HypoKG: Evidence-Disciplined Biomedical Hypothesis Generation Beyond Endpoint Knowledge

Dominic Okonkwo, Adetayo Okunoye, Ismailcem Budak Arpinar
School of Computing
University of Georgia
{dominic.okonkwo, adetayo.okunoye, budak}@uga.edu

Abstract

Large language models (LLMs) can generate biomedical hypotheses, but it remains unclear whether they truly reason from scientific evidence or simply produce convincing-sounding ideas. To study this, we combine three major biological databases: the Kyoto Encyclopedia of Genes and Genomes (KEGG), Rhea, and UniProt, into a unified biochemical knowledge graph and construct a benchmark of 550 paths connecting enzyme sources to rare disease endpoints, yielding 13,200 hypotheses from six LLMs under four conditions varying the biological information each model receives: source enzyme only, full biological path, or source and disease endpoint only. Hypotheses are scored using an expert-derived five-criterion rubric on a 1–5 scale per criterion. We find that models given both the source and disease endpoint often produce the highest-scoring hypotheses, showing that LLMs can generate compelling ideas from minimal information. However, these hypotheses are less grounded in the evidence. In contrast, models given the full biological path generate hypotheses more consistent with known mechanistic relationships. We call this *evidence-disciplined reasoning*. To confirm this effect, we shuffled intermediate path steps while keeping endpoints fixed. Evidence grounding dropped significantly ($\Delta = -0.793$, $p < 0.001$), confirming models genuinely used path structure during reasoning. Our findings show that knowledge graphs support hypothesis generation in two ways: they identify biological endpoint pairs absent from the literature, and their mechanistic paths guide how LLMs reason between them.

1 Introduction

Biomedical hypothesis generation requires more than plausible associations: models must produce mechanistic claims proportionate to the evidence available. Large language models (LLMs) have shown strong scientific reasoning abilities (Brown et al., 2020; Chowdhery et al., 2023) and are increasingly used to generate hypotheses (Yang et al., 2024; Si et al., 2025; Liu et al., 2025; Abdel-Rehim et al., 2025), yet a key question remains unresolved: when LLMs are given knowledge graph (KG) evidence connecting a biological source entity to a disease endpoint, does that structure change the hypothesis produced, or do models exploit parametric memory and generate plausible endpoint-driven explanations?

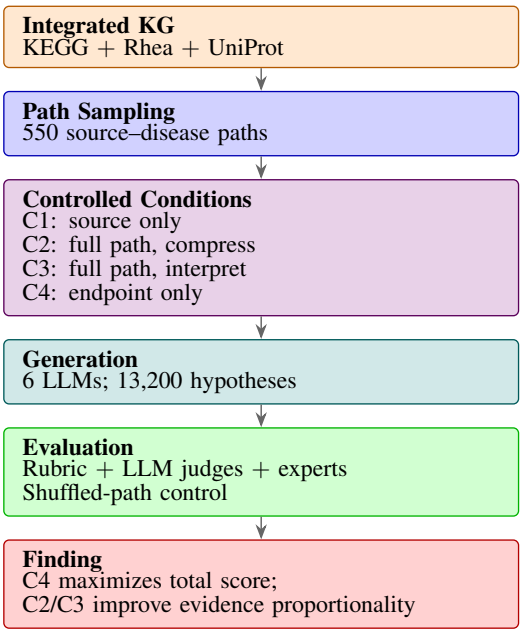


Figure 1: Overview of HypoKG: an integrated KEGG–Rhea–UniProt benchmark for testing whether full KG paths improve evidence-grounded biomedical hypothesis generation beyond endpoint knowledge alone.

This question is especially important in biomedicine, where plausible but weakly supported mechanisms can mislead experimental design. Resources such as KEGG (Kanehisa and Goto, 2000), Rhea (Bansal et al., 2022), and UniProt (UniProt Consortium, 2023) expose multi-hop cross-domain paths connecting enzyme-kinetics source entities to rare disease endpoints, connections often absent from published literature. However, the presence of

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 29262–29278
October 24-29, 2026 ©2026 Association for Computational Linguistics

a KG path does not guarantee that an LLM uses it mechanistically. A model may instead ignore intermediate structure and construct a fluent explanation from the source and disease names alone.

Existing KG-grounded biomedical generation work has focused on factual question answering (Soman et al., 2024), hypothesis generation with hallucination reduction (Xiong et al., 2024), or truthfulness verification (Wen et al., 2024), where evaluation relies on known answers. Hypothesis generation is different: quality is multi-dimensional and KG grounding must be measured against a strong alternative, endpoint-only generation. Prior benchmarks (Xiong et al., 2025; Kumbhar et al., 2025) do not isolate whether full mechanistic paths add value beyond revealing the source–disease endpoint pair.

We introduce HYPOKG, a benchmark for KG-grounded biochemical hypothesis generation built from an integrated KEGG–Rhea–UniProt graph (17,449 nodes; 31,709 edges). We sample 550 multi-hop paths connecting enzyme-kinetics sources to rare disease endpoints (94.5% absent from PubMed co-citation). Six LLMs generate 13,200 hypotheses across four conditions: source-only (C1), full-path compression (C2), full-path interpretation (C3), and endpoint-only (C4), the last testing whether models can generate plausible hypotheses from KG-discovered endpoints without the mechanistic path.

We evaluate hypotheses using an expert-derived five-criterion rubric with a cross-judge LLM-as-judge framework (Pearson $r = 0.776$) validated by three PhD-level domain experts. The results reveal a central tension: endpoint-only prompting achieves the highest total rubric score, but full KG paths uniquely improve evidence proportionality, whether mechanistic claims are supported by path evidence rather than speculation. We call this *evidence-disciplined reasoning*. A shuffled-path control confirms the effect depends on genuine use of intermediate path structure, strongest in larger and domain-specialized models. Our contributions are:

1. **HypoKG benchmark.** A 550-path biochemical hypothesis generation benchmark with 13,200 hypotheses across six LLMs and four controlled grounding conditions, including the first endpoint-only ablation for KG-grounded hypothesis generation.
2. **Evidence-disciplined reasoning.** Full KG paths do not maximize aggregate judged quality; they specifically improve evidence proportionality, constraining models toward grounded mechanistic claims. This finding is validated by expert human evaluation across 55 paths.
3. **Path structure utilization control.** A shuffled-path intervention showing that evidence proportionality gains depend on genuine intermediate path structure, with significant drops when intermediate nodes are permuted while endpoints are held fixed (C2: $\Delta = -0.793$, $p < 0.001$).

2 Related Work

Scientific Hypothesis Generation with LLMs. Recent work has explored LLMs for hypothesis generation across open-domain discovery, data-driven induction, and multi-agent biomedical settings (Yang et al., 2024; Zhou et al., 2024; Xu et al., 2025), with human studies showing LLM-generated ideas are judged more novel than expert-written ones though with feasibility trade-offs (Si et al., 2025). Other approaches combine causal KGs or literature with empirical data, showing that integrating theory-driven and data-driven signals outperforms either alone (Tong et al., 2024; Borrego et al., 2025; Liu et al., 2025). Unlike these studies, HypoKG explicitly controls what grounding information is available to the model, isolating full KG paths from endpoint knowledge alone.

KG-Grounded Biomedical Reasoning. Large-scale biomedical KGs including Hetionet, PrimeKG, SPOKE, and RTX-KG2 have demonstrated the value of knowledge integration for translational reasoning and precision medicine (Himmelstein et al., 2017; Chandak et al., 2023; Morris et al., 2023; Wood et al., 2022). KG-grounded LLM approaches such as KG-RAG have shown benefits for biomedical question answering and explainable generation, while graph-path methods support relation prediction (Soman et al., 2024; Shi et al., 2025; Bakal et al., 2018). However, most work treats KG access as retrieval support rather than testing how different forms of KG evidence affect open-ended hypothesis generation. HypoKG evaluates multi-hop biochemical paths from an integrated KEGG–Rhea–UniProt graph, focusing on whether intermediate path structure changes the mechanistic claims models produce.

Evaluation and Benchmarking. LLM-as-judge evaluation has become common for open-ended

outputs, with Zheng et al. (2023) showing strong agreement between GPT-4 judgments and human preferences while documenting judge biases. Biomedical hypothesis benchmarks including TruthHypo (Xiong et al., 2025), BioVerge (Yang et al., 2025b), and Kumbhar et al. (2025) evaluate truthfulness, agent self-evaluation, and materials discovery respectively. HypoKG differs by testing a distinct question: whether full mechanistic KG paths provide value beyond revealing the same source–disease endpoints. This endpoint-only ablation is central to separating genuine path utilization from plausible endpoint-driven generation.

3 HypoKG Benchmark

3.1 Benchmark Overview

HypoKG evaluates whether structured biochemical knowledge graph (KG) evidence changes the character of LLM-generated biomedical hypotheses. The benchmark is built around 550 multi-hop cross-domain KG paths connecting enzyme-kinetics source entities to rare disease endpoints. For each path, six LLMs generate hypotheses under four controlled prompting conditions, producing 13,200 hypotheses in total. Table 1 summarizes the graph composition, benchmark scale, and generation setup.

Unlike benchmarks that only compare grounded generation against ungrounded prompting (Xiong et al., 2025; Kumbhar et al., 2025), HypoKG separates two roles of the KG: endpoint discovery and mechanistic path grounding. The endpoint-only condition tests whether models can generate plausible hypotheses once the source-disease pair is revealed, while the full-path conditions test whether intermediate KG structure disciplines the mechanism proposed.

3.2 Integrated Biochemical Knowledge Graph

We construct an integrated human biochemical KG from KEGG (Kanehisa and Goto, 2000; Kanehisa et al., 2023), Rhea (Bansal et al., 2022), and UniProt (UniProt Consortium, 2023). KEGG provides pathway topology and gene–metabolite–pathway relationships. Rhea contributes reaction-level precision including directionality and substrate–product relationships. UniProt provides curated protein annotations and enzyme–disease associations from reviewed Swiss-Prot entries.

All data are restricted to human biochemistry.

Component	Details
Integrated KG	KEGG, Rhea, UniProt
KG size	17,449 nodes; 31,709 edges
Domain types	Enzyme kinetics; metabolic pathways; disease mechanisms; drug targets
Benchmark paths	550
Source entities	417 unique enzyme-related entities
Terminal diseases	366 unique disease endpoints
Crossing counts	3–9
Models	6
Conditions	4
Total hypotheses	13,200

Table 1: Summary of the HypoKG benchmark, including graph composition, benchmark scale, and generation setup.

KEGG pathways are filtered to organism code `hsa` and UniProt proteins to taxonomy identifier `9606`, avoiding contamination from model-organism pathways sharing enzyme nomenclature with human homologs. The resulting graph contains 17,449 nodes and 31,709 edges spanning four domain types connected by six relationship types: pathway links, reaction interactions, enzyme-commission relationships, disease associations, enzyme associations, and metabolic pathway membership.

3.3 Cross-Domain Path Construction

A cross-domain path connects an enzyme-kinetics source entity to a terminal disease endpoint through intermediate biochemical entities. A crossing is a traversal step between distinct domain types. We sample 8,000 candidate paths using constrained shortest-path traversal implemented with NetworkX, requiring 3–12 nodes, 1–9 crossings, and terminal nodes restricted to named disease entities. Higher-crossing paths are oversampled to ensure representation of complex reasoning cases.

From these candidates, we select 550 benchmark paths using a greedy stratified filter that maximizes source-entity diversity, terminal disease coverage, and crossing-count coverage simultaneously. When multiple candidate paths were equivalent under the filter objective, tie-breaking was by `hypokg_id` lexicographic order. The final benchmark contains 417 unique source entities and 366 unique terminal diseases, with crossing counts from 3 to 9 and path lengths from 4 to 11 hops.

To characterize path validity, three domain experts in biochemistry and metabolic disease reviewed all 550 paths. No path was judged invalid. Approximately 40% were classified as di-

Condition	Information provided	Purpose
C1: Source-only	Source entity only	Measures parametric generation without endpoint or path information.
C2: Full-path compression	Full KG path	Tests whether models can compress noisy path evidence into a focused mechanistic hypothesis.
C3: Full-path interpretation	Full KG path	Tests whether explicit mechanistic interpretation before hypothesis generation improves path-grounded reasoning.
C4: Endpoint-only	Source and terminal disease only	Tests whether endpoint knowledge alone explains apparent grounding gains.

Table 2: Controlled generation conditions used to isolate the effect of KG path information on hypothesis generation.

rectly supported, with each step reflecting a clear, biologically plausible relationship. The remaining paths were classified as requiring interpretive review: endpoints and database edges were valid, but one or more intermediate steps passed through high-degree annotation hubs, most prominently pyruvate carboxylase deficiency, aceruloplasminemia, and congenital lactase deficiency. These paths are retained by design. HypoKG integrates three databases specifically to surface cross-domain connections that no single database exposes alone. The intermediate ambiguity they carry is not a flaw, but a feature. It reflects genuine uncertainty at the frontier of biochemical knowledge and tests whether LLMs can reason carefully under incomplete evidence.

3.4 Literature Prevalence Annotation

We conduct a two-pass PubMed co-citation analysis to estimate prior literature coverage of each source–disease connection. The first pass uses exact phrase matching of the raw source symbol and terminal disease name. The second pass expands both entities with validated aliases, reviewed to remove broad or ambiguous terms. Each path is assigned to a prevalence stratum: Novel, Sparse, Emerging, Established, or Well-known. The final distribution comprises 520 Novel (94.5%), 19 Sparse (3.5%), 5 Emerging (0.9%), 4 Well-known (0.7%), and 2 Established (0.4%) paths, confirming that HypoKG primarily evaluates hypothesis generation in the low-evidence regime where models cannot rely on well-studied source–disease associations.

3.5 Generation Conditions

For each KG path, models generate a testable mechanistic hypothesis connecting the source entity to a disease mechanism. Four prompting conditions systematically vary the grounding information available. C1 establishes a baseline and exposes disease-drift behavior, models generating hypotheses about familiar diseases unrelated to the assigned endpoint. C2 instructs the model to identify the most biologically meaningful transition rather than narrating the full path. C3 asks the model to reason about meaningful versus weak transitions before committing to a hypothesis. C4 provides only the source and terminal disease, removing all intermediate structure. Table 2 summarizes the purpose of each condition.

C4 is the critical ablation. The source–disease pair in C4 was itself discovered through KG construction, so C4 tests whether the full mechanistic path adds value beyond endpoint discovery, not whether the KG is dispensable. All prompts require specific biological entities, a falsifiable experimental prediction, and grounding in established biochemistry. Full prompts are in Appendix A.

3.6 Model Suite and Generation Protocol

We evaluate six LLMs spanning general-purpose, biomedical-specialized, proprietary, and open-weight systems: Claude Sonnet 4.6 (Anthropic), GPT-4o (OpenAI), Llama-3.3-70B (Grattafiori et al., 2024), Qwen3-235B (Yang et al., 2025a), BioMistral-7B (Labrak et al., 2024), and MedGemma-27B (Sellergren et al., 2025). This suite compares KG grounding behavior across scale, access type, and domain specialization. Each model generates one hypothesis per path per condition, producing 13,200 hypotheses total. All generation runs used greedy decoding where supported (temperature=0), with no few-shot examples and a 512-token output limit. No sampling was applied. API models (Claude Sonnet, GPT-4o, Llama-3.3-70B, Qwen3-235B) were accessed via commercial APIs. HuggingFace models (BioMistral-7B, MedGemma-27B) were run on A100 80GB with 8-bit and 4-bit quantization respectively.

Truncation at the 512-token limit was rare across models. BioMistral-7B had the highest rate at 3.9% (64 of 1,650 responses), consistent with its tendency to repeat text before completing a hypothesis.

MedGemma-27B had a 0.2% rate (4 of 1,650). All other models had zero truncated responses. Truncated responses were included as-is and scored by the judge on whatever text was generated; given the low rates, truncation is unlikely to have meaningfully affected results.

4 Evaluation Framework

4.1 Automated KG Consistency Metrics

We compute three automated grounding metrics for all 13,200 hypotheses. Source grounding measures whether the hypothesis mentions the source entity by symbol or alias. Terminal grounding measures whether it mentions the terminal disease by name or validated alias. Path entity coverage measures the fraction of intermediate KG entities mentioned.

These metrics measure entity use, not scientific quality. A hypothesis can mention many path entities while remaining mechanistically weak, or mention few while making a coherent focused claim. We treat them as complementary grounding diagnostics rather than primary quality measures. Results are reported in Appendix C (Table C3).

4.2 Expert-Derived Rubric

We evaluate hypothesis quality using an expert-derived five-criterion rubric developed from structured sessions with eight PhD-level domain experts in biochemistry and metabolic disease. Experts scored hypotheses from a held-out calibration path and provided written justifications. Analysis of their responses yielded five criteria: path/task relevance, mechanistic specificity and consistency, experimental testability, non-trivial novelty, and evidence proportionality. Each criterion is scored 1–5 with behavioral anchors and worked examples, yielding total scores from 5 to 25. Criteria are grounded in established hypothesis quality literature, experimental testability in Popper (1959) and Banerjee et al. (2009), mechanistic specificity in Gasparyan et al. (2019) and Misra et al. (2021), with evidence proportionality novel to KG evaluation and derived directly from expert consensus signals. Notably, the rubric was derived from expert sessions conducted before the endpoint-only condition (C4) was introduced, ensuring that evaluation criteria were not designed to favor or penalize endpoint-only generation. Full rubric details are in Appendix B.

Evidence proportionality is central to HypoKG. It measures whether the strength of the mechanistic claim is justified by the evidence available to the generator, distinguishing fluent endpoint-driven speculation from hypotheses genuinely constrained by KG path evidence. Among the five criteria, evidence proportionality is the only one that directly measures whether a model used the provided KG path. The remaining four criteria, relevance, mechanistic specificity, testability, and novelty, reflect scientific quality that any good hypothesis should have regardless of whether a path was shown. Evidence proportionality is therefore the natural focus for a study asking whether KG structure changes how models reason.

4.3 Cross-Judge LLM Evaluation

All hypotheses are scored by two LLM judges using the expert-derived rubric, with no model evaluating its own outputs. GPT-4o serves as primary judge for all models except GPT-4o, whose outputs are judged by Qwen3-235B. Gemini 2.5 Flash serves as secondary judge for all hypotheses. Claude Sonnet was excluded as a judge after an initial evaluation showed a 14% content filter refusal rate on biochemical disease hypotheses. Judge prompts include the generation condition context, which is necessary for scoring evidence proportionality correctly; a model should not be penalized for omitting path entities that were never provided.

Across 13,200 hypotheses, primary and secondary judges achieve Pearson $r = 0.776$ agreement on total score with a mean absolute difference of 2.10 points on the 25-point scale. Agreement is highest for experimental testability and lowest for evidence proportionality, reflecting the greater subjectivity of claim-to-evidence calibration.

4.4 Human Expert Validation

Three PhD-level domain experts score a stratified sample of 55 paths, 10% of the benchmark, on evidence proportionality in a blinded within-path design. For each path, experts score all four condition outputs in randomized order with condition labels withheld. Because all four hypotheses share the same source and terminal disease, the comparison controls for path difficulty and disease familiarity. Expert scores are compared with LLM-judge evidence proportionality scores using Kendall's W and Spearman's ρ, and we report the percentage of paths for which experts confirm the main condition ordering.

4.5 Shuffled-Path Control

To test whether evidence proportionality gains reflect genuine intermediate path use rather than endpoint priors, we introduce a shuffled-path control. Source and terminal endpoints are held fixed while intermediate nodes are permuted, disrupting mechanistic order while preserving entity identity. Paths with fewer than three intermediate nodes cannot be meaningfully permuted and are excluded, yielding 518 shuffled paths.

Models generate hypotheses from shuffled paths under full-path conditions. If models use ordered KG structure, shuffling should reduce evidence proportionality while leaving endpoint-driven plausibility and novelty unaffected, indicating that models respond to ordered intermediate structure rather than merely to the presence of biomedical entities. Because the full-path prompts (C2 and C3) include reasoning instructions not present in C4, this control also isolates the contribution of path content from prompt framing: the prompts are held fixed while only the path content changes.

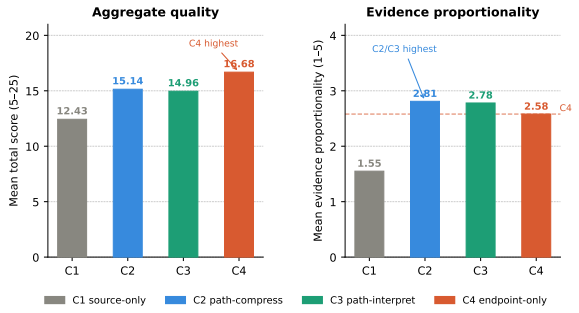


Figure 2: C4 (endpoint-only) achieves the highest total rubric score, but full-path conditions (C2/C3) uniquely improve evidence proportionality - the criterion measuring whether mechanistic claims are supported by path evidence rather than speculation.

5 Results

5.1 Endpoint Knowledge vs. KG Grounding

Figure 2 provides a visual summary of the central pattern: endpoint-only prompting achieves the highest aggregate score, while full-path conditions provide their clearest advantage in evidence proportionality. Table 3 reports the corresponding mean rubric scores by model and condition; all condition comparisons are evaluated over all matched path-model pairs using paired Wilcoxon signed-rank tests with Holm correction ($N = 3{,}300$ per condition, aggregated across all six models), not model-specific subsets.

Endpoint-only prompting achieves the highest total rubric score across all conditions (C4: 16.68), outperforming both full-path conditions (C2: 15.14; C3: 14.96). This shows that LLMs can generate fluent, plausible, and high-scoring hypotheses from source–disease endpoint knowledge alone. However, this does not imply that KG paths are unnecessary: the endpoint pair shown in C4 was itself discovered through KG construction and path sampling, and C4's superiority on four of five criteria reflects not necessarily stronger evidence-grounded reasoning, but greater freedom to construct fluent endpoint-driven mechanisms; models shown only endpoints generate creative, relevance-maximizing hypotheses precisely because they are unconstrained by mechanistic evidence.

The key advantage of full KG paths appears in evidence proportionality. C2 and C3 significantly outperform C4 on this criterion (paired Wilcoxon, Holm corrected; C2 vs C4: $\Delta = +0.23$, $p < 0.001$, $d = 0.256$; C3 vs C4: $\Delta = +0.20$, $p < 0.001$, $d = 0.219$). Evidence proportionality is the criterion most directly measuring whether mechanistic claims are supported by the provided evidence rather than constructed speculatively from endpoint priors. The difference between C2 and C3 on evidence proportionality is not significant after correction ($p = 0.062$), consistent with their close design intent. Thus, full KG paths do not maximize aggregate judged quality; they discipline models toward claims that are more proportionate to available mechanistic evidence. We term this *evidence-disciplined reasoning*.

C1 reveals a complementary failure mode we term *disease-drift bias*. Source-only prompting produces near-minimum path/task relevance (1.01) while retaining high testability (4.04), indicating that models produce well-structured experimental hypotheses even when targeting the wrong disease. Without endpoint or path grounding, models default to familiar high-prevalence diseases, most commonly NAFLD, Alzheimer's disease, and type 2 diabetes, rather than the assigned rare disease endpoint. Path-grounded conditions produce mean total score gains of $+2.71$ (C2) and $+2.53$ (C3) over C1 (both $p < 0.001$, $d > 0.81$), driven primarily by large improvements in path/task relevance ($+1.90$ for C2) and evidence proportionality ($+1.26$ for C2, $d = 1.337$).

Model	Cond.	Rubric Scores					Total	C3–C1	C2/3–C4
		Rel	Mech	Test	Nov	Prop			
Claude Sonnet	C1	1.02	3.24	**4.31**	2.89	1.87	13.32		
	C2	3.05	**3.73**	3.98	3.04	**3.10**	16.89		
	C3	2.93	3.68	4.00	2.89	2.96	16.46		
	C4	**3.75**	3.70	4.28	**3.98**	2.99	**18.70**	+3.14	−2.02
GPT-4o	C1	1.01	**3.86**	**4.36**	2.69	1.01	12.93		
	C2	3.27	3.51	4.00	2.76	2.75	16.29		
	C3	3.39	3.71	3.99	2.94	**2.99**	17.01		
	C4	**4.02**	3.53	4.17	**3.65**	2.00	**17.37**	+4.08	−0.72
Llama-3.3-70B	C1	1.01	2.88	**4.01**	2.74	1.63	12.26		
	C2	2.51	3.02	3.60	2.60	2.60	14.33		
	C3	2.11	2.85	3.35	2.31	2.27	12.88		
	C4	**3.32**	**3.03**	3.89	**3.55**	**2.81**	**16.59**	+0.62	−2.98
Qwen3-235B	C1	1.02	3.42	4.47	2.87	1.91	13.70		
	C2	3.39	3.73	4.05	3.54	**3.41**	18.12		
	C3	2.98	3.66	4.03	3.13	3.04	16.84		
	C4	**3.72**	**3.76**	**4.52**	**4.03**	2.98	**19.01**	+3.14	−1.53
BioMistral-7B	C1	1.01	2.36	**3.11**	2.40	1.37	10.24		
	C2	2.18	2.05	1.69	1.86	1.95	9.74		
	C3	**2.68**	1.92	1.66	1.88	**2.15**	10.29		
	C4	2.44	**2.38**	2.77	**2.72**	2.09	**12.41**	+0.05	−2.39
MedGemma-27B	C1	1.00	2.76	**3.99**	2.84	1.51	12.10		
	C2	3.04	3.02	3.52	2.86	3.02	15.46		
	C3	**3.32**	**3.15**	3.48	3.04	**3.27**	**16.26**		
	C4	2.99	3.01	3.88	**3.54**	2.60	16.02	+4.16	−0.16
All models	C1	1.01	3.09	**4.04**	2.74	1.55	12.43		
	C2	2.91	3.18	3.47	2.78	**2.81**	15.14		
	C3	2.90	3.16	3.42	2.70	2.78	14.96		
	C4	**3.37**	**3.23**	3.92	**3.58**	2.58	**16.68**	+2.53	−1.63

Table 3: Mean rubric scores by model and condition. Rel = Path/Task Relevance, Mech = Mechanistic Specificity, Test = Experimental Testability, Nov = Non-Trivial Novelty, and Prop = Evidence Proportionality. C3–C1 reports the KG-grounding gain; C2/3–C4 reports the gap between path-grounded and endpoint-only generation. Bold values indicate the highest value within each model for each criterion.

5.2 KG Path Utilization Is Model-Dependent

Model-level results in Table 3 reveal a clear capability split. Stronger models, Claude Sonnet, GPT-4o, Qwen3-235B, and MedGemma-27B, show substantial gains from full-path grounding relative to source-only prompting, with C3–C1 improvements ranging from 3.1 to 4.2 points. In contrast, BioMistral-7B and Llama-3.3-70B show near-zero or weak gains (0.05 and 0.62 respectively), suggesting that smaller or less capable models cannot reliably exploit multi-hop KG structure and instead revert to parametric responses regardless of the path shown.

The C2/3–C4 gap further distinguishes models. MedGemma-27B shows the smallest gap (-0.16) and is the only model for which C3 outperforms C4, suggesting that biomedical domain specialization helps translate structured biochemical paths into mechanistically grounded hypotheses that match or exceed what endpoint knowledge alone provides. GPT-4o also shows a relatively small gap (-0.72), while Llama-3.3-70B (-2.98) and BioMistral-7B (-2.39) show large negative gaps, indicating stronger dependence on endpoint-driven generation than path-grounded reasoning.

Together, these results suggest that KG grounding is not automatically useful for all LLMs. Productive use of multi-hop biochemical paths requires sufficient model capacity and, in some cases, domain-specialized pretraining.

5.3 Path Complexity Effects

Table 4 reports total rubric scores by crossing count and condition. KG grounding benefit is strongest at moderate path complexity and weaker for longer, more heterogeneous paths. The C3–C1 gain peaks at crossing count 4 ($+4.72$) and declines at higher crossing counts, with crossing count 8 showing near-zero benefit ($+0.33$). Across path-model pairs, crossing count is negatively correlated with

k	C1	C2	C3	C4	C3–C1
3	12.40	15.43	15.23	**16.67**	+2.83
4	12.00	**17.81**	16.72	16.53	**+4.72**
5	12.46	14.79	14.62	**16.70**	+2.17
6	12.53	15.14	15.04	**16.36**	+2.51
7	12.51	14.09	14.23	**16.87**	+1.72
8	12.67	14.67	13.00	**16.00**	+0.33
9	13.00	14.17	15.11	**17.33**	+2.11

Table 4: Mean total rubric score by path crossing count k and condition. Bold values indicate the highest score within each crossing count.

Metric	Value
C1 mean	1.00
C2 mean	**3.48**
C3 mean	3.18
C4 mean	2.11
Condition ordering	C2 > C3 > C4 > C1
Path-grounded win rate	**90.3%**
C4 speculative flags	90.9%
Kendall's τ	0.703–0.758
Spearman ρ	0.755–0.812

Table 5: Human expert validation of evidence proportionality across 55 paths and three PhD-level domain experts. Scores are mean ratings on a 1–5 scale. Path-grounded win rate denotes C2/C3 outperforming C4; C4 speculative flags were unanimous for 50/55 paths.

C3–C1 grounding gain ($r = -0.223$, $p < 0.001$).

This pattern suggests that moderate-length cross-domain paths provide useful mechanistic scaffolding, while very long paths introduce heterogeneous edge types, intermediate disease nodes, and weak associations that dilute the grounding signal. Qualitative inspection reveals a recurring failure mode we term *local subgraph collapse*: when given long paths, models focus on the strongest local mechanistic transition and ignore either the source entity or the terminal disease, effectively treating a long path as a short one. These results suggest that path complexity should be treated as a benchmark design variable when selecting KG-grounded mechanistic reasoning tasks, since very long paths may dilute the grounding signal with heterogeneous or weakly connected evidence.

5.4 Human Expert Validation

Three PhD-level domain experts evaluated a stratified sample of 55 paths (10% of the benchmark), proportionally sampled across crossing counts 3–9, on evidence proportionality in a blinded within-path design.

Human evaluation confirms the same condition ordering observed in the automated results: C2 > C3 > C4 > C1, matching the LLM-judge ordering exactly across all three experts. Full-path conditions outperformed endpoint-only prompting on the majority of paths, with a mean C2/C3 > C4 win rate of 90.3%. Most strikingly, all three experts unanimously flagged C4 as making speculative mechanistic leaps on 90.9% of paths (50/55), directly confirming that endpoint-only prompting produces plausible but insufficiently grounded mechanisms under independent human evaluation. Inter-expert agreement was strong across all pairs (Kendall's $\tau = 0.703$–0.758, Spearman $\rho = 0.755$–0.812, all $p < 0.001$). Table 5 summarizes the full human validation results.

Metric	Cond.	Orig.	Shuf.	Drop
Total score	C2	14.890	13.601	−1.289
Total score	C3	14.506	13.567	−0.939
Evidence prop.	C2	2.805	2.012	−0.793
Evidence prop.	C3	2.725	2.133	−0.592
Novelty	C2	2.788	2.776	−0.012
Novelty	C3	2.654	2.759	+0.105

Table 6: Original vs. shuffled path scores across all models ($N = 2{,}590$ per condition). All total-score and evidence-proportionality drops are significant at $p < 0.001$; novelty differences are not significant.

5.5 Shuffled-Path Control

Table 6 reports the results of the shuffled-path control described in Section 4.5.

Evidence proportionality drops significantly under shuffled paths (C2: $2.805 \rightarrow 2.012$, $\Delta = -0.793$, $p < 0.001$; C3: $2.725 \rightarrow 2.133$, $\Delta = -0.592$, $p < 0.001$; Wilcoxon, $n = 2{,}590$ per condition). Because endpoints are held fixed, this drop cannot be explained by endpoint knowledge; it isolates the contribution of ordered intermediate path structure. Novelty scores show no significant change under shuffling (C2: $p = 0.215$; C3: $p = 1.000$), demonstrating that path structure specifically disciplines mechanistic grounding rather than general hypothesis creativity.

The drop is strongly model-dependent. Stronger models show large drops (Qwen3-235B C2: $\Delta = -1.110$; Claude Sonnet C2: $\Delta = -0.942$; MedGemma-27B C2: $\Delta = -0.861$), while BioMistral-7B shows a near-zero, non-significant drop under C3 ($\Delta = -0.023$, $p = 0.242$), confirming it was not utilizing intermediate path structure even in the original condition, consistent with its near-zero C3–C1 gains in Table 3. Full per-model results are in Appendix C.

These findings provide causal evidence for the central interpretation of HypoKG: KG paths do not merely reveal useful endpoints; their intermediate structure disciplines stronger models toward more evidence-proportionate mechanistic reasoning.

6 Conclusion

We introduced HypoKG, a benchmark for KG-grounded biochemical hypothesis generation across 550 multi-hop paths, six LLMs, and four controlled grounding conditions. Our results show that endpoint-only prompting achieves the highest aggregate scores, but full KG paths uniquely improve evidence proportionality, disciplining models toward mechanistic claims supported by structured evidence. Human validation and shuffled-path controls confirm that this effect reflects genuine use of intermediate path structure. HypoKG reframes KG grounding as both an endpoint discovery mechanism and an epistemic constraint on biomedical hypothesis generation.

7 Limitations

HypoKG evaluates hypothesis quality through automated LLM judges and a targeted human expert study. Inter-judge agreement is lowest for evidence proportionality ($r = 0.668$), reflecting the subjectivity of claim-to-evidence calibration. Scaling expert evaluation beyond 55 paths was not feasible. The three expert evaluators were PhD-level domain collaborators who participated voluntarily; no personal data was collected and IRB review was not required under our institutional guidelines.

Benchmark paths are drawn from KEGG, Rhea, and UniProt restricted to human biochemistry. While these cover enzyme kinetics, metabolic pathways, and protein-disease associations, other KGs such as SIDER (drug side effects) and HPO (human phenotypes) are not integrated, expanding to these resources is a natural direction for future work. HypoKG also uses constrained shortest-path traversal; other strategies such as fixed-hop or longest-path could surface different mechanistic evidence structures and is worth exploring.

Results may not generalize to models released after our data collection cutoff, and benchmark endpoint pairs may overlap with model pretraining data through individual entity descriptions, though the 94.5% zero-citation rate and the shuffled-path control results reduce this concern. The capability split is based on aggregate rubric scores and may not capture all dimensions of hypothesis quality relevant to practicing scientists. HypoKG measures evidence proportionality as a proxy for grounded reasoning but does not verify whether generated hypotheses are factually correct or experimentally validated.

References

Abbi Abdel-Rehim, Hector Zenil, Oghenejokpeme Orhobor, Marie Fisher, Ross J. Collins, Elizabeth Bourne, Gareth W. Fearnley, Emma Tate, Holly X. Smith, Larisa N. Soldatova, and Ross King. 2025. Scientific hypothesis generation by large language models: Laboratory validation in breast cancer treatment. *Journal of the Royal Society Interface*, 22(227):20240674.

Gokhan Bakal, Preetham Talari, Elijah V. Kakani, and Ramakanth Kavuluru. 2018. Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations. *Journal of Biomedical Informatics*, 82:189–199.

Amitav Banerjee, Udaykumar Bhaskar Chitnis, Sl Jadhav, Js Bhawalkar, and Suprakash Chaudhury. 2009. Hypothesis testing, type i and type ii errors. *Industrial Psychiatry Journal*, 18(2):127–131.

Parit Bansal, Anne Morgat, Kristian B. Axelsen, Venkatesh Muthukrishnan, Elisabeth Coudert, Lucila Aimo, Nevila Hyka-Nouspikel, Elisabeth Gasteiger, Arnaud Kerhornou, Teresa Batista Neto, Monica Pozzato, Marie-Claude Blatter, Alex Ignatchenko, Nicole Redaschi, and Alan Bridge. 2022. Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Research*, 50(D1):D693–D700.

Agustín Borrego, Danilo Dessì, Daniel Ayala, Inma Hernández, Francesco Osborne, Diego Reforgiato Recupero, Davide Buscaldi, David Ruiz, and Enrico Motta. 2025. Research hypothesis generation over scientific knowledge graphs. *Knowledge-Based Systems*, 315:113280.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1877–1901, Red Hook, NY, USA. Curran Associates, Inc.

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10:67.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts,

Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sashank Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Armen Yuri Gasparyan, Lilit Ayvazyan, Ulzhan Mukanova, Marlen Yessirkepov, and George Kitas. 2019. Scientific hypotheses: Writing, promoting, and predicting implications. *Journal of Korean Medical Science*, 34(45):e300.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, and 23 others. 2024. The llama 3 herd of models. *arXiv preprint*.

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

Minoru Kanehisa, Miho Furumichi, Yoko Sato, Masayuki Kawashima, and Mari Ishiguro-Watanabe. 2023. Kegg for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*, 51(D1):D587–D592.

Minoru Kanehisa and Susumu Goto. 2000. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30.

Shrinidhi Kumbhar, Venkatesh Mishra, Kevin Coutinho, Divij Handa, Ashif Iquebal, and Chitta Baral. 2025. Hypothesis generation for materials discovery and design using goal-driven and constraint-guided llm agents. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7539–7570, Albuquerque, New Mexico. Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Haokun Liu, Yangqiaoyu Zhou, Mingxuan Li, Chenfei Yuan, and Chenhao Tan. 2025. Literature meets data: A synergistic approach to hypothesis generation. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.

Durga Prasanna Misra, Armen Yuri Gasparyan, Olena Zimba, Marlen Yessirkepov, Vikas Agarwal, and George Kitas. 2021. Formulating hypotheses for different study designs. *Journal of Korean Medical Science*, 36(50):e338.

John H. Morris, Karthik Soman, Rabia E. Akbas, Xiaoyuan Zhou, Brett Smith, Elaine C. Meng, Conrad C. Huang, Gabriel Cerono, Gundolf Schenk, Angela Rizk-Jackson, Adil Harroud, Lauren Sanders, Sylvain V. Costes, Krish Bharat, Arjun Chakraborty, Alexander R. Pico, Taline Mardirossian, Michael Keiser, Alice Tang, and 8 others. 2023. The scalable precision medicine open knowledge engine (spoke): A massive knowledge graph of biomedical information. *Bioinformatics*, 39(2):btad080.

Karl Raimund Popper. 1959. *The Logic of Scientific Discovery*. Hutchinson & Co.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2025. Medgemma technical report. *arXiv preprint*.

Yucheng Shi, Shaochen Xu, Tianze Yang, Zhengliang Liu, Tianming Liu, Xiang Li, and Ninghao Liu. 2025. Mkrag: Medical knowledge retrieval augmented generation for medical question answering. In *AMIA Annual Symposium Proceedings*, pages 1011–1020.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2025. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *Proceedings of the 13th International Conference on Learning Representations*.

Karthik Soman, Peter W. Rose, John H. Morris, Rabia E. Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, Gabriel Cerono, Yongmei Shi, Angela Rizk-Jackson, Sharat Israni, Charlotte A. Nelson, Sui Huang, and Sergio E. Baranzini. 2024. Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics*, 40(9):btae560.

Song Tong, Kai Mao, Zhen Huang, Yukun Zhao, and Kaiping Peng. 2024. Automating psychological hypothesis generation with ai: When large language models meet causal graph. *Humanities and Social Sciences Communications*, 11:896.

UniProt Consortium. 2023. Uniprot: The universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531.

Yilin Wen, Zifeng Wang, and Jimeng Sun. 2024. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 10370–10388, Bangkok, Thailand. Association for Computational Linguistics.

E. Wood, Amy Glen, Lindsey Kvarfordt, Finn Womack, Liliana Acevedo, Timothy Yoon, Chunyu Ma, Veronica Flores, Meghamala Sinha, Jared Roach, Luis Mendoza, Andrew Hoffman, Eric Deutsch, David Koslicki, and Stephen Ramsey. 2022. Rtx-kg2: A system for building a semantically standardized knowledge graph for translational biomedicine. *BMC Bioinformatics*.

Guangzhi Xiong, Eric Xie, Amir Shariatmadari, Sikun Guo, Stefan Bekiranov, and Aidong Zhang. 2024. Improving scientific hypothesis generation with knowledge grounded large language models. *arXiv preprint*.

Guangzhi Xiong, Eric Xie, Corey Williams, Myles Kim, Amir Hassan Shariatmadari, Sikun Guo, Stefan Bekiranov, and Aidong Zhang. 2025. Toward reliable scientific hypothesis generation: Evaluating truthfulness and hallucination in large language models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 7849–7857.

Qidi Xu, Claudio Soto, Mohammad Shahnawaz, Xiaozhong Liu, Xiaoqian Jiang, and Yejin Kim. 2025. Multi-agent large language models for biomedical hypothesis generation in drug combination discovery. *iScience*, 28.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *arXiv preprint*.

Fuyi Yang, Chenchen Ye, Mingyu Derek Ma, Yijia Xiao, Matthew Yang, and Wei Wang. 2025b. Bioverge: A comprehensive benchmark and study of self-evaluating agents for biomedical hypothesis generation. *LAW 2025: Bridging Language, Agent, and World Models for Reasoning and Planning*. NeurIPS 2025 Workshop.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024. Large language models for automated open-domain scientific hypotheses discovery. In *Findings of the Association for Computational Linguistics (ACL)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. Hypothesis generation with large language models. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 117–139, Miami, FL, USA. Association for Computational Linguistics.

A Prompt Templates

A.1 Generation Prompts

Condition 1 (C1): Source-Only Baseline

System Prompt

You are an expert biochemist specializing in metabolic disease mechanisms.

User Prompt

Source enzyme: {source_entity}
Target domain: disease mechanism

Generate a testable scientific hypothesis connecting this enzyme to a disease mechanism.

Your hypothesis must:

1. Name the specific enzyme; do not generalize to enzyme families.
2. Propose a specific mechanistic connection to a disease.
3. Make a specific, falsifiable prediction: what experimental result would confirm it, and what would refute it.
4. Be grounded in known biochemistry.

Begin with "We hypothesize that" and limit to 150 words.

Condition 2 (C2): Full-Path Compression

System Prompt

You are an expert biochemist.

User Prompt

The following knowledge graph path describes a multi-step mechanistic relationship between biological entities:

{path_text}

Not every edge in this path represents an equally strong biological claim. Treat the path as partial mechanistic evidence of varying reliability.

Your task is mechanistic compression. The path is not a script to retell. Instead, treat it as mechanistic evidence from which a biologically meaningful hypothesis may emerge.

Instructions:

1. Identify the single most biologically specific and meaningful mechanistic transition in the path.
2. Explicitly state why this transition is more biologically meaningful than the adjacent edges you chose to ignore.
3. Prioritize mechanistic coherence over graph completeness. Ignore weak, generic, or incidental edges.
4. Do not narrate the full path step-by-step.
5. Synthesize one focused scientific hypothesis around that specific mechanistic substructure.
6. State one falsifiable experimental prediction with a specific measurable output, such as enzyme activity, metabolite concentration, assay readout, or cellular phenotype.
7. Name the specific entities involved; do not generalize to enzyme families.
8. The hypothesis should sound like a concise scientific claim, not a graph traversal.

Begin with "We hypothesize that" and limit to 150 words.

Condition 3 (C3): Full-Path Interpretation

System Prompt

You are an expert biochemist.

User Prompt

The following knowledge graph path describes a possible mechanistic relationship between biological entities:

{path_text}

Not every edge in this path represents an equally strong biological claim. Treat the path as partial mechanistic evidence of varying reliability.

Your task is mechanistic interpretation. Reason about the path before generating the final hypothesis:

- Which transitions appear biologically meaningful?
- Which steps are weak, generic, or uncertain?
- Where does the path suggest a true mechanistic bottleneck, regulatory switch, or cross-domain interaction?
- Which localized mechanistic theme in the path appears most biologically meaningful, even if the full path may not represent a direct causal chain?

Do not assume every edge represents a direct causal relationship. Do not feel pressure to connect the source entity to the terminal disease if a stronger, more localized biological claim exists within the path. Use the graph as mechanistic guidance, not as a deterministic script.

Then generate:

1. A short mechanistic rationale identifying the most biologically meaningful substructure in the path and why you selected it over other edges.
2. One coherent, falsifiable scientific hypothesis grounded in that strongest mechanistic theme.
3. Include a specific measurable prediction inside the hypothesis.

Format your response as:

RATIONALE: [brief mechanistic interpretation, max 80 words]
HYPOTHESIS: We hypothesize that [concise mechanistic hypothesis including the falsifiable prediction, max 150 words]

Do not add any additional sections beyond RATIONALE and HYPOTHESIS.

Condition 4 (C4): Endpoint-Only Ablation

System Prompt

You are an expert biochemist specializing in metabolic disease mechanisms.

User Prompt

Source enzyme: {source_symbol}
Target disease: {terminal_name}

Your task is to generate a testable scientific hypothesis proposing a specific mechanistic connection between this source enzyme and this target disease.

Your hypothesis must:

1. Name the specific source enzyme; do not generalize to enzyme families.
2. Propose a specific mechanistic pathway or biological process connecting this enzyme to the target disease.
3. Make a specific, falsifiable prediction: what experimental result would confirm it, and what would refute it.
4. Be grounded in known biochemistry.
5. Connect specifically to {terminal_name}; do not connect to a different disease.

Begin with "We hypothesize that" and limit to 150 words.

A.2 Judge System Prompt

The following system prompt was used for GPT-4o (primary judge for all models except GPT-4o) and Gemini 2.5 Flash (secondary judge for all models). Qwen3-235B used the same prompt as primary judge for GPT-4o outputs. No model evaluated its own outputs.

System Prompt

You are an expert scientific evaluator assessing the quality of AI-generated biochemical hypotheses.

User Prompt

You will be given:

1. A source enzyme/gene
2. A target disease
3. A knowledge graph path connecting the source to the disease
4. A generated hypothesis
5. A context line stating what evidence was available to the generator

Your task is to score the hypothesis on five criteria using a 1–5 scale. Return ONLY a valid JSON object with your scores and reasoning. Nothing else.

SCORING SCALE:
1 = Poor | 2 = Weak | 3 = Acceptable / mixed | 4 = Strong | 5 = Excellent

CRITERION 1: Path/Task Relevance (1–5)
Does the hypothesis address the assigned source-to-terminal biological problem?

- Score 5: Directly addresses source gene and terminal disease through a coherent biological mechanism
- Score 4: Addresses source and terminal with minor gaps
- Score 3: Partially addresses the path, mechanism is vague or focuses only on intermediate nodes
- Score 2: Weakly connected, mentions the disease but mechanism does not involve the source gene meaningfully
- Score 1: Addresses a completely different disease or biological question

CRITERION 2: Mechanistic Specificity and Consistency (1–5)
Does the hypothesis propose a specific, biologically coherent mechanism?

- Score 5: Names specific enzyme, substrate, product, cell type, and mechanism
- Score 4: Specific mechanism with minor biochemical gaps
- Score 3: Some specificity but relies on pathway-level language or has one inaccuracy
- Score 2: Vague mechanism, names entities but does not explain how they connect
- Score 1: Biochemically incoherent or uses only generic language

NOTE: Do not reward biochemical-sounding mechanisms if the intermediate reactions are vague or unsupported. The mechanistic chain must be biochemically justified

step by step.

CRITERION 3: Experimental Testability and Operationalization (1–5)
Can the hypothesis be tested with a concrete perturbation, model system, assay, and measurable outcome?

- Score 5: Names perturbation, model system, measurable outcome, and falsification, quantitative claims are justified or appropriately hedged
- Score 4: Clear experimental design with one missing element
- Score 3: Testable in principle but missing key details
- Score 2: Vague prediction, "may affect" without specifics
- Score 1: No testable prediction or purely observational

CRITERION 4: Non-Trivial Novelty / Knowledge Advancement (1–5)
Does the hypothesis generate a new scientific question beyond restating textbook knowledge?

- Score 5: Genuinely new mechanistic connection not established in literature
- Score 4: Extends known biology in a non-obvious direction
- Score 3: Some novel element but mostly builds on well-known mechanisms
- Score 2: Mostly restates established biology
- Score 1: Directly restates a textbook mechanism

CRITERION 5: Evidence-Proportionality and Full-Path Use (1–5)
Does the strength of the claim match the evidence provided to the generator?

CRITICAL RULE:

- If the generator was shown the full KG path: score based on how intelligently it used the path evidence.
- If the generator was NOT shown the KG path: a high score requires that the proposed mechanistic bridge is either (a) explicitly framed as speculative and biologically plausible, or (b) directly supported by well-established published biochemistry. A hypothesis that invents a plausible-sounding but unsupported bridge, without acknowledging uncertainty, should score at most 3 regardless of fluency.
- Score 5: Uses both strong and upstream path elements intelligently, OR if no path shown, bridge is grounded in established biochemistry and explicitly proportionate to evidence
- Score 4: Uses strong path elements well, OR if no path shown, mechanism is plausible but missing explicit uncertainty framing
- Score 3: Uses only the most obvious path element, OR if no path shown, reasonable but unsupported bridge stated with overconfidence
- Score 2: Overclaims beyond evidence, or largely ignores the path shown
- Score 1: Completely ignores the path shown, OR if no path shown, proposes a biochemically implausible mechanism

Return ONLY this JSON structure:

```
{
  "path_task_relevance": <1-5>,
  "mechanistic_specificity": <1-5>,
  "experimental_testability": <1-5>,
  "nontrivial_novelty": <1-5>,
  "evidence_proportionality": <1-5>,
  "total_score": <sum 5-25>,
  "reasoning": "<2-3 sentences
    explaining the scores>"
}
```

Judge input format. Each hypothesis was presented to the judge with the following user prompt:

User Prompt Template
Source enzyme/gene: {source_symbol}
Target disease: {terminal_name}
{context_line}

Knowledge graph path:
{path_text}

Generated hypothesis:
{hypothesis_text}

Score this hypothesis on the five criteria. Return only the JSON object.

Where {context_line} was one of:

- **C1:** "Evidence available to generator: source enzyme name only. The KG path and terminal disease were NOT shown."
- **C2/C3:** "Evidence available to generator: full knowledge graph path shown below."
- **C4:** "Evidence available to generator: source enzyme name and target disease name only. The KG path was NOT shown."

B Expert-Derived Evaluation Rubric

The rubric was derived from structured evaluation sessions with eight PhD-level domain experts in biochemistry and metabolic disease. Experts independently scored hypotheses for a held-out calibration path (CDO1 → Primary Hyperoxaluria Type 1, HKG_0215) under all four conditions and provided written justifications. The five criteria below reflect the qualitative dimensions that consistently drove agreement and disagreement across experts.

Total Score Interpretation
Total score range: 5–25 points.

- **20–25:** Strong hypothesis
- **15–19:** Usable but needs refinement
- **10–14:** Weak or partial
- **Below 10:** Poor

Criterion 1: Path/Task Relevance

Definition
Does the hypothesis address the correct source-to-terminal biological problem? Does it connect to the assigned terminal disease through the given path rather than drifting to a different disease or biological question?

Literature Grounding
Novel to KG evaluation, this work.

Expert Signal

All six experts who saw the C1 disease-drift example flagged the hypothesis as well-formed but path-disconnected, motivating a dedicated relevance criterion.

Behavioral Anchors

Score	Anchor
5	Directly addresses source gene and terminal disease through a coherent biological mechanism
4	Addresses source and terminal with minor gaps in the connection
3	Partially addresses the path, connects to terminal but mechanism is vague, or focuses only on intermediate nodes
2	Weakly connected, mentions the disease but mechanism does not involve the source gene meaningfully
1	Addresses a completely different disease or biological question

Criterion 2: Mechanistic Specificity and Consistency

Definition

Does the hypothesis propose a specific, biologically coherent mechanism involving named enzymes, metabolites, reactions, cell types, or pathway logic? Is the mechanism consistent with known biochemistry?

Literature Grounding

Gasparyan et al. (2019); Misra et al. (2021).

Expert Signal

Experts praised hypotheses that correctly identified AGXT $\rightarrow$ glyoxylate $\rightarrow$ oxalate as the mechanistic anchor, and penalized those using pathway-level language without reaction logic.

Behavioral Anchors

Score	Anchor
5	Names specific enzyme, substrate, product, cell type, and mechanism, all biochemically accurate and supported
4	Specific mechanism with minor biochemical gaps or simplifications
3	Some specificity but relies on pathway-level language, or uses biochemical terminology without explaining actual reaction steps
2	Vague mechanism, names entities but does not explain how they connect
1	Biochemically incoherent, contradicts known biochemistry, or uses only generic language

Criterion 3: Experimental Testability and Operationalization

Definition

Can the hypothesis be tested with a concrete perturbation, model system, assay, and measurable outcome? Is there an explicit falsification condition?

Literature Grounding

Popper (1959); Banerjee et al. (2009).

Expert Signal

Experts preferred hypotheses proposing AGXT knockdown, glyoxylate challenge, and dose-response testing. Multiple experts criticized unjustified numerical precision, such as exact fold-change thresholds stated as fact.

Behavioral Anchors

Score	Anchor
5	Names perturbation, model system, measurable outcome, and falsification, quantitative claims justified or appropriately hedged
4	Clear experimental design with one missing element, or includes numbers framed as approximate predictions
3	Testable in principle but missing key details, or includes unjustified precise numbers stated as fact
2	Vague prediction, "may affect" or "could influence" without specifics
1	No testable prediction, untestable, or purely observational without experiment

Criterion 4: Non-Trivial Novelty and Knowledge Advancement

Definition

Does the hypothesis generate a new scientific question or prediction beyond restating textbook knowledge? Does it propose a new relationship, mechanism, modifier, or experimental angle not already established?

Literature Grounding

Misra et al. (2021) knowledge-gap criterion.

Expert Signal

Five of six experts independently called one calibration hypothesis "correct but trivial," "textbook," or "not a research hypothesis," motivating a dedicated novelty criterion with a low floor for restated biology.

Behavioral Anchors

Score	Anchor
5	Proposes a genuinely new mechanistic connection or modifier not established in existing literature
4	Extends known biology in a non-obvious direction, one credible step beyond established knowledge
3	Has some novel element but mostly builds on well-known mechanisms
2	Mostly restates established biology, the core claim is already known
1	Directly restates a textbook disease mechanism with no new prediction

Criterion 5: Evidence Proportionality and Full-Path Use

Definition

Does the hypothesis use the KG path intelligently, preserving strong edges, treating weak edges cautiously, and attempting to extract value from upstream elements? Does the strength of the claim match the evidence provided to the generator?

Literature Grounding

Novel to KG evaluation, this work.

Expert Signal

Expert 2 explicitly named "proportionality of evidence to claim" and "engagement with the full path" as evaluation dimensions. Experts noted that strong hypotheses correctly identified CDO1 as a weak link but asked whether it could modify the AGXT mechanism, rather than simply abandoning it.

Behavioral Anchors

Score	Anchor
5	Uses both strong and upstream path elements intelligently, or if no path is shown, the proposed bridge is grounded in established published biochemistry and explicitly proportionate to evidence
4	Uses strong path elements well, or if no path is shown, the mechanism is plausible and partially grounded but lacks explicit uncertainty framing
3	Uses only the most obvious path element, or if no path is shown, the mechanism is a reasonable but unsupported bridge stated with overconfidence
2	Overclaims beyond what the evidence supports, or largely ignores the path when one was provided
1	Completely ignores the path when one was shown, or if no path is shown, proposes a biochemically implausible or hallucinated mechanism

Worked Calibration Examples

All five examples use the calibration path: CDO1 → Cyanoamino acid metabolism → AGXT → Primary Hyperoxaluria Type 1.

Example 1: Path-irrelevant but fluent

Context

C1: source enzyme only shown. Terminal disease and path withheld.

Hypothesis

We hypothesize that glutathione reductase deficiency in retinal pigment epithelial cells contributes to age-related macular degeneration through impaired glutathione regeneration, leading to lipofuscin accumulation and complement activation. Reduced GSR activity causes elevated GSSG and decreased GSH. This promotes A2E formation and inflammasome activation. RPE cells with siRNA-mediated GSR knockdown will show increased GSSG/GSH ratios, elevated A2E fluorescence, and increased IL-1β secretion, falsified if knockdown shows no lipofuscin increase.

Scores: Rel=1, Mech=2, Test=4, Nov=3, Prop=1, Total=11. Connects GSR to AMD, not Aminoacylase-1 deficiency. Disease-drift bias example.

Example 2: Correct but textbook

Context

C2/C3: full KG path shown.

Hypothesis

We hypothesize that loss-of-function mutations in AGXT disrupt glyoxylate metabolism in hepatocytes, leading to pathological accumulation of oxalate and causing Primary Hyperoxaluria Type 1. AGXT directly catalyzes transamination of glyoxylate to glycine, preventing oxalate overproduction. The upstream CDO1 link is metabolically tangential. Patients with biallelic AGXT mutations will exhibit reduced hepatic AGXT enzyme activity (less than 15% of wild-type) and elevated urinary oxalate (greater than 0.7 mmol/day), measurable via tandem mass spectrometry.

Scores: Rel=4, Mech=4, Test=2, Nov=1, Prop=3, Total=14. Correct terminal and mechanism but restates textbook AGXT–PH1 biology. Uses strongest path edge but discards CDO1.

Example 3: Strong but imperfect

Context

C2/C3: full KG path shown.

Hypothesis

RATIONALE: The strongest mechanistic connection is between AGXT enzyme dysfunction and Primary Hyperoxaluria Type 1. AGXT catalyzes alanine-glyoxylate aminotransferase activity and its deficiency directly causes oxalate accumulation. The cyanoamino acid connection appears weaker.

HYPOTHESIS: We hypothesize that reduced AGXT enzyme activity creates a metabolic bottleneck in glyoxylate detoxification, leading to increased glyoxylate-to-oxalate conversion. Cells with AGXT knockdown will show a dose-dependent increase in oxalate production with levels increasing 2 to 5 fold at 50% AGXT activity reduction, measurable by oxalate secretion assays.

Scores: Rel=4, Mech=4, Test=4, Nov=3, Prop=3, Total=18. Correct target, testable with dose-response, but cell type unspecified and CDO1 upstream abandoned entirely.

Example 4: Endpoint-only, plausible but overconfident

Context

C4: source and terminal disease only. Path not shown.

Hypothesis

We hypothesize that CDO1 deficiency leads to cysteine accumulation, which is then shunted through alternative metabolic pathways to produce glyoxylate, thereby exacerbating oxalate overproduction in Primary Hyperoxaluria Type 1 patients who already carry AGXT mutations. CDO1 knockdown in AGXT-deficient hepatocytes will show higher urinary oxalate than AGXT knockdown alone.

Scores: Rel=4, Mech=3, Test=3, Nov=3, Prop=2, Total=15. Correctly connects CDO1 to PH1. Mechanism plausible but stated with overconfidence; no path was shown, so the bridge should be framed as speculative.

Example 5: Excellent

Context

C2/C3: full KG path shown.

Hypothesis

We hypothesize that CDO1 activity may modulate the severity of AGXT-deficient Primary Hyperoxaluria Type 1 by influencing cysteine-derived metabolite competition for shared peroxisomal processing capacity and pyridoxal phosphate cofactor availability. When CDO1 is impaired alongside AGXT deficiency, accumulated cysteine metabolites may reduce effective PLP availability or alter peroxisomal metabolic load, potentially amplifying oxalate overproduction beyond what AGXT deficiency alone would produce. We predict that CDO1 knockdown in AGXT-deficient human hepatocytes will produce greater oxalate accumulation than AGXT knockdown alone, measurable by LC-MS oxalate quantification, and that pyridoxine supplementation will partially rescue this amplified phenotype, supporting a PLP-sensitive modifier mechanism.

Note: This example is used to illustrate a high-scoring rubric response, not to claim experimental validation of the proposed mechanism.

Scores: Rel=5, Mech=5, Test=5, Nov=5, Prop=5, Total=25. Connects the source and terminal disease through a novel CDO1-as-modifier mechanism, uses both CDO1 and AGXT intelligently, and remains appropriately hedged. The hypothesis is biochemically specific and experimentally actionable, with a concrete perturbation, LC-MS oxalate readout, and pyridoxine rescue test.

C Additional Results

Table 7: Per-model evidence proportionality drop under path shuffling.

To assess whether individual models genuinely utilized intermediate KG path structure, we report evidence proportionality scores under original and shuffled paths for each model separately. Shuffling permutes intermediate nodes while holding source and terminal endpoints fixed, breaking the biological causal chain while preserving endpoint identity. A significant drop indicates the model was responding to ordered intermediate structure rather than endpoint priors alone.

The model-dependent pattern directly confirms the capability split reported in Section 5.2. Stronger models, Qwen3-235B, Claude Sonnet, MedGemma-27B, and GPT-4o, show large, significant drops under both C2 and C3 (Δ range: 0.68–1.11), indicating genuine utilization of intermediate path structure. Llama-3.3-70B also shows significant drops despite its weak overall grounding gains, suggesting it extracts some path signal even when it cannot translate it into high-quality hypotheses. BioMistral-7B is the clearest outlier: its C3 drop is near-zero and non-significant ($\Delta = -0.023$, $p = 0.242$), mechanistically confirming that this model was not using intermediate path structure even in the original condition.

Table 8: Secondary judge agreement with primary judge.

Inter-judge agreement was computed on per-hypothesis scores for each criterion separately. This breakdown allows assessment of which rubric dimensions are reliably scored by LLM judges and which require greater caution in interpretation.

Agreement is highest for experimental testability ($r = 0.839$) and path/task relevance ($r = 0.813$), both criteria with relatively objective anchors. Agreement is lowest for evidence proportionality ($r = 0.668$), reflecting the greater subjec-

Model	C2 Orig.	C2 Shuf.	C2 Drop	Sig.	C3 Orig.	C3 Shuf.	C3 Drop	Sig.
Qwen3-235B	3.396	2.286	−1.110	***	3.021	2.222	−0.799	***
Claude Sonnet	3.073	2.131	−0.942	***	2.944	2.220	−0.724	***
MedGemma-27B	3.027	2.166	−0.861	***	3.268	2.386	−0.882	***
Llama-3.3-70B	2.593	1.736	−0.857	***	2.251	1.720	−0.531	***
GPT-4o	2.750	2.053	−0.697	***	2.990	2.311	−0.679	***
BioMistral-7B	1.938	1.741	−0.197	***	2.139	2.116	−0.023	ns

Table 7: Per-model evidence proportionality drop under path shuffling. Wilcoxon signed-rank test, $n = 518$ per model per condition. *** $p < 0.001$; ns = not significant. Models are sorted by C2 drop magnitude. Bold values indicate the largest drop within each condition.

Criterion	Pearson r	Mean $\lvert\Delta\rvert$
Experimental Testability	0.839	1.72
Path/Task Relevance	0.813	1.87
Mechanistic Specificity	0.791	1.94
Non-Trivial Novelty	0.774	2.03
Total Score	0.776	2.10
Evidence Proportionality	0.668	2.34

Table 8: Secondary judge agreement with the primary judge across 13,200 hypotheses. Mean $\lvert\Delta\rvert$ is computed on the 1–5 criterion scale, except Total Score, which uses the 5–25 scale.

tivity of assessing whether claim strength is calibrated to the specific evidence provided to the generator.

Table 9: KG consistency metrics by condition.

These automated metrics measure entity mention behavior, whether hypotheses reference the source gene, terminal disease, and intermediate path entities, rather than scientific quality. They serve as complementary grounding diagnostics confirming that conditions differed in the information models used, not just in prompt wording.

Condition	Source	Terminal	Path	Comp.
C1 Source-only	0.821	0.043	n/a	0.821
C2 Path-compress	0.894	0.876	0.312	0.694
C3 Path-interpret	0.887	0.869	**0.341**	0.699
C4 Endpoint-only	**0.912**	**0.961**	n/a	**0.937**

Table 9: KG consistency metrics by condition. Source and terminal grounding are binary mention rates. Path entity coverage is the fraction of named intermediate KG entities mentioned. Composite is the mean of applicable metrics.

Three patterns are worth noting. First, C1's near-zero terminal grounding directly quantifies disease drift without a disease anchor. Second, C4's high terminal grounding confirms that endpoint-only models reliably mention the disease name, but entity mention does not imply mechanistic path use. Third, path entity coverage is highest under C3, consistent with the interpretation instruction encouraging explicit engagement with intermediate path entities.