PertReasonQA: A Knowledge-Grounded Benchmark and Framework for Cell-State–Conditioned Mechanistic Reasoning of Perturbation Effects

Dongkwan Kim [*1] **Yiming Gao** [*1] **Yining Yang** [1] **Yang Shen** [1]

Abstract

Evaluating machine learning in scientific domains requires separating correct predictions from correct reasons under distribution shifts. We introduce PertReasonQA, a knowledge-grounded benchmark for cell-state–conditioned reasoning about perturbation effects. It tests whether models can generate mechanistically faithful explanations and assesses their robustness against complex shifts, such as new cells and unseen perturbations. PertReasonQA combines multiple single-cell genetic and chemical perturbation data with knowledge graphs, and dynamically conditions pathways on cell-specific basal states to avoid generic memorization. Evaluations on state-of-the-art models reveal systematic gaps between outcome prediction and mechanistic reasoning, exhibiting failure modes invisible to standard benchmarks. As a reference probe, we present PertReasonLM, a large language model that aligns outcome predictions with context-specific regulatory reasoning. PertReasonQA thus provides a rigorous diagnostic benchmark for studying faithful, generalizable reasoning in data-rich scientific systems.

1. Introduction

Modeling complex systems remains a fundamental challenge in machine learning. Virtual cell models offer a demanding testbed for assessing whether AI can reason and generalize while accelerating drug discovery and disease modeling. Recent approaches use foundation models to predict transcriptomic changes induced by chemical or genetic perturbations (Cui et al., 2024; Theodoris et al., 2023; Roohani et al., 2024). Despite this progress, their utility remains limited by two issues. First, many models lack **mechanistic reasoning**: by bypassing established biological pathways (Novakovsky et al., 2023; Dimitrov et al., 2026), they cannot fully exploit shared causal mechanisms for generalization to unseen perturbations (Lotfollahi et al., 2023; Wei et al., 2025; Ahlmann-Eltze et al., 2025). Second, **label-centric evaluation** judges models mainly by final predictions (Wu et al., 2025a; Wei et al., 2025; Wu et al., 2025b), making it difficult to distinguish genuine biological reasoning from spurious correlations.

To address these challenges, we introduce PERTREASONQA, a knowledge-grounded question-answering (QA) benchmark for evaluating mechanistic reasoning about perturbation effects. PERTREASONQA pairs cell-specific perturbation outcomes with faithful gene-regulatory pathways derived from knowledge graphs (Türei et al., 2016; Bachman et al., 2023). By dynamically conditioning these pathways on the basal state of each cell, the benchmark avoids generic memorization and ensures that reference reasoning reflects valid, context-specific mechanisms. This design makes it possible to identify cases where models reach correct answers through flawed logic and to test their robustness under distribution shifts.

We also present PERTREASONLM, a reference large language model trained to align outcome predictions with context-specific pathways. Unlike existing methods based on fixed embedding lookups (Adduri et al., 2025) or static graphs (Roohani et al., 2024; Wenkel et al., 2025), PERTREASONLM can translate novel experimental conditions into actionable regulatory states. By learning to generate mechanistically grounded reasoning, it shifts perturbation modeling from numerical regression toward explainable and biologically plausible inference.

We evaluate state-of-the-art models on PERTREASONQA across unseen cells and novel perturbations. Using reasoning-centric metrics, we find that existing approaches, including Retrieval-Augmented Generation baselines (Wu et al., 2025a) and general-purpose language models (Yang et al., 2025), often produce flawed or directionally inconsistent mechanisms even when their final predictions are correct. PERTREASONLM mitigates these failures, suggesting that explicit alignment between predictions and context-specific regulatory reasoning can improve mech-

[1]Department of Electrical and Computer Engineering, Texas A&M University, College Station, Texas, United States. Correspondence to: Yang Shen <yshen@tamu.edu>.

Proceedings of the ICML 2026 3rd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Seoul, Korea. 2026.

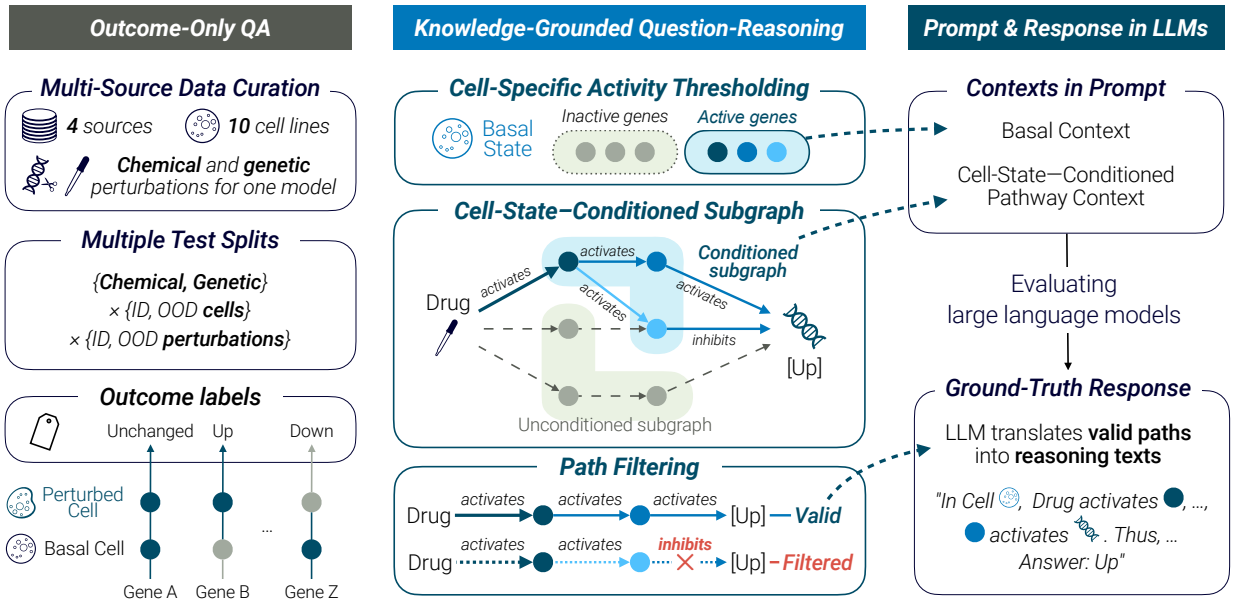

Figure 1. PERTREASONQA integrates multi-source data (left) featuring diverse chemical and genetic perturbations in ID and OOD contexts. It dynamically constructs a cell-state–conditioned subgraph (center) based on basal states to prevent generic memorization. This structured knowledge is then translated into prompts (right) to evaluate faithful mechanistic reasoning that aligns with the ground truth.

anistic faithfulness. Ultimately, PERTREASONQA serves as a rigorous diagnostic tool to foster the development of models that genuinely understand causal biological systems.

Our main contributions are threefold. First, we introduce PERTREASONQA, a knowledge-grounded benchmark for evaluating perturbation reasoning with cell-conditioned pathways (§2). Second, we present PERTREASONLM, a reference language model that aligns perturbation outcomes with context-specific regulatory mechanisms (§3). Third, we show that reasoning-centric evaluation reveals systematic failures in existing models (§4).

2. Benchmarking of Perturbation Reasoning

We present PERTREASONQA, a novel question-answering (QA) benchmark to evaluate knowledge-grounded mechanistic reasoning about cellular responses to perturbations. Further details can be found in Appendix B.

2.1. Background and Pipeline

Cellular systems respond to perturbation p through cascades altering the expression of an effect gene g_e. Because these cascades depend on the cell's basal state, mechanistic reasoning requires identifying a cell-specific causal chain from p to the observed g_e change. We formulate PERTREASONQA as a knowledge-grounded QA task by the following pipeline. First, given a cellular context, p, and g_e, we extract the differential expression label $y \in \{\text{unchanged, up, down}\}$. Second, we pair each outcome with a reference reasoning path $r = [(p, e_1, g_1), \ldots, (g_{k-1}, e_k, g_e)]$, where each e_i denotes a regulatory interaction from knowledge graphs. This evaluates whether valid reasoning supports the correct outcome.

2.2. Constructing Question-Outcome Dataset

We construct outcome-only QA samples by aggregating chemical and genetic (CRISPRi) perturbation data from multiple large-scale sources (Srivatsan et al., 2020; Szał ata et al., 2024; Replogle et al., 2022; Nadig et al., 2024), covering 10 cell lines: K562, MCF7, A549, HepG2, Jurkat, RPE1, B cells, Myeloid cells, NK cells, and T cells. Individual cells are aggregated into pseudobulks and paired with average basal expression profiles. To evaluate out-of-distribution generalization, we partition the data across cell lines and perturbations, using a cluster-based split to reduce pathway leakage based on gene network embeddings (Türei et al., 2016; Liberzon et al., 2015). Finally, following PerturbQA (Wu et al., 2025a), we compute differential expression against controls using the Wilcoxon signed-rank test (Wilcoxon, 1945) with Benjamini-Hochberg correction (Benjamini & Hochberg, 2000), assigning Up or Down-regulation to genes with adjusted p-value < 0.01 and Unchanged to genes with p-value > 0.1.

2.3. Synthesizing Knowledge-Grounded Reasoning

We synthesize knowledge-grounded reasoning paths by linking empirical perturbation outcomes to context-conditioned biological knowledge graphs.

Cell-Specific Basal Activity Thresholding To express cellular context in natural-language reasoning, we discretize basal expression into "low," "medium," and "high"

Table 1. Summary statistics of the PERTREASONQA, including the number of unique perturbations, cell lines, labels (unchanged, up, down), and pathways across training, validation, and all test splits.

		Train ID Cell		Valid ID Cell		Test ID Cell				Test OOD Cell			
		Chem.	Gen.	Chem.	Gen.	ID Chem.	ID Gen.	OOD Chem.	OOD Gen.	ID Chem.	ID Gen.	OOD Chem.	OOD Gen.
Perturbation		135	1172	17	137	15	131	25	518	15	125	25	481
Cell Lines		5	3	5	3	5	3	5	3	2	1	2	1
Outcome Samples	Unch.	62730	212524	7583	24042	8298	24564	11682	91561	3094	7957	4575	30204
	Up	32304	179141	2600	22951	3825	28308	3130	35646	2140	11348	1668	28788
	Down	53412	282868	6653	31775	8140	31910	8963	67015	2493	16421	4114	45947
Reasoning Samples	Unch.	19156	36755	2906	4715	2316	1275	1913	62	809	192	815	24
	Up	17985	28313	1562	4700	1579	1265	628	133	803	396	496	64
	Down	29901	54387	3857	7104	3252	1939	3203	464	1406	692	1935	273
Avg. Pathway Length		5.98	4.91	6.06	4.62	6.25	4.68	5.29	3.95	6.05	4.89	5.07	4.35

states. We define cell-specific thresholds using Kolmogorov-Smirnov (KS) statistics: τ_{low} separates unchanged genes as noise from changed genes as signal, while τ_{high} estimates the saturation point of functionally active genes.

Cell-State–Conditioned Path Extraction Because global knowledge graphs include interactions that may be inactive from gene silencing in a cell, we condition edge weights on the basal expression $\mathbf{b}^{(c)}$ of genes. For each gene edge (u, v), we use $w_{uv}^{(c)} = w_{uv}^{\text{base}} \cdot g_u^{\text{src}} \cdot g_v^{\text{tgt}}$, where $g_\cdot \propto \text{sigmoid}(\mathbf{b}_\cdot^{(c)} - \tau_{\text{low}})$ penalizes edges involving inactive genes. We then convert these activities into inverse traversal costs and run Dijkstra searches on the resulting KG to extract high-confidence paths (Paull et al., 2013).

Reference Paths via Filtering We filter extracted paths to retain mechanisms that support the observed expression outcome. Regulatory signs are assigned from INDRA CoGEx (Bachman et al., 2023), and differentially expressed genes keep only paths whose cumulative sign propagation matches the observed up- or down-regulation (Algorithm 1). For unchanged genes, we retain paths whose signals are blocked by cellular boundary conditions, such as inhibition of already silent genes (Algorithm 2).

Path-to-Text Reasoning Synthesis Finally, we convert validated causal paths into natural-language reasoning using Qwen3-4B (Yang et al., 2025). A post-generation quality control step removes low-quality samples through keyword matching, ensuring that obvious unsupported or refusal-like generations are removed.

Dataset Statistics PERTREASONQA contains 1.4M outcome-only QA samples and 237k reasoning samples. Detailed statistics are summarized in Table 1.

2.4. Evaluation Metrics

We report the balanced accuracy on the 3-way outcome classification (up, down, and unchanged). We also introduce Mechanistic Faithfulness to verify the pathway and biological correctness of context-specific reasoning, specifically,

- **Edge Recall:** We evaluate pathway correctness by the fraction of reference paths correctly covered by the generated reasoning paths. Note that we distinguish surface natural-language reasoning from structured mechanism matching. Our faithfulness metrics are computed on canonical signed triplets extracted from the `<triplets>` block. Thus, Edge Recall measures recovery of structured regulatory edges rather than lexical overlap with LLM-generated prose.
- **Path Connectivity:** To further ensure pathway correctness, this verifies whether the generated intermediate nodes do not form a broken chain.
- **Gene-Ontology Similarity:** To assess biological correctness, we account for generated chains that are functionally similar to references without matching edge-by-edge. We build IDF-weighted Gene-Ontology (GO) Biological Process profiles from intermediate genes and score predicted chains against references using cosine similarity.

3. Models and Experimental Setup

We present PERTREASONLM, a reference probe for reasoning-centric perturbation modeling. PERTREASONLM is trained through supervised fine-tuning (SFT) followed by group relative policy optimization (GRPO). Detailed specifications can be found in Appendix C.1 (SFT) and C.2 (GRPO). This section also describes benchmarking configurations.

3.1. Perturbation Reasoning Model

Supervised Fine-Tuning The SFT stage builds biological knowledge and structured reasoning in PERTREASONLM through a curriculum. We first use outcome-only supervision, where the model predicts differential expression labels from cellular contexts and perturbations, to learn broad domain priors on entities. We then add mechanistic reasoning supervision, training PERTREASONLM to generate natural language Chain-of-Thought (CoT) reasoning (Wei et al., 2022; Chung et al., 2024) that supports the final outcome with context-specific regulatory pathways. Target responses

Table 2. Balanced accuracy for 3-way outcome prediction in ID and OOD test splits of cells and perturbations. GEARS, scGPT, and STATE are evaluated only on the genetic subset by design, and their averages are computed over genetic-only slices (See §D.1).

Model	Average	ID Cell				OOD Cell			
		ID Chem.	ID Gen.	OOD Chem.	OOD Gen.	ID Chem.	ID Gen.	OOD Chem.	OOD Gen.
GEARS	0.356	-	0.357	-	0.405	-	0.341	-	0.320
scGPT	0.359	-	0.374	-	0.390	-	0.350	-	0.323
STATE	0.411	-	0.510	-	0.407	-	0.403	-	0.324
BioMistral 7B	0.309	0.305	0.307	0.313	0.319	0.324	0.294	0.315	0.293
NatureLM 8x7B	0.334	0.336	0.332	0.334	0.333	0.331	0.337	0.335	0.333
SUMMER-4B	0.404	0.444	0.389	0.421	0.360	0.442	0.416	0.392	0.367
Qwen3-4B 0-shot	0.366	0.418	0.330	0.358	0.358	0.400	0.372	0.325	0.370
PERTREASONLM-4B									
Outcome SFT	0.636	0.713	0.690	0.612	0.622	0.653	0.646	0.543	0.611
+ CoT SFT	0.709	0.749	0.722	0.678	0.708	0.705	0.697	0.669	0.743
+ GRPO	0.736	0.772	0.756	0.721	0.745	0.668	0.763	0.680	0.786

are formatted with specialized tags, such as `<thinking>`, `<answer>`, and `<triplets>`, and data mixing (Guo et al., 2025; Mitra et al., 2023) is used to combine reasoning samples with outcome-only data. Finally, relation direction self-supervision masks and predicts causal edge signs (Teru et al., 2020a; Zhang et al., 2024), helping the model learn activation and inhibition patterns from cellular and pathway contexts. This curriculum connects abundant outcome labels with scarce high-fidelity reasoning signals.

Reinforcement Learning with GRPO We apply Group Relative Policy Optimization (GRPO) (Shao et al., 2024) on top of the SFT model as a post-training alignment. It aims to tighten the agreement between the final expression label and the generated causal chain. To do so, the reward combines answer correctness with structured triplet-level feedback, so that the model is encouraged to produce responses that are both label-correct and mechanistically grounded.

Backbone Models We use the Qwen3-4B (Yang et al., 2025) as the primary backbone for PERTREASONLM due to its robust reasoning performance at a moderate scale.

3.2. Baselines

We compare PERTREASONLM with state-of-the-art gene-space and text-space baselines. Detailed descriptions and configurations for all baselines are provided in Appendix D.

Gene-Space Baselines with Numerical Outputs We adopt state-of-the-art models that only generate numerical outputs: (1) GEARS (Roohani et al., 2024), (2) scGPT (Cui et al., 2024), and (3) STATE (SE+ST) (Adduri et al., 2025). We restrict these baselines to the genetic subset for evaluation due to their architecture limitation.

Text-Space Baselines based on LLMs We compare against: (1) the Qwen-3 4B base model (Yang et al., 2025), (2) domain-specific LLMs on biology, BioMistral 7B (Labrak et al., 2024) and NatureLM 8x7B (Xia et al., 2025), (3) SUMMER (Wu et al., 2025a) with Qwen3-4B backbone, a state-of-the-art Retrieval-Augmented Generation (RAG) model for perturbation responses.

4. Results

In this section, we assess PERTREASONQA from both predictive and reasoning standpoints.

Benchmarking Outcome Prediction and Generalization We evaluate models under the cell-conditioned path setting, where predictions must be made through noisy, context-specific regulatory environments. Table 2 reports balanced accuracy across test folds. For gene-space baselines, GEARS and scGPT show unstable behavior across genetic splits; all degrade most clearly under OOD-cell settings. STATE reaches 0.510 balanced accuracy in-domain, yet drops in distribution shifts, showing the limited generalization of non-textual representations. Similarly, Qwen3-4B (0-shot), BioMistral, and NatureLM struggle without explicit alignment to regulatory pathways, often failing to convert biological knowledge into context-specific predictions. PERTREASONLM shows stronger predictive performance and robustness. PERTREASONLM-SFT (Outcome) achieves an average balanced accuracy of 0.636, while adding reasoning in PERTREASONLM-SFT improves it to 0.709. PERTREASONLM-GRPO further increases performance to 0.736 and outperforms SFT in seven out of eight folds, suggesting that RL-based alignment better captures biological logic. Its ID-to-OOD degradation is also small, indicating improved stability across distribution shifts.

Benchmarking Reasoning Faithfulness We test whether predictive gains come from faithful reasoning. Table 3 reports mechanistic faithfulness and outcome metrics aver-

Table 3. Comparing outcome accuracy and mechanism faithfulness across models, using averages over test folds.

Pathway Contexts	Model	Outcome	Reasoning			Failure Modes	
		Bal. Acc.	Edge Recall	Path Conn.	GO Sim.	Answer × Reason ✓	Answer ✓ Reason ×
No Paths	SUMMER-4B	0.374	0.090	0.230	0.326	0.003	0.348
	Qwen3-4B	0.338	0.134	0.873	0.614	0.624	0.266
	PERTREASONLM-4B SFT	0.689	0.345	0.992	0.754	0.294	0.468
	+ GRPO	0.707	0.397	0.992	0.780	0.272	0.427
Cell-Cond. (Default)	SUMMER-4B	0.404	0.837	0.922	0.946	0.502	0.036
	Qwen3-4B	0.366	0.852	0.994	0.897	0.592	0.024
	PERTREASONLM-4B SFT	0.709	0.979	0.978	0.938	0.269	0.037
	+ GRPO	0.736	0.979	0.990	0.938	0.258	0.016

aged across test folds. With cell-conditioned context, pathway information improves reasoning metrics for most models, but PERTREASONLM converts this context into both accurate outcomes and faithful mechanisms. PERTREASONLM-GRPO achieves 0.979 edge recall and 0.938 GO similarity, while also reducing the "right answer with wrong reason" failure mode from 3.7% in PERTREASONLM-SFT to 1.6%. In contrast, Qwen3-4B and the RAG-based SUMMER often extract plausible chains but fail to link them to the correct downstream effect. The gap widens in the no-path setting. Without pathway context, SUMMER and Qwen3-4B show low or unstable edge recall and GO similarity, and Qwen3-4B reaches 62.4% in the "wrong answer with right reason" category. PERTREASONLM retains stronger reasoning faithfulness and outcome accuracy even without explicit contextual clues, indicating that it has internalized transferable regulatory logic rather than simply copying provided pathways.

5. Limitations and Future Work

PERTREASONQA constructs reference mechanisms from curated knowledge graphs and public perturbation datasets, and therefore inherits their incompleteness, coverage bias, and occasional ambiguity; its pathways should be interpreted as biologically grounded mechanistic proxies rather than complete causal explanations or substitutes for experimental validation. In addition, our current pipeline uses Qwen3 both to verbalize filtered structured paths into natural-language reasoning traces and as the primary backbone of PERTREASONLM, while Qwen3 is also included among the evaluated text-space baselines. Although outcome labels, cell-state conditioning, and sign-based path filtering are determined before text generation, this design may introduce a surface-form imitation channel in which PERTREASONLM benefits partly from matching the prose style of the generator used to synthesize reference explanations. Future work should decouple structured path construction from textual realization more strongly by using multiple non-Qwen generators and adding expert-curated subsets for path-level validation.

6. Conclusion

We introduced PERTREASONQA, a knowledge-grounded QA benchmark for evaluating cell-state–conditioned mechanistic reasoning about perturbation effects. By pairing perturbation outcomes with context-specific regulatory pathways, PERTREASONQA moves beyond label-centric evaluation and exposes failure modes that are invisible to endpoint accuracy alone, including predictions that are correct but supported by flawed or context-insensitive mechanisms. We also presented PERTREASONLM as a reference probe for reasoning-centric perturbation modeling, showing that explicit alignment between outcome prediction and structured regulatory evidence can improve both predictive performance and mechanistic faithfulness. Overall, our results suggest that progress in virtual-cell modeling should be assessed not only by how accurately models predict cellular responses, but also by whether they can justify those predictions through biologically plausible and generalizable mechanisms.

Acknowledgements

Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing.

Impact Statement

This work primarily contributes a benchmark and reference modeling framework for evaluating whether machine learning systems predict cellular perturbation effects for the right mechanistic reasons. If used responsibly, PERTREASONQA can help researchers compare virtual-cell models beyond endpoint accuracy and identify cases where a model reaches the correct answer through biologically implausible logic. In that sense, the benchmark may improve transparency and reliability in early-stage drug discovery and disease-modeling workflows, where understanding why a prediction is made is often as important as the prediction itself.

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. *BioRxiv*, pp. 2025–06, 2025.

Ahlmann-Eltze, C., Huber, W., and Anders, S. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8):1657–1661, 2025. doi: 10.1038/s41592-025-02772-6. URL https://doi.org/10.1038/s41592-025-02772-6.

Bachman, J. A., Gyori, B. M., and Sorger, P. K. Automated assembly of molecular mechanisms at scale from text mining and curated databases. *Molecular Systems Biology*, 19(5):MSB202211325, 2023. doi: 10.15252/msb.202211325. URL https://doi.org/10.15252/msb.202211325.

Benjamini, Y. and Hochberg, Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of educational and Behavioral Statistics*, 25(1):60–83, 2000.

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

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 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, 2024.

Dimitrov, D., Schrod, S., Rohbeck, M., and Stegle, O. Interpretation, extrapolation and perturbation of single cells. *Nature Reviews Genetics*, pp. 1–22, 2026.

Du, J., Jia, P., Dai, Y., Tao, C., Zhao, Z., and Zhi, D. Gene2vec: distributed representation of genes based on co-expression. *BMC genomics*, 20(Suppl 1):82, 2019.

Gao, Y., Wang, Z., Chen, J., Antkowiak, M., Hu, M., Kong, J., Pratt, D., Liu, J., Ma, E., Hu, Z., and Xing, E. P. scpilot: Large language model reasoning toward automated single-cell analysis and discovery. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=Vzi96rTe4w.

Grønbech, C. H., Vording, M. F., Timshel, P. N., Sønderby, C. K., Pers, T. H., and Winther, O. scvae: variational auto-encoders for single-cell gene expression data. *Bioinformatics*, 36(16):4415–4422, 2020.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Hetzel, L., Böhm, S., Kilbertus, N., Günnemann, S., Lotfollahi, M., and Theis, F. J. Predicting cellular responses to novel drug perturbations at a single-cell resolution. In *NeurIPS 2022*, 2022.

Kim, H., Hwang, S.-Y., Lim, J., Piao, Y., Oh, Y., Kim, W. Y., Park, C., Ahn, S., and Jeon, J. Progressive multi-agent reasoning for biological perturbation prediction, 2026. URL https://arxiv.org/abs/2602.07408.

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, 2024. URL https://arxiv.org/abs/2402.10373.

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. In *International Conference on Machine Learning*, pp. 27299–27325. PMLR, 2024.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL https://arxiv.org/abs/2005.11401.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6):417–425, 2015.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

Lotfollahi, M., Klimovskaia Susmelj, A., De Donno, C., Hetzel, L., Ji, Y., Ibarra, I. L., Srivatsan, S. R., Naghipourfar, M., Daza, R. M., Martin, B., et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, pp. e11517, 2023.

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), September 2022. ISSN 1477-4054. doi: 10.1093/bib/bbac409. URL http://dx.doi.org/10.1093/bib/bbac409.

Magister, L. C., Mallinson, J., Adamek, J., Malmi, E., and Severyn, A. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1773–1781, 2023.

Mitra, A., Del Corro, L., Mahajan, S., Codas, A., Simoes, C., Agarwal, S., Chen, X., Razdaibiedina, A., Jones, E., Aggarwal, K., et al. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*, 2023.

Nadig, A., Replogle, J. M., Pogson, A. N., McCarroll, S. A., Weissman, J. S., Robinson, E. B., and O'Connor, L. J. Transcriptome-wide characterization of genetic perturbations. *bioRxiv*, 2024. doi: 10.1101/2024.07.03.601903. URL https://www.biorxiv.org/content/early/2024/07/03/2024.07.03.601903.

Ni, S., Kong, X., Zhang, Y., Chen, Z., Wang, Z., Fu, Z., Huo, R., Tong, X., Qu, N., Wu, X., et al. Identifying compound-protein interactions with knowledge graph embedding of perturbation transcriptomics. *Cell genomics*, 4(10), 2024.

Nori, H., King, N., McKinney, S. M., Carignan, D., and Horvitz, E. Capabilities of gpt-4 on medical challenge problems, 2023. URL https://arxiv.org/abs/2303.13375.

Novakovsky, G., Dexter, N., Libbrecht, M. W., Wasserman, W. W., and Mostafavi, S. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*, 24(2):125–137, 2023.

Paull, E. O., Carlin, D. E., Niepel, M., Sorger, P. K., Haussler, D., and Stuart, J. M. Discovering causal pathways linking genomic events to transcriptional states using tied diffusion through interacting events (tiedie). *Bioinformatics*, 29(21):2757–2764, 08 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt471. URL https://doi.org/10.1093/bioinformatics/btt471.

Peidli, S., Green, T. D., Shen, C., Gross, T., Min, J., Garda, S., Yuan, B., Schumacher, L. J., Taylor-King, J. P., Marks, D. S., et al. scperturb: harmonized single-cell perturbation data. *Nature Methods*, 21(3):531–540, 2024.

Phillips, L., Martell, M. B., Misra, A., Stoisser, J. L., Prada-Medina, C. A., Donovan-Maiye, R., and Märtens, K. Synthpert: Enhancing llm biological reasoning via synthetic reasoning traces for cellular perturbation prediction, 2025. URL https://arxiv.org/abs/2509.25346.

Radig, J., Droit, R., Doncevic, D., Li, A., Bui, D. T., Herfurth, L., Kühn, T., and Herrmann, C. Tracking biological hallucinations in single-cell perturbation predictions using scarchon, a comprehensive benchmarking platform. *bioRxiv*, 2025. doi: 10.1101/2025.06.23.661046. URL https://www.biorxiv.org/content/early/2025/06/27/2025.06.23.661046.

Rasley, J., Rajbhandari, S., Ruwase, O., and He, Y. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pp. 3505–3506, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3406703. URL https://doi.org/10.1145/3394486.3406703.

Replogle, J. M., Saunders, R. A., Pogson, A. N., Hussmann, J. A., Lenail, A., Guna, A., Mascibroda, L., Wagner, E. J., Adelman, K., Lithwick-Yanai, G., Iremadze, N., Oberstrass, F., Lipson, D., Bonnar, J. L., Jost, M., Norman, T. M., and Weissman, J. S. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2022.05.013. URL https://www.sciencedirect.com/science/article/pii/S0092867422005979.

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

Rosen, Y., Roohani, Y., Agarwal, A., Samotorčan, L., Consortium, T. S., Quake, S. R., and Leskovec, J. Universal cell embeddings: A foundation model for cell biology. *BioRxiv*, pp. 2023–11, 2023.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Scharli, N., Chowdhery, A., Mansfield, P., y Arcas, B. A., Webster, D., Corrado, G. S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A., and Natarajan, V. Large language models encode clinical knowledge, 2022. URL https://arxiv.org/abs/2212.13138.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S. R., Cole-Lewis, H., Neal, D., Rashid, Q. M., Schaekermann, M.,

Wang, A., Dash, D., Chen, J. H., Shah, N. H., Lachgar, S., Mansfield, P. A., Prakash, S., Green, B., Dominowska, E., Agüera y Arcas, B., Tomašev, N., Liu, Y., Wong, R., Semturs, C., Mahdavi, S. S., Barral, J. K., Webster, D. R., Corrado, G. S., Matias, Y., Azizi, S., Karthikesalingam, A., and Natarajan, V. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025. doi: 10.1038/s41591-024-03423-7. URL https://doi.org/10.1038/s41591-024-03423-7.

Srivatsan, S. R., McFaline-Figueroa, J. L., Ramani, V., Saunders, L., Cao, J., Packer, J., Pliner, H. A., Jackson, D. L., Daza, R. M., Christiansen, L., Zhang, F., Steemers, F., Shendure, J., and Trapnell, C. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020. doi: 10.1126/science.aax6234. URL https://www.science.org/doi/abs/10.1126/science.aax6234.

Su, X., Wang, Y., Gao, S., Liu, X., Giunchiglia, V., Clevert, D.-A., and Zitnik, M. KGARevion: An AI agent for knowledge-intensive biomedical QA. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tnB94WQGrn.

Szał ata, A., Benz, A., Cannoodt, R., Cortes, M., Fong, J., Kuppasani, S., Lieberman, R., Liu, T., Mas-Rosario, J. A., Meinl, R., Nourisa, J., Tumiel, J., Tunjic, T. M., Wang, M., Weber, N., Zhao, H., Anchang, B., Theis, F. J., Luecken, M. D., and Burkhardt, D. B. A benchmark for prediction of transcriptomic responses to chemical perturbations across cell types. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 20566–20616. Curran Associates, Inc., 2024. doi: 10.52202/079017-0650.

Teru, K., Denis, E., and Hamilton, W. Inductive relation prediction by subgraph reasoning. In *International conference on machine learning*, pp. 9448–9457. PMLR, 2020a.

Teru, K. K., Denis, E., and Hamilton, W. L. Inductive relation prediction by subgraph reasoning, 2020b. URL https://arxiv.org/abs/1911.06962.

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., and Ellinor, P. T. Transfer learning enables predictions in network biology. *Nature*, 618:616–624, May 2023.

Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. Omnipath: guidelines and gateway for literature-curated signaling pathway resources. *Nature methods*, 13(12):966–967, 2016.

von Werra, L., Belkada, Y., Tunstall, L., Beeching, E., Thrush, T., Lambert, N., Huang, S., Rasul, K., and Gallouédec, Q. TRL: Transformers Reinforcement Learning, 2020. URL https://github.com/huggingface/trl.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Wei, Z., Wang, Y., Gao, Y., Wang, S., Li, P., Si, D., Gao, Y., Wu, S., Li, D., Dong, K., et al. Benchmarking algorithms for generalizable single-cell perturbation response prediction. *Nature Methods*, pp. 1–14, 2025.

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.

Wenteler, A., Occhetta, M., Branson, N., Curean, V., Huebner, M., Dee, W., Connell, W., Chung, S. P., Hawkins-Hooker, A., Ektefaie, Y., Córdova, C. M. V., and Gallagher-Syed, A. Perteval-scFM: Benchmarking single-cell foundation models for perturbation effect prediction. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=t04D9bkKUq.

Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

Wu, M., Littman, R., Levine, J., Qiu, L., Biancalani, T., Richmond, D., and Huetter, J.-C. Contextualizing biological perturbation experiments through language. In *The Thirteenth 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. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025b. URL https://openreview.net/forum?id=PPPDuyiZaG.

Xia, Y., Jin, P., Xie, S., He, L., Cao, C., Luo, R., Liu, G., Wang, Y., Liu, Z., Chen, Y.-J., Guo, Z., Bai, Y., Deng, P., Min, Y., Lu, Z., Hao, H., Yang, H., Li, J., Liu, C., Zhang, J., Zhu, J., Bi, R., Wu, K., Zhang, W., Gao, K., Pei, Q.,

Wang, Q., Liu, X., Li, Y., Zhu, H., Lu, Y., Ma, M., Wang, Z., Xie, T., Maziarz, K., Segler, M., Yang, Z., Chen, Z., Shi, Y., Zheng, S., Wu, L., Hu, C., Dai, P., Liu, T.-Y., Liu, H., and Qin, T. Nature language model: Deciphering the language of nature for scientific discovery, 2025. URL https://arxiv.org/abs/2502.07527.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

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.

Zhang, Y., Chen, Z., Guo, L., Xu, Y., Zhang, W., and Chen, H. Making large language models perform better in knowledge graph completion. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 233–242, 2024.

Zhao, H., Ma, C., Xu, F., Kong, L., and Deng, Z.-H. Biomaze: Benchmarking and enhancing large language models for biological pathway reasoning, 2025. URL https://arxiv.org/abs/2502.16660.

Zitnik, M., Agrawal, M., and Leskovec, J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466, 06 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty294. URL https://doi.org/10.1093/bioinformatics/bty294.

A. Related Work

Benchmarking and Modeling Cellular Perturbation Prediction The landscape of cellular perturbation prediction encompasses advancements in modeling and benchmarking. On the modeling front, architectures have evolved from gene embeddings (Du et al., 2019) and generative models (Lopez et al., 2018; Grønbech et al., 2020) to latent space arithmetic (Lotfollahi et al., 2023; Hetzel et al., 2022) and single-cell foundation models (Yang et al., 2022; Cui et al., 2024; Theodoris et al., 2023; Rosen et al., 2023; Adduri et al., 2025). To evaluate these systems, standardized benchmarks initially focused on the numerical regression of expression profiles (Peidli et al., 2024; Szał ata et al., 2024; Wei et al., 2025; Wu et al., 2025b; Wenteler et al., 2025). However, optimizing strictly for regression yields black-box models that lack mechanistic reasoning, causing them to frequently underperform linear baselines (Ahlmann-Eltze et al., 2025) or exhibit biological hallucinations (Radig et al., 2025). To incorporate semantic context and structural grounding, recent modeling approaches have explored graph augmentations, synthetic trace imitation, and multi-agent pipelines (Zhao et al., 2025; Phillips et al., 2025; Kim et al., 2026). Concurrently, within benchmarking, PerturbQA (Wu et al., 2025a) pioneered the use of natural language evaluation, though its scope currently remains constrained to discrete factual question-answering rather than tracing complex causal mechanisms. PERTREASONQA addresses these gaps by explicitly shifting the evaluation paradigm from static predictions to the rigorous assessment of mechanistically faithful perturbation reasoning.

Knowledge Graphs and Large Language Models in Biomedical Applications Researchers have used Knowledge Graphs (KGs) to decipher complex interactions between biological entities (Chandak et al., 2023; Zitnik et al., 2018; Roohani et al., 2024; Ni et al., 2024). However, the static nature of predefined KG schemas limits inductive generalization to novel entities (Teru et al., 2020b). Large Language Models (LLMs) bridge this semantic gap by offering transformative biomedical reasoning (Luo et al., 2022; Singhal et al., 2025) and translating transcriptomics into linguistic constructs (Singhal et al., 2022; Nori et al., 2023; Levine et al., 2024). To anchor these models in factual biology, recent frameworks integrate LLMs with external knowledge via Retrieval-Augmented Generation (Lewis et al., 2021; Wu et al., 2025a) or specialized agents (Su et al., 2025; Gao et al., 2025). Despite these synergies, current approaches prioritize discrete fact retrieval over dynamic simulation. To address this, PERTREASONQA evaluates mechanistically faithful reasoning beyond static predictions. Concurrently, PERTREASONLM dynamically aligns flexible language models with context-specific pathways, overcoming the structural limitations of standard graph retrieval.

B. Data Details

B.1. Splits

To evaluate generalizability, we partition cell lines designed to measure out-of-distribution (OOD) performance. Specifically, we set OOD cells as T cells and A549 for chemical perturbations, RPE1 for genetic perturbations.

To rigorously evaluate generalization, we implement a functional splitting strategy for both genetic and chemical perturbations. By partitioning perturbations based on biological function rather than random assignment, we ensure the test set evaluates true extrapolation to novel mechanisms rather than memorization of overlapping pathways. We construct a heterogeneous directed biological graph from OmniPath (Türei et al., 2016) and augment it with Gene Ontology (GO) gene sets from MSigDB (v2023.2) (Liberzon et al., 2015). We then compute Personalized PageRank (PPR) scores across the gene interaction network and aggregate them onto connected GO term nodes, yielding a gene-to-GO continuous vector encoding each gene's functional and structural context.

Based on cosine similarity of these functional vectors, we apply KMeans clustering ($k = 30$) to group genes sharing similar biological roles, then randomly partition entire clusters into ID and OOD sets. Chemical perturbations are aligned to this functional split by mapping drugs to their target genes via INDRA CoGEx (Bachman et al., 2023); a drug is assigned to the OOD test set if more than 10% threshold of its target genes fall into OOD gene clusters. This formulation guarantees that both genetic and chemical OOD splits systematically evaluate the model's capacity to extrapolate to unseen biological mechanisms.

B.2. Knowledge-Grounded Reasoning Details

B.2.1. CELL-SPECIFIC BASAL ACTIVITY THRESHOLDING

A fundamental step in modeling cell-specific mechanisms is interpreting the continuous spectrum of basal gene expression. Raw transcriptomic data frequently suffer from technical noise and systemic batch effects across different experimental

runs, making fixed global thresholds highly unreliable. To robustly categorize expression levels across diverse contexts, we establish adaptive, cell-specific thresholds. Biologically, we identify two critical transition points in the data density. The low threshold (τ_{low}) separates true biological signal from technical background noise, ensuring we only consider functionally active genes. Conversely, the high threshold (τ_{high}) marks the saturation point of active expression, identifying genes that are fully transcribed or operating at maximum biological capacity. This dynamic thresholding mitigates batch effects and provides a normalized context for the PERTREASONQA task.

To determine the low threshold (τ_{low}), which isolates meaningful biological activity from background noise, we employ a maximum separability criterion using a two-sample Kolmogorov-Smirnov (KS) statistic. We treat the basal expression values of genes labeled as unchanged after a perturbation p as a background noise distribution (X_{bg}), and those labeled as changed as a signal distribution (X_{sig}). By evaluating their empirical cumulative distribution functions (ECDFs), $\hat{F}_{\text{bg}}(x)$ and $\hat{F}_{\text{sig}}(x)$, the threshold is defined as the value that maximizes the signed KS statistic:

$$\tau_{\text{low}} = \arg\max_{x} \left(\hat{F}_{\text{bg}}(x) - \hat{F}_{\text{sig}}(x) \right)$$

This objective effectively identifies the precise decision boundary that optimally separates the background distribution from the active signal.

To capture the upper bounds of typical expression dynamics, we define the high threshold (τ_{high}), which represents the biological saturation point of gene expression. For genes exhibiting active expression ($X_{\text{act}} > \tau_{\text{low}}$), we apply a normalization mapping $\phi : X_{\text{act}} \to [0, 1]$:

$$z_i = \phi(x_i) = \frac{x_i - \min(X_{\text{act}})}{\max(X_{\text{act}}) - \min(X_{\text{act}})}$$

Let $\hat{F}_{\text{act}}(z)$ denote the ECDF of these normalized values. We compare this empirical distribution against the cumulative distribution function of a standard Uniform distribution, $F_U(z) = z$, which acts as a constant density baseline. The high threshold is determined by calculating the one-sample Kolmogorov-Smirnov (KS) statistic against this linear baseline:

$$z^* = \arg\max_{z \in [0,1]} \left| \hat{F}_{\text{act}}(z) - z \right|, \quad \tau_{\text{high}} = \phi^{-1}(z^*)$$

In the context of heavy-tailed biological data, this geometric knee point effectively localizes the structural transition from the main body of active expression to a state of saturation.

B.2.2. CELL-STATE-CONDITIONED SUBGRAPH EXTRACTION

A fundamental challenge in biological modeling is that general KGs contain the superset of all possible interactions, many of which are physically impossible in specific cell types due to gene silencing. The static architecture of a general KG must be dynamically refined to reflect the functional connectivity unique to a given transcriptomic state.

Let $\mathcal{G} = (\mathbb{V}, \mathbb{A})$ be the global knowledge graph derived from OmniPath, where each node $v \in \mathbb{V}$ denotes a gene, and each directed edge $(u, v) \in \mathbb{A}$ denotes a regulatory relation. For a specific cell state c, we retrieve its basal gene expression profile $\mathbf{b}^{(c)} \in \mathbb{R}^{|\mathbb{V}|}$.

Using the established low threshold τ_{low}, we implement an asymmetric basal activity gating mechanism that modulates edge traversability based on gene expression context. For each edge (u, v), we define a structural base weight w_{uv}^{base}, which is then modulated using sigmoid-based gating functions with asymmetric treatment of source and target nodes:

$$g_u^{\text{src}} = \max\left(\sigma\big(\beta(\mathbf{b}_u^{(c)} - \tau_{\text{low}})\big), \delta_{\text{src}} \right)$$

$$g_v^{\text{tgt}} = \delta_{\text{tgt}} + (1 - \delta_{\text{tgt}}) \cdot \sigma\big(\beta(\mathbf{b}_v^{(c)} - \tau_{\text{low}})\big)$$

where $\sigma(\cdot)$ denotes the logistic sigmoid function, β is a steepness parameter controlling the gate's sensitivity, and δ_{src} and δ_{tgt} are floor constants that maintain minimum traversability. The final context-conditioned edge weight is:

$$w_{uv}^{(c)} = w_{uv}^{\text{base}} \cdot g_u^{\text{src}} \cdot g_v^{\text{tgt}} + \epsilon_w$$

where ϵ_w is a small constant to prevent weights from becoming exactly zero. This asymmetric design reflects biological reality: the source floor δ_{src} allows signal propagation through weakly-expressed genes (modeling de-repression or inducibility), while the target floor δ_{tgt} ensures downstream genes remain inducible even when they are basally silent.

B.2.3. DIVERSITY-AWARE CAUSAL PATH-FINDING

To construct robust reasoning chains from perturbation sources to differential expression targets, we adopted the TieDIE framework, which was originally designed to connect genomic events to transcriptional states through network diffusion. Our approach extends TieDIE with three key modifications: (1) cell-state conditioned edge weights, (2) explicit path extraction rather than subnetwork identification, and (3) iterative penalty-based diversification.

Given the weighted graph $\mathcal{G}^{(c)} = (\mathbb{V}, \mathbb{A}, \{w_{uv}^{(c)}\})$, we follow influence propagation by estimating a perturbation-centered flow score over nodes using Personalized PageRank (PPR). Let $\mathcal{S}(p)$ denote the set of start nodes associated with perturbation p:

$$\mathcal{S}(p) = \begin{cases} \{t : t \text{ is a known target of chemical } p\}, & \text{if } p \text{ is chemical,} \\ \{p\}, & \text{if } p \text{ is a single-gene perturbation,} \\ \{p_1, \ldots, p_m\}, & \text{if } p = p_1 + \cdots + p_m \text{ is a gene combination.} \end{cases}$$

We define a personalization vector $\mathbf{s}^{(p)} \in \mathbb{R}^{|\mathbb{V}|}$ by assigning equal mass to valid start nodes:

$$s_v^{(p)} = \begin{cases} \frac{1}{|\mathcal{S}(p)|}, & \text{if } v \in \mathcal{S}(p), \\ 0, & \text{otherwise.} \end{cases}$$

Let $\mathbf{P}^{(c)}$ be the row-normalized transition matrix induced by the edge weights $w_{uv}^{(c)}$. The node flow score $\mathbf{p}^{(c,p)}$ is computed as the fixed point of:

$$\mathbf{p}^{(c,p)} = (1-\alpha)\mathbf{s}^{(p)} + \alpha(\mathbf{P}^{(c)})^\top \mathbf{p}^{(c,p)}$$

where α is the damping factor representing the probability of continuing the random walk. The resulting scores define a flow field quantifying each node's mechanistic relevance to the perturbation.

Next, we extract explicit causal paths from $\mathcal{S}(p)$ to the effect gene g_e using Dijkstra search on a PPR-informed cost function. For each traversable edge (u, v), the search cost is defined as:

$$\text{cost}_{uv}^{(c,p)} = \frac{1}{w_{uv}^{(c)}\, \mathbf{p}_u^{(c,p)}\, \mathbf{p}_v^{(c,p)} + \epsilon_c}$$

where ϵ_c is a small cost constant ensuring numerical stability. This cost formulation ensures that paths traverse high-flow regions identified by network diffusion while respecting cell-state structurally plausible edges. The extracted shortest path $\mathcal{P}^\star$ is the one minimizing cumulative resistance $\mathcal{L}(\mathcal{P}) = \sum \text{cost}_{v_i v_{i+1}}^{(c,p)}$, subject to a maximum hop-depth constraint $D_{\max}$.

To capture the multifaceted nature of perturbation mechanisms—such as compensatory signaling axes—we employ iterative penalty-based diversification. After extracting a path, we apply multiplicative penalties to its visited edges in a temporary graph copy, dividing their weight by a penalty factor λ to encourage the exploration of alternative mechanistic routes. Duplicate paths and paths that are complete supersets of other selected paths are pruned. This procedure yields a focused, perturbation-anchored bag-of-paths representation tailored to the queried response gene.

B.2.4. REFERENCE PATHS VIA EVIDENCE AUGMENTATION AND FILTERING

The primary objective of this stage is to filter the extracted causal paths to eliminate mechanistically invalid reasoning trajectories. However, the paths extracted from OmniPath merely establish structural connectivity and frequently contain edges where the regulatory direction is unknown or functionally ambiguous. To enable the rigorous causal inference required for logical filtering, we must first maximize directional certainty. We achieve this by cross-referencing these unspecified edges with directive literature evidence from INDRA CoGEx. By adopting the majority consensus from the literature, we resolve the ambiguous relations into explicit activating ($+1$) or inhibitory (-1) interactions. This evidence augmentation transforms the candidate subgraph into a functionally signed causal network, laying the necessary groundwork for subsequent validation.

Then, we apply deterministic filtering criteria based on the observed differential expression labels. For samples exhibiting significant differential expression (labels `up` or `down`), a biologically valid path must possess a cumulative polarity that matches the observed outcome. We formalize this using a Sign-Reachability criterion (Algorithm 1). Biologically, a perturbation initiates a specific signal cascade. As this signal propagates through the network, inhibitory edges invert the

Algorithm 1 Sign-Reachability Filter for Differential Expression

Input: Extracted path $\pi = (a_1, \ldots, a_L)$, initial perturbation sign $\text{sign}_0(p_i)$, observed label $\ell_i \in \{\texttt{up}, \texttt{down}\}$, ambiguity budget B_0
Output: Validity of the path (True/False)
$\mathbb{S}_0 \leftarrow \{\text{sign}_0(p_i)\}$
$b_{count} \leftarrow 0$
for $t \leftarrow 1$ **to** L **do**
  if $\text{sign}(a_t) = +1$ **then**
    $\mathbb{S}_t \leftarrow \mathbb{S}_{t-1}$
  else if $\text{sign}(a_t) = -1$ **then**
    $\mathbb{S}_t \leftarrow \{-s : s \in \mathbb{S}_{t-1}\}$
  else
    $\mathbb{S}_t \leftarrow \mathbb{S}_{t-1} \cup \{-s : s \in \mathbb{S}_{t-1}\}$
    $b_{count} \leftarrow b_{count} + 1$
  end if
end for
if $b_{count} > B_0$ **then**
  return False // Path exceeds allowable ambiguity
end if
$target\_sign \leftarrow +1$ if $\ell_i = \texttt{up}$ else -1
return $target\_sign \in \mathbb{S}_L$

polarity. A path is retained only if the terminal signal polarity logically aligns with the observed gene expression shift. Paths that accumulate excessive ambiguity or fundamentally contradict the observed differential state are pruned.

Explaining an `unchanged` transcriptomic response presents a unique biological challenge, as a lack of differential expression does not necessarily imply an absence of causal interaction. It frequently stems from biological boundary conditions, such as ceiling or floor effects in gene transcription. To model this phenomenon, we introduce a Sign-Consistency framework (Algorithm 2) that directly integrates the propagated causal signal with the basal transcriptomic state. Utilizing the hysteresis thresholds τ_{low} and τ_{high} defined in previous sections, we discretize the basal expression of the target gene into distinct functional states: silent, active, or saturated. The algorithm then evaluates whether the incoming perturbation signal can physically alter this state. Crucially, if an activating signal reaches a gene already at its saturation point (τ_{high}), or if an inhibitory signal reaches a basally silent gene (τ_{low}), the mechanistic outcome is correctly classified as `unchanged`. This basal context integration prevents the false rejection of valid causal pathways that are simply masked by physiological constraints.

B.2.5. Path-to-Text Reasoning Synthesis

To systematically translate structured causal graphs into natural language, we implement a multi-stage prompt construction and generation pipeline. This process ensures that the resulting text maintains strict fidelity to the underlying biological mechanisms and properly grounds the language model in the defined cellular context.

For a given cell state c, the generation module requires the perturbation p, the effect gene g_e, the observed differential expression label $y \in \{\text{unchanged}, \text{up}, \text{down}\}$, and the filtered causal paths. To present this graph-structured data to the language model, we serialize the pathways using a flat breadth-first search traversal initialized at the perturbation root p. This topological ordering ensures that upstream regulatory events are structurally prioritized in the prompt before the downstream interactions that converge on g_e.

To provide the language model with an accurate representation of intermediate regulator availability, we discretize the continuous basal expression values of the genes along the selected paths. Leveraging the adaptive hysteresis thresholds τ_{low} and τ_{high} defined previously, we map the basal expression of each relevant gene into explicit qualitative states of low, medium, or high availability. This contextual translation allows the model to correctly deduce whether a specific signaling route is mechanistically viable, attenuated, or entirely insufficient to induce a change in g_e within the initial cellular environment.

Algorithm 2 Sign-Consistency Filter for Unchanged Targets

Input: Extracted path $\pi = (a_1, \ldots, a_L)$, perturbation initial state $z_0 \in \{0, 1\}$, target basal expression x_{y_i,c_i}, thresholds $\tau_{\text{low}}, \tau_{\text{high}}$
Output: Consistency of the path (True/False)
if $x_{y_i,c_i} \leq \tau_{\text{low}}$ **then**
  $b \leftarrow 0$ // Basally silent
else if $x_{y_i,c_i} \geq \tau_{\text{high}}$ **then**
  $b \leftarrow 1$ // Basally saturated
else
  $b \leftarrow \varnothing$ // Active state without boundaries
end if
$\mathbb{Z}_0 \leftarrow \{z_0\}$
Propagate binary activity states
for $t \leftarrow 1$ **to** L **do**
  if $\text{sign}(a_t) = +1$ **then**
    $\mathbb{Z}_t \leftarrow \mathbb{Z}_{t-1}$
  else if $\text{sign}(a_t) = -1$ **then**
    $\mathbb{Z}_t \leftarrow \{1 - z : z \in \mathbb{Z}_{t-1}\}$
  else
    $\mathbb{Z}_t \leftarrow \{0, 1\}$
  end if
end for
Evaluate physiological constraints
if $(b = 1 \wedge 1 \in \mathbb{Z}_L) \vee (b = 0 \wedge 0 \in \mathbb{Z}_L)$ **then**
  return True // Ceiling or floor effect observed
end if
return False

The comprehensive prompt integrates five core components: the cell type, the perturbation p, the effect gene g_e, the expanded text of the observed label y, and the serialized path evidence. We apply specific encoding rules for perturbations to maintain mechanistic integrity. The prompt instructs the model to synthesize a single predictive paragraph that derives y purely as a logical consequence of the provided pathways, actively resolving ambiguous regulatory relations into defined polarities without relying on external or unsupported mechanisms.

Text generation is executed using Qwen3-4B configured with bfloat16 precision. Because language models can occasionally fail to ground their outputs in the provided physical constraints, we enforce a strict post-generation quality control protocol. Paragraphs containing semantic markers of mechanistic failure or refusal, such as explicit statements indicating the path does not support the observed label (e.g., "cannot be explained", "not mechanistically supported", or "cannot be logically explained"), are aggressively filtered out. The surviving high-fidelity reasoning samples are assigned deterministic identifiers based on the unique combination of cell type, perturbation, and effect gene, which are subsequently compiled into the final PERTREASONQA dataset.

B.2.6. HYPERPARAMETERS

In our implementation, the structural base weight w_{uv}^{base} was set to a uniform value of 1.0. The gating steepness parameter β was set to 10. To maintain appropriate baseline traversability, the source floor δ_{src} was configured to 0.05, and the target floor δ_{tgt} was configured to 0.5. The small weight constant $\epsilon_w = 10^{-9}$ was added to the final weight computation to prevent exact zeros.

For the PageRank diffusion process, we utilized a damping parameter $\alpha = 0.85$, capped at a maximum of 50 iterations with a tolerance of 10^{-4}. In the path extraction phase, the edge cost formulation included a small constant $\epsilon_c = 10^{-4}$. The maximum hop-depth constraint $D_{\max}$ was restricted to 4 (which is reduced to 3 internally for chemical perturbations after prepending the drug-target edge). To construct the final Bag-of-Paths, the pipeline iteratively extracted up to $K = 4$ diverse paths, where K is the path budget, penalizing visited edges by a factor of $\lambda = 2.0$.

For all path-to-text reasoning generations, we use Qwen's recommended configurations, specifically, temperature 0.7, top-p sampling 0.8, and top-k sampling 20.

C. Model and Training Details

C.1. SFT Training

We cast supervised fine-tuning as a standard causal language modeling objective. Let x denote the input instruction and y the target output. We minimize the negative log-likelihood over response tokens using a completion-only loss strategy, where only the assistant tokens contribute to the gradient:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{|y|} \log P_\theta(y_t \mid x, y_{<t}) \,,$$

where θ represents the trainable Low-Rank Adaptation (LoRA) parameters.

Outcome-Only Supervision The primary objective of this initial stage is to enable PERTREASONLM to acquire domain knowledge regarding the broad associations among a perturbation p, a cellular context c, an effect gene g_e, and the expression outcome y. Because high-quality knowledge graph triplets for reasoning are scarce for many perturbations and effect gene pairs, outcome-only samples are significantly more abundant within PERTREASONQA. Consequently, this phase acts as a crucial knowledge acquisition step, allowing the model to internalize biological entity representations and perturbation response priors. By predicting the final outcome without requiring a full mechanistic explanation, this stage provides a stable initialization that primes the model for the subsequent, more complex reasoning supervision. In practice, the input sequence is constructed using the cell type, the perturbation p, the effect gene g_e, and the basal expression contexts. The corresponding output consists exclusively of the final differential expression label $y \in \{\text{unchanged, up, down}\}$, without any intermediate reasoning paths or retrieved pathway evidence.

Mechanistic Reasoning Supervision Following the initial knowledge acquisition phase, we train PERTREASONLM to generate Chain-of-Thought (CoT) reasoning in natural language (Wei et al., 2022; Chung et al., 2024; Magister et al., 2023) that elucidates the causal pathways from the perturbation p to the effect gene g_e. This supervision forces the model to align the "broad associations" learned in the previous stage with structured, step-by-step "causal mechanisms". By doing so, we mitigate shortcut predictions and ensure that the final concluded outcome y is explicitly justified by established biological pathway evidence. Crucially, the model learns to jointly interpret retrieved pathway knowledge and the basal cell state c to select the correct context-specific regulatory mechanism. We also employ a data mixing that combines complex reasoning samples with the outcome-only samples (Guo et al., 2025; Mitra et al., 2023). The target output is then formatted as a tag-structured response, including a `<thinking>` block of the mechanistic reasoning, a final `<answer>` block explicitly stating the predicted expression change y, and a `<triplets>` block detailing the used causal edges.

Relation Direction Self-Supervision Despite the integration of multiple KG sources, the precise regulatory direction of certain gene-gene interactions often remains unknown or ambiguous. To address this, the final stage of our SFT pipeline introduces a biologically inspired self-supervised learning objective. We formulate a "relation masking and prediction" task designed to teach the model edge-level mechanistic primitives, enabling it to accurately distinguish between activation and inhibition (Teru et al., 2020a; Zhang et al., 2024). Here, the model is explicitly tasked with inferring the causal sign of ambiguous edges based on the cell and KG contexts. To ensure the model learns genuine context-based disambiguation rather than simply copying localized text, we mask the directionality of a subset of well-defined, clear relations during training. This strategy forces PERTREASONLM to actively use the provided cellular and pathway contexts to deduce the correct causal direction, thereby significantly reinforcing its foundational mechanistic understanding.

Multi-Stage SFT Curriculum The comprehensive SFT process is structured as a progressive curriculum that systematically advances from simpler task adaptations to mechanistically richer forms of supervision. Specifically, the training sequence unfolds across four distinct phases: outcome-only adaptation, outcome-plus-reasoning alignment, relation-direction specialization, and a final multi-objective consolidation. This concluding consolidation stage ensures that the three distinct forms of prior supervision do not manifest as isolated capabilities, but are instead seamlessly integrated into a single, cohesive perturbation reasoning within PERTREASONLM. By carefully ordering these learning objectives, our curriculum

Table 4. Hyperparameters for Supervised Finetuning

Hyperparameter	**Value**
Common Settings	
Precision	BFloat16
Learning Rate	2×10^{-4}
LR Scheduler	Cosine
Warmup Ratio	0.03
Max Sequence Length	4096
Training Technique	Gradient Checkpointing, Completion-only Loss
LoRA Rank (r)	32
LoRA Alpha (α)	64
LoRA Dropout	0.05
Stage-Specific Settings	
Outcome-Only Supervision	
Epoch	1
Effective Batch Size	512
Mechanistic Reasoning Supervision	
Epoch	1
Effective Batch Size	128
Relation Direction Supervision	
Epoch	2
Effective Batch Size	128
Clear Relation Masking Ratio	20%

effectively mitigates the inherent data trade-off in PERTREASONQA, bridging the gap between a large but shallow pool of outcome supervision and a small but high-fidelity set of mechanistic reasoning signals.

Due to the diverse nature of PERTREASONQA, the outcome-only, reasoning, and relation-direction samples exhibit significant disparities in both quantity and sequence length. Simple uniform batching across these distributions leads to objective and memory imbalances. To ensure stable optimization, we employ dynamic batch sampling. We construct homogeneous micro-batches containing samples from exclusively one supervision type and then mix these types at the gradient accumulation block level. This strategy maintains the required effective batch size for each specific objective type. Additionally, we apply oversampling based on the perturbation modality and outcome label to mitigate data skewness.

Training is implemented using the `TRL` library (von Werra et al., 2020) with DeepSpeed ZeRO-2 (Rasley et al., 2020) for memory efficiency on two NVIDIA A100 GPUs (40 GB each) with 128 GB system memory. All SFT stages share a common set of hyperparameters, which, along with stage-specific effective batch sizes, epochs, and masking ratios, are detailed in Table 4.

C.2. GRPO Training

Distributed training loop. Each RL epoch runs as a five-phase workflow: (1) **Prepare** — load JSONL data from default-context and no-path-context pools, apply two-stage oversampling, append `/no_think` to disable the backbone's native thinking mode.(2) **Generate** — shard prompts across nodes and decode K=8 completions per prompt with vLLM; (3) **Rewards** — optionally inject the ground-truth response as the final rollout (GT injection), compute scalar rewards, and normalize into group advantages; (4) **Train** — attach a fresh LoRA adapter to the epoch-start merged checkpoint, compute reference log-probabilities with adapters disabled, and optimize with DeepSpeed ZeRO-2; (5) **Merge** — fold the learned adapter back into the checkpoint for the next epoch.

Our GRPO runs on an 8-node cluster with 16 NVIDIA A100-40GB GPUs in total, and each epoch completes in under 10 hours. This hardware configuration supports the distributed rollout, reward, and DeepSpeed training pipeline used for the official model.

Table 5. GRPO implementation settings.

Component	Setting
Initialization	Pre-merged SFT checkpoint
Training data	default-context and no-path-context JSONL pools
Rollouts per prompt	8 completions
Sampling	Temperature 0.85, top-$p = 0.9$, max completion length 768
Objective	Clipped GRPO with KL coefficient $\beta = 0.02$ against epoch-start reference
Reward	$0.8 \times r_{\text{ans}} + 0.2 \times r_{\text{trp}}$
Optimization	AdamW, lr 5×10^{-6}, bf16, DeepSpeed ZeRO-2, gradient checkpointing
LoRA	Rank 16, alpha 32, dropout 0.05; targets: `q/k/v/o_proj`, `gate/up/down_proj`
Stabilizers	GT injection into last rollout; zero-std rescue with baseline ± 1.0

Advantage normalization and stabilization. Group advantages are computed as

$$A_i = \begin{cases} \dfrac{r_i - \mu_x}{\sigma_x}, & \sigma_x > 0, \\ +1.0, & \sigma_x = 0 \text{ and } \mu_x > 0.5, \\ -1.0, & \sigma_x = 0 \text{ and } \mu_x \leq 0.5. \end{cases}$$

The zero-variance rescue prevents zero-signal updates on groups where all completions receive identical rewards, which is common on structured biological outputs. GT injection into the last rollout further reduces the frequency of such groups.

Hyperparameters. Full GRPO settings are listed in Table 5.

Reward Design The official reward is the answer-and-triplet objective

$$r^{\text{sel}} = 0.8\, r_{\text{ans}} + 0.2\, r_{\text{trp}},$$

where r_{ans} is a binary reward on the extracted answer label and r_{trp} is a partial-match reward on causal triplets. The partial-match rule gives full credit when subject, object, and sign are all correct; reduced credit when the entity pair is correct but the sign is neutral; and smaller credit when the entity pair is correct but the sign is wrong.

D. Baseline Details

D.1. Gene-Space Baselines with Numerical Outputs

- **GEARS** (Roohani et al., 2024): A KG-based model for perturbation prediction that uses a Gene Ontology knowledge graph to predict outcomes for unseen single and combinatorial genetic perturbations.
- **scGPT** (Cui et al., 2024): A foundation model pre-trained on massive single-cell transcriptomic data. We fine-tune scGPT on our dataset to represent the capability of generative gene-expression models.
- **STATE (SE+ST)** (Adduri et al., 2025): A large-scale perturbation foundation model trained directly in gene-expression space on Perturb-seq style data. We use the pre-trained SE model's embedding and fine-tune the ST model.

These baselines are evaluated only on the genetic subset. We train them on genetic perturbations and report scores only on matched genetic test slices; we do not merge chemical and genetic splits in their summary statistics. This is the fairest protocol because GEARS and scGPT consume gene-level inputs and predict continuous perturbed expression profiles, rather than operating over a representation that can natively encode both genes and small molecules.

For STATE, while genetic perturbations are represented via ESM2 protein embeddings, allowing the model to generalize to unseen genes via sequence similarity, its chemical perturbation module typically employs embedding lookup by one-hot encodings. This dependency on fixed vocabularies makes chemical OOD extrapolation technically impossible. Consequently, to ensure a fair comparison focused on generalization, we evaluate these numerical baselines exclusively on the genetic subset of PERTREASONQA.

Continuous-to-label conversion. GEARS, scGPT, and STATE output continuous gene-expression values rather than 3-way labels. To compare them against our benchmark, we convert each predicted perturbation profile into differential-expression calls by pairing the predicted perturbed values with empirical control cells from the same AnnData source and cell type, then running Scanpy's Wilcoxon `rank_genes_groups` test for each (c, p) contrast. We apply Benjamini–Hochberg correction across the genes available in that contrast and assign `up` or `down` only when the adjusted p-value passes the evaluation threshold and the estimated log-fold change has the corresponding sign; all remaining cases are mapped to `unchanged`. If a cell-type-specific control pool is unavailable, we fall back to the corresponding source-level controls. This protocol evaluates the induced 3-way endpoint rather than raw regression error, which is the quantity aligned with our QA benchmark.

D.2. Text-Space Baselines based on LLMs

- **Qwen3-4B (base)** (Yang et al., 2025): The backbone of PERTREASONLM and SUMMER (Wu et al., 2025a), evaluated without any fine-tuning to isolate the gains from our training pipeline.
- **BioMistral 7B** (Labrak et al., 2024) and **NatureLM 8x7B** (Xia et al., 2025): Leading open-source biomedical LLMs continuously pre-trained or instruction-tuned on large-scale biomedical corpora (PubMed, medical guidelines). They represent the upper bound of performance for models relying solely on implicit pre-training knowledge without explicit KG-guided reasoning.
- **SUMMER** (Wu et al., 2025a): A state-of-the-art Retrieval-Augmented Generation baseline. It retrieves relevant literature or KG triplets to answer perturbation queries, serving as a strong RAG reference point. We use Qwen3-4B as a backbone model for SUMMER.