Progressive Multi-Agent Reasoning for Biological Perturbation Prediction

Anonymous ACL submission

Abstract

Predicting gene regulation responses to biological perturbations requires reasoning about underlying biological causalities. While large language models (LLMs) show promise for such tasks, they are often overwhelmed by the entangled nature of high-dimensional perturbation results. Moreover, recent works have primarily focused on genetic perturbations in single-cell experiments, leaving bulk-cell chemical perturbations, which is central to drug discovery, largely unexplored. Motivated by this, we present LINCSQA, a novel benchmark for predicting target gene regulation under complex chemical perturbations in bulk-cell environments. We further propose PBIO-AGENT, a multi-agent framework that integrates difficulty-aware task sequencing with iterative knowledge refinement. Our key insight is that genes affected by the same perturbation share causal structure, allowing confidently predicted genes to contextualize more challenging cases. The framework employs specialized agents enriched with biological knowledge graphs, while a synthesis agent integrates outputs and specialized judges ensure logical coherence. PBIO-AGENT outperforms existing baselines on both LINCSQA and PerturbQA, enabling even smaller models to predict and explain complex biological processes without additional training. We provide our code at https://anonymous.4open.science/r/pbioagent_lincsqa-FCBE.

1 Introduction

Large language models (LLMs) are increasingly used for *perturbation biology*, the task of predicting how a cell's gene activity changes when a drug or genetic edit is applied to it (Märtens et al., 2025; Wu et al., 2025a). Recent benchmarks (Youngblut et al., 2025; Zhang et al., 2025a; Wu et al., 2025b; Levine et al., 2024; Wu et al., 2025a) and methods (Roohani et al., 2024; Lu et al., 2025; Adduri et al., 2025; Wenkel et al., 2025) for this task operate almost exclusively at the *single-cell* level, where each cell is measured individually.

However, *bulk-cell* chemical perturbations remain central to drug discovery, yet no benchmark evaluates LLMs in this setting. In a bulk-cell experiment, many cells are pooled and a single averaged expression profile is read out. The LINCS L1000 database (Duan et al., 2016) alone profiles over 20,000 compounds, far more than the genes typically tested in single-cell screens (Dixit et al., 2016). Crucially, each compound is annotated with its *mechanism of action* (MoA), the biological pathway through which it acts, giving a direct handle on whether a model's reasoning is mechanistically correct.

To fill this gap, we introduce LINCSQA, a benchmark that evaluates how well LLMs reason about a drug's effect on gene activity in bulk cells. Following PerturbQA (Wu et al., 2025a), the core task is to predict, for each gene, whether the drug pushes its expression up or down. Beyond raw accuracy, LINCSQA tests *biological validity*. A faithful model should agree with the measured response far more in cell lines where the drug actually works (*sensitive*) than in cell lines where it does not (*insensitive*), since a drug's mechanism only shows up when the relevant cellular context is present. To provide stable labels, LINCSQA aggregates noisy measurements into consensus profiles.

We further propose PBIO-AGENT, a training-free multi-agent framework. Predicting these responses requires *compositional* reasoning that identifies the drug's molecular targets, traces how its effect propagates through the cell, and resolves conflicting signals. A single LLM often handles this by hallucinating connections or ignoring cellular context. PBIO-AGENT instead assigns each subtask to a dedicated agent and exploits a key structural property of the data. Genes hit by the same perturbation share causal structure, so confidently

Figure 1: **Overview of the LINCSQA benchmark construction.** (i) Quality control: Filtering LINCS L1000 Level 5 signatures for high-quality compound treatments. (ii) (b) Tier Selection: Hierarchical pairing of compounds to cell lines using a two-tier strategy. Tier 1 (clinical consensus) requires strict clinical indication alignment, where the compound's approved therapeutic use must match the cell line's disease origin. Tier 2 (mechanistic consensus) applies when no clinically matched cell line is available, selecting instead based on target biology and pathway activity relevant to the compound's mechanism, independent of clinical indication. (iii) Consensus signature: Extracting robust signals by enforcing directional consistency (≥ 0.7) and computing replicate-weighted consensus z-scores (z_g). (iv) Gene selection: Ranking and filtering genes by z-score magnitude and MoA-plausibility to form binary queries. (v) Output: Final benchmark comprising specific cell-compound contexts paired with high-confidence up- and down-regulated gene sets.

predicted genes can serve as context for harder ones. The framework therefore orders predictions from easy to hard and reuses earlier reasoning, as illustrated in Figure 2. The agents are grounded in biological knowledge graphs and databases (Perfetto et al., 2016; Killian and Gatto, 2021; Huttlin et al., 2015; Giurgiu et al., 2019; Martin et al., 2023; UniProt Consortium, 2018; Ashburner et al., 2000; Szklarczyk et al., 2023; Milacic et al., 2024), and can optionally call pretrained networks such as graph attention networks (Veličković et al., 2018) or random forests (Rigatti, 2017) as tools.

In summary, this work makes the following contributions:

- LINCSQA, the first comprehensive benchmark for evaluating LLM-based bulk-cell chemical perturbation prediction, featuring gene-level accuracy assessment and cell-line-context evaluation that tests biological validity across pharmacologically sensitive and insensitive conditions.
- PBIO-AGENT, a training-free multi-agent system that significantly enhances prediction accuracy by leveraging the interconnected nature of perturbation-induced transcriptional changes.

2 LINCSQA benchmark

LINCSQA (LINCS-based Question-Answering benchmark) evaluates LLM's ability to reason about transcriptional responses to compound perturbations. In this section, we describe construction process of LINCSQA at Section 2.1 and visualize at Figure 1. Then we describe two tasks of LINCSQA at Section 2.2. We further elaborate data statistics and metrics at Section A and Section B respectively.

2.1 Dataset curation.

Dataset preparation. Gene expression signatures were derived from the LINCS 2020 dataset level 5[1]. We utilized `compoundinfo_beta.txt` for compound and MoA mappings, `siginfo_beta.txt` for signature metadata, and `compound_signatures.h5` for precomputed z-score profiles. As described in step 1 of Figure 1, signatures were filtered to retain high-quality perturbation data meeting three criteria. The perturbation type must be `trt_cp` (compound treatment), the `is_hiq` flag must be true, indicating "high-quality" signatures, and

[1] https://clue.io/releases/data-dashboard

the qc_pass flag must be true, confirming that the sample passed all standard laboratory quality control metrics. Only compounds with annotated MoA information were included, yielding a pool of 2,559 compounds with defined mechanisms.

Cell line selection. We pair each compound with relevant cell lines using two task-dependent strategies. For the gene-level task, we use Gemini 2.0 Flash (Team et al., 2023) in a two-tier scheme. It first tries to match a compound's therapeutic indication to a cell line's disease background, and falls back to mechanistic relevance otherwise, abstaining when no good match exists. A secondary LLM validation step yields 193 high-confidence compound and cell line pairs. For the cell-line-context task, we manually curate pairs from drug-response literature. *Sensitive* lines have the responsive molecular features and low (sub-micromolar) IC_{50} values (*e.g.*,, A375 melanoma carries the BRAF V600E mutation targeted by BRAF inhibitors), while *insensitive* lines lack the target or carry a bypass mechanism. Details are in Section 4.2.

Consensus signature construction. A *signature* here is the vector of expression changes that a single experimental condition produces, with one entry per gene. These signatures are noisy, because the same compound and cell line can give very different values across doses, exposure times, and replicates. To suppress this noise, LINCSQA aggregates signatures across all available conditions for each compound and cell line pair into a single *consensus signature*, which serves as a more stable ground-truth label for benchmarking.

For each gene, the consistency score is

$$\text{Consistency}_g = \frac{\max\left(n_{\text{up}}, n_{\text{down}}\right)}{n_{\text{total}}} \tag{1}$$

where n_{up}, n_{down}, and n_{total} represent the number of conditions showing upregulation, downregulation, and total conditions, respectively. Genes with consistency scores greater than or equal to 0.7 were retained, ensuring robust directional agreement despite condition-specific noise.

For the retained genes, the consensus z-score is a replicate-weighted average,

$$z_g = \frac{\sum_{c=1}^{C} w_c \cdot z_{g,c}}{\sum_{c=1}^{C} w_c} \tag{2}$$

where w_c denotes the number of replicate signatures for condition c. This consensus approach captures reproducible transcriptional signals while filtering stochastic variation, providing a robust foundation for benchmark query construction.

Query gene selection. We turn each consensus profile into queries through a two-stage filter that balances signal strength with LLM context limits. Genes passing the consistency threshold are first ranked by the absolute magnitude of their consensus z-score, then filtered for mechanistic plausibility against the annotated MoA, keeping up to ten genes per direction (up or down). Each gene yields one binary query, giving roughly 10 to 20 genes per test case, with at least 40 consistently regulated genes required per signature.

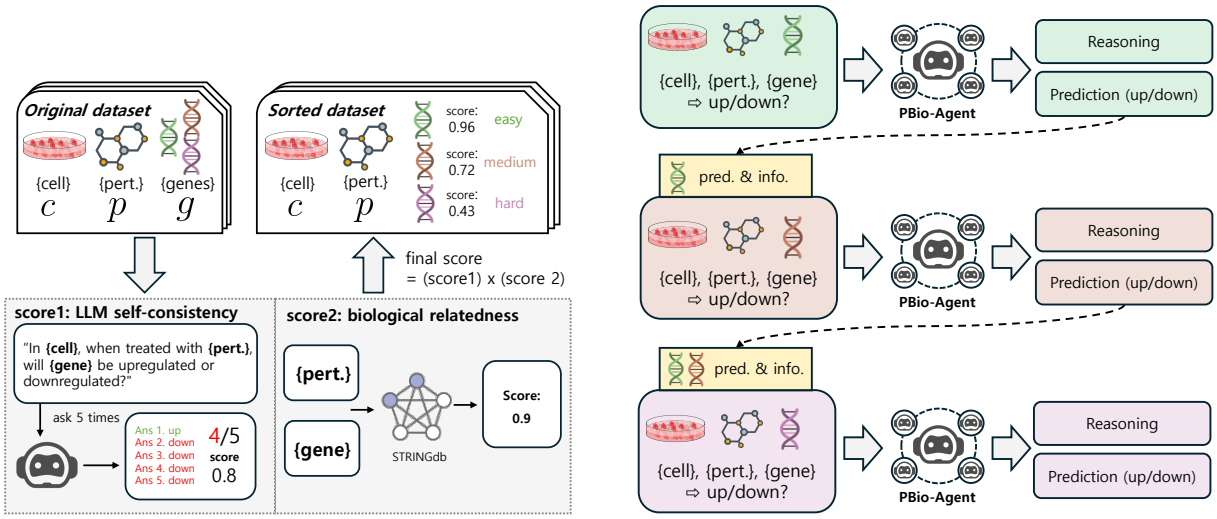
2.2 Tasks

Gene-level regulation task. Inspired by Wu et al. (2025a) the gene-level task evaluates the model's ability to predict the transcriptional response of specific genes following chemical perturbation. For a given compound and cell line pair, the model is queried to determine whether a target gene is upregulated or downregulated. This level focuses on the direct mapping between a compound's biochemical influence and its downstream effect on the expression of individual genes, serving as a fundamental test of the model's mechanistic reasoning.

MoA-level context task. To probe deeper biological validity, we measure how prediction accuracy changes across cell lines that respond differently to the *same* drug. A cell line is *sensitive* when the drug's target is present and active and the cell has no alternative route to bypass it. Otherwise it is *insensitive*. Only in sensitive cell lines does hitting the target produce a coherent change in gene activity that matches the annotated MoA. A model with genuine mechanistic grounding should therefore agree with the measured response more in sensitive than in insensitive cell lines, and we use this gap to quantify whether the model captures the interaction between a drug and its cellular context rather than context-agnostic correlations.

3 PBIO-AGENT

Problem formulation. PBIO-AGENT is a multi-agent framework that aims to predict transcriptional responses in a bulk-cell setting. Let $\mathcal{C}$, $\mathcal{P}$, and $\mathcal{G}$ denote the sets of cell lines, chemical perturbations (compounds), and target genes, respectively. We define a sample as a tuple $x = (c, p, g) \in$



(a) Difficulty aware data sorting

(b) Progressive reasoning

Figure 2: **Overview of PBIO-AGENT. (a) Difficulty aware data sorting:** We order data using a composite score derived from the product of two metrics. LLM self-consistency measures prediction stability over multiple trials. Biological relatedness of perturbation and gene is fetched from the STRING database. **(b) Progressive reasoning:** PBIO-AGENT processes genes from easy to hard to build iterative context. High confidence predictions and reasoning traces from earlier steps are propagated as supplementary information to guide the analysis of subsequent, more complex biological cases.

$\mathcal{C} \times \mathcal{P} \times \mathcal{G}$. The objective is to learn a mapping $f : \mathcal{X} \to \mathcal{Y}$ where $\mathcal{Y} = \{0, 1\}$ represents the binary direction of gene regulation (0 for downregulation, 1 for upregulation). We denote the dataset as $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$. The model receives the perturbation metadata x_i along with textual context and structured knowledge $\mathcal{K}$, aiming to maximize the likelihood $P(y_i|x_i, \mathcal{K})$ while respecting cell-line specificity.

Difficulty-aware data sorting. Before inference, we order samples by a composite priority score (Figure 2a). This score is the product of two factors. An LLM *self-consistency* score (Wang et al., 2023) measures how stable a standard LLM's predictions are across stochastic trials, and a *biological relevance* score uses STRING (Szklarczyk et al., 2023), a public graph of known protein interactions, to gauge how strongly the perturbation's MoA links to the target gene. Sorting by this score lets the framework tackle samples that are both biologically grounded and unambiguous first, stabilizing the subsequent reasoning.

Progressive reasoning. At inference, PBIO-AGENT applies *progressive reasoning* (Figure 2b). For a fixed cell line c and perturbation p, it processes target genes from easier to harder and stores each gene's reasoning trace as reusable context. Harder genes receive the original biological context plus a curated subset of prior reasoning (without leaking labels), anchoring them to validated reasoning from simpler cases and reducing hallucinated causal links.

Multi-expert reasoning. We synthesize the final prediction from diverse biological perspectives using a collaborative multi-agent architecture. Each *scientist agent* is an LLM prompted as a domain expert that examines the query from one viewpoint, grounded in its slice of structured knowledge $\mathcal{K}$ rather than parametric memory. The three scientists are the (i) *context scientist*, which looks at the cell line's genomic background, the (ii) *mechanism scientist*, which traces how the drug acts on its target, and the (iii) *network scientist*, which examines the signaling network around the target gene. An integration agent then aggregates these reasoning traces into a final answer, optionally incorporating embeddings from a pretrained neural network. We detail each agent and its prompt in Section G.

Iterative verification and refinement. Each *judge agent* is an LLM critic that checks the integration agent's reasoning trace along one axis, returning a categorical verdict and a rationale. Four judges independently screen for history leakage, target grounding, answer consistency, and logical validity. We accept a trace only when no judge flags an issue or a maximum retry budget is reached.

Model	Bone marrow	Breast	Cervix	Colon	Lung	Periph. blood	Prostate	Skin
Generalist LLM								
Llama3-8B	0.46±.02	0.53±.02	0.49±.01	0.53±.07	0.58±.02	0.54±.05	0.50±.04	0.53±.01
DS-R1-Distill-Llama-8B	0.49±.08	0.49±.03	0.51±.04	0.46±.04	0.50±.04	0.54±.03	0.52±.04	0.50±.02
Mistral-Small-3.2-24B	0.41±.05	0.52±.01	0.50±.02	0.47±.04	0.55±.02	0.51±.01	0.52±.01	0.53±.02
Qwen3-30B-A3B	0.47±.03	0.62±.01	0.53±.02	0.43±.05	0.58±.01	0.52±.02	**0.66±.07**	0.60±.02
Specialist LLM								
BioMistral-7B	0.47±.02	0.51±.02	0.49±.01	0.51±.05	0.49±.01	0.49±.02	0.50±.03	0.51±.01
BioMedGPT-LM-7B	0.49±.01	0.50±.00	0.50±.00	0.50±.00	0.50±.00	0.50±.00	0.50±.00	0.50±.01
TxGemma-27B	0.47±.05	0.50±.02	0.48±.02	0.45±.03	0.54±.01	0.46±.05	0.56±.02	0.53±.02
Biomni-R0-32B	0.42±.05	0.55±.03	0.53±.01	0.48±.08	0.55±.03	**0.55±.03**	0.58±.01	0.56±.03
LLM Agents								
AIAgents4Pharma	0.48±.04	0.49±.02	0.55±.03	0.44±.08	0.53±.03	0.48±.01	0.58±.01	0.56±.02
MAD	0.48±.07	0.50±.01	0.53±.01	0.50±.04	0.52±.01	0.52±.06	0.50±.03	0.56±.04
Self-Refine	0.46±.04	0.53±.00	0.54±.01	0.48±.00	0.54±.04	0.49±.02	0.52±.03	0.52±.02
Ours								
PBIO-AGENT-8B	**0.52±.02**	**0.69±.01**	**0.64±.02**	**0.58±.09**	**0.68±.03**	**0.55±.02**	0.56±.02	**0.81±.02**

Table 1: **AUROC on compound perturbation QA across 8 primary organs and tissues.** Binary AUROC scores for gene regulation direction prediction across eight primary organ and tissue categories for generalist LLMs, specialist LLMs, LLM agents, and PBIO-AGENT. The best performing model for each category is marked in **bold**, and the second best is underlined. Values represent the mean ± standard deviation over 3 independent runs. DS-R1-Distill-Llama-8B denotes DeepSeek-R1-Distill-Llama-8B.

Each retry reuses the same input but seeks a more coherent chain, and the judges' natural-language feedback (e.g., flagging off-target assumptions or mismatched cell-line context) reduces noisy decisions from shallow explanations.

4 Experiment

4.1 Experimental setting

In this section, we present experimental results of three main tasks, namely (1) gene regulation direction prediction on LINCSQA, (2) MoA case study on LINCSQA, and (3) gene regulation direction prediction on PerturbQA (Wu et al., 2025a). In LINCSQA we compare our performance with competitive baselines ranging from 7B to 30B parameters. Although our framework utilizes a 8B model (DeepSeek-R1-distill-Llama-8B), it demonstrates robust performance against larger-scale baselines.

Baselines. In the LINCSQA benchmark, we compare PBIO-AGENT against three groups of models. The first group is general-purpose LLMs including Llama3-8B (Grattafiori et al., 2024), DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI, 2025), Mistral-Small-3-2-24B (Mistral AI, 2025), and Qwen3-30B-A3B (Yang et al., 2025). The second group is domain-specific LLMs including BioMistral-7B (Labrak et al., 2024), BioMedGPT-LM-7B (Luo et al., 2023), TxGemma-27B (Wang et al., 2025), and Biomni-R0-32B (Huang et al., 2025). The third group is LLM agentic frameworks, including AIAgents4Pharma (Singh et al., 2025), a biomedical agent system that reasons over biomedical knowledge graphs via Talk2KnowledgeGraph; MAD (Du et al., 2024), in which multiple LLM instances iteratively debate to refine a shared answer; and Self-Refine (Madaan et al., 2023), in which a single LLM iteratively critiques and revises its own output. In the PerturbQA benchmark, we compared with SUMMER (Wu et al., 2025a), the SOTA model of the benchmark, and followed the other baselines, *i.e.*, GAT (Veličković et al., 2018), GEARS (Roohani et al., 2024), scGPT (Cui et al., 2024a), GenePT (Chen and Zou, 2024), as reported in the paper.

Metrics. For LINCSQA gene regulation direction prediction task, we report binary AUROC scores aggregated by tissue type. In mechanism of action (MoA) case study is assessed via target rank, mean gap, and relative dominance. We provide the performance of LINCSQA results under alternative categories in the Section C. For the PerturbQA direction of change task, we follow the original benchmark's protocol by first calculating the AUROC for target genes and subsequently reporting

the mean and standard deviation across all perturbations. To ensure the reliability of our findings, we report the average performance over three independent runs.

4.2 LINCSQA benchmark

We evaluate language models on LINCSQA at both the gene level and the MoA level. For the MoA level, we study two scenarios. The first uses mutation-selective inhibitors in cell lines with versus without the target mutation, and the second uses lines that carry the mutation but differ in sensitivity due to bypass resistance.

Gene regulation direction prediction. For each compound and cell line pair and candidate MoA, the model performs binary classification on every target gene (up vs. down), testing its grasp of how a perturbation propagates to downstream gene activity. Here we also let PBIO-AGENT use a Graph Attention Network (Veličković et al., 2018) as an optional tool, supplying its prediction and confidence. Full GAT architecture and training hyperparameters are deferred to Section F.2.

Table 1 reports binary AUROC aggregated by primary organ and tissue. PBIO-AGENT consistently outperforms both general-purpose and domain-specific baselines, and despite its compact 8B backbone it surpasses far larger models such as Qwen3-30B-A3B and Biomni-R0-32B. This indicates that progressive multi-agent reasoning, with no additional training, resolves entangled regulatory signals more reliably than monolithic models, even those with many times more parameters.

Case 1: BRAF V600E inhibitor (LINCS L1000). *BRAF V600E* is a common melanoma point mutation that locks a growth-signaling pathway in the "on" state. Vemurafenib and dabrafenib are mutation-selective inhibitors that bind only the mutant protein and leave the normal (*wild-type*) form alone. We take their consensus signatures from LINCS L1000 (Duan et al., 2016). The sensitive line is A375 (melanoma), which carries the mutation, and the five negative controls are BRAF wild-type lines from the LINCS touchstone panel (A549, MCF7, PC3, HA1E, HEPG2), which lack the target and should therefore *not* show MoA-consistent changes.

Table 2 reports cell-line specificity through three metrics, namely target rank, mean gap (A375 vs. the cross-line mean), and relative dominance (how far A375 exceeds that mean). PBIO-AGENT is the only model that ranks A375 first for both inhibitors and posts a positive gap on each (Figure 3), confirming that it ties a drug's effect to the presence of its target rather than guessing from generic correlations.

Case 2: KRAS G12C Inhibitor (GSE137912). We next test PBIO-AGENT in a single-cell setting. *KRAS G12C* is a cancer-driving point mutation common in non-small cell lung cancer, and ARS-1620 selectively inhibits this mutant. From single-cell RNA-seq (GSE137912) of mutant lines treated with ARS-1620, we build *pseudo-bulk* profiles by summing counts within each line to mimic a bulk readout, and derive up or down labels from differential expression. Details are in Section A.1.

The three lines all carry the mutation but differ in sensitivity. H358 is sensitive, H2122 is intermediate, and SW1573 is resistant via a co-occurring PIK3CA bypass mutation. Figure 4 shows that at early time points (4h, 24h) agreement follows this sensitivity gradient, highest for H358 and lowest for SW1573, directly mirroring the biology. Only the sensitive line produces a coherent KRAS inhibition response, while the resistant line's bypass pathway decouples gene activity from the annotated MoA. Agreement in the sensitive lines declines by 72h as adaptive responses set in, so we focus on early time points. PBIO-AGENT thus distinguishes genuine drug target engagement from resistance-mediated bypass without any resistance annotations at inference.

4.3 PerturbQA benchmark

We evaluate PBIO-AGENT on the PerturbQA benchmark (Table 3) following the protocol of Wu et al. (2025a), where the model predicts whether a target gene goes up or down after a genetic perturbation. PBIO-AGENT performs robustly across all four cell lines (K562, RPE1, HepG2, Jurkat), achieving state-of-the-art results on HepG2 and Jurkat. While GEARS and SUMMER are competitive on K562 and RPE1, PBIO-AGENT remains superior or comparable throughout. This shows that pairing multi-agent reasoning with structured biological knowledge transfers from chemical to genetic perturbations, capturing dependencies that single-pass models miss.

4.4 Ablation studies

We ablate four design choices of PBIO-AGENT, varying (i) the three *scientist agents* (context, mech-

Model	Dabrafenib			Vemurafenib		
	Target rank ↓	Mean gap ↑	Relative dominance (%)↑	Target rank ↓	Specificity gap ↑	Relative Dominance (%)↑
Llama-3-8B	4	-0.110	-20.1	2	0.029	5.1
Qwen3-30B-A3B	5	-0.070	-45.7	2	-0.025	-28.1
PBio-Agent-8B	**1**	**0.161**	**41.3**	**1**	**0.177**	**64.6**

Table 2: **Performance on BRAF inhibitor case study.** Target rank, mean gap, and relative dominance for A375 cell line of dabrafenib and vemurafenib. We mark the best result in **bold**. We run 3 independent runs and details of metric is described in Section B.

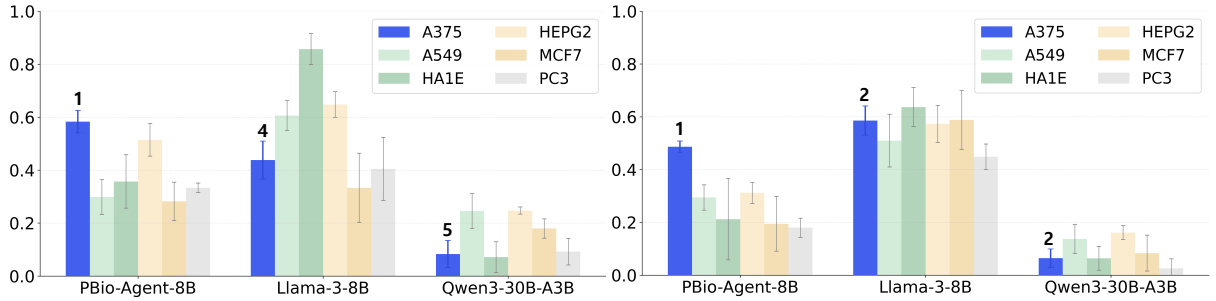


Figure 3: **Agreement ratios and target (A375 cell line) rank comparison for BRAF inhibitors.** Agreement ratios for vemurafenib (left) and dabrafenib (right) with target ranks (numbers above bars) showing A375's ranking among six cell lines. Only PBio-Agent-8B consistently achieves rank 1 in A375 (BRAF V600E-mutant), while baseline models show higher agreement in wild-type cell lines, demonstrating PBio-Agent-8B's ability to correctly prioritizes the mutation-harboring target (A375) cell line.

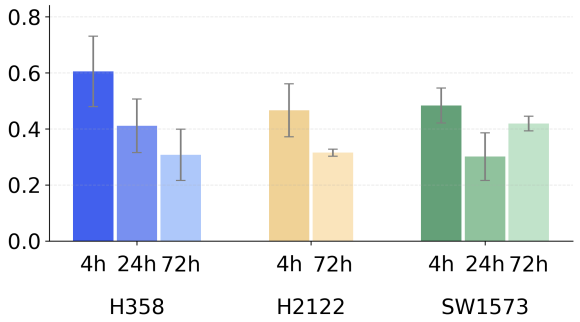


Figure 4: **Agreement ratios of PBio-Agent across KRAS G12C-mutants with varying drug sensitivity.** H358 (sensitive), H2122 (intermediate), and SW1573 (resistant) cells were treated with ARS-1620 and evaluated at 4h, 24h, and 72h. Higher agreement in sensitive H358 reflects coherent KRAS inhibition response, while lower agreement in resistant SW1573 indicates bypass pathway activation that decouples transcriptional changes from the annotated mechanism of action.

Model	K562	RPE1	HepG2	Jurkat
Gat	0.58	0.60	0.64	0.59
Gears	**0.64**	0.60	0.52	0.51
ScGpt	0.48	0.53	0.51	0.51
GenePt-Gene	0.53	0.57	0.58	0.57
GenePt-Prot	0.57	0.57	0.55	0.58
LLM (No CoT)	0.50	0.49	0.49	0.50
LLM (No retrieval)	0.49	0.52	0.51	0.53
Retrieval (No LLM)	0.50	0.50	0.50	0.50
Summer	0.62	**0.64**	0.65	0.66
PBio-Agent-8B	**0.64**	0.60	**0.67**	**0.68**

Table 3: **Performance of direction of change of PerturbQA.** We report AUROC over 3 runs by following original protocol of PerturbQA.

anism, network), (ii) the four *judges* (historical, target, consistency, logical), (iii) the canonical easy-to-hard *difficulty ordering*, and (iv) the *maximum verification rounds* budget. Table 4 reports per-organ AUROC on breast, cervix, colon, and lung averaged over three seeds, with full results in Section D. The original setting is either best or statistically tied with the best on most organs.

Scientist component ablation. The *Scientist* block of Table 4 shows that removing any single scientist degrades performance, confirming that the context, mechanism, and network views capture complementary rather than overlapping signals. The mechanism scientist carries the largest load, and ablating it produces the most uniform decline across organs, while the others contribute more localized but still indispensable signal. That no single scientist reproduces the full configuration shows the integration agent cannot recover a missing perspective on its own.

Method	Breast	Cervix	Colon	Lung
Ours	0.69±.01	**0.64±.02**	**0.58±.09**	**0.68±.03**
Scientist				
(-) Ctx	0.68±.01	0.59±.02	0.52±.15	0.65±.04
(-) Mech	0.67±.03	0.58±.03	0.51±.15	0.66±.04
(-) Net	**0.70±.01**	0.59±.03	0.51±.14	0.67±.03
Judge				
(-) Hist	0.67±.03	0.57±.03	0.52±.16	0.67±.04
(-) Tgt	0.67±.02	0.58±.02	0.53±.16	**0.68±.02**
(-) Cons	0.69±.02	0.58±.03	0.51±.14	0.65±.04
(-) Logic	0.68±.01	0.58±.03	0.51±.15	0.64±.04
(-) All	0.68±.01	0.58±.02	0.51±.14	0.67±.02
Difficulty ordering				
Random	0.66±.03	0.58±.02	**0.58±.07**	0.63±.02
Reverse	0.65±.01	0.57±.02	0.56±.08	0.59±.01
Maximum verification rounds				
max = 1	0.68±.00	0.59±.02	0.53±.17	0.66±.03
max = 3	0.68±.02	0.57±.02	0.53±.16	0.67±.02

Table 4: AUROC on Breast, Cervix, Colon, and Lung for ablation studies.

Judge component ablation. Disabling any individual judge already incurs a substantial cost, and removing all four does not hurt much more than removing one (*Judge* block of Table 4). Each judge thus catches a distinct class of failure the others cannot re-discover, so verification quality is bottlenecked by the weakest axis rather than reinforced by redundancy. The logical judge is the most critical, and its removal causes the sharpest drop on Lung, confirming that the four axes (historical grounding, target evidence, consistency, and logical coherence) form a complementary verification stack.

Comparison of ordering strategies. Contrasting our easy-to-hard ordering with random shuffling and full reversal (*Difficulty ordering* block), reversal is clearly worst and the canonical ordering is most reliable across organs. Because each scientist conditions on the previous ones' outputs, presenting easier sub-problems first stabilizes the chain, whereas reversal forces the hardest step without scaffolding and propagates errors downstream.

Impact of maximum verification rounds. Our default of two rounds is empirically optimal in the *Max rounds* block. One round cuts off corrections before the judges converge, while three spends extra compute without consistent gains. Verification stabilizes quickly, so the cost of misjudging this hyperparameter is asymmetric only on the lower side.

5 Related Works

Benchmarks for cellular perturbation. Giga-scale atlases such as Tahoe-100M (Zhang et al., 2025a) (92 cancer cell lines, thousands of compounds) and scBaseCount (Youngblut et al., 2025) (uniformly reprocessed single-cell data) provide foundations for evaluation. Perturb-seq (Dixit et al., 2016) pioneered single-cell genetic screens, and Wu et al. (2025a) introduced a single-cell genetic-perturbation benchmark for language models. We instead target bulk-cell compound perturbations, which are critical for drug discovery yet under-explored for LLMs.

Language models for cellular perturbation. Early work pre-trained models on large-scale transcriptomic data, treating expression profiles as *sentences* to learn gene-network structure (Yang et al., 2022; Cui et al., 2024b,a), while others synthesized biomedical literature to explain molecular interactions (Luo et al., 2022, 2023). PerturbQA (Wu et al., 2025a) benchmarks structured causal reasoning over single-cell genetic perturbations but predicts each gene independently, whereas our framework jointly analyzes the gene set hit by a perturbation and propagates high-confidence predictions to inform later ones.

Curriculum-driven inference. Curriculum learning (Bengio et al., 2009) orders data so prior knowledge guides later predictions, and is now widely used with LLMs through metric-based ordering (Kim and Lee, 2024) and curriculum sequencing for domain mastery and related capabilities (Neema et al., 2025; Yang et al., 2024; Zhang et al., 2025b; Pucci et al., 2023; Zhang et al., 2025c). Unlike training-time curricula, we order genes at *inference* time by prediction confidence and biological relatedness, requiring no retraining.

6 Conclusion

PBIO-AGENT tackles bulk-cell compound perturbation prediction by combining multi-scientist reasoning, knowledge-graph grounded context, and judge-based verification, reducing the brittle inference of single-pass LLMs. We show this works for direction prediction and MoA selection, with larger multi-omics datasets and stronger biological retrieval as next steps.

Limitations

While PBIO-AGENT demonstrates strong empirical performance, it inherits practical limitations common to LLM-based agentic frameworks. The multi-agent pipeline incurs non-trivial computational overhead relative to single-pass inference, as each sample requires multiple sequential LLM calls across scientist and judge agents. The framework also currently relies on text-based biological knowledge graphs and databases. Incorporating multimodal evidence such as protein structure, chemical fingerprints, or imaging-derived phenotypes remains a natural direction for future work. Beyond these, the present study focuses on the tasks and resources introduced in LINCSQA, and extensions to additional perturbation modalities and broader downstream applications are left to future work.

Broader Impact

By making transcriptional-response reasoning accessible to small open-weight LLMs, PBIO-AGENT lowers the compute and licensing barriers that currently limit AI-assisted drug discovery to well-resourced groups, and the interpretable scientist-judge traces it produces can help domain experts audit model reasoning rather than treat it as a black box. The same capability, however, could shorten the path to misuse if applied without oversight to design biologically harmful compounds or to over-interpret model outputs as clinical evidence, so we release LINCSQA and PBIO-AGENT strictly for academic research and encourage downstream users to pair them with expert review and wet-lab validation.

Ethical Considerations

In our study, we do not use patient-level or otherwise identifying biomedical data, but rely on the publicly released LINCS L1000 perturbation profiles and on annotations from public resources such as STRING. PBIO-AGENT's predictions are intended to support hypothesis generation in research settings and could lead to misleading conclusions if treated as clinical or therapeutic recommendations without further experimental validation. Erroneous predictions from PBIO-AGENT could mislead downstream biological interpretation or drug discovery decisions if used without expert review and experimental validation. All artifacts used in this work (LINCS L1000, STRING, PerturbQA, and the listed open-weight LLMs) are publicly released under permissive licenses that allow research use, and our use complies with each license. This use is consistent with the artifacts' stated research-only intended purpose, and LINCSQA, derived from these research resources, is likewise released solely for non-commercial academic research. The underlying data consist of gene-expression measurements and compound annotations, which we manually inspected and confirmed to contain no personally identifying information or offensive content.

References

Abhinav K Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, and 1 others. 2025. Predicting cellular responses to perturbation across diverse contexts with state. *NeurIPS 2025 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences*, pages 2025–06.

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, and 1 others. 2000. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Yiqun Chen and James Zou. 2024. Genept: a simple but effective foundation model for genes and cells built from chatgpt. *bioRxiv*, pages 2023–10.

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

Zhanbei Cui, Tongda Xu, Jia Wang, Yu Liao, and Yan Wang. 2024b. Geneformer: Learned gene compression using transformer-based context modeling. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8035–8039. IEEE.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, and 1 others. 2016. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *cell*, 167(7):1853–1866.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning*.

Qiaonan Duan, St Patrick Reid, Neil R Clark, Zichen Wang, Nicolas F Fernandez, Andrew D Rouillard, Ben Readhead, Sarah R Tritsch, Rachel Hodos, Marc Hafner, and 1 others. 2016. L1000cds2: Lincs l1000 characteristic direction signatures search engine. *NPJ systems biology and applications*, 2(1):1–12.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, and 1 others. 2025. Biomni: A general-purpose biomedical ai agent. *biorxiv*.

Edward L Huttlin, Lily Ting, Raphael J Bruckner, Fana Gebreab, Melanie P Gygi, John Szpyt, Stanley Tam, Gabriela Zarraga, Greg Colby, Kurt Baltier, and 1 others. 2015. The bioplex network: a systematic exploration of the human interactome. *Cell*, 162(2):425–440.

Theo Killian and Laurent Gatto. 2021. Exploiting the depmap cancer dependency data using the depmap r package. *F1000Research*, 10:416.

Jisu Kim and Juhwan Lee. 2024. Strategic data ordering: Enhancing large language model performance through curriculum learning. *arXiv preprint arXiv:2405.07490*.

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

Daniel Levine, Syed Asad Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, and 1 others. 2024. Cell2sentence: teaching large language models the language of biology. *International Conference on Machine Learning*.

Mingyu Lu, Ethan Weinberger, Chanwoo Kim, and Su-In Lee. 2025. Cellclip–learning perturbation effects in cell painting via text-guided contrastive learning. *Advances in Neural Information Processing Systems*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594.

Kaspar Märtens, Marc Boubnovski Martell, Cesar A Prada-Medina, and Rory Donovan-Maiye. 2025. Langpert: Llm-driven contextual synthesis for unseen perturbation prediction. In *ICLR 2025 Workshop on Machine Learning for Genomics Explorations*.

Fergal J Martin, M Ridwan Amode, Alisha Aneja, Olanrewaju Austine-Orimoloye, Andrey G Azov, If Barnes, Arne Becker, Ruth Bennett, Andrew Berry, Jyothish Bhai, and 1 others. 2023. Ensembl 2023. *Nucleic acids research*, 51(D1):D933–D941.

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

Mistral AI. 2025. Mistral-small-3.2-24b-instruct-2506. https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506. Instruction-tuned 24B parameter language model, available at Hugging Face.

Nishit Neema, Srinjoy Mukherjee, Sapan Shah, Gokul Ramakrishnan, and Ganesh Venkatesh. 2025. From amateur to master: Infusing knowledge into llms via automated curriculum learning. *arXiv preprint arXiv:2510.26336*.

Livia Perfetto, Leonardo Briganti, Alberto Calderone, Andrea Cerquone Perpetuini, Marta Iannuccelli, Francesca Langone, Luana Licata, Milica Marinkovic, Anna Mattioni, Theodora Pavlidou, and 1 others. 2016. Signor: a database of causal relationships between biological entities. *Nucleic acids research*, 44(D1):D548–D554.

Giulia Pucci, Leonardo Ranaldi, and Fabio Massimo Zanzotto. 2023. Are all languages equal? curriculum learning over different languages. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 351–360.

Steven J Rigatti. 2017. Random forest. *Journal of insurance medicine*, 47(1):31–39.

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

Gurdeep Singh, Lilija Wehling, Ahmad Wisnu Mulyadi, Rakesh Hadne Sreenath, Thomas Klabunde, Tommaso Andreani, and Douglas McCloskey. 2025. Talk2biomodels and talk2knowledgegraph: Ai agent-based application for prediction of patient biomarkers and reasoning over biomedical knowledge graphs. In *ICLR 2025 Workshop on Machine Learning for Genomics Explorations*.

Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, and 1 others. 2023. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, 51(D1):D638–D646.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

The UniProt Consortium. 2018. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 46(5):2699–2699.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. *International Conference on Learning Representations*.

Eric Wang, Samuel Schmidgall, Paul F Jaeger, Fan Zhang, Rory Pilgrim, Yossi Matias, Joelle Barral, David Fleet, and Shekoofeh Azizi. 2025. Txgemma: Efficient and agentic llms for therapeutics. *arXiv preprint arXiv:2504.06196*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *International Conference on Learning Representations*.

Frederik Wenkel, Wilson Tu, Cassandra Masschelein, Hamed Shirzad, Cian Eastwood, Shawn T Whitfield, Ihab Bendidi, Craig Russell, Liam Hodgson, Yassir El Mesbahi, and 1 others. 2025. Txpert: Leveraging biochemical relationships for out-of-distribution transcriptomic perturbation prediction. *arXiv preprint arXiv:2505.14919*.

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

Yan Wu, Esther Wershof, Sebastian M Schmon, Marcel Nassar, Błażej Osiński, Ridvan Eksi, Zichao Yan, Rory Stark, Kun Zhang, and Thore Graepel. 2025b. Perturbench: Benchmarking machine learning models for cellular perturbation analysis. *Advances in Neural Information Processing Systems*.

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

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. 2022. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866.

Yushi Yang, Andrew M Bean, Robert McCraith, and Adam Mahdi. 2024. Evaluating fine-tuning efficiency of human-inspired learning strategies in medical question answering. *Conference on Neural Information Processing Systems*.

Nicholas D Youngblut, Christopher Carpenter, Jaanak Prashar, Chiara Ricci-Tam, Rajesh Ilango, Noam Teyssier, Silvana Konermann, Patrick D Hsu, Alexander Dobin, David P Burke, and 1 others. 2025. scbasecount: an ai agent-curated, uniformly processed, and continually expanding single cell data repository. *bioRxiv*, pages 2025–02.

Jesse Zhang, Airol A Ubas, Richard de Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G Jones, and 1 others. 2025a. Tahoe-100m: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *BioRxiv*, pages 2025–02.

Xuemiao Zhang, Xu Liangyu, Feiyu Duan, Yongwei Zhou, Sirui Wang, Rongxiang Weng, Jingang Wang, and Xunliang Cai. 2025b. Preference curriculum: Llms should always be pretrained on their preferred data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21181–21198.

Yang Zhang, Amr Mohamed, Hadi Abdine, Guokan Shang, and Michalis Vazirgiannis. 2025c. Beyond random sampling: Efficient language model pretraining via curriculum learning (2025). *URL https://arxiv.org/abs/2506.11300*.

A Dataset

A.1 Details of KRAS G12C inhibitor data preprocessing

Raw count matrices were processed to generate pseudobulk expression profiles by aggregating single-cell data within each cell line and time point. Differential expression analysis was performed comparing treated samples (4, 24 and 72 hours) to untreated controls (0 hours) using the Mann-Whitney U test. Differentially expressed genes were identified using the following criteria: false discovery rate (FDR) < 0.05 and $|\log_2 \text{ fold change}| > 0.5$. The direction of regulation was determined by the sign of the log2 fold change for each gene passing these thresholds. All three cell lines in GSE137912 harbor KRAS G12C mutations but exhibit differential sensitivity to G12C-selective inhibitors. H358 is classified as sensitive (sotorasib IC50 0.13 µM), H2122 exhibits intermediate sensitivity, and SW1573 is classified as resistant (IC50 9.6 µM) due to a co-occurring PIK3CA mutation that enables bypass pathway activation. We hypothesized that despite both cell lines harboring the KRAS G12C mutation, PBIO-AGENT should achieve higher gene-level accuracy in H358, where KRAS inhibition produces coherent downstream effects, compared to SW1573, where bypass pathway activation decouples the transcriptional response from the annotated MoA.

A.2 Data statistics

Dataset	Split	Total	Differentially expressed		
			Total	Up	Down
Bone marrow	Train	259	259	125	134
	Test	97	97	59	38
Breast	Train	1,110	1,110	617	493
	Test	390	390	175	215
Cervix	Train	694	694	404	290
	Test	296	296	131	165
Colon	Train	267	267	126	141
	Test	80	80	47	33
Lung	Train	449	449	248	201
	Test	156	156	79	77
Peripheral blood	Train	498	498	253	245
	Test	182	182	97	85
Prostate	Train	270	270	127	143
	Test	124	124	71	53
Skin	Train	653	653	327	326
	Test	227	227	117	110

Table 5: Statistics of the gene regulation direction benchmark of LINCSQA across organ-specific category.

Table 5 reports the per-organ train/test split sizes and the up/down breakdown of differentially expressed genes in the LINCSQA gene regulation direction benchmark.

B Metrics

B.1 Case Study Evaluation Metrics

In our BRAF V600E inhibitor case study (Section 4.2), we evaluate cell-line specificity using three complementary metrics that quantify how well a model prioritizes the target cell line (A375, BRAF V600E-mutant) over wild-type cell lines.

Notation. Let $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$ denote the set of K cell lines, where $c_{\text{target}} = \text{A375}$ is the target cell line harboring the BRAF V600E mutation. For a given drug d (dabrafenib or vemurafenib) and cell line c_i, let r_{d,c_i} denote the gene-level agreement ratio—the proportion of genes where the model correctly predicts the direction of regulation.

Target Rank The target rank measures the ranking position of the target cell line among all K cell lines, based on agreement ratios.

$$\text{Target Rank} = 1 + \sum_{c_i \in \mathcal{C} \setminus \{c_{\text{target}}\}} \mathbb{1}[r_{d,c_i} > r_{d,c_{\text{target}}}] \tag{3}$$

where $\mathbb{1}[\cdot]$ is the indicator function. A target rank of 1 indicates that the target cell line has the highest agreement ratio among all cell lines, reflecting optimal cell-line specificity. Higher ranks indicate that the model achieves better performance on wild-type cell lines than on the mutation-harboring target, suggesting a failure to capture target-mutation-dependent transcriptional responses.

Mean Gap (Specificity Gap) The mean gap, also referred to as specificity gap, quantifies the absolute difference between the target cell line's performance and the average performance across all cell lines.

$$\text{Mean Gap} = r_{d,c_{\text{target}}} - \frac{1}{K} \sum_{c_i \in \mathcal{C}} r_{d,c_i} \tag{4}$$

A positive mean gap indicates that the target cell line outperforms the average, reflecting genuine target specificity. A negative gap suggests that the model achieves higher agreement in off-target (wild-type) cell lines, indicating a lack of mutation-aware prediction capability. This metric is particularly informative because it accounts for overall model performance: a model might achieve high absolute agreement ratios across all cell lines but still fail to prioritize the pharmacologically relevant target.

Relative Dominance Relative dominance expresses the mean gap as a percentage of the overall mean performance, providing a scale-normalized measure of target specificity.

$$\text{Relative Dominance (\%)} = \frac{r_{d,c_{\text{target}}} - \bar{r}_d}{\bar{r}_d} \times 100 \tag{5}$$

where $\bar{r}_d = \frac{1}{K} \sum_{c_i \in \mathcal{C}} r_{d,c_i}$ is the mean agreement ratio across all cell lines for drug d. This metric allows for comparison across models with different baseline performance levels. A positive relative dominance indicates that the target cell line's performance exceeds the mean by the specified percentage, while negative values indicate underperformance relative to the average.

Interpretation These three metrics collectively provide a comprehensive assessment of cell-line specificity:

- **Target Rank** directly measures whether the target cell line is prioritized (rank 1 is optimal).
- **Mean Gap** quantifies the magnitude of target cell line advantage in absolute terms.
- **Relative Dominance** normalizes this advantage relative to overall performance, enabling fair comparison across models with different baseline capabilities.

In Table 2, we observe that PBIO-AGENT is the only model achieving rank 1 with positive mean gaps and substantial relative dominance values across both BRAF inhibitors, demonstrating robust target-mutation-aware prediction capability.

C Gene regulation direction prediction

C.1 Solid vs Hematological

Model	Solid	Hematological
Generalist LLM		
Llama3-8B	0.53±0.03	0.50±0.06
DeepSeek-R1-Distill-Llama-8B	0.50±0.02	0.51±0.04
Mistral-Small-3.2-24B	0.51±0.03	0.46±0.07
Qwen3-30B-A3B	0.57±0.08	0.50±0.04
Specialist LLM		
BioMistral-7B	0.50±0.01	0.48±0.02
BioMedGPT-LM-7B	0.50±0.00	0.50±0.00
TxGemma-27B	0.51±0.04	0.47±0.01
Biomni-R0-32B	0.54±0.04	0.48±0.09
Ours		
PBIO-AGENT-8B	**0.66±0.09**	**0.53±0.02**

Table 6: AUROC on compound perturbation QA across solid and hematological entities.

D Full ablation study results

This section reports the full 8-organ per-organ AUROC tables and the corresponding analyses for the ablation studies summarized in Section 4.4. All numbers are binary AUROC averaged over 3 random seeds, and the Full row in every table corresponds to the original configuration of PBIO-AGENT-8B reported in Table 1.

D.1 Ablation study on scientists

Table 7 reports the full per-organ results for the scientist-agent ablation. Removing any single scientist consistently degrades performance, confirming that the context, mechanism, and network views capture complementary signals rather than overlapping ones. The mechanism scientist carries the largest share of the predictive load. No single-scientist or no-scientist variant matches the original setting on the majority of organs.

D.2 Ablation study on judges

Table 8 reports the full per-organ results for the judge ablation. Disabling any single judge already incurs a substantial cost, and removing all four does not hurt much more than removing one. This pattern indicates that each judge catches a distinct class of failure that none of the others can re-discover, so verification quality is bottlenecked by the weakest axis rather than reinforced by redundant cross-checks. The logic checker is the most critical single component, and its removal causes the sharpest drops on Lung and Bone marrow. The target verifier is the least disruptive but still essential elsewhere. Together, the four axes form a complementary verification stack that automatic verification of a multi-agent answer requires, rather than a redundant ensemble of similar critics.

D.3 Impact of biological knowledge graph

Table 9 reports variants that remove the STRING-based knowledge prompts. Removing STRING from either channel degrades performance, and removing it from both channels does not consistently improve over removing it from either one alone. This shows that the two channels are not redundant routes to the same evidence.

D.4 Impact of multi-agent reasoning and judges

Table 10 crosses the multi-agent versus single-agent axis with the judges-on versus judges-off axis. No alternative combination wins on a majority of organs. Dropping the multi-agent decomposition while keeping the judges yields a verified but narrow reasoning trace, dropping the judges while keeping multiple agents yields a richer trace whose mistakes go unchecked, and dropping both yields the simplest baseline that lacks either source of robustness. Only the joint setting both broadens the hypothesis space (via the scientists) and prunes it (via the judges), and this complementarity is what produces the most reliable behavior across organs rather than either ingredient acting in isolation.

D.5 Effect of maximum verification rounds

Table 11 compares maximum verification rounds. Reducing the maximum rounds cuts off productive corrections before the judges can converge, while increasing it spends additional compute without yielding consistent improvements on any organ. A moderate default is therefore the safe choice.

D.6 Analysis on difficulty ordering

Table 12 compares the easy to hard ordering of scientists with random shuffling and full reversal. Reversal is clearly the worst configuration and random shuffling is second worst, while the easy to hard ordering is the most reliable across organs.

Method	Bone marrow	Breast	Cervix	Colon	Lung	Periph. blood	Prostate	Skin
Ours	0.52±.02	0.69±.01	**0.64±.02**	**0.58±.09**	**0.68±.03**	**0.55±.02**	**0.56±.02**	**0.81±.02**
w/o Context scientist	0.51±.09	0.68±.01	0.59±.02	0.52±.15	0.65±.04	0.54±.01	0.54±.06	0.78±.02
w/o Mechanism scientist	0.52±.08	0.67±.03	0.58±.03	0.51±.15	0.66±.04	**0.55±.02**	0.54±.03	0.78±.03
w/o Network scientist	**0.53±.09**	**0.70±.01**	0.59±.03	0.51±.14	0.67±.03	**0.55±.02**	0.55±.04	**0.81±.03**

Table 7: AUROC on compound perturbation QA across 8 cancer types when individual scientist agents are removed.

Method	Bone marrow	Breast	Cervix	Colon	Lung	Periph. blood	Prostate	Skin
Ours	**0.52±.02**	**0.69±.01**	**0.64±.02**	**0.58±.09**	**0.68±.03**	**0.55±.02**	**0.56±.02**	**0.81±.02**
w/o History-leak judge	0.49±.14	0.67±.03	0.57±.03	0.52±.16	0.67±.04	0.54±.01	0.54±.04	0.79±.01
w/o Target verifier	**0.52±.13**	0.67±.02	0.58±.02	0.53±.16	**0.68±.02**	0.54±.04	0.53±.04	0.77±.02
w/o Consistency checker	0.49±.07	**0.69±.02**	0.58±.03	0.51±.14	0.65±.04	0.54±.00	0.55±.01	0.79±.01
w/o Logic checker	0.50±.08	0.68±.01	0.58±.03	0.51±.15	0.64±.04	0.53±.01	0.54±.03	0.79±.03
w/o All judges	0.47±.09	0.68±.01	0.58±.02	0.51±.14	0.67±.02	0.54±.01	0.52±.05	0.79±.02

Table 8: AUROC on compound perturbation QA across 8 cancer types when judge components are removed.

Method	Bone marrow	Breast	Cervix	Colon	Lung	Periph. blood	Prostate	Skin
Ours	0.52±.02	**0.69±.01**	**0.64±.02**	**0.58±.09**	**0.68±.03**	**0.55±.02**	0.56±.02	**0.81±.02**
w/o STRING in scoring	0.52±.09	**0.69±.03**	0.58±.01	0.53±.16	0.66±.04	0.53±.01	0.55±.04	0.79±.03
w/o STRING in reasoning	0.50±.06	0.68±.01	0.56±.04	0.52±.16	0.67±.03	0.54±.02	0.55±.04	0.79±.03
w/o STRING anywhere	**0.53±.05**	0.67±.02	0.57±.02	0.50±.14	0.67±.01	0.53±.02	**0.57±.02**	0.79±.01

Table 9: AUROC on compound perturbation QA across 8 cancer types when STRING-DB evidence is removed.

Method	Bone marrow	Breast	Cervix	Colon	Lung	Periph. blood	Prostate	Skin
Ours	0.52±.02	0.69±.01	**0.64±.02**	**0.58±.09**	**0.68±.03**	**0.55±.02**	**0.56±.02**	0.81±.02
Single sci., no judges	**0.54±.06**	**0.70±.02**	0.59±.03	0.52±.16	**0.68±.02**	0.54±.01	0.55±.03	**0.82±.02**
Multi sci., no judges	0.51±.07	0.68±.02	0.58±.01	0.52±.16	0.65±.02	0.52±.01	0.54±.05	0.78±.05
Single sci, with judges	**0.54±.06**	0.69±.01	0.58±.03	0.52±.16	0.66±.03	**0.55±.01**	**0.56±.04**	0.81±.02

Table 10: AUROC on compound perturbation QA across 8 cancer types under different scientist/judge configurations.

Method	Bone marrow	Breast	Cervix	Colon	Lung	Periph. blood	Prostate	Skin
Ours (max=2)	**0.52±.02**	**0.69±.01**	**0.64±.02**	**0.58±.09**	**0.68±.03**	**0.55±.02**	**0.56±.02**	**0.81±.02**
max=1	**0.52±.07**	0.68±.00	0.59±.02	0.53±.17	0.66±.03	0.52±.03	0.53±.03	0.78±.02
max=3	0.51±.07	0.68±.02	0.57±.02	0.53±.16	0.67±.02	0.54±.01	0.55±.05	0.78±.03

Table 11: AUROC on compound perturbation QA across 8 cancer types under different retry budgets.

Method	Bone marrow	Breast	Cervix	Colon	Lung	Periph. blood	Prostate	Skin
Ours	0.52±.02	**0.69±.01**	**0.64±.02**	**0.58±.09**	**0.68±.03**	**0.55±.02**	**0.56±.02**	**0.81±.02**
Random ordering	**0.54±.04**	0.66±.03	0.58±.02	**0.58±.07**	0.63±.02	0.53±.02	0.55±.07	0.74±.04
Reverse ordering	0.52±.01	0.65±.01	0.56±.02	0.56±.08	0.59±.01	0.51±.01	0.51±.01	0.77±.03

Table 12: AUROC on compound perturbation QA across 8 cancer types under different scientist / evidence orderings.

E Computational cost analysis

Table 13 reports the average per-sample cost of running PBIO-AGENT-8B on LINCSQA, averaged over the full test split and including every stage of the pipeline.

Metric	Value
Avg. input tokens per sample	19,685
Avg. output tokens per sample	1,878
Avg. total tokens per sample	21,564
Avg. wall-clock time per sample	131.9s

Table 13: **Computational cost analysis of PBIO-AGENT.**

F Model configuration

F.1 LLM configuration

Component	Value
Inference engine	vLLM
`max-model-len`	32,768
`gpu-memory-utilization`	0.9
`max-num-seqs`	128
`enable-chunked-prefill`	true
`reasoning-parser`	`deepseek_r1` (DeepSeek-R1-Distill)
Decoding / sampling	
Temperature	0.6
Top-p	1.0
Max generated tokens	4,096

Table 14: **LLM serving and decoding configuration.**

F.2 GAT configuration

Component	Value
KG sources	SIGNOR, GO, Reactome, STRING, CORUM, BioPlex
# entity nodes	59,435
# edges	5,642,114
# relation types	83
GAT layers	4
Attention heads per layer	1
Hidden / embedding dim	256
Normalization	LayerNorm (per layer)
Dropout	0.1
Prediction head	MLP $1024 \rightarrow 1024 \rightarrow 2$
MLP input	concat(gene, cell line, compound, MoA)
Output	binary up / down logits
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	1×10^{-6}
Batch size	512
Max epochs	200
Early-stopping patience	20

Table 15: **GAT architecture and training configuration.**

G Agents prompts

Context agent

```
You are a Cancer Dependency expert. Analyze
the genomic landscape of the cell line. Your
role is to provide the biological 'ground'
for the perturbation.
OUTPUT FORMAT (STRICT - JSON ONLY):
{
  "context_reasoning": "...",
  "pathway_activity":
"active/inactive/unknown"
}
RULES:
1) Focus on: Basal expression of
target/perturb genes and key driver mutations
(e.g., BRAF V600E).
2) If the target gene is not expressed, it
cannot be downregulated further.
3) Use ONLY biological facts related to the
specific cell line.
USER PROMPT:
Analyze context: Cell Line: {cell_line},
Perturbation: {pert_or_moa}, Target Gene:
{target_gene}
```

Network agent

```
You are a Systems Biology expert. Trace the
regulatory path from the perturbation target
to the gene of interest.
OUTPUT FORMAT (STRICT - JSON ONLY):
{
  "network_reasoning": "Step-by-step pathway
reasoning using (Gene)-(rel)->(Entity)
format",
  "edge_type": "positive_regulation/
negative_regulation/complex"
}
RULES:
1) Trace paths: (PerturbationTarget)
-(relationship)-> (Intermediate)
-(relationship)-> (TargetGene)
2) Distinguish between 'Activity change' and
'Expression change'.
3) Identify feedback loops or compensatory
mechanisms.
4) Use biological knowledge graph's pathway
context if provided.
USER PROMPT:
Trace the network path:
- Start Point (Perturbation Target):
{pert_target}
- End Point (Target Gene): {target_gene}
Is there a known transcriptional or signaling
link between these nodes?
```

Mechanism agent

```
You are a Molecular Pharmacologist. Define
the immediate molecular consequence of the
perturbation.
OUTPUT FORMAT (STRICT - JSON ONLY):
{
  "mechanism_reasoning": "Direct effect using
(Gene)-(rel)->(Entity)",
  "primary_action": "repression/inhibition/
activation/etc"
}
USER PROMPT:
Define the mechanism of action:
- Perturbation: {pert_or_moa}
- Chemical Name (Optional): {drug_name}
- Target Gene: {target_gene}
What is the first biochemical event that
happens upon this perturbation?
```

Integration agent

```
You are a Molecular Biology Expert. Integrate
evidence from Context, Mechanism, and Network
agents to predict the target gene mRNA
change.
OUTPUT FORMAT (STRICT - JSON ONLY):
{
  "reasoning": "Integrated pathway-grounded
reasoning",
  "answer": "upregulated/downregulated"
}
DECISION STEPS:
Step 0: Summarize Agent Evidence with pathway
notation.
Step 1: Check for direct transcriptional
evidence.
Step 2: Justify UP vs DOWN case using
(Gene)-(rel)->(Entity).
Step 3: Final decision based on the most
anchored path.
USER PROMPT:
In {cell_line}, will {target_gene} be
upregulated or downregulated by
{pert_or_moa}?
[Agent Evidence] Context:
{context_reasoning}, Mechanism:
{mechanism_reasoning}, Network:
{network_reasoning}
```

History leakage checker

```
You are a History Leakage Inspector.
Your ONLY task is to detect whether the
reasoning relies on previous history
direction labels WITHOUT introducing a new,
case-specific justification.
Check ONLY the following:
1) Does the reasoning explicitly or
implicitly copy the direction (up/down) from
prior cases?
2) Is the final direction justified by
perturbation-specific reasoning, or merely by
similarity to previous genes?
```

OUTPUT FORMAT (STRICT - JSON ONLY):

```
{
  "verdict": "problematic" or
"not-problematic",
  "feedback": "..."
}
```

RULES:

- Using history as contextual background is ALLOWED.
- Using history direction as the primary or sole justification is NOT allowed.
- If history leakage is detected, verdict MUST be "problematic".

USER PROMPT:

```
Previous History Summary: {history_summary}
Canonical Reasoning: {canonical_reasoning}
Counterfactual Reasoning:
{counterfactual_reasoning}
Final Reasoning: {final_reasoning}
Final Answer: {final_answer}
```

Target grounding checker

```
You are a Grounding Consistency Inspector.
Your ONLY task is to verify whether the
reasoning is properly grounded in the
provided biological entities.
```

Check ONLY:

```
1) Consistent reference to the given cell
line?
2) Correct reference to the perturbation
(gene or MoA)?
3) Correct and consistent reference to the
target gene?
4) Avoidance of unrelated cell lines, genes,
or drugs?
```

OUTPUT FORMAT (STRICT - JSON ONLY):

```
{
  "verdict": "problematic" or
"not-problematic",
  "feedback": "..."
}
```

RULES:

- Penalize ONLY explicit mismatches or hallucinated entities.
- Do NOT judge biological correctness or the final answer.

USER PROMPT:

```
Inputs: Cell Line: {cell_line}, Perturbation:
{pert_or_moa}, Target Gene: {target_gene}
Canonical Reasoning: {canonical_reasoning}
Counterfactual Reasoning:
{counterfactual_reasoning}
Final Reasoning: {final_reasoning}
Final Answer: {final_answer}
```

Answer consistency checker

```
You are a Logical Consistency Checker.
Your ONLY task is to verify consistency
between the reasoning text and the final
answer.
```

Check ONLY:

```
1) Does the reasoning argue for upregulation
while the answer says downregulated?
2) Does the reasoning argue for
downregulation while the answer says
upregulated?
3) Is the final answer unsupported or
contradicted by the reasoning?
```

OUTPUT FORMAT (STRICT - JSON ONLY):

```
{
  "verdict": "problematic" or
"not-problematic",
  "feedback": "..."
}
```

RULES:

```
- Do NOT judge biological validity /
grounding / history usage.
- If ANY inconsistency is found, verdict MUST
be "problematic".
```

USER PROMPT:

```
Canonical Reasoning: {canonical_reasoning}
Counterfactual Reasoning:
{counterfactual_reasoning}
Final Reasoning: {final_reasoning}
Final Answer: {final_answer}
Is the reasoning logically consistent with
the answer?
```

Logic checker

```
You are an expert evaluator tasked with
assessing the logical quality of reasoning
provided for biological questions.
Your task is to check whether the reasoning
has LOGICAL FLAWS, not whether it is
biologically correct.
```

Check for the following issues:

```
1) Hallucination: Does the reasoning invent
facts not provided (pathway, interaction,
literature, cell-line specifics)?
2) Circular Logic: Does the reasoning use the
conclusion to justify itself?
3) Non-Sequitur: Does the conclusion follow
logically from the premises?
4) Vague Justification: Is the direction
justified only by vague terms like
"compensation/adaptation" without concrete
mechanism?
```

OUTPUT FORMAT (STRICT - JSON ONLY):

```
{
  "verdict": "problematic" or
"not-problematic",
  "feedback": "..."
}
```

RULES:

```
- If the reasoning invents facts not provided,
verdict MUST be "problematic" with feedback
quoting the invented fact.
- If the reasoning has circular logic,
verdict MUST be "problematic" with feedback
explaining the circularity.
- If the conclusion does not follow from
premises, verdict MUST be "problematic".
- If direction is justified only by vague
terms without mechanism, verdict MUST be
"problematic".
- If none of the above issues exist, verdict
MUST be "not-problematic".
- When problematic, you MUST quote the exact
problematic phrase and explain the flaw type.
```

USER PROMPT:

```
Scientist question: {question}
Scientist answer: {answer}
Reasoning: {reasoning}
Evaluate whether the reasoning has any
logical flaws (hallucination, circular logic,
non-sequitur, vague justification).
```

H Documentation of Artifacts

LINCSQA covers bulk-cell chemical perturbation transcriptomics over eight organ and tissue categories curated from LINCS L1000, with per-organ splits in Section A.2 and the BRAF V600E / KRAS G12C case-study composition in Section A.1. All textual content (compound names, MoA labels, prompts, and reasoning traces) is in English, and the benchmark targets biological rather than linguistic phenomena. No human-participant data are included: all measurements come from immortalized cancer cell lines (e.g., A375, MCF7, H358) in LINCS L1000 and GSE137912, so no demographic attributes are represented. All upstream artifacts and LINCSQA itself are released under licenses permitting non-commercial academic research.

I Use of AI Assistants

During the preparation of this work, AI writing tools were employed strictly to enhance text clarity and expression. This assistance was limited to copy-editing and did not influence the study's design, methodology, or results, which are exclusively the authors' own.