semantic guidance can direct retrieval toward meaningful mechanisms, but it can also mislead search when the proposed mechanism is incomplete or irrelevant. Its usefulness may also differ across queries. Adaptive RAG work has shown that retrieval behavior can depend on confidence and query properties (Jiang et al., 2023b; Asai et al., 2024; Jeong et al., 2024), but this issue remains less understood when semantic guidance operates inside graph-constrained retrieval.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This setting raises a specific design question: how should an intermediate mechanistic hypothesis enter graph retrieval? Its contribution depends on three decisions made at different points in the pipeline. A relevant intermediate entity must first be available among the candidates; among those candidates, semantic plausibility need not imply value for retrieving the reference mechanism; and even useful guidance need not have the same influence for every query. These are decisions about coverage, downstream utility, and influence, not different names for a single relevance score.

We therefore introduce Confidence-Gated Mechanistic Bridge Retrieval, a graph-constrained framework built around these three decisions. A Mechanistic Bridge is a graph-linked intermediate hypothesis that provides a semantic signal during path search. The framework constructs candidate Bridges from complementary biological views to broaden access to intermediate mechanisms. It then estimates each candidate's reliability from its marginal contribution to downstream retrieval, rather than from textual similarity alone. Finally, a query-level controller determines how strongly the selected Bridges affect path ranking. All search steps remain graph-valid: Bridges prioritize existing paths but do not add edges to the graph. This decomposition also gives each component a clear empirical role: candidate coverage can be measured before search, Bridge reliability can be evaluated under matched candidate and search conditions, and query-specific influence can be compared against constant or reassigned guidance strengths. The resulting experiments therefore isolate what each retrieval decision contributes to mechanistic evidence recovery without requiring every stage to produce a large, monotonic gain in aggregate endpoint accuracy. In particular, matched retrieval conditions separate Bridge utility from candidate access, while reassigned guidance strengths distinguish query-specific influence from average semantic guidance.

Our contributions are:

- **Separating the decisions behind semantic guidance.** We identify a novel three-part decomposition of graph guidance: whether intermediate candidates are available, whether they improve downstream retrieval, and how strongly they should influence each query. This makes the three sources of failure independently testable.
- **Retrieval-aware Mechanistic Bridge guidance.** We introduce a novel graph-constrained framework that constructs Bridges from complementary biological views, estimates their retrieval-derived reliability, and controls their influence per query. Bridges alter path ranking while all retrieved paths remain valid in the original graph.
- **Isolating the effects of each retrieval decision.** A novel set of matched interventions separates candidate access, Bridge usefulness, and query-specific influence. The experiments show broader coverage, stronger mechanistic evidence recovery, and benefits from matching guidance to queries, while exposing limits across retrieval metrics.

2 PRELIMINARIES

2.1 TASK SETUP

We represent a biomedical knowledge graph as $G = (V, E, \mathcal{R})$, where $V = \{v_1, \ldots, v_n\}$ denotes the set of biomedical entities, $\mathcal{R}$ denotes the relation types, and $E \subseteq V \times \mathcal{R} \times V$ denotes directed biomedical relations (Himmelstein et al., 2017; Chandak et al., 2023). Each graph fact is represented as a triple $(v_i, r, v_j) \in E$, and explicit graph structure supports retrieval and reasoning over entities, relations, paths, and query-conditioned graph regions (Das et al., 2018; Sun et al., 2024; Ma et al., 2025b; Yuan et al., 2025; Edge et al., 2024; He et al., 2024). Recent work further explores knowledge-graph-guided retrieval, differentiable KG retrieval, multi-strategy graph retrieval, structured-data reasoning, and the broader integration of large language models with knowledge graphs (Zhu et al., 2025; Gao et al., 2025; Mavromatis et al., 2025; Jiang et al., 2023a; Ma et al., 2025a). Together, these settings motivate retrieval procedures that exploit explicit graph structure while preserving the distinction between graph evidence and auxiliary semantic guidance.

Given a biomedical mechanistic question q, we identify a source entity $e_s \in V$, a target entity $e_t \in V$, and a reasoning intent I_q, and denote the resulting structured query as $Q = (q, e_s, e_t, I_q)$. Our primary setting is endpoint-conditioned retrieval, in which both endpoints are available at retrieval time and the target therefore defines the retrieval endpoint rather than an entity to be predicted. A graph-valid path is written as $p = (e_s, r_1, v_1, \ldots, r_L, e_t)$, where L denotes the number of relations,

and $\mathcal{P}(e_s, e_t)$ denotes the set of graph-valid candidate paths between the endpoints (Das et al., 2018; Lin et al., 2018; Sun et al., 2019; Jiménez Gutiérrez et al., 2025; Guo et al., 2025). The retrieved evidence $G_E \subseteq G$ may contain one or more supporting paths together with their associated entities and relations. Throughout the paper, semantic information may alter the ranking of graph-valid candidates, but it does not create virtual edges or modify graph connectivity.

2.2 Problem Statement

We study retrieval of intermediate mechanistic evidence between known biomedical endpoints. Unlike relation prediction or answer recovery (Das et al., 2018; Lin et al., 2018; Sun et al., 2019), this setting evaluates the biological entities and relations along a retrieved connection. DrugMechDB provides such ordered intermediate relations as curated reference mechanisms (Gonzalez-Cavazos et al., 2023). THis explicitly trace how source and target entities are mechanistically connected

Definition 1 (Mechanistic Bridge). For a query $Q = (q, e_s, e_t, I_q)$ on G, a Mechanistic Bridge h is a query-relevant hypothesis about an intermediate biomedical entity that can be linked to a node of G. A Bridge identifies a candidate waypoint for retrieval; it is neither a graph edge nor evidence that a complete mechanism exists. Let $H_0(Q)$ be the set of grounded Bridge candidates available for Q.

Problem 1 (Mechanistic Evidence Retrieval). Given G, Q, and a retrieval budget B, return an evidence subgraph $G_E \subseteq G$ comprising graph-valid paths from e_s to e_t within B. The objective is to recover intermediate entities and relations relevant to the mechanism in q, rather than endpoint connectivity alone. When reference mechanisms are available, we assess this objective using path success and mechanistic edge recall. Bridge candidates may guide which valid paths are explored and ranked, but cannot change the feasible graph paths.

This formulation separates three choices in semantic guidance: which intermediate hypotheses are available, which of them improve downstream retrieval, and how much their guidance should affect a particular query. The method addresses these choices through candidate construction, retrieval-aware selection, and query-specific influence control, respectively.

3 An Exploratory Study

Graph grounding rules out invented edges, but it does not establish that a grounded intermediate hypothesis will lead search toward the desired mechanism. A Bridge can name a real biological entity and still favor a plausible path that misses the reference relations, despite remaining fully valid in the graph. This creates a question left open by the task formulation: when graph connectivity and the search budget are held fixed, can changing only the intermediate guidance turn a recoverable mechanism into a missed one? A second question follows even if the Bridge is informative: should it exert the same influence on every query?

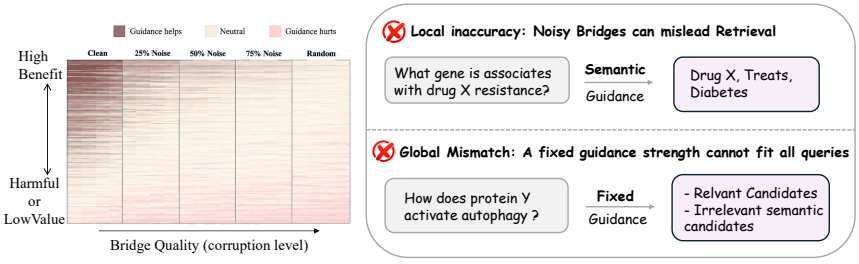

Figure 1: Semantic guidance is conditional: degrading Bridge quality reduces retrieval benefit, while fixed guidance produces heterogeneous responses across queries.

We examine these questions through a controlled corruption study. Starting from reference-aligned Clean Bridges, we progressively replace intermediate entities with type-compatible distractors drawn from the same biomedical graph vocabulary, preserving the form of the guidance while reducing its mechanistic accuracy. Random contains only distractor guidance, and No Bridge provides the graph-only reference. Across these conditions, the graph, source–target pair, and search budget are

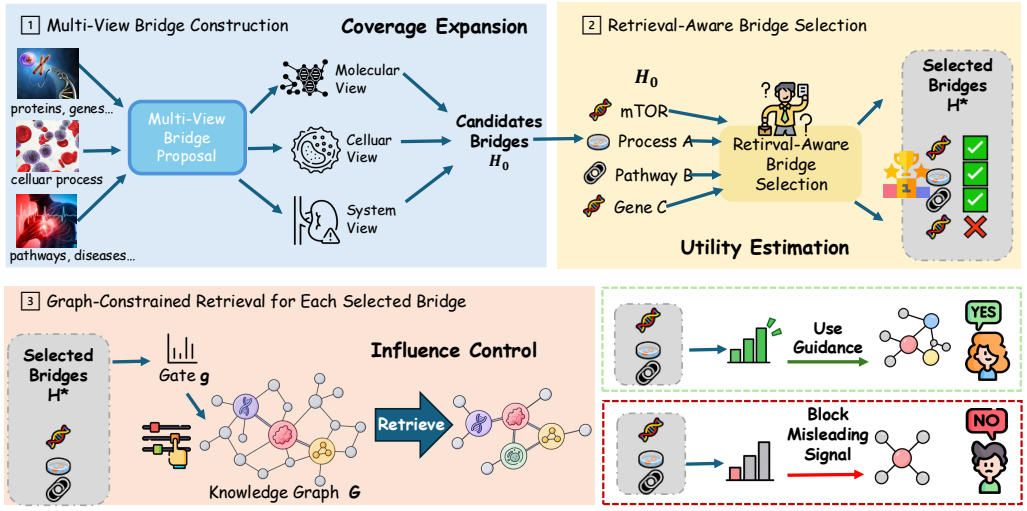

Figure 2: Overview of the three-stage retrieval framework. Multi-view Bridge construction forms H_0, retrieval-aware selection yields H^*, and confidence-gated graph retrieval uses g to control query-specific semantic influence while preserving graph-valid search.

matched. The comparison therefore attributes changes in retrieved evidence to the supplied Bridge information rather than to altered connectivity or search capacity.

Figure 1 reveals two distinct behaviors. First, guidance utility deteriorates with Bridge quality: Clean Bridges benefit more retrieval instances, whereas progressive corruption shifts queries toward neutral or harmful outcomes. Because the graph and search budget are matched, this change isolates the effect of the supplied intermediate information and shows that unreliable Bridges can redirect search away from otherwise recoverable evidence. Second, guidance effects remain heterogeneous across queries even under the same fixed coefficient: some instances benefit from the supplied semantic information, whereas others receive little or negative benefit. This intervention varies supplied guidance rather than candidate generation, so it does not measure whether useful Bridges are available in realistic proposals. Instead, it shows why grounded candidates must be assessed for downstream utility and why their influence should vary across queries.

4 Confidence-Gated Mechanistic Bridge Retrieval

4.1 Overview

We present the overall architecture in Figure 2. Given Q and G, the framework implements the three guidance choices introduced in (Section 2.2). Stage I constructs a graph-linked candidate pool H_0 from complementary molecular, cellular, and system-level views to address candidate coverage (Section 4.2); Stage II estimates retrieval-aware Bridge reliability and reduces H_0 to a reliable and complementary subset H^* to address downstream utility (Section 4.3); and Stage III maps query-level reliability to a continuous guidance strength g that controls the influence of H^* during graph-constrained path ranking (Section 4.4). The resulting search proceeds from candidate availability, to candidate usefulness, to query-specific influence, yielding an evidence subgraph G_E composed only of graph-valid supporting paths. Semantic information changes only the ranking of valid graph candidates and never alters graph connectivity or introduces non-graph edges.

4.2 Mechanistic Bridge Construction

Multi-view candidate construction. Stage I addresses candidate-coverage failure by constructing graph-linked Mechanistic Bridges from complementary biological views. We use three view groups, $\mathcal{V} = \{\text{molecular}, \text{cellular}, \text{system}\}$, so that candidate generation is not restricted to a single abstraction level. In our implementation, the molecular view covers entity types such as proteins, chemical substances, molecular activities, and gene families. The cellular view covers cells, cellular

components, macromolecular complexes, and gross anatomical structures. The system-level view covers biological processes, pathways, phenotypic features, and organism-level entities. For each view $v \in \mathcal{V}$, a view-specific proposer produces a candidate set $\tilde{H}^{(v)} = \mathrm{Prop}_v(Q)$. We combine the candidates as $\tilde{H} = \bigcup_{v \in \mathcal{V}} \tilde{H}^{(v)}$, then normalize surface forms and remove duplicates before graph grounding. The candidate pool is $H_0 = \mathrm{Link}(\mathrm{Dedup}(\mathrm{Normalize}(\tilde{H})), G)$. Only candidates that can be mapped to entities already present in G are retained. Stage I therefore expands the set of available intermediate hypotheses without adding new graph edges or changing graph topology.

Controlled view partition. The same biological-view grouping is used in the controlled Stage-I analysis in Section 5.3, but that partition serves only to isolate candidate coverage. In this setting, held-out Clean candidates are assigned to molecular, cellular, or system-level groups using the same entity-type mapping. Retrieval-aware selection and adaptive influence control are disabled, and all retained Bridges are assigned equal confidence under a fixed semantic guidance strength. The multi-view condition uses the union of the three grounded view-specific candidate sets under the same retrieval procedure. This controlled construction separates the effect of candidate availability from later reliability and influence mechanisms. It therefore provides the methodological basis for evaluating whether Stage I broadens access to potentially useful intermediate evidence.

4.3 Retrieval-Aware Bridge Reliability

Bridge representation. Stage II addresses utility mismatch by estimating whether each candidate Bridge can improve downstream mechanistic retrieval, rather than whether it is only semantically related to the query. For each $h \in H_0$, we construct a representation $x_h = [x_h^{\mathrm{sem}}; x_h^{\mathrm{graph}}]$. The semantic component is built from the task label, query text, Bridge name, and Bridge type using word-level TF–IDF features with 1–2 grams and character-level TF–IDF features with 3–5 grams. The graph component describes pre-search structural context. It includes log-transformed in-degree, out-degree, and total degree, source adjacency, source out-degree, target in-degree, normalized incidence over the 30 most frequent relation types, and node-type indicators. Text vocabularies and numeric standardization statistics are fitted on the training split only. We do not use reference-path membership, target-path oracle information, or post-retrieval outcomes as input features.

Retrieval-derived reliability supervision. We define Bridge usefulness by its marginal effect on retrieval utility $M(H) = \omega\,\mathrm{PathSuccess}(H) + (1-\omega)\,\mathrm{EdgeRecall}(H)$. The default weight is $\omega = 0.7$; its label sensitivity is examined in Appendix C. The marginal contribution of candidate h is $\Delta_h = M(H \cup \{h\}) - M(H)$, and a candidate is labeled useful when $\Delta_h > 0$. This supervision therefore reflects whether adding h improves retrieval under matched graph and search conditions. It does not use semantic plausibility as the target. Given x_h, we estimate the probability of useful guidance as $c_h = \sigma(w_c^\top x_h + b_c)$. In implementation, the estimator is a class-balanced logistic model trained only on the training split. The resulting score $c_h \in [0,1]$ is used as retrieval-aware Bridge reliability.

Reliability-aware Bridge selection. High-confidence candidates can still provide redundant guidance, so Stage II does not select Bridges using c_h alone. Given the currently selected set H, each remaining candidate is scored by $\mathrm{Gain}(h \mid H) = c_h - \mu\,\mathrm{Red}(h, H) + \beta\,\mathrm{TypeBonus}(h, H)$. The redundancy term penalizes overlap with already selected Bridges, while the type bonus favors complementary mechanistic types. Starting from an empty set, candidates are added greedily according to this gain until the Bridge budget is reached. This produces the selected set $H^* \subseteq H_0$, which is passed to Stage III for query-specific influence control.

4.4 Confidence-Gated Graph Retrieval

Query-specific influence control. Stage III addresses influence mismatch by determining how strongly the selected semantic guidance should affect each query during retrieval. Given the selected Bridge set H^*, we construct a query-level representation x_Q from the task label, query text, Bridge names and types, together with graph information available before search. Structural features include normalized candidate-set size, log source out-degree, log target in-degree, candidate-type proportions, source-adjacency rate, and mean log candidate degree. These features describe the query and its semantic context without using shortest-path or post-retrieval information. Query-level supervision is derived directly from retrieval outcomes. For each training instance, we define

$z = \mathbb{I}[\max_{g'>0} M(g'; q, H^*) > M(0; q, H^*)]$, where $M(g'; q, H^*)$ is the retrieval utility under positive semantic guidance and $M(0; q, H^*)$ is the corresponding graph-only utility. Thus, $z = 1$ indicates that positive guidance can improve retrieval for the query. We estimate query-level reliability as $p_{\text{rel}} = \sigma(w_g^\top x_Q + b_g)$ and map it to a continuous strength $g = g_{\max} p_{\text{rel}}^\gamma$. The parameters $g_{\max}$ and γ are selected using validation retrieval utility. Lower reliability therefore reduces semantic influence without imposing a binary guidance decision.

Graph–Bridge path scoring. For each graph-valid path p, we combine structural transitions with agreement between p and the selected Bridges. The graph score $S_{\text{graph}}(p)$ averages local scores $s_{\text{local}}(v_i, r_i) = \lambda q_{\text{rel}}(r_i) + (1 - \lambda)(1 - \tilde{d}(v_i))$. Relation quality favors mechanistic edges, while the degree term discounts hubs. The Bridge score is $S_{\text{bridge}}(p, H^*) = \text{clip}_{[0,1]}[\rho(p) C(p, H^*)]$. The summary C combines Bridge coverage, mean reliability, current endpoint reliability, and recent Bridge hits. Coverage has the largest contribution; the endpoint and recency terms receive equal, smaller contributions. The factor $\rho(p)$ rises as the path reaches more distinct selected Bridges. It gives repeated mechanistic support more weight than one incidental hit. This distinction matters when several paths contain a Bridge but only some reach multiple selected mechanisms. Appendix B specifies the fixed scorer values, and Appendix C reports their local sensitivity.

Graph-constrained retrieval. The path-ranking score is $S(p) = (1 - g) S_{\text{graph}}(p) + g S_{\text{bridge}}(p, H^*)$. The gate changes how graph and Bridge evidence rank valid paths. Search begins at the source and expands only edges present in G. A fixed beam width and expansion budget limit this process. After each step, we retain the highest-scoring frontier states. We exclude repeated nodes to prevent cycles. Search stops at the target or when the budget is exhausted. The retained path then defines the evidence subgraph G_E. The graph and feasible paths remain fixed as the gate changes. The controller does not differentiate through beam search and receives no answer-generation loss. Bridge guidance can prioritize one valid path over another, but it cannot add edges or make an unreachable target reachable. Matched controls use identical graphs and search budgets. They also retain the same reachable candidates and the same stopping rule across guidance settings. This separation matters for interpreting the ablations: improved retrieval cannot be attributed to expanded reachability or a more permissive search. The resulting differences in recovered mechanistic evidence can therefore be attributed to path ranking under a common graph and search budget.

5 Experimental Evaluation

We evaluate the framework through four research questions. **RQ1:** How do semantic guidance, retrieval-aware Bridge selection, and query-specific influence affect the full retrieval pipeline? **RQ2:** Does multi-view Bridge construction reduce candidate-coverage failure? **RQ3:** Can retrieval-aware supervision distinguish downstream-useful Bridges from merely plausible candidates? **RQ4:** Does query-specific influence control outperform uniform semantic guidance by matching guidance strength to each query? Unless otherwise stated, model selection and hyperparameter tuning use training and validation data only, with held-out test sets reserved for final evaluation.

5.1 Experimental Setup

Datasets. We evaluate the core framework on DrugMechDB (Gonzalez-Cavazos et al., 2023), using the BTE-RAG benchmark family (Joy & Su, 2026) of DMDB-Gene, DMDB-Metabolite, and DMDB-Process. Gene and Metabolite evaluate recovery of intermediate mechanistic entities from drug–disease contexts, while Process is reformulated as drug+disease → biological process to avoid answer leakage. EpiKG, released with EpiGraph (Dai et al., 2026), provides an independent biomedical graph used to test whether retrieval-aware Bridge reliability remains informative under a different graph schema and candidate distribution. This setting complements DrugMechDB by changing the graph structure and supervision while preserving the evaluation of graph-grounded intermediate guidance. Additional transfer evaluations on OpenBioLink (Breit et al., 2020), BioHopR (Kim et al., 2025), and ComplexWebQuestions (CWQ) (Talmor & Berant, 2018) are reported in Appendix D. These experiments probe whether the proposed guidance principles remain informative across alternative biomedical graphs, multi-hop retrieval settings, and a general-domain graph. They are used as supplementary evidence rather than as primary benchmarks for the mechanistic retrieval task.

Evaluation Metrics. We use benchmark-appropriate metrics while preserving the same evaluation principle within each task. DrugMechDB reports Path Success (PS), Edge Recall (ER), and Answer Hit (AH): PS measures recovery of a graph-valid source–target path, ER measures recovery of directed edges in the reference mechanism, and AH indicates whether the benchmark answer entity appears in the retrieved evidence. Graph-recovery evaluations use the corresponding path- or edge-recovery metrics, while target-agnostic ranked-answer settings use Hits@K and MRR under their native evaluation protocols. Bridge-level and query-level reliability estimators are evaluated using AUROC, AUPRC, and Brier score. Paired method comparisons use non-parametric bootstrap resampling with 95% confidence intervals. Because task formulations and metric scales differ, scores are interpreted within each benchmark rather than compared directly across datasets.

Baselines. We use controlled variants that isolate structural retrieval, semantic guidance, retrieval-aware Bridge selection, and query-specific influence. Graph Only uses only graph-based retrieval signals, following graph-constrained retrieval settings studied in prior KG reasoning systems (Sun et al., 2024; Luo et al., 2025). Semantic+Fixed adds semantic Bridge guidance with a single fixed influence, reflecting the use of semantic relevance to guide structured retrieval (He et al., 2024; Zhu et al., 2025). Utility+Fixed uses retrieval-aware Bridge selection while retaining the same semantic influence. Full combines retrieval-aware selection with the proposed query-specific continuous controller. For Stage-III analysis, we additionally compare against global, mean-matched, and query-shuffled gate controls, following prior work on adaptive retrieval (Jiang et al., 2023b; Asai et al., 2024; Jeong et al., 2024). Oracle variants are used only as non-deployable upper-bound references. Additional implementation details and search hyperparameters are provided in Appendix B.

5.2 Overall End-to-End Retrieval Context

Full-pipeline context. To answer RQ1, we examine the aggregate retrieval behavior of the complete pipeline through test set before isolating the three retrieval decisions studied in Sections 5.3–5.5. We compare Graph Only, Semantic+Fixed, Utility+Fixed, and Full under the same end-to-end candidate generation, Bridge-selection, and search pipeline, using Path Success (PS), Edge Recall (ER), and Answer Hit (AH). This experiment provides overall pipeline context and is distinct from the controlled Stage-III intervention in Sections 5.5, where candidate access and the selected Bridge set are fixed to isolate query-specific influence.

Table 1: Overall retrieval gains.

Method	PS	ER	AH	ΔER
Graph Only	27.02	33.97	70.56	–
Semantic+Fixed	39.11	58.56	91.53	+24.59
Utility+Fixed	39.52	58.94	91.13	+24.97
Full	**39.52**	**59.46**	**91.94**	**+25.49**

Overall observations. We make three observations. (1) Introducing mechanistic Bridge guidance produces the largest aggregate change over Graph Only, showing that intermediate semantic evidence substantially improves graph-constrained retrieval. This gain is driven primarily by the introduction of reliable Bridges: under the same Clean condition, Fixed Bridge already reaches 58.56% Edge Recall, compared with 59.46% for the full pipeline, indicating that the later reliability and control stages provide smaller refinements rather than accounting for the entire improvement. The same overall advantage is also observed on the independent EpiKG benchmark (Table 1). (2) The subsequent transitions from semantic guidance to retrieval-aware selection and query-specific control produce substantially smaller aggregate changes, indicating that endpoint metrics alone do not isolate the decisions addressed by the individual stages. (3) The full-pipeline results therefore provide overall retrieval context rather than a monotonic stage-wise improvement claim. Scorer interventions (Appendix C) further show that relation quality and multi-Bridge rewards affect mechanistic edge recovery in the frozen-search replay while graph connectivity and candidate access remain unchanged.

5.3 Complementarity and Coverage of Multi-View Mechanistic Bridges

Coverage and controlled retrieval. To answer RQ2, we compare molecular, cellular, system-level, and multi-view Bridge construction while disabling retrieval-aware selection and adaptive influence control. We first evaluate candidate availability and useful-candidate coverage across views, then match the final Bridge budget to separate broader candidate access from fixed-capacity retrieval effects. We further compare task-level retrieval behavior across Gene, Metabolite, and Process queries. Results are summarized in Figure 3.

Stage-I observations. Multi-view construction primarily reduces candidate-coverage failure. Combining biological views increases useful-candidate coverage from 40.32% for the strongest single-view proposal to 61.29%, while the advantage becomes small once the final Bridge budget is matched. This indicates that Stage I mainly broadens access to potentially useful intermediate evidence rather than uniformly improving fixed-budget retrieval. The task-level results further show that the strongest single view varies across Gene, Metabolite, and Process queries, supporting multi-view construction as a coverage and robustness mechanism.Accordingly, Stage I should be interpreted primarily as reducing proposal failure rather than as optimizing retrieval performance in isolation. The remaining gap after candidate access is broadened motivates Stage II, where available Bridges are further distinguished by their downstream retrieval utility.

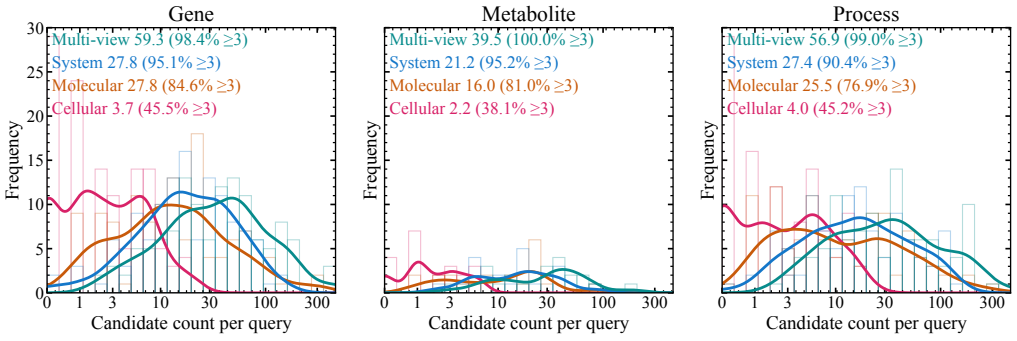

Figure 3: Stage-I candidate availability across Gene, Metabolite, and Process queries. Multi-view construction provides the broadest coverage.

5.4 Retrieval Utility of Learned Bridge Reliability

Usefulness discrimination and controlled selection. To answer RQ3, we compare Lexical Similarity, a local Graph Heuristic, and retrieval-aware reliability on the same held-out Bridge candidates. We evaluate usefulness discrimination on the full and Clean candidate sets, followed by a matched single-Bridge intervention that fixes candidate access, search conditions, and semantic influence while disabling Stage III. We further evaluate the same Stage-II reliability formulation on EpiKG to test whether the learned utility signal remains informative under a different biomedical graph schema and candidate distribution. Because the released EpiBench questions are only partially grounded in the released graph, this evaluation uses a frozen graph-grounded masked-edge setting with disjoint source–target pairs across train, validation, and test splits. Results are summarized in Tables 2 and 3.

Table 2: Stage-II Bridge usefulness and controlled retrieval.

Method	Bridge usefulness				Controlled retrieval (%)		
	All conditions		Clean		Matched intervention		
	AUROC	AUPRC	AUROC	AUPRC	PS	ER	AH
Lexical Similarity	0.594	0.166	0.515	0.257	**29.92**	38.19	75.32
Graph Heuristic	0.391	0.089	0.509	0.238	27.66	36.15	72.90
Retrieval-aware Reliability	**0.795**	**0.393**	**0.815**	**0.655**	29.19	**39.92**	**76.61**

Table 3: Stage-II reliability transfer on EpiKG.

EpiBench grounding		EpiKG Bridge reliability	
Criterion	Value	Metric	Value
Questions audited	1,000	Useful-Bridge prevalence	1.05%
Correct option exactly mapped	10.9%	AUROC	0.881
All four options exactly mapped	4.3%	AUPRC	0.104
At least one option exactly mapped	32.8%	Random AUPRC reference	0.0105
Mean mapped options per question	0.705	AUPRC / prevalence	9.9×

Stage-II observations. Retrieval-aware reliability distinguishes downstream-useful Bridges more clearly than semantic similarity or local structural relevance, including within the Clean candidate subset. This persistence after removing corrupted candidates suggests that the model captures differ-

ences in downstream usefulness among plausible Bridges, rather than merely separating clean from degraded inputs. Under matched retrieval conditions, its advantage translates most consistently into stronger mechanistic evidence recovery. Path success does not improve in parallel, however, so better Bridge ranking should not be read as a uniform gain across retrieval endpoints. The distinction is consistent with the supervision target: Stage II learns which intermediate candidates contribute useful evidence, while complete path recovery also depends on the fixed graph and search constraints. On EpiKG, the utility signal remains informative despite a different biomedical schema and a more imbalanced candidate distribution. Together, these results support retrieval-aware reliability as a way to prioritize useful Bridges across settings, with its clearest effect on the mechanistic evidence recovered by retrieval under matched candidate access, search constraints, and guidance strength.

5.5 Adaptive Control of Semantic Guidance

Query-specific influence control. To answer RQ4, we isolate Stage III on the held-out DrugMechDB test set, fixing the graph, selected Bridges, candidate access, and search constraints. Graph Only removes semantic influence; Best Global α uses the validation-selected coefficient; and Validation-mean Constant α applies the validation mean gate ($\alpha = 0.114$) to every test query. Full uses learned query-specific gates with a similar test-set mean ($\bar{g} = 0.109$). Shuffled Full randomly reassigns these gates across test queries, preserving their distribution while breaking query–gate correspondence. These controls isolate influence after Bridge selection, unlike the end-to-end comparison in Section 5.2 reported on the same held-out split.

Table 4: Stage-III controls for query-specific semantic influence.

Query-specific gate control				Constant-guidance control			
Method	**ER (%)**	**Utility**	**Mean g**	**Method**	**ER (%)**	**Utility**	**Mean g**
Graph Only	33.97	0.2910	0.000	Best Global α	36.35	0.3151	0.050
Shuffled Full gates	37.18	0.3198	0.109	Validation-mean Constant	37.29	0.3179	0.114
Full	**38.05**	**0.3287**	0.109	Full	**38.05**	**0.3287**	0.109

Left: Full exceeds shuffled assignments in 97.6% of runs for ER and 96.0% for utility. Right: the validation-mean constant uses nearly the same average semantic guidance as Full.

Stage-III observations. The controls provide a progressively stronger test of query-specific influence. Full improves over Graph Only and the validation-selected global coefficient, showing that semantic guidance remains useful after Bridge selection is fixed. Its advantage over the validation-mean constant is modest despite similar average guidance strength, so it should not be read as a large gain from stronger guidance. The shuffled-gate control is more diagnostic: reassigning the same learned values across queries reduces mechanistic edge recovery and retrieval utility while preserving the gate distribution. Across repeated random reassignments, correctly matched gates outperform shuffled gates on both measures. Together, these controls support Stage III as a query-specific influence mechanism: matching guidance strength to each query matters under otherwise identical retrieval and search conditions.

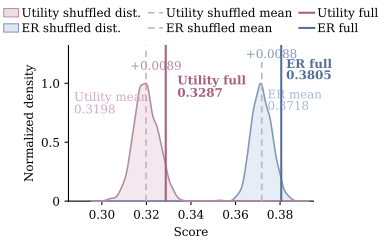

Figure 4: Shuffled Full gates: 500 query reassignments preserve the gate distribution. Correct matching improves Edge Recall and utility.

6 Conclusion

We study how intermediate semantic information can guide biomedical knowledge-graph retrieval while preserving graph-valid search. Our results support a failure-oriented view of semantic guidance: useful intermediate candidates must first be available, their downstream retrieval value must be distinguished from semantic plausibility, and their influence should depend on the query. Confidence-Gated Mechanistic Bridge Retrieval addresses these coverage, utility, and influence failures through multi-view construction, retrieval-aware Bridge reliability, and query-specific control, respectively. Controlled analyses show that these stages correspond to distinct retrieval behaviors rather than a monotonic sequence of accuracy gains, providing a more precise account of when semantic guidance helps and where its limitations arise.

AI USE STATEMENT

Generative AI tools were used to assist with literature retrieval and discovery, as well as editing and debugging software code. Retrieved references were manually verified against their original sources, and AI-assisted code modifications were reviewed and tested by the authors for correctness. Generative AI was not used to generate experimental results or mathematical claims. The authors reviewed all AI-assisted content and take full responsibility for the final content of this work.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. The study does not involve human subjects, clinical interventions, or the collection of personally identifiable information. Our experiments use existing research datasets and biomedical knowledge resources in accordance with their intended research purposes. The proposed method is intended for research on biomedical knowledge retrieval and should not be interpreted as providing clinical or medical advice. We recognize that errors in biomedical knowledge retrieval may propagate incomplete or incorrect information; accordingly, the system is evaluated as a research method rather than as a tool for clinical decision-making.

REPRODUCIBILITY STATEMENT

We provide the information necessary to reproduce the experiments reported in this work. The main paper specifies the proposed method, datasets, evaluation protocols, and experimental settings, while additional implementation details and configurations are documented in the Appendix. Controlled variants and evaluation procedures are described alongside the corresponding experiments to enable direct verification of the reported comparisons. These details are intended to support independent reproduction and verification of our findings.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Anna Breit, Simon Ott, Asan Agibetov, and Matthias Samwald. OpenBioLink: A benchmarking framework for large-scale biomedical link prediction. *Bioinformatics*, 36(13):4097–4098, 2020. doi: 10.1093/bioinformatics/btaa274.

Jackson Callaghan, Colleen H. Xu, Jiwen Xin, Marco Alvarado Cano, Anders Riutta, Eric Zhou, Rohan Juneja, Yao Yao, Madhumita Narayan, Kristina Hanspers, Ayushi Agrawal, Alexander R. Pico, Chunlei Wu, and Andrew I. Su. BioThings Explorer: A query engine for a federated knowledge graph of biomedical apis. *Bioinformatics*, 39(9):btad570, 2023. doi: 10.1093/bioinformatics/btad570.

Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67, 2023. doi: 10.1038/s41597-023-01960-3.

Yuyang Dai, Zheng Chen, Jathurshan Pradeepkumar, Yasuko Matsubara, Jimeng Sun, Yasushi Sakurai, and Yushun Dong. Epigraph: Building generalists for evidence-intensive epilepsy reasoning in the wild. *arXiv preprint arXiv:2605.09505*, 2026.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, Luke Vilnis, Ishan Durugkar, Akshay Krishnamurthy, Alex Smola, and Andrew McCallum. Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning. In *International Conference on Learning Representations*, 2018.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Guangze Gao, Zixuan Li, Chunfeng Yuan, Jiawei Li, Wu Jianzhuo, Yuehao Zhang, Xiaolong Jin, Bing Li, and Weiming Hu. D-RAG: Differentiable retrieval-augmented generation for knowledge graph question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35398–35417. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1793.

Adriana Carolina Gonzalez-Cavazos, Anna Tanska, Michael Mayers, Denise Carvalho-Silva, Brindha Sridharan, Patrick A. Rewers, Umasri Sankarlal, Lakshmanan Jagannathan, and Andrew I. Su. Drugmechdb: A curated database of drug mechanisms. *Scientific Data*, 10:632, 2023. doi: 10.1038/s41597-023-02534-z.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 10746–10761. Association for Computational Linguistics, 2025.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V. Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L. Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E. Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6:e26726, 2017. doi: 10.7554/eLife.26726.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 7036–7050. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.389.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Xin Zhao, and Ji-Rong Wen. StructGPT: A general framework for large language model to reason over structured data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9237–9251. Association for Computational Linguistics, 2023a. doi: 10.18653/v1/2023.emnlp-main.574.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.emnlp-main.495.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From rag to memory: Non-parametric continual learning for large language models, 2025.

Janet Joy and Andrew I. Su. Federated knowledge retrieval elevates large language model performance on biomedical benchmarks. *GigaScience*, 15:giag007, 2026. doi: 10.1093/gigascience/giag007.

Yunsoo Kim, Yusuf Abdulle, and Honghan Wu. Biohopr: A benchmark for multi-hop, multi-answer reasoning in biomedical domain. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 12894–12908. Association for Computational Linguistics, 2025.

Mufei Li, Siqi Miao, and Pan Li. Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation. In *International Conference on Learning Representations*, 2025.

Xi Victoria Lin, Richard Socher, and Caiming Xiong. Multi-hop knowledge graph reasoning with reward shaping. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3243–3253. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1362.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *International Conference on Learning Representations*, 2024.

Linhao Luo, Zicheng Zhao, Gholamreza Haffari, Yuan-Fang Li, Chen Gong, and Shirui Pan. Graph-constrained reasoning: Faithful reasoning on knowledge graphs with large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 41540–41565. PMLR, 2025.

Linhao Luo, Zicheng Zhao, Junnan Liu, Zhangchi Qiu, Junnan Dong, Serge Panev, Chen Gong, Thuy-Trang Vu, Gholamreza Haffari, Dinh Phung, Alan Liew, and Shirui Pan. G-reasoner: Foundation models for unified reasoning over graph-structured knowledge. In *International Conference on Learning Representations*, 2026.

Chuangtao Ma, Yongrui Chen, Tianxing Wu, Arijit Khan, and Haofen Wang. Large language models meet knowledge graphs for question answering: Synthesis and opportunities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24578–24597. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.emnlp-main.1249.

Shengjie Ma, Chengjin Xu, Xuhui Jiang, Muzhi Li, Huaren Qu, Cehao Yang, Jiaxin Mao, and Jian Guo. Think-on-graph 2.0: Deep and faithful large language model reasoning with knowledge-guided retrieval augmented generation. In *International Conference on Learning Representations*, 2025b.

Nicholas Matsumoto, Jay Moran, Hyunjun Choi, Miguel E. Hernandez, Mythreye Venkatesan, Paul Wang, and Jason H. Moore. Kragen: a knowledge graph-enhanced rag framework for biomedical problem solving using large language models. *Bioinformatics*, 40(6):btae353, 2024. doi: 10.1093/bioinformatics/btae353.

Costas Mavromatis and George Karypis. GNN-RAG: Graph neural retrieval for efficient large language model reasoning on knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16682–16699. Association for Computational Linguistics, 2025.

Costas Mavromatis, Soji Adeshina, Vassilis N. Ioannidis, Zhen Han, Qi Zhu, Ian Robinson, Bryan Thompson, Huzefa Rangwala, and George Karypis. BYOKG-RAG: Multi-strategy graph retrieval for knowledge graph question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27881–27898. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1417.

Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4498–4507. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.412.

Haitian Sun, Tania Bedrax-Weiss, and William W. Cohen. Pullnet: Open domain question answering with iterative retrieval on knowledge bases and text. In *Proceedings of EMNLP-IJCNLP*, 2019.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *International Conference on Learning Representations*, 2024.

Alon Talmor and Jonathan Berant. The web as a knowledge-base for answering complex questions. In *Proceedings of NAACL-HLT*, 2018.

Kaehyun Um, KyuHwan Yeom, Haerim Yang, Minyoung Choi, Hyeongjun Yang, and Kyong-Ho Lee. RPO-RAG: Aligning small LLMs with relation-aware preference optimization for knowledge graph question answering. *arXiv preprint arXiv:2601.19225*, 2026.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. Medical graph RAG: Evidence-based medical large language model via graph retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 28443–28467. Association for Computational Linguistics, 2025.

Fan Yang, Zhilin Yang, and William W. Cohen. Differentiable learning of logical rules for knowledge base reasoning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Xujie Yuan, Yongxu Liu, Shimin Di, Shiwen Wu, Libin Zheng, Rui Meng, Lei Chen, Xiaofang Zhou, and Jian Yin. A pilot empirical study on when and how to use knowledge graphs as retrieval augmented generation. *arXiv preprint arXiv:2502.20854*, 2025.

Nan Zhang, Prafulla Kumar Choubey, Alexander Fabbri, Gabriel Bernadett-Shapiro, Rui Zhang, Prasenjit Mitra, Caiming Xiong, and Chien-Sheng Wu. Sirerag: Indexing similar and related information for multihop reasoning. In *International Conference on Learning Representations*, 2025.

Sen Zhao, Lincheng Zhou, Yue Chen, and Ding Zou. Topology of reasoning: Retrieved cell complex-augmented generation for textual graph question answering. In *International Conference on Learning Representations*, 2026.

Zhuoping Zhou, Davoud Ataee Tarzanagh, Sima Didari, Wenjun Hu, Baruch Gutow, Oxana Verkholyak, Masoud Faraki, Heng Hao, Hankyu Moon, and Seungjai Min. Query-aware flow diffusion for graph-based rag with retrieval guarantees. In *International Conference on Learning Representations*, 2026.

Xiangrong Zhu, Yuexiang Xie, Yi Liu, Yaliang Li, and Wei Hu. Knowledge graph-guided retrieval augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 8912–8924. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.449.

Luyao Zhuang, Shengyuan Chen, Yilin Xiao, Huachi Zhou, Yujing Zhang, Hao Chen, Qinggang Zhang, and Xiao Huang. Linearrag: Linear graph retrieval augmented generation on large-scale corpora. In *International Conference on Learning Representations*, 2026.

A Related Work

Knowledge-graph reasoning has progressed from embedding-based relational inference and multi-hop path search (Bordes et al., 2013; Yang et al., 2017; Das et al., 2018; Saxena et al., 2020) toward retrieval systems that combine explicit graph structure with language-model reasoning. Think-on-Graph and Think-on-Graph 2.0 perform iterative relation and entity selection during graph traversal (Sun et al., 2024; Ma et al., 2025b), whereas GNN-RAG and SubgraphRAG learn query-conditioned relevance over graph elements or subgraphs (Mavromatis & Karypis, 2025; Li et al., 2025). Other recent approaches emphasize graph-valid reasoning, relation-aware retrieval, and structured medical evidence (Luo et al., 2024; 2025; Um et al., 2026; Wu et al., 2025). Biomedical knowledge graphs make these retrieval decisions especially important because relevant evidence spans heterogeneous entity and relation types. Hetionet and PrimeKG integrate broad biomedical knowledge (Himmelstein et al., 2017; Chandak et al., 2023), DrugMechDB represents curated mechanisms through ordered intermediate relations (Gonzalez-Cavazos et al., 2023), and BioThings Explorer and BTE-RAG support evidence-grounded biomedical retrieval (Callaghan et al., 2023; Joy & Su, 2026; Matsumoto et al., 2024). In contrast to methods that primarily determine which graph elements or regions should be retrieved, our work studies when intermediate semantic hypotheses should influence an otherwise graph-constrained retrieval process. Adaptive retrieval provides a complementary perspective. FLARE uses model confidence to determine when additional evidence should be retrieved (Jiang et al., 2023b), Self-RAG learns reflection signals for adaptive retrieval and evidence use (Asai et al., 2024), and Adaptive-RAG selects retrieval strategies according to estimated query complexity (Jeong et al., 2024). These methods establish that retrieval behavior should vary across inputs, but they primarily adapt whether or how retrieval is invoked. Our setting instead keeps graph-valid retrieval active and adapts the influence of already selected semantic guidance.

B Implementation Details

Retrieval configuration. All controlled variants operate on the same underlying graph representation and use matched search constraints within each benchmark. Retrieval expands only graph-valid outgoing transitions, and semantic guidance affects path ranking without introducing virtual edges or modifying graph connectivity. For the endpoint-conditioned DrugMechDB experiments, search is initialized from the source entity

and terminates when the target is reached or the search budget is exhausted. We use beam size 2, a maximum of 8 hops, a 14-node budget, and a 20-edge expansion budget. Repeated nodes within a path are excluded. Candidate paths are ranked after each expansion step, and only the highest-scoring beam states are retained. The same search configuration is used across matched DrugMechDB comparisons so that differences between variants arise from semantic guidance rather than changes in search capacity. The exploratory corruption study in Section 3 uses 1,000 retrieval instances and five random seeds under these matched constraints.

Scorer values. The fixed settings are $\lambda = 0.82$, $C(p, H^*) = 0.45C_{\text{cov}} + 0.25C_{\text{mean}} + 0.15C_{\text{cur}} + 0.15C_{\text{recent}}$, and $\rho(p) \in \{0.65, 0.85, 1.0\}$ in ascending order of distinct Bridge hits.

Reliability estimators and feature preprocessing. The Bridge-level reliability estimator in Stage II uses the semantic and structural features defined in Section 4.3. Semantic features are constructed with word-level TF–IDF using 1–2 grams and character-level TF–IDF using 3–5 grams. Both vocabularies are fitted on the training split only. Numeric graph features are standardized using statistics estimated from the training split. Bridge reliability is learned with a class-balanced logistic regression model. The query-level estimator in Stage III follows the same train-only preprocessing principle. Its textual representation combines the task, query, and Bridge-set context, while its numeric representation contains only pre-search graph and candidate-set statistics. Neither estimator uses reference-path membership, shortest-path oracle information, or held-out post-retrieval outcomes as input features.

Model selection and continuous-gate tuning. All learned models and controller parameters are selected using training and validation data only. For the continuous gate, we evaluate $g_{\max} \in \{0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50\}$ and $\gamma \in \{0.5, 0.75, 1.0, 1.25, 1.5, 2.0\}$ on the validation split. Candidate settings are compared using the same retrieval utility used by the method, $0.7\,\mathrm{PathSuccess} + 0.3\,\mathrm{EdgeRecall}$, and the selected pair is then fixed for held-out evaluation. The constrained search is discrete and is not differentiated through, and no downstream answer-generation loss is used to train either reliability estimator or the retrieval controller. Held-out test outcomes are never used for model selection, feature fitting, or hyperparameter tuning.

C Sensitivity of Utility and Search Coefficients

Stage-II utility weight. We recompute usefulness labels from recorded candidate-level retrieval outcomes for $\omega \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$ without refitting the reliability model. Across 3,460 held-out candidates, including 692 Clean candidates, positive rates and frozen-score AUROC/AUPRC are unchanged: 0.795/0.393 overall and 0.815/0.655 on Clean. This post hoc check establishes local label stability, not a uniquely optimal weight or a full retrieval rerun.

Protocol. We replay DrugMechDB Clean with the main experiment's graph, Bridge confidences, query gates, and search budgets. The default replay reproduces all recorded outcomes for the 248 test queries. We vary one score component at a time without retraining or test-set selection. Table 5 reports Path Success (PS), Edge Recall (ER), and paths changed from the default on 239 validation and 248 test queries.

Table 5: Scoring-component interventions on DrugMechDB Clean. PS and ER are percentages; changed paths are counts. Interventions are post hoc, not hyperparameter selection.

Configuration	Validation PS / ER	Test PS / ER	Changed paths (val / test)
Default	29.71 / 54.28	39.52 / 59.46	0 / 0
No relation prior ($\lambda = 0$)	31.38 / 50.63	38.31 / 55.20	107 / 97
No hub term ($\lambda = 1$)	28.45 / 53.41	37.10 / 59.17	49 / 38
No one-Bridge reward ($\rho_1 = 0$)	29.71 / 50.40	37.10 / 54.22	24 / 29
No two-Bridge reward ($\rho_2 = 0$)	23.01 / 35.41	31.05 / 39.11	138 / 152
Mean-confidence term only	28.03 / 48.48	35.48 / 51.85	49 / 66

Interpretation. Removing relation quality or Bridge rewards changes retrieval, most strongly when the two-Bridge reward is removed. Nearby coefficients are less identifiable: changing 0.82 to 0.819, 0.81, or 0.83 changes no path on either split. One-at-a-time ± 0.01 changes to Bridge weights or the first two saturation factors leave PS and ER unchanged; ± 0.05 changes shift ER by at most 0.084 validation and 0.081 test percentage points. Thus the components matter, but no default coefficient is uniquely optimal.

The original $0.012|p|$ penalty is rank-inert: paths within each beam step have equal length, and search stops upon reaching the target. Replays at zero and 0.2 change no path; Section 4.4 states the effective score. This frozen-search analysis excludes retraining and graph transfer.

Table 6: Transfer retrieval under matched within-benchmark comparisons. Scores are percentages; gains over Graph Only are percentage points. Bold marks the best score within each benchmark.

Benchmark	Retrieval variant	Hits@1	Hits@5	MRR	ΔH@1	ΔMRR
BioHopR	Graph Only	81.33	89.88	85.30	—	—
	Semantic Only	**82.29**	89.18	85.84	+0.96	+0.54
	Full	**82.29**	**90.14**	**86.06**	+0.96	+0.76
CWQ	Graph Only	0.03	0.14	0.51	—	—
	Semantic Only	0.25	2.10	1.68	+0.23	+1.17
	Fixed Bridge	**4.64**	**12.02**	**8.24**	+4.62	+7.73
	Full target-agnostic	4.47	10.95	7.75	+4.44	+7.24

D TRANSFER BEYOND DRUGMECHDB

Scope and protocol. We examine whether graph-valid retrieval with semantic Bridge guidance remains useful when the graph or answer format changes. These experiments test transfer of a retrieval principle, rather than transfer of DrugMechDB's biomedical mechanism labels or its path-success metric. BioHopR (Kim et al., 2025) uses a PrimeKG-based, target-agnostic two-hop entity-ranking task; ComplexWebQuestions (CWQ) (Talmor & Berant, 2018) uses question-specific Freebase subgraphs and target-agnostic multi-hop answer ranking. Gold answers and paths are excluded from test-time scoring in both settings. Within each benchmark, variants share the same graph and candidate access, and model selection uses training and validation data. Hits@K and MRR are reported as percentages, so only within-benchmark comparisons are meaningful.

BioHopR has 1,146 held-out test questions (seed 17); CWQ has 3,531, and its table entries average seeds 17, 29, and 43.

What transfers. On BioHopR, Full improves Hits@1 by 0.96 points and MRR by 0.76 points over structural retrieval. This is evidence that semantic guidance can contribute on a different biomedical graph and an answer-ranking task, but the gain is modest and does not establish a broad mechanistic-reasoning benefit. On CWQ, Bridge-guided retrieval substantially improves over Graph Only despite the shift to general-domain Freebase reasoning: mean Hits@1 rises from 0.03% to 4.47%. A paired bootstrap on seed 17 gives a 95% interval of [3.82, 5.21] percentage points for Full minus Graph Only. However, the fixed-Bridge variant is stronger than Full on all three reported CWQ metrics; the corresponding Hits@1 interval for Full minus Fixed Bridge is [-0.31, 0.08] points. Thus CWQ supports transfer of graph-constrained Bridge guidance, not a consistent advantage for learned query-specific influence. The low absolute CWQ accuracy further limits any claim of competitive general-domain question answering.

OpenBioLink and limits of the evidence. OpenBioLink (Breit et al., 2020) changes the biomedical graph schema and relation distribution and would provide a complementary graph-transfer test. We do not report a numerical OpenBioLink comparison here because a matched run and its per-query outputs are not available for verification. Consequently, the present appendix does not establish robustness across all three proposed transfer settings. The BioHopR and CWQ results support a narrower conclusion: semantic Bridges can remain useful after a change in graph and task format, whereas the incremental value of reliability-controlled influence depends on the target setting. Neither result extends the DrugMechDB mechanistic-path claim to arbitrary graphs or answer-ranking tasks.