PERTURBOS: AN END-TO-END AGENTIC FRAMEWORK FOR PREDICTION-GUIDED SINGLE-CELL PERTURBATION DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Identifying interventions that reshape cellular behavior can reveal disease mechanisms and guide therapeutic research, but testing large genetic and chemical spaces with single-cell assays is costly. Existing perturbation datasets and predictive models offer a way to focus these experiments. However, their availability alone does not establish whether they cover the interventions and cellular context specified by a new discovery question. Using resources without assessing that fit can mislead intervention selection. To turn existing resources into discovery-specific guidance, we introduce PerturbOS, an end-to-end agentic framework that assesses resource suitability before prioritizing interventions. Given a natural-language question, PerturbOS selects and executes a suitable analysis workflow, then uses its quantitative output alongside biological knowledge to refine candidate interventions. In retrospective drug prioritization for increasing TOP2A expression in A549 cells, 49% of PerturbOS's shortlisted drugs were among those with the largest measured increases, compared with 9% for an agentic baseline without predictive feedback. PerturbOS also outperformed LLM baselines in selecting appropriate workflows across 37 biological scenarios. By improving both workflow selection and intervention prioritization, PerturbOS offers a foundation for more systematic, question-driven biological discovery guided by existing perturbation resources.

1 INTRODUCTION

A central challenge in biology and drug discovery is to identify gene or chemical interventions that reshape cellular behavior, revealing disease mechanisms and informing therapeutic strategies (Rood et al., 2024). However, single-cell experiments are difficult to scale across either genetic or chemical intervention spaces. The human genome contains approximately 19,400 protein-coding genes, whereas representative targeted Perturb-seq experiments assay only tens to hundreds of genes at a time (Frankish et al., 2021; Yao et al., 2024). Chemical perturbation studies face a parallel constraint: under a conventional non-multiplexed design, testing a single drug against a control with three replicates costs approximately US$9,900 in library reagents alone (De Simone et al., 2025). Exhaustive experimental screening is therefore costly and often impractical. Substantial progress has been made in addressing these constraints: targeted and multiplexed assays expand screening throughput (Schraivogel et al., 2020; Srivatsan et al., 2020); predictive models estimate unmeasured perturbation responses (Roohani et al., 2024); and sequential methods prioritize interventions for subsequent testing (Roohani et al., 2025; Hao et al., 2025).

The first interventions selected for a new study are especially consequential: they consume scarce experimental resources and determine which measurements will be available to guide later rounds. Although agentic systems such as BioDiscoveryAgent (Roohani et al., 2025) and PerTurboAgent (Hao et al., 2025) can refine recommendations using feedback from earlier experimental rounds, that feedback is unavailable when the first interventions are chosen. Existing perturbation studies could provide quantitative guidance at this stage, provided their data are relevant to the new biological question and can be used with an appropriate predictive model. Establishing this match is difficult. Datasets differ in intervention coverage, cellular context, structure, and metadata (Peidli et al., 2024), while models such as scGen (Lotfollahi et al., 2019), chemCPA (Hetzel et al., 2022), GEARS (Roohani et al., 2024), and STATE (Adduri et al., 2026) have distinct input requirements, assumptions, and generalization

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

capabilities (Wei et al., 2026). Together, these differences make it difficult to identify a suitable source of evidence for a new biological question and use its results to prioritize interventions (Wei et al., 2026; Ahlmann-Eltze et al., 2025). This creates a need for an approach that assesses whether existing resources can address a new question and, when they can, uses their outputs to guide the first intervention choices.

To meet this need, we present **PerturbOS**, an agentic framework that transforms a natural-language biological question into evidence-grounded intervention prioritization. This goal is achieved through four stages: scenario construction, registry-grounded dataset–model routing, validated execution, and iterative discovery. To make resource suitability assessable, a scenario agent translates the question into explicit requirements for the intervention, cellular context, prediction target, and generalization setting. To compare heterogeneous resources against these requirements, agent-managed, evidence-backed registries capture dataset characteristics alongside model capabilities and input requirements. Routing agents use this information to distinguish questions answerable from existing observations from those requiring prediction, selecting a suitable dataset–model pair in the latter case. To improve candidate selection with quantitative feedback, an iterative discovery agent combines the resulting analysis or predictions with biological knowledge, updating its hypotheses and prioritizing interventions likely to produce the requested effect. By drawing on existing perturbation datasets and models, PerturbOS can provide quantitatively informed recommendations during experimental planning without collecting new task-specific wet-lab measurements.

In summary, our main contributions are:

- **Problem formulation.** We highlight an upstream gap in agentic perturbation discovery: the availability of perturbation datasets and models does not establish whether they are suitable for a particular biological question. Mismatched resources can mislead quantitative guidance. We formulate resource suitability as a question-conditioned problem requiring biological relevance, technical compatibility, and coverage of the requested intervention and cellular context before quantitative prioritization.
- **Methodological design.** We introduce PerturbOS, an end-to-end agentic framework that turns existing perturbation resources into question-specific quantitative guidance for intervention prioritization. It makes resource suitability part of discovery and uses quantitative feedback only when evidence supports the biological question, grounding candidate selection in more than LLM recommendations alone.
- **Evaluation and analysis.** Across retrospective discovery and routing benchmarks, PerturbOS improves intervention prioritization and workflow selection. Its top-60 drug hit rate reaches 46–53% across LLM backbones, versus 4–16% without predictive feedback. It also identifies appropriate analysis routes and suitable dataset–model pairs with greater accuracy than LLM baselines. Together, these results demonstrate its ability to turn existing perturbation resources into quantitative guidance for intervention prioritization.

2 Related Work

Single-cell perturbation-response prediction. A growing range of models predicts how cellular expression changes following perturbation. scGen and CPA learn perturbation effects in latent or compositional spaces (Lotfollahi et al., 2019; 2023), while chemCPA (Hetzel et al., 2022) incorporates chemical structure to support prediction for unseen compounds. GEARS (Roohani et al., 2024) focuses on genetic perturbations using prior knowledge of gene relationships, whereas STATE (Adduri et al., 2026) predicts population-level responses across chemical, genetic, and cytokine perturbations. These models make different assumptions about the perturbation modality, available annotations, training data, and generalization setting. Their performance also varies considerably with the dataset, biological context, evaluation split, and metric used. This dependence is reflected in PerturBench (Wu et al., 2026) and the large-scale evaluations of Wei et al. (2026) and Li et al. (2026), which compare methods across multiple datasets, unseen-context and unseen-perturbation splits, and expression-, differential-expression-, and distribution-level metrics. Together, these benchmarks show that model suitability is context-dependent. PerturbOS builds on this insight by coordinating existing predictors as components of a broader perturbation-discovery process.

Agentic virtual-cell systems and experimental design. Existing perturbation-focused agentic systems largely pursue two directions: building response predictors and recommending interventions (Table 1). CellForge (Tang et al., 2025) uses collaborating agents to design and implement a predictor for a supplied single-cell dataset and research objective, while HarmonyCell (Huang et al., 2026) harmonizes inconsistent dataset metadata and searches for architectures suited to distribution shifts. Both advance task-specific predictor construction, but begin with a dataset rather than selecting suitable resources for a biological question. BioDiscoveryAgent (Roohani et al., 2025) recommends genetic perturbations using biological knowledge and experimental feedback, without using quantitative response predictions to screen candidates before testing. PerTurboAgent (Hao et al., 2025) adds model-based guidance through a lightweight surrogate, but builds that surrogate from measurements collected in preceding experimental rounds. Together, these systems leave unresolved how to identify a suitable source of evidence for a new biological question and use its quantitative results to prioritize interventions before new experimental measurements are available.

Table 1: **Scope comparison of related agentic systems.** PerturbOS combines perturbation design with registry search, model selection, direct evidence, and an executable workflow; the compared systems cover only subsets of these capabilities.

System	Perturbation design	Registry search	Model selection	Model synthesis	Direct evidence	Executable workflow
BioDiscoveryAgent (Roohani et al., 2025)	✓	–	–	–	–	○
PerTurboAgent (Hao et al., 2025)	✓	–	–	–	–	○
CellForge (Tang et al., 2025)	–	–	–	✓	–	✓
HarmonyCell (Huang et al., 2026)	–	–	–	✓	–	✓
PerturbOS	✓	✓	✓	–	✓	✓

✓: central capability; ○: partial or supporting capability; –: outside the stated scope.

3 METHODOLOGY

PerturbOS formulates perturbation discovery as a question-conditioned resource-suitability problem. Given a biological question and optional user-specified control cells, PerturbOS determines whether the available resources in its agent-managed registry can support the requested analysis before using them to prioritize interventions. PerturbOS addresses this through three components: scenario-grounded routing, model preparation when prediction is needed, and iterative discovery guided by perturbation effects. Figure 1 provides an overview, with details in Sections 3.1 to 3.3.

3.1 SCENARIO-GROUNDED ROUTING

Scenario agent. The scenario agent translates a research question into a structured description of its species, cell context, perturbation modality and candidates, target phenotype, desired output, and required generalization regime. It preserves explicit constraints, records the provenance of inferred fields, and interactively asks the user to clarify unresolved ambiguities. This produces a question-specific structured specification to better assess dataset relevance and model suitability.

Dataset and model registry agents. To identify resources that meet these requirements, PerturbOS maintains agent-managed registries of perturbation datasets and predictive models. Dataset and model registry agents inspect publications, repositories, GEO studies, or user-provided descriptions to create records for public or private resources. Dataset records specify species, cell context, modality, measured perturbations, and available artifacts; model records specify data requirements, generalization regimes, and implementation details. Each property is linked to its supporting source.

Routing agent. The routing agent combines the scenario with the registry and returns one of three outcomes. It selects a *direct* route when relevant observations support the requested analysis without a predictive model. Otherwise, it considers an *indirect* route, translating gaps between the scenario and available data into required model capabilities, such as unseen-perturbation, cross-context, or cross-batch prediction. If neither route is supported, the outcome is *unsupported*.

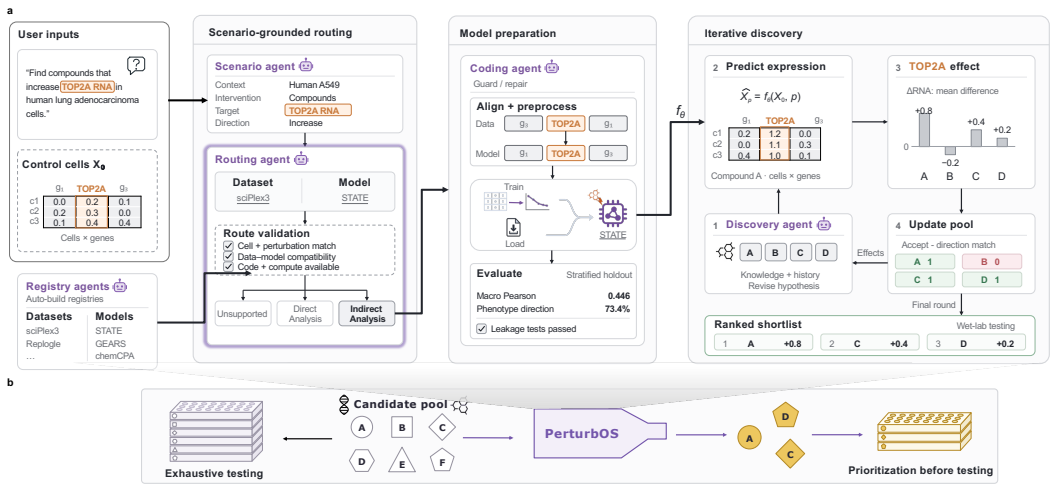

Figure 1: **PerturbOS overview. (a)** The user supplies a research question and optional control-cell data for scenario-grounded routing. Registry agents provide dataset and model records; the scenario agent identifies the intervention, cell context, and target; and the routing agent validates the choice of unsupported, direct, or indirect analysis. For an indirect route, model preparation aligns and preprocesses the data, trains the selected model, and evaluates it on a held-out set. During iterative discovery, the discovery agent proposes compounds, the model predicts their expression responses, and estimated target-gene effects guide updates to the candidate pool and subsequent recommendations. The process produces a ranked shortlist. **(b)** PerturbOS prioritizes candidates from a larger pool before testing, in contrast to exhaustive testing of the full pool.

For indirect routing, datasets and models are assessed jointly for the particular scenario. The agent evaluates biological priors, such as parent–subtype relationships, while deterministic checks exclude pairs of datasets and models with incompatible modalities, feature spaces, or unsupported generalization capabilities. Eligible pairs of datasets and models are then ranked by biological match, model compatibility, evidence strength, and computational efficiency. PerturbOS then checks data artifacts and model adapters in rank order and selects the highest-ranked executable pair. Section A provides more details regarding the scenario representation, provenance constraints, registry validation, and joint-pair ranking procedure.

3.2 Model Preparation

Once an indirect route is selected, PerturbOS creates an execution handoff specifying the dataset–model pair and the details needed for model preparation. The workflow is revalidated against this handoff before execution. Model-specific adapters provide a common interface for data preparation, training, validation, held-out evaluation, and inference. If execution fails, a coding agent diagnoses and repairs the implementation. The successful checkpoint is evaluated on held-out data and its performance metrics are reported to the user before predictions are made for iterative prioritization.

3.3 Iterative Discovery

PerturbOS extends BioDiscoveryAgent (BDA) (Roohani et al., 2025) with model-derived perturbation feedback and a persistent discovery history. At each round, the agent uses the complete history to recommend a batch from the permitted candidate universe, excluding all previously explored candidates.

At round t, the discovery agent recommends a batch R_t of candidate perturbations. The selected workflow estimates a task-relevant effect $\widehat{\delta}_t(p)$ for each $p \in R_t$. A deterministic binary rule a_q, defined by the research question q, accepts candidates whose estimated effects meet the objective and rejects invalid outputs. Starting from $C_0 = \varnothing$, the accepted pool is updated as

$$C_t = C_{t-1} \cup \left\{ p \in R_t : a_q\left(p, \widehat{\delta}_t(p)\right) = 1 \right\}. \tag{1}$$

After each round, PerturbOS stores the recommendations, predicted effects, acceptance decisions, reflection, and cumulative accepted pool, and provides this complete history to subsequent rounds. The exact recommendation and history updates are given in Section A. After the final round, the accepted pool C_T is locked. If the pool contains no more candidates than the user-specified maximum shortlist size, all accepted candidates are returned. Otherwise, a final LLM-based selection chooses that number of candidates from C_T using the accumulated discovery history.

4 Experimental Setups

We evaluate PerturbOS at three levels, ordered from the overall discovery outcome to its supporting components. First, a retrospective TOP2A drug prioritization study evaluates the performance of the pipeline's final shortlists. Second, a held-out perturbation-response study tests whether the executed workflow provides an informative signal for prioritization. Third, a routing benchmark tests whether the system selects the correct analysis route and, when prediction is needed, an appropriate dataset–model pair. Experimental environment and implementation settings are provided in Section B.

4.1 TOP2A Drug Prioritization

Setup and baselines. We evaluate the task "Identify drugs likely to increase TOP2A RNA expression in human A549 lung adenocarcinoma cells." PerturbOS routes the question to a dataset–STATE workflow and prepares the model as described in Section 4.2. Its predictions are then used for TOP2A drug prioritization. Across the evaluated LLM backbones, we run five replicates per strategy: PerturbOS with matched predictive feedback, BDA with feedback permuted across drugs, iterative BDA without feedback, and one-round BDA. Five uniform-random shortlists are shared across backbones. Each iterative strategy recommends 10 unique drugs per round for five rounds. Prediction-feedback arms return up to 20 accepted drugs; no-feedback BDA, one-round BDA, and uniform random return exactly 20. Every observed shortlist has 20 drugs except one DeepSeek–PerturbOS shortlist of 18.

Metrics. Within each study, all trajectories and final shortlists were locked before scoring against measured responses. Hit rate is the fraction of shortlisted drugs among the K drugs with the largest positive measured TOP2A effects, for $K \in \{5, 10, 20, 30, 40, 50, 60\}$ (Section D). The measured-positive fraction is the fraction of shortlisted drugs with measured all-dose effects > 0, regardless of rank. For all-dose effects, we report mean $\pm$ sample SD across five runs using actual shortlist sizes.

4.2 Routed Model Preparation and Held-Out Prediction

Case-study setup. We run PerturbOS on the same request as in Section 4.1 to identify drugs that increase *TOP2A* RNA expression in A549 lung adenocarcinoma cells. *TOP2A* is chosen as it is associated with cell-cycle activity and non-small-cell lung cancer progression (Wu et al., 2024). Model fitting, preprocessing, and checkpoint selection use the E09 source batch (Srivatsan et al., 2020); only F10 control cells are supplied for target-context inference, while treated F10 profiles remain hidden until evaluation. We evaluate transcriptome-wide recovery and TOP2A effect ranking, error, direction, and top-effect overlap across compounds.

4.3 Registry-Grounded Routing

Setup and comparisons. We construct a benchmark to evaluate PerturbOS's routing capability. It contains 37 biological questions, 16 candidate datasets, and three models: GEARS, STATE, and chemCPA. The reference routes comprise four direct, 28 indirect, and five unsupported cases. Hidden annotations specify the reference route, required model capabilities, and the acceptable and preferred models and dataset–model pairs.

We compare PerturbOS with three LLM baselines across different LLM backbones: direct selection from candidate names, source lookup, and a registry-context ablation. Within each backbone, methods share the scenarios and candidate universe; outputs are locked before scoring.

Metrics. The primary metric is routing accuracy across all 37 cases, defined as the proportion assigned the correct category: direct, indirect, or unsupported. We separately evaluate whether the

agent's inferred capability set matches the reference annotations. For the 28 indirect cases, we report the proportion for which the system selects an acceptable dataset–model pair—one that meets the question's biological and technical requirements—and the proportion for which it selects the preferred pair, the reference annotation's best choice among acceptable pairs.

5 Experimental Results and Analysis

We analyze PerturbOS from its final recommendations back to the decisions that support them. We first assess the quality of its TOP2A drug shortlists and the contribution of matched predictive feedback. We then evaluate whether the held-out response predictions provide a useful signal for prioritization. Finally, we examine whether registry-grounded routing selects suitable analysis routes and dataset–model pairs.

5.1 Overall TOP2A Drug Prioritization

PerturbOS improves final shortlist quality. We evaluate performance against the 60 largest positive measured TOP2A effects as the primary reference set. The effect at this cutoff corresponds to approximately 2.79 control-versus-control standard deviations, representing a meaningful increase over control. Across all five LLM backbones, PerturbOS achieves a 46–53% hit rate, compared with 2–20% for single-round BDA (Table 2). Thus, the complete workflow selects substantially more high-effect candidates than initial LLM recommendations alone, supporting quantitatively informed prioritization for subsequent experimental validation.

Table 2: **TOP2A shortlist performance (%).** Top-60 hit rates use the 60 largest positive measured effects in the all-dose view; the measured-positive fraction uses all-dose effects. Values are five-run means ± sample SD. Bold marks the highest mean per backbone and metric.

Strategy	Claude Opus 4.7	Claude Opus 5	GPT-5.6 Sol	GPT-6 Sol	DeepSeek Flash
Top-60 hit rate: all doses (primary)					
BDA permuted feedback	28.0±7.6	21.0±13.9	18.0±6.7	26.0±8.9	17.0±6.7
BDA no feedback	16.0±7.4	9.0±4.2	13.0±9.1	15.0±7.1	4.0±5.5
One-round BDA	10.0±5.0	12.0±5.7	20.0±11.7	10.0±6.1	2.0±2.7
Uniform random	30.0±10.6	30.0±10.6	30.0±10.6	30.0±10.6	30.0±10.6
PerturbOS	**53.0**±8.4	**49.0**±4.2	**47.0**±2.7	**46.0**±2.2	**52.0**±2.7
Measured-positive fraction: all doses					
BDA permuted feedback	57.0±8.4	47.0±16.4	44.0±4.2	51.0±12.9	43.0±8.4
BDA no feedback	33.0±13.5	28.0±9.7	28.0±9.1	27.0±5.7	25.0±3.5
One-round BDA	30.0±3.5	29.0±8.2	41.0±12.9	27.0±7.6	16.0±4.2
Uniform random	71.0±5.5	71.0±5.5	71.0±5.5	71.0±5.5	71.0±5.5
PerturbOS	**97.0**±2.7	**100.0**±0.0	**97.0**±2.7	**100.0**±0.0	**100.0**±0.0

Matched quantitative feedback is critical for improvement. PerturbOS outperforms iterative BDA both without feedback and with permuted feedback across all backbones (Table 2). Its final shortlists also contain a higher fraction of measured-positive drugs (Table 2). Iteration without feedback yields no consistent gain over single-round BDA (Table 2). Permuted feedback performs somewhat better than no feedback, but remains substantially inferior to PerturbOS. These comparisons show that matching quantitative feedback to candidates matters for final shortlist quality.

Performance gains extend across evaluation thresholds. PerturbOS generally separates further from the baselines as the reference set size K increases (Figure 2A–E). For the Claude backbones, however, its advantage is less consistent among the highest-ranked drugs at small K. This pattern is consistent with stronger directional than fine-grained ranking accuracy for the predictive model (Section 5.2), suggesting that prediction quality may limit enrichment for the most extreme responders.

Prior biological knowledge can misguide initial directional prioritization. In the first discovery round, the Claude and DeepSeek Flash backbones produce fewer than 25% measured-positive

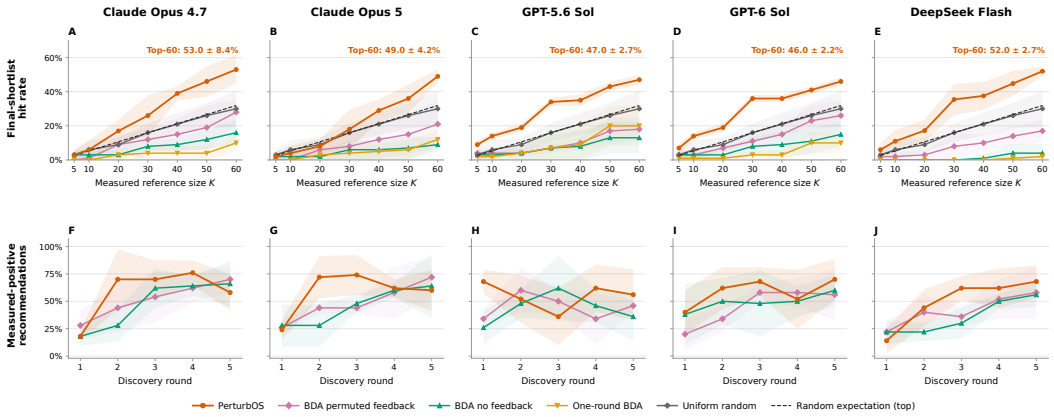

Figure 2: **All-dose TOP2A prioritization across five backbones. A–E:** Final-shortlist hit rates against positive measured top-K sets; dashed lines show $K/188$. **F–J:** Fractions of each round's ten new drugs with measured TOP2A effects > 0, including rejected candidates. Lines show five-run means; bands show ± 1 sample SD, clipped to $[0, 1]$. Uniform random shortlists are shared across backbones.

recommendations, compared with approximately 60% expected under random selection (Figure 2F–J; Table 3). Analysis of their stated rationales suggests that initial recommendations favor biologically related compounds even when the LLM acknowledges the requested direction of TOP2A change. At the primary endpoint, PerturbOS is the only strategy that exceeds uniform random sampling across all backbones (Table 2; Figure 2A–E), underscoring the difficulty of directional prioritization without correctly matched quantitative feedback.

Table 3: **Measured-positive recommendations: round 1 versus rounds 2–5.** Fractions with measured all-dose TOP2A effects > 0, including rejected candidates. Values are five-run means $\pm$ sample SD.

Backbone	Strategy	Round 1 (%)	Rounds 2–5 (%)	Change (pp)
Claude Opus 4.7	BDA permuted feedback	**28.0**±13.0	57.5±7.9	+29.5
	BDA no feedback	18.0±8.4	55.0±5.0	+37.0
	PerturbOS	18.0±4.5	**68.5**±11.0	**+50.5**
Claude Opus 5	BDA permuted feedback	26.0±5.5	54.5±6.0	+28.5
	BDA no feedback	**28.0**±19.2	50.0±3.1	+22.0
	PerturbOS	24.0±8.9	**67.0**±4.5	**+43.0**
GPT-5.6 Sol	BDA permuted feedback	34.0±8.9	47.5±6.8	+13.5
	BDA no feedback	26.0±15.2	48.0±9.4	**+22.0**
	PerturbOS	**68.0**±11.0	**51.5**±8.2	−16.5
GPT-6 Sol	BDA permuted feedback	20.0±14.1	51.5±3.4	**+31.5**
	BDA no feedback	38.0±27.7	52.0±9.9	+14.0
	PerturbOS	**40.0**±20.0	**63.0**±4.5	+23.0
DeepSeek Flash	BDA permuted feedback	**22.0**±8.4	46.5±8.0	+24.5
	BDA no feedback	**22.0**±11.0	39.5±3.3	+17.5
	PerturbOS	14.0±11.4	**59.0**±5.5	**+45.0**

Change is the difference in means (percentage points); bold marks the highest value per backbone and column.

Matched quantitative feedback corrects the direction of selection. For the two Claude backbones and DeepSeek Flash, the measured-positive fraction of PerturbOS recommendations increases by approximately 43–51 percentage points from round 1 to rounds 2–5 (Figure 2F–J; Table 3). GPT-5.6 Sol starts higher, at 68%, and shows no corresponding increase (Table 3). Although later rounds also improve directional selection in some control trajectories, PerturbOS shows the largest immediate

correction among the backbones with weak initial selection (Figure 2F–J), supporting the value of matched predictive feedback.

5.2 Routed Model Evaluation

Held-out predictions recover perturbation-response structure. STATE, trained through PerturbOS on E09, recovered perturbation-response structure in the held-out F10 treatments across pooled doses. Transcriptome-wide predictions achieved a macro-averaged Pearson correlation of 0.775, with a top-100 overlap of 0.759 between genes ranked by predicted and measured absolute effects. Together, these results indicate that the routed workflow captures quantitative response patterns and prominent gene-level changes under batch variation, providing a basis for the subsequent prioritization study.

TOP2A predictions support directional prioritization. Predicted compound effects tracked measured *TOP2A* responses (Spearman $\rho = 0.766$) and distinguished increases from decreases with a balanced accuracy of 0.822. Biological priors may favor plausible interventions that act in the wrong direction, making this directional discrimination important for prioritization. For the strongest increases, the predicted and measured top-10 lists had an overlap of 0.20, compared with 0.053 expected at random. Although this represents nearly fourfold enrichment, the incomplete recovery of the strongest responders may explain why prioritization gains are clearer for larger top-K sets than for the smallest ones (Figure 2A–E).

5.3 Registry-Grounded Routing Improves Route Selection

PerturbOS improves decisions in workflow choice and resource selection. Route accuracy and dataset–model pair accuracy are central to downstream execution: routing determines whether to answer directly, construct a predictive workflow, or mark a question as unsupported, while pair selection determines which dataset–model pair to use for training. Selecting a suitable model alone is insufficient if the dataset cannot support its execution or is biologically incompatible with the question. PerturbOS achieves the highest route accuracy for both Claude backbones, exceeding the strongest baseline by 16.2 percentage points with Opus 4.7 and 5.4 points with Opus 5 (Table 4). Improvements in acceptable dataset–model pair accuracy are larger, reaching 17.9 and 21.4 points, respectively. These results demonstrate gains in both routing and resource selection, with the strongest benefit in identifying suitable dataset–model combinations for downstream execution.

Table 4: **Registry-grounded routing performance (%).** Route and capability use all 37 cases; selection uses 28 indirect cases, counting incorrect routes and invalid choices as failures. Registry-context LLM is the bundled ablation.

Method	Route Accuracy ↑	Capability Exact ↑	Model selection Accept. ↑	Model selection Prefer. ↑	Dataset–model pair Accept. ↑	Dataset–model pair Prefer. ↑
Claude Opus 4.7						
Direct-selection LLM	43.2	16.2	21.4	21.4	21.4	21.4
Source-lookup LLM	40.5	18.9	21.4	21.4	17.9	14.3
Registry-context LLM	45.9	21.6	25.0	25.0	25.0	25.0
PerturbOS	**62.2**	**32.4**	**50.0**	**39.3**	**42.9**	**32.1**
Δ vs. best baseline	+16.2	+10.8	+25.0	+14.3	**+17.9**	**+7.1**
Claude Opus 5						
Direct-selection LLM	43.2	**27.0**	21.4	21.4	21.4	21.4
Source-lookup LLM	45.9	21.6	21.4	21.4	17.9	14.3
Registry-context LLM	48.6	24.3	25.0	25.0	25.0	25.0
PerturbOS	**54.1**	24.3	**50.0**	**46.4**	**46.4**	**35.7**
Δ vs. best baseline	+5.4	−2.7	+25.0	+21.4	**+21.4**	**+10.7**

Higher is better (↑). Accept./Prefer. mean acceptable/preferred selection. Bold marks the best result within each backbone; Δ is PerturbOS minus the strongest baseline for each metric, in percentage points before rounding.

Structured registry information and suitability analysis improve pair selection. The Registry-context LLM achieves higher dataset–model pair accuracy than both direct selection and source

lookup across the two Claude backbones (Table 4). This shows the benefit of presenting relevant dataset and model attributes in a structured form rather than requiring the LLM to extract them from primary sources. PerturbOS further outperforms the Registry-context LLM through suitability analysis, which assesses those attributes against the question's biological and technical requirements. The progression suggests complementary benefits: the registry makes suitable resources easier to identify, while suitability analysis helps select an appropriate pair for execution.

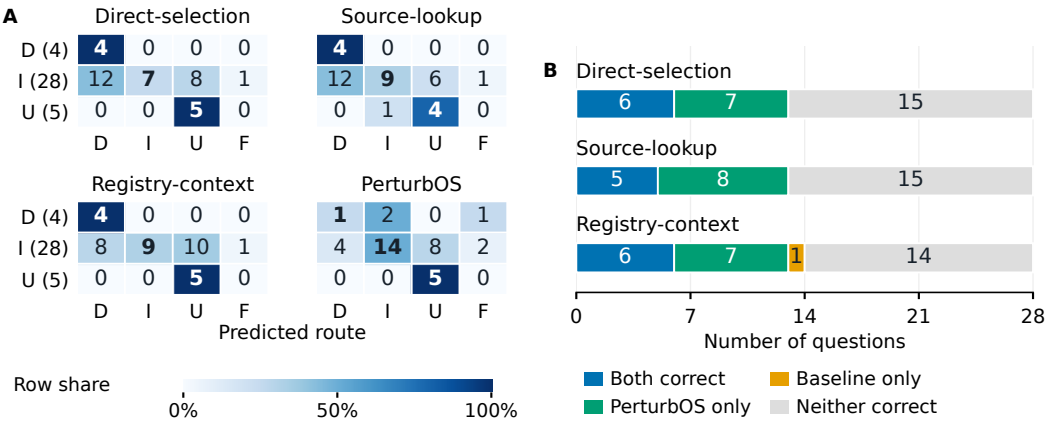

Figure 3: **Opus 5 routing and pair selection. A:** Confusion counts for 37 questions (rows: reference; columns: prediction; D/I/U/F: direct/indirect/unsupported/failed). **B:** PerturbOS versus each baseline on acceptable dataset–model pairs for 28 questions labeled indirect.

PerturbOS helps resolve indirect cases. The Opus 5 confusion matrices show that all three LLM baselines frequently classify questions requiring an indirect workflow as either direct or unsupported (Figure 3A). We suspect that this is because an indirect route requires the additional step of identifying and validating a usable dataset–model pair. PerturbOS correctly routes more indirect cases, although it also misroutes three of the four direct cases; its routing gain should therefore not be read as a uniform improvement across route types. At the pair level, PerturbOS solves nearly every case solved by each baseline, as well as seven or eight cases that the respective baseline misses (Figure 3B). The improvement thus largely extends baseline coverage rather than arising from an entirely different set of successes. Further analysis identified the greatest routing weakness (9 out of 15 cases): matching required gene perturbations to dataset coverage. If all datasets had locally accessible artifacts, the registry agent could extract artifact-level information from H5AD files to enrich registry records and address this weakness.

6 Discussion and Conclusion

Limitations. Our routing benchmark evaluates end-to-end decisions but does not provide step-wise labels for diagnosing errors in unseen-gene perturbation scenarios. To our knowledge, no existing benchmark for question-conditioned perturbation dataset–model routing provides the intermediate annotations needed to distinguish errors in interpreting generalization requirements from errors in matching them to model capabilities.

Future work. We will refine the benchmark by adding expert-validated labels and written justifications for intermediate decisions, including the generalization requirements identified from each question and the capabilities required of a suitable model. These annotations will enable step-wise error analysis and help identify which routing decisions need refinement.

Conclusion. We introduced **PerturbOS**, an end-to-end agentic framework that turns existing perturbation datasets and models into quantitative guidance for intervention prioritization. By linking question-conditioned resource selection, validated execution, and iterative candidate refinement, PerturbOS improves workflow selection and drug prioritization in retrospective evaluations, enabling higher-quality, quantitatively guided choices of initial interventions.

AI USE STATEMENT

We used generative AI tools to draft benchmark questions, provide feedback on research methodology and experiments, clean and reformat datasets, implement methods, summarize existing literature, and support qualitative and thematic data analysis. Within PerturbOS, LLMs propose and refine intervention hypotheses during iterative discovery, as described in Section 3.3. We also used generative AI tools to edit manuscript language. We did not use generative AI to generate biological data, develop theoretical models or conceptual frameworks, translate text, or interpret experimental results. Formulating mathematical claims, providing ingredients for their proofs, and writing proofs were not applicable to this work. We reviewed all AI-assisted work and checked the quality of the LLM-generated benchmark. We take responsibility for the final content, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

PerturbOS aims to support responsible biological discovery by assessing the suitability of existing perturbation data and models before prioritizing interventions. Our experiments use publicly available single-cell datasets and involve no new human participants or wet-lab experiments. We recognize that agent-generated recommendations may be incorrect or misused, especially if predicted responses are treated as established biological effects. Our retrospective results do not establish prospective performance; proposed interventions require expert review and independent experimental validation before use. We also acknowledge the computational and environmental costs of training and inference and seek to reuse existing resources where suitable. We report the system's decisions, evaluation protocol, and limitations to support transparent and reproducible research.

REPRODUCIBILITY STATEMENT

Section 3 and Section A describe scenario construction, joint dataset–model scoring, and discovery-state updates. Section B reports computing infrastructure, software versions, LLM settings, preprocessing, training hyperparameters, and inference artifacts. Section 4 specifies the evaluation tasks, LLM backbones, baselines, candidate budgets, held-out data separation, and metrics. Section C details registry construction, dataset selection and sources, and question construction; Section D provides the formal hit-rate definition. The prioritization study reports means and sample standard deviations across five runs, with trajectories and final shortlists locked before scoring against measured responses. These descriptions support verification of the experimental design and reported comparisons.

REFERENCES

Britt Adamson, Thomas M Norman, Marco Jost, Min Y Cho, James K Nuñez, Yuwen Chen, Jacqueline E Villalta, Luke A Gilbert, Max A Horlbeck, Marco Y Hein, et al. A multiplexed single-cell crispr screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882, 2016.

Abhinav K Adduri, Dhruv Gautam, Beatrice Bevilacqua, Mohsen Naghipourfar, Alishba Imran, Rohan Shah, Noam Teyssier, Rishi Verma, Christopher Carpenter, Basak Eraslan, et al. Predicting cellular responses to perturbation across diverse contexts with state. *Cell*, 2026.

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8): 1657–1661, 2025.

Alexandre F Aissa, Abul BMMK Islam, Majd M Ariss, Cammille C Go, Alexandra E Rader, Ryan D Conrardy, Alexa M Gajda, Carlota Rubio-Perez, Klara Valyi-Nagy, Mary Pasquinelli, et al. Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer. *Nature communications*, 12(1):1628, 2021.

Paul Datlinger, André F Rendeiro, Christian Schmidl, Thomas Krausgruber, Peter Traxler, Johanna Klughammer, Linda C Schuster, Amelie Kuchler, Donat Alpar, and Christoph Bock. Pooled crispr screening with single-cell transcriptome readout. *Nature methods*, 14(3):297–301, 2017.

Marco De Simone, Jonathan Hoover, Julia Lau, Hayley M Bennett, Bing Wu, Cynthia Chen, Hari Menon, Amelia Au-Yeung, Sean Lear, Samir Vaidya, et al. A comprehensive analysis framework for evaluating commercial single-cell rna sequencing technologies. *Nucleic Acids Research*, 53(2): gkae1186, 2025.

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

Chris J Frangieh, Johannes C Melms, Pratiksha I Thakore, Kathryn R Geiger-Schuller, Patricia Ho, Adrienne M Luoma, Brian Cleary, Livnat Jerby-Arnon, Shruti Malu, Michael S Cuoco, et al. Multimodal pooled perturb-cite-seq screens in patient models define mechanisms of cancer immune evasion. *Nature genetics*, 53(3):332–341, 2021.

Adam Frankish, Mark Diekhans, Irwin Jungreis, Julien Lagarde, Jane E Loveland, Jonathan M Mudge, Cristina Sisu, James C Wright, Joel Armstrong, If Barnes, et al. Gencode 2021. *Nucleic acids research*, 49(D1):D916–D923, 2021.

Minsheng Hao, Yongju Lee, Hanchen Wang, Gabriele Scalia, and Aviv Regev. Perturboagent: An llm-based agent for designing iterative perturb-seq experiments. In *MLCB*, pp. 44–64, 2025.

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

Wenxuan Huang, Mingyu Tsoi, Yanhao Huang, Xinjie Mao, Xue Xia, Hao Wu, Jiaqi Wei, Yuejin Yang, Lang Yu, Cheng Tan, et al. Harmonycell: Automating single-cell perturbation modeling under semantic and distribution shifts. *arXiv preprint arXiv:2603.01396*, 2026.

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

Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular systems biology*, 19(6):MSB202211517, 2023.

James M McFarland, Brenton R Paolella, Allison Warren, Kathryn Geiger-Schuller, Tsukasa Shibue, Michael Rothberg, Olena Kuksenko, William N Colgan, Andrew Jones, Emily Chambers, et al. Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nature communications*, 11(1):4296, 2020.

Thomas M Norman, Max A Horlbeck, Joseph M Replogle, Alex Y Ge, Albert Xu, Marco Jost, Luke A Gilbert, and Jonathan S Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019.

Efthymia Papalexi, Eleni P Mimitou, Andrew W Butler, Samantha Foster, Bernadette Bracken, William M Mauck III, Hans-Hermann Wessels, Yuhan Hao, Bertrand Z Yeung, Peter Smibert, et al. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nature genetics*, 53(3):322–331, 2021.

Stefan Peidli, Tessa D Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J Schumacher, Jake P Taylor-King, Debora S Marks, et al. scperturb: harmonized single-cell perturbation data. *Nature Methods*, 21(3):531–540, 2024.

Joseph M Replogle, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575, 2022.

Jennifer E Rood, Anna Hupalowska, and Aviv Regev. Toward a foundation model of causal cell and tissue biology with a perturbation cell and tissue atlas. *Cell*, 187(17):4520–4545, 2024.

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

Yusuf Roohani, Andrew Lee, Qian Huang, Jian Vora, Zachary Steinhart, Kexin Huang, Alexander Marson, Percy Liang, and Jure Leskovec. Biodiscoveryagent: An ai agent for designing genetic perturbation experiments. In *International Conference on Learning Representations*, volume 2025, pp. 26417–26466, 2025.

Daniel Schraivogel, Andreas R Gschwind, Jennifer H Milbank, Daniel R Leonce, Petra Jakob, Lukas Mathur, Jan O Korbel, Christoph A Merten, Lars Velten, and Lars M Steinmetz. Targeted perturb-seq enables genome-scale genetic screens in single cells. *Nature methods*, 17(6):629–635, 2020.

Eric Shifrut, Julia Carnevale, Victoria Tobin, Theodore L Roth, Jonathan M Woo, Christina T Bui, P Jonathan Li, Morgan E Diolaiti, Alan Ashworth, and Alexander Marson. Genome-wide crispr screens in primary human t cells reveal key regulators of immune function. *Cell*, 175(7):1958–1971, 2018.

Sanjay R Srivatsan, José L McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A Pliner, Dana L Jackson, Riza M Daza, Lena Christiansen, et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020.

Artur Szałata, Andrew Benz, Robrecht Cannoodt, Mauricio Cortes, Jason Fong, Sunil Kuppasani, Richard Lieberman, Tianyu Liu, Javier A Mas-Rosario, Rico Meinl, et al. A benchmark for prediction of transcriptomic responses to chemical perturbations across cell types. *Advances in Neural Information Processing Systems*, 37:20566–20616, 2024.

Xiangru Tang, Zhuoyun Yu, Jiapeng Chen, Yan Cui, Daniel Shao, Weixu Wang, Fang Wu, Yuchen Zhuang, Wenqi Shi, Zhi Huang, et al. Cellforge: agentic design of virtual cell models. *arXiv preprint arXiv:2508.02276*, 2025.

Zhiting Wei, Yiheng Wang, Yicheng Gao, Shuguang Wang, Ping Li, Duanmiao Si, Yuli Gao, Siqi Wu, Danlu Li, Kejing Dong, et al. Benchmarking algorithms for generalizable single-cell perturbation response prediction. *Nature Methods*, 23(2):451–464, 2026.

Jiatao Wu, Wenjuan Li, Xueying Zhang, Fan Shi, Qianhao Jia, Yufei Wang, Yuqi Shi, Shiwu Wu, and Xiaojing Wang. Expression and potential molecular mechanism of top2a in metastasis of non-small cell lung cancer. *Scientific Reports*, 14(1):12228, 2024.

Yan Wu, Esther Wershof, Sebastian Schmon, Marcel Nassar, Błażej Osiński, Ridvan Eksi, Zichao Yan, Rory Stark, Kun Zhang, and Thore Graepel. Perturbench: Benchmarking machine learning models for cellular perturbation analysis. *Advances in Neural Information Processing Systems*, 38, 2026.

Douglas Yao, Loic Binan, Jon Bezney, Brooke Simonton, Jahanara Freedman, Chris J Frangieh, Kushal Dey, Kathryn Geiger-Schuller, Basak Eraslan, Alexander Gusev, et al. Scalable genetic screening for regulatory circuits using compressed perturb-seq. *Nature biotechnology*, 42(8): 1282–1295, 2024.

Ronghui Zhu, Emma Dann, Jun Yan, Justine Reyes Retana, Ryunosuke Goto, Reese C Guitche, Lillian Brixi, Mineto Ota, Austin Hartman, Theodore L Roth, et al. Genome-scale perturb-seq in primary human cd4+ t cells maps context-specific regulators of t cell programs and human immune traits. *Cell*, 2025.

A Methodological Details

Scenario representation. For a research question q and optional user-provided controls C, the scenario-construction agent $\mathcal{A}_{\text{scenario}}$ returns a structured scenario s and provenance map π:

$$(s, \pi) = \mathcal{A}_{\text{scenario}}(q, C). \tag{2}$$

The scenario is

$$s = \left(s_{\text{species}}, s_{\text{cell}}, s_{\text{modality}}, s_{\text{candidates}}, s_{\text{phenotype}}, s_{\text{output}}, s_{\text{generalization}}\right), \tag{3}$$

whose components describe the species, cell context, perturbation modality, candidate perturbations, target phenotype, desired output, and required generalization regime. Scenario construction is subject to

$$\text{Preserve}(s, q) = 1, \qquad \pi(f) \neq \varnothing \quad \forall f \in \text{Added}(s; q). \tag{4}$$

The first condition preserves the question's intent and explicit constraints; the second requires every added or inferred field to identify its supporting source. Registry identifiers and associated artifacts are validated before an entry is added, and unsupported properties are recorded as unknown.

Joint routing score. For indirect routing, the candidate set is

$$\mathcal{P}(s) = \mathcal{D}_s \times \mathcal{M}, \tag{5}$$

where $\mathcal{D}_s \subseteq \mathcal{D}$ contains datasets relevant to s, and $\mathcal{M}$ is the model registry. Eligible pairs are ranked by

$$S(p) = 0.50B(p) + 0.30K(p) + 0.15E(p) + 0.05F(p), \tag{6}$$

where B, K, E, and F denote biological match, model compatibility, evidence strength, and computational efficiency. Each component uses a fixed rubric scaled from 0 to 100.

Discovery-state updates. Let $\mathcal{U}$ be the permitted candidate universe and H_{t-1} the complete history before round t. The agent recommends previously unexplored candidates as

$$R_t = \mathcal{A}_{\text{BDA}}\left(q, H_{t-1}, \mathcal{U} \setminus \bigcup_{\tau<t} R_\tau\right). \tag{7}$$

After prediction and the acceptance update in Equation (1), PerturbOS records

$$H_t = H_{t-1} \parallel \left(R_t, \widehat{\boldsymbol{\delta}}_t, D_t, \phi_t, C_t\right), \tag{8}$$

where D_t contains the decisions and ϕ_t the agent's reflection.

B Experimental Setup Details

B.1 Computing Infrastructure

System implementation. PerturbOS separates registry management, scenario construction and routing, model execution, and discovery into Python packages, exposed through a FastAPI service. Registry records, execution handoffs, and run manifests use structured JSON; expression matrices and predictions use AnnData/H5AD. The orchestration environment specifies Python 3.11. LLM agents access remote model endpoints, while predictive-model training and inference run locally in Docker containers. The evaluated executable adapter is STATE; GEARS and chemCPA also appear as candidate models in the routing benchmark.

Hardware and software configuration. The local server has two AMD EPYC 9654 processors (96 physical cores each; 384 logical CPUs in total), approximately 1 TiB of system memory, and eight NVIDIA L40 GPUs (48 GB nominal memory per GPU). The selected STATE training run uses one GPU, eight data-loading workers, and 32 GiB of container shared memory. CPU numerical-library thread counts are set to one per process. The host deployment uses Ubuntu 24.04 LTS. The scientific runtime is a separate Ubuntu 22.04 container, whose exact dependency versions are recorded in Table 5. The runtime combines a CUDA 12.4.1 base image with distinct CUDA 12.8 libraries bundled with the installed PyTorch wheel. A STATE patch in the frozen image enables the configured Softplus output activation.

Table 5: **Scientific runtime used for the reported STATE workflow.** Versions are verified from the frozen image and saved dependency inventory.

Component	Version
Python	3.10.12
STATE (`arc-state`)	0.10.2
PyTorch / bundled CUDA	2.10.0+cu128 / 12.8
PyTorch Lightning	2.6.6
AnnData / Scanpy	0.11.4 / 1.11.5
NumPy / SciPy / pandas	2.2.6 / 1.15.3 / 2.3.3
`cell-load` / `cell-eval`	0.10.3 / 0.6.8
h5py	3.16.0

B.2 LLM Configuration

The experiments use five LLM backbones, listed in Table 6. Predictive-model computation uses the local infrastructure described above. Each agent-based discovery strategy is evaluated with five independent LLM trajectories per backbone. Prediction-feedback strategies share a frozen STATE checkpoint. Iterative strategies propose ten compounds per round for five rounds. The LLM endpoints do not support sampling seeds in the recorded protocols.

Table 6: **LLM backbones used in the discovery experiments.** Model identifiers are the names submitted to the remote endpoints. DeepSeek Flash additionally enables thinking.

Backbone	Request identifier	Reasoning effort
Claude Opus 4.7	`claude-opus-4-7`	High
Claude Opus 5	`claude-opus-5`	High
GPT-5.6 Sol	`gpt-5.6-sol`	Medium
GPT-6 Sol	`gpt-6-sol`	High
DeepSeek Flash	`deepseek-flash`	High

B.3 Preprocessing and Data Partitions

The shared standardizer retains cells with 200–5,000 expressed genes, 500–20,000 counts, and at most 25% mitochondrial counts. E09 cells are partitioned by a deterministic stratified split (seed 13), using input, perturbation, cell type, and available batch and dose metadata. Validation and evaluation each target 10% of cells within eligible strata; strata smaller than ten cells remain in the fitting partition. The saved partition contains 66,113 fitting, 7,852 validation, and 7,852 evaluation cells. Physical cell memberships are disjoint, although STATE's native loader reuses validation controls as basal training references; held-out evaluation expression is accessed only after checkpoint locking. Normalization uses full-source library sizes and the fitting-set median target of 1,202 counts, followed by $\log(1+x)$. Scanpy's Seurat-flavor highly variable gene selection is fitted on the training partition and retains 4,000 genes, including TOP2A. Evaluation and inference reuse this transformation and gene order.

B.4 Predictive-Model Hyperparameters

STATE is trained from scratch with one-hot perturbation representations and sampled basal cells, using the configuration summarized in Table 7. The checkpoint with minimum validation loss is retained and hash-locked before held-out evaluation. These settings describe the single predictive-model fit reused by the discovery study; the five runs per strategy replicate LLM discovery.

B.5 Inference and Reproducibility

The frozen checkpoint receives all 2,404 admitted F10 control cells for each of 188 candidate compounds, using seed 42 for each inference shard. No F10 feature selection, normalization fitting, or parameter tuning is performed. Predictions are dose-agnostic; we evaluate all-dose effects. Discovery feedback includes the predicted TOP2A effect, its units, the number of finite gene effects, the response-vector norm, and the eight most upregulated and downregulated genes. A candidate is

Table 7: **STATE architecture and training hyperparameters.**

Hyperparameter	Value
Architecture	
Cell-set length	512
Hidden / feed-forward dimension	768 / 3,072
Transformer layers / attention heads	8 / 12
Attention	Bidirectional
Input-encoder / output-decoder layers	1 / 1
Residual prediction / Softplus	Enabled / enabled
Optimization and training	
Distributional objective	Energy loss
Batch size	8 cell sets
Learning rate	10^{-4}
Gradient clipping	10
Precision	bfloat16 mixed precision
Training seed	42
Optimizer steps	10,000
Validation / checkpoint interval	2,000 steps
Checkpoint selection	Minimum validation loss

accepted when its disclosed TOP2A effect is finite and positive; the agent's reflection is advisory. Per-run artifacts preserve prompts, responses, recommendations, feedback, and the final shortlist. Runtime image and checkpoint digests, feature and split manifests, source snapshots, dependency inventories, and prediction and trajectory locks bind the reported results to their execution inputs.

C Benchmark Construction

The routing benchmark contains 37 natural-language biological questions to assess whether PerturbOS can select an appropriate analysis route and, when prediction is required, a compatible dataset–model pair. Reference annotations distinguish direct analysis of observed responses, indirect prediction with a registered model, and requests unsupported by the registered resources. The annotations define the routing targets evaluated in Section 4.3.

Registry construction. Dataset records describe organism, cell context, perturbation modality and coverage, controls, and the intended data subset, with source publications, repositories, and accessions where available. The A549, K562, and MCF7 components of sci-Plex3 are separate entries. The registry also includes a resting CD4 T-cell CRISPRi subset of GSE314342. Entries identify defined data subsets, with overlapping resources recorded explicitly. Perturbation coverage is assessed at the registered-subset level, including for focused single-cell follow-ups of genome-wide screens. Proposed reductions in cell counts preserve eligible perturbation labels and controls. Artifact preparation is tracked separately from registry descriptions. Model records separately describe intended capabilities, required inputs, limitations, and implementation status.

Dataset selection and sources. We selected these datasets to span complementary human genetic and chemical perturbations, targeted and genome-scale screens, cell lines and primary cells, and distinct stimulation and drug-response states. Most source studies are also curated in the community resource scPerturb (Peidli et al., 2024), while the OpenProblems dataset supplies a dedicated perturbation-prediction benchmark (Szałata et al., 2024). Table 8 cites all 16 candidate entries and gives a short rationale for each. The datasets were selected to provide diversity in routing-relevant biological contexts and perturbation modalities.

Question construction. Questions were designed to cover varied routing scenarios. Each question specifies a cellular context, an expression readout gene, and a desired direction of change, while leaving the intervention open for recommendation. For example, "I'm planning to identify a compound that reduces TOP2A expression in human A549 cells" names a biological objective without prescribing a compound, dataset, or model. The named gene defines the expression readout; the intervention remains open for recommendation. Questions may specify a perturbation modality or

Table 8: **Dataset sources and selection rationale for the routing benchmark.** IDs match the frozen workbook. The 16 registry entries come from 14 studies; the three sci-Plex3 cell-line subsets share one publication. Entries define the candidate resource pool for routing.

ID	Dataset and primary source	Selection rationale
G02	Adamson K562 Perturb-seq (Adamson et al., 2016)	Contrasts focused CRISPRi with genome-scale screens.
G03	Dixit K562 Perturb-seq (Dixit et al., 2016)	Separates knockout from repression and activation.
G05	Datlinger CROP-seq (Datlinger et al., 2017)	Contrasts unstimulated and stimulated Jurkat cells.
G07	Norman K562 CRISPRa (Norman et al., 2019)	Adds CRISPRa through the single-gene activation arm.
G12	Frangieh Perturb-CITE-seq (Frangieh et al., 2021)	Adds melanoma and immune-interaction contexts.
G13	Papalexi ECCITE-seq (Papalexi et al., 2021)	Adds monocytic, multimodal immune-checkpoint data.
G15	Replogle K562 Perturb-seq (Replogle et al., 2022)	Supplies genome-scale CRISPRi coverage in K562.
G16	Shifrut primary T-cell follow-up (Shifrut et al., 2018)	Adds a focused knockout subset in primary T cells.
G17	GSE314342, resting CD4 T cells (Zhu et al., 2025)	Adds multi-donor, genome-scale primary-cell CRISPRi.
C02A	sci-Plex3 A549 (Srivatsan et al., 2020)	Provides broad chemical coverage in lung cancer cells.
C02K	sci-Plex3 K562 (Srivatsan et al., 2020)	Adds a hematopoietic context for the same drug panel.
C02M	sci-Plex3 MCF7 (Srivatsan et al., 2020)	Adds a breast cancer context for the same drug panel.
C03	ComboSciPlex (Lotfollahi et al., 2023)	Tests single-agent scope within a combination study.
C04	McFarland MIX-Seq (McFarland et al., 2020)	Broadens chemical responses across cancer cell lines.
C07	Aissa drug-tolerance scRNA-seq (Aissa et al., 2021)	Distinguishes drug-sensitive and drug-tolerant states.
C09	OpenProblems PBMC drug data (Szałata et al., 2024)	Adds a primary-immune-cell community benchmark.

compound class and the availability of a matched control H5AD. The benchmark covers single-gene and single-compound interventions. Routing and scoring consider biological states, stimulation conditions, donors, and batches; dose and treatment time are handled during subsequent model preparation and analysis.

D Formal Hit-Rate Definition

A drug's measured effect is its mean TOP2A expression in treated cells minus the mean in control cells, using the training-defined normalization and $\log(1+x)$ transformation. Within each measured view, let T_K contain the K drugs with the largest strictly positive measured effects, breaking ties by exact name. An empty shortlist has zero overlap and an undefined hit rate. For a nonempty final shortlist S, the hit rate uses its actual size:

$$\mathrm{HitRate}(K) = \frac{|S \cap T_K|}{|S|}. \tag{9}$$