BIOME-Bench: A Benchmark for Biomolecular Interaction Inference and Multi-Omics Pathway Mechanism Elucidation from Scientific Literature

Anonymous ACL submission

Abstract

Multi-omics studies often rely on pathway enrichment to interpret heterogeneous molecular changes, but pathway enrichment (PE)-based workflows inherit structural limitations of pathway resources, including curation lag, functional redundancy, and limited sensitivity to molecular states and interventions. Although recent work has explored using large language models (LLMs) to improve PE-based interpretation, the lack of a standardized benchmark for end-to-end multi-omics pathway mechanism elucidation has largely confined evaluation to small, manually curated datasets or ad hoc case studies, hindering reproducible progress. To address this issue, we introduce **BIOME-Bench**, constructed via a rigorous four-stage workflow, to evaluate two core capabilities of LLMs in multi-omics analysis: **B**iomolecular **I**nteraction Inference and end-to-end Multi-**O**mics Pathway **M**echanism **E**lucidation. We develop evaluation protocols for both tasks and conduct comprehensive experiments across multiple strong contemporary models. Experimental results demonstrate that existing models still exhibit substantial deficiencies in multi-omics analysis, struggling to reliably distinguish fine-grained biomolecular relation types and to generate faithful, robust pathway-level mechanistic explanations. Code and datasets are available on GitHub.[1]

1 Introduction

Multi-omics profiling, including proteomics, metabolomics, and single-cell transcriptomics, among others, has become central to studying complex biological systems and disease mechanisms (Sanches et al., 2024; Baião et al., 2025). By capturing coordinated molecular changes across regulatory layers, multi-omics can support mechanistic hypotheses that connect perturbations to pathway-level phenotypes. However, translating heterogeneous multi-omics signals into coherent, causally grounded explanations remains a major bottleneck (Mohr et al., 2024).

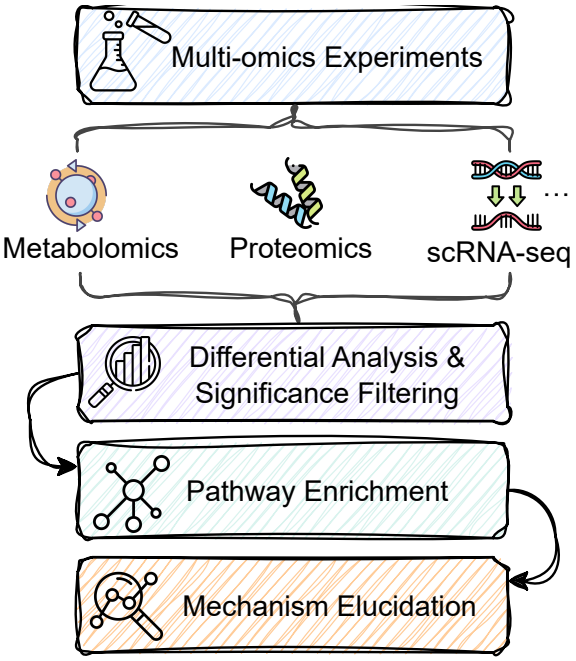

Figure 1: Overview of a pathway enrichment-based multi-omics mechanism elucidation workflow. Multi-omics experiments (e.g., metabolomics, proteomics, and scRNA-seq) are followed by differential analysis with significance filtering to identify perturbed entities, which are then mapped to biological pathways via enrichment analysis to support downstream mechanistic interpretation.

A common approach is pathway enrichment (PE) interpretation (Zhao and Rhee, 2023; Ryan V et al., 2025; Mubeen et al., 2022). As shown in Figure 1, practitioners typically perform differential analysis and significance filtering to identify perturbed entities, map them to pathways through enrichment, and interpret the enriched pathways to derive mechanistic narratives. PE therefore serves as a widely used interface between entity-level changes and higher-level biological processes (Elizarraras et al., 2024).

[1] https://anonymous.4open.science/r/BIOME-Bench

However, many limitations of PE-based interpretation arise from the structure of pathway resources. Pathway knowledge bases are curated and versioned, which introduces curation lag and can omit newly discovered or context-specific mechanisms (Agrawal et al., 2024). Enrichment results also exhibit substantial functional redundancy due to overlapping gene sets, often yielding long lists of near-duplicate terms that are difficult to prioritize (Balestra et al., 2023; Ge et al., 2025; Ozisik et al., 2022). More fundamentally, enrichment scores are largely context-insensitive. They do not represent molecular states (e.g., phosphorylation or activation), intervention directionality, or the multi-hop causal structure required to connect perturbed entities to pathway-level phenotypes. Consequently, downstream interpretation often relies on post hoc narrative stitching rather than state-aware mechanistic reasoning (Slobodyanyuk et al., 2024). In practice, researchers may obtain an interpretable list of pathway names, but often still need additional analysis to determine how the observed perturbations give rise to the phenotype.

Recent advances in large language models (LLMs) provide opportunities to improve PE-based interpretation, for example by reducing redundancy in enrichment outputs and using LLMs to generate mechanistic explanations (Ge et al., 2025; Zhou et al., 2024). However, progress in this direction is difficult to measure reliably because existing biomedical benchmarks rarely evaluate the end-to-end capability required to produce mechanistic interpretations from multi-omics observations. Many literature-grounded datasets focus on question answering (Jin et al., 2019) or local relation extraction (Zhang et al., 2019; Luo et al., 2022), where relations are often assessed in isolation from pathway context and phenotype-level consequences. Meanwhile, existing pathway-reasoning benchmarks typically emphasize navigation or subgraph operations rather than generating state-aware mechanistic chains that connect perturbed entities to pathway-level phenotypes (Zhao et al., 2025). In the absence of such end-to-end benchmarks, evaluations often fall back on small manually curated datasets or ad hoc case studies (Ge et al., 2025). This limits coverage and reproducibility, and it leaves open how well current models can directly generate coherent pathway-mechanism explanations on perturbed entities and pathway context.

Motivated by these limitations, we consider a complementary formulation that more closely aligns with how scientists reason from observations to hypotheses: *end-to-end multi-omics pathway mechanism elucidation*. Under this formulation, a model receives only (i) a set of significantly perturbed entities derived from multi-omics measurements and (ii) a pathway context describing the relevant biological process. The model must then generate a coherent mechanistic explanation without relying on explicit pathway retrieval and graph traversal. This end-to-end perspective can reduce redundancy by shifting the unit of interpretation from overlapping pathway labels to mechanistic hypotheses. It also better reflects research practice, where the goal is to elucidate mechanisms from perturbed entities conditioned on pathway context. The task is nontrivial because it requires rich biological knowledge and multi-step causal reasoning to connect perturbed entities to pathway-level phenotypes (see Appendix B), which motivates standardized, instance-level supervision for reliable evaluation.

To address this gap, we introduce BIOME-Bench, a literature-grounded benchmark designed to evaluate LLMs on two core capabilities: biomolecular interaction inference and end-to-end multi-omics pathway mechanism elucidation. To construct high-quality evaluation instances, we develop a dedicated data construction workflow. As shown in Figure 2, the workflow transforms pathway information and evidence from biomedical literature into structured, validated knowledge representations through three sequential phases: (i) Literature Search and Relevance Filtering, (ii) Information Extraction and Standardization, and (iii) Knowledge Structuring and Validation. The resulting knowledge representations are then used to formulate the benchmark tasks. Our contributions are as follows:

- We formulate end-to-end multi-omics pathway mechanism elucidation as a benchmarkable task. Given perturbed entities and pathway context, models must generate coherent, state-aware, and intervention-consistent mechanistic explanations without external retrieval or graph traversal.

- We construct BIOME-Bench, a literature-grounded benchmark with instance-level supervision that captures key mechanistic elements, including entities, molecular states, biomolecular relations, and pathway-level phenotypes,

enabling evaluation beyond surface narrative quality.

- We design evaluation protocols that measure mechanistic correctness at multiple granularities, including structured knowledge graph evaluation and holistic mechanism evaluation, enabling diagnosis of common failure modes in current LLMs.

- We benchmark a diverse set of LLMs and analyze their deficiencies, motivating future research toward reliable end-to-end pathway mechanism elucidation from multi-omics observations.

2 Related Work

2.1 Pathway Knowledge Bases

KEGG, Reactome, and WikiPathways provide curated pathway representations as molecular interaction and reaction networks (Kanehisa et al., 2025; Milacic et al., 2024; Agrawal et al., 2024). KEGG organizes pathway maps using the KO system to support cross-species mapping (Kanehisa et al., 2025). Reactome provides a consistent, manually curated human-centric pathway model with disease and drug context (Milacic et al., 2024). WikiPathways supports community-driven pathway curation (Agrawal et al., 2024). However, these databases are not designed as end-to-end benchmarks for multi-omics mechanism elucidation. They typically lack benchmark-ready, instance-level supervision, such as explicit interventions, molecular states, relations, and phenotype consequences, and they may lag behind newly published mechanistic findings. Accordingly, we start from KEGG pathways and ground benchmark instances in scientific literature to derive state-aware structured supervision and mechanistic explanations. Notably, our method is applicable to other pathway resources.

2.2 Literature-Grounded Benchmarks for Pathway Reasoning and Relation Extraction

Many biomedical benchmarks are derived from scientific literature, but they do not evaluate end-to-end pathway mechanism elucidation from multi-omics observations. PubMedQA focuses on evidence-based question answering over PubMed abstracts (Jin et al., 2019). DrugProt (Luo et al., 2022) and ChemProt (Zhang et al., 2019) evaluate local relation extraction from abstracts or sentences. However, these benchmarks do not assess biomolecular interaction inference under pathway context.

More recently, BioMaze benchmarks intervention-centric pathway reasoning and introduces pathway subgraph navigation agents such as PathSeeker (Zhao et al., 2025). However, it emphasizes navigation and subgraph operations rather than generating state-aware mechanistic chains that connect perturbed entities to pathway-level phenotypes. In contrast, BIOME-Bench provides literature-grounded, instance-level supervision with explicit molecular states, relations, and pathway phenotypes, enabling fine-grained evaluation of both interaction types and end-to-end mechanistic explanations.

2.3 LLM-Based Systems for Multi-Omics and Pathway Interpretation

Recent LLM-based systems such as AutoBA (Zhou et al., 2024) and MAPA (Ge et al., 2025) demonstrate how LLMs can support practical multi-omics workflows, including analysis planning, tool execution, and interpretation of pathway analysis outputs. However, due to the lack of end-to-end benchmarks for multi-omics pathway mechanism elucidation, model performance in this setting is often evaluated through ad hoc case studies or small manually curated datasets. This limits coverage and reproducibility, and it makes it difficult to diagnose whether models can, given only perturbed entities and pathway context, directly generate accurate pathway-mechanism explanations. To address this gap, BIOME-Bench provides literature-grounded, instance-level supervision and standardized evaluation for this end-to-end setting.

3 Methodology

As shown in Figure 2, we propose a data construction workflow that converts pathway information and supporting evidence from biomedical literature into structured, validated knowledge representations in three sequential phases (detailed prompts are provided in Appendix C). Building on these representations, we introduce **BIOME-Bench**, a literature-grounded benchmark for evaluating LLMs on two tasks: Biomolecular Interaction Inference and Multi-Omics Pathway Mechanism Elucidation.

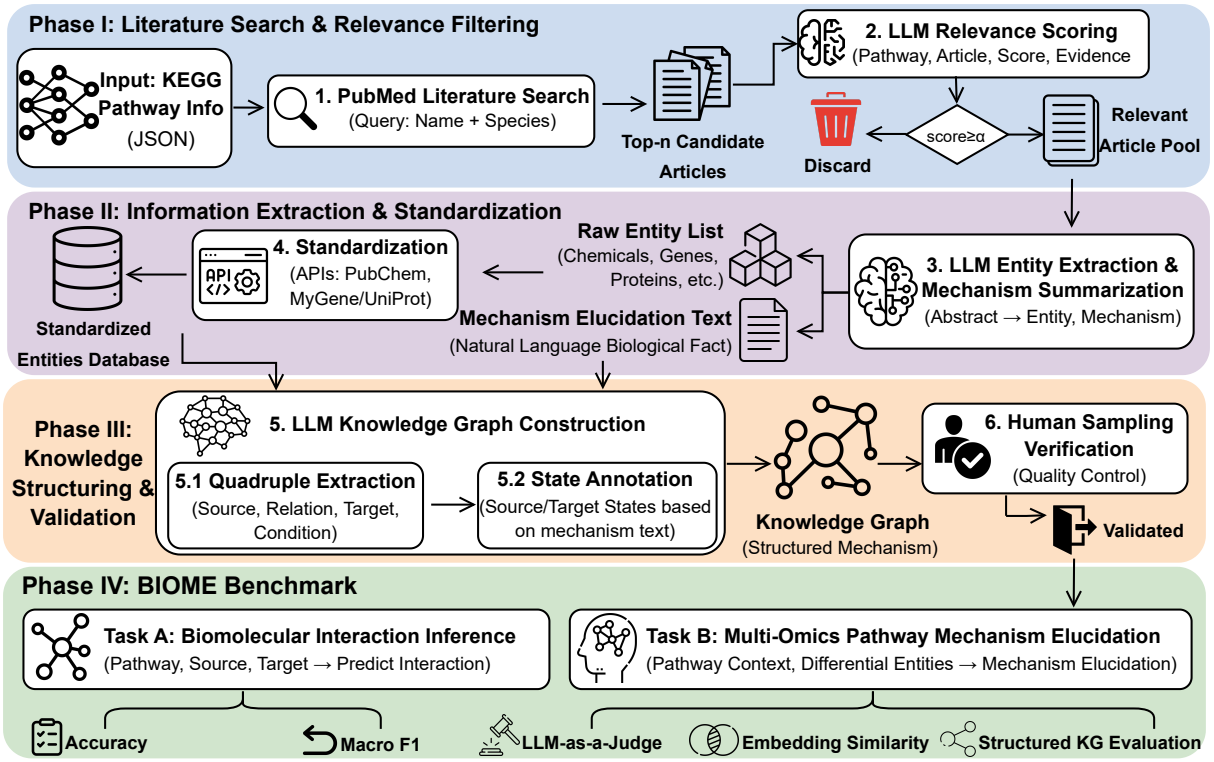

Figure 2: Workflow for constructing BIOME-Bench: (I) MeSH-guided PubMed retrieval with LLM relevance filtering; (II) LLM-based entity extraction and standardization; (III) state-aware knowledge graph construction with human sampling verification; and (IV) benchmark formulation with two tasks—biomolecular interaction inference and multi-omics pathway mechanism elucidation.

3.1 Phase I: Literature Retrieval and Relevance Filtering

To ensure strong biological validity, the construction process begins with a rigorous literature acquisition stage. Let

$$\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$$

denote a predefined set of target KEGG pathways. Each pathway p_i is characterized by its pathway name N_{p_i} and associated species S_{p_i}.

MeSH-guided Literature Retrieval. For each pathway p_i, we perform a structured literature search on the PubMed (White, 2020) database using Medical Subject Headings (MeSH) (National Library of Medicine (US)) to improve recall precision and semantic consistency. Specifically, the pathway name N_{p_i} is mapped to a set of MeSH descriptors, denoted as $\mathrm{MeSH}(N_{p_i})$, and the species S_{p_i} is mapped to the corresponding MeSH organism term, denoted as $\mathrm{MeSH}(S_{p_i})$.

The final PubMed query is constructed as a conjunction of pathway-related MeSH terms and species constraints:

$$Q(p_i) = \mathrm{MeSH}(N_{p_i}) \ \wedge \ \mathrm{MeSH}(S_{p_i}). \quad (1)$$

Executing $Q(p_i)$ yields an initial candidate document set:

$$D_{\mathrm{cand}}(p_i) = \{d_1, d_2, \ldots, d_m\}, \quad (2)$$

where each document d_j is indexed by PubMed and annotated with curated MeSH terms.

LLM-based Semantic and Mechanistic Relevance Scoring. MeSH-guided retrieval yields a controlled, high-recall candidate set, but MeSH annotations alone do not ensure that an article contains pathway-specific mechanistic evidence. To identify literature suitable for mechanism-level benchmarking, we use an LLM-based semantic evaluator with parameters θ.

Given a document and pathway pair (d, p_i), the evaluator assigns a relevance score

$$s = f_\theta(d, p_i), \quad s \in [0, 10],$$

where f_θ aggregates multiple biologically motivated dimensions:

$$f_\theta(d, p_i) = g_\theta(\mathbf{S}), \quad \mathbf{S} = \begin{pmatrix} S_{\mathrm{subj}} \\ S_{\mathrm{spec}} \\ S_{\mathrm{mol}} \\ S_{\mathrm{ctx}} \end{pmatrix}. \quad (3)$$

S_{subj} measures whether the pathway's biological process is the primary focus of the article, as opposed to a background mention. S_{spec} measures whether the organism studied matches the pathway species, and it also credits appropriate model organisms used to study human physiology or disease while penalizing biologically unrelated species. S_{mol} measures whether the article mentions pathway-defined molecular entities, such as genes, enzymes, or metabolites. S_{ctx} measures whether the article describes pathway regulation, such as activation, inhibition, or other modulatory effects, rather than merely reporting pathway presence.

We retain a document only if it exceeds a strict relevance threshold:

$$D_{\text{rel}}(p_i) = \{d \in D_{\text{cand}}(p_i) \mid f_\theta(d, p_i) \geq \alpha\}. \quad (4)$$

In this work, we set $\alpha = 8$ to prioritize articles in which the target pathway is central and supported by explicit molecular and regulatory evidence.

3.2 Phase II: Information Extraction and Entity Standardization

3.2.1 LLM-based Mechanistic Extraction

For each document $d \in D_{\text{rel}}(p_i)$, we process the abstract using a LLM to produce two complementary outputs:

1. **Raw Entity Set** E_{raw}: a collection of mentioned biological entities categorized into Chemicals, Genes/Proteins, and Phenotypes.

2. **Mechanism Description** M_{text}: a coherent natural language explanation describing the molecular interactions and regulatory mechanisms reported in the document. This text later serves as ground truth for generative evaluation.

3.2.2 Entity Normalization and Ontology Mapping

To ensure interoperability with external biological resources, we normalize each raw entity $e \in E_{\text{raw}}$ to a canonical identifier using an ontology resolution function $\phi(e)$. Specifically,

- **Chemical entities** are mapped to PubChem (Kim et al., 2016) compound identifiers (CIDs) using PubChemPy.

- **Genes and proteins** are mapped to NCBI Gene (Brown et al., 2015) identifiers or UniProt (Consortium, 2015) accessions via MyGene.info.

To improve benchmark quality, we discard a candidate document if any entity cannot be resolved to a valid identifier. Only documents for which all entities are successfully normalized are retained, yielding the standardized entity set:

$$E_{\text{std}} = \{\phi(e) | e \in E_{\text{raw}} \wedge \forall e' \in E_{\text{raw}}, \phi(e') \neq \emptyset\} \quad (5)$$

3.3 Phase III: Knowledge Structuring and Validation

This phase leverages an LLM to convert the extracted mechanistic information and standardized entities into a fine-grained, state-aware knowledge graph representation.

3.3.1 Interaction Quadruple Extraction

We first extract the core interaction structure from M_{text}. Each interaction is represented as a quadruple:

$$T_{\text{core}} = (e_s, r, e_t, c), \quad (6)$$

where $e_s, e_t \in E_{\text{std}}$ denote the source and target entities, $r \in \mathcal{R}$ is a relation type drawn from a controlled biological vocabulary, and c specifies the biological condition under which the interaction occurs.

3.3.2 Biological State Annotation

To capture dynamic molecular behavior, we further annotate entity-specific biological states. Let σ_s and σ_t denote the states of the source and target entities, respectively (e.g., mutated, overexpressed). Incorporating state information yields a state-aware hexaplet representation:

$$T_{\text{final}} = (e_s, \sigma_s, r, e_t, \sigma_t, c). \quad (7)$$

This formulation enables the benchmark to distinguish subtle yet critical mechanistic differences, such as changes in protein abundance versus post-translational modifications.

3.3.3 Human Expert Verification

To establish a high-confidence gold standard, we perform human-in-the-loop validation. A randomly sampled subset of the constructed knowledge graph entries is reviewed by domain experts in molecular biology and systems biology, who cross-check each entry against the supporting literature to verify its accuracy and grounding (see Appendix A for details).

3.4 Phase IV: BIOME-Bench Task Formulation

Based on the curated and validated knowledge representations, BIOME-Bench defines two complementary evaluation tasks.

3.4.1 Task A: Biomolecular Interaction Inference

This task evaluates an LLM's ability to infer precise molecular relationships within a pathway context. Given a pathway p_i, a source entity e_s with state σ_s, a target entity e_t with state σ_t, and a biological condition c, the model is required to predict the correct interaction relation from a finite controlled vocabulary $\mathcal{R}$:

$$\hat{r} = \arg\max_{r \in \mathcal{R}} P\big(r \mid p_i, e_s, \sigma_s, e_t, \sigma_t, c\big). \quad (8)$$

Model performance is evaluated using Accuracy and Macro-F1 over relation labels in $\mathcal{R}$ (with invalid or unrecognized predictions treated as errors for the corresponding ground-truth label).

3.4.2 Task B: Multi-Omics Pathway Mechanism Elucidation

This task simulates realistic omics-driven pathway analysis scenarios. The model is provided with a pathway context p_i and a set of differentially observed entities

$$E_{\text{diff}} \subseteq E_{\text{std}},$$

and is required to generate a coherent mechanistic explanation $\hat{Y}$ that elucidates the biological interactions, regulatory relationships, and molecular processes connecting these entities within the given pathway context.

We adopt a multi-dimensional evaluation strategy to comprehensively assess the quality of the generated explanations:

- **LLM-as-a-Judge**: Given the model-generated explanation $\hat{Y}$ and the literature-derived ground truth M_{text}, judge model evaluates the output across four biologically motivated dimensions (scale: 1-5): Phenotype Coverage, Causal Reasoning, Factuality, and Hallucination.

- **Structured Knowledge Graph Evaluation**: Based on the literature-derived knowledge graph, we adopt a closed-set evaluation protocol. Specifically, a extraction model is provided with the standardized knowledge graph and is only allowed to select supporting knowledge tuples from this graph based on the generated explanation $\hat{Y}$. As a result, the predicted tuple set satisfies

$$\mathcal{T}_{\text{pred}} \subseteq \mathcal{T}_{\text{GT}},$$

ensuring that no out-of-graph or hallucinated knowledge can be introduced.

Under this constraint, factual completeness is quantified using coverage, defined as:

$$\text{Coverage} = \frac{|\mathcal{T}_{\text{pred}}|}{|\mathcal{T}_{\text{GT}}|}.$$

- **Semantic Embedding Similarity**: We compute the cosine similarity between vector representations of $\hat{Y}$ and the ground truth mechanism text M_{text} using an LLM-based embedding model, providing a complementary measure of semantic alignment.

4 Experiments

4.1 Benchmark Statistics and Characteristics

Species	Number of Pathways	Number of Entities	Number of Processes and Phenotypes	Task A Biomolecular Interaction Inference	Task B Multi-Omics Pathway Mechanism Elucidation
hsa	80	1,349	1,781	4,032	490
mmu	80	1,356	1,860	4,162	496
rno	80	1,141	1,265	3,384	361
Total	**240**	**3,846**	**4,906**	**11,578**	**1,347**

Table 1: Benchmark statistics of BIOME-Bench across species.

BIOME-Bench is a multi-species benchmark that covers three commonly used organisms: hsa (human), mmu (mouse), and rno (rat). Table 1 summarizes the core statistics, including the numbers of curated pathways, standardized entities, process and phenotype terms, mechanism analysis instances, and knowledge graph relations. Overall, it includes 1,347 instances for multi-omics mechanism elucidation and 11,578 instances for biomolecular interaction inference, both evaluated under consistent pathway contexts.

4.2 Experimental Setup

We evaluate a range of LLMs, including Qwen3-14B, 32B and 235B (Yang et al., 2025), DeepSeek-V3.2-R1 (Liu et al., 2025), GLM-4.6 (GLM et al., 2024), Gemini3-Pro (Google DeepMind), GPT-5.2 (OpenAI), Doubao-Seed-1.8 (Seed et al., 2025),

Model	Biomolecular Interaction Inference		Multi-Omics Pathway Mechanism Elucidation						Avg.
	Acc	Macro-F1	LLM-as-a-Judge				Similarity	Coverage	
			Phenotype Coverage	Causal Reasoning	Factuality	Hallucination			
Qwen3-14B	47.43	43.72	3.12	3.31	3.97	4.64	78.73	42.38	64.13
Qwen3-32B	41.84	40.51	3.00	3.26	3.89	4.79	**78.98**	**45.43**	63.20
Qwen3-235B	51.41	46.21	3.66	4.32	4.54	4.40	77.34	42.22	69.45
DeepSeek-V3.2-R1	53.10	47.52	3.28	4.31	4.20	4.10	75.12	40.76	66.79
GLM-4.6	53.60	50.08	3.50	4.14	4.32	4.18	76.89	39.95	67.92
Gemini3-Pro	52.34	46.54	3.60	4.57	4.59	4.54	77.21	41.13	69.74
GPT-5.2	54.66	**50.70**	3.68	4.58	4.69	4.62	71.38	37.49	70.70
Doubao-Seed-1.8	**55.42**	50.40	3.81	**4.69**	4.69	4.57	74.92	39.72	71.96
Intern-S1-235B	54.15	50.36	3.96	4.28	4.75	**4.92**	78.71	44.49	73.24
S1-Base-671B	54.68	50.41	**4.02**	4.48	**4.76**	4.83	77.36	44.45	**73.59**

Table 2: Performance comparison of large language models on BIOME-Bench. Similarity refers to the cosine similarity between the embeddings of the generated answer and the reference answer. Coverage refers to the knowledge graph coverage derived using Structured Knowledge Graph Evaluation. **Avg.** is the arithmetic mean of all metrics after normalizing each score to the 0–100 range.

Intern-S1-235B (Bai et al., 2025) and S1-Base-671B (ScienceOne). We deploy Qwen3-14B, 32B and 235B with vLLM (Kwon et al., 2023), and access the remaining models through their respective APIs. For all models, we conducted a single experimental run with `temperature` $= 0$ and `max_tokens` $= 10240$.

For LLM-as-a-Judge and information extraction for Structured Knowledge Graph Evaluation, we use Qwen3-32B with `temperature` $= 0$ and `max_tokens` $= 10240$. For embedding-based similarity, we use Qwen3-Embedding-8B (Zhang et al., 2025). All experiments run on a server with $8\times$ NVIDIA A100-SXM4-80GB GPUs.

4.3 Main Results

Table 2 summarizes the overall performance of contemporary LLMs on BIOME-Bench. For biomolecular interaction inference, most models fall within a relatively narrow range (Acc 41.84% to 55.42%, Macro-F1 40.51% to 50.70%), indicating that the task remains challenging and that current progress is largely incremental. Doubao-Seed-1.8 achieves the highest accuracy (55.42%), whereas GPT-5.2 attains the best Macro-F1 (50.70%). The discrepancy suggests that leading systems are comparable in overall correctness but differ in robustness on minority classes.

For multi-omics pathway mechanism elucidation, the LLM-as-a-Judge dimensions exhibit a consistent pattern. Factuality and hallucination control are generally strong, typically above 4, while phenotype coverage is lower, around 3 to 4. This suggests that current models often produce plausible and reasonably grounded narratives, but they frequently omit required process or phenotype elements, making phenotype-level reasoning a primary bottleneck. Doubao-Seed-1.8 and GPT-5.2 score highest on causal reasoning (4.69 and 4.58) while maintaining high factuality (both 4.69). Intern-S1 performs particularly well on evidence alignment, achieving the best hallucination control (4.92) together with reliable phenotype coverage (3.96). In contrast, S1-Base-671B attains the highest phenotype coverage (4.02) and factuality (4.76), indicating an advantage in producing more complete and reliable mechanistic narratives.

In addition, similarity and knowledge graph coverage do not fully reflect judge-assessed mechanistic quality. Qwen3-32B attains the highest similarity (0.7898) and KG coverage (45.43), but it does not lead on causal reasoning or factuality. This indicates that similarity and KG coverage primarily capture proximity to reference phrasing and the breadth of entity linking, rather than the correctness and sufficiency of multi-step causal explanations. In addition, similarity can be sensitive to verbosity: stronger models often produce longer responses with auxiliary background or contextual text beyond the core mechanistic trace, which can reduce surface-form similarity despite improving explanatory content. Overall, strong BIOME performance requires balancing mechanistic completeness, causal coherence, and evidence grounding. Improvements in similarity or KG coverage therefore do not necessarily translate into better judge-assessed mechanistic reasoning.

Finally, as scientific foundation models, Intern-S1-235B and S1-Base-671B achieve stronger average performance than general-purpose models, suggesting that continued scientific-domain training has a positive effect on multi-omics pathway mech-

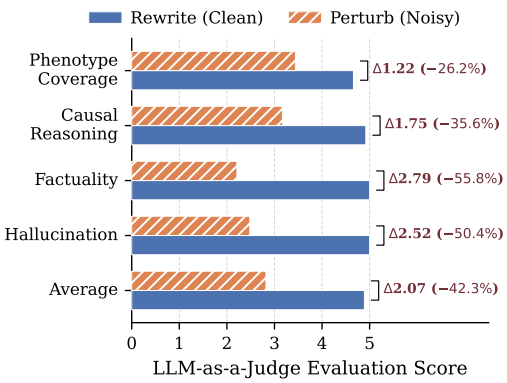

Figure 3: Sensitivity of Qwen3-32B judge to semantic perturbations. Scores are reported for rewrite and perturb . Drop% denotes the relative score decrease from rewrite to perturb.

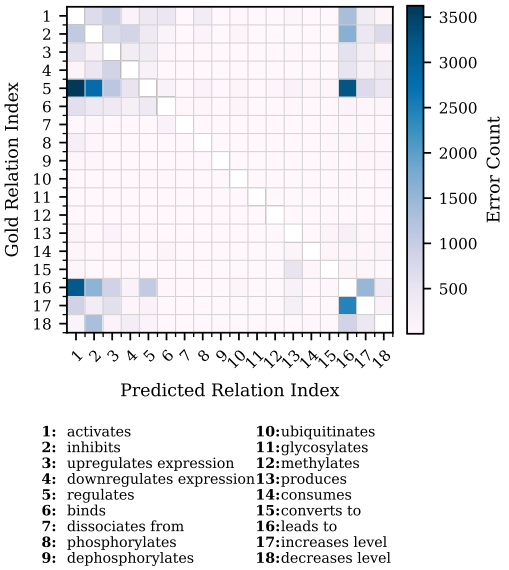

Figure 4: Error confusion matrix for Biomolecular Interaction Inference. Rows are gold relation types and columns are predicted types. Color encodes the count of misclassified gold→predicted relations.

anism elucidation. Nevertheless, the two BIOME-Bench tasks remain highly challenging for current models. Existing models still struggle to reliably distinguish fine-grained relation types and to derive robust mechanistic explanations directly from perturbed entities.

4.4 Validity of LLM-as-a-Judge

To ensure the reliability of automated evaluation in BIOME, we conducted experiments to assess the effectiveness of the LLM-as-a-judge. For each test instance, we construct two candidate answers from the gold reference. The first is a rewrite version that paraphrases the reference while preserving entities, relations, and causal semantics. The second is a perturb version that introduces targeted semantic errors by replacing key entities and/or interaction relations, while keeping the text fluent. A valid judge should remain insensitive to semantics-preserving rewrites, yet substantially penalize perturbations that degrade mechanistic correctness.

Figure 3 shows that Qwen3-32B consistently distinguishes semantics-preserving rewrites from mechanistically corrupted perturbations. The rewrite answers obtain near-ceiling scores of $4.66/4.92/5.00/5.00$ on Phenotype Coverage, Causal Reasoning, Factuality, and Hallucination, with an overall average of 4.89. In contrast, the perturb answers drop to $3.44/3.17/2.21/2.48$ and an average of 2.82. This corresponds to relative score decreases of 26.2%, 35.6%, 55.8%, and 50.4% for the four dimensions, and 42.3% on average. The large and consistent drops indicate that the judge is highly sensitive to entity and relation perturbations even when the text remains fluent, supporting the validity of Qwen3-32B as an LLM-as-a-Judge under our reference-grounded rubric.

4.5 Interaction Type Confusion in Biomolecular Inference

To analyze failure modes in *Biomolecular Interaction Inference*, we aggregate all misclassified relations and summarize the confusion matrix in Figure 4. Errors are dominated by coarse regulatory and causal labels. Across all mistakes, predictions most often fall into `leads_to` (relation 16, 11,079 cases), `activates` (1, 10,469), and `inhibits` (2, 7,930), indicating a strong tendency to default to generic causality or signed regulation under uncertainty.

Two confusions account for most of this mass. First, the underspecified label `regulates` (5) is frequently polarized or rewritten as causation, most often to `activates` (3,626), `inhibits` (2,783), or `leads_to` (3,222). Second, models blur pathway-level causation and direct regulation: gold `activates`/`inhibits` are often predicted as `leads_to` (1,325/1,643), while gold `leads_to` is over-interpreted as `activates` or `inhibits` (3,174/1,573).

Overall, the results indicate that current models still fall short in distinguishing fine-grained interaction types and separating direct regulation from pathway-level causation, making it difficult to recover complete, verifiable mechanistic chains.

Limitations

BIOME-Bench has three main limitations. First, each instance conditions on a single pathway context; extending to multi-pathway settings with explicit crosstalk and compositional reasoning is an important direction. Second, our supervision is largely organized as a one-to-one mapping between a pathway and a supporting paper; future work should build multi-pathway, multi-document mechanistic graphs that integrate dispersed evidence and enable reasoning over interconnected mechanisms. Third, our current instruction set has limited stylistic and structural diversity, while modern LLMs can be highly prompt-sensitive due to training-data distribution effects; as a result, benchmark performance may vary with prompt phrasing and introduce evaluation bias. Expanding prompt templates and reporting robustness across prompts would help mitigate this issue.

Ethics Statement

Our benchmark construction relies on multiple public biomedical resources, including KEGG, PubMed, PubChem, NCBI Gene, and UniProt. For all resources, we strictly adhere to their respective usage policies and access them solely for academic research purposes. We also control API request rates to avoid excessive traffic and ensure responsible use. In addition, we provide explicit citations for all public biomedical resources used in this work.

Large language models are used primarily to support data processing tasks. The generated content is limited to biomedical mechanistic analysis grounded in the provided literature evidence and pathway context, and is not intended to produce offensive, hateful, or otherwise harmful content. We further incorporate validation procedures and human checks to mitigate the risk of unverified or misleading statements.

We did not employ third-party annotators. The two validators involved in expert verification are collaborators on this project based in China, and their verification efforts constitute one component of their contribution to this work. Both validators are domain experts in Bioinformatics with Ph.D. degrees, ensuring the professional quality of the annotations.

This study does not involve human-subject research and does not collect personal data. All information used is derived exclusively from public biomedical databases and published scientific literature.

References

Ayushi Agrawal, Hasan Balcı, Kristina Hanspers, Susan L Coort, Marvin Martens, Denise N Slenter, Friederike Ehrhart, Daniela Digles, Andra Waagmeester, Isabel Wassink, and 1 others. 2024. WikiPathways 2024: next generation pathway database. *Nucleic acids research*, 52(D1):D679–D689.

Lei Bai, Zhongrui Cai, Yuhang Cao, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, and 1 others. 2025. Intern-S1: A scientific multimodal foundation model. *arXiv preprint arXiv:2508.15763*.

Ana R Baião, Zhaoxiang Cai, Rebecca C Poulos, Phillip J Robinson, Roger R Reddel, Qing Zhong, Susana Vinga, and Emanuel Gonçalves. 2025. A technical review of multi-omics data integration methods: from classical statistical to deep generative approaches. *Briefings in bioinformatics*, 26(4):bbaf355.

Chiara Balestra, Carlo Maj, Emmanuel Müller, and Andreas Mayr. 2023. Redundancy-aware unsupervised ranking based on game theory: Ranking pathways in collections of gene sets. *Plos one*, 18(3):e0282699.

Garth R Brown, Vichet Hem, Kenneth S Katz, Michael Ovetsky, Craig Wallin, Olga Ermolaeva, Igor Tolstoy, Tatiana Tatusova, Kim D Pruitt, Donna R Maglott, and 1 others. 2015. Gene: a gene-centered information resource at NCBI. *Nucleic acids research*, 43(D1):D36–D42.

UniProt Consortium. 2015. UniProt: a hub for protein information. *Nucleic acids research*, 43(D1):D204–D212.

John M Elizarraras, Yuxing Liao, Zhiao Shi, Qian Zhu, Alexander R Pico, and Bing Zhang. 2024. WebGestalt 2024: faster gene set analysis and new support for metabolomics and multi-omics. *Nucleic acids research*, 52(W1):W415–W421.

Yifei Ge, Feifan Zhang, Yijiang Liu, Chao Jiang, Peng Gao, Nguan Soon Tan, Sai Zhang, Yuchen Shen, Qianyi Zhou, Xin Zhou, and 1 others. 2025. Leveraging large language models for redundancy-aware pathway analysis and deep biological interpretation. *bioRxiv*, pages 2025–08.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. *arXiv preprint arXiv:2406.12793*.

Google DeepMind. Gemini 3.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577.

Minoru Kanehisa, Miho Furumichi, Yoko Sato, Yuriko Matsuura, and Mari Ishiguro-Watanabe. 2025. KEGG: biological systems database as a model of the real world. *Nucleic acids research*, 53(D1):D672–D677.

Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, and 1 others. 2016. PubChem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. DeepSeek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, and Zhiyong Lu. 2022. A sequence labeling framework for extracting drug-protein relations from biomedical literature. *Database*, 2022:baac058.

Marija Milacic, Deidre Beavers, Patrick Conley, Chuqiao Gong, Marc Gillespie, Johannes Griss, Robin Haw, Bijay Jassal, Lisa Matthews, Bruce May, and 1 others. 2024. The Reactome pathway knowledgebase 2024. *Nucleic acids research*, 52(D1):D672–D678.

Alex E Mohr, Carmen P Ortega-Santos, Corrie M Whisner, Judith Klein-Seetharaman, and Paniz Jasbi. 2024. Navigating challenges and opportunities in multi-omics integration for personalized healthcare. *Biomedicines*, 12(7):1496.

Sarah Mubeen, Alpha Tom Kodamullil, Martin Hofmann-Apitius, and Daniel Domingo-Fernandez. 2022. On the influence of several factors on pathway enrichment analysis. *Briefings in bioinformatics*, 23(3):bbac143.

National Library of Medicine (US). Medical subject headings (mesh) 2025.

OpenAI. GPT-5.2.

Ozan Ozisik, Morgane Térézol, and Anaïs Baudot. 2022. orsum: a Python package for filtering and comparing enrichment analyses using a simple principle. *BMC bioinformatics*, 23(1):293.

William G Ryan V, Smita Sahay, John Vergis, Corey Weistuch, Jarek Meller, and Robert E McCullumsmith. 2025. Pathway analysis interpretation in the multi-omic era. *BioTech*, 14(3):58.

Pedro H Godoy Sanches, Nicolly Clemente de Melo, Andreia M Porcari, and Lucas Miguel de Carvalho. 2024. Integrating molecular perspectives: Strategies for comprehensive multi-omics integrative data analysis and machine learning applications in transcriptomics, proteomics, and metabolomics. *Biology*, 13(11):848.

ScienceOne. S1-Base-671B.

ByteDance Seed, Jiaze Chen, Tiantian Fan, Xin Liu, Lingjun Liu, Zhiqi Lin, Mingxuan Wang, Chengyi Wang, Xiangpeng Wei, Wenyuan Xu, and 1 others. 2025. Seed1.5-thinking: Advancing superb reasoning models with reinforcement learning. *arXiv preprint arXiv:2504.13914*.

Mykhaylo Slobodyanyuk, Alexander T Bahcheli, Zoe P Klein, Masroor Bayati, Lisa J Strug, and Jüri Reimand. 2024. Directional integration and pathway enrichment analysis for multi-omics data. *Nature Communications*, 15(1):5690.

Jacob White. 2020. Pubmed 2.0. *Medical reference services quarterly*, 39(4):382–387.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Yijia Zhang, Hongfei Lin, Zhihao Yang, Jian Wang, and Yuanyuan Sun. 2019. Chemical-protein interaction extraction via contextualized word representations and multihead attention. *Database*, 2019:baz054.

Haiteng Zhao, Chang Ma, Lingpeng Kong, and Zhi-Hong Deng. 2025. Benchmarking and enhancing large language models for biological pathway reasoning. *OpenReview*.

Kangmei Zhao and Seung Yon Rhee. 2023. Interpreting omics data with pathway enrichment analysis. *Trends in Genetics*, 39(4):308–319.

Juexiao Zhou, Bin Zhang, Guowei Li, Xiuying Chen, Haoyang Li, Xiaopeng Xu, Siyuan Chen, Wenjia He, Chencheng Xu, Liwei Liu, and 1 others. 2024. An ai agent for fully automated multi-omic analyses. *Advanced Science*, 11(44):2407094.

Appendices

A Human Expert Verification Protocol

To evaluate the reliability of our data construction workflow and to establish a high-confidence gold standard, we performed a human-in-the-loop verification study with two domain experts in molecular biology and systems biology.

Sampling and Materials. We randomly sampled 50 *pathway–paper* pairs from the constructed dataset. For each pair, we provided (i) the paper abstract and the intermediate artifacts produced by our pipeline, including (ii) the LLM-extracted mechanistic text, (iii) the LLM-based relevance score with its abstract-grounded evidence and rationale, and (iv) the normalized entity list. For each normalized entity, we additionally supplied multiple database-derived aliases and synonyms to facilitate matching to the abstract. We also provided the structured outputs derived from the abstract, including annotated entity states and relations in the resulting knowledge graph.

Verification Steps and Criteria. Experts evaluated each instance using a four-step checklist. They were instructed to use the abstract as the evidence source, which matches the evidence constraints of our construction pipeline:

1. **Pathway–paper relevance.** Based on the abstract, determine whether the paper is highly relevant to the specified pathway context.

2. **Hallucination check for mechanistic text.** Determine whether the LLM-extracted mechanistic text contains unsupported claims, defined as statements not substantiated by the abstract.

3. **Entity presence validation.** Verify whether each normalized entity is present in the abstract, allowing matches via the provided aliases and synonyms.

4. **State and relation grounding.** Verify whether the annotated molecular states and inter-entity relations in the knowledge graph are consistent with the abstract.

We labeled an instance as **reliable** only if it satisfied all four criteria. Otherwise, it was labeled as unreliable.

Results. All 50 sampled instances satisfied the checklist criteria (100% pass rate). This result supports that our workflow produces robust intermediate artifacts and accurate structured supervision.

B Case Study: Multi-Omics Pathway Mechanism Elucidation

Legend. **Entity** *Relation* **Phenotype/Process** [State/Observation].

We present a representative end-to-end instance of multi-omics pathway mechanism elucidation for the KEGG pathway *Osteoclast differentiation* (hsa04380), grounded in a linked mechanistic study (PMID: 41266497). The instance is specified as follows:

Input. **Pathway Name:** Osteoclast differentiation – Homo sapiens (human).

Pathway Description: The osteoclasts, multinucleared cells originating from the hematopoietic monocyte-macrophage lineage, are responsible for bone resorption. Osteoclastogenesis is mainly regulated by signaling pathways activated by RANK and immune receptors, whose ligands are expressed on the surface of osteoblasts. Signaling from RANK changes gene expression patterns through transcription factors like NFATc1 and characterizes the active osteoclast.

Observed Entities: Gsk-650394, SGK1, STAT3 (dephosphorylated), MYCL (expression reduced), CTSK (transcription promoted), NFATC1.

Instruction: Based on the pathway context and biological principles, generate a coherent mechanism analysis explaining how these entities interact to produce the observed phenotype.

Output. Mechanistically, **SGK1** *phosphorylates* **STAT3** at Tyr705, and Tyr705-phosphorylated **STAT3** *upregulates* **MYCL** expression. **MYCL** directly *binds* the **CTSK** promoter and *promotes* **CTSK** transcription, thereby supporting **osteoclast differentiation** and **bone resorption**. Consistent with this signaling axis, pharmacological inhibition of **SGK1** by **GSK650394** reduces **STAT3** Tyr705 phosphorylation, leading to reduced **MYCL** expression and impaired osteoclastogenesis. Importantly, **MYCL** overexpression can *rescue* the osteoclast differentiation defect induced by **SGK1** inhibition, providing functional support for the causal

SGK1→**STAT3**→**MYCL**→**CTSK** pathway in osteoclastogenesis.

This instance specifies a chemical perturbation (**GSK650394**) and heterogeneous molecular alterations spanning signaling and transcriptional regulation, including **SGK1**, **STAT3** (dephosphorylated at Tyr705), **MYCL** (reduced expression), and **CTSK** (increased transcription), together with the pathway-context transcription factor **NFATC1**. It is challenging because the model must integrate these observations into a single, causally constrained explanation that links a targeted intervention to pathway-level phenotypes **osteoclast differentiation** and **bone resorption**. In particular, the model must (i) perform state-aware grounding to align entities with explicit molecular states (e.g., Tyr705 phosphorylation of **STAT3**) rather than treating them as state-free symbols, (ii) propagate intervention effects with correct directionality under **SGK1** inhibition, (iii) construct a multi-step causal chain from upstream kinase signaling to transcriptional control and phenotype rather than enumerating isolated facts, and (iv) incorporate pathway priors such as **NFATC1** as contextual constraints without asserting unsupported causal edges. Although these requirements reflect routine mechanistic reasoning in biology, contemporary models often fail to produce end-to-end explanations that are simultaneously state-consistent, intervention-consistent, and evidence-grounded from the provided perturbed entities and pathway context.

C Prompts for Data Construction Workflow

Prompt for Semantic and Mechanistic Relevance Scoring

 # Role
You are an expert Biocurator and Molecular Biologist specializing in pathway analysis and literature mining. Your task is to evaluate the relevance between a specific KEGG Pathway and a scientific article (PubMed).

 # Task
Analyze the provided "KEGG Pathway Info" and "Literature Info" to determine if the article provides meaningful evidence, context, or experimental data related to the pathway.

 # Input Data
1. KEGG Pathway Info
{{KEGG_PATHWAY_JSON}}
2. Literature Info
{{PAPER_JSON}}

 # Evaluation Criteria & Steps

1. **Subject Matching**: Does the article primarily discuss the biological process described in the pathway (e.g., Glycolysis, Gluconeogenesis)? Differentiate between core focus vs. background mention.

2. **Species Consistency**: Check if the species in the pathway (e.g., "hsa" for Human) matches the organism model in the paper.
    - Note: If the pathway is Human but the paper uses a model organism (e.g., Mouse/Murine) to simulate human physiology/disease, this is considered **Relevant**.
    - Note: If the species are completely unrelated (e.g., Plant pathway vs. Human study), penalize the score.

3. **Molecular Evidence**: Look for specific mentions of the genes, enzymes, or metabolites described in the pathway description.

4. **Directionality/Context**: Does the paper discuss the activation, inhibition, or regulation of this pathway?

 # Scoring Standard (0–10)

- **0–1 (Irrelevant)**: No meaningful connection. The terms might appear only in references or unrelated contexts.
- **2–4 (Low)**: Pathway is mentioned as a keyword or broad concept, but not investigated. Major species mismatch without translational value.
- **5–7 (Medium)**: The pathway is part of the results (e.g., "we observed changes in glycolysis"). Valid species match or relevant model.
- **8–10 (High)**: The pathway is the central topic. The paper elucidates mechanisms, regulation, or disease implications of this specific pathway. High species alignment.

 # Output Format
Provide the result in a valid JSON object strictly adhering to the following structure. Do not output markdown backticks (```) or extra text.

```
{
  "relevance_score": <int, 0-10>,
  "relevance_level": "<High/Medium/Low/Irrelevant>",
  "species_check": "<Briefly state if species match or if a valid model organism is used>",
  "evidence_summary": "<Extract 1-2 key sentences from the abstract that support the link>",
  "reasoning": "<Concise explanation of the score, focusing on biological mechanisms and study focus>"
}
```

Prompt for Entity Extraction and Mechanism Summarization

 # Role
You are an expert Systems Biologist and Biomedical Literature Curator. Your task is to extract, categorize, and standardize key molecular information to construct a high-quality "Ground Truth" benchmark for multi-omics pathway analysis.

 # Goal
You will be provided with a **Target Pathway** and a scientific paper's details. Your goal is to generate two structured outputs:

1. **Significant Entities (Classified & Normalized)**: A structured list of metabolites, genes, proteins, or phenotypes explicitly mentioned in the text. You must classify them by type and normalize them (expand abbreviations) to facilitate downstream database mapping.

2. **Mechanism Analysis**: A concise, coherent biological explanation of how these entities interact. Crucially, this must be written as a direct fact or expert interpretation, not as a summary of a study.

 # Instructions

- **Entity Classification & Normalization**
  - **Chemicals/Metabolites**: Target for PubChem mapping. Provide the original text found in the abstract and a standard_name (full chemical name, expand abbreviations like 5-ALA to 5-aminolevulinic acid).
  - **Genes/Proteins**: Target for UniProt/NCBI mapping. Provide the original text and a standard_name (official symbol or full protein name).
  - **Processes/Phenotypes**: Target for GO/MeSH. Biological outcomes or processes (e.g., Ferroptosis, Oxidative Stress). Keep these separate as they are not valid targets for chemical/gene databases.
- **Tone & Style (Critical)**: Write the mechanism_analysis as an objective biological fact. Do not use phrases such as "This study reveals", "The authors found", or "We observed". Start directly with the biological subject (e.g., "Elevated levels of X cause . . . ").
- **Contextualize**: Use the Pathway Description and Evidence Summary to filter for relevance.
- **Precision**: Extract only entities explicitly mentioned in the text.

 # Constraints

1. **No Hallucinations**: Do not invent standard names if an abbreviation is ambiguous. Use the most likely biological expansion based on context.
2. **Relevance**: Focus on the **Target Pathway**.
3. **No Meta-Language**: Strictly ban words referring to the source material in the mechanism analysis.
4. **Format**: Return strict JSON only.

 # Input
Target Pathway: {{pathway_name}}
Pathway Description: {{pathway_description}}
Title: {{title}}
Abstract: {{abstract}}
Evidence Summary: {{evidence_summary}}

 # Output JSON Format
Do not output markdown backticks (```) or extra text.

```
{
  "significant_entities": {
    "chemicals": [
      { "original": "Text in abstract", "standard_name": "Full Standardized Name" }
    ],
    "genes_proteins": [
      { "original": "Text in abstract", "standard_name": "Official Symbol/Name" }
    ],
    "processes_phenotypes": [
      "Phenotype 1",
      "Phenotype 2"
    ]
  },
  "mechanism_analysis": "A direct biological explanation describing the interaction of these entities and the resulting outcome."
}
```

Prompt for Interaction Quadruple Extraction

 # Role
You are an expert Biological Knowledge Graph Constructor. Your task is to convert a biological mechanism text into a structured list of interactions (quadruplets) using a strict vocabulary.

 # Task
Extract all biological interactions from the "Mechanism Text" and map them to the following JSON structure: {"source": "Entity A", "relation": "Relation_Type", "target": "Entity B", "condition": "Context/Prerequisite"}.

 # Constraints

1. **Entity Mapping**: Use the exact names provided in the "Standardized Entities List" for source and target.
    - Note the **[Type]** tags provided in the input (e.g., [Chemical], [Gene], [Phenotype]) to understand the biological context of each entity.
    - If an entity is missing from the list but critical for the logic, use its name from the text.

2. **Relation Vocabulary (Strict)**: Use only the specific relation strings defined in the lists within the following JSON schema. Use the inline comments (//) as guidance.

```
{
  "Regulatory": [
    "activates",                // covers: activation
    "inhibits",                 // covers: inhibition
    "upregulates_expression",   // covers: expression
    "downregulates_expression", // covers: repression
    "regulates"                 // covers: indirect effect, state change
  ],
  "Physical_Interaction": [
    "binds",                    // covers: binding/association
    "dissociates_from"          // covers: dissociation
  ],
  "Modification": [
    "phosphorylates",           // covers: phosphorylation
    "dephosphorylates",         // covers: dephosphorylation
    "ubiquitinates",            // covers: ubiquitination
    "glycosylates",             // covers: glycosylation
    "methylates"                // covers: methylation
  ],
  "Metabolic": [
    "produces",                 // covers: compound (enzyme -> product)
    "consumes",                 // covers: metabolic_reaction (substrate -> enzyme)
    "converts_to"               // covers: metabolic_reaction (substrate -> product)
  ],
  "Causal (Phenotypic)": [
    "leads_to",                 // e.g., leads to Apoptosis
    "increases_level",          // e.g., increases ROS levels
    "decreases_level"
  ]
}
```

3. **Compound Handling**: If "Gene A increases Metabolite B to activate Gene C", split it into:
    - (Gene A, increases_level, Metabolite B)
    - (Metabolite B, activates, Gene C)

    Do not use compound as a relation.

4. **Condition Extraction**: Extract a short, specific phrase describing when, where, or how the interaction happens (e.g., "upon light activation", "in HO-1 deficient cells", "under hypoxia", "when combined with DCA"). If the interaction is a general biological fact with no specific constraint, use "General".

 # Input
Mechanism Text: {{mechanism_text}}
Standardized Entities List: {{formatted_entity_str}}

 # Output Format
Return a strict JSON object following this exact schema:

```
[
  {
    "source": "Entity A",
    "relation": "Relation_Type_From_Vocabulary",
    "target": "Entity B",
    "condition": "Context string or 'General'"
  }
]
```

Prompt for Biological State Annotation

 # Role
You are an expert Biological Graph Annotator.

 # Task
You will be provided with:

1. **Mechanism Text**: A biological description.
2. **Existing Interactions**: A list of structured interactions (quadruplets) extracted from the text.

Your job is to annotate the biological state (e.g., increased, mutated, added, phosphorylated) for the source and target entities in each interaction, based strictly on the text.

 # Instructions

1. For each interaction in the list, add two new fields: source_state and target_state.
2. Extract the state as a short phrase from the text.
    - Examples: "elevated levels", "administration", "mutated", "deficiency", "accumulation", "overexpression".
3. **Treatment/Drug Handling**: If an entity is a drug or treatment added to the system, use terms such as "administration", "added", or "treated with".
4. **Default**: If the text does not specify a change or state for that entity in that specific context (i.e., it is simply a pathway component), use "Present" or "Endogenous".
5. **Constraint**: Do not change the original source, relation, target, or condition values. Keep them exactly as provided.

 # Input
Mechanism Text: {{mechanism_text}}
Existing Interactions to Annotate: {{existing_graph_json}}

 # Output Format
Return a strict JSON object with a single key "annotated_kg" containing the updated list.

 # Example Output Schema

```
{
  "annotated_kg": [
    {
      "source": "Entity A",
      "source_state": "elevated levels",
      "relation": "increases_level",
      "target": "Entity B",
      "target_state": "accumulation",
      "condition": "General"
    }
  ]
}
```