CROSSPERTURB: A CROSS-CONTEXT BENCHMARK AND EVIDENCE-GROUNDED FRAMEWORK FOR LLM-BASED VIRTUAL CELL PERTURBATION PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting cellular responses to drug perturbations is fundamental to virtual cell modeling and therapeutic discovery. While large language models (LLMs) show growing promise for biological reasoning, existing benchmarks evaluate them solely within the same cellular context. Consequently, simple voting baselines achieve comparable performance to complex LLM reasoning frameworks, obscuring whether models perform genuine biological reasoning or merely exploit majority labels. This limitation highlights the urgent need for more rigorous evaluation benchmarks that test whether LLMs can perform mechanistic reasoning in virtual cell modeling. To bridge this gap, we introduce CrossPerturb, a comprehensive benchmark for evaluating drug perturbation prediction across diverse cellular contexts. Derived from the LINCS L1000 and Tahoe datasets, CrossPerturb assesses transcriptional outcomes across diverse cell lines, drug dosage, and exposure durations. Alongside the benchmark, we introduce ARCS, an evidence-grounded framework that **A**nalyzes biological pathways underlying perturbations, **R**etrieves cross-context evidence, **C**ritiques candidate evidence, and **S**ynthesizes the final response score through mechanism-constrained aggregation. Extensive experiments demonstrate that our evidence integration framework surpasses current state-of-the-art methods, whereas existing LLM reasoning frameworks struggle to achieve robust performance, confirming that cross-context biological reasoning remains an open challenge for current LLMs.

1 INTRODUCTION

Understanding complex cellular processes requires systematic intervention followed by comprehensive profiling of cellular responses (Beck et al., 2024). While high-throughput screening generates massive profiling data, experimental constraints and prohibitive costs restrict comprehensive coverage across the vast space of cellular conditions and therapeutic interventions (Dixit et al., 2016; Roohani et al., 2024; Li et al., 2024). These empirical bottlenecks have motivated virtual cell modeling (Roohani et al., 2025), which leverages computational approaches to simulate cellular behavior in response to chemical or genetic perturbations, thereby accelerating biological discovery and therapeutic development (Noutahi et al., 2025).

Recent advancements in Large Language Models (LLMs) offer a transformative perspective for virtual cell modeling by recasting perturbation prediction as a natural-language reasoning task (Wu et al., 2025). This paradigm relies on the capabilities of LLMs to perform complex reasoning (Ke et al., 2025) and interpret structured biological knowledge, such as drug mechanisms and pathway relationships (Chandak et al., 2023; Wei et al., 2026). Consequently, LLMs offer an interpretable, knowledge-guided approach that extends beyond purely data-driven prediction.

Despite this progress, existing language-based perturbation benchmarks are confined to within-context evaluation, where retrieved evidence and test perturbations share the same biological context. Consequently, they fail to evaluate whether LLMs can generalize across varying cellular contexts (e.g., cell lines, drug dosage, and exposure durations), which is a core requirement for reliable virtual cell modeling (Dibaeinia et al., 2026). More critically, recent studies show that current LLM frameworks can be easily outperformed by simple statistical baselines (Yuan et al., 2026). As il-

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

lustrated in Figure 1, our evaluation on PerturbQA reveals that simple voting outperforms advanced LLM framework SUMMER. Under within-context evaluation, genes frequently exhibit repetitive response tendencies, allowing models to exploit shortcut patterns rather than learning true perturbation specificity. Together, these findings indicate that plausible biological explanations often mask a lack of genuine reasoning, causing current benchmarks to drastically overrate LLM capabilities.

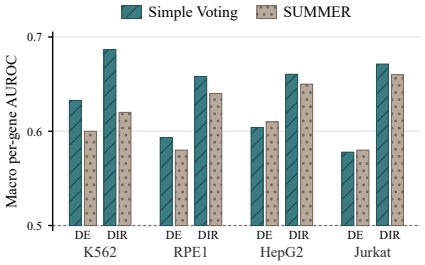


Figure 1: Prediction performance on PerturbQA, showing that simple voting strategy performs better than SUMMER on differential expression (DE) and direction of change (DIR) tasks.

To overcome these challenges, we introduce CrossPerturb, a comprehensive drug perturbation benchmark designed to evaluate how LLMs reason across diverse cellular contexts. Encompassing over a thousand compounds under varying experimental conditions, CrossPerturb assesses gene differential expression and regulation direction within diverse cellular contexts, providing a rigorous evaluation framework for LLM-based biological reasoning. Complementing the benchmark, we develop ARCS, an LLM-guided evidence integration framework with three steps. ARCS first analyzes the queried drug and gene using textual descriptions and structured biological relations to identify candidate mechanisms and relevant evidence sources. It then retrieves cross-context observations and critiques their compatibility with the target context. Finally, ARCS combines the selected evidence with a target-context prior through bounded aggregation to predict the final perturbation response. Our main contributions can be summarized as follows:

- We find that evaluation within observed cellular contexts can overestimate model generalization by exploiting gene-specific response tendencies rather than context-sensitive evidence, enabling simple statistical baselines to surpass advanced LLM frameworks.
- We introduce CrossPerturb, the first cross-context benchmark for virtual cell perturbation prediction. CrossPerturb systematically evaluates how models generalize across diverse experimental conditions, establishing that cross-context biological reasoning remains a critical open challenge for current LLMs.
- We propose ARCS, a evidence-grounded framework that integrates reliable evidence for robust perturbation prediction, achieving state-of-the-art performance across diverse evaluation settings on the CrossPerturb benchmark.

2 Related Work

Virtual Cell Modeling for Perturbation Response Prediction. Virtual cell modeling (Roohani et al., 2025) integrates mathematical modeling and artificial intelligence to simulate intracellular biochemical, genetic, and physiological processes in silico, overcoming the cost-intensive and time-consuming limitations of wet-lab experiments. Advances in machine learning and single-cell profiling (Li et al., 2026) have expanded this concept to encompass models that forecast cellular behaviors and formulate experimental hypotheses. Existing computational methods for perturbation modeling can be broadly categorized into three paradigms: generative response models that learn conditional transformations from untreated states to post-intervention profiles, like chemCPA (Hetzel et al., 2022) and CellFlow (Klein et al., 2025); knowledge-informed models that explicitly incorporate molecular relationships or regulatory networks, including GEARS (Roohani et al., 2024); and large-scale foundation models that acquire transferable cellular representations through extensive pretraining, such as scGPT (Cui et al., 2024), and State (Adduri et al., 2025). Despite this progress, most evaluations emphasize prediction within familiar experimental settings, leaving the reliability of virtual cell models across distinct cellular contexts insufficiently characterized.

LLM-based Reasoning Frameworks for Biological Discovery. Propelled by rapid advances in prompt engineering (Sahoo et al., 2024) and agent frameworks (Zhao et al., 2025), large language models (LLMs) are increasingly applied to interpretable biological discovery, leveraging their robust comprehension capabilities to uncover underlying biological mechanisms. For instance, in-context learning frameworks like SUMMER (Wu et al., 2025) and VCWorld (Wei et al., 2026) integrate structured contextual prompts with extensive prior knowledge bases to predict perturbation effects.

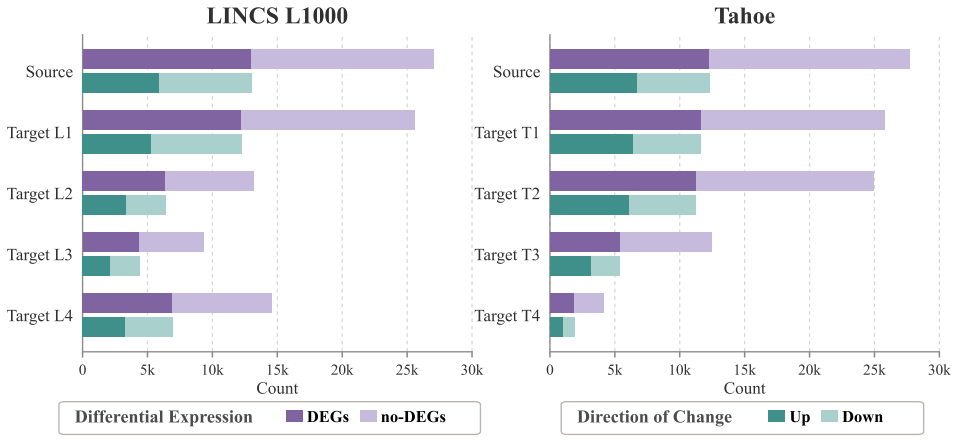


Figure 2: Benchmark composition of CrossPerturb across source observations and target regimes for LINCS L1000 and Tahoe. Stacked bars show count of DEGs and no-DEGs for Differential Expression task, and upregulated and downregulated genes for Direction of Change task.

Furthermore, advanced agent frameworks such as PantheonOS (Xu et al., 2026) and PerturbAgent (Pei et al., 2026) orchestrate autonomous execution and iterative reasoning to analyze perturbation response and decode complex regulatory pathways. However, Yuan et al. (2026) recently showed that a simple gene-frequency baseline outperforms existing LLM-based frameworks on standard benchmarks. This indicates that while LLMs produce biologically plausible explanations, they fail to capture true perturbation-specific effects. This vulnerability arises because current benchmarks evaluate models within the same cellular contexts, where genes frequently exhibit consistent perturbation responses that favor simple aggregation strategies. Consequently, establishing cross-context benchmarks is essential to rigorously evaluate true biological reasoning.

Biological Knowledge Bases and Perturbation Benchmarks. Structured biological databases offer essential priors concerning molecular interactions and functional organization. STRING (Mering et al., 2003) and CORUM (Giurgiu et al., 2019) curate protein interactions and complex formations, while the Gene Ontology (GO) (Ashburner et al., 2000) and Reactome (Croft et al., 2010) delineate high-level pathway hierarchies. Additionally, specialized pharmacological and regulatory databases are crucial for characterizing drug actions. Specifically, DrugBank (Knox et al., 2024), and PrimeKG (Chandak et al., 2023) offer structured insights into drug-target interactions, binding affinities, and pharmacological properties, whereas TRRUST (Han et al., 2018) captures transcription factor and target gene regulations. In parallel, benchmarks such as PerturbQA (Wu et al., 2025) and GeneTak (Wei et al., 2026) support standardized evaluation of perturbation reasoning. However, existing benchmarks are restricted to the same cellular context, neglecting the cross-context capability essential for virtual cell modeling. Therefore, we construct a novel cross-context benchmark by leveraging LINCS L1000 (Subramanian et al., 2017) and Tahoe (Zhang et al., 2025), two small-molecule drug perturbation datasets encompassing rich experimental conditions.

3 CrossPerturb Benchmark

To facilitate rigorous evaluation of LLMs for virtual cell perturbation prediction, we introduce CrossPerturb, a novel benchmark centered on cross-context reasoning. CrossPerturb comprises two complementary high-throughput resources based on LINCS L1000 (Subramanian et al., 2017) and Tahoe (Zhang et al., 2025). Each experimental record is indexed by the administered drug, the resulting gene responses, and its cellular context, which is defined by the cell line, drug dosage, and exposure duration in this work. Encompassing diverse experimental settings, CrossPerturb provides a robust testbed for evaluating whether LLMs can generalize to diverse cellular contexts.

Dataset Curation. CrossPerturb processes expression profiles by retaining 973 landmark genes for L1000 and selecting the 2,000 most highly variable genes for Tahoe, following standard preprocessing protocols in perturbation modeling. The LINCS L1000 component comprises 1,178 compounds across 11 cell lines, spanning seven standardized drug-dose levels (0.04, 0.12, 0.37, 1.0,

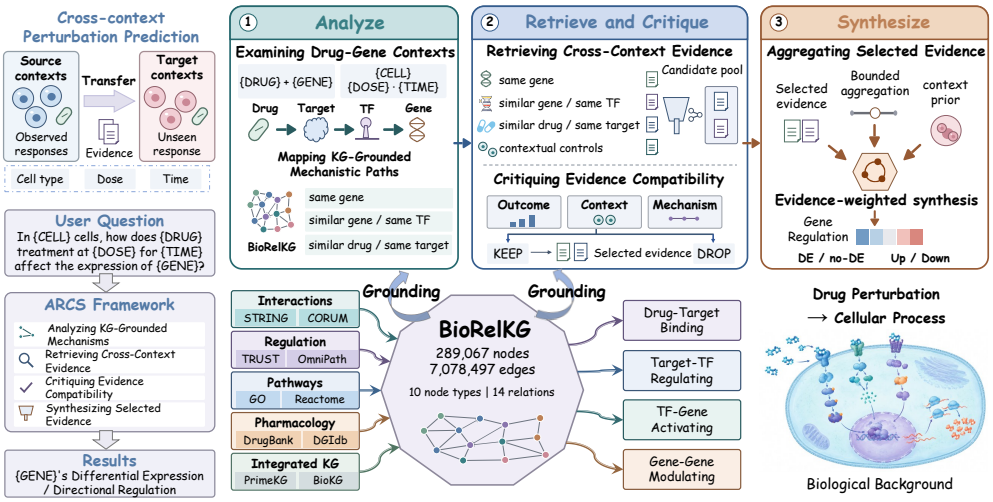


Figure 3: **Overview of the ARCS framework.** Leveraging the BioRelKG knowledge graph, ARCS performs mechanistic analysis, cross-context evidence retrieval and critique, and bounded aggregation for robust cross-context drug perturbation prediction.

3.3, 5.0, and 10.0 μM) and two exposure durations (6 and 24 h) from Phase I and Phase II experiments. The Tahoe component comprises 168 compounds across 12 cell lines, with three drug-dose levels (0.05, 0.5, and 5 μM) measured after 24 h of exposure. Together, these components provide complementary coverage of cellular and experimental contexts while maintaining sufficient sample support for perturbation-response evaluation. Labels are generated for perturbation-context-gene triplets from pseudobulk expression contrasts. Differential expression is assessed using moderated t-statistics implemented in the limma package (Ritchie et al., 2015). Genes satisfying the significance criteria are labeled as DEGs, with their regulation directions determined by the sign of the expression change. The Differential Expression task includes both DEGs and sampled no-DEGs from the same cellular context, whereas the Direction of Change task includes only DEGs and distinguishes upregulation from downregulation.

Evaluation Protocol. Within each dataset, compounds are partitioned into disjoint source and target sets at a 3:7 ratio, following the few-shot evaluation design adopted by VCWorld (Wei et al., 2026) for assessing model generalization. Target queries are assigned to four dataset-specific context settings. For LINCS L1000, Target L1 corresponds to within-context evaluation, where the cell line, drug dosage, and exposure duration are all represented in the source observations. Target L2 tests generalization to an unseen cell line, Target L3 examines an unseen exposure duration, and Target L4 evaluates an unseen drug dose. For Tahoe, Target T1 serves as the within-context setting, whereas Target T2 focuses on an unseen cell line, Target T3 considers an unseen drug dosage, and Target T4 combines an unseen cell line with an unseen drug dosage. Overall, L1 and T1 evaluate within-context generalization, while the remaining settings measure cross-context generalization along one or more context dimensions. In all target settings, the target compounds are held out from the source compound set. Figure 2 summarizes the label composition and query distribution of the complete benchmark. Complete preprocessing procedures and split statistics are provided in Appendix A.

4 Method

4.1 Task Formulation

We formulate CrossPerturb as a gene-centric prediction problem under an explicit cellular context. Each prediction target is defined as a representative triplet $q = (p, g, \mathbf{s})$, where $p \in \mathcal{P}$ denotes a drug, $g \in \mathcal{G}$ denotes a response gene, and $\mathbf{s} = (c, d, t) \in \mathcal{S}$ specifies the cell line, drug dosage, and exposure duration. Each triplet is assigned a binary label $l \in \{0, 1\}$ across two fine-grained classification tasks: (i) **Differential Expression (DE)**, where $l = 1$ indicates that g is differentially expressed under drug p and context $\mathbf{s}$, and $l = 0$ otherwise; (ii) **Direction of Change (DIR)**, where $l = 1$ indicates upregulation and $l = 0$ indicates downregulation. The DIR task is evaluated exclusively for genes identified as differentially expressed in the corresponding experimental contrast.

These tasks are cast as knowledge-grounded inference, where a Large Language Model (LLM) $\mathcal{L}$ serves as the reasoning engine. For each triplet q, similar labeled perturbation examples $\mathcal{E}_q \subseteq \mathcal{E}_{\text{train}}$ are retrieved from the observed source data, while query-specific biological evidence $\mathcal{K}_q \subseteq \mathcal{K}$ is retrieved from external knowledge bases, including drug-target relation, signaling pathways and protein-protein interaction databases. The resulting text-based inference is formulated as $\hat{l}^\tau = \mathcal{L}(q, \mathcal{E}_q, \mathcal{K}_q; \tau)$, where $\tau \in \{\text{DE}, \text{DIR}\}$ denotes the prediction task.

4.2 BioRelKG: A Heterogeneous Biological Knowledge Graph

LLM-based biological reasoning relies on structured evidence connecting drugs, genes, and underlying molecular processes. For perturbation prediction, such relational evidence is essential to move beyond surface-level textual similarity, particularly when generalizing to unseen drugs or cellular contexts. While existing biomedical knowledge graphs offer valuable biological priors, they are often sparse or fragmented across individual sources, leaving many drug-gene pairs without traversable paths through shared targets or regulatory networks. To bridge these gaps, we construct **BioRelKG**, a heterogeneous biomedical knowledge graph designed to expand relational evidence and supply interpretable paths for query-specific retrieval.

BioRelKG represents diverse biological entities as typed nodes, including drugs, genes, proteins, pathways, functional terms, and protein complexes, with edges encoding drug-target interactions, transcriptional regulation, signaling pathways, and functional annotations. To ensure comprehensive coverage, BioRelKG integrates multiple foundational resources: DrugBank (Knox et al., 2024) for pharmacological and drug-target relations, STRING (Mering et al., 2003) for protein-protein interactions, TRRUST (Han et al., 2018) and OmniPath (Türei et al., 2026) for regulatory networks, Gene Ontology (Ashburner et al., 2000) and Reactome (Croft et al., 2010) for pathway annotations, CORUM (Giurgiu et al., 2019) for protein complexes, and PrimeKG (Chandak et al., 2023) for integrated biomedical associations. Depending on graph topology, retrieved evidence spans direct associations, local neighborhoods, or multi-hop paths. Appendix B details the complete resource provenance, relation types, and graph statistics.

4.3 ARCS Framework

As shown in Figure 3, we propose ARCS, an LLM-guided framework that integrates structured biological knowledge with empirical evidence for cross-context perturbation prediction. It consists of three key stages: (1) analyzing KG-grounded mechanisms and prioritizing relevant evidence; (2) retrieving cross-context observations and critiquing their compatibility with the target query; and (3) synthesizing the selected evidence with a target-context prior through bounded aggregation.

4.3.1 Analyzing KG-Grounded Perturbation Mechanisms via LLMs

To provide explicit biological context for subsequent mechanism analysis, we construct textual representations that capture the molecular functions and interaction profiles of the queried drug p and response gene g. We collect a functional summary F_v for each entity $v \in \{p, g\}$ from curated biomedical databases, such as NCBI (Sayers et al., 2021), and retrieve its complete one-hop neighborhood $\mathcal{N}_1(v)$ from BioRelKG. This neighborhood contains all directly connected entities and their typed relations, such as drug-target associations, transcriptional regulation, and signaling relations. We integrate these complementary sources via LLM-based summarization to obtain textual representations that capture each entity's biological functions and molecular interactions:

$$D_v = \mathcal{L}_{\text{Desc}}\big(v, F_v, \mathcal{N}_1(v)\big), \qquad v \in \{p, g\}, \tag{1}$$

where D_p and D_g represent the drug and response gene, respectively. Then, we formulate structured and mechanistic KG-grounded drug-to-gene regulatory hypotheses and prioritize the most informative evidence sources via LLM reasoning:

$$\big(H_q, \Pi_q\big) = \mathcal{L}_{\text{Analyze}}\big(q, D_p, D_g, \mathcal{N}_1(p), \mathcal{N}_1(g)\big). \tag{2}$$

Here, H_q captures candidate regulatory paths involving drug targets, transcriptional regulators, or pathways, while Π_q prioritizes evidence from the same gene, similar genes sharing a regulator, and pharmacologically similar drugs. This prioritization guides subsequent retrieval toward more relevant observations. $\mathcal{L}_{\text{Desc}}$ and $\mathcal{L}_{\text{Analyze}}$ denote the LLM-based summarization and mechanism-analysis functions, respectively, with their prompt templates provided in Appendix J.

4.3.2 RETRIEVING AND CRITIQUING CROSS-CONTEXT EVIDENCE

The observed source data are defined as $\mathcal{E}_{\text{train}} = \{e_i = (q_i, l_i)\}_{i=1}^{M}$, where $q_i = (p_i, g_i, \mathbf{s}_i)$ and $\mathbf{s}_i = (c_i, d_i, t_i)$ denote the drug, response gene, and cellular context of the i-th perturbation example. For a query q and the query q_i associated with a candidate example, we define their similarity as:

$$\text{Sim}(q, q_i) = \alpha \, \text{Sim}_{\text{mol}}\big(\mathbf{d}_{p,g}, \mathbf{d}_{p_i,g_i}\big) + (1-\alpha) \, \text{Sim}_{\text{con}}\big(\mathbf{s}, \mathbf{s}_i\big), \tag{3}$$

where Sim_{mol} measures similarity between the drug–gene representations, and Sim_{con} evaluates compatibility between their cellular contexts. The molecular representations are obtained by encoding entity descriptions as $\mathbf{d}_v = f(D_v)$, while context similarity evaluates cell-line, dosage, and exposure-time compatibility. Complete similarity criteria are provided in Appendix E.

Given the evidence priorities Π_q, candidate examples are selected under three distinct KG-guided matching criteria: the same response gene, genes associated with prioritized transcription factors or pathways, and drugs associated with prioritized targets or pharmacological similarity. Let $\mathcal{R}_{\Pi_q}$ denote the selected criteria; the final candidate retrieval pool is then constructed by retaining the top-k most relevant examples under each criterion:

$$\mathcal{C}(q) = \cup_{r \in \mathcal{R}_{\Pi_q}} \text{TopK}_k \left(\mathcal{E}_r(q; \Pi_q), \text{Sim}(q, q_i) \right), \tag{4}$$

where $\mathcal{E}_r(q; \Pi_q)$ denotes candidate examples satisfying criterion r under priorities Π_q.

Similarity-based retrieval identifies potentially relevant observations, but textual similarity alone does not guarantee that an observation is suitable for supporting the target prediction. An LLM-based critique therefore examines each example's observed response, retrieval basis, and relevant biological relations before assigning a binary decision:

$$v_i = \mathcal{L}_{\text{Critique}}\big(q, D_p, D_g, H_q, e_i\big), \qquad v_i \in \{\text{KEEP}, \text{DROP}\}. \tag{5}$$

4.3.3 SYNTHESIZE

Finally, ARCS synthesizes the evidence retained by the critique with the empirical response tendency of the target gene under the query context. Let $\mathcal{S}(q) = \{e_i \in \mathcal{C}(q) \mid v_i = \text{KEEP}\}$ denote the retained evidence. The evidence-based response score is defined as follows:

$$f(q) = \frac{1}{|\mathcal{S}(q)|} \sum_{e_i \in \mathcal{S}(q)} l_i, \qquad l_i \in \{0, 1\}, \tag{6}$$

where l_i denotes the binary label for differential expression or direction of change.

To account for gene-specific response tendencies observed in existing perturbation benchmarks, we estimate $p_{\text{prior}}(q)$ from source observations of the same or similar response gene under the target cellular context $\mathbf{s} = (c, d, t)$. We use exact context matches when available. When no exact context matches exist, we use the closest available context based on cell line, drug dosage, and exposure duration. The perturbation response score is defined as follows:

$$p(q) = \lambda f(q) + (1 - \lambda) p_{\text{prior}}(q). \tag{7}$$

Here, $f(q)$ reflects responses of the selected cross-context examples, whereas $p_{\text{prior}}(q)$ captures the context-specific response tendency of the target gene. Their weighted aggregation allows selected evidence to refine the empirical prior while reducing sensitivity to heterogeneous retrieved examples. The final binary perturbation prediction is calculated as:

$$\hat{l}(q) = \mathbb{I}[p(q) \geq 0.5]. \tag{8}$$

Overall, rather than relying on complex chain-of-thought reasoning, ARCS performs LLM-guided evidence prioritization and assessment, followed by principled evidence synthesis for robust cross-context biological perturbation prediction under distribution shifts.

5 EXPERIMENTS

5.1 BASELINES

We benchmarked performance against several statistical, generative, and language-based baselines on CrossPerturb. **Simple Voting** predicts the response of a query gene by aggregating empirical

Table 1: Population-level AUROC for Differential Expression (DE) and Direction of Change (DIR) prediction across the four context settings. Bold indicates the best performance.

Task	Method	LINCS L1000				Tahoe			
		L1	L2	L3	L4	T1	T2	T3	T4
DE	Simple Voting	0.60±.00	0.51±.00	0.53±.00	0.46±.00	0.55±.00	0.50±.00	0.51±.00	0.51±.00
	GAT	0.57±.03	**0.52±.01**	0.52±.02	0.52±.01	0.54±.02	0.50±.00	0.51±.01	0.51±.03
	scVI	0.51±.02	0.50±.01	0.51±.00	0.50±.01	0.61±.03	**0.60±.02**	0.59±.01	**0.60±.02**
	ChemCPA	0.50±.01	0.50±.01	0.46±.02	0.44±.02	0.50±.01	0.49±.01	0.50±.01	0.48±.02
	State	0.53±.01	0.51±.03	0.51±.01	0.50±.02	0.54±.00	0.51±.01	0.50±.03	0.50±.02
	VCWorld	0.52±.01	0.46±.00	0.48±.01	0.44±.00	0.56±.01	0.53±.00	0.51±.00	0.51±.00
	ARCS	**0.60±.02**	0.51±.03	**0.54±.01**	**0.53±.01**	**0.63±.03**	0.52±.02	**0.60±.02**	0.54±.01
DIR	Simple Voting	0.68±.00	0.61±.00	0.63±.00	0.65±.00	0.70±.00	0.60±.00	0.68±.00	0.58±.00
	GAT	0.72±.02	**0.64±.02**	0.63±.01	0.69±.03	0.39±.01	0.42±.02	0.40±.03	0.41±.01
	scVI	0.52±.01	0.52±.02	0.54±.01	0.54±.02	0.55±.01	0.57±.01	0.56±.02	0.59±.03
	ChemCPA	0.51±.01	0.50±.02	0.51±.01	0.51±.00	0.50±.00	0.50±.01	0.50±.00	0.50±.02
	State	0.54±.00	0.52±.00	0.52±.00	0.48±.00	0.58±.00	0.53±.00	0.52±.00	0.51±.00
	VCWorld	0.64±.02	0.53±.03	0.56±.00	0.48±.04	0.52±.01	0.51±.01	0.52±.02	0.50±.01
	ARCS	**0.78±.02**	0.63±.01	**0.65±.01**	**0.73±.00**	**0.81±.03**	**0.67±.02**	**0.78±.00**	**0.63±.01**

labels from same gene perturbations with similar cellular context, providing a non-parametric baseline based on gene-level response tendencies. **GAT** uses biological relations from BioRelKG to construct a graph attention network and directly predicts binary perturbation outcomes. **scVI** is a conditional variational autoencoder that models cellular states, enabling generation of perturbation-specific expression responses (Lopez et al., 2018). **ChemCPA** recognizes novel perturbations by learning disentangled and linearly combined latent representations of cellular states and perturbations (Hetzel et al., 2022). **State** is a pretrained model for predicting cellular responses to chemical perturbations from large-scale single-cell data (Adduri et al., 2025). **VCWorld** is an LLM-based virtual-cell framework that incorporates biological context into perturbation reasoning (Wei et al., 2026). Detailed descriptions of the baselines are provided in Appendix C.

5.2 Evaluation Metrics

We assess model performance from two complementary perspectives: population-level and gene-level perturbation prediction. To ensure a fair comparison, models that produce continuous expression profiles are converted into binary perturbation labels using the same differential expression procedure applied to the benchmark annotations, whereas models that directly predict binary outcomes are evaluated using their native predictions. We report accuracy, AUROC, and F1 score under both aggregation schemes. For population-level evaluation, each perturbation-context-gene triplet is treated as an individual prediction instance. For gene-level evaluation, scores are first computed separately for each gene across its associated perturbation queries and subsequently averaged across genes. Detailed metric definitions and aggregation procedures are provided in Appendix D. VCWorld and ARCS adopt Qwen3-8B (Yang et al., 2025) as the underlying reasoning model, with complete inference settings and hyperparameter analysis detailed in Appendix E.

5.3 Population-level Performance

We evaluate population-level performance by aggregating all perturbation instances within each target-context setting. We adopt AUROC as the primary metric to assess class discrimination across the complete cell population, while Accuracy and F1 scores are reported in Appendix F. L1–L4 and T1–T4 denote the four settings (see Section 3) for LINCS L1000 and Tahoe, respectively.

Differential Expression Task Performance. Table 1 shows that all methods perform better in within-context settings (L1 and T1) than in cross-context settings, with the most substantial performance declines observed under unseen cell line settings. ARCS achieves highly competitive performance across all settings, although scVI outperforms it in several Tahoe cross-context scenarios. Simple voting and GAT outperform deep generative models, including ChemCPA and State, in most cross-context settings, suggesting that empirical gene response tendencies and structured biological

Table 2: Overall macro per-gene AUROC, Accuracy, F1 on DE and DIR tasks.

Task	Method	LINCS L1000			Tahoe		
		AUROC	Accuracy	F1	AUROC	Accuracy	F1
DE	Simple Voting	0.51±.00	0.54±.00	0.45±.00	0.46±.00	0.54±.00	0.26±.00
	GAT	0.51±.01	0.53±.02	0.50±.01	0.50±.02	0.57±.01	0.10±.01
	scVI	0.50±.01	0.53±.02	0.30±.00	**0.53±.02**	**0.60±.02**	0.38±.01
	ChemCPA	0.50±.00	0.55±.02	0.02±.00	0.50±.00	0.60±.03	0.03±.00
	State	0.50±.01	0.49±.02	**0.51±.03**	0.50±.02	0.60±.02	0.08±.01
	VCWorld	0.43±.00	0.40±.00	**0.51±.00**	0.50±.00	0.52±.00	**0.40±.00**
	ARCS	**0.52±.01**	**0.55±.02**	0.48±.01	**0.53±.02**	**0.60±.02**	0.35±.01
DIR	Simple Voting	0.47±.00	0.63±.00	0.42±.00	0.41±.00	0.57±.00	0.44±.00
	GAT	0.51±.01	0.66±.02	0.33±.01	0.50±.02	0.39±.03	0.09±.00
	scVI	0.51±.02	0.52±.01	0.45±.03	0.53±.01	0.49±.02	0.19±.01
	ChemCPA	0.50±.01	0.51±.01	0.01±.00	0.50±.00	0.41±.02	0.03±.00
	State	0.50±.01	0.50±.02	0.44±.00	0.50±.02	0.41±.01	0.06±.00
	VCWorld	0.47±.01	0.55±.01	0.35±.02	0.52±.01	0.52±.02	0.54±.01
	ARCS	**0.56±.01**	**0.70±.02**	**0.47±.02**	**0.56±.01**	**0.71±.03**	**0.64±.02**

priors provide more reliable signals than direct response generation. Moreover, VCWorld demonstrates limited effectiveness in cross-context settings, indicating that LLM-based reasoning alone does not guarantee robust transfer across cellular contexts. Additional detailed analysis regarding VCWorld's performance on CrossPerturb is presented in Appendix G.

Directional Change Task Performance. Overall, methods achieve stronger discrimination on DIR, with ARCS maintaining an AUROC of at least 0.63 across all eight settings and consistently outperforming every baseline. GAT ranks second on L1000, where the knowledge graph has broader coverage, but fails to distinguish regulatory direction on Tahoe, yielding an AUROC below 0.50 across all settings. The simple LLM-free voting strategy also consistently outperforms VCWorld, raising questions about whether existing LLM-based reasoning frameworks capture genuine biological signals across cellular contexts. These results further indicate that cross-context virtual cell modeling remains a critical open challenge for current LLMs.

5.4 Gene-level Performance

Gene-level evaluation complements population-level evaluation by assigning equal weight to each target gene. This design mitigates performance inflation caused by genes with strong response tendencies or imbalanced numbers of perturbation instances, providing a more rigorous assessment of per-gene predictive capability. Because the number of target genes varies across context settings, we compute overall gene-level metrics aggregated across all settings.

As shown in Table 2, simple voting exhibits a substantial drop from its population-level performance, suggesting that its strong population-level results are largely driven by target genes with inherent dominant response tendencies and highly uneven instance frequencies. Other baselines, such as GAT, ChemCPA, and State, also provide limited gene-level discrimination despite occasional advantages in macro accuracy, whereas VCWorld does not show a consistent or robust gene-level advantage across datasets or tasks. ARCS demonstrates a clear advantage on the DE task, achieving the highest macro AUROC and accuracy. On the DIR task, ARCS performs substantially better, reaching the highest macro AUROC, accuracy, and F1 score across both datasets. These results indicate that ARCS effectively integrates context-specific evidence beyond the inherent response tendencies exploited by simple voting, highlighting the true ability of an evidence-grounded LLM to accurately model context-dependent perturbation mechanisms.

5.5 Ablation Study

As illustrated in Figure 4, we systematically evaluate the contribution of the two LLM-guided components in ARCS across two tasks and two datasets: L1000-DE, L1000-DIR, Tahoe-DE, and Tahoe-DIR. We compare the full framework with three variants that remove the Analyze module, the Cri-



Figure 4: Ablation study across the DE and DIR tasks on L1000 and Tahoe. Dashed lines indicate the simple voting baseline, and outlined segments show performance drops relative to ARCS.

tique module, or both components. Simple voting strategy is included as a non-LLM reference. All results are measured using macro AUROC score, with dashed outlines clearly indicating the performance drop between each ablated variant and the full ARCS framework.

The full framework outperforms the simple voting baseline across all four settings, with the largest gains observed on the DIR task in Tahoe dataset. Removing either module generally causes performance degradation, although omitting the Critique component yields a slight improvement on L1000-DE, suggesting that when retrieved evidence credibility is inherently limited, strict filtering becomes less impactful. The Analyze module makes a particularly strong contribution to the DIR prediction task, while the Critique module also provides consistent gains in the Tahoe dataset. Removing both components produces the most severe degradation, reducing macro AUROC to 0.50, 0.46, 0.48, and 0.43 across the four settings, respectively. These results indicate that the Analyze and Critique components jointly support ARCS by prioritizing relevant evidence and filtering context-incompatible observations, enabling more reliable perturbation prediction than either partial configuration or the non-LLM reference.

5.6 Efficiency Analysis and Case Study

Efficiency Analysis. To assess practical inference efficiency, we compare ARCS with VCWorld using the same Qwen3-8B backbone. As shown in Table 12, although ARCS consumes 24.1% more input tokens than VCWorld, it generates 70.7% fewer output tokens. This reduction leads to substantially lower latency and inference cost, dropping the latency from 1.09 s to 0.31 s (detailed efficiency analysis are provided in Appendix H). These results demonstrate that ARCS provides a more favorable efficiency profile for repeated cross-context perturbation prediction.

Case Study. To illustrate the auditable reasoning process, we provide a representative case from the DIR task on the L1000 dataset in Appendix I. The log records the candidate evidence pool, observed outcomes, critique decisions, and the final score. This example highlights how ARCS produces an inspectable prediction from traceable evidence rather than merely outputting a final label.

5.7 Conclusion

In this work, we reveal that current language-based perturbation benchmarks suffer from a critical limitation: their confinement to within-context evaluations enables models to exploit gene-specific response tendencies and shortcut patterns rather than performing genuine context-sensitive biological reasoning. To bridge this gap, we introduced **CrossPerturb**, the first comprehensive benchmark designed to systematically evaluate virtual cell perturbation prediction across diverse cross-context settings, including unseen cell lines, drug dosage, and exposure durations. Complementing the benchmark, we proposed **ARCS**, an evidence-grounded framework that integrates heterogeneous biological evidence through principled mechanism analysis, cross-context retrieval, compatibility critique, and bounded evidence synthesis. Extensive experiments across the LINCS L1000 and Tahoe datasets demonstrate that ARCS consistently outperforms existing state-of-the-art baselines and LLM frameworks while offering superior inference efficiency. Furthermore, our analysis highlights that while existing LLM-based reasoning frameworks can generate plausible biological explanations, they often fail to achieve true generalization under distribution shifts. Our findings establish that cross-context generalization remains a vital open challenge for virtual cell modeling, and we believe CrossPerturb will serve as a reliable foundation for advancing robust, generalizable, and mechanism-grounded AI-driven biological discovery.

AI USE STATEMENT

Generative AI tools were used for language editing and polishing, and the illustration depicting the cellular response mechanism to drug treatment in the bottom-right corner of Figure 3 was generated using AI. The authors reviewed all AI-assisted revisions and take responsibility for the final content of the submission.

ETHICS STATEMENT

This work presents a benchmark and inference framework for drug perturbation prediction. All datasets used in this study are publicly available and contain no personally identifiable or sensitive information. Our work does not involve human subjects, animal experiments, or wet-lab procedures. The proposed framework is intended solely as a computational tool for generating hypotheses about drug-induced transcriptional responses. It is not designed to guide clinical decision-making, drug development, or any therapeutic intervention. Any biological insights derived from our framwork should be validated through independent experimental studies.

REFERENCES

Abhinav K Adduri, Dhruv Gautam, Beatrice Bevilacqua, Mohsen Naghipourfar, Alishba Imran, Rohan Shah, Noam Teyssier, Rishi Verma, Christopher Carpenter, Basak Eraslan, et al. Predicting cellular responses to perturbation across diverse contexts with state. *Cell*, 2025.

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

Martin Beck, Roberto Covino, Inga Hänelt, and Michaela Müller-McNicoll. Understanding the cell: Future views of structural biology. *Cell*, 187(3):545–562, 2024.

Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific data*, 10(1):67, 2023.

David Croft, Gavin O'kelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy, Phani Garapati, Gopal Gopinath, Bijay Jassal, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, 39(suppl_1):D691–D697, 2010.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 21(8):1470–1480, 2024.

Payam Dibaeinia, Sudarshan Babu, Mei Knudson, Ali ElSheikh, Yibo Wen, Han Liu, Jason Perera, and Aly A Khan. Virtual cells need context, not just scale. *bioRxiv*, 2026.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *cell*, 167(7): 1853–1866, 2016.

Madalina Giurgiu, Julian Reinhard, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and Andreas Ruepp. Corum: the comprehensive resource of mammalian protein complexes—2019. *Nucleic acids research*, 47(D1):D559–D563, 2019.

Heonjong Han, Jae-Won Cho, Sangyoung Lee, Ayoung Yun, Hyojin Kim, Dasom Bae, Sunmo Yang, Chan Yeong Kim, Muyoung Lee, Eunbeen Kim, et al. Trrust v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic acids research*, 46 (D1):D380–D386, 2018.

Leon Hetzel, Simon Boehm, Niki Kilbertus, Stephan Günnemann, Fabian Theis, et al. Predicting cellular responses to novel drug perturbations at a single-cell resolution. *Advances in Neural Information Processing Systems*, 35:26711–26722, 2022.

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese, et al. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems. *arXiv preprint arXiv:2504.09037*, 2025.

Dominik Klein, Jonas Simon Fleck, Daniil Bobrovskiy, Lea Zimmermann, Sören Becker, Alessandro Palma, Leander Dony, Alejandro Tejada-Lapuerta, Guillaume Huguet, Hsiu-Chuan Lin, et al. Cellflow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, pp. 2025–04, 2025.

Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun Chin, Seth A Strawbridge, et al. Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic acids research*, 52(D1):D1265–D1275, 2024.

Chen Li, Haoxiang Gao, Yuli She, Haiyang Bian, Qing Chen, Kai Liu, Lei Wei, and Xuegong Zhang. Benchmarking ai models for in silico gene perturbation of cells. *bioRxiv*, pp. 2024–12, 2024.

Lanxiang Li, Yue You, Yunlin Fu, Wenyu Liao, Xueying Fan, Shihong Lu, Ye Cao, Bo Li, Wenle Ren, Jiaming Kong, et al. A systematic comparison of single-cell perturbation response prediction models. *Science Advances*, 12(37):eaed3414, 2026.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

Christian von Mering, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. String: a database of predicted functional associations between proteins. *Nucleic acids research*, 31(1):258–261, 2003.

Emmanuel Noutahi, Jason Hartford, Prudencio Tossou, Shawn Whitfield, Alisandra K Denton, Cas Wognum, Kristina Ulicna, Michael Craig, Jonathan Hsu, Michael Cuccarese, et al. Virtual cells: Predict, explain, discover. *arXiv preprint arXiv:2505.14613*, 2025.

Kanglu Pei, Sisi Qu, Philip Torr, Jonathan G Hedley, and Christian Schroeder de Witt. Perturbagent: An agentic ai system for analysis and prediction of genetic perturbations. In *Second Workshop on XAI4Science: From Understanding Model Behavior to Discovering New Scientific Knowledge*, 2026.

Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 42(6):927–935, 2024.

Yusuf H Roohani, Tony J Hua, Po-Yuan Tung, Lexi R Bounds, Feiqiao B Yu, Alexander Dobin, Noam Teyssier, Abhinav Adduri, Alden Woodrow, Brian S Plosky, et al. Virtual cell challenge: Toward a turing test for the virtual cell. *Cell*, 188(13):3370–3374, 2025.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.

Eric W Sayers, Jeffrey Beck, Evan E Bolton, Devon Bourexis, James R Brister, Kathi Canese, Donald C Comeau, Kathryn Funk, Sunghwan Kim, William Klimke, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 49(D1):D10–D17, 2021.

Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.

Dénes Türei, Jonathan Schaul, Nicolàs Palacio-Escat, Balázs Bohár, Yunfan Bai, Francesco Ceccarelli, Elif Çevrim, Macabe Daley, Melih Darcan, Daniel Dimitrov, et al. Omnipath: integrated knowledgebase for multi-omics analysis. *Nucleic Acids Research*, 54(D1):D652–D660, 2026.

Zhijian Wei, Runze Ma, Zichen Wang, Zhongmin Li, Shuotong Song, and Shuangjia Zheng. Vcworld: A biological world model for virtual cell simulation. In *International Conference on Learning Representations*, volume 2026, pp. 72795–72823, 2026.

Menghua Rachel Wu, Russell Littman, Jacob Levine, Lin Qiu, Tommaso Biancalani, David Richmond, and Jan-Christian Huetter. Contextualizing biological perturbation experiments through language. In *International Conference on Learning Representations*, volume 2025, pp. 73598–73624, 2025.

Weize Xu, Erwin Poussi, Quan Zhong, Zehua Zeng, Christopher Zou, Xuehai Wang, Yifan Lu, Miao Cui, Daiji Okamura, Cinlong Huang, et al. Pantheonos: An evolvable multi-agent framework for automatic genomics discovery. *bioRxiv*, pp. 2026–02, 2026.

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

Xinyu Yuan, Xixian Liu, Jianan Zhao, Yashi Zhang, Hongyu Guo, and Jian Tang. Plausibility is not prediction: Contrastive evidence for llm-based cellular perturbation reasoning. *arXiv preprint arXiv:2606.01042*, 2026.

Jesse Zhang, Airol A Ubas, Richard De Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G Jones, et al. Tahoe-100m: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *BioRxiv*, pp. 2025–02, 2025.

Bingxi Zhao, Lin Geng Foo, Ping Hu, Christian Theobalt, Hossein Rahmani, and Jun Liu. Llm-based agentic reasoning frameworks: A survey from methods to scenarios. *arXiv preprint arXiv:2508.17692*, 2025.

Appendix for CrossPerturb

A Detailed Dataset Curation

Gene Processing. For the L1000 dataset, the original 978-gene landmark panel was aligned with available source observations. Five landmark symbols lacking usable source records were excluded, yielding a final set of 973 genes. For Tahoe, pseudogenes and non-coding genes were filtered out prior to constructing the variable-gene panel. Gene variability was quantified from control pseudobulk profiles using the Seurat dispersion estimation procedure, following counts-per-million (CPM) normalization and log-transformation. The 2,000 genes exhibiting the highest dispersion scores were retained, ensuring that gene selection remained strictly decoupled from drug-treated profiles. Finally, the Tahoe panel was restricted to genes possessing valid differential expression statistics across all selected cell lines.

Cell-Line Selection. For L1000, candidate cell lines were systematically ranked by their cumulative compound coverage across Phase I and Phase II experiments. We carefully selected 11 cell lines that jointly maximize chemical space representation while enabling robust context-specific evaluation. For Tahoe, 12 out of 48 cell lines with available pseudobulk differential expression profiles were chosen based on a rigorous multi-criteria filtering, including drug-condition coverage, response yield, overlap with L1000, and tissue-origin diversity. This filtered Tahoe collection comprises 9 source cell lines alongside 3 distinct cell lines strictly reserved for unseen-cell generalization evaluations. Comprehensive and transparent details regarding cell-line annotations, tissue origins, coverage metrics, and data split assignments are provided in Table 3.

Label Construction. For each drug-context-gene triplet $(p, \mathbf{s}, g)$, binary ground-truth labels were derived from pseudobulk treatment-versus-control contrasts. Differential expression statistics were computed across biological replicates via moderated t-statistics implemented in the *limma* package. A gene was designated as differentially expressed if its Benjamini-Hochberg adjusted p-value satisfied $q \leq 0.05$ and its absolute log-fold change met $|\log_2 \mathrm{FC}| \geq 0.36$. Experimental conditions yielding fewer than 15 differentially expressed genes were excluded to ensure statistical reliability. For the directional change (DIR) task, we retained up to 20 genes exhibiting the highest positive fold changes and up to 20 with the most negative fold changes. For the differential expression (DE) task, we additionally sampled up to 40 non-responsive background genes from the identical cellular context. In shifted-context scenarios, up to 20 hard negatives were incorporated when a gene responded to a given drug in an alternate cellular context yet remained non-responsive in the target context.

Drug and Context Splits. Compounds were partitioned at the drug level into disjoint source and target sets using a 30:70 ratio. The L1000 split contains 353 source compounds and 825 target compounds, while the Tahoe split contains 52 source compounds and 116 target compounds. Drug identities were normalized before partitioning, and no compound appears in both sets.

For L1000, target queries are divided into four regimes according to the relationship between their cellular contexts and the source observations: L1 contains observed contexts, L2 contains an unseen cell line, L3 contains an unseen exposure duration, and L4 contains an unseen drug dosage. The L3 regime evaluates a 6 h exposure duration, which is absent from the source observations, while source contexts primarily use a 24 h exposure duration. For Tahoe, the corresponding regimes are T1 for observed contexts, T2 for unseen cell lines, T3 for an unseen drug dosage of 0.05 μM, and T4 for the joint combination of an unseen cell line and an unseen drug dosage. In all regimes, target drugs are absent from the source compound set. The numbers of queries for the differential-expression (DE) and direction-of-change (DIR) tasks are summarized in Table 4.

B BioRelKG Construction

BioRelKG is constructed as a highly structured and heterogeneous knowledge graph representing biomedical entities and their typed relationships. Entities encompass compounds, genes and proteins, functional terms, pathways, protein complexes, diseases, phenotypes, target roles, anatomies,

Table 3: Cell-line composition and experimental coverage in CrossPerturb. Statistics are computed over the complete benchmark, combining training and test task files. Compounds denote the number of distinct drugs measured in each cell line. Drug-dose and exposure-time report the numbers of standardized drug-dose and exposure-duration levels represented for each cell line.

Dataset	Cell line	Compounds	Drug-dose	Exposure-time
L1000	A549	20	6	1
	MCF7	182	6	1
	PC3	190	6	1
	A375	231	7	2
	HT29	169	6	1
	VCaP	619	6	2
	HA1E	217	6	1
	HeLa	122	5	1
	FIBRNPC	11	1	1
	Primary human hepatocytes	42	1	1
	SKB	31	1	1
Tahoe	A549	50	3	1
	A-172	140	3	1
	HEC-1-A	110	3	1
	HT-29	44	3	1
	HS-578T	101	3	1
	J82	83	3	1
	LoVo	53	3	1
	PANC-1	63	3	1
	RKO	40	3	1
	SW480	64	3	1
	A498	88	3	1
	HOP62	88	3	1

Table 4: Number of differential-expression (DE) and direction-of-change (DIR) queries in each cross-context evaluation regime. All target drugs are unseen during source observation.

Dataset	Regime	Context shift	DE queries	DIR queries
L1000	L1	Observed context	25,603	12,240
	L2	Unseen cell line	13,165	6,356
	L3	Unseen exposure duration (6 h)	9,354	4,380
	L4	Unseen drug dosage	14,589	6,911
Tahoe	T1	Observed context	25,770	11,610
	T2	Unseen cell line	24,912	11,272
	T3	Unseen drug dosage (0.05 μM)	12,429	5,389
	T4	Unseen cell line and drug dosage	4,165	1,885

and drug classes. The graph incorporates pharmacological, molecular interaction, regulatory, signaling, pathway-based, functional, and complex-level relations, with transparent and reliable source provenance preserved for every edge to ensure full traceability.

The current release of BioRelKG comprises 289,067 nodes and 7,078,497 edges across 10 node types and 14 relation types. Each edge is associated with 22 attributes, among which the type triple, predicate, sign, direction flag, mechanism slot, evidence tier, and provenance list are utilized by downstream modules (Table 6). Edges are carefully aggregated from five primary database categories: pharmacology and drug-target interactions (DrugBank, ChEMBL, DGIdb, STITCH, DrugCentral, GtoPdb), molecular interactions and complexes (STRING, BioPlex, CORUM), regulation and signaling (TRRUST, OmniPath, SIGNOR, DoRothEA), pathway and functional annotations (Gene Ontology, Reactome, KEGG), and integrated biomedical associations (PrimeKG, CTD, Hetionet, BioKG). Among all edges, 1,922,642 (27.2%) contain an explicit directional sign, while the remaining 5,155,855 record standard undirected associations.

Table 5: Biomedical resources and source families integrated into BioRelKG. The table summarizes the main provenance categories represented in the final graph.

Resource	Main content	Relations incorporated into BioRelKG
DrugBank	Drug annotations and pharmacological information	Drug-target associations and pharmacological target relations
ChEMBL	Bioactivity and mechanism-of-action annotations	Drug-target, binding, and mechanism-related relations
STITCH	Chemical-protein interaction evidence	Chemical-protein interaction and target relations
DrugCentral	Drug-target and pharmacological annotations	Drug-target and pharmacological relations
STRING	Protein-protein interaction and functional association evidence	Protein-protein interaction relations
BioPlex	Proteomic protein-protein interaction evidence	Protein-complex and protein-protein interaction relations
TRRUST	Transcription factor regulatory interactions	Transcription factor-gene regulatory relations
SIGNOR	Causal signaling relationships	Signed regulatory and signaling relations
OmniPath	Signaling and regulatory network resources	Signaling, enzyme-substrate, and regulatory relations
DoRothEA	Transcription factor activity and regulons	Transcription factor-gene regulatory relations
Gene Ontology	Molecular function, biological process, and cellular component annotations	Functional annotations and ontology relations
Reactome	Curated biological reactions and pathways	Reaction, pathway, and pathway-membership relations
KEGG	Metabolic and signaling pathways	Pathway membership and pathway relations
CORUM	Mammalian protein complexes	Protein-complex membership relations
CTD	Chemical, gene, disease, and phenotype associations	Chemical-gene, chemical-phenotype, and disease-related associations
Hetionet	Integrated biomedical entity associations	Cross-domain drug, gene, disease, and phenotype relations
BioKG	Integrated biomedical associations	Cross-domain biomedical relations
PrimeKG	Integrated drug, gene, protein, disease, pathway, and phenotype associations	Drug effects, indications, contraindications, protein interactions, pathway annotations, and functional relations

Semantics of Edge Signs. A sign value of 0 indicates that the source database does not specify a regulatory direction, rather than denoting a neutral or bidirectional effect. This distinction is critical: the directional prediction task exclusively permits edges that are both signed and directed as supportive evidence, whereas unsigned edges serve solely as mechanism vicinity indicators. This prevents unannotated structural relations from being misconstrued as active regulatory increases.

Provenance and Multi-Source Integration. Entity identifiers are normalized within their respective namespaces prior to graph integration. Duplicate relations are merged while retaining complete source provenance. Relations connecting identical entity pairs with disparate source records are consolidated into a single edge accompanied by a multi-source provenance list (comprising 79,098 multi-source edges, or 1.1% of the build). This design avoids conflating database mirroring frequency with independent biological mechanisms. In cases where source databases report conflicting signs for the same relation, both records are preserved and the edge is explicitly flagged as conflicting, avoiding heuristic voting mechanisms.

Identifier Normalization. Normalization is executed per entity type to prevent artifactual collapse between case-sensitive symbols. Compound names undergo multi-step cleaning to match binding tables: for instance, exact lowercase lookups resolve 50.0% of compounds on the Tahoe set, which increases to 68.1% upon stripping trailing parentheticals, and reaches 71.6% after handling

Table 6: Node and edge statistics of BioRelKG. The graph contains 289,067 typed nodes and 7,078,497 typed edges.

Category	Type	Count
Nodes	Chemical and drug entities	99,328
Nodes	Gene and protein entities	89,017
Nodes	Gene Ontology terms	38,351
Nodes	Reaction and pathway entities	21,177
Nodes	Protein complexes	19,000
Nodes	Diseases	13,637
Nodes	Phenotypes	7,106
Nodes	Target-role entities	706
Nodes	Anatomical entities	400
Nodes	Drug classes	345
Edges	Interacts with	3,512,745
Edges	Activates	1,030,059
Edges	Inhibits	869,775
Edges	Member of	834,124
Edges	Modifies	228,545
Edges	Binds	188,509
Edges	Has effect on	133,272
Edges	Associated with	115,247
Edges	Indicated for	66,638
Edges	Regulates	34,271
Edges	Contraindicated for	31,115
Edges	Targets	30,454
Edges	Off-label use for	2,678
Edges	Has target role	1,065

salt and hydrate forms using an explicit whitelist (e.g., distinguishing `sodium butyrate` from `butyrate`). This deterministic key mapping enables seamless joining between query and database tables without heuristic pairwise matching.

B.1 Edge Representation

To ensure deterministic and uniform input representations across all inference modules, graph relations are converted into natural language sentences using fixed templates rather than raw triples. Each relation type corresponds to a strict template mapping defined by its predicate, mechanism slot, sign, and evidence tier (Table 7). Three core design principles govern the template architecture:

First, **directionality is explicitly tied to the verb and restricted to annotated sources**. Directional templates are triggered only when the relation supplies both polarity and a mechanism slot (e.g., expression, activity, abundance, or complex-level action). For instance, an `activates` edge mapped to an expression mechanism is rendered as "X increases expression of Y", whereas the same predicate mapped to an activity mechanism is rendered as "X increases activity of Y".

Second, **undirected relations incorporate explicit hedges**. Binding, targeting, modification, physical interaction, and pathway membership are rendered with explicit qualifying phrases (e.g., "no direction-of-effect annotated" or "undirected") embedded directly into the template, preventing downstream modules from misinterpreting structural associations as regulatory actions.

Third, **predicted relations carry stronger epistemic hedging than curated ones**. Text-mined or transferred associations are rendered conditionally (e.g., "X may physically interact with Y (predicted association; undirected)"), enabling models to distinguish empirical assertions from computational hypotheses at inference time.

Relations lacking a dedicated template fall back to a generic descriptive form rather than being omitted or misassigned. Furthermore, each rendered edge is appended with a four-part ledger detailing its relation type, sign/direction, evidence tier, and scope statements (specifying what the relation supports versus what it does not establish). This serves as an explicit safeguard against

Table 7: Template rendered into the prompt for each relation type. "Dir." marks templates that assert a direction; the remaining templates carry an explicit hedge. X is the subject and Y the object of the edge; the mechanism slot selects the verb within a predicate.

Relation (mechanism)	**Dir.**	**Template**
`activates`, `expression`	yes	"X increases expression of Y"
`inhibits`, `expression`	yes	"X decreases expression of Y"
`activates`, `activity`	yes	"X increases activity of Y"
`inhibits`, `activity`	yes	"X decreases activity of Y"
`activates`, `quantity`	yes	"X up-regulates Y" (weakest signed form)
`inhibits`, `quantity`	yes	"X down-regulates Y" (weakest signed form)
`activates`, `complex_formation`	yes	"X promotes formation of the protein complex Y"
`inhibits`, `complex_formation`	yes	"X blocks formation of the protein complex Y"
`activates`, `complex_activity`	yes	"X increases the activity of the protein complex Y"
`inhibits`, `complex_activity`	yes	"X decreases the activity of the protein complex Y"
`modifies`	no	"X modifies Y (no direction annotated)"
`binds`	no	"X binds Y (no direction-of-effect annotated)"
`targets`	no	"X targets Y ($\langle$role$\rangle$) (no direction-of-effect annotated)"
`interacts_with`, curated	no	"X and Y physically interact (undirected, no regulatory direction)"
`interacts_with`, predicted	no	"X may physically interact with Y (predicted association; undirected)"
`member_of`, pathway	no	"X is a member of the Y pathway"
`member_of`, biological process	no	"X participates in the biological process Y"
`member_of`, molecular function	no	"X has the molecular function Y"
`member_of`, complex	no	"X is a subunit of the protein complex Y (purified from $\langle$cell line$\rangle$ cells)"
no dedicated template	no	generic form naming the relation between X and Y; never approximated by another template

over-interpretation; for example, a `binds` edge supports physical association but explicitly disclaims regulatory direction or cell-line-specific differential expression.

B.2 KG QUERY

For each individual CrossPerturb query, BioRelKG retrieves relevant heterogeneous evidence associated with the target drug and response gene, yielding direct associations, local relational neighborhoods, or multi-hop paths depending on graph topology. Cellular contexts are supplied by the query specifications, while BioRelKG provides context-independent and static biological relations. Target-context expression measurements and ground-truth evaluation labels are strictly excluded from both the retrieved graph evidence and source observations.

Graph coverage naturally varies across query instances. Direct edges between queried drugs and genes are relatively sparse, occurring in 0.12% of L1000-DE, 0.15% of L1000-DIR, 0.05% of Tahoe-DE, and 0.08% of Tahoe-DIR test rows. Consequently, retrieval predominantly relies on extended local neighborhoods and multi-hop paths, synthesizing up to a maximum of six named targets and three routed paths per query. When no pathway is retrieved, the module explicitly reports a knowledge base coverage gap rather than assuming a negative biological effect, preventing data sparsity from being misconstrued as empirical non-responsiveness.

Role Assignment. Prior to the retrieval and prompting phases, each edge is clearly assigned to one of seven functional roles: pharmacological target, target activity, gene regulator, shared biological process, observed response, ADME relation, or association only. Among edges connecting compounds and genes, 2,170,808 (86.6%) represent literature-mined expression or activity changes, whereas only 42,749 (1.7%) denote classical pharmacological actions, with the two sets being kept strictly disjoint. This role separation ensures that text-mined expression correlations are not conflated with underlying upstream pharmacological mechanisms of action.

B.3 Leakage Control

To effectively prevent information leakage from measured transcriptional responses into the retrieval and reasoning pipeline, these two rigorous safeguards are enforced. First, external database sources are restricted strictly to mechanism and interaction repositories that exclude perturbation screen results. Second, all loaded knowledge is filtered against an allowlist of trusted source families and a strict forbidden-token pattern specifically covering experimental outcome vocabularies, with any violation triggering an immediate system compilation error.

These controls are validated through structural audits comparing the graph against benchmark labels. If response labels leaked into the graph, queried drug–gene pairs would exhibit dense, highly aligned direct edges. However, direct edges appear in fewer than 0.2% of evaluation rows, and among the 48 signed direct edges present, their signs align with gold labels in 29 cases ($\sim$60%), matching expected biological baseline rates rather than the near-perfect alignment indicative of label contamination.

C Baseline Details

We compare ARCS against a diverse suite of baselines, spanning statistical, graph-based, deep generative, and large language model (LLM) paradigms. To ensure a rigorous and fair comparison, all methods operate under identical compound-disjoint source and target partitions and are evaluated on the same Differential Expression (DE) and Direction of Change (DIR) queries.

- **Simple Voting.** This non-parametric baseline retrieves historical source observations for the target response gene, ranks them by cellular-context similarity, and assigns the majority label among the top five observations with available outcomes. Simple Voting strategy relies purely on empirical data heuristics without parameterized learning or explicit biomedical knowledge graphs.
- **GAT.** The graph attention network (GAT) baseline models drugs, genes, and intermediate biomedical entities as nodes within BioRelKG, utilizing typed biological relations to establish graph connectivity. It aggregates neighborhood topologies to predict binary outcomes for each perturbation–context–gene query.
- **scVI.** We implement an scVI-style conditional variational autoencoder (CVAE) that learns probabilistic latent representations of basal and perturbed cellular states. The decoded perturbation-induced expression shifts are subsequently used to derive the specific response of each queried gene (Lopez et al., 2018).
- **ChemCPA.** ChemCPA effectively disentangles basal cellular states, chemical perturbations, and covariate factors within a unified shared latent space. It flexibly composes drug representations and dose-dependent effects with cellular states to accurately forecast post-perturbation expression profiles (Hetzel et al., 2022).
- **State.** State is a pretrained perturbation-response foundation model that predicts post-perturbation cellular states based on representations learned from large-scale single-cell repositories (Adduri et al., 2025).
- **VCWorld.** VCWorld retrieves labeled perturbation examples and integrates them with biological context via LLM-based reasoning to predict gene-level perturbation outcomes (Wei et al., 2026). To ensure a strictly controlled comparison, VCWorld and ARCS utilize the identical Qwen3-8B backbone and decoding configuration.

For methods that generate continuous expression profiles (scVI, ChemCPA, and State), we systematically and carefully translate their predictions into binary DE and DIR labels using the same standardized differential expression criteria applied during the construction of CrossPerturb. Methods explicitly designed for direct binary classification are evaluated using their native decision outputs. Crucially, no baseline leverages target-query labels during training, retrieval, or model selection, thereby ensuring an unbiased and rigorous comparative framework.

D EVALUATION METRICS

To rigorously assess model performance across multiple complementary dimensions, we report Accuracy, AUROC, and F1 scores for both differential expression (DE) and directional (DIR) prediction tasks at both population and gene levels. Let $\mathcal{Q} = \{q_i\}_{i=1}^{N}$ denote the set of evaluation queries, with ground-truth labels $y_i \in \{0, 1\}$ and binary predictions $\hat{y}_i \in \{0, 1\}$. For methods predicting continuous expression profiles, predictions are mapped to binary decisions using the standardized thresholding procedure described in the main text. Abstentions or invalid outputs are carefully and systematically treated as errors in all primary evaluations. Note that the reported AUROC is computed directly from these binary decisions, rendering it numerically equivalent to balanced accuracy and ensuring a robust, fair comparison across both direct classification and profile-generating frameworks under a unified output representation.

Population-level evaluation. At the population level, every individual perturbation–context–gene query contributes equally to the global evaluation pool, capturing the overall macro-performance across the entire test distribution. Accuracy and F1 scores are aggregated across all instances within a given context setting:

$$\mathrm{Acc}_{\mathrm{pop}} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i), \qquad \mathrm{F1}_{\mathrm{pop}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{9}$$

The population-level AUROC is formulated as:

$$\mathrm{AUROC}_{\mathrm{pop}} = Pr(\hat{y}^+ > \hat{y}^-) + \frac{1}{2}Pr(\hat{y}^+ = \hat{y}^-), \tag{10}$$

where $\hat{y}^+$ and $\hat{y}^-$ represent predictions for randomly selected positive and negative instances, respectively. Population-level metrics are reported separately for L1–L4 and T1–T4 settings to reflect macro-averaged efficacy over the full instance population. However, because these metrics pool all queries together, genes associated with a higher frequency of target queries exert a disproportionately heavy influence on the final aggregate score.

Gene-level evaluation. To mitigate potential evaluation biases arising from unequal target-gene frequencies and inherent gene-specific response propensities, we introduce a complementary gene-level evaluation protocol. Specifically, we combine the four context settings within each dataset and compute each metric independently for every individual response gene. Let $\mathcal{Q}_g \subseteq \mathcal{Q}$ denote the subset of queries corresponding to gene g, and let M_g represent the Accuracy, F1, or AUROC computed exclusively on $\mathcal{Q}_g$. The macro-averaged gene-level metric is defined as:

$$M_{\mathrm{gene}} = \frac{1}{|\mathcal{G}_M|}\sum_{g \in \mathcal{G}_M} M_g, \tag{11}$$

For Accuracy and F1, the gene set $\mathcal{G}_M$ encompasses all genes possessing at least one valid query, with any undefined F1 denominators conservatively mapped to zero. For AUROC, $\mathcal{G}_M$ is restricted to genes containing at least one positive and one negative instance, ensuring that ranking metrics remain mathematically well-defined. This micro-to-macro aggregation assigns equal weight to every eligible gene regardless of its query frequency, preventing heavily-represented genes from overshadowing rare or tightly regulated targets. Under this unified framework, Accuracy measures overall classification correctness, F1 evaluates the precision-recall trade-off for the positive class under potential class imbalance, and gene-wise AUROC captures intrinsic class discrimination capacity within individual genetic contexts.

E EXPERIMENT AND HYPERPARAMETER SETTINGS

Experimental Setup. Both ARCS and VCWorld utilize Qwen3-8B (Yang et al., 2025) as the core reasoning backbone, while Qwen3-72B is deployed offline to synthesize comprehensive drug and gene descriptions that capture higher-order semantic relationships. All baseline evaluations are averaged across three independent runs. For LLM text generation, the sampling temperature and top-p parameter are set to 0.6 and 0.9, respectively, balancing generation diversity with structural stability. Comprehensive system prompts for description generation, mechanism analysis, and candidate critique are detailed in Appendix J.

Table 8: Sensitivity of ARCS to the retrieval weighting parameter α and the evidence fusion parameter λ. Results represent macro-averaged gene-level AUROC across the four context settings. Bold values indicate the default configurations utilized in the main experiments.

Parameter	Value	L1000–DE	L1000–DIR	Tahoe–DE	Tahoe–DIR
α	0.00	0.36	0.41	0.45	0.44
	0.25	0.50	0.49	0.50	0.51
	0.50	**0.52**	0.54	**0.53**	0.54
	0.75	**0.52**	**0.56**	**0.53**	**0.56**
	1.00	0.51	0.53	0.51	0.52
λ	0.00	0.51	0.47	0.46	0.41
	0.25	0.52	**0.56**	**0.53**	**0.56**
	0.50	**0.53**	0.55	**0.53**	0.54
	0.75	0.52	0.55	**0.53**	0.53
	1.00	0.51	0.54	0.52	0.54

Retrieval Configuration. The retrieval scoring function integrates molecular and cellular context similarities via a weighting parameter $\alpha = 0.75$, prioritizing the core drug–gene relationship while incorporating context compatibility into the ranking metric. Given a target context $\mathbf{s} = (c, d, t)$ and a candidate context $\mathbf{s}_i = (c_i, d_i, t_i)$, the context similarity is defined as:

$$\mathrm{Sim}_{\mathrm{con}}(\mathbf{s}, \mathbf{s}_i) = w_c\, u_{\mathrm{cell}}(c, c_i) + w_d\, \phi\left(\left|\log \frac{d}{d_i}\right|; \tau_d, \sigma_d\right) + w_t\, \phi\left(\left|\log \frac{t}{t_i}\right|; \tau_t, \sigma_t\right), \tag{12}$$

subject to the constraint

$$w_c + w_d + w_t = 1, \tag{13}$$

where $u_{\mathrm{cell}}(c, c_i) \in [0, 1]$ denotes the LLM-derived semantic similarity between cell-line descriptions. Dosage and exposure-duration deviations are evaluated in log-space and mapped through a piecewise compatibility kernel:

$$\phi(\Delta; \tau, \sigma) = \begin{cases} 1, & 0 \le \Delta \le \tau, \\ \exp\left[-\dfrac{(\Delta - \tau)^2}{2\sigma^2}\right], & \Delta > \tau, \end{cases} \tag{14}$$

which preserves a plateau of high compatibility within a defined tolerance region and smoothly downweights larger contextual discrepancies. All weights and decay parameters are held fixed across experiments. For each knowledge-graph-guided matching criterion, ARCS retrieves at most $k = 5$ candidate examples uniformly across datasets, tasks, and context regimes.

Evidence Fusion. The aggregated evidence score is combined with the context-specific response prior using an interpolation weight of $\lambda = 0.25$. The response prior is estimated empirically from historical observations of the target gene under the query context; if an exact contextual match is unavailable, the nearest observed context is selected based on cell-line, dosage, and exposure-duration proximity. The resulting continuous score is mapped to a binary prediction using a standard decision threshold of 0.5.

Hyperparameter Sensitivity Analysis. To evaluate the robustness of ARCS, we perform sensitivity analysis on the retrieval weighting parameter α and the evidence fusion parameter λ. Table 8 reports the macro gene-level AUROC across each dataset and task.

The empirical results reveal distinct operational regimes for both parameters. For the retrieval weight α, the extreme cases represent severely isolated and incomplete information sources. Setting $\alpha = 0.0$ corresponds to pure context matching, which relies exclusively on cell-line, dose, and time similarities while completely discarding underlying drug-gene relational associations. Conversely, setting $\alpha = 1.0$ represents pure molecular affinity, which evaluates drug-gene pairs based solely on knowledge graph topology while entirely ignoring essential cellular context and experimental conditions. Both extremes lead to substantial performance degradation, confirming that effective retrieval requires a harmonized and synchronized integration of drug-gene mechanisms and operational environments, with $\alpha = 0.75$ achieving the optimal balance.

Similarly, tuning the evidence fusion parameter λ exposes the interplay between external knowledge and empirical data. Setting $\lambda = 0.0$ relies entirely on the empirical response prior derived from historical observations while completely discarding the structured relational evidence retrieved from BioRelKG. On the other end, setting $\lambda = 1.0$ depends exclusively on the graph-derived reasoning score while ignoring historical baselines. The optimal configuration of $\lambda = 0.25$ indicates that ARCS performs best when anchored primarily by structured relational reasoning, with empirical priors serving as a calibrated stabilizer. These trends confirm that ARCS exhibits stable performance across a reasonable neighborhood around the selected configurations.

F Additional Population-Level Results

Tables 9 and 10 present the population-level Accuracy and F1 scores across the four target-context settings, complementing the AUROC performance reported in Table 1. While Accuracy measures overall classification correctness, it can be skewed by class imbalance. Conversely, F1 evaluates positive-class recovery by balancing precision and recall, ensuring that each instance involving a perturbation, context, and gene is weighted equally.

For DE prediction, performance varies without a single method dominating all settings. Although ARCS achieves competitive Accuracy on both datasets, particularly in subsets L1, L3, T1, and T3, this advantage does not strictly translate into superior F1 scores. This is especially evident on Tahoe, where scVI exhibits higher reliability in recovering positive DE instances. This divergence highlights that DE prediction remains highly sensitive to class imbalance and the selection of decision thresholds for identifying responsive genes.

In contrast, the advantages of ARCS are significantly more pronounced for DIR prediction. ARCS achieves top Accuracy across all eight settings and secures the best or tied-best F1 score in seven of them, with L3 as the sole exception. These gains are particularly striking on Tahoe, where baseline methods such as GAT, scVI, ChemCPA, and State experience sharp declines in F1. While Simple Voting remains a competitive baseline, ARCS demonstrates superior robustness in directional forecasting under cross-context shifts in cell lines, dosages, and exposure durations.

Analyzing Accuracy alongside F1 exposes critical model behaviors that single-metric evaluations might obscure. For instance, ChemCPA and State frequently attain moderate Accuracy paired with suppressed F1 scores, indicating a tendency to collapse toward the dominant class under discrete evaluation. ARCS strikes a robust balance between global correctness and positive-class recovery, especially for DIR. Furthermore, its nuanced DE performance underscores an intrinsic biological challenge: detecting whether a transcriptional response occurs is fundamentally more difficult than determining its direction once context-relevant evidence is available.

G Additional Analysis of VCWorld on CrossPerturb

To gain deeper insight into VCWorld's performance on CrossPerturb, we examine its decision coverage and output distributions across the complete evaluation sets. Table 11 disaggregates valid binary decisions from explicit abstentions and parsing failures, alongside gold positive rates to reveal systematic prediction biases.

As shown, unanswered cases stem almost exclusively from explicit abstentions rather than structural parsing failures, with parsing errors accounting for fewer than 1% of queries across all sets. This indicates that VCWorld successfully adheres to the required output format, and its limited coverage is a direct consequence of the framework's internal decision policy under ambiguous or sparse evidence, rather than a failure of response decoding.

However, coverage varies dramatically across datasets. While VCWorld leaves 14.1% of L1000-DE and 7.8% of L1000-DIR queries unanswered, these abstention rates surge to 41.1% and 39.4% on Tahoe. This sharp increase highlights that unconstrained LLM reasoning struggles to synthesize conclusive decisions when transferred to more complex perturbation spaces. Although this diagnostic does not isolate whether the deficit stems from knowledge gaps, evidence retrieval sparsity, or contextual divergence, it confirms that massive coverage collapse is a primary driver of VCWorld's performance degradation on Tahoe.

Table 9: Population-level Accuracy for Differential Expression (DE) and Direction of Change (DIR) prediction on CrossPerturb. L1–L4 and T1–T4 denote the four target-context settings for LINCS L1000 and Tahoe, respectively. Each evaluated instance is weighted equally; bold indicates the best performance and higher is better.

Task	Method	LINCS L1000				Tahoe			
		L1	L2	L3	L4	T1	T2	T3	T4
DE	Simple Voting	**0.59**	0.49	0.52	0.46	0.56	0.51	0.53	0.54
	GAT	0.56	**0.51**	0.50	0.50	0.56	0.52	0.54	0.54
	scVI	0.52	**0.51**	0.48	0.52	0.61	**0.60**	0.60	**0.60**
	ChemCPA	0.52	0.52	**0.53**	0.53	0.55	0.55	0.57	0.55
	State	0.50	0.50	0.51	0.50	0.55	0.55	0.57	0.55
	VCWorld	0.43	0.43	0.41	0.42	0.52	0.51	0.52	0.52
	ARCS	**0.59**	0.50	**0.53**	**0.52**	**0.64**	0.54	**0.62**	0.58
DIR	Simple Voting	0.75	0.61	0.63	0.65	0.70	0.60	0.68	0.59
	GAT	0.68	0.52	0.55	0.55	0.36	0.39	0.35	0.39
	scVI	0.52	0.52	0.54	0.54	0.50	0.54	0.48	0.57
	ChemCPA	0.57	0.48	0.52	0.53	0.45	0.45	0.41	0.47
	State	0.52	0.48	0.52	0.48	0.46	0.46	0.41	0.48
	VCWorld	0.66	0.48	0.57	0.49	0.52	0.52	0.54	0.50
	ARCS	**0.80**	**0.63**	**0.65**	**0.73**	**0.82**	**0.68**	**0.79**	**0.64**

Table 10: Population-level F1 score for Differential Expression (DE) and Direction of Change (DIR) prediction on CrossPerturb. L1–L4 and T1–T4 denote the four target-context settings for LINCS L1000 and Tahoe, respectively. Each evaluated instance is weighted equally; bold indicates the best performance and higher is better.

Task	Method	LINCS L1000				Tahoe			
		L1	L2	L3	L4	T1	T2	T3	T4
DE	Simple Voting	0.61	0.58	0.58	0.47	0.50	0.42	0.44	0.39
	GAT	**0.65**	**0.60**	0.60	**0.58**	0.32	0.27	0.29	0.28
	scVI	0.29	0.15	**0.61**	0.25	**0.59**	**0.55**	**0.56**	**0.47**
	ChemCPA	0.32	0.21	0.22	0.22	0.12	0.08	0.12	0.07
	State	0.56	0.57	0.55	0.55	0.25	0.35	0.24	0.34
	VCWorld	0.57	0.56	0.55	0.53	0.48	0.48	0.45	0.45
	ARCS	0.63	0.54	0.58	0.54	**0.59**	0.43	0.52	0.36
DIR	Simple Voting	0.70	0.63	0.59	0.63	0.72	0.63	0.72	0.62
	GAT	0.72	0.59	**0.61**	0.64	0.17	0.17	0.19	0.17
	scVI	0.47	0.54	0.52	0.53	0.20	0.31	0.26	0.36
	ChemCPA	0.24	0.22	0.22	0.14	0.36	0.21	0.23	0.18
	State	0.47	0.51	0.50	0.48	0.36	0.27	0.34	0.36
	VCWorld	0.55	0.42	0.49	0.34	0.57	0.56	0.61	0.55
	ARCS	**0.75**	**0.63**	0.58	**0.73**	**0.84**	**0.73**	**0.82**	**0.70**

Beyond coverage deficits, answered decisions exhibit pronounced class-specific biases. Among valid DE predictions, VCWorld predicts the positive class for 86.6% of L1000 cases and 79.0% of Tahoe cases, heavily overshooting the true gold positive rates of 47.66% and 44.82%. A reverse pattern emerges for DIR, where downregulation dominates 68.0% of L1000 and 71.7% of Tahoe predictions, despite true rates of 47.03% and 55.41%. Consequently, VCWorld's performance bottleneck cannot be attributed to abstentions alone; its active prediction subset suffers from severe distribution skew.

Table 11: VCWorld decision coverage and output distribution on CrossPerturb. Unanswered queries comprise explicit abstentions and parsing failures. The gold positive rate denotes the proportion of positive labels in each evaluation set; the positive class for DIR is Increase.

Dataset–Task	n	Positive decision	Negative decision	Abstain	Parse failure	Unanswered	Gold positive
L1000–DE	62,711	46,633 (DE)	7,217 (No-DE)	8,471	390	8,861 (14.1%)	47.66%
L1000–DIR	29,887	8,814 (Increase)	18,728 (Decrease)	2,288	57	2,345 (7.8%)	47.03%
Tahoe–DE	67,276	31,308 (DE)	8,315 (No-DE)	27,580	73	27,653 (41.1%)	44.82%
Tahoe–DIR	30,156	5,176 (Increase)	13,084 (Decrease)	11,889	7	11,896 (39.4%)	55.41%

Crucially, this affirmative bias persists even when the evidence retrieval strategy strictly adheres to VCWorld by balancing positive analogue cases and negative contrast cases. This systemic skew aligns with prior evaluations on the GeneTAK benchmark, where LLM-based reasoners inherently lean toward positive predictions. Such persistence indicates that balanced context retrieval alone is insufficient to counteract the intrinsic affirmative tendency of generative language models. When synthesizing complex biological narratives, LLMs tend to favor plausible functional associations over strict statistical calibration, exposing a fundamental vulnerability in unconstrained reasoning paradigms for virtual cell modeling.

Implications for LLM-Based Perturbation Reasoning. These findings offer critical insights into the limitations of unadapted LLM-based reasoning paradigms. Although such models maintain stable execution protocols under dense evidence regimes, their surging abstention rates and severe class skews under cross-context shifts demonstrate that textual plausibility does not inherently yield calibrated out-of-distribution generalization. Furthermore, while scaling up foundation model sizes or employing more complex reasoning backbones may mitigate superficial coverage deficits, empirical evidence suggests that raw model scaling alone cannot overcome structural calibration bottlenecks. This highlights a core takeaway for virtual cell modeling, indicating that achieving robust cross-context prediction requires principled architectural inductive biases, such as mechanism-constrained evidence aggregation, rather than relying solely on the emergent reasoning capabilities of general-purpose language models.

H Efficiency Analysis

We access Qwen3-8B directly through the official Qwen API and evaluate ARCS and VCWorld using the same standardized backbone and pricing configuration. Neither method imposes any additional truncation or strict method-specific token budget on the input prompts or generated responses. We report the comprehensive average input tokens, output tokens, and latency for a single query. Input and output prices are given in USD per one million tokens, and inference costs are calculated from the corresponding empirical per-query token usage.

As detailed in Table 12, ARCS and VCWorld exhibit distinct computational and financial profiles. Although ARCS processes a slightly larger input token count (5,300 tokens versus 4,300 tokens for VCWorld) to incorporate comprehensive evidence integration, it generates significantly fewer output tokens (400 tokens compared to 1,300 tokens for VCWorld). This efficiency stems from the structured and targeted evidence synthesis design of ARCS, which bypasses verbose chain-of-thought generation in favor of concise, mechanism-constrained aggregation. Consequently, ARCS achieves a substantially lower per-query latency of 0.31 seconds, representing nearly a 3.5-fold speedup over VCWorld's 1.09 seconds.

Furthermore, this compact output footprint translates directly into operational cost savings. Based on the standard API pricing of \$0.50 per million input tokens and \$2.00 per million output tokens, ARCS incurs an average cost of approximately \$0.00345 per query, whereas VCWorld costs \$0.00475 per query. These findings demonstrate that ARCS provides a highly scalable and cost-effective solution for virtual cell modeling, maintaining superior predictive performance while substantially reducing both inference time and token expenditure.

Table 12: Efficiency comparison between ARCS and VCWorld using the Qwen3-8B backbone. Input and output prices are reported in USD per 1M tokens.

Model	Input Price	Output Price	Input Tokens	Output Tokens	Latency (s)
ARCS	0.50	2.0	5300	400	0.31
VCWorld	0.50	2.0	4300	1300	1.09

Table 13: Target contexts and observed responses for the representative L1000/DIR case. Only the cell line changes between the two queries.

Query	Cell line	Drug	Gene	Dose	Time	Response
q_{712}	MCF-7	molindone	CDKN2A	1.11 μM	24 h	Up ($\log FC = +2.83$)
q_{719}	HT-29	molindone	CDKN2A	1.11 μM	24 h	Down ($\log FC = -2.88$)

I Case Study

We examine a context-dependent direction-flip instance involving molindone and CDKN2A ($p16^{INK4a}$) within the L1000 dataset. This query pair shares identical drug, response gene, dosage, and exposure duration parameters, while varying exclusively in cell line, where the transcriptional response shifts from upregulation in MCF-7 to downregulation in HT-29. This pair serves as an ideal testbed to evaluate whether the framework dynamically traces context-specific evidence rather than collapsing onto static, global drug–gene associations.

For the MCF-7 query, the Analyze module allocates nominal quotas of 10, 15, 5, and 5 to the same-gene, similar-profile, same-mechanism, and distant-context sources, respectively. Conversely, for the HT-29 query, it assigns conservative quotas to the first three sources and omits the distant-context source. Post-retrieval, the MCF-7 and HT-29 queries yield 17 and 8 candidate observations, respectively, demonstrating that the framework dynamically adapts its retrieval strategy despite identical input parameters.

The Critique module subsequently filters the retrieved candidates based on cellular compatibility. For MCF-7, it retains two observations: tofacitinib in MCF-7 at $1.11\,\mu$M and ibudilast in MCF-7 at $3.33\,\mu$M, both exhibiting CDKN2A upregulation. For HT-29, it retains three observations: novobiocin at $0.37\,\mu$M, dutasteride at $10\,\mu$M, and pimobendan at $10\,\mu$M (all measured in HT-29 cells), all showing downregulation. Consequently, the retained evidence strictly conforms to the target cellular context across the shift.

Finally, the evidence synthesis stage integrates these observations to compute context-aware predictions. For MCF-7, the retained instances yield $\mathrm{frac} = 1.0$ and $p_{\mathrm{prior}} = 0.4791$, producing a synthesized score of $0.25 \times 1.0 + 0.75 \times 0.4791 = 0.6093$, which correctly translates to an upregulation prediction. For HT-29, they yield $\mathrm{frac} = 0.0$ and $p_{\mathrm{prior}} = 0.4863$, resulting in $0.25 \times 0.0 + 0.75 \times 0.4863 = 0.3647$, corresponding to a downregulation prediction. Both predictions precisely recover the experimental ground truth.

J Prompt Templates

The following blocks show the prompt components used by the reported pipeline. Braced expressions denote runtime slots. Drug and gene descriptions are generated offline and inserted into the same knowledge block used by Analyze and Critique. Analyze does not see observed outcomes and produces a mechanism hypothesis and evidence priorities; Critique sees the retrieved outcomes and assigns KEEP or DROP decisions to individual records. The wording below is lightly edited for presentation and aligned with the method description in the main paper.

Drug description prompt

```
You are a molecular pharmacologist studying how small molecules alter
    cellular states and gene expression.
```

```
Task: Write a brief overview of the small-molecule drug {pert}. Focus
    on its molecular targets, mechanism of action, and molecular or
    cellular processes that may be affected by perturbing cells with {
    pert}.

Inputs:
- A curated description of {pert}, which may include its
    pharmacological class, known mechanism, molecular targets, or
    established uses.
- Knowledge entries connecting {pert} to genes, proteins, pathways, or
     protein complexes. Entries may come from curated pharmacological
    or molecular-interaction resources, or from explicitly marked
    predicted associations.

Format:
- Write no more than five sentences as one brief overview.
- First describe the pharmacological class, primary targets, and
    mechanism of action of {pert}, when supported by the inputs.
- Then describe affected pathways, cellular processes, or downstream
    responses only when they are supported by the supplied description
    or relations.

Constraints:
- Use only information supported by the description and knowledge
    entries. Do not add targets, mechanisms, pathways, cellular effects
    , or clinical claims.
- Preserve the type and direction of each relation. State activation
    or inhibition only when explicitly annotated. Binding, targeting,
    physical interaction, pathway membership, and complex membership do
     not establish a direction of effect or gene-expression change.
- Mark predicted associations as predicted, putative, or possible
    rather than established.
- Do not infer that a molecular interaction causes differential
    expression of a downstream gene.
- If the inputs do not specify a mechanism or downstream response,
    omit the claim.
- Avoid generic statements such as "this drug affects cellular
    processes."
- Use a professional scientific tone. Do not add notes, citations,
    references, bullet points, headings, or confidence scores.

Description of drug {pert}:
{summary}

Knowledge entries relating {pert} to genes, proteins, pathways, or
    protein complexes:
{entries}

{drug_block}

Additional guidance: You may qualify target interactions using the
    supplied affinity or potency values. If no curated pharmacological
    target is provided, do not name a specific target or mechanism.

Brief overview of drug {pert}:
```

The runtime target block is rendered as follows when curated target evidence is available:

```
Curated pharmacological target annotations:
```

```
- {target} ({predicate}); affinity pAct {value} ({Action}); target
    family {family}
- {gene} (BindingDB {KI|IC50|KD|EC50} {value} nM)
```

When no curated target is available, the block is replaced by:

```
No curated pharmacological target was found for this drug in KG-
    curated, DrugCentral, ChEMBL, IUPHAR, BindingDB or activity panels.
```

The relation renderer maps `inhibits` to `down-regulates`, `activates` to `up-regulates`, and `binds`/`targets` to an explicitly direction-free relation. Expression relations retain their annotated direction; other relations are rendered as associations and carry a source-level suffix.

Gene description prompt

```
You are a molecular biologist studying how genes respond to molecular
    and cellular perturbations.

Task: Write a brief overview of the human gene {name}. Focus on its
    molecular function and on perturbations that may affect its
    expression. Relevant information may include upstream regulators,
    signaling pathways, feedback or compensatory mechanisms, and
    cellular responses explicitly connected to {name} in the supplied
    knowledge.

Inputs:
- A curated description of {name}, which may include its molecular
    function, protein activity, subcellular role, or participation in
    biological processes.
- Knowledge entries connecting {name} to upstream regulators, genes or
     proteins, pathways, biological processes, or protein complexes.
    Entries may come from curated gene-function, gene-regulation,
    pathway, or molecular-interaction resources, or from explicitly
    marked predicted associations.

Format:
- Write no more than five sentences as one brief overview.
- First describe the principal molecular function of {name}.
- Then describe upstream regulators or perturbations that may affect
    the expression of {name}, when explicitly supported by the inputs.
- Use pathway, protein-interaction, or complex-membership information
    as biological context, not as direct evidence of expression
    regulation.

Constraints:
- Use only information supported by the description and knowledge
    entries. Do not add regulators, perturbations, pathways, expression
     effects, or disease claims.
- Preserve causal direction. If {name} regulates another entity, this
    describes a downstream function of {name}; it does not show that
    the other entity regulates {name}'s expression.
- Preserve the biological level of each relation. A change in protein
    activity or abundance does not necessarily imply a change in gene
    expression.
- Treat binding, physical interaction, pathway membership, biological-
    process membership, and complex membership as functional context,
    not as evidence that a perturbation changes {name}'s expression.
- Mark predicted associations as predicted, putative, or possible
    rather than established.
```

```
- If no upstream regulator or expression-related perturbation is
    identified, focus on the gene's function and biological context.
- Summarize repeated regulatory themes concisely and mention
    differences only when they affect how {name} may be regulated.
- Avoid claims about the gene's broad importance and generic
    statements such as "this gene is involved in cellular processes."
- Use a professional scientific tone. Do not add notes, citations,
    references, bullet points, headings, or confidence scores.

Description of gene {name}:
{summary}

Knowledge entries relating {name} to regulators, genes, proteins,
    pathways, biological processes, or protein complexes:
{entries}

Additional regulatory knowledge:
- {g} is a transcription factor with {n} curated high-confidence (A/B-
    level) target genes.
- It is reported to be regulated by transcription factors {regulators
    }.
- It is reported to be phosphorylated by kinases {kinases}.

Additional guidance: You may mention upstream transcription factors or
     kinases only when they are listed in the provided knowledge.

Brief overview of gene {name} and perturbations that may affect its
    expression:
```

The regulatory block is omitted when all three of its fields are unavailable. The gene slot is an uppercase gene symbol; the summary is selected from the available curated function summaries and truncated to the configured limit.

Analyze system prompt

```
You are a biomedical researcher analyzing a drug-response-gene query.
    Use the provided entity descriptions and BioRelKG relations to
    formulate a KG-grounded mechanism hypothesis and prioritize the
    evidence used for retrieval. Do not make the final prediction.

Inputs include the drug description and drug-target relations, the
    response gene description, relations involving transcriptional
    regulators or pathways, recorded drug-to-gene relations, and the
    target cellular context. Response labels and candidate examples are
     not provided at this stage.

Constraints:
1. Use only the supplied descriptions and KG relations.
2. Describe candidate regulatory paths from the drug to the response
    gene
   through drug targets, transcriptional regulators, or pathways.
3. Prioritize observations for the same response gene, genes
    associated with
   the same transcription factor or pathway, and pharmacologically
    similar
   drugs sharing the same target.
4. Treat a KG relation as evidence of biological relevance, not as
    proof of
   an expression response or its direction.
5. Return the mechanism hypothesis and evidence priorities only; do
    not infer
   the final outcome.
```

Analyze output contract

```
Return the following two fields and no final prediction:

Mechanism hypothesis: H_q
Evidence priorities: Pi_q

First, summarize the candidate KG-grounded regulatory paths. Second,
    list the prioritized observations in the following order when
    supported by the supplied relations: the same response gene, genes
    associated with the same transcription factor or pathway, and
    pharmacologically
similar drugs sharing the same target.
```

Analyze user prompt

```
TARGET CONTEXT
  compound: {compound}
  gene: {gene}
  cell line: {cell_line}
  dose: {dose} micromolar
  time: {time} hours after treatment

CURATED MECHANISM NOTE FOR THE TARGET COMPOUND
{mechanism_note | none recorded}

{knowledge block | drug and gene descriptions, drug-target relations,
transcriptional regulators, pathways, and recorded drug-to-gene
    relations}

{two-field output contract}
```

Critique system prompt

```
You are a biomedical evidence auditor.

You will receive one target experiment and records retrieved from a
    perturbation database. Each record measures the same gene in a
    DIFFERENT experimental context, using a different compound, cell
    line, dose, or time. Each record has its own observed result.

Decide for each record whether its result can be transferred to the
    target context.

WHY THESE RECORDS ARE WORTH READING AT ALL

Each record comes from a different experiment and usually uses a
    different compound. This is the premise of the task: the database
    contains no measurement of the target gene under the target
    compound in the exact target context. A different context is
    therefore not, by itself, a reason to drop a record.

Judge whether the recorded result provides evidence about the gene in
    the target context. A record should be retained because of a
    biological connection to the target, not solely because its context
     is numerically close.

WHAT MAKES A RECORD CARRY OVER
```

```
1. Shared mechanism. If the record's compound reaches the gene through
     a route recorded in the knowledge block, or shares a target with
    the target compound, the response may transfer even when the
    compound, cell line, or exact dose differs.
2. A neighbouring step in the same pathway. An upstream or downstream
    gene may respond in the same direction at a different time because
    it belongs to the same regulatory chain.

WHAT STOPS A RECORD FROM CARRYING OVER

3. A large dose gap. Responses are often monotonic, but biphasic
    responses can reverse direction at low and high doses. A difference
     of about one order of magnitude is usually transferable; a gap of
    two or more orders may reverse the direction. Assess whether the
    direction is likely to change, not merely whether the doses differ.
4. A biologically relevant cell-line difference. Cell lines differ in
    receptor and transporter expression and pathway dependence, so the
    same compound, dose, and time can produce opposite directions. A
    different cell line calls for caution but is not, by itself,
    sufficient reason to drop a record.
5. A non-significant result is not evidence for the opposite direction
    . It reflects the result observed at that experiment's statistical
    power. It is still informative about whether the gene responds, so
    do not discard it just because it lacks a direction.

HOW TO ANSWER

Write exactly one line for every record, in index order. Use the
    record index in square brackets, followed by KEEP or DROP, a hyphen
    , and one short sentence explaining why the result should or should
     not transfer. Naming a context difference alone is not sufficient;
     state how it affects the result.

[1] KEEP - the compound shares a target with the target compound, and
    the gene moved in the same direction at a comparable dose.
[2] KEEP - the gene did not respond to a different compound in another
     cell line, which provides evidence about whether it responds in
    the target context.
[3] DROP - the gene moved in the opposite direction at a dose two
    orders of magnitude above the target dose, where the response may
    reverse.

Then, on its own final line, list the indices you kept:

Selected: 1, 3, 4

RULES YOU MUST FOLLOW

- Keep between {lo} and {hi} records. This is a hard requirement; an
    answer outside this range is discarded and replaced by a
    deterministic fallback.
- Choose the records by comparing their biological relevance. Do not
    meet the quota by selecting records at random.
- Write one line for every record. An omitted record is treated as
    kept.
- The Selected line must list exactly the records you marked KEEP.
- Use only the indices you were given. Do not invent one.
- Base the decision on the experiment and the supplied biology. Do not
     infer the target experiment's unobserved result.
```

Critique user prompt

```
--- the target experiment ---

Compound: {compound}
Gene measured: {gene}
Cell line: {cell_line}
Dose: {dose} micromolar
Time: {time} hours

{task question}

[Mechanism note for the target compound: {mechanism_note}]

[{knowledge block -- the same knowledge block used by Analyze}]

--- the analysis used to prioritize evidence ---
Mechanism hypothesis: {H_q}
Evidence priorities: {Pi_q}
Judge the records using this hypothesis and these priorities. Context
similarity alone is not sufficient to retain a record.

--- the {N} retrieved records ---

[1]  retrieved as: {channel}
     compound {c} | cell line {cl} | dose {d} micromolar | time {t}
    hours
     distance from the target: {same cell line|different cell line} |
    dose step {x} | time step {y}
     result of this measurement: {outcome rendered as a word}
     what this record says about the question: {bearing}
```

For the response task, the question is whether the gene responds in the target experiment. Both significant increases and decreases count as responses; a non-significant result counts as no response. For the direction task, the question is whether the gene moves UP or DOWN. A non-significant result carries no direction and therefore supports neither answer. The per-record bearing field is generated from the observed outcome and task.