Progressive Multi-Agent Reasoning for Biological Perturbation Prediction

Hyomin Kim [1 2] **Sang-Yeon Hwang** [2] **Jaechang Lim** [2] **Yinhua Piao** [1 3] **Yunhak Oh** [1 2] **Wooyoun Kim** [1 2] **Chanyoung Park** [1] **Sungsoo Ahn** [1] **Junhyeok Jeon** [2]

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.

1. Introduction

Large language models have shown distinct benefits in biological reasoning and 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 efforts in the computational biology and machine learning communities have developed 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 predicting these gene-expression (*i.e.*,, transcriptional) responses at the single-cell level.

Single-cell studies, in which each cell is measured one by one, have received much attention. However, *bulk-cell* chemical perturbations, where many cells are pooled and a single average expression profile is read out, remain central to drug discovery. The LINCS L1000 database (Duan et al., 2016) contains expression profiles for over 20,000 compounds, far more than the number of genes typically tested in Perturb-seq (Dixit et al., 2016) experiments. Each compound is annotated with its mechanism of action, which lets us directly check a model's mechanistic reasoning. In addition, bulk signatures average the response over many cells, much like the tissue-level measurements used in early-stage drug studies. Despite this practical importance, no benchmark currently evaluates LLMs on bulk-cell chemical perturbation prediction.

To bridge this gap, we introduce LINCSQA, a benchmark to evaluate how LLMs reason through biological responses to drugs in bulk cell. Inspired by PerturbQA (Wu et al., 2025a), LINCSQA evaluates gene regulation direction (up/down) prediction given a hypothesized *mechanism of action* (MoA), which is the biological pathway the drug uses to produce its effect. Crucially, we assess whether LLM predictions exhibit *biological validity*. That is, for a given compound and its known MoA, the model's agreement with observed expression changes should be much higher in cell lines where the drug is active (*sensitive*) than in cell lines where the target pathway is missing or worked around by alternative pathways (*insensitive*). This design directly tests whether LLMs understand that a drug's mechanism only manifests transcriptionally when the relevant biological context is present. Finally, LINCSQA uses consensus signatures to filter out experimental noise, providing a stable ground truth for evaluating transcriptional changes.

In addition to the benchmark, we introduce PBIO-AGENT,

Work conducted during HITS internship by H. Kim and Y. Oh. [1]Korea Advanced Institute of Science and Technology (KAIST) [2]HITS, Seoul, Republic of Korea [3]AI Co-Research & Education for innovative Drug Institute. Correspondence to: Hyomin Kim <hyomin126@kaist.ac.kr>, Junhyeok Jeon <hyeok797@gmail.com>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)

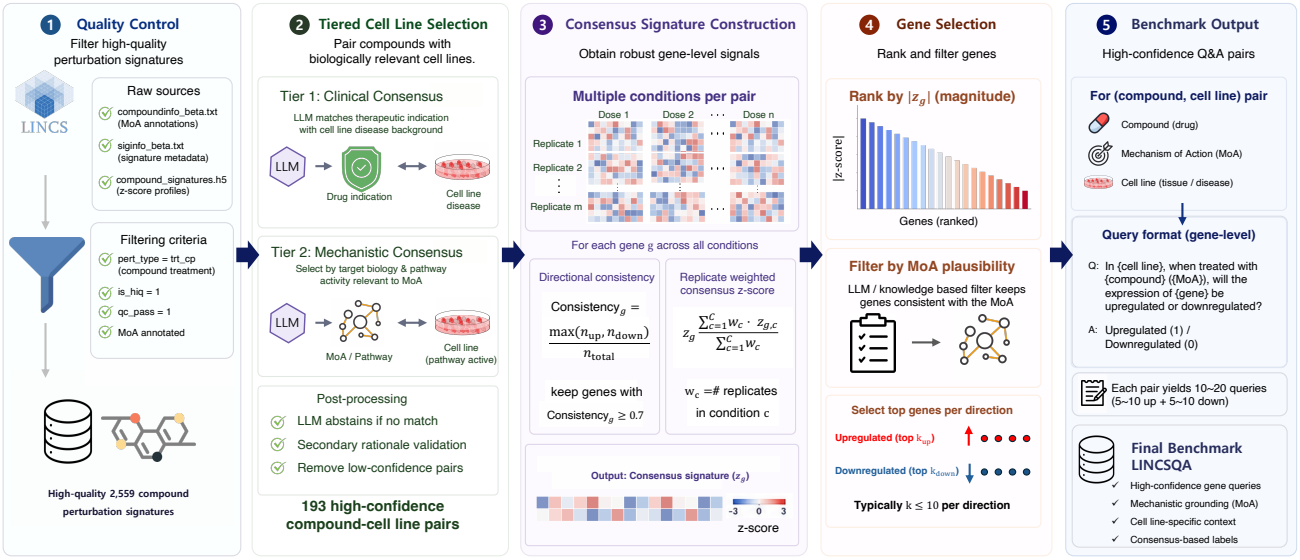


Figure 1. **Overview of the LINCSQA benchmark construction.** (i) Quality control: Filtering LINCS L1000 Level 5 signatures for high-quality compound treatments. (ii) (b) Tier Selection: Hierarchical pairing of compounds to cell lines using a two-tier strategy. Tier 1 (clinical consensus) requires strict 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.

a multi-agent framework that decomposes biological reasoning across specialized agents with access to structured knowledge. Predicting gene regulation from chemical perturbations requires compositional reasoning: retrieving the drug's molecular targets, tracing the cell-type-specific signaling cascade (the chain of protein interactions that carries the drug's effect through the cell), and resolving conflicting regulatory signals. Monolithic LLMs frequently hallucinate pathway connections or ignore cellular context when these subtasks are handled implicitly. Our framework addresses this by assigning each subtask to dedicated agents, and further exploits the shared causal structure among genes affected by the same perturbation through progressive reasoning, as illustrated in Figure 2.

Our multi-agent system integrates specialized reasoning agents that leverage biological knowledge graphs and databases (Perfetto et al., 2016; Killian & 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) to trace regulatory pathways and identify molecular mechanisms. Optionally, the framework can incorporate pre-trained neural networks as tools, such as graph attention networks (Veličković et al., 2018) or random forests (Rigatti, 2017) for pathway inference, enabling ensemble predictions.

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–MoA mappings, `siginfo_beta.txt` for signature metadata, and

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

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 the following criteria: (i) perturbation type trt_cp (compound treatment), (ii) is_hiq flag set to true, indicating "high-quality" signatures and (iii) qc_pass flag set to 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. Cell line selection follows two task-dependent strategies: gene-level prediction and cell-line-context evaluations. For gene-level prediction, we employed a two-tier strategy to pair compounds with biologically relevant cell lines. The first-tier prioritized clinical context, using Gemini 2.0 Flash (Team et al., 2023) to match a compound's therapeutic indication with a cell line's disease background. For compounds without direct disease matches, a second tier focused on mechanistic relevance was applied. The model was instructed to abstain if no appropriate match existed. After a secondary LLM validation of the clinical rationale, 193 high-confidence compound–cell line pairs were retained for the final benchmark.

For the cell-line-context evaluation, we manually curated cell lines based on molecular features and drug sensitivity. Sensitive cell lines were selected based on known responsive features and sub-micromolar IC_{50} values (*e.g.*,, A375 melanoma cells carries the BRAF V600E mutation for BRAF inhibitors). Conversely, insensitive cell lines were chosen based on the absence of the drug target or the presence of bypass resistance mechanisms. These annotations were grounded in published IC_{50} data and drug response literature. We further provide details 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–cell line pair into a single *consensus signature*, which serves as a more stable ground-truth label for benchmarking.

For each gene, we computed a directional consistency score:

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

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, a consensus z-score was computed as 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. Genes used for benchmark queries were selected through a two-stage process designed to balance biological relevance with LLM context constraints. In the first stage, genes that passed the consistency threshold were ranked by the absolute magnitude of their consensus z-scores. In the second stage, the top-ranked genes were further filtered based on mechanistic plausibility with respect to the annotated MoA, with up to ten genes selected per direction (upregulated and downregulated). Each selected gene yielded a single binary query. The resulting query set typically comprised 10–20 genes per test case, subject to a minimum requirement of 40 consistently regulated genes per signature.

2.2. Tasks

Gene-level regulation task. Inspired by Wu et al. (2025a) gene-level task evaluates the model's ability to predict the transcriptional response of specific genes following chemical perturbation. For a given compound–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 assess deeper biological validity, we evaluate prediction accuracy across cell lines that respond differently to the same drug. We call this strength of response *pharmacological sensitivity*. A cell line is *sensitive* when the drug's intended target is present and active, when the surrounding pathway needs that target to work, and when the cell has no alternative route to compensate. Otherwise, it is *insensitive*. In sensitive cell lines, hitting the intended target travels through the signaling network and produces coherent expression changes. In insensitive cell lines, this link can be weakened because the target is barely expressed, other pathways take over, or feedback restores normal pathway output despite the drug. We hypothesize

that a transcriptional-response–based MoA prediction framework with genuine mechanistic grounding should exhibit higher predictive agreement with observed transcriptional changes in sensitive cell lines compared to insensitive ones. By measuring this performance gap, we aim to quantify the extent to which the model captures the interaction between a drug's mode of action and its cellular environment, rather than relying on 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 $x = (c, p, g) \in \mathcal{C} \times \mathcal{P} \times \mathcal{G}$. 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. Prior to inference, we optimize the learning trajectory by implementing a curriculum-based data organization strategy as described in Figure 2a. We compute a composite priority score for each sample, defined as the product of a LLM's self-consistency (Wang et al., 2023) score and biological relevance score. The biological score measures how strongly the perturbation's MoA and the target gene are linked using STRING (Szklarczyk et al., 2023), a public graph of known protein interactions, so that the model focuses on biologically plausible connections first. Simultaneously, the consistency score measures the stability of a standard LLM's predictions over multiple stochastic trials, effectively filtering out noisy samples where the signal is ambiguous. By sorting the dataset based on this metric, the framework prioritizes samples that are both biologically grounded and suited for reasoning, stabilizing the subsequent multi-agent optimization.

Progressive reasoning. At inference time, we further enhance reasoning by applying a progressive reasoning as described in Figure 2b. Specifically, it loops over target genes conditioned on a fixed cell line c and perturbation p. Targets are ordered by a difficulty score and processed from easier to harder, and the reasoning traces for earlier targets are stored as reusable context. For subsequent harder targets, we supply both the original biological context and a curated subset of prior reasoning, enabling evidence reuse without leaking labels. This staged conditioning reduces hallucinated causal links and improves consistency for challenging genes by anchoring them to validated reasoning from simpler cases.

Multi-expert reasoning. We deploy a collaborative multi agent architecture to synthesize a final prediction from diverse biological perspectives. Each *scientist agent* is an LLM prompted to act as a domain expert that examines the query from one biological viewpoint. Each agent is supplied with the corresponding slice of structured knowledge $\mathcal{K}$ so that its reasoning remains grounded in evidence rather than parametric memory.

The three specialized scientists are the (i) *context scientist*, which looks at the cell line's genomic background; (ii) *mechanism scientist*, which traces how the drug acts on its target at a biochemical level; and (iii) *network scientist*, which examines the structure of the signaling network around the target gene. These disparate reasoning traces are then aggregated by an integration agent, which synthesizes the agents insights and knowledge to formulate a final response to the user question. Optionally, this agent can incorporate embeddings from a pretrained neural network. We further details the explanation and prompt of each agents at Section F.

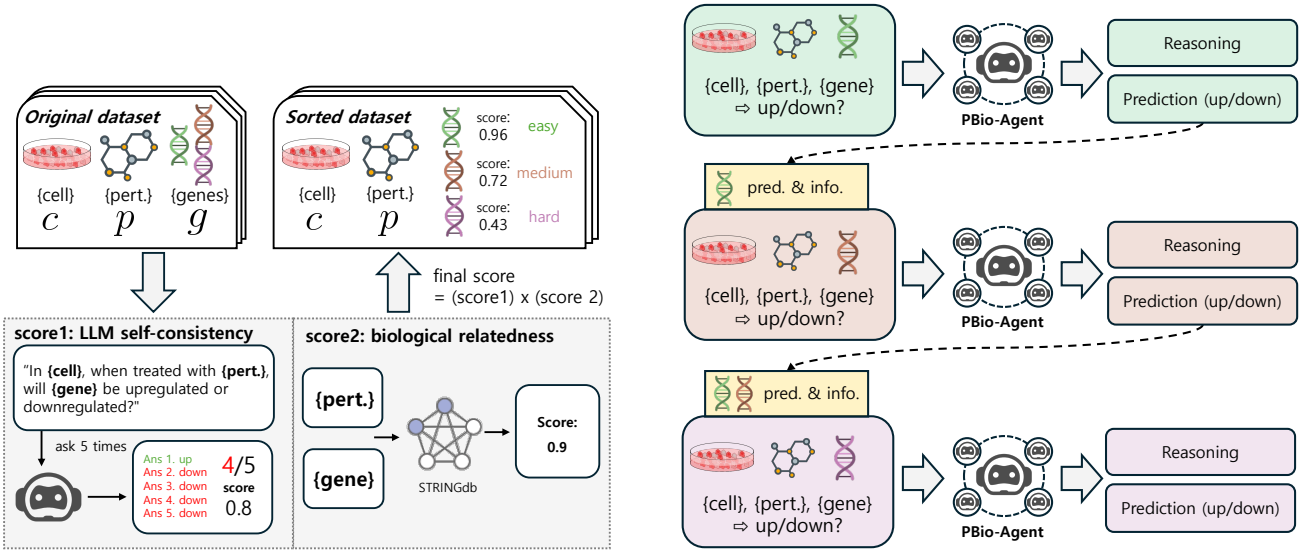
Iterative verification and refinement. Each *judge agent* is an LLM-based critic that checks the integration agent's reasoning trace along one verification axis, returning a categorical verdict together with a natural language rationale. We apply four judges that independently screen for history leakage, target grounding, answer consistency, and logical validity. The pipeline counts problematic verdicts and retries when any judge flags an issue. Formally, we accept a reasoning trace only if the problematic count is zero or once a predefined maximum number of retries is reached.

The judges also provide natural language feedback for later analysis. This feedback helps identify systematic errors such as off-target assumptions or mismatched cell-line context. The retry loop uses the same input but seeks a more coherent reasoning chain. In practice, this filter reduces noisy decisions that stem from shallow explanations.

4. Experiment

4.1. Experimental setting

In this section, we present experimental results of three main tasks: (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.



(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.

Baselines. In LINCSQA benchmark, we compare PBIO-AGENT against two distinct groups of LLMs: (i) 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) and (ii) 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). In 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 & 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

In this section, we evaluate the performance of various language models on the proposed LINCSQA benchmark for gene-level and MoA-level. We first present the results for gene-level regulation direction prediction, which assesses a model's ability to identify transcriptional changes under specific chemical perturbations. Subsequently, we analyze MoA prediction performance, where the models must recognize correct gene regulation direction when correct MoA and cell line pair is given. Specifically, we tested two scenarios: (1) mutation-selective inhibitors in cell lines with versus without the target mutation, and (2) mutation-selective inhibitors in cell lines harboring the target mutation but exhibiting differential drug sensitivity due to bypass resistance mechanisms.

Gene regulation direction prediction. Gene regulation direction prediction evaluates the model's ability to classify whether a specific gene's expression is upregulated or downregulated following a compound treatment in a particular cellular context. For a given compound-cell line pair and a candidate MoA, the model performs binary classification for each target gene, predicting its regulation state as either increased (1) or decreased (0). This task tests the model's fundamental understanding of signaling cascades and downstream transcriptional consequences induced by chemical perturbations. In this setting, we incorporate the prediction and confidence of neural network, *i.e.*, Graph Attention Net-

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.

work (Veličković et al., 2018), as a tool that PBIO-AGENT can utilize.

The performance is reported at Table 1. To ensure biological robustness, we report the performance using binary AUROC scores aggregated over primary organ and tissue level. Notably, LINCSQA consistently outperforms both general-purpose and domain-specific baselines, demonstrating that a multi-agent approach can effectively resolve entangled regulatory signals without the need for additional training. Despite its compact 8B parameter size, our framework achieves superior accuracy compared to significantly larger models, such as Qwen3-30B-A3B and Biomni-R0-32B, as well as established specialist LLMs. This performance gain suggests that progressive reasoning provides a more stable foundation for capturing complex transcriptional dependencies than monolithic architectures, even those with substantially higher parameter counts.

Case 1: BRAF V600E inhibitor (LINCS L1000). *BRAF V600E* is a common cancer-driving point mutation in melanoma that keeps a growth-promoting signaling cascade switched on. Vemurafenib and dabrafenib are mutation-selective inhibitors that bind only this mutant form of BRAF, and leave the unmutated (*wild-type*) protein largely alone. Gene expression signatures for both inhibitors were extracted from the LINCS L1000 dataset (Duan et al., 2016), and consensus signatures were constructed following the procedure in Section 2.1. For the sensitive cell line, we selected A375 (melanoma), which carries BRAF V600E and is well known to respond to BRAF inhibitors.

As negative controls, we selected five BRAF wild-type cell lines from the LINCS touchstone panel: A549 (lung adenocarcinoma), MCF7 (breast cancer), PC3 (prostate cancer), HA1E (kidney), and HEPG2 (hepatocellular carcinoma). Because the inhibitors bind only mutant BRAF, these wild-type cells do not have the relevant target and should therefore *not* produce MoA-consistent transcriptional changes.

The performance results for the BRAF inhibitor case study are reported in Table 2. To evaluate cell-line specificity, we utilize three complementary metrics: target rank, mean gap (difference between A375 performance and overall mean), and relative dominance (percentage by which A375 exceeds the mean). Notably, PBIO-AGENT achieves rank 1 for the A375 cell line in both dabrafenib and vemurafenib as described in Figure 3. It records mean gaps of 0.161 and 0.177 respectively, making it the only model to demonstrate positive gaps across both inhibitors. This translates to relative dominance values of 41.3% and 64.6%, confirming that the framework correctly prioritizes the BRAF V600E-mutant cell line over wild-type controls.

Case 2: KRAS G12C Inhibitor (GSE137912). In this task, we evaluate PBIO-AGENT in a single-cell setting. *KRAS G12C* is another cancer-driving point mutation, common in non-small cell lung cancer (NSCLC), and ARS-1620 is a selective inhibitor of this mutant form. Single-cell

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.

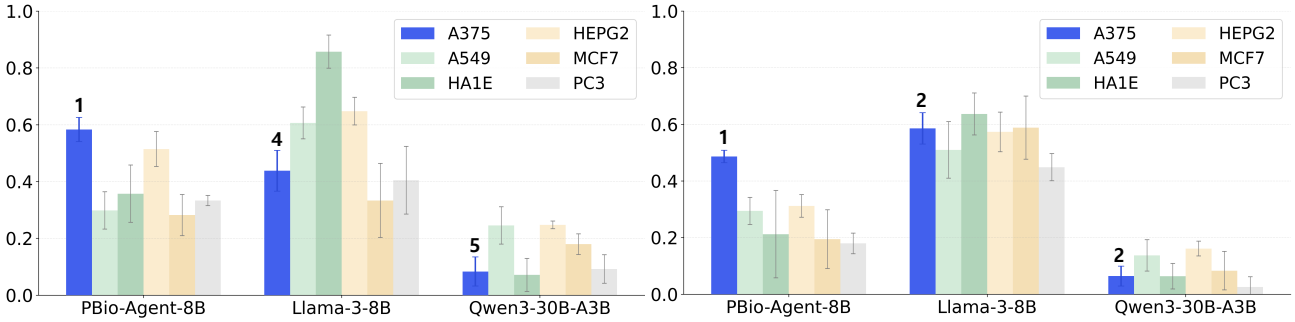


Figure 3. **Agreement ratios and target (A375 cell line) rank comparison for BRAF inhibitors.** Agreement ratios for vemurafenib (left) and dabrafenib (right) with target ranks (numbers above bars) showing A375's ranking among six cell lines. Only PBio-Agent-8B consistently achieves rank 1 in A375 (BRAF 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.

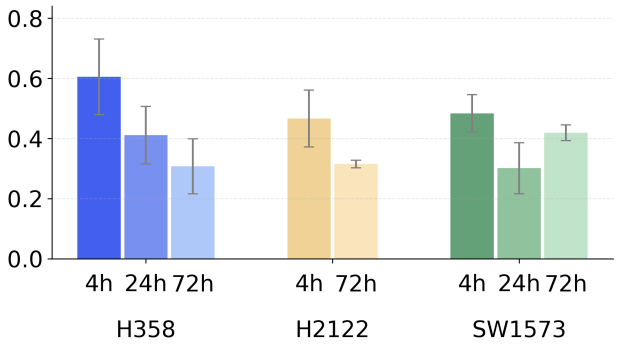


Figure 4. **Agreement ratios of PBIO-AGENT across KRAS G12C-mutants with varying drug sensitivity.** H358 (sensitive), H2122 (intermediate), and SW1573 (resistant) cells were treated with ARS-1620 and evaluated at 4h, 24h, and 72h. Higher agreement in sensitive H358 reflects coherent KRAS inhibition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

RNA-seq data were obtained from GSE137912, comprising KRAS G12C-mutant NSCLC cell lines treated with ARS-1620 (10 μM) at 0, 4, 24, and 72 hours. To build a stable evaluation set, we constructed *pseudo-bulk* profiles by summing single-cell counts within each cell line, which mimics a bulk readout. Ground-truth up/down labels came from differential expression analysis, which flags genes whose expression changes significantly between treated and untreated cells. Preprocessing details are in Section A.1.

We hypothesized that even though both cell lines carry the KRAS G12C mutation, PBIO-AGENT should achieve higher gene-level accuracy in H358, where blocking KRAS produces coherent downstream effects, than in SW1573, where an alternative (*bypass*) pathway makes up for the inhibition, so the transcriptional response no longer reflects the annotated MoA. We visualize the agreement ratios in Figure 4. We evaluated PBIO-AGENT across three KRAS G12C-mutant NSCLC cell lines with differential sensitivity to ARS-1620: H358 (sensitive), H2122 (intermediate), and SW1573 (resistant). Transcriptional responses were measured at 4, 24, and 72 hours post-treatment. At early time points (4h and 24h), agreement ratios showed a clear sensitivity-dependent gradient: H358 achieved the highest values, H2122 exhibited intermediate agreement, and SW1573 showed substantially lower concordance. This pattern directly reflects the underlying biology—H358 exhibits coherent KRAS inhibition-driven transcriptional responses, whereas SW1573's co-occurring PIK3CA mutation enables bypass pathway activation that decouples transcriptional output from the annotated mechanism of action. At 72 hours, agreement ratios declined in the sensitive and intermediate cell lines (H358 and H2122), consistent with the emergence of adaptive responses that diverge from immediate mechanistic signatures. We therefore focus our interpretation on early time points, which more faithfully capture direct MoA-driven transcriptional changes. These results demonstrate that PBIO-AGENT can distinguish genuine drug-target engagement from resistance-mediated pathway bypass without

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.

requiring explicit resistance annotations during inference.

4.3. PerturbQA benchmark

In this section, we evaluate the reasoning capabilities of PBIO-AGENT on the PerturbQA benchmark reported at Table 3. Following the experimental protocols established by Wu et al. (2025a), we conduct a comparative analysis against various baseline models to assess their proficiency in predicting transcriptional responses to genetic perturbations. Gene regulation direction prediction within the PerturbQA framework requires models to identify whether a target gene is upregulated or downregulated following a specific genetic perturbation.

As shown in the Table 3, PBIO-AGENT demonstrates robust performance across multiple cell lines, including K562, RPE1, HepG2, and Jurkat. Specifically, our framework achieves state-of-the-art results on HepG2 (0.67) and Jurkat (0.68). While models like GEARS and SUMMER show competitive performance in specific contexts such as K562 and RPE1, PBIO-AGENT maintains superior or comparable accuracy across the entire suite of tested cell lines. These results indicate that integrating multi-agent reasoning with structured biological knowledge effectively captures the complex dependencies inherent in genetic perturbation data.

4.4. Ablation studies

We ablate four design choices of PBIO-AGENT to confirm that every component of our original setting contributes to the final performance. Concretely, we vary (i) the three *scientist agents* that decompose the reasoning into context, mechanism, and network views, (ii) the four *judges* that verify the integration agent's answer along historical, target, consistency, and logical axes, (iii) the canonical easy-to-hard *difficulty ordering* of the scientists, and (iv) the *maximum verification rounds* budget that controls how many verification rounds are permitted. Table 4 reports per-organ

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.

AUROC on breast, cervix, colon, and lung averaged over three seeds. Across every ablation, the original setting is either the best or statistically tied with the best configuration on the majority of organs, and full per-organ results for all variants are deferred to Section D.

Scientist component ablation. As shown in the *Scientist* block of Table 4, 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: ablating it produces the most uniform decline across organs, while the context and network scientists contribute more localized but still indispensable signal. Although dropping the network scientist marginally helps Breast, the corresponding loss on the remaining organs is substantially larger, indicating that its role is to stabilize reasoning across heterogeneous tissues rather than to specialize in any single one. The fact that no single scientist alone reproduces the full configuration shows that the integration agent cannot recover a missing perspective on its own.

Judge component ablation. The *Judge* block of Table 4 shows that disabling any single judge incurs substantial cost, yet removing all four does not hurt much more than removing one, indicating each judge catches a distinct failure class the others cannot recover. The logical judge is the most critical, with its removal causing the sharpest drop on Lung,

while the target judge is least disruptive but still essential elsewhere. Together, the four axes form a complementary verification stack where historical grounding, target-level evidence, internal consistency, and logical coherence each guard against a different failure mode.

Comparison of ordering strategies. The *Difficulty ordering* block of Table 4 contrasts our easy-to-hard ordering with random shuffling and full reversal. Reversal performs worst, random shuffling second worst, and canonical ordering most reliably across organs. This confirms that difficulty progression serves as an implicit curriculum the multi-agent design relies on, not an arbitrary scheduling convention.

Impact of maximum verification rounds. The *Max rounds* block of Table 4 varies the verification budget around our default value. Reducing the budget cuts off productive corrections before the judges can converge, while increasing it spends additional compute without yielding consistent improvements on any organ. This indicates that the verification stage stabilizes quickly: the default budget already reaches the regime where further retries neither rescue residual errors nor degrade already-correct answers, so the cost of misjudging this hyperparameter is asymmetric only on the lower side.

5. Related Works

Benchmarks for cellular perturbation. The emergence of giga-scale atlases has provided a robust foundation for evaluating model fidelity across diverse chemical and genetic landscapes. Tahoe-100M (Zhang et al., 2025a) study compiled a massive perturbation atlas covering 92 cancer cell lines and thousands of compounds, enabling deep analysis of context-dependent gene functions. To address data consistency issues across disparate studies, scBaseCount (Youngblut et al., 2025) utilized AI agents to uniformly reprocess vast single-cell datasets, establishing a unified pipeline that minimizes analytical variability from different reference genomes.

Perturb-seq (Dixit et al., 2016) revolutionized the single-cell gene perturbation studies by combining pooled genetic screening with single-cell RNA sequencing. Recently, Wu et al. (2025a) introduced a single-cell gene perturbation benchmark for language models. In this work, we present a benchmark for bulk-cell compound perturbations. While critical for drug discovery, this area remains under-explored, and the reasoning capabilities of LLMs in this context have not yet been fully investigated.

Language models for cellular perturbation. Early studies applied the language models by pre-training them on large-scale transcriptomic data. The studies view gene expression profiles as *sentences* and learn the hidden structure of gene networks (Yang et al., 2022; Cui et al., 2024b;a). Alongside these data-driven models, LLMs focused on biomedical literature which can synthesize a lot of prior biological knowledge (Luo et al., 2022; 2023). They predict and explain complex molecular interactions effectively.

Building on this groundwork, PerturbQA (Wu et al., 2025a) benchmarks structured reasoning over experimental single-cell genetic perturbation data. PerturbQA demonstrates LLMs' potential for interpretable reasoning by focusing on causal relationships rather than retrieval. However, unlike PerturbQA's independent target gene prediction, our framework jointly analyzes target gene sets affected by the same perturbation. Furthermore, our confidence-aware ordering propagates high-confidence information to inform subsequent steps.

Curriculum-driven inference. Curriculum-based data ordering enables prior knowledge to guide subsequent predictions, originating from curriculum learning (CL) proposed by (Bengio et al., 2009). This structured ordering is now actively employed in LLMs, where metric-based ordering outperforms simple length-based sorting (Kim & Lee, 2024), and curriculum-based sequencing facilitates domain mastery (Neema et al., 2025; Yang et al., 2024) and optimizes diverse capabilities including preference alignment, multilingual proficiency, and convergence speed (Zhang et al., 2025b; Pucci et al., 2023; Zhang et al., 2025c).

Leveraging these insights, our work adopts this progressive strategy for cell perturbation analysis by sequentially accumulating predicted results at inference time, mirroring how step-by-step gene regulation analysis builds a more accurate understanding of cellular responses. Unlike training-time curriculum learning, our approach dynamically prioritizes genes based on prediction confidence and biological relatedness without requiring model retraining.

6. Conclusion

We present PBIO-AGENT for bulk-cell compound perturbation prediction. The method combines multi-scientist reasoning, context grounded in knowledge graphs, and judge-based verification in one pipeline. This design ties together complementary biological viewpoints and reduces the brittle inference common in single-pass LLM systems. Our evaluation demonstrates the feasibility of combining agentic reasoning with biological constraints for both direction prediction and MoA selection. Larger multi-omics datasets and richer pathway resources will be needed to scale the method. Future work should explore lighter judge ensembles and stronger biological retrieval.

Impact Statement

This work introduces LINCSQA, a novel benchmark for evaluating the reasoning capabilities of large language models (LLMs) in predicting transcriptional responses to chemical perturbations in bulk-cell environments. By proposing PBIO-AGENT, a multi-agent framework that utilizes progressive reasoning and structured biological knowledge, this research enhances the accuracy and interpretability of computational drug discovery. The methodology does not involve human participants or the use of sensitive personal data, and thus we do not anticipate direct ethical risks from the experiments presented. However, the development of more reliable AI systems for predicting biological causalities may significantly accelerate the discovery of new therapeutics and the understanding of disease mechanisms. We advocate for the responsible application of these methods in drug development and encourage practitioners to adhere to established safety, legal, and professional guidelines to mitigate potential societal risks associated with automated biological reasoning.

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2025-02304967, AI Star Fellowship(KAIST), Institute for Information & communications Technology Planning & Evaluation(IITP)grant funded by the Korea government(MSIT) (RS-2019-II190075, Artificial Intelligence Graduate School Program(KAIST)), the High-Performance Computing Support Project, funded by the Government of the Republic of Korea (Ministry of Science and ICT)(RQT-25-070211).

References

Adduri, A. K., Gautam, D., Bevilacqua, B., Imran, A., Shah, R., Naghipourfar, M., Teyssier, N., Ilango, R., Nagaraj, S., Dong, M., et al. 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*, pp. 2025–06, 2025.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

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

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 21(8):1470–1480, 2024a.

Cui, Z., Xu, T., Wang, J., Liao, Y., and Wang, Y. Geneformer: Learned gene compression using transformer-based context modeling. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8035–8039. IEEE, 2024b.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *cell*, 167(7):1853–1866, 2016.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning*, 2024.

Duan, Q., Reid, S. P., Clark, N. R., Wang, Z., Fernandez, N. F., Rouillard, A. D., Readhead, B., Tritsch, S. R., Hodos, R., Hafner, M., et al. L1000cds2: Lincs l1000 characteristic direction signatures search engine. *NPJ systems biology and applications*, 2(1):1–12, 2016.

Giurgiu, M., Reinhard, J., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Ruepp, A. Corum: the comprehensive resource of mammalian protein complexes—2019. *Nucleic acids research*, 47 (D1):D559–D563, 2019.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Huang, K., Zhang, S., Wang, H., Qu, Y., Lu, Y., Roohani, Y., Li, R., Qiu, L., Li, G., Zhang, J., et al. Biomni: A general-purpose biomedical ai agent. *biorxiv*, 2025.

Huttlin, E. L., Ting, L., Bruckner, R. J., Gebreab, F., Gygi, M. P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K., et al. The bioplex network: a systematic exploration of the human interactome. *Cell*, 162(2):425–440, 2015.

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

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

Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., and Dufour, R. Biomistral: A collection of open-source pretrained large language models for medical domains. *Findings of the Association for Computational Linguistics*, 2024.

Levine, D., Rizvi, S. A., Lévy, S., Pallikkavaliyaveetil, N., Zhang, D., Chen, X., Ghadermarzi, S., Wu, R., Zheng, Z., Vrkic, I., et al. Cell2sentence: teaching large language models the language of biology. *International Conference on Machine Learning*, 2024.

Lu, M., Weinberger, E., Kim, C., and Lee, S.-I. Cellclip–learning perturbation effects in cell painting via text-guided contrastive learning. *Advances in Neural Information Processing Systems*, 2025.

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., and Liu, T.-Y. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409, 2022.

Luo, Y., Zhang, J., Fan, S., Yang, K., Wu, Y., Qiao, M., and Nie, Z. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*, 2023.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Märtens, K., Martell, M. B., Prada-Medina, C. A., and Donovan-Maiye, R. Langpert: Llm-driven contextual synthesis for unseen perturbation prediction. In *ICLR 2025 Workshop on Machine Learning for Genomics Explorations*, 2025.

Martin, F. J., Amode, M. R., Aneja, A., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., et al. Ensembl 2023. *Nucleic acids research*, 51(D1):D933–D941, 2023.

Milacic, M., Beavers, D., Conley, P., Gong, C., Gillespie, M., Griss, J., Haw, R., Jassal, B., Matthews, L., May, B., et al. The reactome pathway knowledgebase 2024. *Nucleic acids research*, 52(D1):D672–D678, 2024.

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

Neema, N., Mukherjee, S., Shah, S., Ramakrishnan, G., and Venkatesh, G. From amateur to master: Infusing knowledge into llms via automated curriculum learning. *arXiv preprint arXiv:2510.26336*, 2025.

Perfetto, L., Briganti, L., Calderone, A., Cerquone Perpetuini, A., Iannuccelli, M., Langone, F., Licata, L., Marinkovic, M., Mattioni, A., Pavlidou, T., et al. Signor: a database of causal relationships between biological entities. *Nucleic acids research*, 44(D1):D548–D554, 2016.

Pucci, G., Ranaldi, L., and Zanzotto, F. M. Are all languages equal? curriculum learning over different languages. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, pp. 351–360, 2023.

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

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

Singh, G., Wehling, L., Mulyadi, A. W., Sreenath, R. H., Klabunde, T., Andreani, T., and McCloskey, D. 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*, 2025.

Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A. L., Fang, T., Doncheva, N. T., Pyysalo, S., et al. 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, 2023.

Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., Cohan, A., and Gerstein, M. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 599–621, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.33. URL https://aclanthology.org/2024.findings-acl.33/.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. Graph attention networks. *International Conference on Learning Representations*, 2018.

Wang, E., Schmidgall, S., Jaeger, P. F., Zhang, F., Pilgrim, R., Matias, Y., Barral, J., Fleet, D., and Azizi, S. Txgemma: Efficient and agentic llms for therapeutics. *arXiv preprint arXiv:2504.06196*, 2025.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *International Conference on Learning Representations*, 2023.

Wenkel, F., Tu, W., Masschelein, C., Shirzad, H., Eastwood, C., Whitfield, S. T., Bendidi, I., Russell, C., Hodgson, L., Mesbahi, Y. E., et al. Txpert: Leveraging biochemical relationships for out-of-distribution transcriptomic perturbation prediction. *arXiv preprint arXiv:2505.14919*, 2025.

Wu, M., Littman, R., Levine, J., Qiu, L., Biancalani, T., Richmond, D., and Huetter, J.-C. Contextualizing biological perturbation experiments through language. *International Conference on Learning Representations*, 2025a.

Wu, Y., Wershof, E., Schmon, S. M., Nassar, M., Osiński, B., Eksi, R., Yan, Z., Stark, R., Zhang, K., and Graepel, T. Perturbench: Benchmarking machine learning models for cellular perturbation analysis. *Advances in Neural Information Processing Systems*, 2025b.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yang, F., Wang, W., Wang, F., Fang, Y., Tang, D., Huang, J., Lu, H., and Yao, J. 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, 2022.

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

Youngblut, N. D., Carpenter, C., Prashar, J., Ricci-Tam, C., Ilango, R., Teyssier, N., Konermann, S., Hsu, P. D., Dobin, A., Burke, D. P., et al. scbasecount: an ai agent-curated, uniformly processed, and continually expanding single cell data repository. *bioRxiv*, pp. 2025–02, 2025.

Zhang, J., Ubas, A. A., de Borja, R., Svensson, V., Thomas, N., Thakar, N., Lai, I., Winters, A., Khan, U., Jones, M. G., et al. Tahoe-100m: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *BioRxiv*, pp. 2025–02, 2025a.

Zhang, X., Liangyu, X., Duan, F., Zhou, Y., Wang, S., Weng, R., Wang, J., and Cai, X. Preference curriculum: Llms should always be pretrained on their preferred data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21181–21198, 2025b.

Zhang, Y., Mohamed, A., Abdine, H., Shang, G., and Vazirgiannis, M. Beyond random sampling: Efficient language model pretraining via curriculum learning (2025). *URL https://arxiv. org/abs/2506.11300*, 2025c.

A. Dataset

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.

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 $|\log_2$ 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.

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. Cancer type

Model	Breast	Cervical	Leukemia	Melanoma	Lung	Prostate	Colorectal	Lymphoma
Generalist LLM								
Llama3-8B	0.53±0.00	0.49±0.00	0.50±0.06	0.53±0.00	0.58±0.00	0.50±0.00	0.53±0.00	0.50±0.06
DeepSeek-R1-Distill-Llama-8B	0.49±0.00	0.51±0.00	0.51±0.04	0.50±0.00	0.50±0.00	0.52±0.00	0.46±0.00	0.51±0.04
Mistral-Small-3.2-24B	0.52±0.00	0.50±0.00	0.46±0.07	0.53±0.00	0.55±0.00	0.52±0.00	0.47±0.00	0.46±0.07
Qwen3-30B-A3B	0.62±0.00	0.53±0.00	0.50±0.04	0.60±0.00	0.58±0.00	**0.66±0.00**	0.43±0.00	0.50±0.04
Specialist LLM								
BioMistral-7B	0.51±0.00	0.49±0.00	0.48±0.02	0.51±0.00	0.49±0.00	0.50±0.00	0.51±0.00	0.48±0.02
BioMedGPT-LM-7B	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00
TxGemma-27B	0.50±0.00	0.48±0.00	0.47±0.01	0.53±0.00	0.54±0.00	0.56±0.00	0.45±0.00	0.47±0.01
Biomni-R0-32B	0.55±0.00	0.53±0.00	0.48±0.09	0.56±0.00	0.55±0.00	0.58±0.00	0.48±0.00	0.48±0.09
Ours								
PBIO-AGENT-8B	**0.69±0.00**	**0.64±0.00**	**0.53±0.02**	**0.81±0.00**	**0.68±0.00**	0.56±0.00	**0.58±0.00**	**0.53±0.02**

Table 6. AUROC on compound perturbation QA across 8 cancer types.

C.2. 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 7. 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 8 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.

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

D.2. Ablation study on judges

Table 9 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.

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

D.3. Impact of biological knowledge graph

Table 10 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.

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

D.4. Impact of multi-agent reasoning and judges

Table 11 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.

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

D.5. Effect of maximum verification rounds

Table 12 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.

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 12. AUROC on compound perturbation QA across 8 cancer types under different retry budgets.

D.6. Analysis on difficulty ordering

Table 13 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**
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 13. AUROC on compound perturbation QA across 8 cancer types under different scientist / evidence orderings.

D.7. Comparison with more agentic frameworks

Table 14 compares PBIO-AGENT against four published multi-agent baselines: AIAgents4Pharma (Singh et al., 2025) , MAD (Du et al., 2024), MedAgents (Tang et al., 2024), Self-Refine (Madaan et al., 2023). All run has the same 8B backbone on the same task. PBIO-AGENT wins on the large majority of organ categories and remains the most reliable across the panel, while the alternative frameworks cluster well below it.

Method	Bone marrow	Breast	Cervix	Colon	Lung	Periph. blood	Prostate	Skin
Ours (Full pipeline)	**0.52±.02**	**0.69±.01**	**0.64±.02**	**0.58±.09**	**0.68±.03**	**0.55±.02**	0.56±.02	**0.81±.02**
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 (Multi-Agent Debate)	0.48±.07	0.50±.01	0.53±.01	0.50±.04	0.52±.01	0.52±.06	0.50±.03	0.56±.04
MedAgents	0.45±.02	0.54±.05	0.59±.04	0.46±.02	0.52±.04	0.50±.01	0.54±.03	0.57±.02
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

Table 14. AUROC on compound perturbation QA across 8 cancer types comparing our full pipeline against alternative multi-agent / refinement baselines (same backbone, same task).

E. Computational cost analysis

Table 15 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. LLM calls per sample	14.1
Avg. wall-clock time per sample	131.9s

Table 15. **Computational cost analysis of PBIO-AGENT.**

F. Agents prompts

Table 16. **Prompts of agents in PBIO-AGENT**

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)-¿ (Target-Gene)
	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?

Table 17. **Prompts of agents in PBIO-AGENT**

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}

Table 18. **Prompts of agents in PBIO-AGENT**

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?

Table 19. **Prompts of agents in PBIO-AGENT**

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).